

differently by the causes of death, specific causes may need to be targeted to reduce inequalities.

## References

1. Preston S.H. (1974). Evaluation of Postwar Mortality Projections in the United States, Canada, Australia, New Zealand. *World Health Statistics Report*, 27(1): 719-745.
2. Mackenbach, JP., Looman CWN. Life expectancy and national income in Europe, 1900-2008: an update of Preston's analysis. *International Journal of Epidemiology* 2013;42:1100-1110
3. Mehta, N.K., Abrams, L.R., Myrskylä, M. (2020). US life expectancy stalls due to cardiovascular disease, not drug deaths. *Proceedings of the National Academy of Sciences*, 117(13):6998-7000.
4. Aburto, J. M., van Raalte, A. (2018). Lifespan Dispersion in Times of Life Expectancy Fluctuation: The Case of Central and Eastern Europe. *Demography*, 55: 2071-2096.
5. Vallin, J., Meslé, F. (2009). The Segmented Trend Line of Highest Life Expectancies. *Population and development review*, 35 (1): 159-187.
6. White, Kevin M. 2002. Longevity advances in high-income countries, 1955-1996. *Population and Development Review*, 28: 59-76.
7. Caussinus, H. and Courgeau, D. (2010). Estimating Age without Measuring it: A New Method in Paleodemography. *Population (English Edition)*, 2002, 65(1):117-144.
8. Łukasik, S., Bijak, J., Krenz-Niedbała, M., Liczbinska, G., Sinika, V., and Piontek, J. (2017). Warriors Die Young: Increased Mortality in Early Adulthood of Scythians from Glinoe, Moldova, Fourth through Second Centuries BC. *Journal of Anthropological Research*, 73(4):584-616.
9. Human Mortality Causes of Death Database 2021. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded on 01/01/2021.
10. Oeppen, J., Vaupel, J.W.: Broken Limits to Life Expectancy. *Science* **296** (5570): 1029-1031 (2002)
11. Riley J. *Rising Life Expectancy: A Global History*. Cambridge. Cambridge University Press (2001)
12. Shkolnikov V.M., Andreev E.M. Tursun-zade R., Leon D.A. (2019). Patterns in the relationship between life expectancy and gross domestic product in Russia in 2005-15: a cross-sectional analysis. *The Lancet Public Health*.
13. World Bank. GDP (current US).2021. <https://data.worldbank.org/>

# Locally sparse functional regression with an application to mortality data

## *Regressione funzionale localmente sparsa con un'applicazione a dati di mortalità*

Mauro Bernardi, Antonio Canale, Marco Stefanucci

**Abstract** A novel method for functional regression with functional response and functional covariate is discussed. The method is particularly useful when the regression surface is non-zero only on a subset of its bivariate domain, allowing for a local relation between the response and predictor variable. By means of a tensor product splines representation of the unknown functional coefficient and an overlap group lasso penalty we are able to effectively estimate the regression function. The model performance is illustrated through its application to the well-known Swedish Mortality dataset, clearly showing the local nature of the relation between the mortality at consecutive years.

**Abstract** Viene discusso un nuovo metodo per implementare la regressione funzionale penalizzata per una risposta funzionale ed una covariata funzionale. Il metodo è particolarmente utile quando la superficie di regressione è diversa da zero solo in un sottoinsieme del suo dominio bivariato, permettendo una relazione locale tra la variabile risposta e la variabile di previsione. Grazie ad una rappresentazione del coefficiente funzionale incognito con splines prodotto-tensoriali e ad una penalità lasso a gruppi sovrapposti siamo in grado di stimare efficacemente la funzione di regressione. Tale procedura di stima viene quindi applicata al noto set di dati Swedish Mortality, mostrando chiaramente la natura locale della relazione tra la mortalità in anni consecutivi.

**Key words:** Functional Data Analysis; Function-on-function regression; Sparsity; Swedish Mortality

---

Mauro Bernardi, Antonio Canale, Marco Stefanucci,  
mbernardi@stat.unipd.it, canale@stat.unipd.it, stefanucci@stat.unipd.it,  
Università di Padova, Via Cesare Battisti, 241, 35121 Padova

## 1 Introduction

Functional linear model investigates the indirect relation between two or more variables, where at least one is functional in nature. The function-on-scalar regression model focuses scalar response variable and functional explanatory variables [1]. The converse applies to the scalar-on-function model [3]. In this short paper we study the function-on-function regression model, i.e.

$$y(t) = \int \beta(s,t)x(s)ds + \varepsilon(t), \quad (1)$$

where both the response and explanatory variable are defined over a continuum,  $s \in \mathcal{S}, t \in \mathcal{T}$  and  $\varepsilon(t)$  is functional noise. This framework has increased in popularity thanks to the famous book by Ramsay and Silverman [5], where a penalized estimation approach based on B-splines is discussed. Real data applications involving the function-on-function regression model include —but are not limited to— chemometrics, pharmacology, neuroscience, demography and meteorology. In the aforementioned book, Ramsay and Silverman discussed the interesting special case of the Historical Functional Model (HFM) where the influence of the explanatory variable  $x(s)$  on the response  $y(t)$  is confined to a specific interval of the domain, namely  $s < t$ , and the resulting model is

$$y(t) = \int_{s < t} \beta(s,t)x(s)ds + \varepsilon(t).$$

With this formulation only values of  $s$  that are lower than  $t$  are used in the prediction of  $y(t)$  and the estimated  $\beta(s,t)$  is an upper-triangular matrix. Considering the well-known Canadian Weather dataset discussed in [5] as illustrative application,  $x(s)$  is the temperature observed from January 1st to December 31st and  $y(t)$  is the level of precipitation observed in the same period ( $\mathcal{S} = \mathcal{T}$ ). Of course only the temperature *before* the time  $t$  can be predictive of the precipitation at time  $t$ , and the HFM results in an elegant way to force this constraint. In this application the interval of integration is motivated by the phenomenon under study and hence is chosen by the user. However this aspect poses some limitations:

- often the choice of the interval of integration is not obvious. Beyond simple cases, the user could be unable to make this crucial choice.
- the choice is not data-dependent. Specifically, integration on a restricted domain acts like a model selection, similarly to an *a priori* subset selection not derived by using the data.

From these reasons we propose a methodology able to automatically detect the regions of sparsity in the unknown operator, without any restriction on the domain of integration. The method is based on a tensor product splines representation of the regression function and a penalized approach that makes use of a modified lasso penalty [4]: the Overlap Group Lasso (OGL).

## 2 Model

We define  $\{\eta_j(s)\}_{j=1}^L$  and  $\{\theta_k(t)\}_{k=1}^M$  as B-splines bases over  $\mathcal{S}$  and  $\mathcal{T}$ . Then we have  $x_i(s) = \sum_{j=1}^L a_{ij} \eta_j(s)$  and  $y_i(t) = \sum_{k=1}^M c_{ik} \theta_k(t)$  for each  $i = 1, \dots, n$ . The bivariate function  $\beta(s, t)$  can be expressed as  $\beta(s, t) = \sum_{j=1}^L \sum_{k=1}^M b_{jk} \eta_j(s) \theta_k(t)$ . If  $L$  and  $M$ , the dimensions of the two spline bases, are small enough, model (1) can be fitted computing the

$$\operatorname{argmin}_{\mathbf{B}} \|\mathbf{C}\boldsymbol{\Theta} - \mathbf{A}\mathbf{N}^T \mathbf{B}\boldsymbol{\Theta}\|_F^2. \quad (2)$$

where  $\boldsymbol{\Theta}$  and  $\mathbf{N}$  are matrices of basis functions and  $\mathbf{A}, \mathbf{B}, \mathbf{C}$  are matrices of coefficients. However, a penalized estimation is often preferable. Consistently with this the dimensions of the spline bases are much higher and to regularize the solution a penalty is added to (2). A popular choice for the penalty term is the *ridge* penalty defined as  $\int \beta^2(s, t) ds dt$ . In this case the coefficient function can be obtained by

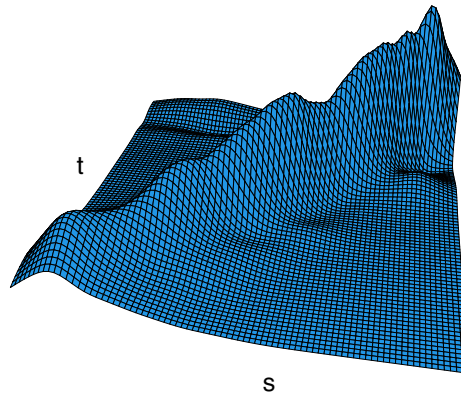
$$\operatorname{argmin}_{\mathbf{B}} \|\mathbf{C}\boldsymbol{\Theta} - \mathbf{A}\mathbf{N}^T \mathbf{B}\boldsymbol{\Theta}\|_F^2 + \lambda \operatorname{pen}(\mathbf{B}), \quad (3)$$

where the parameter  $\lambda$  controls the balance between the two terms. The main limitation of this particular penalty function and, in general, of all smooth penalties, is that they are not suited for identify sparsity and the  $\hat{\beta}(s, t)$  obtained in (2) and (3) will result in a overall smooth function. A possible way to detect sparsity is the use of the lasso penalty [7] on the coefficients of the expansion of  $\beta(s, t)$  but this comes at a price: sacrifice smoothness. The main idea behind our locally sparse model is to borrow strength from the two approaches in order to construct an estimator able to identify both the smooth and sparse regions of the unknown function.

We achieve this exploiting the OGL penalty and a feature of the B-splines representation of  $\beta(s, t)$  that allows  $\beta(s, t)$  to be exactly zero on a region of its domain if a block of adjacent coefficients  $b_{jk}$  of suitable dimension depending on the B-spline order is jointly zero.

## 3 Application

We apply the method described in previous section to the well-known Swedish Mortality dataset, where log-hazard rates are observed for the Swedish female population between years 1751 and 1894. The aim is to predict the log-hazard function  $y_i(t)$  at a specific calendar year  $i$  by using the log-hazard function at previous year  $x_i(s) = y_{i-1}(t)$ . Existing studies [6, 2] show that the hazard function at year  $i$  and age  $t$  is mainly influenced by the hazard function at the previous year  $i - 1$  at age  $s = t - 1$ , resembling a quasi-concurrent relation, following the terminology of [5]. However, none of these studies report the total absence of relation when  $s$  and  $t$  are far away. Figure 2 shows the  $\hat{\beta}(s, t)$  obtained by the proposed methodology. It



**Fig. 1** Estimated  $\beta(s, t)$  with the proposed locally sparse approach for the Swedish mortality database

is easy to see that smoothness is preserved along the main diagonal and that the function is flat when  $s$  and  $t$  are far enough.

## References

1. Cardot, H., Ferraty, F., Sarda, P. (1999), Functional linear model. *Statistics and Probability Letters* **45**, 11–22.
2. Chiou, J.M. and Muller, H.G. (2009), Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association* **104** (486), 572–585.
3. Faraway, J.J. (1997), Regression analysis for a functional response. *Technometrics* **39**, 254–261.
4. Jenatton, R., Audibert, J.-Y., and Bach, F. (2011), Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, **12**, 2777–2824.
5. Ramsay, J.O. and Silverman B.W. (2005), *Functional Data Analysis*, 2nd edition. Springer, New York
6. Ramsay, J.O., Hooker, G., and Graves, S. (2009), *Functional data analysis with R and MATLAB*. Springer Science & Business Media.

Locally sparse functional regression with an application to mortality data

7. Tibshirani, R. (1996), Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267–288.
8. Wu, T.T. and Lange, K. (2010), The MM alternative to EM. *Statistical Science*, **25**, 492–505.

## 4.12 Environmental statistics

# A Distribution-Free Approach for Detecting Radioxenon Anomalous Concentrations

## *Un Approccio Distribution-Free per Individuare Concentrazioni Anomale di Radioxeno*

M. Scagliarini, R. Gualdi<sup>2</sup>, G. Ottaviano<sup>3</sup>, A. Rizzo<sup>4</sup> and F. Padoani<sup>5</sup>

**Abstract** The detection of anomalous radioxenon atmospheric concentrations plays a key role for revealing both underground nuclear explosions and radioactive emissions from nuclear power plants and medical isotope production facilities. For this purpose, the method currently used by the International Data Center of the CTBTO is based on descriptive thresholds. In this work we propose a statistical inference-based method, that allows to distinguish between the typical background of atmospheric radioxenon and anomalous values above background. We used a non-parametric methodology that does not require any assumption on the phenomenon distribution. In such a way we overcome the problem due to the non-normality of the radioxenon data.

**Abstract** *L'identificazione di concentrazioni anomale di radioxeno nell'atmosfera è fondamentale per rilevare sia esplosioni nucleari sotterranee sia rilasci di radioattività da impianti nucleari o impianti di produzione di radiofarmaci. Il metodo attualmente usato dall' International Data Center è basato su soglie descrittive. In questo lavoro l'obiettivo è proporre un metodo, basato su criteri inferenziali, che consenta di distinguere tra la radioattività di fondo e concentrazioni anomale. La scelta di un metodo non parametrico ha consentito di superare il problema legato alla non normalità distributiva dei dati a disposizione.*

**Key words:** abnormal concentration; permutation; statistical monitoring algorithms; radioactivity; statistical process control

---

<sup>1</sup> Michele Scagliarini, Department of Statistical Sciences, University of Bologna; e-mail: michele.scagliarini@unibo.it

<sup>2</sup> Rosanna Gualdi, Department of Statistical Sciences, University of Bologna; e-mail: rosanna.gualdi@studio.unibo.it

<sup>3</sup> Giuseppe Ottaviano, ENEA, Bologna Research Centre; e-mail: giuseppe.ottaviano@enea.it

<sup>4</sup> Antonietta Rizzo, ENEA, Bologna Research Centre; e-mail: antonietta.rizzo@enea.it

<sup>5</sup> Franca Padoani, ENEA, Bologna Research Centre, e-mail: franca.padoani@enea.it

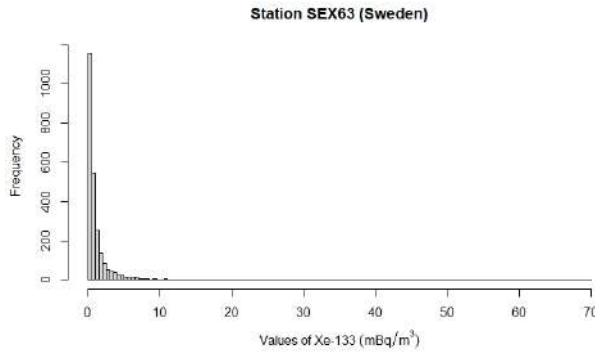


## 1 Introduction and Background

The “Comprehensive Nuclear-Test-Ban Treaty Organization” (CTBTO) supports and promotes the “Entry Into Force” (EIF) of the “Comprehensive Nuclear Test-Ban Treaty”, a treaty that outlaws nuclear test explosions. The core of the CTBTO is a verification regime based on three pillars, the International Monitoring System (IMS), the International Data Centre (IDC) and the On-Site Inspections (OSI). The IMS, when completed, will consist of 337 stations worldwide to monitor the planet for signs of nuclear explosions. Around 90 percent of the facilities are already up and running. Eighty-eight fission and activation products are considered relevant by CTBTO in order to reveal a nuclear explosion. Among these, four radioisotopes of Xenon noble gas (radioxenon:  $^{131m}\text{Xe}$ ,  $^{133m}\text{Xe}$ ,  $^{133}\text{Xe}$ ,  $^{135}\text{Xe}$ ) are the only fission products that allow to reveal an underground nuclear explosion (UNE), because they are the only fission products able to migrate from the subsurface to the atmosphere and therefore to be detected from the IMS stations. However, their signatures can be difficult to recognise due to the anthropogenic releases of radioxenon into the atmosphere from nuclear power plants (NPPs) and from medical isotope production facilities (MIPFs), as known as civil sources. Making conclusions about a suspected underground nuclear explosion or radioactive emissions from civil sources, considering measurements of radioxenon, may prove to be a very complex task due to several factors [3]. One of the objectives of the IDC of the CTBTO is to analyse and categorize the radioxenon observations and therefore a classification system has been defined in which a measured activity concentration is classified as anomalous if it is above the “abnormal activity concentration” value. The abnormal activity concentration value, or “abnormal limit”, used by the IDC, is defined as:  $L_A = Q_2 + \lambda_A [Q_3 - Q_1]$  where  $Q_1$  is the first quartile,  $Q_2$  is the median,  $Q_3$  the third quartile and  $\lambda_A=3$ . The radioxenon concentrations are measured daily and the  $L_A$  value is constantly updated after each observation since the last 365 observations are used for computing the quartiles. Summarising, an activity concentration is classified as anomalous if it is above the “typical background” of the monitoring station established by applying an inter-quartile filter, updated after each observation, to the data. The interquartile filter method in use by the IDC is essentially based on descriptive thresholds and it is the only procedure used to classify the radioxenon concentrations. Therefore, it could be suitably integrated with a method based on statistical inference. In this paper the objectives are: to propose an inference-based method that allows to distinguish between the typical background and anomalous values; to discuss the advantages offered by the joint use of the two methods. Within this framework we focus our attention on statistical monitoring algorithms, such as control charts, since they are able to distinguish between common causes of variation, resulting in a steady background random distribution, and assignable causes of variation, resulting in shifts in process location and/or scale.

## 2 The Data and the Methodology

The data considered in this study consist of the  $^{133}\text{Xe}$  activity concentrations ( $\text{mBq}/\text{m}^3$ ) measured from December 2013 to August 2018 at the IMS station SEX63 (Stockholm-Sweden). Figure 1 shows the histogram of the 2641 valid data at our disposal. Due to the strong asymmetry of the data distribution we appropriately adopted a non-parametric approach. In greater detail, we used a distribution-free control chart, based on recursive segmentation and permutation RS/P [1]. The most relevant methodological aspects of the RS/P procedure are reported in what follow, while for further details we refer to [1]. Let us denote with  $x_{ij}$  the  $j$ -th observation ( $j=1,2,\dots,n$ ) from the  $i$ -th subgroup ( $i=1,2,\dots,m$ ). Note that the case of individual observations can be managed by setting  $n=1$ . When the phenomenon is undisturbed, or in-control, observations  $x_{ij}$  are assumed independent with an unknown but common distribution function  $F_0(x)$ . We assume that the out-of-control case (the scenario of a disturbed radioactivity background) can be described by the following model:  $x_{ij} \sim F_0(x)$  if  $0 < i \leq \tau_1$ ,  $x_{ij} \sim F_1(x)$  if  $\tau_1 < i \leq \tau_2, \dots$ ,  $x_{ij} \sim F_k(x)$  if  $\tau_k < i \leq m$ , where  $0 < i < \tau_1 < \tau_2 < \dots < \tau_k < m$  denote  $k$  unknown change points and  $F_r(\cdot)$ ,  $r=0,1,\dots,k$ , are unknown distribution functions.



**Figure 1:** Histogram of the  $^{133}\text{Xe}$  activity concentrations

Within this framework we are interested in testing  $H_0 : k = 0$  the process was in-control, or in other words no change occurred in the monitored phenomenon, versus  $H_1 : k > 0$  the process was out-of-control, or in other words at least one change occurred in the “typical background” of the  $^{133}\text{Xe}$  activity concentrations. The methodology consists in computing a set of test statistics designed for detecting  $1,2,\dots,K$  shifts, where  $K$  denotes the maximum number of change points for which we want to search. In case of individual data the RS/P method allows to detect step shifts in the process location by computing the statistics  $T_k$ ,  $k=1,2,\dots,K$ , that are designed for testing  $H_0$  versus and  $H_{1,k} : E(x_{ij}) = \mu_0$  if  $0 < i \leq \tau_1$ ;  $E(x_{ij}) = \mu_1$  if

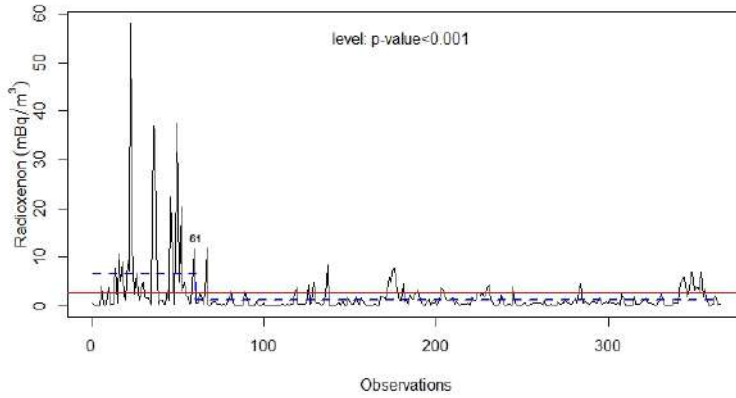
$\tau_1 < i \leq \tau_2; \dots; E(x_{ij}) = \mu_k$  if  $\tau_k < i \leq m$ . The statistics  $T_k$  and the possible change points are computed using a forward recursive-segmentation approach. The algorithm starts with  $k=0$  and proceeds in  $K$  successive stages. At the beginning of stage  $k$ , the interval  $[1, m]$  is partitioned into  $k$  subintervals, each having a length greater or equal to  $L_{MIN}$ . The quantities  $K$  and  $L_{MIN}$  can be appropriately chosen by the researcher [1]. At stage  $k$ , one of these subintervals is split, adding a new potential change point. The new change point is selected maximizing the function  $f_k = \sum_{i=1}^{k+1} (\hat{\tau}_i - \hat{\tau}_{i-1}) (\bar{x}(\hat{\tau}_{i-1}, \hat{\tau}_i) - \bar{\bar{x}})^2$  conditionally on the results of the previous stages and the control statistic  $T_k$  is equal to the attained maximum value of function  $f_k$ . Here,  $0 = \hat{\tau}_0 < \hat{\tau}_1 < \dots < \hat{\tau}_k < \tau_{k+1} = m$  are the boundaries of the new partition,  $\bar{x}(a, b) = \sum_{i=a+1}^b \bar{x}_i / (b-a)$ ,  $\bar{x}_i = \sum_{j=1}^n x_{ij} / n$  and  $\bar{\bar{x}} = \sum_{i=1}^m \bar{x}_i / m$ . Then, these statistics are standardized and aggregated to obtain an overall statistic and the  $p$ -value is computed using a permutation approach [1,4]. To detect shifts in the variability of the monitored phenomenon the function  $f_k$  used for the recursive segmentation is defined as  $f_k = \sum_{i=1}^{k+1} (\hat{\tau}_i - \hat{\tau}_{i-1}) \log(s^2/s^2(\hat{\tau}_{i-1}, \hat{\tau}_i))$  where  $s^2(a, b) = \sum_{i=a+1}^b s_i^2 / (b-a)$ ,  $s_i^2 = \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 / n$  and  $s^2 = \sum_{i=1}^m s_i^2 / m$ . The other steps of the procedure remain unchanged [1].

### 3 Results and Concluding Remarks

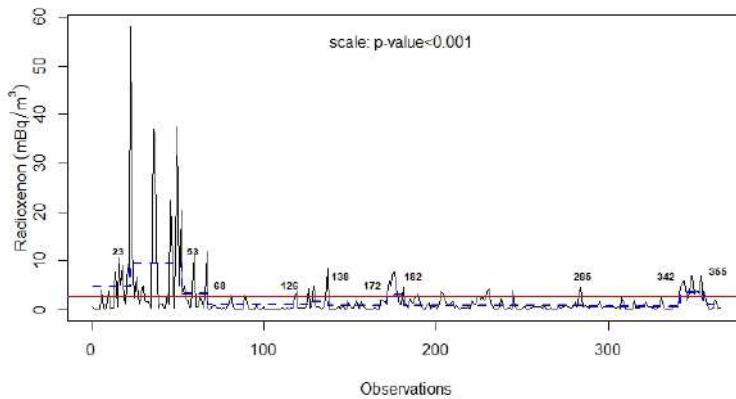
To be consistent with the inter-quartile filter method we applied the RS/P methodology to the 365 valid observations available from 19/04/17 to 26/01/18. We used the *R* package *dfphase1* [2] and Figure 2 shows the RS/P results for testing the stability over time of the  $^{133}\text{Xe}$  mean level. In Figure 2 are shown the original observations (continuous line), the abnormal limit computed by the IDC (the red continuous line at  $2.59 \text{ mBq/m}^3$ ), an estimate of the possibly time-varying process means (dashed line) and the estimated change points. The  $p$ -value suggests to reject the hypothesis of stability of the  $^{133}\text{Xe}$  location and in the case in question one change point is estimated at observation 61. Figure 3 illustrates the application of the RS/P procedure for monitoring the variability (scale) of the  $^{133}\text{Xe}$  activity concentrations. In Figure 3 are displayed the original observations (continuous line), the abnormal limit (red line), an estimate of the possible time-varying process variability [1] and the estimated change points. Also, in this case we reject the hypothesis of stability over time of the  $^{133}\text{Xe}$  variability and ten possible change points are estimated at observations 23, 53, 68, 126, 138, 172, 182, 285, 342 and 355, respectively. Let us now discuss the overall results in order to evaluate the appropriateness of the RS/P methods to detect anomalous values. To do this we base

A distribution-free approach for detecting radioxenon anomalous concentrations

our considerations also looking results of at the interquartile-filter used by the IDC. The results indicate that the RS/P method detects possible changes in mean level and variability often in correspondence of observations that exceed the abnormal limit threshold.



**Figure 2:**  $^{133}\text{Xe}$  activity concentrations and results of RS/P method for detecting level shifts



**Figure 3:**  $^{133}\text{Xe}$  activity concentrations and results of RS/P method for detecting scale shifts

It can be noted that there is an initial period, up to observation 68, in which one shift in the mean and three changes in the variability of the  $^{133}\text{Xe}$  activity concentrations are estimated. In the same period, the number of observations above the  $L_A$  threshold were twenty-nine. By considering the data from the 126-th to the 182-th observation, the RS/P method estimates four possible shifts in the variability. In the same period ten exceedances of the abnormal limit were observed (observations 126, 129, 137, 172, 173, 174, 175, 176, 177, 181). The same kind of considerations can be made for the data ranging from the 285-th to the 355-th observation, where four changes in

variability are estimated. In the same period fourteen concentrations exceeded the  $L_A$  threshold (observations 285, 342-345, 347-355). However, for the sake of completeness it is important to note that there are thirteen values of  $^{133}\text{Xe}$  above the threshold  $L_A$  that do not belong to time periods in which shifts in location or scale are detected. More precisely there are six single values (observations 81, 89, 190, 227, 245 and 283) and seven grouped values (observations 118-119, 203-205 and 230-231). The single values above  $L_A$  most likely were caused by random oscillations of the phenomenon, since emissions due to UNE or civil sources should lead to sequences of high  $^{133}\text{Xe}$  activity concentrations. As far as the remaining seven grouped observations is concerned no reliable conclusions can be drawn. Summarizing, the RS/P method provides outcomes with an appreciable level of agreement with the interquartile-filter, but with the advantage of associating a significance level to the results. In such a way, it provides an objective criterion with which to distinguish between random oscillations of the phenomenon and significant shifts in mean and/or in variability. The joint use of the interquartile-filter and the RS/P method might provide an important contribution to the detection of radionuclide anomalous values. This research is the result of a collaboration between the Department of Statistical Sciences, University of Bologna and the National Data Centre - Radionuclides (NDC-RN) of the ENEA, Bologna Research Centre. The data used in this work were obtained within the framework of a two years' contract called "virtual Data Exploitation Centre" (vDEC) signed with the Preparatory Commission of the CTBTO. The views expressed herein are those of the authors and not necessarily reflect the views of the CTBTO Preparatory Commission.

## References

1. Capizzi, G., Masarotto, G.: Phase I Distribution-Free Analysis of Univariate Data. *J. Qual. Technol.* **45**, 273-284 (2013)
2. Capizzi, G., Masarotto, G.: *dfphase1*: Phase I Control Charts (with Emphasis on Distribution-Free Methods). R package version 1.1.1. <https://CRAN.R-project.org/package=dfphase1> (2017)
3. Kalinowski, M., Axelsson, A., Bean, M., Blanchard, X., Bowyer, T., Brachet, G., Hebel, S. McIntyre, J., Peters, J., Pistner, C., Raith, M., Ringbom, A., Saey, P., Schlosser, C., Stocki, T., Taffary, T., and Ungiar, K.: Discrimination of Nuclear Explosions against Civilian Sources Based on Atmospheric Xenon Isotopic Activity Ratios. *Pure Appl. Geophys.* **167**, 517-539, (2010).
4. Pesarin, F.: *Multivariate Permutation Tests: With Applications in Biostatistic*. Wiley, New York (2001)

# **Ecosud Car, a novel approach for the predictive control of the territory**

## ***Ecosud Car, un approccio innovativo per il controllo predittivo del territorio***

Giacomo Iula<sup>1</sup>, Massimo Dimo<sup>1</sup>, Saverio Gianluca Crisafulli<sup>2</sup>, Marco Vito Calciano<sup>1</sup>, Vito Santarcangelo<sup>2</sup>, Massimiliano Giacalone<sup>3</sup>

**Abstract** The paper shows the path of research and development conducted by the Lucanian company Ecosud in the year 2019 and 2020, aimed at the conception of new processes of territorial monitoring, in a CSR key, through the intelligent use of its corporate fleet, declined in the form of Ecosud Car, and thanks to the active collaboration of institutions and citizens as environmental ambassadors. The work therefore introduces the Ecosud company and its corporate mission up to the exposition of the monitoring process of the territory connected to the new technique at the base of the Ecosud Car in line with the areas of action of ISO 26000. Finally, the fuzzy results achieved are presented.

**Abstract** *Il paper illustra il percorso di ricerca e sviluppo condotto dalla società lucana Ecosud nell'anno 2019 e 2020, mirato alla ideazione di nuovi processi di monitoraggio del territorio, in chiave CSR, mediante l'utilizzo intelligente della propria flotta aziendale, declinata nella forma di Ecosud Car, e grazie alla collaborazione attiva di istituzioni e cittadini in qualità di ambasciatori ambientali. Il lavoro introduce dunque l'azienda Ecosud e la relativa mission aziendale fino ad arrivare all'esposizione del processo di monitoraggio del territorio collegato alla nuova tecnica alla base della Ecosud Car in linea con le aree di azioni della ISO 26000. Infine vengono esposti i risultati "fuzzy" ottenuti.*

**Key words:** ecosud car, intelligent cartography, environmental responsibility, green fleet, ISO 26000

---

<sup>1</sup> G. Iula, M. Dimo, M. V. Calciano, Ecosud Srl, Pisticci (Italy), massimo.dimo@ecosud.net:

<sup>2</sup> V. Santarcangelo, S.G. Crisafulli, iInformatica Srl, Trapani (Italy); vito@iinformatica.it;

<sup>3</sup> M. Giacalone, Università degli Studi Federico II, Napoli (Italy); massimiliano.giacalone@unina.it

## **1 Introduction**

The innovative mission that characterizes the Lucanian company Ecosud S.r.l. has led to an evolution of the continuous and predictive monitoring processes of the territory through the design of a cartographic system of real-time environmental monitoring thanks to the use of the innovative fleet of Ecosud Car and the analysis carried out through HMI by environmental ambassadors (employees, stakeholders and volunteers of the territory), connected to the new technique object of the patents of invention of the company in the last two years.

Ecosud S.r.l. was founded in 1984 as a chemical/physical analysis laboratory and as a study of design and management of purification plants, gradually changing its physiognomy over the years, orienting its services to the growing demands of the environmental market. In 1998 Giacomo Iula decided to invest in an activity oriented to the protection of the environment and which would allow the growth of young local professionals, stimulate the growth of the territory and add commercial value to the other companies he owned.

Today, Ecosud S.r.l. mainly addresses its activity to the ecological sector and to the industrial technical services, providing support to the companies or bodies that request it with the aim of reconciling the need for industrial development with the protection of the environmental heritage in accordance with the regulations in force, offering high quality design and environmental consulting services, such as geognostic and geotechnical investigations, chemical-physical/bacterial analyses, environmental site investigations and related emergency safety measures, geological consultancy, air quality monitoring, microclimatic investigations and quality of working environments, risk analysis, characterization and remediation of contaminated sites pursuant to D. lgs. 152/06 (ex D.M. 471/99).

## **2 Mission, CSR and ISO 26000**

The mission characterizing the company Ecosud S.r.l., described by the values of its ethical code, focuses on the problematic of environmental monitoring. In this regard, research activities have focused on the objective of improving monitoring processes through the use of data of the territory related to geological information collected by Ecosud Car, which allow a predictive simulation of the territory, with the active collaboration of institutions and citizens (environmental ambassadors) in view of social responsibility and ISO 26000.

Corporate Social Responsibility (CSR) is defined as the sense of responsibility that a company or any other business entity demonstrates towards the community and the environment, understood both as the natural and geographical environment and as the social context in which it operates, and has become so relevant that it is now considered a priority in business strategies, with the primary objective of serving customers well, taking care of employees, treating suppliers fairly and

Ecosud Car, a novel approach for the predictive control of the territory contributing effectively to the welfare of society and the preservation of the planet, thus creating a shared value [1].

The concept of corporate social responsibility has actually changed over time, because if in the beginning CSR projects and initiatives were unrealistic or the result of some form of self-regulation in the sector, over time there have been regulatory obligations and impositions that have pushed companies towards a greater and necessary social compliance. It is difficult to summarize these obligations and regulatory impositions in a unitary way because, they vary greatly from country to country, also in consideration of market conditions and aspects linked to corporate culture. Sometimes it is the company's mission itself, as in the case of Ecosud S.r.l., that implies its involvement in themes and issues of a certain social relevance but, depending on the business involved, social responsibility operations can be inherent in the same company processes and assets. Other times, however, it is only in order to have a good impact on one's stakeholders that one chooses this path. In all cases, a good corporate social responsibility plan is always an important business card for the company, since it can give a real competitive advantage and is truly appreciated by customers, increasing brand reputation.

UNI ISO 26000 was published on 1 November 2010 with the aim of filling a gap in the ISO system in the field of Corporate Social Responsibility of Organizations; it is in fact the result of the broadest international consensus reached among experts and key stakeholders and was developed with the intention of encouraging the adoption of best practices in the field of social responsibility worldwide [2].

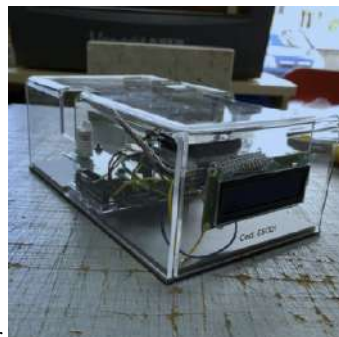
It should be pointed out that ISO 26000 is a guideline and not a standard, therefore it will not be certifiable by a third party on the model of quality, environment health and safety management systems. A company or organization wishing to adopt these new guidelines cannot rely on an external company to certify its commitment to social responsibility, but must consult with its stakeholders, first and foremost the trade unions in terms of working relationships and conditions, so that they can assess whether or not they are complying with the contents of ISO 26000. The standard provides a new definition of social responsibility, "responsibility by an organization for the impacts of its decisions and activities on society and the environment, through ethical and transparent behavior that: contributes to sustainable development, including the health and well-being of society; takes into account the expectations/interests of stakeholders; is in compliance with applicable law and consistent with international standards of behavior; and is integrated throughout the organization and put into practice in its relationships". The action areas of ISO 26000 have been well implemented by the research and development journey conducted by the company [3].

### **3 Ecosud Car prototype and data results**

The Ecosud Car, a prototype realized during the research and development path undertaken by the company, is nothing but a car of the company fleet characterized



by an embedded device that allows to carry out a predictive outdoor monitoring activity of the territory. The device, together with the relative method, are the subject of an industrial patent.



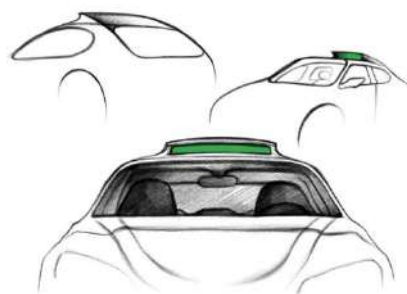
**Figure 1:** First embedded prototype of the Ecosud Car



**Figure 2:** HMI box with example data

The embedded device is represented by a box installed on the company's fleet of cars, the Ecosud Car in fact, which to all intents and purposes represents a mobile detection station that allows precise and objective monitoring of the territory in environmental terms [6].

Ecosud Car, a novel approach for the predictive control of the territory



**Figure 3:** Ecosud Car with its monitoring box

The device embedded in the body of the company's vehicle is controlled by means of an intelligent mirror, which records all environmental parameters and related information acquired by the embedded. The embedded device is equipped with appropriate sensors, able to detect various environmental parameters, such as temperature, humidity, environmental pollution (presence or absence of harmful gases in the area), acoustic, electromagnetic and radioactive pollution, associating to the measurements made the relevant GPS coordinates and consequently creating a dynamic fuzzy cartography of all the places crossed by the fleet every day in relation to the environmental analysis carried out to provide a daily dynamic environmental feedback in view of social responsibility. In addition to the environmental parameters, the device allows to detect and monitor, through visual and photographic feedback, the route taken by the fleet, in order to perform a cartographic environmental monitoring, monitoring that will allow to simulate the evolution of polluted areas [5].



**Figure 4:** Ecosud Car smart mirror in action

The parameters detected by the fleet are flanked by those obtained from the measurements of Ecosud employees or even citizens themselves; in fact, an app has been created through which users can manually carry out environmental cartographic monitoring. After logging in, the user will be automatically geolocalized and will be able to choose whether to leave an evaluation on the degree of perceived pollution, by means of a questionnaire, or whether to make a report on possible problems encountered. The ratings left by users will be visible as dots on the map and can be green (positive), yellow (average) or red (polluted area). In this way, users become to all intents and purposes "environmental ambassadors".

Considering the dataset collected during the year 2020 by Ecosud employers and stakeholders (environmental ambassadors) the fuzzy feedbacks are distributed as 80% for green, 6,7% for yellow and 13,3% from red (considering Basilicata and

Calabria). The text mining semantic analysis applied on textual feedback is focused mainly on the concept “bad smells” (related to amines) and “plastic waste”. All data received by the company fleet and ambassadors are stored on servers and temporally certified in blockchain [4]. The data collected by Ecosud Car, combined with the data collected by the environmental ambassadors, together with the environmental remediation data managed by the company and obtainable from entities, through reports obtained through OSINT, are weighted with appropriate TOPSIS ranking in fuzzy optics [7], supervised by technicians and incorporated into the cloud data warehouse in order to have the most precise and accurate information of the monitored environment, planning consequently the activities to be carried out, with the final objective of acquiring as much data and information as possible, thus obtaining a quality map of the surrounding environment in which fleet and ambassadors operate in synergy and putting it at the service of the company and the territory. This map related semantically to the stratigraphic information of the territory and the typology of pollutant detected allows to carry out a predictive simulation of the scenario useful to determine a priority of attention and intervention.

#### 4 Conclusion and future development

The research and development project has therefore opened the door to a new process/service/product approach aimed at carrying out environmental monitoring from a social responsibility point of view [8], allowing with low costs to transform any company fleet or means of public or private transport in a green point of view as a possible monitoring system from a social responsibility point of view, which we hope could become a valid collective contribution to the growth of our territory.

#### References

1. Maggiolini, Piercarlo. "ISO 26000: una guida alla responsabilità sociale." (2012): 163-192.
2. Serrano, Maryelis Montero. "La responsabilidad social y la norma ISO 26000." *Revista de Formación Gerencial* 11.1 (2012): 102-119.
3. Moratis, Lars, and Timo Cochius. *ISO 26000: The business guide to the new standard on social responsibility*. Routledge, 2017.
4. M.Dimo, B.Iula (2019), Sistema fuzzy e relativo metodo per il monitoraggio predittivo del territorio, UIBM
5. M.Dimo, B.Iula (2020), Dispositivo dinamico e relativo metodo per la ispezione cartografica e per il trasferimento di know-how, UIBM
6. M.Dimo, B.Iula (2020), Box monitora ambiente per automobili, UIBM
7. V. Santarcangelo (2020), Combination of AHP and TOPSIS methods for the ranking of information security controls to overcome its obstructions under fuzzy environment. *J. Intell. Fuzzy Syst.* 38(5): 6075-6088 (2020)
8. M.Giacalone, V.Santarcangelo (2021), Big data for corporate social responsibility: blockchain use in Gioia del Colle DOP, *Quality & Quantity*, Springer

# Effect of ties on the empirical copula methods for weather forecasting

## *L'effetto di dati ripetuti nei metodi di meteorologia basati su copule empiriche*

Elisa Perrone, Fabrizio Durante, and Irene Schicker

**Abstract** Weather forecasts are typically in the form of an ensemble of forecasts obtained through different runs of physical models. Ensemble forecasts are often biased and affected by errors and need to be statistically post-processed to be corrected. Here, we focus on the empirical copula based techniques for the statistical post-processing of multivariate forecasts. We present the methodology and discuss its pros and cons, especially when ties appear in the ensemble. We consider a case study of joint temperature forecasts for three locations in Austria. We analyze various ways of dealing with ties and show that, in general, the current practice of breaking them at random may not be the optimal solution for forecasting purposes.

**Abstract** *Le previsioni meteorologiche sono spesso espresse sotto forma di un insieme di scenari ottenuti tramite diversi modelli fisici. Queste sono di solito distorte e viziate da errori e devono essere corrette mediante opportune tecniche statistiche. Qui si analizzano le tecniche di correzione basate su copule empiriche. Si presentano i principali aspetti di tale approccio e si discute il caso in cui l'insieme di previsioni presenti dei dati ripetuti (ties). In particolare, si considera un caso di studio riguardante le previsioni di temperatura in tre stazioni meteo localizzate in Austria. Analizzando vari modi di trattare i dati ripetuti si mostra che, in generale, la loro randomizzazione non sempre fornisce la migliore previsione.*

**Key words:** Empirical copulas, Weather forecasting, Statistical post-processing, Ensemble Copula Coupling, Ties.

---

Elisa Perrone  
Eindhoven University of Technology, Groene Loper 3, 5612 AE Eindhoven (The Netherlands)  
e-mail: e.perrone@tue.nl

Fabrizio Durante  
Università del Salento, Centro Ecotekne, 73100 Lecce (Italy)  
e-mail: fabrizio.durante@unisalento.it

Irene Schicker  
Zentralanstalt für Meteorologie und Geodynamik, Hohe Warte 38, 1190 Vienna (Austria)  
e-mail: irene.schicker@zamg.ac.at

## 1 Introduction

Accounting for the right amount of uncertainty is key to the quality of weather predictions. The complexity of the physical atmospheric phenomena makes it hard to achieve so through a single deterministic forecast. As a consequence, meteorologists often use probabilistic numerical weather prediction (NWP) models, which represent the forecast probability distribution as an ensemble of possible forecasts obtained by multiple runs of deterministic physical models. Still, the uncertain initial and boundary model conditions are reflected in the ensemble forecasts, making them biased and affected by errors. As a consequence, raw ensemble forecasts are often statistically post-processed to account for such errors and gain accuracy. In this respect, commonly used approaches consist of two steps. First, we obtain a univariate corrected (parametric) distribution for every single variable of the forecasting problem. Then, we reconstruct the dependence structure from the rank structure of a reference sample. The most popular approaches of this type are *Schaake Shuffle* [2], *Ensemble Copula Coupling* (ECC) [21], and *Sim Schaake* [20]. Each method differs for the chosen reference sample: ECC obtains a reference sample from the raw ensemble forecasts, while Schaake Shuffle and Sim Schaake use past observations. We notice that these methods are all based on empirical copulas, which are mathematical tools to describe the dependence structure of a multivariate sample through their associated ranks [15]. As such, the procedure is implicitly requiring that data can be uniquely ranked and no ties, i.e. repeated observations, appear in the reference sample. Since this is often not the case in practice, how do we handle situations when ties appear in the reference sample?

This problem has a long history, as it can also be noticed from the seminal contribution [10]. Nowadays, the influence of ties in copula methods and models has largely been recognized in the literature, especially for their effects on the statistical estimation and goodness-of-fit tests; see, for instance, [4, 11, 14]. Usually, the practitioner's solution of jittering the data (i.e. add a small error term) to guarantee a unique ranking of the observations should be done with extreme care (see [17]). However, to the best of our knowledge, the influence of ties has not been addressed in the weather forecasting literature, where ties are simply solved at random.

In this work, we study this aspect by focusing on ECC. Starting with a case study of temperature forecasts in Austria originally presented in [18], we construct some realistic simulation scenarios to evaluate the effect of ties on the corrected multivariate forecasts. Several ways of breaking ties are hence considered.

## 2 Statistical post-processing in weather forecasting

In this section, we introduce the general setting and methodology. For a detailed description of the methods and the data, we refer the reader to [18] and references in there. As mentioned in the introduction, we focus on ECC, which is a statistical

post-processing technique based on the ranks of original raw ensemble forecasts. In particular, it consists of the following main steps.

- Step 1 For each variable, we obtain a univariate corrected distribution by applying any univariate statistical postprocessing method (e.g., the Ensemble Model Output Statistics (EMOS) [6]).
- Step 2 We construct a new sample from the corrected univariate distributions (e.g., by taking uniform quantiles).
- Step 3 We reorder each univariate corrected sample according to the rank structure of the raw ensemble forecasts.

We now give a simple illustration of these steps. In general, we assume an ensemble system of  $M \in \mathbb{N}$  members, with  $d = J$ , and  $J \in \mathbb{N}$ , univariate raw margins of the form  $(x_1^{(j)}, \dots, x_m^{(j)})$ , where  $j \in \{1, \dots, J\}$  is a location. We also denote by  $d^*$  the number of variables of our interest. As an example, we consider an ensemble system of size  $M = 6$ , and we aim to correct the temperature forecasts of three stations at a fixed lead-time. Thus, the multivariate forecast of our interest has dimension  $d^* = 3$ . In this scenario, the raw ensemble forecast is a  $(6 \times 3)$ -matrix: each matrix column represents the raw forecasts of the temperature of a particular station, and each matrix row represents a (raw) three-dimensional multivariate forecast. For example, the raw forecasts  $\mathbf{R}$  and its corresponding reference dependence structure  $\mathbf{D}$  might be as follows:

$$\mathbf{R} = \begin{bmatrix} 262.34 & 263.80 & 266.82 \\ 263.14 & 263.88 & 267.73 \\ 262.55 & 263.03 & 269.31 \\ 263.15 & 263.62 & 267.39 \\ 264.92 & 261.22 & 267.10 \\ 260.18 & 265.57 & 265.13 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} (2) & (4) & (2) \\ (4) & (5) & (5) \\ (3) & (2) & (6) \\ (5) & (3) & (4) \\ (6) & (1) & (3) \\ (1) & (6) & (1) \end{bmatrix} \quad (1)$$

where  $(\cdot)$  indicates a rank, and  $\mathbf{D}$  is obtained by transforming the raw forecasts into their corresponding ranks. The first step of ECC is to correct the forecasts of each station. For this aim, we use EMOS, which is a Gaussian-based approach originally presented in [6]. Specifically, EMOS is a regression method that uses the ensemble forecasts as covariates and optimizes the parameters of a Gaussian response distribution to adapt for errors in the mean and uncertainty of the forecasts.

We apply EMOS to each column of  $\mathbf{R}$  and obtain the (corrected) univariate distributions  $F_1 \sim \mathcal{N}(259.5, 1.05)$ ,  $F_2 \sim \mathcal{N}(270, 1.13)$ , and  $F_3 \sim \mathcal{N}(273, 0.8)$ . Then, we construct three samples of size  $M = 6$  from  $F_1$ ,  $F_2$ , and  $F_3$ . Namely, the vectors:

$$s_1 = \begin{bmatrix} 258.38 \\ 258.91 \\ 259.31 \\ 259.69 \\ 260.09 \\ 260.62 \end{bmatrix} \quad s_2 = \begin{bmatrix} 268.79 \\ 269.36 \\ 269.80 \\ 270.20 \\ 270.64 \\ 271.21 \end{bmatrix} \quad s_3 = \begin{bmatrix} 272.15 \\ 272.55 \\ 272.86 \\ 273.14 \\ 273.45 \\ 273.85 \end{bmatrix}$$

A multivariate sample with corrected margins is given by the following matrix  $\mathbf{C}$ .

$$\mathbf{C} = \begin{bmatrix} 258.38 & 268.79 & 272.15 \\ 258.91 & 269.36 & 272.55 \\ 259.31 & 269.80 & 272.86 \\ 259.69 & 270.20 & 273.14 \\ 260.09 & 270.64 & 273.45 \\ 260.62 & 271.21 & 273.85 \end{bmatrix} \quad (2)$$

We notice that  $\mathbf{C}$  does not account for any special dependence structure of the random vector  $(F_1, F_2, F_3)$ , and it might not be representative of the actual distribution of  $(F_1, F_2, F_3)$ . This issue is fixed in the last step of ECC by reordering the entries of each column of  $\mathbf{C}$  according to the ranks in  $\mathbf{D}$ . In our illustrative example, the reordering step results in the matrix  $\tilde{\mathbf{C}}$  given below.

$$\begin{bmatrix} (2) & (4) & (2) \\ (4) & (5) & (5) \\ (3) & (2) & (6) \\ (5) & (3) & (4) \\ (6) & (1) & (3) \\ (1) & (6) & (1) \end{bmatrix} \rightarrow \tilde{\mathbf{C}} = \begin{bmatrix} 258.91 & 270.64 & 272.55 \\ 259.69 & 271.21 & 273.45 \\ 259.31 & 269.36 & 273.85 \\ 260.09 & 269.80 & 273.14 \\ 260.62 & 268.79 & 272.86 \\ 258.38 & 271.21 & 272.15 \end{bmatrix}$$

The post-processed three-dimensional sample  $\tilde{\mathbf{C}}$  of size 6 has the same empirical dependence structure of the raw forecasts. Thus,  $\tilde{\mathbf{C}}$  represents a corrected multivariate sample of  $(F_1, F_2, F_3)$  with a reconstructed inter-variable and spatial dependence. We notice that any ties in the columns of the raw forecasts  $\mathbf{R}$  impact the methodology since the assigned ranks to tied values are arbitrary. In the next section, we focus on this point and compare various ways of breaking ties in a case study setting.

### 3 A case study of temperature forecasts in Austria

We now consider a case study of joint temperature forecasts for three stations in Austria, namely, *Sonnblick*, *Kolm Saigurn*, and *Rauris*. Our setup is similar to the one discussed in [18], where the authors also provide a thorough description of the seventeen-member Austrian ensemble system ALADIN-LAEF. In this work, we consider a temporal period from January 2014 through May 2018, when we have both the raw forecasts and the corresponding true observations.

As the scope of this paper is to analyze the effect of ties, we examine a simulated scenario from this real data situation. Specifically, we introduce ties artificially by simply rounding the values of the raw ensemble forecasts to the closest integer. Table 1 reports five simple tie-breaking rules, which suffice to show the impact of solving ties in our case study. Our choice here is arbitrary and motivated by the exploratory goal of this work. Other criteria that apply to more than three stations are, of course, possible but beyond the scope of this work.

For each day of the testing period, we perform our analysis as follows.

1. We apply EMOS to obtain a corrected distribution from the raw forecasts of each station and construct a sample of size  $M = 17$  by taking uniform quantiles.
2. We reorder the corrected univariate samples according to Table 1.
3. We compute standard scoring rules, namely, the Energy Score [5] and the Variogram Score [22], to quantify the quality of the corrected multivariate forecasts.

We report the results in Table 2. The scores are averaged over the entire testing period, and a better forecast corresponds to a lower score. The comparison also includes the sample of uniform quantiles with no reordering, corresponding to matrix **C** in the example discussed in Section 2. Such a sample, named EMOS-Q, is unrealistic and can be used as a baseline for the other methods. Looking at the scores, we notice that ECC generally improves the forecasts if compared with EMOS-Q. This indicates that, despite the presence of ties, the partial rank structure of the raw ensemble forecasts is still useful to obtain a more accurate multivariate prediction. We now analyze the performance of each ECC method, corresponding to a different way of solving ties. From Table 2, we conclude that ECC 4 shows the best performance both in terms of Energy score and Variogram score. The difference between ECC 4 and ECC 3, and ECC 1 and ECC 2, respectively, reflects the effect of the non-tied values on the final empirical structure of the multivariate corrected forecast. This suggests the importance of choosing the most effective tie-breaking rule for a specific partial rank structure. Overall, we notice that only ECC 1 has higher scores than ECC 5, which suggests that, in this case, randomization does not result in the best multivariate forecast.

**Conclusions.** In this work, we discuss the impact of tied raw forecasts on the performance of ECC. In the future, we plan to provide a more general and effective way of solving ties in this context by accounting for the partial rank structure of the untied raw forecasts.

**Table 1** The five ways of breaking ties for each corresponding station.

	Station 1	Station 2	Station 3
ECC 1	Ascending order	Ascending order	Ascending order
ECC 2	Descending order	Descending order	Descending order
ECC 3	Ascending order	Descending order	Ascending order
ECC 4	Descending order	Ascending order	Descending order
ECC 5	Random order	Random order	Random order

**Table 2** Variogram score and energy score averaged over the period Jan 2014 – May 2018.

	ECC 1	ECC 2	ECC 3	ECC 4	ECC 5	EMOS-Q
Variogram Score	0.381	0.368	0.362	0.350	0.379	0.429
Energy Score	2.710	2.686	2.689	2.668	2.705	2.787

**Acknowledgements** FD has been partially supported by MIUR-PRIN 2017, Project “Stochastic Models for Complex Systems” (No. 2017JFFHSH).



## References

1. V. J. Berrocal, A. E. Raftery, and T. Gneiting. Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135(4):1386–1402, 2007.
2. M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby. The Schaake Shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1):243–262, 2004.
3. F. Durante and C. Sempì. *Principles of Copula Theory*. CRC/Chapman & Hall, Boca Raton, FL, 2015.
4. C. Genest, J. Nešlehová, and B. Rémillard. Asymptotic behavior of the empirical multilinear copula process under broad conditions. *J. Multivariate Anal.*, (159):82–110, 2017.
5. T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
6. T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.
7. T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.
8. T. Gneiting, L. I. Stanberry, E. P. Gritti, L. Held, and N. A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211, 2008.
9. H. Joe. *Dependence Modeling with Copulas*. Chapman and Hall/CRC, 2nd edition, 2014.
10. M. G. Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251, 1945.
11. I. Kojadinovic. Some copula inference procedures adapted to the presence of ties. *Comput. Statist. Data Anal.*, 112(C):24–41, 2017.
12. A. Kolesárová, R. Mesiar, J. Mordelová, and C. Sempì. Discrete Copulas. *IEEE Transactions on Fuzzy Systems*, 14(5):698–705, 2006.
13. S. Lerch, S. Baran, A. Möller, J. Groß, R. Scheffzik, S. Hemri, and M. Graeter. Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Processes in Geophysics Discussions*, 27:1–30, 2020.
14. Y. Li, Y. Li, Y. Qin, and J. Yan. Copula modeling for data with ties. *Stat. Interface*, 13 (1):103–117, 2020.
15. R. Mesiar. Discrete copulas-what they are. In *Proceedings of EUSFLAT- LFA Conference (Barcelona, Spain)*:927–930, 2005.
16. R. B. Nelsen. *An Introduction to Copulas (Springer Series in Statistics)*. Springer, 2nd edition, 2006.
17. R. Pappadà, F. Durante, and G. Salvadori. Quantification of the environmental structural risk with spoiling ties: is randomization worthwhile? *Stoch. Environ. Res Risk Assess.*, 31(10):2483–2497, 2017.
18. E. Perrone, I. Schicker, and M. N. Lang. A case study of empirical copula methods for the statistical correction of forecasts of the ALADIN-LAEF system. *Contributions to Atmospheric Sciences*, 29 (4):277–288, 2020.
19. L. Rüschendorf. On the distributional transform, Sklar’s theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139:3921–3927, 2009.
20. R. Scheffzik. A similarity-based implementation of the Schaake Shuffle. *Monthly Weather Review*, 144(5):1909 – 1921, 2016.
21. R. Scheffzik, T. L. Thorarindottir, and T. Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4):616–640, 2013.
22. M. Scheuerer and T. M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.
23. A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l’Institut de Statistique de Paris*, 8:229–231, 1959.

# Spatio-temporal regression with differential penalization for the reconstruction of partially observed signals

*Regressione spazio-temporale con penalizzazione differenziale per la ricostruzione di segnali parzialmente osservati*

Eleonora Arnone, Laura M. Sangalli

**Abstract** We study a spatio-temporal regression technique with differential penalization, that efficiently handles sparse space-time data and partially observed functional data with spatial dependence. The pattern of observation of these data can be of various types. In the simplest case, the datum is observed uniformly in space and time, in more complex cases, the missing data are clustered in sub-regions. The proposed methodology is suited for dealing with signals that exhibit complex local features or defined over complicated spatial domains. Finally, we consider an application to the study of surface water temperature of Lake Victoria, from data having a high proportion of missing values in a complex pattern.

**Abstract** Studiamo una tecnica di regressione spazio-temporale con penalizzazione differenziale, e analizziamo dati parzialmente osservati con dipendenza spazio-temporale. Il disegno con cui osserviamo questi dati può essere di diversi tipi: nel caso più semplice, il dato è osservato uniformemente nello spazio e nel tempo, mentre in casi più complessi le osservazioni mancanti sono concentrate in sotto regioni. La metodologia proposta è adatta per trattare segnali caratterizzati da variazioni locali o definiti su domini spaziali complicati. Consideriamo infine un'applicazione alla stima della temperatura superficiale del Lago Vittoria, a partire da dati con molti valori mancanti distribuiti in un pattern complesso.

**Key words:** Penalized regression, Smoothing, Partially observed functional data

---

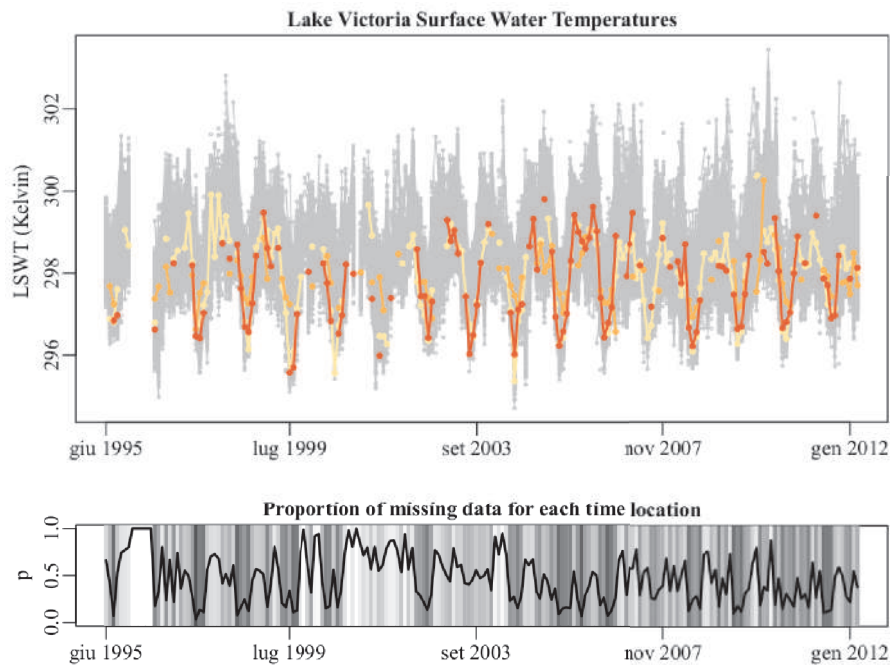
Eleonora Arnone  
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy. e-mail: eleonora.arnone@polimi.it

Laura M. Sangalli  
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy. e-mail: laura.sangalli@polimi.it

## 1 Spatio-Temporal regression for partially observed data

We propose a spatio-temporal regression model with differential penalization (ST-PDE), that efficiently handles sparse space-time data and partially observed functional data with spatial dependence. These data have recently attracted an increasing interest, they are in fact frequently encountered in applications, especially in geosciences and environmental sciences. For example, air pollution data often exhibit a high rate of missing values.

We can think of spatio-temporal data as curves sampled in scattered spatial locations or surfaces observed at some time instants. As an example of the first visualization on the data considered in next section, see Figure 1. Each curve represents the evolution of the quantity of interest over time, for a fixed spatial location. The second visualization is considered in Figure 2, where the spatial field is shown for fixed time instants. We observe from both figures, that there are many time instants and spatial locations where the observations are missing, moreover, the available



**Fig. 1** Temporal profiles of Lake Victoria surface water temperatures (monthly averages) observed over a fine uniform grid covering the lake. Values observed at consecutive time instants are linked by a line. The temporal profiles measured at three random locations over the lake are highlighted in color. The bottom panels shows the proportion of missing data along time, pointing out the high proportion of missing data at various months, including an interval of some consecutive months where no datum is recorded.

observations are not scattered all over the domain, and there are sub-regions where there are no data.

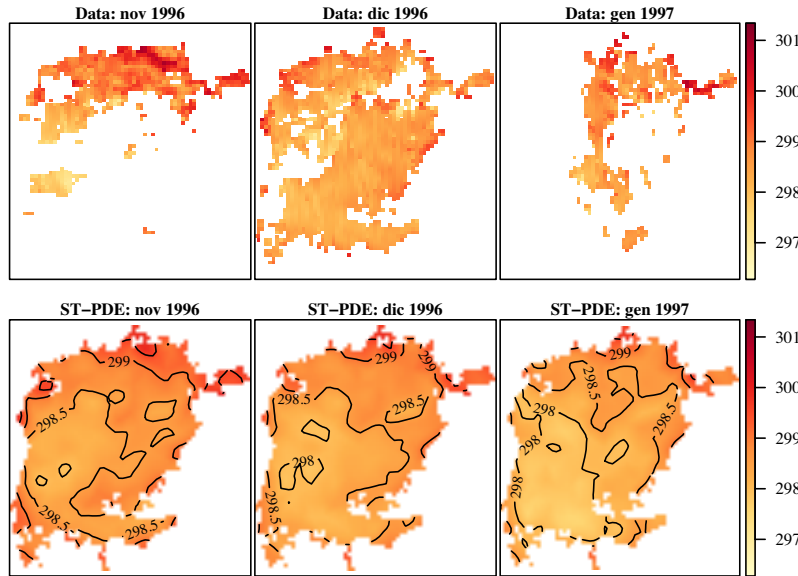
The proposed method constitutes an addition to the class of semiparametric and nonparametric regression models with partial differential equation regularization, reviewed in [7]; see also [8], [3]. With respect to other models that consider a simpler roughness penalty such as the ones in [9], [6], [2], [4] and [1], we here consider a more flexible regularization, which may involve general forms of Partial Differential Equations (PDEs). The estimation functional is discretized using a finite element basis over a triangulation of the spatial domain of interest. This basis is unstructured; therefore the method is able to deal with data observed over domains with irregular shapes and can accurately capture complicated fields with localized features.

The model estimates the spatio-temporal field by minimizing a penalized sum of square error functional, where the considered penalty is:

$$\lambda_S \int_T \int_{\Omega} (Lf - u)^2 d\mathbf{p} dt + \lambda_T \int_{\Omega} \int_T \left( \frac{\partial^2 f}{\partial t^2} \right)^2 dt d\mathbf{p}$$

where  $\lambda_S > 0$  and  $\lambda_T > 0$  are two positive smoothing parameters and  $Lf = u$  is a PDE that formalise the problem-specific information about the unknown spatial field, when available. The differential operator  $L$  has the form:

$$Lf = -\text{div}(K\nabla f) + \mathbf{b} \cdot \nabla f + cf$$



**Fig. 2** On the top row the LWST data of Lake Victoria and on the bottom the corresponding ST-PDE estimate, for three consecutive months.

with  $K \in \mathbb{R}^2$  symmetric and positive definite diffusion tensor,  $\mathbf{b} \in \mathbb{R}^2$  transport vector and  $c \in \mathbb{R}^+$  reaction term. When no prior information is available, isotropic smoothing can be obtained considering null transport and reaction and setting  $K$  to the identity matrix.

## 2 Surface water temperature estimation for Lake Victoria

We here consider the monthly lake surface water temperature (LSWT) of Lake Victoria, that are part of the ARC-Lake database [5]. Figure 1 and the first row of Figure 2 show the available observations. Each datum is an average of LWST over a pixel of  $0.05^\circ$  longitude by  $0.05^\circ$  latitude, and over a month time (not considering nights); the datum is assigned to the spatial coordinates of the centre of the pixel, resulting in 2313 locations over the lake. The observation period consists of 203 months, from June 1995 to April 2012.

The data set is characterized by a large proportion of missing values, about 45%. The bottom panel of Figure 1 highlights the high proportion of missing data in several months, including an interval of some consecutive months where no datum is available, across the whole lake.

The bottom panels of Figure 2 show the estimated field with ST-PDE, for three time instants. Since no prior information is available, we consider isotropic smoothing in the penalization. We observe that the method produces reasonable estimates when there is an high proportion of missing values, as in the left and right panels. From the central panel we can see that the method is able to capture very well the features of the signal where data are available. Moreover, the method is able to deal with a spatial domain characterized by a complicated shape, as shown in Figure 2.

## References

- [1] M. C. Aguilera-Morillo, M. Durbán, and A. M. Aguilera. Prediction of functional data with spatial dependence: a penalized approach. *Stochastic environmental research and risk assessment*, 31(1):07–22, 2017.
- [2] N. H. Augustin, V. M. Trenkel, S. N. Wood, and P. Lorance. Space-time modelling of blue ling for fisheries stock management. *Environmetrics*, 24(2):109–119, 2013.
- [3] L. Azzimonti, L. M. Sangalli, P. Secchi, M. Domanin, and F. Nobile. Blood flow velocity field estimation via spatial regression with pde penalization. *Journal of the American Statistical Association*, 110(511):1057–1071, 2015.
- [4] M. S. Bernardi, L. M. Sangalli, G. Mazza, and J. O. Ramsay. A penalized regression model for spatial functional data with application to the analysis of the production of waste in venice province. *Stochastic environmental research and risk assessment*, 31(1):23–38, 2017.

- [5] S. N. MacCallum and C. J. Merchant. Arc-lake algorithm theoretical basis document—arc-lake v1. 1, 1995–2009 [dataset], 2011.
- [6] G. Marra, D. L. Miller, and L. Zanin. Modelling the spatiotemporal distribution of the incidence of resident foreign population. *Statistica Neerlandica*, 66(2):133–160, 2012.
- [7] L. M. Sangalli. A novel approach to the analysis of spatial and functional data over complex domains. *Quality Engineering*, 32(2):181–190, 2020.
- [8] L. M. Sangalli, J. O. Ramsay, and T. O. Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681–703, 2013.
- [9] M. Ugarte, T. Goicoa, A. Militino, and M. Durbán. Spline smoothing in small area trend estimation and forecasting. *Computational Statistics & Data Analysis*, 53(10):3616–3629, 2009.

# Sea Surface Temperature Effects on the Mediterranean Marine Ecosystem: a Semiparametric Model Approach

## *Effetti della Temperatura Superficiale del Mare sull'Ecosistema Marino del Mediterraneo: un Approccio Basato su Modelli Semiparametrici*

Claudio Rubino, Giacomo Milisenda, Antonino Abbruzzo, Giada Adelfio, Mar Bosch-Belmar, Francesco Colloca, Manfredi Di Lorenzo, Vita Gancitano

**Abstract** Ocean warming is a worldwide phenomenon. The mean temperature of the catch (MTC) is becoming one of the leading indicators to assess the impact of sea surface temperature on fish communities. In this study, we apply a semi-parametric regression approach to the MTC of the catches from MEDITS bottom trawl program in the Strait of Sicily (Central Mediterranean Sea) for the period 1995 to 2018 to evaluate the effects of climate change on continental shelf fish community. All covariates included in the model have a significant impact on the MTC level. Notably, the sea surface temperature (SST) effect on the MTC depends on depth, being positive near the surface and negative at the bottom.

**Abstract** *Il riscaldamento degli oceani è un fenomeno osservabile nei mari di tutto il mondo. La temperatura media delle catture (MTC) rappresenta un indicatore principale per studiare gli effetti dell'aumento della temperatura superficiale del mare sulle comunità ittiche. In questo studio, l'MTC delle catture a strascico del programma MEDITS nello Stretto di Sicilia, per il periodo 1995-2018, viene modellato attraverso un approccio di regressione semiparametrica, per valutare gli effetti del cambiamento climatico sulla comunità di pesci della piattaforma continentale. Le covariate incluse nel modello presentano un impatto significativo sul livello di MTC e in particolare, l'effetto della temperatura superficiale del mare (SST) sembra dipendere dalla profondità.*

---

C. Rubino (*corresponding author*), A. Abbruzzo, G. Adelfio  
Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo  
e-mail: claudio.rubino@unipa.it

G. Milisenda, F. Colloca  
Dipartimento di Ecologia Marina Integrata, Stazione Zoologica Anton Dohrn, Palermo e-mail:  
giacomo.milisenda@szn.it

M. Di Lorenzo, V. Gancitano  
Istituto per le Risorse Biologiche e le Biotecnologie Marine, Consiglio Nazionale delle Ricerche,  
Mazara del Vallo

M. Bosch-Belmar  
Dipartimento di Scienze della Terra e del Mare, Università degli Studi di Palermo

**Key-words:** Marine Ecosystem, Climate Change, GAM, semiparametric approach

## 1 Introduction

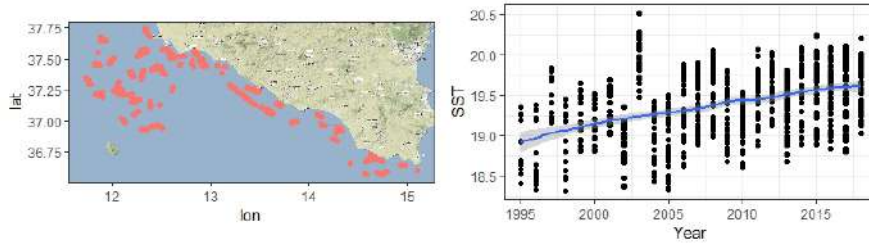
Climatic phenomena and global warming are recognized to be the main driver for sea temperature increase [7]. Such temperature increase may affect the biological characteristics of a population, including somatic growth [2] and some reproduction aspects (e.g. onset and duration of spawning [11]; length/age at maturity [12]). In the Mediterranean Sea, it is also well documented that global warming can drive the distribution and abundance of marine fish populations [10]. In this area, the temperature increase is most evident in the marine faunal composition, which has been altered by alien species from the Suez Canal. It has been estimated that about 400 alien species entered the Mediterranean Sea [8]. [3] recently proposed the mean temperature of the catch (MTC) as an index for evaluating sea warming on fisheries catches. This index is calculated as the average inferred of the temperature preference of the exploited species weighted by their annual catch, that is  $MTC_t = \frac{\sum_{i=1}^n T_i C_{it}}{\sum_{i=1}^n C_{it}}$ , where  $C_{it}$  are the catches of species  $i$  for year  $t$ ,  $T_i$  is the median temperature preference of species  $i$  and  $n$  is the total number of species in the annual catch. Globally, MTC increased at a rate of 0.19°C per decade and is positively related to the change in sea surface temperature (SST) in most large marine ecosystems of the world [3]. Our work's primary goal is to test the effect of sea temperature increase on the central Mediterranean Sea's catch composition, analyzing the MTC index measured on data from scientific surveys.

## 2 Materials and methods

The Strait of Sicily, the south-central Mediterranean Sea, is a transition area connecting the Western Eastern Mediterranean basins. Along the southern coast of Sicily (south Italy), the continental shelf is characterized by two wide and shallow (100 m depth) banks in the western (Adventure Bank) and eastern sectors (Malta Bank), separated by a narrow shelf in the middle part. We collected georeferenced biomass data of fish within the demersal trawl surveys MEDITS (Mediterranean International Trawl Survey program [1]) performed in the study area 1995 to 2018. The MEDITS survey is carried out annually in late spring-early summer, providing a long-term dataset of fishery-independent indices relating to demersal species abundance, demographic structure, and spatial distribution. The sampling design includes hauls between -50 and -800 meters. In the present work, only the hauls located on the continental shelf are considered, assuming that the organisms that inhabit this area have a greater probability of being influenced by changes in the sea surface temperature. At each trawl station, fish species are sorted, weighed, counted and measured, and their relative abundance is expressed as kg/km<sup>2</sup>. Each species'



preferred temperature (median, 25th and 75th percentile) is acquired from the on-line database FishBase (<http://www.fishbase.org>). The MTC is calculated for each haul, according to the MTC index. Several parameters, assumed to be related to MTC level, Surface Sea Temperature (SST, Celsius degrees), depth, categorized in three levels (low: [0-60m], medium: (60-100m], high: (100-200m]), month of the catch, and the spatial (latitude and longitude) and temporal (year) coordinates were used to assess the temporal and spatial changes of the MTC in the Strait of Sicily. Fig. 1 shows the locations of the hauls in the study area (left) and the increasing overall trend of the SST, during the studied period (right).



**Fig. 1** Locations of the catch points in the study area (left) and smoothing of the trend of the SST yearly mean during the time interval 1995-2018.

To assess the dependence of the MTC indicator on the selected covariates, the following semiparametric generalised additive model [5] is applied:

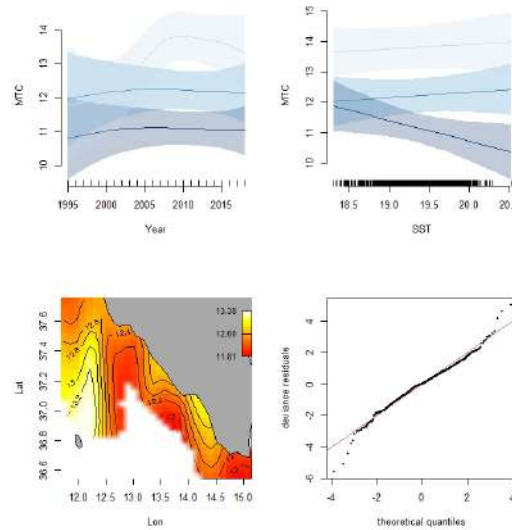
$$\mathbb{E}[\text{MTC}_{s,t} | \mathbf{x}_{s,t}, \mathbf{s}, t] = \beta_0 + \boldsymbol{\beta}' \mathbf{x}_{s,t} + f(\mathbf{s}) + f(t, \text{Depth}) + f(\mathbf{s}, t) \quad (1)$$

where  $\text{MTC}_{s,t} | \mathbf{x}_{s,t}, \mathbf{s}, t \sim \text{Gaussian}$ ,  $\mathbf{s}$  is a  $2 \times 1$  vector of spatial coordinates (latitude and longitude),  $t$  indicates year,  $\beta_0$  is the intercept,  $\boldsymbol{\beta}$  is a  $k \times 1$  vector of regression coefficients,  $f(\mathbf{s})$  is a spatial smoothing,  $f(t, \text{Depth})$  is a smooth function of time for each level of depth,  $f(\mathbf{s}, t)$  is a tensor product structure [13] which aims to model nonlinear interactions between space and time, and the other covariates in vector  $\mathbf{x}$  enter the model parametrically, with SST interacting with depth. The tensor product combines, productwise, the basis functions smoothing space and time; therefore it has to be interpreted as a non parametric interaction, which aims to improve the accuracy of the predictions.

### 3 Results and discussion

The sea surface temperature yearly average has been centred around the sample mean value (19.4 degrees) (Table 1). All covariates included in the model have a significant effect on the MTC level. In particular, the MTC level is higher at shallow depth and lower at the bottom. Also, the parametric effect of the SST on the MTC

depends on the level of depth (Fig. 2, top right): it is positive near the surface, and it becomes negative at the bottom. The Q-Q plot in Fig. 2 (bottom right), which compares deviance residuals of the GAM model with theoretical quantiles of a Gaussian distribution, suggests a reasonably good fit of the model.

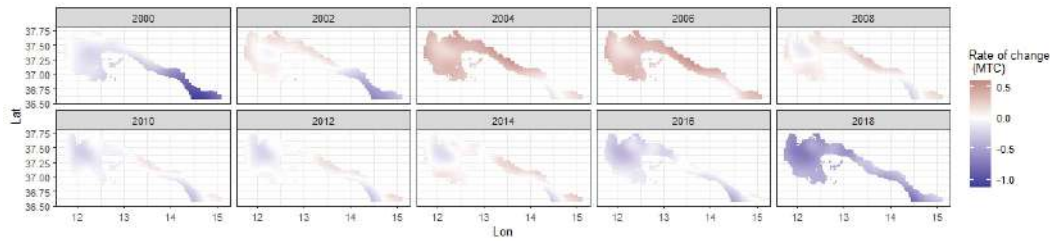


**Fig. 2** Overall trend (top left) and SST effect (top right) for each level of depth (darker shade of blue corresponds to higher levels of depth), spatial effects (bottom left) and QQ-plot of residuals (bottom right).

To highlight both the direction and rate of the change in the MTC level, the difference between the  $t$  and the  $t - 1$  time values, predicted from the model (1), has been computed for a set of locations and time points (Fig. 3).

The proposed analysis shows that the MTC increases in the catch composition in the Central Mediterranean Sea starting in 2002, especially in the shallow waters (Fig. 2, top left). Ismail et al. [6] described the trend of temperature and salinity (from 1995 to 2009) both of the surface water masses (Atlantic waters) and the intermediate ones (which flow from east to west), which cross the Strait of Sicily, highlighting a constant increase of temperature, due to climatic changes. Furthermore, the authors found a significant increase in the growth rate in salinity and temperature from 2003 onwards. This trend agrees with our study’s MTC trends, highlighting the possible composition modifications of demersal communities to environmental changes. This result is in line with recent papers carried out in different areas of the Mediterranean Sea [3].

Fig. 2 (bottom left) shows the average effect of the spatial smoothing; grey areas represent the mainland of Sicily. Spatial heterogeneity, unexplained by the parametric part of the model, is still present. This heterogeneity could be related to the



**Fig. 3** Rate of change of the MTC

particular hydrographic structure of the Strait of Sicily, characterized by important up-welling and by the constant arrival of waters at lower temperatures from the Atlantic Ocean [4].

The observed MTC increase suggests an alteration in the relative catch proportions of species; the thermophilic species (those that prefer warmer temperatures) increased in proportion in the catches over the time series, while psychrophilous (those that prefer colder temperatures) decreased (until 2015). Such change could be due to the displacement of the thermophilous species to a higher latitude and the shift in mean latitude or depth or both psychrophilous species [9]. Furthermore, our results show a negative relationship between the sea surface temperature and the MTC of the deeper areas, suggesting a probable shift of the psychrophilous species from the shallowest to the deeper areas following the increase in SST.

## 4 Conclusion

Ocean warming, along with overfishing, habitat degradation and pollution, is having a large impact on Central Mediterranean marine fisheries, and this study shows that the MTC should be used as a valid proxy to examine and quantify ocean warming impacts. As future work, we will analyze the species-specific composition changes of the species along the time series, wondering which organisms are responsible for the observed MTC increase. This would help to improve our view on the climatic phenomena that are taking place, and to produce management plans for important commercial caught species that also take into account future temperature rises.

<b>A)</b> Parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	12.1164	0.1649	73.4747	< 0.0001
SST (centered)	-0.6809	0.3377	-2.0160	0.0441
Depth Medium (ref: High)	1.0915	0.1628	6.7041	< 0.0001
Depth Low	2.1936	0.2379	9.2221	< 0.0001
Month 6	-0.2941	0.1418	-2.0744	0.0384
Month 7	-0.5617	0.1428	-3.9334	0.0001
Month 8	-0.3992	0.1989	-2.0073	0.0451
Month 9	-0.8724	0.2272	-3.8406	0.0001
Month 10	-0.5658	0.5730	-0.9875	0.3237
Month 11	-0.7665	0.2371	-3.2330	0.0013
Month 12	-0.3745	0.2819	-1.3283	0.1845
SST yearly mean:Depth Medium	0.8606	0.4070	2.1142	0.0348
SST yearly mean:Depth Low	0.8355	0.4115	2.0302	0.0427
<b>B)</b> Smooth terms	edf	Ref.df	F-value	p-value
s(Lon,Lat)	15.6607	20.2640	2.3368	0.0008
s(Year):Depth High	1.0001	1.0002	0.3708	0.5428
s(Year):Depth Medium	1.0014	1.0026	0.8859	0.3470
s(Year):Depth Low	4.2162	5.2041	5.2535	0.0001
s(Lon,Lat,Year)	19.2493	26.6827	1.6384	0.0218

**Table 1** Estimates of the parametric coefficients **A)** and smooth terms **B)** of the GAM model.

## References

- Bertrand, J.A., Gil De Sola L., Papaconstantinou C., Relini G., Souplet A.: The general specifications of the MEDITS surveys. *Scientia Marina* **66**, 9–17 (2002).
- Brander, K. M.: The effect of temperature on growth of Atlantic cod (*Gadus morhua* L.). *ICES J. Mar. Sci.* **52**, 1–10 (1995).
- Cheung, W. W. L., Watson, R., Pauly, D.: Signature of ocean warming in global fisheries catch. *Nature* **497**, 365–369 (2013).
- Di Lorenzo, M., Sinerchia, M., Colloca, F.: The North sector of the Strait of Sicily: a priority area for conservation in the Mediterranean Sea. *Hydrobiologia* **821**(2), 235–253 (2018).
- Hastie, T., Tibshirani, R.: Generalized additive models (with discussion). *Statistical Science* **1**, 297–318 (1986).
- Ismail, S.B., Schroeder, K., Sammari, C., Gasparini, G.P., Borghini, M., Aleya, L.: Interannual variability of water mass properties in the Tunisia–Sicily Channel. *Journal of Marine Systems* **135**, 14–28 (2014).
- Levitus, S., Antonov, J. I., Boyer, T. P., Stephens, C.: Warming of the world ocean. *Science* **287**(5702), 2225–2229 (2000).
- Nunes, A.L., Katsanevakis, S., Zenetos, A., Cardoso, A.C.: Gate ways to alien invasions in the European seas. *Aquat. Invasions* **9**, 133–144 (2014).
- Perry, A., Low, P.J., Ellis, J.R., Reynolds, J.D.: Climate Change and Distribution Shifts in Marine Fishes. *Science* **324**, 1912–1915 (2005).
- Tsikliras, A. C.: Climate-related geographic shift and sudden population increase of a small pelagic fish (*Sardinella aurita*) in the eastern Mediterranean Sea. *Mar. Biol. Res.* **4**, 477–481 (2008).
- Tsikliras, A. C., Antonopoulou, E., Stergiou, K. I.: Spawning period of Mediterranean marine fishes. *Rev. Fish Biol. Fish.* **20**, 499–538 (2010).
- Tsikliras, A. C., Stergiou, K. I.: Age at maturity of Mediterranean marine fishes. *Mediterr. Mar. Sci.* **16**, 5–20 (2015).
- Wood, S. N.: *Generalized Additive Models: An Introduction with R*. Boca Raton, Chapman & Hall/CRC (2017).

## 4.13 Functional data analysis

# Remote Analysis of Chapas Stops in Maputo from GPS data: a Functional Data Analysis Approach

*Analizzare in remoto le fermate chapas a Maputo tramite dati GPS: un approccio basato sull'analisi dei dati funzionali*

Agostino Torti<sup>1,2</sup>, Davide Ranieri<sup>1</sup> and Simone Vantini<sup>1</sup>

**Abstract** SAFARI is an interdisciplinary research project of Politecnico di Milano which aims at fighting the inefficiency of transportation services in African Sub-Saharan urban areas. Applying tools from Functional Data Analysis, we develop a fully replicable and scalable approach to study how people move inside an urban area starting from only GPS data, therefore providing a flexible, automatic and self-adaptive algorithm that can be applied, with minimal costs, to any reality. We focus on Maputo, the capital city of Mozambique, as study pilot, and we demonstrate the capabilities of our approach studying the access of general population to the chapas service (note that vans and minibuses by which people mostly move are commonly know as chapas). The obtained results highlight different spatio-temporal usage patterns across the city and provide useful insights to the mobility managers.

**Abstract** SAFARI è un progetto di ricerca interdisciplinare del Politecnico di Milano nato con lo scopo di combattere l'inefficienza del servizio di trasporti nelle aree urbane dell'Africa subsahariana. Applicando l'analisi dei dati funzionali a dati GPS, ci proponiamo di sviluppare un approccio totalmente scalabile per lo studio dei flussi di mobilità urbani. L'obiettivo ultimo è quello di sviluppare un software automatico che possa essere applicato, con costi minimi, a qualsiasi realtà. In questo lavoro ci focalizziamo su Maputo, la capitale del Mozambico, e ne studiamo i flussi di mobilità compiuti tramite il servizio di chapas (i bus e mini-bus con cui la maggior parte della popolazione è solita muoversi in città). I risultati ottenuti evidenziano i pattern spatio-temporali di utilizzo e forniscono utili linee guida ai mobility manager.

**Key words:** Mobility, GPS Data, Functional Data Analysis, Maputo

---

<sup>1</sup>MOX - Department of Mathematics, Politecnico di Milano, Italy

<sup>2</sup>Center for Analysis Decisions and Society, Human Technopole, Milano

## 1 Introduction

Mobility in African Sub-Saharan urban areas is supported by informal transport systems because many cities have not developed efficient public collective mobility systems. In details, 80% of the urban mobility is supplied by paratransit, i.e. informal transportation services that supplement public mass transit by providing rides without fixed routes or timetables. This inefficiency of transportation services has many disadvantages: road accidents are the second reason of death in African cities and social and economic development of cities is being prevented. Facing these critical issues is the final objective of Safari Njema project (SAFARI), an interdisciplinary research project of Politecnico di Milano launched in March 2019 after the attainment of Polisocial Award (the program of Politecnico which honours the research projects of commitment and social responsibility with fundings from 5x1000 contributions). The idea is to make use of still underexploited resource, as mobile phone data and statistical modelling, so to restructure the current mobility offer and prove safer, reliable, economically sustainable solutions to both people demand and business models in a fully replicable and scalable perspective. By operating in this way, we aim at defining a flexible procedure to study, with minimal costs, mobility flow in any urban area around the world, achievement that would be extremely expensive and almost impossible to fulfill in the old fashioned way.

To pursue this objective, a pilot project is designed in Maputo, the capital city of Mozambique. Maputo is a city of 1.77 million of inhabitants with a surface of 346  $km^2$ , where people mostly move using privately operated vans and minibuses which are commonly know as chapas. The project is based on the exploitation of BigData Analysis to understand how the city can afford mobility policies to improve informal mobility. Among the pursued results from SAFARI, in this work, the focus is on the analysis of the chapas stops: we aim at defining a scalable and replicable approach to study the access of general population to the chapas service, i.e. departures, arrivals and waiting times at each chapas stop. Analysing GPS location data, we first profile each chapas stop with a set of curves - describing both the departure and arrival daily profiles of people at each stop, along with the distribution of the waiting times - and then segment the city in different activity areas by highlighting different usage patterns. By doing so, we aim at providing useful insights to the municipality so to handle with mobility management and make feasible and efficient future plans.

## 2 Methods and Analyses

To study the access of general population to the chapas service in Maputo, we develop a complete pipeline based on tools from Functional Data Analysis (FDA), the branch of statistics dealing with curves, surfaces or anything else varying over a continuum ([4]). Notice that the usage of FDA to study mobility data is nowadays well established in the literature (e.g., [1], [6])

The input of our model are GPS data provided by Cuebiq company through their "Data for Good" program. Cuebiq provides anonymous, privacy-compliant location-based data to academic research and humanitarian initiatives related to human mobility. These de-identified data are collected from users who opt-in to share their data anonymously for research purposes, through a GDPR-compliant framework. Cuebiq then applies additional privacy preservation techniques to remove sensitive locations from the dataset, and to obfuscate personal areas such as home locations by "upleveling" them to 600m x 600m geohash tiles.

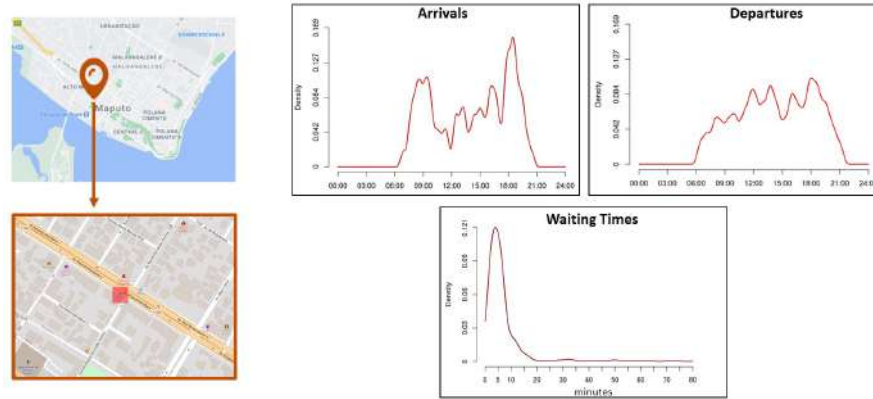
The first step of our approach is a preprocessing step in which, for each chapas stop, we discern between arrivals, departures, people waiting and people passing. To do so, for each chapas stop, a 20 meters square is designed around it and all the GPS trajectories - i.e. a sequence of GPS points which records the spatial track with related timestamps of a moving person - passing through it are taken into account. Then, a modal splitting procedure is employed by estimating the speed of travel at each point of the trajectory, therefore distinguish between people waiting, moving on foot or moving by vehicle. In the end, logically looking at the obtained information for each point of each trajectory, we can deduce who is getting on/off the chapas and who is waiting at the chapas stop.

The next step of our approach is the estimation of the functional data associated to each chapas stop. We apply a kernel density estimation smoothing technique ([2]) to estimate the daily density functions of both departures and arrivals, together with the density function of the waiting times. In details, fixed a chapas stop and a period of interest (e.g., one month), we define, for that chapas stop, a multivariate functional datum composed from three curves:

- a curve representing the departure density profile of people which arrived on foot at the chapas stop and are departing by chapas at time  $t \in (0, 24)$ ;
- a curve representing the arrival density profile of people who arrived by chapas at the chapas stop and are getting off at time  $t \in (0, 24)$ ;
- a curve representing the density function of the waiting times, in minutes, at the chapas stop.

As example, we show the obtained functional data for a chapas stop in Belita district, one of the main commercial area of Maputo. Note that, the time span of the analysed data goes from July 2019 until March 2020. In Figure 1, we display both the map of Maputo, highlighting the selected chapas stop, and the related functions. By simply looking at the graph, various information concerning the usage of the chapas stop are highlighted: the arrivals reveal two main peaks of usage during the day, respectively, in the morning and afternoon rush hours, while the departures reveal a flatter behavior during the day; the waiting times show a high positive skewness with a mode centered around five minutes.





**Fig. 1** Left: map of Maputo highlighting the selected chapas stop. Right: the curves of Arrivals, Departures and Waiting Times at the chapas stop.

The final step of the developed procedure is the within chapas stop and between chapas stops comparison. All of this is possible by setting different periods of interest (e.g. working vs non-working days) and applying both depth measures and clustering techniques from FDA framework ([3], [5]). Proceeding in this way, it is possible to highlight different spatio-temporal patterns of how people move by chapas through the city of Maputo.

### 3 Conclusions

This work is part of the SAFARI interdisciplinary research project whose aim is to improve informal mobility in Sub-Saharan African urban areas. Analysing GPS data in Maputo through tools from FDA, we developed a fully flexible and scalable approach to understand the access of general population to the chapas service, namely looking at departures, arrivals and waiting times at each chapas stop. During the presentation of this work, we will show how the developed methodology allows to segment the city in different activity areas, highlighting different spatio-temporal usage patterns and hence providing useful insights to the municipality, allowing them to best handle mobility flow and make feasible and efficient future plans. In conclusion, we remark that the the developed approach can be applied, with minimal costs, to any urban area, so to study its intrinsic mobility patterns, achievement that would be almost impossible to fulfill and extremely expensive in the old fashioned way.

## Acknowledgment

The authors acknowledge all partners of the "Safari Njema Project", funded by Polisocial Award 2018 - Politecnico di Milano.

## References

1. Charles Bouveyron, Etienne Côme, Julien Jacques, et al. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760, 2015.
2. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
3. Francesca Ieva and Anna M Paganoni. Depth measures for multivariate functional data. *Communications in Statistics-Theory and Methods*, 42(7):1265–1276, 2013.
4. James O Ramsay. *Functional data analysis*, volume 4. Wiley Online Library, 2004.
5. Laura M Sangalli, Piercesare Secchi, Simone Vantini, and Valeria Vitelli. K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233, 2010.
6. Agostino Torti, Alessia Pini, and Simone Vantini. Modelling time-varying mobility flows using function-on-function regression: Analysis of a bike sharing system in the city of milan. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(1):226–247, 2021.

# A Conformal approach for functional data prediction

## *Un approccio Conforme per la previsione di dati funzionali*

Jacopo Diquigiovanni, Matteo Fontana and Simone Vantini

**Abstract** The contribution deals with the key challenge of creating prediction sets in the functional data framework. Starting from the investigation of the literature concerning this topic, we propose an innovative approach building on top of Conformal Prediction able to overcome the main drawbacks characterizing the existing approaches. The nonparametric method proposed is able to construct finite-sample either valid or exact prediction bands under minimal distributional assumptions. Different specifications of the method are compared in terms of efficiency in some simulated scenarios.

**Abstract** Il contributo affronta il tema cruciale della creazione di insiemi previsivi nel contesto dei dati funzionali. A partire dall'esplorazione della letteratura riguardante questo argomento, proponiamo un approccio innovativo basato sulla Conformal Prediction che è capace di superare le limitazioni principali che contraddistinguono gli approcci esistenti. Il metodo non parametrico che proponiamo è capace di costruire bande di previsione valide o esatte per ogni dimensione campionaria facendo minime assunzioni distributive. Confrontiamo alcune specificazioni del metodo in termini di efficienza prendendo in considerazione diversi scenari simulati.

**Key words:** Conformal prediction, Distribution-free prediction set, Exact prediction set, Functional data, Prediction band, Valid prediction band

---

Jacopo Diquigiovanni

Dept. of Statistical Sciences, University of Padova, Via Cesare Battisti 241, Padova, 35121, Italy, e-mail: jacopo.diquigiovanni@phd.unipd.it

Matteo Fontana

MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy. Now at Joint Research Centre - European Commission, Ispra (VA), Italy e-mail: matteo.fontana@ec.europa.eu

Simone Vantini

MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy, e-mail: simone.vantini@polimi.it

## 1 Introduction

Functional Data Analysis ([5] [12]) is a field of paramount interest in Statistics. Its purpose is to develop new methodological approaches to deal with functions characterized by some degree of smoothness. Moving from the manuscript written by Jim O. Ramsay in the nineteen-eighties [11], several authors have proposed many interesting findings in this field: Functional Boxplots [14], Functional Principal Component Analysis [12] and Functional Linear Regression [12] are only a few examples of works concerning this ebullient topic. However, many issues are unfortunately still open research problems given the intrinsic complexity of the framework considered. Among these, we deal with the issue of creating prediction sets for independent and identically distributed functional data. Specifically, given a nominal confidence level  $1 - \alpha$ , the purpose is to develop a method able to output either exact - i.e. guaranteeing a coverage exactly equal to  $1 - \alpha$  - or valid - i.e. guaranteeing a coverage close to  $1 - \alpha$ , but never less than the nominal confidence level - prediction sets. Only few manuscripts have addressed this crucial challenge in the functional data analysis framework. These works can be generally divided into two groups: the first group is made up of approaches based on parametric bootstrapping techniques [3, 2], while the second group is characterized by approaches based on dimensionality reduction techniques [7, 1]. Both groups of approaches carry some drawbacks: first of all, both of them are either based on non-trivial distributional assumptions and/or asymptotic statements. In addition, the first class of techniques is computationally demanding, whereas the second class is obviously affected by the approximation induced by the dimensionality reduction. In view of this, this contribution will focus on presenting a procedure able to overcome these shortcomings by means of a new approach in the field of Conformal Prediction [15].

## 2 Short Outline

The first part of the contribution will focus on Conformal Prediction [15, 13, 16], an innovative nonparametric approach to create prediction sets proposed in the Machine Learning framework as a method to construct prediction intervals for Support Vector Machines [6] and used in the functional context as an exploration tool via the use of a truncated basis method [9]. Specifically, we consider the Split Conformal approach and the Smoothed Split Conformal approach. The first approach constructs finite-sample valid prediction sets under the assumption of exchangeable data, whereas the second one is able to generate finite-sample exact prediction sets under the same assumption. The core of both approaches is the choice of the *non-conformity measure*, namely any measurable function which takes values in  $\bar{\mathbb{R}}$  and whose aim is to score how different the observation we aim to predict is with respect to the observed sample  $y_1, \dots, y_n$ .

The second part of the contribution will focus on the definition of the desirable properties that a prediction set must satisfy in the functional framework. An aspect

of practical interest concerns the shape of the sets: in particular, we will show that a desirable prediction set must be a prediction band when data are functions [10, 9], since it allows an easy visualization of the prediction set in parallel coordinates [8]. Indeed, the prediction band, which can be defined as the Cartesian product of infinitely many intervals (i.e. one interval for each point of the domain), by definition coincides with its envelope and is not merely a subset of it as happens when different shapes are considered.

The third part of the contribution will focus on the definition of a new group of nonconformity measures based on the supremum metric. All the nonconformity measures belonging to this group build prediction sets that, in addition to be distribution-free, finite-sample either valid or exact according to the Conformal approach considered, and in addition to be bands as required in the functional framework, can be found in closed form. Some emphasis will be also placed on the computational cost characterizing the method, since the procedure is highly scalable as the computational effort required by the procedure increases only linearly with the sample size  $n$ .

Finally, the fourth part of the contribution will focus on the efficiency of the method. In the Conformal Prediction framework, the term efficiency is related to the size of the prediction sets returned by a given procedure. Different nonconformity measures - belonging to the aforementioned group of nonconformity measures - will be compared in different scenarios through simulation studies and a real-world application. A detailed description of the method, simulations and application can be found in [4].

## References

1. Antoniadis, A., Brossat, X., Cugliari, J., Poggi, J.: A prediction interval for a function-valued forecast model: Application to load forecasting. *International Journal of Forecasting*. **32**(3), 939-947 (2016)
2. Cao, G., Yang, L., Todem, D.: Simultaneous Inference For The Mean Function Based on Dense Functional Data. *Journal of Nonparametric Statistics*. **24**(2), 359-377 (2012)
3. Degras, D. A.: Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*. **21**(4) (2011)
4. Diquigiovanni, J., Fontana, M., Vantini, S.: The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data. arXiv:2102.06746 (2021)
5. Ferraty, F., Vieu, P.: *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media (2006)
6. Gammernan, A., Vovk, V., Vapnik, V.: Learning by Transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. 148-155 (1998)
7. Hyndman, R. J., Shahid Ullah, M.: Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*. **51**(10), 4942-4956 (2007)
8. Inselberg, A.: The plane with parallel coordinates. *The visual computer*. **1**(2), 69-91 (1985)
9. Lei, J., Rinald, A., Wasserman, L.: A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*. **74**(1-2), 29-43 (2015)
10. López-Pintado, S., Romo, J.: On the concept of depth for functional data. *Journal of the American Statistical Association*. **104**(486), 718-734 (2009)

11. Ramsay, J. O.: When the data are functions. *Psychometrika*. **47**(4), 379–396 (1982)
12. Ramsay, J. O., Silverman, B. W.: *Functional data analysis*. Springer series in statistics. Second edition (2005)
13. Shafer, G., and Vovk, V.: A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*. **9**, 371–421 (2008)
14. Sun, Y., Genton, M. G.: Functional Boxplots. *Journal of Computational and Graphical Statistics*. **20**(2), 316–334 (2011)
15. Vovk, V., Gammerman, A., Shafer, G.: *Algorithmic learning in a random world*. Springer Science & Business Media (2005)
16. Zeni, G., Fontana, M., Vantini, S.: Conformal Prediction: a Unified Review of Theory and New Challenges. arXiv: 2005.07972 (2020)

# Block testing in covariance and precision matrices for functional data analysis

## *Test per blocchi della matrice di covarianza o precisione per dati funzionali*

Marie Morvan, Alessia Pini, Madison Giacomci, Valerie Monbet

**Abstract** We propose a method to test dependence or conditional dependence between parts of the domain of functional data. The tests are based on permutation procedure that tests if suitable blocks of the covariance or precision matrix of basis expansion coefficients are equal to zero. We show that the procedure is able to identify the true structure of conditional dependence.

**Abstract** *Proponiamo un metodo per testare indipendenza o indipendenza condizionale tra parti del dominio di dati funzionali. Utilizziamo test di permutazione per verificare se blocchi della matrici di covarianza o precisione dei coefficienti di una espansione in base sono uguali a zero. Mostriamo che la procedura in grado di identificare in maniera corretta la vera struttura di indipendenza condizionale.*

**Key words:** functional data analysis, independence, conditional independence

## 1 Introduction

Functional data analysis (FDA) is a particularly lively area of statistical research (see [10, 11, 8] and references therein). One of the current lines of research in FDA is local inference - i.e., methods where the null hypothesis is tested locally by defining a function that assigns a  $p$ -value to each point of the domain (e.g., [13, 9]). Most

---

Marie Morvan

Department of Mathematics, Universit Rennes 1 e-mail: marie.morvan@univ-rennes1.fr

Alessia Pini

Department of Statistical Sciences, Universit Cattolica del Sacro Cuore, Milan e-mail: alessia.pini@unicatt.it

Madison Giacomci

Universit Rennes 2 e-mail: joyce.giacofci@univ-rennes2.fr

Valerie Monbet

Department of Mathematics, Universit Rennes 1 e-mail: valerie.monbet@univ-rennes1.fr

of those methods are meant to deal with inference on the mean, group-comparison, or regression. In this work we focus instead on the covariance of data. Our aim is to identify a sparse structure on it by testing independence between different parts of the functions' domain. For Gaussian multivariate data, independence or conditional independence can be easily tested by making inference on the elements of the correlation or of the precision (the inverse of the covariance) matrices.

In FDA, instead, this problem poses many methodological challenges, since data are infinite-dimensional, so the covariance is an infinite dimensional operator. Furthermore, the sample covariance operator is non-invertible, and testing for conditional independence is particularly challenging. We will assume here that data can be described by means of a (possibly high-dimensional) B-splines basis expansion. In such case, coefficients of the basis expansion are directly related to the parts of the domain where the support of basis functions is strictly positive. The covariance structure between functional data is then univocally identified by the covariance matrix of basis expansion coefficients. Since the basis is high-dimensional, the number of basis functions is possibly higher than the sample size, so the sample estimator of the covariance matrix has a high variability, and usually cannot be directly inverted. In addition, unlike multivariate data, basis components are naturally ordered along the domain, and it is possible to exploit this information for inference.

We further assume that the domain can be partitioned into regions of interest. This is usually the case when it is possible to identify landmarks on functional data (regions of interests are intervals whose endpoints are landmarks), or where some information is available on the domain. In such a case, we expect the covariance (precision) matrix to have a block structure, where blocks correspond to elements of the partition. So, to infer about which areas of the domain - that are related to components of the partition - are independent (conditionally independent) between each other, we could focus on blocks of the covariance (precision) matrix.

A possible solution would be to estimate the precision matrix using a penalization like in the graphical lasso (Glasso [1]). However, Glasso penalization does not induce a block structure. Here, we focus instead on the problem on an inferential perspective, since we aim at controlling precisely the probability of making an incorrect decision (that is, selecting two parts of the domain as dependent where they are not). In the case of conditional dependence, this is done by [3] by performing a test on each block and adjusting for multiplicity for controlling false discovery rate (FDR). Instead of the FDR, we are interested here in controlling the family-wise error rate. In addition, for performing the multiplicity adjustment, we will directly exploit the information about proximity between blocks that is not used in [3].

## 2 Methodology

We assume to observe Gaussian functional data  $(\mathbf{X}_1, \dots, \mathbf{X}_n)$  defined on the domain  $D \subset \mathbb{R}$ . We call  $(\mathbf{C}_1, \mathbf{C}_n)$  the vectors of the  $p$  coefficients of the basis expansion  $\mathbf{X}_i(t) = \sum_{j=1}^p \phi_j(t)C_{ij}$ ,  $i = 1, \dots, n$ . We also assume that the coefficients



$\{C_{ij}\}$  are partitioned into  $M$  blocks, associated to different portions of the curves. Let  $\mathcal{J}_1, \dots, \mathcal{J}_M \subset \{1, \dots, p\}$  be a prespecified partition of  $\{1, \dots, p\}$  which index block membership. We also assume that blocks are ordered according to index  $m = 1, \dots, M$ , that is that the first block is associated to a first portion of the domain, the second block is associated to a second portion, and so on.

We denote as  $\mathbf{R} = (r_{ij})_{i,j=1,\dots,p}$  the  $p \times p$  correlation matrix of basis coefficients and  $\mathbf{\Omega} = (\omega_{ij})_{i,j=1,\dots,p}$  the  $p \times p$  precision matrix. We specify two different sets of tests, one on the correlation matrix  $\mathbf{R}$  and one on the precision matrix  $\mathbf{\Omega}$ . In both cases, the blocks that we have introduced are associated to submatrices of  $\mathbf{R}$  and  $\mathbf{\Omega}$ .

In both cases, since we are jointly performing  $M(M - 1)/2$  tests (one for each couples of different blocks), so we need to adjust for the multiplicity of tests, controlling the family-wise error rate. We start by describing the method that we propose for performing the test on a given submatrix, and then we specify how to adjust such tests for multiplicity.

### 2.1 Tests on the Correlation Matrix

On the correlation matrix, we wish to test the hypothesis of independence between blocks. In the Gaussian case, this is equivalent to test for zero correlation. In particular for all couples of blocks  $\mathcal{J}_m, \mathcal{J}_{m'}$  with  $1 \leq m < m' \leq M$ , we test:

$$H_{0,m,m'} : \mathbf{R}_{\mathcal{J}_m \times \mathcal{J}_{m'}} = 0 \quad \text{versus} \quad H_{1,m,m'} : \mathbf{R}_{\mathcal{J}_m \times \mathcal{J}_{m'}} \neq 0. \quad (1)$$

We use permutation tests for testing hypotheses (1). We need to define a set of permutations preserving the likelihood under the null hypothesis, and a test statistic that is stochastically greater under the alternative hypothesis.

Since functional data  $\mathbf{X}_i$  are iid, also the vectors  $\mathbf{C}_i$  of coefficients of the basis expansion are iid (and thus exchangeable with respect to units). In addition, under the null hypothesis, subvectors  $\mathbf{C}_{i_{\mathcal{J}_m}}$  and  $\mathbf{C}_{i_{\mathcal{J}_{m'}}$  are independent between each other for all  $i$ , and thus exchangeable. Hence,  $n!$  likelihood-invariant permutations can be found by permuting the units in the first vector keeping the second one fixed, (or vice-versa).

As test statistic, we use the sum of the correlations between units of  $\mathcal{J}_m$  and  $\mathcal{J}_h$ . The unadjusted  $p$ -value of the test is defined as the number of permutations (out of the total  $n!$ ) leading to a test statistic greater or equal to the one observed with the non-permuted data.

Since the permutations are likelihood-invariant under  $H_{0,m,m'}$ , the test is exact. Since  $\hat{r}_{ij}$  is a consistent estimator of  $r_{ij}$ , the test statistic is stochastically greater under  $H_{1,m,m'}$  than under  $H_{0,m,m'}$ , so the test is also consistent [6].

## 2.2 Tests on the Precision Matrix

We wish to test the null hypothesis of conditional independence between blocks. In the Gaussian case, this is equivalent to test if the entries of the precision matrix are equal to zero on the given submatrix. In particular, for all couples of blocks  $\mathcal{J}_m, \mathcal{J}_{m'}$  with  $1 \leq m < m' \leq M$ , we test:

$$H_{0,m,m'} : \boldsymbol{\Omega}_{\mathcal{J}_m \times \mathcal{J}_{m'}} = 0 \quad \text{versus} \quad H_{1,m,m'} : \boldsymbol{\Omega}_{\mathcal{J}_m \times \mathcal{J}_{m'}} \neq 0. \quad (2)$$

We use permutation tests for testing hypotheses (2). In this case, the null hypothesis  $\boldsymbol{\Omega}_{\mathcal{J}_m \times \mathcal{J}_{m'}} = 0$  means that the subvectors  $\mathbf{C}_{i_{\mathcal{J}_m}}$  and  $\mathbf{C}_{i_{\mathcal{J}_{m'}}}$  are conditionally independent given  $\mathbf{C}_{i_{\{1,\dots,p\} \setminus (\mathcal{J}_m \cup \mathcal{J}_{m'})}}$ . Subvector  $\mathbf{C}_{i_{\mathcal{J}_m}}$  is no longer exchangeable with respect to units while keeping  $\mathbf{C}_{i_{\mathcal{J}_{m'}}}$  fixed.

Since data are assumed to be Gaussian, conditional independence is equivalent to conditional linear independence:

$$\mathbf{C}_{\mathcal{J}_m} = \mathbf{C}_{\{1,\dots,p\} \setminus (\mathcal{J}_m \cup \mathcal{J}_{m'})} \mathbf{A} + \boldsymbol{\varepsilon}_{\mathcal{J}_m} \quad (3)$$

$$\mathbf{C}_{\mathcal{J}_{m'}} = \mathbf{C}_{\{1,\dots,p\} \setminus (\mathcal{J}_m \cup \mathcal{J}_{m'})} \mathbf{A}' + \boldsymbol{\varepsilon}_{\mathcal{J}_{m'}} \quad (4)$$

where  $\mathbf{A}$  and  $\mathbf{A}'$  are two matrices of dimension  $n \times p - |\mathcal{J}_m| - |\mathcal{J}_{m'}|$  (where we used the notation  $|\cdot|$  for the cardinality of a set), and  $\boldsymbol{\varepsilon}_{\mathcal{J}_m}$  and  $\boldsymbol{\varepsilon}_{\mathcal{J}_{m'}}$  are mutually independent residuals. We propose to compute uncorrelated residuals of models (3)-(4) and permute them, according to the method proposed by [2]. This will lead to an asymptotically exact test, since residuals are only asymptotically exchangeable.

Similarly to the previous case of testing independence, the test statistic is the sum of squared elements of the estimated precision matrix. The unadjusted  $p$ -value of the test is defined as the number of permutations (out of the total  $n!$ ) leading to a test statistic greater or equal to the one observed with the non-permuted data.

## 2.3 Multiple Testing of Submatrices

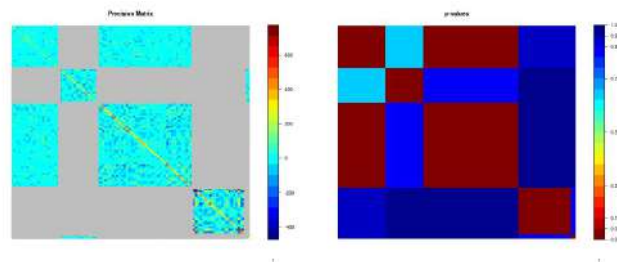
Once the test on each submatrix is done, it is important to adjust results to take into account multiplicity. Most multiplicity adjustment methods (e.g., Bonferroni, Holm [4], Benjamini-Hochberg [5], the test proposed by [3]) are specifically designed for multivariate data, so they do not take into account the specific structure of functional data, that is, the fact that basis coefficients (and hence blocks in our case) are naturally ordered along the domain. In our work, we extend the interval-wise testing procedure (IWT) first proposed by [7], that is a multiple testing method that takes into account the ordered structure of functional data.

We propose to define the adjusted  $p$ -value based on the following procedure. For simplicity we describe it for the test of independence (1), but the procedure can be applied in the same way to the test of conditional independence (2).

1. Perform the tests (1) of independence between each couple of blocks.
2. Perform a tests of independence between each couple of non-overlapping intervals of blocks. Such tests can also be performed with the same permutation test described in the last subsection, it is only necessary to change the tested blocks to non-overlapping intervals of blocks.
3. For each couple of blocks  $\mathcal{I}_m$  and  $\mathcal{I}_{m'}$ , compute the adjusted  $p$ -value as the maximum between all  $p$ -values of tests of intervals of blocks including  $\mathcal{I}_m$  and  $\mathcal{I}_{m'}$  (including blocks  $\mathcal{I}_m$  and  $\mathcal{I}_{m'}$  themselves).

### 3 Numerical Example

In this section we describe a simulation study for the test of conditional independence. Let us consider a covariance matrix  $M = P^{-1}DP$ , with  $D$  the diagonal matrix of eigenvalues of  $M$  and  $P$  a matrix of eigenvectors. The sparse block structure of the covariance matrix  $M$  is inherited from the block structure of  $P$ . We consider a case where the precision matrix has 5 blocks, with 3 large blocks and two small ones (see Figure 1). The non-zero blocks couples are fixed in the eigenvector matrix as random rotation matrices. In such way, we obtain a block-diagonal matrix  $P$ , with some non-zero off-diagonal blocks. The resulting matrix is positive definite, easily invertible with a sparse structure in blocks. The sparse precision matrix is finally obtained as:  $M^{-1} = P^{-1}D^{-1}P$ . Left panel of Figure 1 shows a precision matrix with  $p = 100$  and 5 blocks. The grey areas correspond to the zero blocks. We set all off-diagonal blocks to zero except of two blocks, a bigger and a smaller one.



**Fig. 1** Precision matrix (left panel), and Adjusted p-values (right panel) for an experiment with 5 blocks, 100 variables, 100 observations. The grey color refers to zeros.

The right panel of Figure 1 shows the adjusted p-value for one instance of simulated data, with a sample size  $n = 100$ . The test is able to correctly identify the bigger block, while the smaller one is not identified, possibly due to its smaller size.

Data with precision matrix displayed in Figure 1 are simulated 50 times, to assess the performances of the procedure in this setting. At each repetition, sensitivity (proportion of correctly detected non-zero entries), specificity (proportion of correctly

detected zero entries) and accuracy on the precision matrix are computed. With a rather low sample size ( $n = 30$ ), the mean accuracy (sd) was equal to 0.87(0.07), that is a quite good value. The test tends to be rather conservative: indeed its mean (sd) specificity was 1.00(0.00), while the mean sensitivity is lower, being 0.82(0.09). This is due to the relatively strong control of the FWER that is imposed. However, the number of truly detected blocks is quite high as well. When  $n$  increases, the performances are improved; for instance with  $n = 100$ , the mean accuracy grows up to 0.94(0.08) and sensitivity to 0.91 (0.10).

Preliminary simulations (not fully reported here for brevity) suggest that the power of the procedure is affected by the sample size: when increasing the sample size, both sensitivity and specificity tend to increase, that is quite natural. Interestingly, the power is also affected by the number and size of blocks: smaller blocks tend to be more difficult to detect, leading to a decrease in sensitivity (as in the example shown in Figure 1). Further, if the number of blocks increase, the procedure also tends to be less powerful, with in particular a lower specificity.

## References

1. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3) 432-441 (2008)
2. Freedman, D., Lane, D.: A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Stat.*, **1**(4), 292–298 (1983)
3. Xia Y., Cai T., Cai T.T.: Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions *J. Am Stat. Assoc.*, **113**(521), 328–339 (2018)
4. Holm, S.: A simple sequentially rejective multiple test procedure *Scand. J. Stat.*, **6**(2), 65–70 (1979)
5. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. Roy. Stat. Soc. B*, **57** 289–300 (1995)
6. Pesarin, F., Salmaso, L.: *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons (2010)
7. Pini, A. and Vantini, S.: The interval testing procedure: a general framework for inference in functional data analysis *Biometrics* **72**(3) 835–845 (2016)
8. Ramsay, J. O. and Silverman, B. W.: *Functional data analysis*. Springer, New York (2005)
9. Abramowicz, K., Häger, C.K., Pini, A., Schelin, L., Sjöstedt de Luna, S., Vantini, S.: Non-parametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scand. J. Stat.* **45**(4) (2018)
10. Aneiros, G., Cao, R., Fraiman, R., Genest, C., Vieu, P.: Recent advances in functional data analysis and high-dimensional statistics. *J. Multivariate Anal.* **170**, 3–9 (2019)
11. Goia, A., Vieu, P.: An introduction to recent advances in high/infinite dimensional statistics. *J. Multivariate Anal.* **146**, 1–6 (2016)
12. Horváth, L., Kokoszka, P.: *Inference for functional data with applications*, vol. 200. Springer Science & Business Media (2012)
13. Pini, A., Vantini, S.: Interval-wise testing for functional data. *J. Nonparametr. Stat.* **29**(2), 407–424 (2017)

# **Analysing contributions of ages and causes of death to gender gap in life expectancy using functional data analysis**

## ***Analisi dei contributi per età e cause di morte alla differenza di genere nell'aspettativa di vita attraverso l'analisi dei dati funzionali***

Alessandro Feraldi, Virginia Zarulli, Stefano Mazzuco and Cristina Giudici

**Abstract** The work consists of application of functional data analysis (FDA) to demographic data: it analyses the contribution of ages and causes of death to gender gap in life expectancy in 14 European and non-European countries between 1998 and 2016. Causes-of-death data and life tables were retrieved from the Human Causes-of-Death Database (HCD) and from the Human Mortality Database (HMD). Our analysis allows to identify two main components that capture most of the variability and which captures the extent of the cause-specific gender differences and the age pattern, respectively. Over time, an increase in the most relevant contributions is observed, especially around the modal age and a shift of the contributions towards older age.

**Abstract** Il lavoro consiste in un'applicazione dell'analisi dei dati funzionali (FDA) a dati demografici: si analizza il contributo delle età e cause di morte alle differenze di genere nella speranza di vita in 14 paesi europei ed extraeuropei nel periodo compreso tra il 1998 e il 2016. I dati sui decessi per causa provengono dallo Human Cause-of-Death Database (HCD), mentre le tavole di mortalità sono tratte dallo Human Mortality Database (HMD). L'analisi consente di individuare due componenti principali che colgono gran parte della variabilità e che descrivono rispettivamente l'entità delle differenze di genere specifiche per causa e i contributi età-specifici. Nel tempo, si osserva un aumento dei contributi più rilevanti soprattutto intorno all'età modale ed uno spostamento degli stessi verso l'età avanzata.

**Key words:** gender gap, life expectancy, causes of death, functional data analysis

---

Alessandro Feraldi, Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro, 5 (00185) Rome, e-mail: [alessandro.feraldi@uniroma1.it](mailto:alessandro.feraldi@uniroma1.it)

Virginia Zarulli, Interdisciplinary Centre on Population Dynamics, CPOP, University of Southern Denmark. Campusvej 55. Odense M DK-5230, e-mail: [vzarulli@sdu.dk](mailto:vzarulli@sdu.dk)

Stefano Mazzuco, Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy, email: [stefano.mazzuco@unipd.it](mailto:stefano.mazzuco@unipd.it)

Cristina Giudici, MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161 Rome, Italy, e-mail: [cristina.giudici@uniroma1.it](mailto:cristina.giudici@uniroma1.it)

## 1 Introduction

On average and worldwide, women live longer than men and the absolute difference between male and female mortality risk reaches its maximum at old ages. From the beginning of the 1920s, in most industrialized countries the gap in life expectancy between the two sexes widened until the 1970s, when the difference started to narrow (Austad (2006); Zarulli et al. 2020)). The literature shows that where some convergence has taken place, men have experienced more rapid gains in survival than women (Meslé & Vallin (2011)).

To explain the recent narrowing in the sex gap in LE, several studies have focused on the causes of death that contributed to the gender gap in mortality rates and, thereby, either narrowed or widened this gap (Klenk et al. (2016)). In most of the studies, the topic of the variability in the gender gap in mortality has usually been tackled comparing age trajectories of cause-specific death rates between men and women by fitting specific parametric models on cause-specific life table death rate for women and men separately (Horiuchi et al. (2013)). Other studies used life table or aggregated mortality indicators to provide summary measures of mortality levels (e.g. life expectancy) and dispersion (e.g. lifespan variation) both separately for men and women and with the decomposition of the difference between sexes according to age and causes of death (Trias-Llimós & Janssen (2018)).

Although numerous studies have decomposed the sex gap in life expectancy according to age and causes of death, they did not study the principal components of the contributions of age and causes of death from a functional perspective, which has been shown to be more informative approach (Ramsay & Silverman (2002); Léger & Mazzuco (2020)). To fill this gap, we study absolute and relative contributions of age and causes of death to the gender gap in life expectancy (GGLE) for several countries, using the Functional Data Analysis (FDA) (Ramsay & Silverman (2002)). Following this approach, we consider age- and cause-specific contributions as functions, and therefore we analyse curves rather than scalar data. More specifically we propose a Functional Principal Component Analysis (FPCA) of the contribution profiles of several countries, in order to identify the main components of the distribution of age-specific contributions according to causes of death. To the best of our knowledge, this is the first study analysing age- and cause-specific contributions to the gender gap in life expectancy with a functional data analysis approach.

## 2 Data and Method

### 2.1 Data

Cause-specific mortality data were retrieved by gender, 5-year age interval and year from the Human Cause-of-Death Database (HCD (2019)) and life tables were taken from the Human Mortality Database (HMD (2019)). HMD and HCD are open-source projects: the former offers harmonized data on constructed series of mortality rates,

## Analysing contributions of age and cause of death to the gender gap in life expectancy using functional data analysis

life tables, death counts and population exposures; the latter contains reconstructed long-term trends in cause-specific mortality for sixteen countries over time. Since this study aims at analysing patterns of causes of death across several countries, we focussed on the last 15 years available for each country, within the period 1998-2016. Countries were grouped according to geographical areas: Eastern Europe (EE) (i.e. Russia, Ukraine, Poland, the Czech Republic, Estonia, Latvia, Lithuania and Belarus); Western Europe (WE) (i.e. France, Spain, Germany and the United Kingdom (UK)), and extra-European countries (i.e. Japan and the United States (US)). Romania and Moldova were excluded from the analysis because no data were available in the HMD. Within this time frame all the causes of death are coded in each country according to the 10th version of International Classification of Disease (ICD-10). This allows us to avoid problems related to differences in the classification and to obtain comparable information for all the countries under study. We further restricted our focus on the short list of the ICD-10, in which all the causes are grouped into sixteen major categories, each including a set of similar diseases (e.g. heart diseases, neoplasms, external causes, respiratory diseases etc.) (HCD (2019)). Finally, cause-specific mortality data for all the ages above 85 were grouped in the open-end age interval 85+, to avoid problems related to the data quality, which are particularly common at very old ages.

### 2.2 Analysis

Age- and cause-specific contributions to the GGLE (female - male) were obtained for each country and over time applying Arriaga's age- and cause-specific decomposition technique combining life tables from the HMD and cause-specific mortality data from the HCD (Arriaga (1984)). FDA was applied to the age-specific relative contributions to the GGLE, separately for each cause of death. Discrete age-specific relative contribution data  $x(t_1), \dots, x(t_N)$  were assumed to be independent realizations drawn from the same continuous stochastic process  $X(t)$  (Ramsay & Silverman (2002)). To obtain the functional representation, each  $X(t)$  was approximated by using a basis expansion of cubic B-splines functions (1) and the B-splines basis coefficients were estimated by the ordinary least squares method minimizing the sum of squared residuals (Léger & Mazzuco (2020)). Therefore,

$$X(t) = \sum_{j=1}^p \gamma_j \psi_j(t) \quad (1)$$

where  $\psi_j$  are  $p$  known basis functions and  $\gamma_j$  are the corresponding coefficients to be estimated. In order to maintain the data structure, we used a sequence of  $p = 19$  equally distributed knots (i.e. one for 5-years age interval). Afterwards, we performed FPCA separately for each cause of death in order to synthesize the variability of the curves and to identify the main components of the distributions of age-specific contributions according to causes of death, across countries. FPCA is the extension of the more classical multivariate PCA to functional data: for a generic curve  $x_i(t)$  we can obtain the approximation (2)

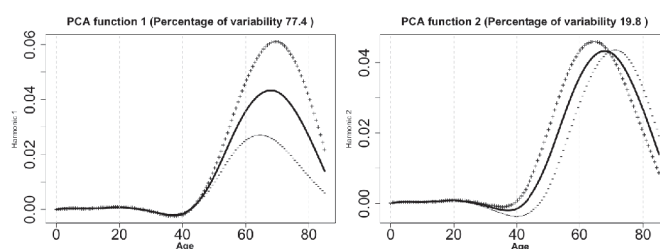
$$x_i(t) = \sum_{k=1}^{\infty} c_{i,k} \phi_k(t) \quad (2)$$

where  $c_{i,k}$  are the principal component scores and  $\phi_k(t)$  are the eigenfunctions or harmonics. Therefore, the information on the curve  $x_i(t)$  were then synthesized by the first  $q$  terms. All the analyses were conducted using the R package *fda* (Ramsay et al. (2011)).

### 3 Results

Decomposition results confirmed that neoplasms, heart diseases and external causes of death made the largest contributions to the GGLE in all the countries, explaining together more than two third of the overall gap. Additionally, the largest contributions to the GGLE were given by old ages for most of the causes of death over the entire period (Meslé & Vallin (2011); Trias-Llimós & Janssen (2018)). Results of FPCA focus on the three most relevant causes of death.

Most of the variability in the age-specific contributions is explained by the first two principal components for each cause of death (e.g. 97%, 94% and 95% for neoplasm, heart diseases and external causes, respectively). The first FPC mainly captures the extent of the cause-specific gender differences, while the second FPC captures the age pattern. A classical way to interpret the FPCs is to plot the group mean function (solid curve in Figure 1) as well as the functions obtained by adding (+ curve) and subtracting (- curve) to the mean function twice the square root of the principal component variance (Ramsay & Silverman (2002); Léger & Mazzuco (2020)). Regarding neoplasm (Figure 1), for the first FPC, the variability is concentrated at 40 years' age and older and especially around the modal age. A high score on this component suggests an above-average contribution. The second FPC corresponds to a shift of the curves with respect to the overall contribution mean towards older ages. The (+) curve has a higher contribution than the (-) curve with respect to the mean curve before 70 years, lower afterwards. A low score on this component suggests an above-average shift of the distribution towards older ages.

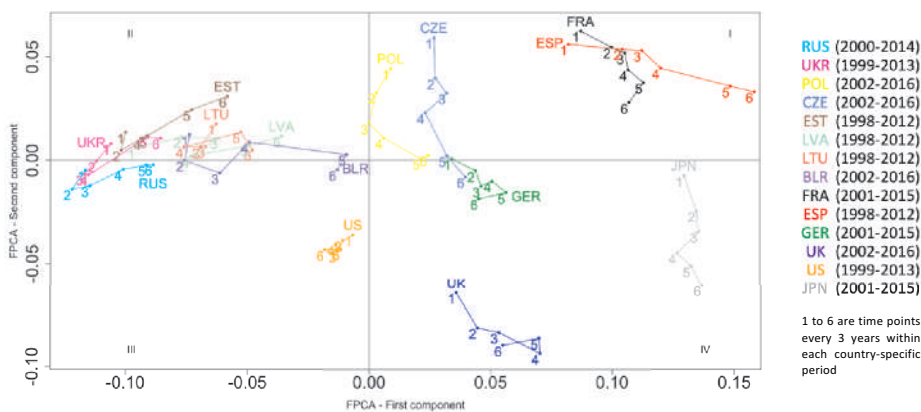


**Figure 1.** Effect of the first two FPCs on age-specific contributions to GGLE for neoplasm: overall mean and mean  $\pm$  a suitable multiple of the principal component weight function.



**Analysing contributions of age and cause of death to the gender gap in life expectancy using functional data analysis**

In order to study the evolutions of age-cause contributions over time, the scores of the two first components for each country are plotted at every 3 years. With regards to neoplasm (Figure 2), the first axis indicates that, throughout the whole period, relative contributions to the GGLE were higher in France, Spain (I quarter) and in Japan (IV quarter) than in the other countries, followed by UK, Germany, Poland, Czech Republic and US. The lowest contributions were shown in Ukraine, Estonia, Latvia, Lithuania, Belarus and in Russia (II quarter). The second axis denotes that the distribution of age-specific contributions to the GGLE was more concentrated at older ages in UK than in the other countries. Furthermore, distributions were more concentrated at older ages in US and in Japan (scores of the second FPC < 0) than in the overall mean distribution (i.e. for all the countries).



**Figure 2.** Principal component subspace - neoplasm.

On the contrary, in the remaining countries distributions were more concentrated at younger ages than in the overall mean distribution. Moreover, the increasing trends in the first FPC scores, especially in Spain, together with the decreasing trends in second FPC scores showed also in France and in UK, indicate increasing relative contribution of neoplasm over time as well as a shift of the distribution towards older ages. Similarly, in Poland, Czech Republic and in Japan, although the relative contributions of neoplasm to the GGLE stagnated over time (small variations in first FPC scores), the decreasing trends of the second FPC scores denote that the distributions of age-specific contributions shifted towards older ages over time. A small increase in the score of the first component in Estonia and in Belarus (especially at the second half of the period) suggests a slight increase in the relative contribution of neoplasm to the GGLE. Finally, stagnation in both FPC scores was reported in Germany, Russia, Ukraine and in US over time. Therefore, independent from trends in relative contributions of neoplasm to the GGLE, in most of the countries the distributions shifted towards older ages over time.

For the sake of brevity, we do not display the results of FPCA for the other two main causes, however the analysis shows similar patterns across countries and over time.

## Conclusion

This work gives a deeper insight of the main contributing factors to the gender differences in life expectancy, analysing components of the relative age-specific contributions according to the most relevant causes of death in determining the gap. The study also aims at illustrating the demographic application of FDA, a new method for demographic analysis but which could prove useful to deepen our understanding of complex demographic phenomena. Our results allow to identify two main components which capture most of the variability. The first component mainly captures the extent of the cause-specific gender differences, while the second FPC captures the age patterns. Over time, an increase in the most relevant contributions is observed, especially around the modal age and a shift of the contributions towards older age. The analysis confirms that FDA allows to highlight country-specific patterns in the context of the epidemiological transition which need to be further analysed. Further analyses include functional cluster analysis to group countries according to age-cause contributions and study the evolution of the contributions in each cluster over time. Finally, following the FDA approach we also apply other statistical analyses (i.e. regression and hypothesis tests) to the same data and suggest to increase the use of such approach in population studies.

## References

1. Arriaga, E. E. (1984). Measuring and explaining the change in life expectancies. *Demography*, 21(1), 83–96.
2. Austad, S. N. (2006). Why women live longer than men: Sex differences in longevity. *Gender Medicine. Official Journal of the Partnership for Gender-Specific Medicine at Columbia University*, 3(2), 79–92.
3. Horiuchi, S., Ouellette, N., Cheung, S. L. K., & Robine, J.-M. (2013). Modal age at death: Lifespan indicator in the era of longevity extension. *Vienna Yearbook of Population Research*, 37–69.
4. Human Cause-of-Death Database. (2019). French Institute for Demographic Studies (France) and Max Planck Institute for Demographic Research (Germany) (downloaded on 2019).
5. Human Mortality Database. (2019). Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany) (downloaded on 2019).
6. Klenk, J., Keil, U., Jaensch, A., Christiansen, M. C., & Nagel, G. (2016). Changes in life expectancy 1950–2010: Contributions from age- and disease-specific mortality in selected countries. *Population Health Metrics*, 14, 20.
7. Léger, A. E., & Mazzuco, S. (2020). What can we learn from functional clustering of mortality data? An application to HMD data. *arXiv preprint arXiv:2003.05780*.
8. Meslé, F., & Vallin, J. (2011). Historical trends in mortality. In *International handbook of adult mortality* (pp. 9–47). Springer.
9. Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. Springer Series in Statistics, Springer-Verlag.
10. Ramsay, J., Wickham, H., Graves, S., and Hooker, G. (2011). *fda: Functional Data Analysis*. R package.
11. Trias-Llimós, S., & Janssen, F. (2018). Alcohol and gender gaps in life expectancy in eight Central and Eastern European countries. *European Journal of Public Health*, 28(4), 687–692.
12. Zarulli, V., Lindahl-Jacobsen, R., & Vaupel, J. W. (2020). Onset of the old-age gender gap in survival. *Demographic Research*, 42, 727–740.

# Supervised classification of ECG curves via a combined use of functional data analysis and random forest to identify patients affected by heart disease

*Classificazione supervisionata delle curve ECG tramite un uso combinato dell'analisi dei dati funzionali e delle foreste casuali per identificare pazienti affetti da malattie cardiache*

Fabrizio Maturo, Rosanna Verde

**Abstract** Cardiovascular disease (CVD) is a significant cause of death and disability for individuals from different background, age, gender, race, income groups, and countries. Indeed, the Global Burden of Disease Study (2013) calculated that CVD causes approximately 30% of all deaths worldwide, and thus CVD prevention is a major public health problem. One of the most used methods for monitoring the heart is the electrocardiogram (ECG) and, in recent years, many apps have also been developed for the constant tracking of the cardiac condition of patients. All these approaches produce an electrical signal that can provide indispensable indications to prevent heart attacks and strokes. This study proposes a supervised classification method that is based on the joint use of functional data analysis and random forest to identify and classify patients at risk.

**Abstract** *Le malattie cardiovascolari (CVD) sono una causa significativa di morte e disabilità per individui provenienti da background, età, sesso, razza, gruppi di reddito e paesi diversi. Infatti, il Global Burden of Disease Study (2013) ha calcolato che la CVD causa circa il 30% di tutti i decessi nel mondo, e quindi la prevenzione delle CVD è un importante problema di salute pubblica. Uno dei metodi più utilizzati per il monitoraggio del cuore è l'elettrocardiogramma (ECG) e, negli ultimi anni, sono state sviluppate anche molte app per il monitoraggio costante delle condizioni cardiache dei pazienti. Tutti questi approcci producono un segnale elettrico in grado di fornire indicazioni indispensabili per prevenire infarti ed ictus. Questo articolo propone un metodo di classificazione supervisionato che si basa sull'uso congiunto dell'analisi dei dati funzionali e delle foreste casuali per identificare e classificare i pazienti a rischio.*

---

Fabrizio Maturo

University of Campania Luigi Vanvitelli, Caserta, Italy e-mail: [fabrizio.maturo@unicampania.it](mailto:fabrizio.maturo@unicampania.it)

Rosanna Verde

University of Campania Luigi Vanvitelli, Caserta, Italy e-mail: [rosanna.verde@unicampania.it](mailto:rosanna.verde@unicampania.it)

**Key words:** FDA, functional random forest, supervised classification, ECG curves, heart disease.

## 1 Introduction

*Cardiovascular disease (CVD)* (often used interchangeably with the term “*Heart disease*”) is one of the most significant determinants of morbidity and death among both worldwide women and men. Therefore, the prediction of cardiovascular disorders is one of the most critical problems in medicine and biostatistics. Indeed, identifying in advance patients affected by heart disease can prevent serious consequences, e.g. stroke and heart attack. There are many methods of controlling people with suspected heart disease. However, the goal of this article is not to provide a complete understanding of the phenomenon and its determinants, and thus we focus only on ECG to propose a methodological approach to treat this data. ECG reads heart’s electrical impulses and may be used to diagnose a heart attack or abnormal heart rhythms (called ‘*arrhythmias*’). An ECG is just a representation of the electrical activity of the heart muscle, habitually printed on paper for more straightforward interpretation. A primary characteristic of the ECG is that the electrical activity of the heart is shown as it varies with time. Thus, it provides a graph, plotting electrical activity on the vertical axis against time on the horizontal axis. It is clear that the resulting data is not a simple set of scalar observations, but a real function that depends on time. Hence, treating ECG signals with functional data analysis (FDA) [1, 2], which considers this function as a single entity, appears to be the most natural way to deal with kind of data, and moreover, may also provide useful additional indications when considering the derivatives of the original signals. The smart idea of using FDA to treat this type of data was developed in different previous works with diverse purposes (e.g. see [3, 4, 5]).

The novelty of this contribution is to suggest the joint use of FDA and machine learning in order to make a supervised classification of the ECG curves. In particular, FDA is used in combination with Random Forest (RF). RF [6] is one of the most efficient machine learning algorithms and is a particular case of bagging for decision trees. It consists of applying bagging to the data and bootstrap sampling to the predictor variables at each split. This implies that at each splitting step of the tree algorithm, a random sample of  $n$  predictors is selected as split candidates from the full set of the predictors. This leads to an improvement of the classic bagging because it allows to obtain a classifier that is not strongly influenced by the correlation among trees, which otherwise would all be dominated by the most discriminating variable. RF can be adapted to the FDA framework, both in the case that the functions are obtained by smoothing high frequency data in the time domain and in the event that the functions depend on other specific parameters. In this work, we propose Functional Random Forest (FRF) considering ECG curves as functional predictors of heart disease.

## 2 Material and Methods

FDA has become widespread during the last decades and now is a primary research area in statistics. The basic idea of this method is to handle data functions as single objects. Nevertheless, in practical applications, functional data are often observed as series of point data, and thus the function expressed by  $z = f(x)$  reduces to record of discrete observations that are denoted by the  $T$  pairs  $(x_j; z_j)$  where  $x \in \mathfrak{X}$  and  $z_j$  are the values of the function computed at the points  $x_j$ ,  $j = 1, 2, \dots, T$  [1]. Generalizing the reference framework, we consider that a functional variable  $X$  is a random variable assuming values in a functional space  $\xi$ , and a functional data set is a sample  $x_1, \dots, x_N$ , also denoted  $x_1(t), \dots, x_N(t)$ , drawn from a functional variable  $X$  [2]. The first step in FDA is to convert the observed values  $z_{i1}, z_{i2}, \dots, z_{iT}$  for each unit  $i = 1, 2, \dots, N$  to a functional form. The most common approach to estimate the functional datum is the basis approximation. The basic idea is that functions can be obtained using a finite representation in a fixed basis [1]. Limiting our attention to the  $\mathcal{L}_2$  context (see [1, 7, 8] for more details), a function  $x(t)$  can be expressed by a linear combination of these basis functions as follows:

$$x(t) = \sum_{j \in \mathbb{N}} c_j \phi_j(t) \approx \sum_{j=1}^K c_j \phi_j(t) = \hat{x}(t) \quad (1)$$

where  $c$  is the vector of coefficients defining the linear combination and  $\phi_j(t)$  is the  $j$ -th basis function, from a subset of  $K < \infty$  functions that can be used to approximate the full basis expansion.

Exploiting the coefficients of a fixed basis system like those in Equation 1, the Decision Tree (DT) and RF approaches can be extended to the case of functional data of the form  $\{y_i, x_i(t)\}$ , with a predictor curve  $x_i(t)$ ,  $t \in J$ , and  $y_i$  being the (scalar) response value observed at sample  $i = 1, \dots, n$ . The response variable could be either numeric or categorical, leading to regression or classification trees, respectively; however, here we focus on the case of a binary dependent variable and thus we concentrate on functional classification trees, particularly on the scalar-on-function classification problem. Classification trees consist in recursive binary partitions of the feature space into rectangular regions (terminal nodes or leaves). To build the tree, an optimal binary partition is selected at each step of the algorithm, based on an optimization criterion. The algorithm begins with the full data set composed of the coefficients obtained in Equation 1 and continues until the leaves are obtained. Having obtained the best split in one node, the data are partitioned into the respective regions and we replicate the rule of finding the most suitable binary separation on all resulting nodes. Typically, a huge tree is produced at the beginning, which is then pruned according to an optimization criterion.

Therefore, the coefficients of the linear combination are used as new features to predict the response. The interpretation is slightly different with respect to the classical DT because the values of the splits of  $c_j$  should be interpreted according to the part of the domain that the single b-spline  $\phi_j(t)$  mostly represent. Hence, the joint read of the coefficients and of the plot of  $\phi_j(t)$  can help interpreting the classi-

fication tree. The great problem with a single tree is that its predictive performance is usually not persuasive, and modest changes in the data may lead to very diverse trees. A useful technique to reduce this kind of variance is to create an ensemble of trees using the RF approach [6]. The idea of FRF is quite recent. Few papers are available in the literature and the approaches are considerably different. For example, Möller et al. [10] propose an approach based on the mean of the function within fixed intervals of the domain whereas El Haouij et al. [11] and also Gregorutti et al. [12] focus on the wavelet basis decomposition. Our approach is quite different and is based on the b-spline decomposition.

Assume the FRF consists of  $H$  trees  $\tau_h$ ,  $h = 1, \dots, H$ , where  $H$  is chosen to be a large number, such as  $H = 200$ . The  $h$ -th tree  $\tau_h$  is grown on a random subset of the training set, obtained from the original data  $D = \{(y_i, x_i(t)), i = 1, \dots, n\}$  by drawing, with replacement, a bootstrap sample  $D_h^* = \{(y_s^{(h)}, x_s^{(h)}(t)), s = 1, \dots, n\}$  of the same size  $n$  as the original data set. It is straightforward to replace  $x_s^{(h)}(t)$  using its expansion in term of b-spline basis as in Equations 1. Thus, the data points  $s = 1, \dots, n$  present in the  $h$ -th bootstrap sample  $D_h^*$  are called an in-bag sample, on which the  $h$ -th tree will be grown. Instead, the out-of-bag (OOB) sample is composed of the remaining data points  $\{y_i, x_i(t)\}$  that are not present in  $D_h^*$ . Thus, we construct  $H$  decision trees using  $H$  bootstrapped training sets, and we average the resulting predictions. Because each tree is grown deep and is not pruned, each tree has low bias, but high variance. Averaging these  $H$  trees diminishes the variance. This is what we can call the phase of “*functional bagging*” (FB) and gives gains in accuracy with respect to a single DT because it combines hundreds or thousands of trees. For a given test observation, we register the class predicted by each of the  $H$  trees, and take a “majority vote”. Consequently, the overall prediction is the most commonly occurring class among the  $H$  forecasts. Expanding the number of trees  $H$  will not lead to overfitting. In practice, we want to use a value of  $H$  that is large enough for the test error to have settled down. Now, suppose that after the expansion computed using Equation 1, we observe that there are some moderately strong predictors but there is one very strong predictor in the training data; in our case, for example, one basis can explain a specific part of the domain and be dominant in resolving the final classification of the original curves. In FB, most or all of the individual trees will use this powerful predictor in the top split. Consequently, all bagged trees will look quite similar to each other, so the predictions from these DTs will be highly correlated. Averaging highly correlated scores leads to a smaller decrease in variance than averaging uncorrelated quantities. Therefore, FB will not lead to a tangible reduction in variance over a single tree.

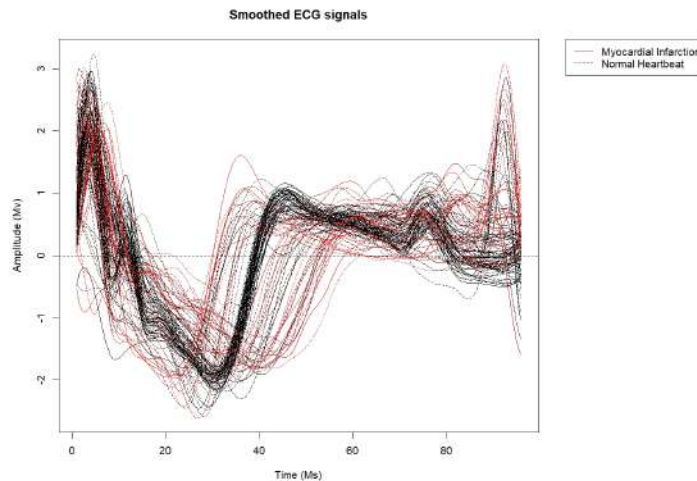
Functional Random Forest (FRF) gives an improvement over FB because it involves a small tweak that decorrelates the trees. At each split in the tree-building process, we consider a random sample of  $\pi$  predictors,  $\pi < K$ , as candidates for the split, where  $K$  is the total number of b-spline basis (see Equation 1). A new sample of  $\pi$  predictors is taken at each split, for example of size  $\pi \approx \sqrt{K}$ . Therefore, at each split in the tree, the algorithm is not even allowed to consider a majority of the available b-spline coefficients. Indeed, on average,  $\frac{\pi-K}{\pi}$  of the splits will not even contemplate some predictors. In this way, FRF decorrelates the trees, making

Supervised classification of ECG curves via a combined use of FDA and RF...

the average of the trees less variable and hence more reliable. Thus, the difference between FB and FRF depends on the choice of  $\pi$ . When  $\pi = K$ , FRF is equivalent to FB.

### 3 Application and Conclusions

The proposed approach is applied to a real dataset formatted by R. Olszewski as part of his thesis at Carnegie Mellon University, 2001. Each series traces the electrical heart activity of 200 patients recorded during one heartbeat [9]. Our goal is to build a model to predict the two classes that are a normal heartbeat and a Myocardial Infarction (MI). Figure 1 illustrates the smoothed versions of the original signals computed using Equation 1.



**Fig. 1** Smoothed ECG signals of patients.

The black signal are the patients with a normal heartbeat whereas the red curves are those who have had a diagnosis of heart disease.

For the sake of brevity, we limit ourselves to saying that the sample was divided into a training sample and a test sample both of equal size. The percentage of cases correctly classified is 85% on the training set and 80% on the test set, respectively.

We are aware of the limitations of the study due to the approximation in identifying the disease. A more in-depth study that considers some covariates or a multivariate FRF would be desirable. Nonetheless, this article is a first step towards defining an advanced method for classifying patients with the disease in question. Future re-

search will focus on the use of the functional principal components decomposition of the original functional data as a fixed basis system to implement FRF as well.

## References

1. Ramsay J., Silverman B.: Functional data analysis. Springer, New York (2005).
2. Ferraty F., Vieu P.: Nonparametric functional data analysis. Springer, New York (2006).
3. Zhou, Y., Sedransk, N.: Functional data analytic approach of modeling ECG T-wave shape to measure cardiovascular behavior. *The Annals of Applied Statistics*, 3, (2010).
4. Ieva, F., Paganoni, A., Pigoli, D., Vitelli, V.: Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 62(3), 401-418, (2013).
5. Cammarota, C. Curione, M.: Trend Extraction in Functional Data of Amplitudes of R and T Waves in Exercise Electrocardiogram. *Fluctuation and Noise Letters*, 16(02), (2017).
6. Ho T.K.: The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832-844 (1998).
7. Aguilera, A., Aguilera-Morillo, M.: Penalized PCA approaches for b-spline expansions of smooth functional data. *Applied Mathematics and Computation* 219, pp. 7805-7819, (2013).
8. Febrero-Bande, M., de la Fuente M.: Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51, pp. 1-28, (2012).
9. Olszewsk, R.: UCR Time Series Classification Archive (accessed on 1st of March, 2020), [https://www.cs.ucr.edu/~eamonn/time\\_series\\_data/](https://www.cs.ucr.edu/~eamonn/time_series_data/).
10. Moller, A., Tutz, G., Gertheiss, J.: Random forests for functional covariates. *Journal of Chemometrics*, 30(12), 715-725, (2016).
11. El Haouij, N., Poggi, J.M., Ghazi, R., Sevestre-Ghalila, S.: Jaïdane, M. Random forest-based approach for physiological functional variable selection for driver's stress level classification. *Statistical Methods & Applications*, 28(1), 157-185, (2018).
12. Gregorutti, B., Michel, B., Saint-Pierre, P.: Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90, 15-35, (2015).



## 4.14 Mixture models

# Alternative parameterizations for regression models with constrained multivariate responses

## *Parametrizzazioni alternative per modelli di regressione con risposte multivariate vincolate*

Roberto Ascari, Agnese Maria Di Brisco, Sonia Migliorati, and Andrea Ongaro

**Abstract** The extended flexible Dirichlet regression model has been recently proposed as a tool for modeling multivariate constrained responses. Being a special finite mixture, it displays a far richer dependence structure and a wider variety of shapes than the Dirichlet regression model. Moreover, it defines several group regression curves - one for each mixture component - which improve interpretation issues. Nonetheless, these curves may display non-smooth shapes which real datasets generally do not show. For this reason, we propose two alternative parameterizations able to fix this aspect, and we compare them both from an analytic and from a simulative point of view.

**Abstract** *Il modello di regressione basato sulla distribuzione extended flexible Dirichlet è stato recentemente proposto per modellare risposte multivariate vincolate. Essendo una speciale mistura finita, mostra una struttura di dipendenza ed una varietà di forme più ricche rispetto al modello Dirichlet. Inoltre, esso consente di costruire differenti curve di regressione per i gruppi individuati dalle componenti della mistura, curve molto utili a fini interpretativi. Tuttavia, tali curve presentano dei punti angolosi che difficilmente si riscontrano nei dati reali. Per questo motivo, due parametrizzazioni alternative vengono proposte e confrontate sia da un punto di vista analitico sia tramite uno studio simulativo.*

**Key words:** simplex, mixture model, compositional regression, bayesian inference.

---

Ascari Roberto, Sonia Migliorati, and Andrea Ongaro  
Univerisità Milano-Bicocca, Department of Economics, Management and Statistics.  
e-mail: roberto.ascari@unimib.it, sonia.migliorati@unimib.it, andrea.ongaro@unimib.it

Di Brisco Agnese Maria  
Univerisità del Piemonte Orientale, Department of Studies in Economics and Business.  
e-mail: agnese.dibrisco@uniupo.it

## 1 Introduction

Compositional data, namely proportions of some whole, are defined on the  $D$ -part simplex  $\mathcal{S}^D = \{\mathbf{Y} : Y_j > 0, j = 1, \dots, D, \sum_{j=1}^D Y_j = 1\}$ . A novel distribution for a  $D$ -dimensional vector on  $\mathcal{S}^D$  is the extended flexible Dirichlet (EFD) distribution [6], which can be expressed as a structured finite mixture with Dirichlet components. Indeed, the distribution function of an EFD-distributed random vector  $\mathbf{Y}$  admits the following representation:

$$\text{EFD}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{r=1}^D p_r \text{Dir}(\mathbf{y}; \boldsymbol{\alpha} + \tau_r \mathbf{e}_r), \quad (1)$$

where  $\text{Dir}(\cdot; \cdot)$  denotes the Dirichlet distribution function,  $\mathbf{y}$  and  $\mathbf{p}$  lie in  $\mathcal{S}^D$ ,  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)'$ ,  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_D)'$ ,  $\alpha_r > 0$ ,  $\tau_r > 0$ , and  $\mathbf{e}_r$  is a vector of zeros except for the  $r$ -th element which is equal to one. Its probability density function (p.d.f.) can be written as:

$$f_{\text{EFD}}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \left( \prod_{r=1}^D \frac{y_r^{\alpha_r-1}}{\Gamma(\alpha_r)} \right) \sum_{h=1}^D p_h \frac{\Gamma(\alpha_h) \Gamma(\alpha^+ + \tau_h)}{\Gamma(\alpha_h + \tau_h)} y_h^{\tau_h}, \quad (2)$$

where  $\alpha^+ = \sum_{r=1}^D \alpha_r$ . The EFD distribution contains the Dirichlet as an inner point when  $\tau_r = 1$  and  $p_r = \alpha_r / \alpha^+$  for every  $r = 1, \dots, D$ . Relevant properties of the EFD distribution include the large variety of shapes of its p.d.f., including uni- and multi-modal ones, as well as the flexible modelization of the dependence structure of the composition, thereby overcoming the drawbacks entailed by the Dirichlet distribution [1, 6]. To define the extended flexible Dirichlet regression (EFDReg) model [7] it is useful to choose an alternative parameterization that explicitly includes the mean vector  $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}$ . This can be obtained by noting that the  $r$ -th mixture component in (1) has a mean vector  $\boldsymbol{\lambda}_r$  equal to:

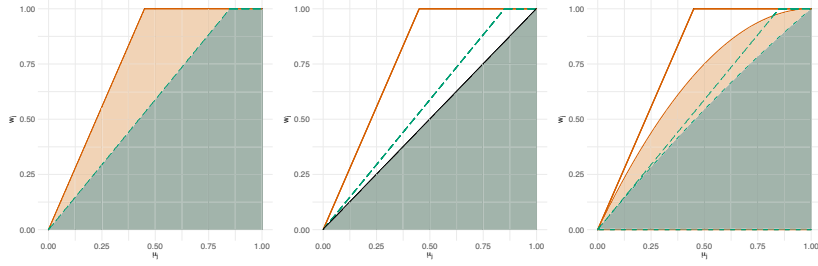
$$\boldsymbol{\lambda}_r = (1 - w_r) \bar{\boldsymbol{\alpha}} + w_r \mathbf{e}_r, \quad (3)$$

where  $\bar{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}}{\alpha^+}$  and  $w_r = \frac{\tau_r}{\alpha^+ + \tau_r}$ . Thus, the generic element  $\mu_j$  of the EFD mean vector  $\boldsymbol{\mu}$  takes the form:

$$\mu_j = \mathbb{E}[Y_j] = \sum_{r=1}^D p_r \lambda_{rj} = \bar{\alpha}_j \sum_r p_r (1 - w_r) + p_j w_j \quad j = 1, \dots, D. \quad (4)$$

Note that the parameterization based on  $\mu_j$ ,  $p_j$ , and  $w_j$  is not variation independent, i.e. some constraints exist among these parameters, as shown in the left panel of Fig. 1. This could cause inferential problems, in particular within a Bayesian approach (in the prior elicitation phase as well as in the Markov Chain Monte Carlo (MCMC) based posterior computation). Indeed, it is possible to show that  $w_j < \min \left\{ \frac{\mu_j}{p_j}, 1 \right\}$  (see [7] for details), so that we can define the normalized version

Title Suppressed Due to Excessive Length



**Fig. 1** Parameter space of  $w_j$  as  $\mu_j$  varies for  $p_j = 0.45$  (red area) and  $p_j = 0.85$  (green area). EFDReg (left panel), linear (middle panel) and quadratic (right panel) parametrizations.

of  $w_j$  as:

$$\tilde{w}_j = \frac{w_j}{\min\left\{\frac{\mu_j}{p_j}, 1\right\}} \Rightarrow w_j = \tilde{w}_j \cdot \min\left\{\frac{\mu_j}{p_j}, 1\right\}, \text{ for } j = 1, \dots, D. \quad (5)$$

We can now derive a regression model based on the EFD distribution parameterized by  $\boldsymbol{\mu} \in \mathcal{S}^D$ ,  $\mathbf{p} \in \mathcal{S}^D$ ,  $\tilde{w}_j \in (0, 1)$  ( $j = 1, \dots, D$ ), and  $\alpha^+ > 0$ . Let us denote the response and the covariate vectors for unit  $i$  ( $i = 1, \dots, n$ ) by  $\mathbf{Y}_i \in \mathcal{S}^D$  and  $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iK})'$  respectively. Following a GLM strategy [5], we adopt a multinomial logit link function to link the mean vector  $\boldsymbol{\mu}_i$  to the linear predictor as follows:

$$g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{\mu_{iD}}\right) = \mathbf{x}_i' \boldsymbol{\beta}_j, \quad j = 1, \dots, D; i = 1, \dots, n, \quad (6)$$

where  $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jK})'$  is a vector of regression coefficients. It is worth noting that the  $D$ -th category is fixed as baseline, therefore  $\beta_{Dk} = 0$  for  $k = 0, 1, \dots, K$ . As a consequence, we get:

$$\mu_{ij} = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}_j) = \begin{cases} \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta}_j)}{1 + \sum_{r=1}^{D-1} \exp(\mathbf{x}_i' \boldsymbol{\beta}_r)}, & \text{for } j = 1, \dots, D-1 \\ \frac{1}{1 + \sum_{r=1}^{D-1} \exp(\mathbf{x}_i' \boldsymbol{\beta}_r)}, & \text{for } j = D. \end{cases} \quad (7)$$

Very interestingly, the EFDReg model enables the construction of  $D$  group regression curves given by (3).

Inferential issues for the EFDReg model are dealt with by a Bayesian approach. We adopt the Hamiltonian Monte Carlo (HMC) [8] algorithm, which is a generalization of the Metropolis algorithm combining MCMC and deterministic simulation methods. The HMC algorithm is easily implemented in the Stan modeling language [9]. Concerning priors elicitation, we propose to take advantage of non- or weakly informative priors to induce the minimum impact on the posteriors [2]. We set a multivariate normal prior for the regression parameters  $\boldsymbol{\beta}_j$  with zero mean vector and diagonal covariance matrix with “large” values of the variances to induce

vagueness. Furthermore, we select a Uniform(0, 1) prior for  $\tilde{w}_j$ ,  $j = 1, \dots, D$ , and a Dirichlet prior with hyperparameter equal to  $\mathbf{1}$  for the vector  $\mathbf{p}$ . Last, we adopt a Gamma( $g, g$ ) prior (with  $g$  taking “small” values) for  $\alpha^+$ .

## 2 Alternative parameterizations

In order to give rise to smooth group regression curves, we propose two normalization approaches alternative to (5). More specifically, we aim at finding polynomial functions that approximate the behavior of  $\min\left\{\frac{\mu_j}{p_j}, 1\right\}$  as better as possible when  $\mu_j$  and  $p_j$  belong to the unit interval. We consider both linear and quadratic functions, imposing the following constraints:

- when  $\mu_j = 0$  (or 1) their value is 0 (or 1) for any fixed  $p_j$ ,
- they are increasing functions,
- the selected quadratic function must be uniformly higher than any other quadratic function for  $\mu_j \in (0, 1)$  and fixed  $p_j$ .

With some algebra, we can obtain the following expressions for  $w_j$ :

$$(i) w_j = \tilde{w}_j^L \cdot \mu_j \quad (ii) w_j = \begin{cases} \tilde{w}_j^Q \left( \frac{1}{p_j} \mu_j + \left(1 - \frac{1}{p_j}\right) \mu_j^2 \right), & \text{if } p_j \geq 0.5 \\ \tilde{w}_j^Q \left( 2\mu_j - \mu_j^2 \right), & \text{if } p_j < 0.5 \end{cases} \quad (8)$$

where  $\tilde{w}_j^L$  and  $\tilde{w}_j^Q$  are the normalized version of  $w_j$  under the linear and quadratic constraints, respectively.

As we can see in the middle and right panels of Fig. 1, the quality of the approximations heavily depends on the values of  $\mu_j$  and  $p_j$ . In particular, when  $p_j$  goes to one, (i) and (ii) are equal to (5) for any  $\mu_j$ . On the other hand, the lower  $p_j$ , the worst (i) approximates the original constraint (5), since the former does not depend on  $p_j$ . Even if the expression of (ii) does not depend on  $p_j$  when  $p_j < 0.5$ , it is uniformly higher than the linear approximation (i). Finally, when  $0.5 < p_j < 1$ , (ii) still does provide a better approximation than (i). However, it is easy to note that also the quadratic expression does not cover the entire parameter space of  $w_j$ , thus leading to a parameterization that excludes a subset of the original parameter space. Though, such an exclusion does not necessarily imply a worse fit in general, as it will be shown. With the aim of comparing the fit of the EFDReg model under the three competing parameterizations of  $w_j$ , we show the results of a simulation study based on the following data generating process and on 500 replications. For each observation, we generated a quantitative covariate  $x_i$  from a uniform distribution on  $(0.5, -0.5)$ , and a response vector  $\mathbf{y}_i$  ( $i = 1, \dots, 150$ ) from a Dirichlet distribution with  $D = 3$ , precision  $\alpha^+ = 50$  and mean vector given by Eq. (7) with  $\boldsymbol{\beta}_1 = (1, 2)'$  and  $\boldsymbol{\beta}_2 = (0.5, -3)'$ . Then, we randomly selected 15 observations and we modified them through the perturbation operation defined as  $\mathbf{y} \oplus \boldsymbol{\delta} = \mathcal{C}\{y_1 \cdot \delta_1, \dots, y_D \cdot \delta_D\} \in \mathcal{S}^D$ , where  $\mathbf{y}$  and  $\boldsymbol{\delta} \in \mathcal{S}^D$ , and  $\mathcal{C}\{\cdot\}$  is the closure

Title Suppressed Due to Excessive Length

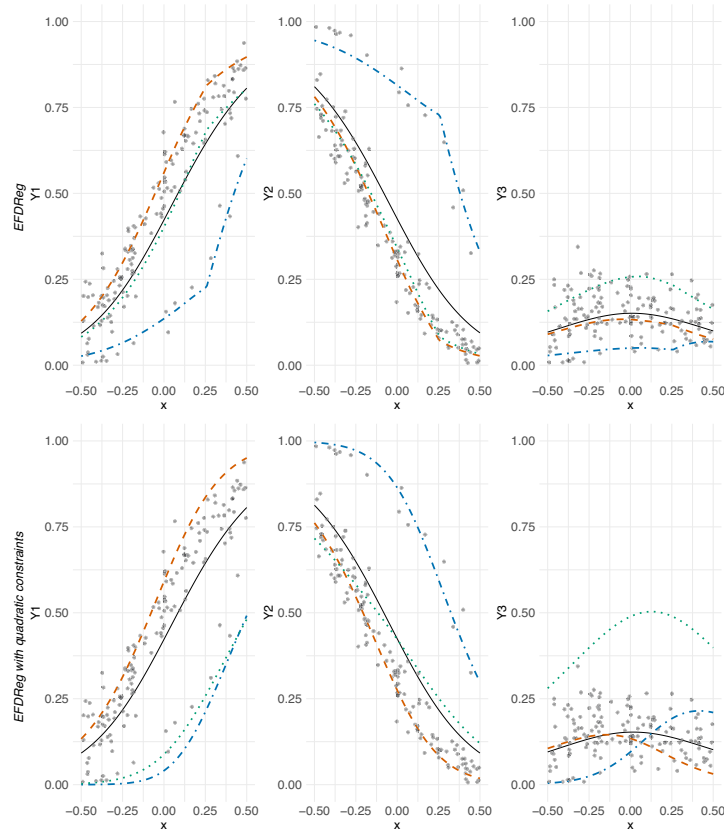
operator defined as  $\mathcal{C}\{\mathbf{q}\} = \left\{ q_1 / \sum_{j=1}^D q_j, \dots, q_D / \sum_{j=1}^D q_j \right\}$ . In this simulation we selected  $\boldsymbol{\delta} = (0.07, 0.86, 0.07)'$ .

Parameter	EFDReg	EFDReg lin.	EFDReg quad.
$\beta_{01}$	0.991 (0.036)	0.982 (0.034)	0.987 (0.035)
$\beta_{11}$	1.923 (0.129)	1.825 (0.140)	1.939 (0.131)
$\beta_{02}$	0.904 (0.146)	0.812 (0.042)	0.969 (0.091)
$\beta_{12}$	-2.342 (0.195)	-2.790 (0.188)	-2.417 (0.167)
$\alpha^+$	35.492 (4.774)	23.202 (2.536)	35.642 (4.819)
$p_1$	0.560 (0.283)	0.840 (0.053)	0.684 (0.159)
$p_2$	0.153 (0.098)	0.146 (0.008)	0.205 (0.037)
$p_3$	0.287 (0.281)	0.014 (0.051)	0.111 (0.147)
$\tilde{w}_1$	0.283 (0.084)	0.678 (0.054)	0.304 (0.088)
$\tilde{w}_2$	0.706 (0.058)	0.986 (0.002)	0.964 (0.015)
$\tilde{w}_3$	0.268 (0.102)	0.467 (0.016)	0.301 (0.070)
WAIC	-813.2	-764.2	-834.15

**Table 1** Monte Carlo approximations of the estimators' mean and standard error (in parenthesis) for the model parameters under each competing parameterization.

Table 1 reports the Monte Carlo approximations of the means and standard errors of the regression model parameters under the three parametrizations, together with the Watanabe-Akaike information criterion (WAIC) values, which allow to compare the goodness of fit of the models [4], a lower WAIC denoting a better fit. Looking at Table 1, we can see that the estimates of  $\tilde{w}_j$ 's are higher under the "linear" and "quadratic" versions of the EFDReg. This is quite reasonable since, due to the restricted parameter space, larger values of  $\tilde{w}_j$ 's are necessary to achieve the same value of  $w_j$  (as far as possible). The EFDReg and the quadratic EFDReg are characterized by a larger estimate of the precision parameter  $\alpha^+$  than the linear EFDReg, and the WAIC's values confirm that these two models show a better fit to data than the linear one. In particular, the quadratic EFDReg has the best performance. Clearly this latter result depends on the structure of the data, but it has been confirmed by further simulation studies. In many (although not all) of the considered scenarios, the quadratic EFDReg performs as well as (or better than) the standard EFDReg, the linear EFDReg being the worst.

Let us now graphically appreciate the advantage of the quadratic model in terms of smoothness. Fig. 2 shows the EFDReg's estimated regression curves (upper panels) and the quadratic EFDReg's curves (lower panels) for one of the datasets artificially generated for the simulation study. It is possible to observe a non smooth behavior of  $\lambda_2$  when  $x \approx 0.25$  in the upper panels, and this is exactly due to the way we normalize  $w_j$  in (5). Contrarily, the quadratic EFDReg succeeds in leading to smoother regression curves.



**Fig. 2** Fitted regression curves under the EFDReg (upper panels) and the quadratic EFDReg (lower panels) for an artificially generated dataset. Curves are referred to  $\mu$  (black solid),  $\lambda_1$  (orange dashed),  $\lambda_2$  (blue dot-dashed), and  $\lambda_3$  (green dotted).

## References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. The Blackburn Press (2003)
2. Albert, J.: Bayesian computation with R. Springer Science & Business Media (2009)
3. Frühwirth-Schnatter, S.: Finite mixture and Markov switching models. Springer Science & Business Media (2006)
4. Gelman, A., et al.: Bayesian Data Analysis, 3rd edn. CRC Press, London (2013)
5. McCullagh, P., Nelder, J.: Generalized linear models. Chapman & Hall, London (1989)
6. Ongaro, A., Migliorati, S., Ascari, R. (2020). A new mixture model on the simplex. *Statistics and Computing*. <https://doi.org/10.1007/s11222-019-09920-x>
7. Di Brisco, A.M., Ascari, R., Migliorati, S., and Ongaro, A. (2019). A new regression model for bounded multivariate responses. In *Book of Short Papers SIS2019* (pp.817-822).
8. Neal, R.M.: An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm. *J. Comput. Phys.*, **111**(1), 194–203 (1994)
9. Stan Development Team. 2020. Stan Modeling Language Users Guide and Reference Manual, 2.25. <http://mc-stan.org/>

# Spatially dependent mixture models with a random number of components

## *Modelli mistura spazio-dipendenti con un numero aleatorio di componenti*

Matteo Gianella, Mario Beraha and Alessandra Guglielmi

**Abstract** In finite mixture models, the choice of the number of components is crucial. From the Bayesian perspective, the correct approach is assuming such number unknown and random. In this work, we set such a prior on a finite mixture model for areal data, assuming that, within each area, data are iid from area-specific densities and we introduce spatial dependence in their joint distribution. We propose a transdimensional sampler via reversible jump which exploits optimal proposals that improve chain mixing and sampler efficiency. The approach is validated on a simulated scenario.

**Abstract** *Nei modelli mistura finiti, la scelta del numero di componenti è fondamentale. Da un punto di vista bayesiano, l'approccio corretto è assumere tale numero incognito e aleatorio. In questo lavoro, definiamo tale prior in un modello mistura finito per dati areali, assumendo che, in ogni area, i dati siano iid da una densità ad essa specifica ed introduciamo una dipendenza spaziale nella loro distribuzione congiunta. Proponiamo un sampler transdimensionale per mezzo di reversible jump che sfrutta l'introduzione di proposte ottimali che migliorano il mixing della catena e l'efficienza del sampler. L'approccio è validato su dati simulati.*

**Key words:** Bayesian model selection, reversible jump MCMC, spatial mixtures

## 1 Introduction

Mixture models provide a natural framework for model-based clustering as well as for approximating densities that are not suitably modeled by standard parametric families. For a recent review, see [2]. Even though mixtures are often used under

---

Matteo Gianella<sup>1</sup>, Mario Beraha<sup>1,2</sup> and Alessandra Guglielmi<sup>1</sup>

<sup>1</sup>Department of Mathematics, Politecnico di Milano, Milano, Italy

<sup>2</sup>Department of Computer Science, Università degli Studi di Bologna, Bologna, Italy

e-mail: matteo1.gianella@mail.polimi.it, {mario.beraha, alessandra.guglielmi}@polimi.it



the assumption of exchangeable samples from a unique unknown distribution, there are cases in which such models have been adopted to model data that show spatial dependence. A novel work in this field is [1], where the problem of modeling areal data is considered. In particular, [1] assumes a finite mixture with a fixed number of components  $H$  for each area and introduce spatial dependence via a suitable prior on the weights of the mixtures, i.e. the *logistic multivariate CAR prior*.

A common issue with finite mixtures is the choice of an appropriate number of mixture components, particularly important when the analysis requires an interpretation of the clusters induces by the mixture. Two strategies are commonly adopted to deal with such problem. The first one consists in fixing  $H$  to a reasonably large upper bound and assume a *sparse* prior on the weights, so that, asymptotically, only  $k < H$  components result allocated (see [8]). However, such *sparse* priors have been studied only for classical mixture models, i.e., when data are exchangeable from a single unknown distribution.

The second strategy is straightforward under the Bayesian approach and is the one we assume here: the number of components  $H$  is unknown and considered random. Despite its conceptual simplicity, this latter approach is characterized by computational difficulties since a *transdimensional* Markov Chain Monte Carlo (MCMC) algorithm should be designed for posterior inference. Examples of such transdimensional MCMC algorithms include the reversible jump MCMC in [7] and the MCMC based on birth-and-death processes in [9]. More recently, by exploiting the notion of exchangeable partition probability function (EPPF) [5] has proposed a “marginal” MCMC sampling scheme.

In this work, we extend the the spatial mixture model defined in [1] by assuming a prior on the number  $H$  of components and we propose a transdimensional sampler via a reversible jump MCMC algorithm. This sampling strategy is forced by the model itself, since theoretical results about *sparse* priors in [8] are not available in this more complex setting, and EPPF of our model is not known in analytical form so far.

The rest of this article is organized as follows. In Section 2 we introduce the model. Then, Section 3 describes and motivates the reversible jump move step. Finally, Section 4 presents a simulation study to check the correctness and efficiency of our algorithm.

## 2 The Bayesian model

In this section we introduce a model that extends the Bayesian mixtures in [1] by assuming a prior on the number of components.

**Likelihood for areal data.** Consider data  $\mathbf{y}_1, \dots, \mathbf{y}_I$ , where  $\mathbf{y}_i = (y_{i1}, \dots, y_{iN_i})^T$  are exchangeable observations from the areal unit  $i$ , for  $i = 1, \dots, I$ . Assume that a neighbouring structure  $G$  between the  $I$  different areal units is known. We assume  $G$  as a  $I \times I$  matrix, where its entries  $G_{ij}$  indicates whether  $i$  and  $j$  are neighbours ( $G_{ij} = 1$ ) or not ( $G_{ij} = 0$ ). For each  $i$  in  $1, \dots, I$ , the conditional distribution of our data is

Spatially dependent mixture models with a random number of components

specified as follows:

$$y_{ij} | \mathbf{w}_i, \boldsymbol{\tau}, H \stackrel{\text{iid}}{\sim} \sum_{h=1}^H w_{ih} \mathcal{N}(\cdot | \boldsymbol{\tau}_h) \quad j = 1, \dots, N_i, \quad (1)$$

where  $\mathbf{w}_i = (w_{i1}, \dots, w_{iH})^T$  is a  $H$ -dimensional vector in the simplex  $S^H$ , i.e.,  $w_{ih} \geq 0$  and  $\sum_h w_{ih} = 1$  and  $\mathcal{N}(\cdot | \boldsymbol{\tau}_h)$  denotes the Gaussian density with parameters  $\boldsymbol{\tau}_h = (\mu_h, \sigma_h^2)$ . Observe how in (1) the  $(\mu_h, \sigma_h^2)$ 's are shared across all the spatial locations. Thus, this model allows to introduce dependency between mixtures associated to different areas only through the prior for the weights  $(\mathbf{w}_1, \dots, \mathbf{w}_I)$ .

**Logistic MCAR prior.** The prior introduced to induce spatial dependence among mixtures from different areas is defined through a multivariate CAR distribution on a transformation of the weights. In particular, the weights are transformed via the additive log ratio map, defined from  $S^H$  to  $\mathbb{R}^{H-1}$  and such that:

$$\tilde{w}_h = \log(w_h/w_H) \quad h = 1, \dots, H-1.$$

Once the transformed weights have been defined, then a multivariate CAR prior is imposed on  $\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_I)$  as:

$$\tilde{\mathbf{w}}_i | \tilde{\mathbf{w}}_{-i}, \boldsymbol{\Sigma}, \rho, H \sim \mathcal{N}_{H-1}(\rho \sum_j G_{ij} \tilde{\mathbf{w}}_j, \boldsymbol{\Sigma}) \quad i = 1, \dots, I. \quad (2)$$

This model defines a unique joint distribution for  $\tilde{\mathbf{w}}$  when  $\rho \in (-1, 1)$ ; see [3].

**Prior on  $\mathbf{H}$ .** Given the likelihood for the data and the prior for the weights, the Bayesian model is then extended adding priors on the hyperparameters. In our context, we model the atoms  $\boldsymbol{\tau}_h$  conditioning to  $H$ , independently from a *Normal – InvGamma* $(\mu_0, a, b, \lambda)$ , the matrix  $\boldsymbol{\Sigma}$  is assumed to be diagonal, i.e.  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$  with  $\sigma^2 \sim \text{InvGamma}(\alpha, \beta)$ ,  $\rho$  has a uniform prior in  $(0, 1)$ . Finally, we assume a shifted Poisson distribution on the number of components  $H$ , i.e.,  $H-1 \sim \text{Poi}(\lambda)$ .

### 3 Reversible Jump computation via recursive auxiliary priors

The reversible jump MCMC sampler [4] provides a general framework for transdimensional simulation schemes. It can be viewed as an extension of the Metropolis-Hastings algorithm. As it happens in standard Metropolis-Hastings, given the current state of the chain  $\boldsymbol{\theta} = (H, \boldsymbol{\theta}_H)$  (where we make explicit the ‘‘dimension’’  $H$ ), the next state  $\boldsymbol{\theta}' = (H', \boldsymbol{\theta}_{H'})$  is (i) sampled from a proposal distribution  $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ , and (ii) accepted with probability  $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$  equal to

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}' | \mathbf{y}) q(\boldsymbol{\theta}', \boldsymbol{\theta})}{\pi(\boldsymbol{\theta} | \mathbf{y}) q(\boldsymbol{\theta}, \boldsymbol{\theta}')} \right\}.$$

Usually, the proposal distribution is defined in two steps. If  $\boldsymbol{\theta}_H \in \mathbb{R}^{n_H}$  and  $\boldsymbol{\theta}' \in \mathbb{R}^{n_{H'}}$ , with  $n_{H'} > n_H$  and  $d = n_{H'} - n_H$ , first a random vector  $\mathbf{u} \in \mathbb{R}^d$  is sampled from a distribution  $q_d(\mathbf{u})$  and then  $\boldsymbol{\theta}_{H'}$  is defined as  $g_{H \rightarrow H'}(\boldsymbol{\theta}_H, \mathbf{u})$  for a suitable mapping function  $g_{H \rightarrow H'}$ . Since both the proposal distribution  $q_d(\mathbf{u})$  and the mapping function  $g_{H \rightarrow H'}$  are arbitrary, the definition of a suitable reversible jump move is usually a difficult task. The approach we follow to design a reversible jump move for the model described in Section 2 is based on [6], where the author defines optimal auxiliary priors and proposals for generic nested models indexed by an integer  $H \in \mathbb{N}^+$  with unknown parameter  $\boldsymbol{\theta}_H$  and prior of the form  $\pi(\boldsymbol{\theta}_H | H)\pi(H)$ . Let us introduce key ideas in [6], that we will adapt to our context.

Since the models are nested, the unknown parameters are nested as well, i.e., if  $H' > H$ , the first  $H$  elements of  $\boldsymbol{\theta}_{H'}$  correspond to vector  $\boldsymbol{\theta}_H$ , the one indexed by  $H$ . Given the current state  $(\boldsymbol{\theta}_H, H)$ , consider moving to  $(\boldsymbol{\theta}_{H'}, H')$  with  $H' = H + 1$  (the case  $H' = H - 1$  is identical but with  $H$  and  $H'$  swapped). The joint distribution for  $(\mathbf{y}, \boldsymbol{\theta}_{H'}, H)$  is defined as:

$$f(\mathbf{y}, \boldsymbol{\theta}_{H'}, H) = \tilde{\pi}_H([\boldsymbol{\theta}_\infty]_{H'} | \boldsymbol{\theta}_H, \mathbf{y})f(\mathbf{y} | H, \boldsymbol{\theta}_H)\pi(\boldsymbol{\theta}_H)\pi(H),$$

where  $[\boldsymbol{\theta}_\infty]_{H'}$  represents the  $H'$ -th element of  $\boldsymbol{\theta}_\infty$  and  $\tilde{\pi}_H([\boldsymbol{\theta}_\infty]_{H+1} | \boldsymbol{\theta}_H, \mathbf{y})$  needs to be defined. Choosing such quantity as the conditional posterior

$$\pi([\boldsymbol{\theta}_\infty]_{H+1} | \mathbf{y}, H + 1, \boldsymbol{\theta}_H) \propto f(\mathbf{y} | H + 1, \boldsymbol{\theta}_{H+1})\pi(\boldsymbol{\theta}_{H+1} | H + 1) \quad (3)$$

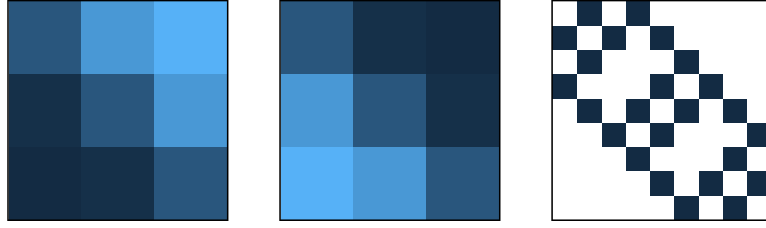
guarantees optimal conditions in terms of overall chain mixing and minimization of the estimated variance. Nonetheless, this optimal posterior is not known a priori, so we need to estimate it. We see how this is possible in our specific case, i.e. for the spatial mixture model.

First of all, at a fixed dimension  $H$ , the unknown parameter vector  $\boldsymbol{\theta}_H$  is  $\text{vec}(\tilde{\mathbf{w}}, \boldsymbol{\tau})$ , where  $\tilde{\mathbf{w}} = \text{vec}(\{\tilde{\mathbf{w}}_i\}_{i=1:H})$  and  $\text{vec}$  indicates the vectorization of the given quantity. Thus, in case the algorithm propose to add a new component, it is required to sample  $[\boldsymbol{\theta}_\infty]_{H+1} = (w_{1H+1}, \dots, w_{IH+1}, \boldsymbol{\mu}_{H+1}, \sigma_{H+1}^2)$ . The great novelty of this approach is the fact that the posterior distribution  $\pi([\boldsymbol{\theta}_\infty]_{H+1} | \mathbf{y}, H + 1, \boldsymbol{\theta}_H)$  is a direct proposal distribution for the new added component. In this way, we side step the artificial construction of proposal distributions and mapping functions, whose definition is totally arbitrary and does not ensure any particular property in terms of sampling performance. The acceptance rate  $\alpha[(H, \boldsymbol{\theta}_H), (H', \boldsymbol{\theta}_{H'})]$  is given by  $\min\{1, A\}$ , with  $A$  equal to

$$A = \frac{f(\mathbf{y} | \text{alr}^{-1}(\tilde{\mathbf{w}})_{i=1:H}, \boldsymbol{\tau}, H')\pi(\boldsymbol{\tau} | H')\pi(\tilde{\mathbf{w}}_{i=1:H} | \rho, \boldsymbol{\Sigma}, H')\pi(H')}{f(\mathbf{y} | \text{alr}^{-1}(\tilde{\mathbf{w}})_{i=1:H}, \boldsymbol{\tau}, H)\pi(\boldsymbol{\tau} | H)\pi(\tilde{\mathbf{w}}_{i=1:H} | \rho, \boldsymbol{\Sigma}, H)\pi(H)} \times \left( \frac{\chi_{\{H'=H+1\}}}{\tilde{\pi}([\boldsymbol{\theta}_\infty]_{H'} | \boldsymbol{\theta}_H, \mathbf{y})} + \chi_{\{H'=H-1\}}\tilde{\pi}([\boldsymbol{\theta}_\infty]_H | \boldsymbol{\theta}_{H'}, \mathbf{y}) \right). \quad (4)$$

Note that the marginal priors for  $\boldsymbol{\Sigma}$  and for  $\rho$  do not appear in (4) since they does not depend on the number of components  $H$ . Finally, since the proposal posterior distribution is not known, in case of addition of a new component we approximate it with a multivariate Gaussian centered in  $\boldsymbol{\theta}^* = \arg \max \pi([\boldsymbol{\theta}_\infty]_{H+1} | \mathbf{y}, H + 1, \boldsymbol{\theta}_H)$

Spatially dependent mixture models with a random number of components



**Fig. 1** Simulation from spatially dependent weights, from left to right:  $w_{i1}$ ,  $w_{i2}$  and  $G$ .

(see (3)) and variance equal to the Hessian of  $\pi([\boldsymbol{\theta}_\infty]_{H+1} | \mathbf{y}, H+1, \boldsymbol{\theta}_H)$  evaluated in  $\boldsymbol{\theta}^*$ . On the other hand, when we reduce the dimension of the state, the parameters of the approximated optimal posterior are computed using as maximizer the selected component to drop.

#### 4 Simulation study

We present a simple simulation study to assess the performance of our reversible jump MCMC algorithm. Code for posterior simulation has been implemented in C++ and is available as an R package at <https://github.com/TeoGiane/SPMIX>. We run the MCMC chain for a total of 10,000 iterations, discarding the first 5,000 as burn-in and thinning the chain every two iterations, so that the final sample size is 2,500.

We consider 9 areas numbered in lexicographical order as in Figure 1 (left). In the  $i$ -th area, we draw a sample from

$$y_{ij} \stackrel{\text{iid}}{\sim} w_{i1} \mathcal{N}(-5, 1) + w_{i2} \mathcal{N}(0, 1) + w_{i3} \mathcal{N}(5, 1) \quad j = 1, \dots, 25.$$

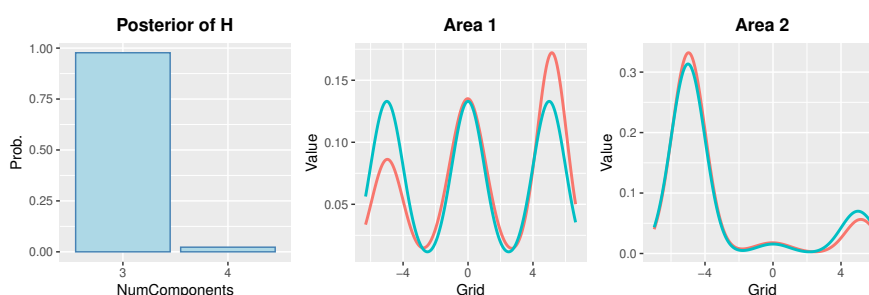
The weights are computed as  $alr^{-1}(\tilde{\mathbf{w}})$  and  $\tilde{\mathbf{w}}$  is defined as

$$\tilde{w}_{i1} = 3(x_i - \bar{x}) + 3(y_i - \bar{y}) \quad \tilde{w}_{i2} = -3(x_i - \bar{x}) - 3(y_i - \bar{y}), \quad (5)$$

being  $(x_i, y_i)$  and  $(\bar{x}, \bar{y})$  the coordinates of the center of area  $i$  and of the grid center. In this context,  $G$  is defined so that two areas are close if they share a common edge, as in Figure 1 (right).

Note that the number of samples in each location is extremely small, so that the sharing of information between neighboring mixtures is essential. Moreover, observe that (5) induces a different kind of spatial dependency from (2).

Figure 2 (left) shows the posterior distribution of the number of components. The remaining panels in Figure 2 shows the estimated and true densities in two areas. The sampler is extremely effective in retrieving the correct number of components and in estimating densities. It also provides remarkable results even with few observations in each area.



**Fig. 2** Some results obtained for the considered scenario. To the left, the posterior traceplot of  $H$  and the comparison between theoretical (light blue) and estimated (pink) density for two areas.

## 5 Conclusions

In this paper, we have extended the model introduced in [1] by adding a prior on the number of components. From the computational point of view, the model becomes very challenging and we have proposed a reversible jump algorithm. We have selected a suitable reversible jump move to directly build the optimal proposal and avoid the artificial construction of mapping functions and proposal distributions. We have set up an efficient sampling scheme able of sampling from such model to show the goodness of fit of our model in retrieving the right number of components and capturing the spatial dependency among different areas.

## References

1. Beraha, M., Pegoraro, M., Peli, R., Guglielmi, A.: Spatially dependent mixture models via the Logistic Multivariate CAR prior. arXiv preprint arXiv:2007.14961 (2020)
2. Fruhwirth-Schnatter, S., Celeux, G., Robert, C.P.: Handbook of mixture analysis. CRC press (2019)
3. Gelfand, A.E., Vounatsou, P.: Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**(1), 11–15 (2003)
4. Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**(4), 711–732 (1995)
5. Miller, J.W., Harrison, M.T.: Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**(521), 340–356 (2018)
6. Norets, A.: Optimal auxiliary priors and reversible jump proposals for a class of variable dimension models. *Econometric Theory* p. 1–33 (2020)
7. Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(4), 731–792 (1997)
8. Rousseau, J., Mengersen, K.: Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5), 689–710 (2011)
9. Stephens, M.: Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of statistics* pp. 40–74 (2000)

# Finite mixtures of regression models for longitudinal data

## *Miscugli finiti di modelli di regressione per dati longitudinali*

Marco Alfò and Roberto Rocci

**Abstract** Individual-specific, time-constant, random effects are often introduced in model specification to account for dependence and/or omitted covariates in regression models for longitudinal data. This approach has been frequently criticized as it would not be robust to the presence of correlation between the observed and the unobserved covariates. Often, this is felt as a reason to choose the *fixed effect* estimator instead. Starting from the so-called *correlated effect approach*, we argue that the conditional random effect distribution may be estimated non-parametrically by using a discrete distribution, leading to a general solution to the problem. The effectiveness of the proposed approach is shown via a large scale simulation study.

**Abstract** *La specificazione di un modello di regressione per dati longitudinali è spesso basata su effetti casuali costanti nel tempo e specifici dell'individuo, che permettono di tener conto della dipendenza tra le osservazioni e dell'omissione di potenziali covariate. Questo approccio è spesso criticato perchè non sarebbe robusto rispetto alla presenza di correlazione tra le covariate osservate e quelle omesse; tale argomento è frequentemente utilizzato per suggerire l'uso dello stimatore a effetti fissi. Partendo dal cosiddetto approccio ad effetti correlati, argomentiamo che una soluzione ancora più generale può essere ottenuta stimando in modo non-parametrico la distribuzione condizionata degli effetti casuali con una discreta. L'efficacia della proposta è illustrata con l'applicazione a dati simulati.*

**Key words:** Longitudinal data, omitted variables, dependence, random effect models, correlation bias, nonparametric MLE.

---

Marco Alfò

La Sapienza University, Dept. of Statistical Science, e-mail: marco.alfò@uniroma1.it

Roberto Rocci

La Sapienza University, Dept. of Statistical Science, e-mail: roberto.rocci@uniroma1.it

## 1 Introduction

In longitudinal data, the same individuals are repeatedly observed over a, usually short, time window; such data often present two key features: individuals are heterogeneous, and measurements from the same individual are likely to be dependent. Individual-specific random effects are often added to the linear predictor to account for unobserved heterogeneity and dependence, as the repeated measurements are assumed to share some common individual-specific unobservables. The so-called *fixed effect* estimator is frequently advocated due to the claim that the approach based on random effects would not lead to a consistent estimator in the presence of dependence between the observed covariates and the random effects. It is worth noticing, however, that both estimators may be based on the same working hypotheses and what really makes the difference is the use of a conditional rather than a marginal maximum likelihood approach. According to the thoughtful review in [22], the distinction between fixed and random effects makes no sense; rather, we have to talk about individual-specific effects and approaches to estimation that can be made conditional or unconditional (marginal) to the individual effects. The Hausman test [11] is often adopted to choose between the two estimators; the test is essentially a tool for verifying whether the working hypothesis of independence between the random effects and the observed covariates, that is usually employed in the marginal approach, is supported by the observed data.

We approach this issue, starting from the proposal of the *auxiliary equation* made by [16] and extended by [3, 4], where the so-called *correlated* random effect estimator is defined. The topic has been discussed in the statistical world by several authors; just to cite a few, see eg [17], [18], [19], and, quite recently and in a very different context, [14]. An interesting contribution to the debate has been given by [13], who discussed an approach based on the so-called *QP decomposition*, extended to glms by [17]. Here,  $\mathbf{P}$  and  $\mathbf{Q}$  are matrices projecting the covariates vector onto the space spanned by the individual means (over the analyzed period) and by the orthogonal counterpart (time-specific deviation from the individual means), respectively. This result is simple and sheds some light on the Mundlak approach. While all these alternatives provide reliable estimates, we think that a less parameterized approach, based on finite mixture and a proper, conditional, representation of the prior masses could be an efficient and general solution to the problem. The paper is structured as follows. In section 2, the problem and our proposal is presented. In section 3, we briefly outline the scheme and the results of a large-scale simulation study. Last section gives some concluding remarks.

## 2 The model

Longitudinal studies refer to a two-stage sample  $\{y_{it}, \mathbf{x}_{it}\} i = 1, \dots, n, t = 1, \dots, T$ , where the same units  $i = 1, \dots, n$  have been observed at a number of (usually common) time occasions belonging to a discrete set, i.e.  $t \in \{1, \dots, T\}$ . Here,  $y_{it}$  repre-

sents the observed value of the response  $Y_{it}$  for the  $i$ -th individual at the  $t$ -th time occasion, and  $\mathbf{x}_{it}$  denotes a  $p$ -dimensional set of observed covariates, used to represent the *observed* heterogeneity. According to [7], “(...) the major advantage of the former (longitudinal study) is its capacity to separate what in the context of population studies are called cohort and age effects”. That is, by employing a longitudinal study, we may distinguish between the impact of natural, between-individual, heterogeneity (the *cohort* effect) and the impact of time (the *age* effect), measured by the dynamics in the observed covariates.

We translate these concepts into modelling by assuming a parametric conditional distribution for the observed response, e.g. a member of the exponential family

$$Y_{it} \mid \mathbf{x}_{it}, \mathbf{u}_i \sim \text{EF}(\theta_{it}), \quad i = 1, \dots, n, t = 1, \dots, T$$

where the canonical parameter is described via a random effect model

$$\theta_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \mathbf{w}'_{it} \mathbf{u}_i.$$

In the following, we will refer to the case  $\mathbf{w}_{it} = 1$  and  $\mathbf{u}_i = u_i$ , defining a random intercept model, but the approach could be easily adapted to higher dimension random effects as well. The model is specified by the (conditional) distribution we adopt for the longitudinal response, and by an appropriate specification for the random effect distribution  $U_i \mid \mathbf{x}_i \sim g(\cdot)$ . The random effect  $u_i$  is meant to represent individual-specific *unobserved* heterogeneity, while *observed* heterogeneity is summarized by the covariates  $\mathbf{x}_i = \text{vec}(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) \in \mathcal{M}(T(p+1), 1)$ .

Assuming independence of the repeated measurements from the same individual, conditional on the individual-specific latent effects, the marginal likelihood is

$$L(\cdot) = \prod_{i=1}^n \left\{ \int_{\mathcal{U}} \left[ \prod_{t=1}^T f(y_{it} \mid \mathbf{x}_{it}, u_i) \right] g(u_i \mid \mathbf{x}_i) du_i \right\},$$

where the individual-specific latent effects are integrated out. Obviously, we may also resume to a conditional approach, when, for example, a sufficient statistic for the  $u_i$ s does exist, and adopt the fixed effect estimator, see [3] and [2].

The key point in the marginal likelihood approach is that it is usually assumed  $g(u_i \mid \mathbf{x}_i) = g(u_i)$ , leading to the so-called *random effect* estimator (the GLS estimator in the Gaussian case). This is a strong hypothesis that can be verified by the Hausman test, which is used to choose between a conditional (leading to fixed effect-type estimators) and a marginal, potentially misspecified, approach (leading to the random effect estimator). However, rejection of such an hypothesis does not necessarily mean that a marginal approach cannot be taken, but rather that the assumption of independence is not supported by the data. It is interesting to note that, at least in the statistical field, a huge literature has been focused on the choice of an appropriate form for the random effect marginal distribution, while relatively few papers have dealt with the possible consequences of a wrong assumption of independence how to avoid it.



It follows that we may still rely on the marginal approach, but we have to solve the issue of how  $g(u_i | \mathbf{x}_i)$  should be handled. A first attempt is to parametrically model this dependence via the *auxiliary* regression approach due to [16]. By projecting the random effects onto the space spanned by the  $\mathbf{x}_i$ s, we may reparameterize the random intercept as follows

$$U_i = [E(U_i | \mathbf{x}_i) + U_i^*] = \frac{1}{T} \sum_t \mathbf{x}'_{it} \gamma_t + U_i^*$$

In the simplest case  $\gamma_t = \gamma, t = 1, \dots, T$  and  $\frac{1}{T} \sum_t \mathbf{x}'_{it} \gamma = \bar{\mathbf{x}}'_i \gamma$ . Therefore, for a (conditionally) Gaussian response, we get the following model structure

$$E(Y_{it} | \mathbf{x}_{it}, u_i) = \mathbf{x}'_{it} \beta + u_i, \quad u_i = \bar{\mathbf{x}}'_i \gamma + u_i^*, \quad U_i^* \sim g(\cdot | \phi).$$

This parameterization has been extended by [4], and the corresponding estimator is usually referred to as the *correlated random effect* estimator.

Our idea is to estimate the general model semiparametrically, i.e. without assuming a particular parametric form for  $g(\cdot)$ . In this case, the (bounded) likelihood is maximized wrt  $g(\cdot)$  by a discrete distribution with at most  $K \leq n_d$  support points, where  $n_d$  denotes the number of distinct individual covariates profiles in the sample. Indeed, the likelihood function is approximated by the following finite mixture [15]

$$L(\cdot) = \prod_{i=1}^n \int_{\mathcal{U}} \left[ \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, u_i) \right] g(u_i | \mathbf{x}_i) d(u_i) \simeq \prod_{i=1}^n \sum_{k=1}^K \left[ \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \zeta_k) \right] \pi_k(\mathbf{x}_i)$$

where the conditional distribution of  $U_i | \mathbf{x}_i$  is discrete over  $K$  locations  $\zeta_k$  with mass  $\pi_k(\mathbf{x}_i)$ , as in the *concomitant variable* models [5]. Model identifiability is discussed in [21], where full rank condition for the covariates are given, while more stringent conditions are provided by [8], [12], [9]. DerSimonian [6] details the algorithm for Maximum Likelihood estimation.

This would prevent that dimension of the parameter vector increases with the dimension of the random effects, leading to a simpler interpretation, and accounting for *nonlinearities* in the association between the random effects and the observed covariates vector. It is worth noticing that we may use this parameterization anytime some of the covariates (say observed explanatory variables, autoregressive components, indicators of missing information, etc.) that are included in the model can be thought of as being, potentially, endogenous.

### 3 Simulation Study

To study the empirical behaviour of the proposed modelling approach, we have designed the following simulation study. 250 samples have been drawn from either a Gaussian or a Bernoulli population:

Finite mixtures of regression models for longitudinal data

- A - Gaussian case  $Y_{it}|x_{it}, u_i \sim N(\mu_{it}, \sigma_e^2), \mu_{it} = \beta_0 + \beta_1 x_{it} + u_i$
- B - Bernoulli case  $Y_{it}|x_{it}, u_i \sim \text{Bin}(1, \pi_{it}), \text{probit}(\pi_{it}) = \beta_0 + \beta_1 x_{it} + u_i$

where regression coefficients have been uniformly drawn from the following ranges  $\beta_0 \in [-0.6, -0.2]$  and  $\beta_1 \in [0.25, 0.75]$ . As far as the covariates values are concerned, they have been drawn from a multivariate ( $T$ -dimensional) Gaussian density with unit variances and constant correlations. For each of the two populations (A: Gaussian, B: Bernoulli) we considered 3 scenarios:

- [Scenario 1] Gaussian individual-specific random effects with constant correlation  $\rho$  with  $\mathbf{x}_{it}, t = 1, \dots, T$ , where  $\rho \in (0, 0.2), \rho \in (0.2, 0.5), \rho \in (0.5, 0.8)$ ;
- [Scenario 2]  $u_i = \exp(\gamma_0 + \gamma_1 \bar{x}_i) + \varepsilon$ , where  $\varepsilon \sim N(0, 1)$ ;
- [Scenario 3]  $K = 3$ , and  $\pi_k(\bar{x}_i) \propto \exp(\phi_{0k} + \phi_{1k} \bar{x}_i)$ , with  $\phi_0 = (0, 0.5, -3.5), \phi_1 = (0, -3.5, 3), \zeta = (-2, 0, 1)$ .

For each sample, we have considered a standard finite mixture (Plain) model, a finite mixture model with the covariates-dependent prior where  $\pi_k(\bar{x}_i) \propto \exp(\alpha_{0k} + \alpha_{1k} \bar{x}_i)$  (Cov), a finite mixture model with QP decomposition (QP), a standard parametric mixed model without (PlainPar) or with (QPPar) a QP decomposition. In the case of binary responses, we have also considered a fixed effect estimator (FE), and a bias-corrected fixed effect estimator (FEbc), according to [10], available in the R package `bife`, see [20]. The FE estimator has not been taken into account in the Gaussian case, as it can be shown to be equal to the estimator obtained by QPPar, see [16] or [1]. For all the finite mixture models, we have considered three different methods to select the number of components: maximum likelihood (with minimum threshold  $\varepsilon = 10^{-07} * npar(K)$  between two subsequent values  $K$  and  $(K + 1)$ ), where  $npar(K)$  represents the number of parameters for a model with  $K$  components, AIC or BIC. Further, we have considered different values for the number of individuals  $n \in \{100, 250, 500\}$  and the number of occasions  $T = 5, 10$  for each individual.

The results of the simulation study, not shown here for sake of space, show that the proposed approach offers in all analysed scenarios good performance in terms of both bias and MSE of model parameter estimates. When we move from the linear to the nonlinear case (Bernoulli with probit link) it outperforms all the others, including the fixed effect and the bias corrected fixed effect estimator.

## 4 Concluding remarks

In this paper, we have described a semiparametric approach to deal with covariates endogeneity in random effect models for longitudinal responses. While we focus on the simplest case of discrete time and common measurement occasions, the approach we propose can be readily extended to studies in continuous time with (at least partially) non common and unequally spaced time occasions. Moreover, while we discuss, for sake of simplicity, only balanced designs, the approach works just as fine in the case of unbalanced studies.

## References

1. Baltagi, B.H. *A Companion to Econometric Analysis of Panel Data*. Chichester: John Wiley & Sons, 2009.
2. M.R. Conaway. Analysis of repeated categorical measurements with conditional likelihood methods *Journal of the American Statistical Association*, 84:53–62, 1989.
3. G. Chamberlain. Analysis of covariance with qualitative data *Review of Economic Studies*, 47:225–238, 1980
4. G. Chamberlain. Panel Data In Z. Griliches and M.D. Intriligator, eds, *Handbbook of Econometrics, Volume II*. Elsevier Science Publicher BV, 1984.
5. C.M. Dayton and G.B. MacReady. Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83:173–178, 1988.
6. R. DerSimonian. Algorithm AS 221: Maximum Likelihood Estimation of a Mixing Distribution *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 35:302–309
7. P. Diggle, P. Heagerty, K-Y Liang, S. Zeger *Analysis of Longitudinal Data*, Oxford: Oxford University Press, 2002.
8. D.A. Follmann and D. Lambert. Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference*, 27:375–381, 1991.
9. B. Grün and F. Leisch. Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, 25:225–247, 2008.
10. J. Hahn, and W. Newey Jackknife and analytical bias reduction for nonlinear panel models *Econometrica*, 72:1295–1319, 2004.
11. J.A. Hausman Specification tests in econometrics *Econometrica*, 46:1251–1271, 1978.
12. C. Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17:273–296, 2000.
13. J. Krishnakumar Time invariant variables and panel data models: a generalized Frisch-Waugh theorem and its implications In B. Baltagi ed. *Panel Data Econometrics, Volume 274: Theoretical Contributions and Empirical Applications (Contributions to Economic Analysis)*, Emerald Group Publishing Limited, 2006.
14. A.E. Lamont, J.K. Vermunt and M. Lee Van Horn Regression Mixture Models: Does Modeling the Covariance Between Independent Variables and Latent Classes Improve the Results? *Multivariate Behavioral Research*, 51:35–52, 2016
15. G.J. McLachlan, D. Peel *Finite Mixture Models*. John Wiley & Sons
16. Y. Mundlak On the pooling of time series and cross section data *Econometrica*, 46:69–85, 1978
17. J.M. Neuhaus and J.D. Kalbfleisch Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54, 638–645, 1998
18. J.M. Neuhaus and C.E. McCulloch Separating between- and within-cluster covariate effects by using conditional and partitioning methods *Journal of the Royal Statistical Society, Series B*, 68, 859–872, 2006.
19. A. Skrondal, S. Rabe-Heskett. Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *Journal of the Royal Statistical Society, series C*, 63, 211–237, 2014.
20. A. Stammann and F. Heiss and D. McFadden Estimating Fixed Effects Logit Models with Large Panel Data. *EconStor Working paper*, 2016.
21. P. Wang, M.L. Puterman and I. Cockburn, and N. Le. Mixed Poisson regression models with covariate dependent rates. *Biometrics*, 52:381–400, 1996.
22. J.M. Wooldridge. Correlated random effects models with unbalanced panels Working Paper, Michigan State University, Department of Economics, 2009

# Mixtures of regressions for size estimation of heterogeneous populations

## *Misture di regressioni per la stima della numerosità di popolazioni eterogenee*

Gianmarco Caruso

**Abstract** We propose a capture-recapture model which exploits finite mixtures of logistic regressions to account for latent heterogeneity between groups of individuals, in order to better understand their different propensities to the capture as well as different behavioural patterns. The additional behavioural variation in capture probabilities among individuals within a group is expressed by a suitable time-dependent covariate, which summarises the past individual experience [3]. A real data example and a simulation study illustrate how the proposed model performs.

**Abstract** *Si propone un modello cattura-ricattura che sfrutta le misture finite di regressioni logistiche per spiegare l'eterogeneità latente tra gruppi di individui, al fine di comprendere meglio le loro differenti propensioni alla cattura. La variabilità tra le probabilità di cattura di individui appartenenti ad uno stesso gruppo viene espressa mediante un'adeguata covariata tempo-dipendente, che riassume l'esperienza individuale passata [3]. Le potenzialità del modello proposto vengono illustrate attraverso un esempio basato su dati reali e uno studio di simulazione.*

**Key words:** capture-recapture, population size estimation, finite mixtures of GLM, logit regression.

## 1 Introduction

Capture-recapture methods are widely employed in estimating the size of elusive populations, whose units are subject to multiple captures across several occasions.

The main idea behind these techniques is to account for the number of unobserved individuals by suitably modelling and exploiting the capture histories of the observed units. One assumes that a closed population of unknown size  $N$  is sampled

---

Gianmarco Caruso  
Dipartimento di Scienze Statistiche, La Sapienza Università di Roma, Italy  
e-mail: gianmarco.caruso@uniroma1.it

$t$  times, with independence between individuals. For example, in the common case of wildlife populations, animals that are captured for the first time are marked and then released, so that they can be recognizable in future trapping occasions. Supposing that  $M$  distinct individuals have been captured across  $t$  occasions, data are collected on a  $M \times t$  matrix,  $\mathbf{X} = [x_{ij}]$ : in particular,  $x_{ij} = 1$  if individual  $i$  is captured on occasion  $j$ , otherwise  $x_{ij} = 0$ . The  $i$ -th row of the matrix reports the capture history of the  $i$ -th individual. If there are  $N$  individuals in the population, then one can add  $N - M$  rows of zeros to the matrix in order to include all the uncaptured individuals. In the following, one supposes to deal with closed populations, where there are no births, no deaths and no migrations: this assumption seems to be meaningful if the first and the last capture occasions are not too far in time and the range where the population lives is well bounded.

## 2 The model

One considers a model which allows capture probabilities to vary among individuals and across capture occasions. In addition, here one considers the presence of unobserved heterogeneity between groups of individuals, in the sense that different groups may exhibit different responses to captures. Finite mixtures of logistic regressions are thus exploited to account for latent heterogeneity and to better understand different responses by heterogeneous groups of individuals. The additional variation in capture probabilities among individuals within each group may be expressed by a suitable time-dependent covariate, which summarises the past individual experience [6, 3].

In the following, one considers a heterogeneous population of  $N$  individuals which can be partitioned in  $G$  subpopulations (or groups),  $\mathcal{P}_1, \dots, \mathcal{P}_G$ ; namely, the  $N$  individuals are supposed to come from  $G$  different subpopulations of unknown proportions,  $\pi_1, \dots, \pi_G$ , which are non-negative and add up to 1. The proportion  $\pi_g$  represents the *a priori* probability for an individual to belong to the  $g$ -th subpopulation. The observed response  $x_{ij}$  is therefore supposed to be generated by a finite mixture of logistic regressions [11], where the mixture is assumed to be formed by  $G$  components: hence, each mixture component identifies a different group.

Conditional to the group  $g$ , the response at occasion  $j$  for individual  $i$  is given by

$$x_{ij}|p_{ij}^{(g)} \sim \text{Bern}(x_{ij}|p_{ij}^{(g)}), \quad (1)$$

where  $p_{ij}^{(g)}$  is the probability of being captured at occasion  $j$  for the  $i$ -th individual belonging to the  $g$ -th cluster ( $i \in \mathcal{P}_g$ ).

If  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_G)$  is the vector of mixture weights, the unconditional probability distribution of  $x_{ij}$  is given by

$$h(x_{ij}|\boldsymbol{\pi}, \{p_{ij}^{(g)}\}_{g=1, \dots, G}) = \sum_{g=1}^G \pi_g \text{Bern}(x_{ij}|p_{ij}^{(g)}), \quad (2)$$

Mixtures of regressions for capture-recapture modeling

for  $i = 1, \dots, N$  and  $j = 1, \dots, t$ .

The capture probability  $p_{ij}^{(g)}$  depends on the group-specific regression parameters  $\alpha_g$  and  $\beta_g$  and on the value of the covariate  $z_{ij}$ , according to a linear logistic model, namely

$$p_{ij}^{(g)} = \frac{\exp(\alpha_g + \beta_g z_{ij})}{1 + \exp(\alpha_g + \beta_g z_{ij})}, \quad (3)$$

$\forall i \in \mathcal{D}_g, g = 1, \dots, G, j = 1, \dots, t$  [9, 2]. The heterogeneity between groups of individuals is given by differences in the group-specific regression parameters which connect the covariate to the conditional expected value of the response: thus, same levels of the covariate affect the probabilities of recapture of individuals in distinct groups in different ways.

The time-varying covariate matrix  $\mathbf{Z} = [z_{ij}]$  can be derived by exploiting the class of memory-related summaries introduced by [3], so that

$$z_{ij} = g\lambda(x_{i1}, \dots, x_{ij-1}) = \sum_{h=1}^{j-1} \frac{\lambda^{h-1}}{\sum_{k=1}^{j-1} \lambda^{k-1}} x_{ih}, \quad (4)$$

which takes values in  $[0, 1]$ . Notice that  $z_{ij} = 0$  for all partial capture histories such that  $(x_{i1}, \dots, x_{ij-1}) = (0, \dots, 0)$  and, conventionally, for  $j = 1$  (i.e. the first column of the matrix  $\mathbf{Z}$  is composed by all zeros).

As discussed by [3],  $z_{ij}$  represents a weighted average of the past trapping experience for the individual  $i$  based on the first  $j - 1$  occasions. In particular, for  $\lambda = 1$ , all past captures has the same impact on the summary, while, for  $\lambda > 1$ , most recent captures have a greater impact on the summary. A positive value of  $\beta_g$  accounts for trap-happiness type of response to capture while a negative value accounts for trap-shyness.

### 3 Unconditional maximum likelihood estimation

Following [10], if  $\mathbf{P} = [p_{ij}^{(g)}]$  is the matrix of capture probabilities, the unconditional likelihood for the model (2) is

$$L(N, \mathbf{P}, \boldsymbol{\pi}) = \frac{N!}{(N-M)!} \prod_{i=1}^N \prod_{j=1}^t \sum_{g=1}^G \pi_g [p_{ij}^{(g)}]^{x_{ij}} [1 - p_{ij}^{(g)}]^{1-x_{ij}}. \quad (5)$$

Once the number of mixture components  $G$  is fixed, inference on  $N$  is made through iterative fitting of the mixture of logistic regressions for each  $N \in \{M, \dots, N_{\max}\}$ , where  $N_{\max}$  is a high fixed upper bound for the population size [3]. The unconditional MLE (UMLE) for  $N$  is then the maximizer of the profile likelihood function

$$\hat{L}(N) = L(N, \hat{\mathbf{P}}(N), \hat{\boldsymbol{\pi}}(N)) = \sup_{\mathbf{P}, \boldsymbol{\pi}} L(N, \mathbf{P}, \boldsymbol{\pi}), \quad (6)$$

where the matrix  $\mathbf{P}$  is function of the regression parameters  $\alpha_1, \dots, \alpha_G, \beta_1, \dots, \beta_G$ . Details about fitting of finite mixtures of GLMs are available in [7].

## 4 Illustration

A real data example and a simulation study are presented in the following, in order to show how the proposed model performs.

### 4.1 Real data example

One considers a data set coming from a survey in which snowshoe hares (*Lepus americanus*) were repeatedly captured during 6 consecutive days of trapping by using animal-baited traps. At the end of the sixth day, the number of observed individual hares was 68. The considered dataset has already been analysed by some authors (e.g. [1, 5]) and it is available in R package `Rcapture`.

The proposed model is fitted to hares' capture histories for different numbers of mixture components ( $G = 1, 2, 3$ ) and for different values of  $\lambda$  (i.e.  $\lambda = 1, 2$ ). The choice of  $\lambda = 1$  yields a time-dependent covariate which represents the relative frequency of the previous capture occurrences, while  $\lambda = 2$  yields to a time-dependent covariate which enjoys a connection with Markov models [3]. For fixed  $G$  and  $\lambda$ , several finite mixtures of logit regressions are fitted for a set of candidate values of  $N$ , by using the functions in the R package `flexmix`: in particular, the function `initFlexmix` allows to repeat the EM algorithm with different starting values and chooses the solution which maximizes the likelihood.

The results displayed in Table 1 show that the models associated with the lowest values of the AIC are the ones corresponding to  $G = 2$  components. This is somewhat expected since other authors - like [5] - have already shown the presence of groups of hares with different capture rates. The model with  $G = 2$  and  $\lambda = 1$  yields  $\hat{\alpha}_1 = -1.45$ ,  $\hat{\beta}_1 = 4.12$ ,  $\hat{\alpha}_2 = -0.75$  (all of them associated to a  $p$ -value smaller than  $7 \times 10^{-3}$ ) and  $\hat{\beta}_2 = -0.75$ , which appears not to be significantly different by 0 ( $p = 0.28$ ). These results suggest that initial trap-happiness characterises the first group of hares, while for the second group no sufficient evidence of behavioural effects is provided. This indicates that a more parsimonious two-components mixture model with only one group manifesting behavioural effects could be further elaborated.

The 90% profile likelihood confidence intervals are built following [4], who highlights their advantages in a mark-recapture context. Notice that as the number of components increases, the confidence intervals tend to get wider, due to the flatter shape of the corresponding profile log-likelihood. This feature is probably due to the fact that the information provided by the data is insufficient to establish any upper bound on the number of animals, above all when a complex model is fitted on data coming from a relative low number of occasions [8].

Table 1: Unconditional maximum likelihood estimates for the population size, 90% confidence intervals and AIC index associated with alternative fitted models for different values of  $G$  and  $\lambda$ .

$G$	$\lambda$	$\hat{N}$	$(N_{\text{low}}, N_{\text{upp}})$	AIC
1	1	80	(73, 94)	81.53
	2	78	(72, 89)	83.20
2	1	79	(71, 197)	<b>75.72</b>
	2	75	(70, 111)	76.69
3	1	80	(72, 178)	81.39
	2	76	(70, 146)	82.09

### 4.2 Simulation study

Motivated by the results of the previous example, a simulation study is carried out in order to assess the ability of the proposed model in estimating the population size. Capture histories are generated for two subpopulations of individuals (thus  $G = 2$ ) and collected binary entries matrix with  $N = 100$  rows and  $t$  columns, where  $N - M$  rows have zero entries. The probability that an individual belongs to the first group is  $\pi_1 = 0.33$  and the regression parameters are set to  $\alpha_1 = -3$ ,  $\beta_1 = -2$ ,  $\alpha_2 = -3$  and  $\beta_2 = 4$ . Since the probability of first capture is completely determined by the value of the intercept  $\alpha$ , one is implicitly assuming that the first capture probability is the same for all the individuals of the population, regardless of the group they belong to. The replication of 20 simulated datasets has been carried out, for different time-dependent covariate specifications ( $\lambda = 1, 2$ ) and for different number of occasions ( $t = 15, 30$ ). For each data set, the true data-generating process is fitted to the data. From the results reported in 2, it appears that the empirical confidence intervals coverage is consistent with its theoretical counterpart. The population size seems to be slightly overestimated, though the bias decreases with the number of occasions, as expected.

Table 2: Simulation study with 20 simulated data sets for several model specifications, determined by different numbers of occasions ( $t = 15, 30$ ) and different values of  $\lambda$ . The table contains: average and median of the UML estimates of  $N$  (respectively,  $N_{\text{ave}}$  and  $N_{\text{med}}$ ), root mean square error (*rmse*), percentage of 95% confidence intervals coverage (*CI coverage*), average length of the confidence intervals ( $l_{CI}$ ).

$t$	$\lambda$	$N_{\text{ave}}$	$N_{\text{med}}$	$l_{CI}$	CI coverage	rmse
15	1	110.0	88.0	122.6	0.95	49.0
	2	113.9	120.0	71.5	0.95	31.7
30	1	104.7	104.5	56.2	0.90	14.5
	2	108.9	100.5	46.4	0.95	22.7



## 5 Final remarks and further developments

The proposed model appears a flexible extension of the one proposed in [3], allowing for the presence of latent heterogeneity between groups of individuals by means of group-specific regression parameters. Some possible further developments should involve a more in-depth study of the groups composition, along with a more flexible and parsimonious model which accounts for the possibility that some groups are not subject to behavioural effects, as suggested from the real data example. Moreover, a more extensive simulation study should be carried out, mainly in order to assess whether a model misspecification could be correctly identified when the population is composed by heterogeneous groups. Still through simulation studies, it can be interesting to investigate whether the better performances (in terms of AIC) of the proposed model on real data are indeed reliable; or whether, on the other hand, the AIC may tend to favour one model against the other. A Bayesian alternative might be proposed too, in order to overcome possible annoying problems due to the flatness of the profile likelihood when  $G$  is large.

## References

- [1] Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, pages 494–500.
- [2] Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635.
- [3] Alunni Fegatelli, D. and Tardella, L. (2016). Flexible behavioral capture-recapture modeling. *Biometrics*, 72(1):125–135.
- [4] Cormack, R. M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics*, pages 567–576.
- [5] Dorazio, R. M. and Andrew Royle, J. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59(2):351–364.
- [6] Farcomeni, A. (2011). Recapture models under equality constraints for the conditional capture probabilities. *Biometrika*, 98(1):237–242.
- [7] Grün, B. and Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in  $r$ . *Computational Statistics & Data Analysis*, 51(11):5247–5252.
- [8] Hirst, D. (1994). An improved removal method for estimating animal abundance. *Biometrics*, pages 501–505.
- [9] Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140.
- [10] Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56(2):434–442.
- [11] Wedel, M. and DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of classification*, 12(1):21–55.

# Finite mixtures of regressions with random covariates using multivariate skewed distributions

## *Misture di regressione a covariate casuali con distribuzioni multivariate asimmetriche*

Salvatore D. Tomarchio, Michael P.B. Gallagher, Antonio Punzo and Paul D. McNicholas

**Abstract** Finite mixtures of regressions with random covariates (CWMs) are a common model-based clustering methodology. Despite a lot of distributions have been considered for both the responses and covariates, skewed distributions have not yet been considered in this framework. Here, a family of 24 novel CWMs is introduced, where both the covariates and response variables are modeled by using one of the four considered skewed distributions, or the Gaussian distribution. A simulated data example is illustrated.

**Abstract** *Le misture di regressione a covariate casuali (CWMs) sono una nota metodologia di model-based clustering. Nonostante diverse distribuzioni siano state considerate sia per le variabili risposta che per le covariate, le distribuzioni asimmetriche non sono ancora state usate in tale contesto. Quindi, una famiglia di 24 CWMs viene qui introdotta, dove sia le variabili risposta che le covariate sono modellate usando una delle quattro distribuzioni asimmetriche considerate, oppure usando la distribuzione Gaussiana. Viene fornito un esempio su dati simulati.*

**Key words:** Finite mixture of regressions, Random covariates, Skewed distributions

---

Salvatore D. Tomarchio,  
Department of Economics and Business, University of Catania, Catania, Italy,  
e-mail: daniele.tomarchio@unict.it.

Michael P.B. Gallagher,  
Department of Statistical Science, Baylor University, Waco, Texas, USA.

Antonio Punzo,  
Department of Economics and Business, University of Catania, Catania, Italy.

Paul D. McNicholas,  
Department of Mathematics and Statistics, McMaster University, Ontario, Canada.

## 1 Introduction

Finite mixture models are one of the most prevalent model-based clustering technique. When no exogenous variables explain the location and the variability of each component, they are also called unconditional mixture models. However, when there is a regression relationship between the variables, important insight can be gained by accounting for functional dependencies between them. In such a framework, finite mixtures of regression models with fixed covariates (FMR) have been proposed in the literature (see, e.g. [3]; [4]). As in traditional regression analysis, FMRs assume that the covariates are fixed, and therefore they do not explicitly use the distribution of the covariates for clustering, i.e., they assume the so-called assignment independence [12].

As an alternative to this approach, finite mixtures of regression models with random covariates (CWM) offer far more flexibility, since the distribution of the covariates is taken into account [7]. Specifically, for each mixture component, CWMs decompose the joint distribution of responses and covariates into the product between the marginal distribution of the covariates and the conditional distribution of the responses given the covariates. The vast majority of the existing CWMs consider a univariate response variable and a set of covariates, modeled by a univariate and a multivariate distribution, respectively (see, e.g. [8, 9, 11, 10]). To our knowledge, only [1] consider a set of response variables and covariates, both modeled via multivariate Gaussian distributions. In this work, we extend this branch of the literature by considering four multivariate skewed distributions for both the responses and the covariates. By including also the Gaussian distribution, and by considering all the possible combinations, a novel family of 24 CWMs is obtained. In this way, we endow enough flexibility to CWMs for considering scenarios where both the responses and the covariates are skewed, or in which one of the two sets of variables is Gaussian distributed and the other is skewed. Section 2 gives some details concerning the CWM and the four skewed distributions used herein, whereas Section 3 illustrates an example on simulated data.

## 2 Methodology

Let  $\mathbf{Y}_i$  be a  $p$ -dimensional random vector of response variables and  $\mathbf{X}_i$  a  $d$ -dimensional random vector of covariates, for a sample of  $N$  observations, with  $i \in \{1, \dots, N\}$ . Let also assume that the sample can be partitioned into  $G$  groups. In a CWM framework, the joint density of  $\mathbf{Y}_i$  and  $\mathbf{X}_i$  is written as

$$p(\mathbf{x}_i, \mathbf{y}_i \mid \vartheta) = \sum_{g=1}^G \pi_g p_{\mathbf{X}}(\mathbf{x}_i \mid \phi_g) p_{\mathbf{Y}}(\mathbf{y}_i \mid \mathbf{x}_i, \theta_g), \quad (1)$$

where  $p_{\mathbf{X}}(\cdot)$  is the density function for  $\mathbf{X}_i$  parameterized by  $\phi_g$ ,  $p_{\mathbf{Y}}(\cdot)$  is the density function of  $\mathbf{Y}_i \mid \mathbf{X}_i$  parameterized by  $\theta_g$  and  $\vartheta$  contains all the parameters of

Title Suppressed Due to Excessive Length

the model. Furthermore, and in each group, the conditional expectation  $E(\mathbf{Y}_i | \mathbf{X}_i)$  is a linear function of  $\mathbf{X}_i$  depending on a  $\mathbf{B}_g$  matrix of coefficients of  $(1+d) \times p$  dimensionality.

As already mentioned in Section 1, [1] use multivariate Gaussian distributions both for  $p_{\mathbf{X}}(\cdot)$  and  $p_{\mathbf{Y}}(\cdot)$ . However, this assumption is too restrictive and can lead to overfitting issues when asymmetry is present in the data. For this reason, other than the multivariate Gaussian distribution, and being  $\mathbf{Z}$  a continuous random vector, we allow the following skewed distributions to be used in (1):

- the skew- $t$ , denoted by  $ST(\mu, \alpha, \Sigma, \nu)$ , with pdf

$$f_{ST}(\mathbf{z} | \vartheta) = \frac{2 \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \exp\{(\mathbf{z} - \mu)' \Sigma^{-1} \alpha\}}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\delta(\mathbf{z}; \mu, \Sigma) + \nu}{\rho(\alpha, \Sigma)}\right)^{-\frac{\nu+p}{4}} \\ \times K_{-\frac{\nu+p}{2}}\left(\sqrt{[\rho(\alpha, \Sigma)] [\delta(\mathbf{z}; \mu, \Sigma) + \nu]}\right),$$

where

$$\delta(\mathbf{z}; \mu, \Sigma) = (\mathbf{z} - \mu)' \Sigma^{-1} (\mathbf{z} - \mu), \quad \rho(\alpha; \Sigma) = \alpha' \Sigma^{-1} \alpha,$$

and  $\nu > 0$ ;

- the generalized hyperbolic, denoted by  $GH(\mu, \alpha, \Sigma, \lambda, \omega)$ , with pdf

$$f_{GH}(\mathbf{z} | \vartheta) = \frac{\exp\{(\mathbf{z} - \mu)' \Sigma^{-1} \alpha\}}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} K_{\lambda}(\omega)} \left(\frac{\delta(\mathbf{z}; \mu, \Sigma) + \omega}{\rho(\alpha, \Sigma) + \omega}\right)^{\frac{(\lambda - \frac{p}{2})}{2}} \\ \times K_{(\lambda - p/2)}\left(\sqrt{[\rho(\alpha, \Sigma) + \omega] [\delta(\mathbf{z}; \mu, \Sigma) + \omega]}\right),$$

$\lambda \in \mathbb{R}$ ,  $\omega \in \mathbb{R}^+$ ;

- the variance gamma, denoted by  $VG(\mu, \alpha, \Sigma, \gamma)$ , with pdf

$$f_{VG}(\mathbf{z} | \vartheta) = \frac{2\gamma^{\gamma} \exp\{(\mathbf{z} - \mu)' \Sigma^{-1} \alpha\}}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} \Gamma(\gamma)} \left(\frac{\delta(\mathbf{z}; \mu, \Sigma)}{\rho(\alpha, \Sigma) + 2\gamma}\right)^{\frac{(\gamma - p/2)}{2}} \\ \times K_{(\gamma - \frac{p}{2})}\left(\sqrt{[\rho(\alpha, \Sigma) + 2\gamma] [\delta(\mathbf{z}; \mu, \Sigma)]}\right),$$

where  $\gamma \in \mathbb{R}^+$ .

- the normal inverse Gaussian, denoted by  $NIG(\mu, \alpha, \Sigma, \kappa)$ , with pdf

$$f_{NIG}(\mathbf{z} | \vartheta) = \frac{2 \exp\{(\mathbf{z} - \mu)' \Sigma^{-1} \alpha + \kappa\}}{(2\pi)^{\frac{p+1}{2}} |\Sigma|^{\frac{1}{2}}} \left(\frac{\delta(\mathbf{z}; \mu, \Sigma) + 1}{\rho(\alpha, \Sigma) + \kappa^2}\right)^{-(1+p)/4} \\ \times K_{-(1+p)/2}\left(\sqrt{[\rho(\alpha, \Sigma) + \kappa^2] [\delta(\mathbf{z}; \mu, \Sigma) + 1]}\right),$$

where  $\kappa \in \mathbb{R}^+$ .

Parameter estimation is carried via the expectation-maximization (EM) algorithm [2]. Closed form expressions are derived for all the parameters involved, with the exclusion of  $\nu$ ,  $\lambda$ ,  $\omega$  and  $\gamma$  which are numerically estimated. The EM algorithm is initialized in two different ways: 10 times using a random soft initialization and once with a  $k$ -means hard initialization. The solution providing the highest log-likelihood value is selected.

### 3 A simulated data example

For illustrative purposes, we consider the ST-ST CWM, ST-N CWM and N-ST CWM (notice that the label “N” identifies the Gaussian distribution). In this way, we are able to cover the following different scenarios

1.  $p_{\mathbf{X}}$  and  $p_{\mathbf{Y}}$  are the same skewed density;
2.  $p_{\mathbf{X}}$  is skewed and  $p_{\mathbf{Y}}$  is Gaussian;
3.  $p_{\mathbf{X}}$  is Gaussian and  $p_{\mathbf{Y}}$  is skewed.

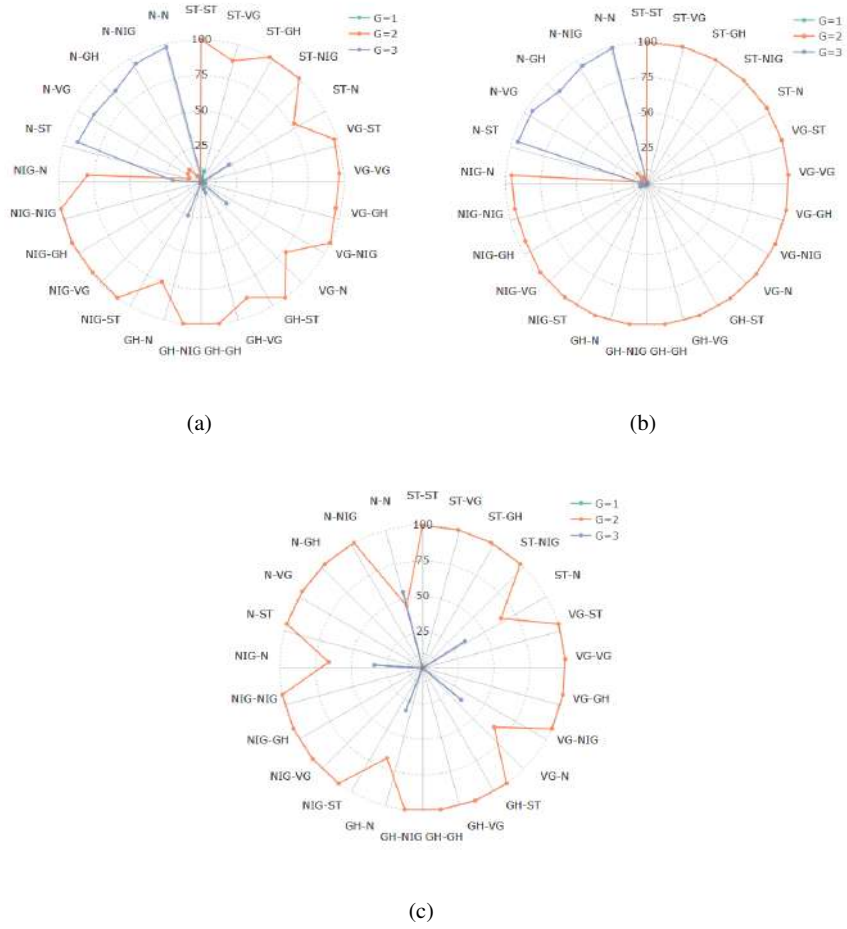
We set  $p = 2, d = 3, N = 400$  and  $G = 2$ . Then, for each of these 3 models, 100 datasets are generated and all the 24 novel CWMs, along with the N-N CWM are fitted for  $G \in \{1, 2, 3\}$ .

The results are illustrated in the radar plots of Fig. 1. Specifically, each sub-plot illustrates the number of times each  $G$  is chosen by the BIC for each model over the 100 datasets. Starting with Fig. 1(a), we notice that when the data are generated by the ST-ST CWM, all the CWMs for which either  $p_{\mathbf{X}}$ ,  $p_{\mathbf{Y}}$ , or both are assumed to be Gaussian, face problems in detecting the true  $G$  in the data. As mentioned in Section 2, when the Gaussian distribution is used for modeling skewed data, it has a tendency to overfit the true number of groups. This is confirmed by our results, but it is also interesting to note that this issue has a different magnitude depending on which one of  $p_{\mathbf{X}}$  or  $p_{\mathbf{Y}}$  is modeled using the Gaussian distribution. Specifically, when  $p_{\mathbf{X}}$  is assumed to be skewed and  $p_{\mathbf{Y}}$  assumed to be Gaussian, most of the time  $G = 2$  is properly selected, even if it is still not as accurate as the CWMs where both  $p_{\mathbf{X}}$  and  $p_{\mathbf{Y}}$  are assumed to be skewed. On the contrary, when  $p_{\mathbf{X}}$  is assumed Gaussian and  $p_{\mathbf{Y}}$  assumed skewed,  $G = 3$  is nearly always chosen. This seems to suggest that the consequences are greater if  $p_{\mathbf{X}}$  is misspecified as Gaussian when it should be skewed, compared to the case when  $p_{\mathbf{Y}}$  is misspecified. Clearly, the N-N CWM is the worst model as both  $p_{\mathbf{X}}(\cdot)$  and  $p_{\mathbf{Y}}(\cdot)$  are misspecified.

When the datasets are generated from a ST-N CWM, the only models facing issues are those for which the covariates are assumed to be Gaussian distributed, as shown in Fig. 1(b). Because of their greater flexibility, all the CWMs that assume a skewed density for  $p_{\mathbf{Y}}$  are able to accurately model symmetric data. The results for the N-ST CWM are displayed in Fig. 1(c). Here, the only CWMs that present issues are those for which  $p_{\mathbf{Y}}$  is assumed Gaussian.

It is clear that such issues have also an effect on the underlying data classifications, that are not reported here for the sake of brevity. In particular way, the CWMs

Title Suppressed Due to Excessive Length



**Fig. 1** Radar plots of the number of times each  $G$  is chosen by the BIC, for the CWMs, when the data are generated from (a) ST-ST CWM, (b) ST-N CWM and (c) N-ST CWM. Each sub-plot refers to 100 datasets.

assuming the Gaussian distribution for  $p_X$  have bad performances when the data are generated from the ST-N and ST-ST models. Conversely, all the other CWMs (N-N CWM excluded, which is the worst) produce good classifications under all the three data generating models considered.

## 4 Conclusions

A novel family of 24 CWMs has been introduced, where the covariates and response variables have modeled by using the multivariate skew- $t$ , generalized hyperbolic, variance gamma, normal inverse Gaussian and Gaussian distributions. The main risks of using only the Gaussian distribution in CWMs, as it has done so far in the literature, rely on the overfitting tendency of such distribution and the destruction of the underlying group structure when data are skewed. Conversely, our models have proven to be flexible enough to adequately model scenarios where both the responses and the covariates are skewed or in which one of the two sets of variables is normally distributed and the other is skewed. An interesting point for further extensions could be the generalization of our CWMs to the matrix-variate paradigm [5, 6].

## References

1. Dang, U. J., Punzo, A., McNicholas, P. D., Ingrassia, S., & Browne, R. P. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, 34(1), 4-34.
2. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1-22.
3. DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, 5(2), 249-282.
4. Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.
5. Gallagher, M.P.B. & McNicholas P. D. (2017). A matrix variate skew- $t$  distribution. *Stat*, 6(1), 160-170.
6. Gallagher, M.P.B. & McNicholas P. D. (2018). Three skewed matrix variate distributions. *Statistics and Probability Letters*, 145, 103-109.
7. Gershenfeld, N. (1997). Nonlinear Inference and Cluster-Weighted Modeling. *Annals of the New York Academy of Sciences*, 808(1), 18-24.
8. Ingrassia, S., Minotti, S. C., & Vittadini, G. (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of classification*, 29(3), 363-401.
9. Ingrassia, S., Punzo, A., Vittadini, G., & Minotti, S. C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, 32(2), 327-355.
10. Počuča, N., Jevtić, P., McNicholas, P. D., & Miljkovic, T. (2020). Modeling frequency and severity of claims with the zero-inflated generalized cluster-weighted models. *Insurance: Mathematics and Economics*, 94, 79-93.
11. Punzo, A., & Ingrassia, S. (2016). Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, 31(3), 989-1013.
12. Punzo, A., & McNicholas, P. D. (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification*, 34(2), 249-293.

## 4.15 New applications of regression models



# The Shapley-Lorenz decomposition approach to mitigate cyber risks

## *La scomposizione di Shapley-Lorenz per la mitigazione dei rischi informatici*

Paolo Giudici and Emanuela Raffinetti

**Abstract** Artificial Intelligence (AI) methods are gaining an increasing relevance, especially in all the contexts where data are available only on ordinal scale, as it happens in the cyber risk management area. However, AI methods are not suitable in the most regulated fields, due to their black box nature resulting in automated decision processes. In order to avoid that wrong actions can be taken as a consequence of automatic decision making, eXplainable Artificial Intelligence (XAI) methods are required to clearly explain the reasons underlying the detection of certain factors impacting on the predictions. In this contribution, we present a novel XAI approach which is the result of a combination between the Shapley value-based approach and the Lorenz Zonoid tool. The proposed methodology appears agnostic, in the sense that it does not depend on the nature of data and on the adopted model. Moreover, due to the Lorenz Zonoid properties, it provides results which are easy to be interpreted.

**Abstract** *I metodi di Intelligenza Artificiale stanno assumendo una rilevanza sempre maggiore, soprattutto in tutti i contesti in cui i dati siano disponibili solo su scala ordinale, come accade nell'area del cyber risk management.*

*Uno dei principali svantaggi legati all'impiego dei metodi di Intelligenza Artificiale riguarda l'impossibilità di comprendere completamente le intricate architetture di apprendimento automatico che li contraddistinguono. Al fine di evitare che azioni dannose possano essere intraprese come conseguenza di un processo decisionale automatico, i metodi di Intelligenza Artificiale devono essere "spiegabili", consentendo l'individuazione dei fattori che influiscono maggiormente sulle previsioni generate dai modelli di Machine Learning.*

*In questo contributo illustriamo un nuovo metodo di Intelligenza Artificiale spiega-*

---

Paolo Giudici

Department of Economics and Management, Via San Felice 5, 27100 Pavia (Italy), e-mail: paolo.giudici@unipv.it

Emanuela Raffinetti

Department of Economics and Management, Via San Felice 5, 27100 Pavia (Italy), e-mail: emanuela.raffinetti@unipv.it

*bile, che nasce dalla combinazione tra l'approccio basato sui valori di Shapley e l'approccio basato sugli Zonoidi di Lorenz. La metodologia proposta risulta agnostica, in quanto non dipende dalla natura dei dati e dal modello di Machine Learning adottato. Inoltre, grazie alle proprietà di cui godono gli Zonoidi di Lorenz, i risultati ottenuti appaiono facilmente interpretabili.*

**Key words:** Cyber risks; Lorenz Zonoids; Rank regression models; Shapley values

## 1 Introduction

Currently, we are assisting to an explosion of IT (Information and Technology) systems, especially due to the globalisation of financial services and technology advancements. The use of IT systems may cause cyber risks intended as any risk which can compromise the availability and integrity of data. In the last few years the number of cyber attacks on information technology systems has surged with a cyber attack growth of about 30% between 2014 and 2017. The trend in 2018 follows a similar behavior, with 730 cyber attacks observed only in the first half of the year (see [1]). Thus, cybersecurity has become a serious concern for businesses.

It is worth noting that even if literature on the operational risk quantitative measurement, based on losses data, represents a large body (see [2]; [6]), literature on the cyber risk measurement is limited especially due to the lack of data which are typically not disclosed or if disclosed, they are expressed through an ordinal scale denoting the levels of severity, such as “low”, “medium” or “high” severity.

The aim of this contribution is twofold: 1) in order to deal with the ordinal nature of the target variable describing the cyber attack severity degree, a rank regression model is employed with the purpose of detecting the main factors impacting on the severity degree; 2) a new eXplainable Artificial Intelligence (XAI) method, resulting from the combination of the Shapley value-based approach together with the Lorenz Zonoid tool, is proposed in order to measure the effect of any factor in explaining the cyber attack severity degree over all the possible machine learning model configurations.

The paper is structured as follows: Section 2 formalizes our proposed methodology; Section 3 focuses on the application of our proposal to real losses data, organised in terms of severity levels; finally Section 4 concludes the paper.

## 2 Methodological approach

In order to deal with the ordinal nature of the target variable, representing the cyber attack severity degree, we follow the approach suggested by [5], devoted to an appropriate extension of the linear regression models.

Let  $Y$  be an ordinal variable expressed according to  $h$  ordered categories. Based on

The Shapley-Lorenz decomposition approach to mitigate cyber risks

what proposed by [5], variable  $Y$  has to be transformed into a quantitative discrete variable through the employment of the rank tool. Thus, a rank  $r_1 = 1$  is assigned to the smallest ordered category of  $Y$  and a rank  $r_j = (r_{j-1} + n_{j-1})$  is assigned to the  $j$ -th ordered category, where  $r_{j-1}$  and  $n_{j-1}$  are the rank and the absolute frequency associated with the  $(j-1)$ -th category (with  $j = 2, \dots, h$ ). Based on this adjustment, the  $Y$  variable can be rewritten in terms of its ranks  $R$ , i.e.:

$$R = \left\{ \underbrace{r_1, \dots, r_1}_{n_1}, \underbrace{r_2, \dots, r_2}_{n_2}, \dots, \underbrace{r_h, \dots, r_h}_{n_h} \right\}. \quad (1)$$

Given  $p$  explanatory variables  $(X_1, \dots, X_p)$ , a regression model for  $R$  can be specified as follows

$$\hat{R} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p, \quad (2)$$

whose unknown parameters can be estimated by the classical OLS method. The need of fulfilling the requirements advanced by the most regulated fields (e.g., finance and health) leads to integrate the machine learning-based approaches with Artificial Intelligence methods which have to be explainable. Explainability means that the crucial drivers of a model decision have to be fully understood by the interested stakeholders. In this perspective, we propose to extend the Shapley-Lorenz decomposition, recently introduced by [4] in the case of a continuous target variable, to cover the case of an ordinal target variable. The Shapley-Lorenz decomposition results from the combination between the Shapley-value based approach and the Lorenz Zonoid. As discussed by [3], the Lorenz Zonoid represents a generalisation of the ROC curve in the multidimensional context and, therefore, the Shapley-Lorenz decomposition has the advantage of combining predictive accuracy and explainability performance into one single diagnostics (e.g., [4]). In addition, the Lorenz Zonoid appears as a measure of the mutual variability and, consequently, it is more robust in presence of outlying observations.

Let us suppose that our purpose is to study a phenomenon through a machine learning model defined as a function of  $K$  predictors. In statistical data analysis, the model specification is typically followed by a procedure aimed at selecting the explanatory variables which mainly impact on the target variable. In [3], the Lorenz Zonoid tool was exploited to formalize partial dependence measures that allow to detect the additional contribution of a new predictor into an existing model according to a stepwise selection procedure. The Shapley-Lorenz decomposition leads to detect the marginal contribution provided by the inclusion of the generic covariate  $X_k$  into an existing model over all the possible model configurations.

More formally, given a predictor  $X_k$ , with  $k = 1, \dots, K$ , the additional contribution in explaining the rank transformed response variable ordering equals to

$$LZ^{X_k}(\hat{R}) = \sum_{X' \subseteq \mathcal{C}(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [LZ(\hat{R}_{X' \cup X_k}) - LZ(\hat{R}_{X'})], \quad (3)$$

where  $LZ(\hat{R}_{X' \cup X_k})$  and  $LZ(\hat{R}_{X'})$  measure the marginal contribution provided by the inclusion of variable  $X_k$ ;  $K$  is the number of available predictors;  $\mathcal{C}(X) \setminus X_k$  is the set of all the possible model configurations which can be obtained with  $K - 1$  variables, excluding variable  $X_k$ ;  $|X'|$  denotes the number of variables included in each possible model.

It is worth noting that  $LZ(\hat{R}_{X' \cup X_k})$  and  $LZ(\hat{R}_{X'})$  in equation (3) can be expressed as function of the covariance operators, i.e.,

$$LZ(\hat{R}_{X' \cup X_k}) = \frac{2}{nE(\hat{R}_{X' \cup X_k})} Cov(\hat{R}_{X' \cup X_k}, r(\hat{R}_{X' \cup X_k})) \quad \text{and}$$

$$LZ(\hat{R}_{X'}) = \frac{2}{nE(\hat{R}_{X'})} Cov(\hat{R}_{X'}, r(\hat{R}_{X'})),$$

where:  $n$  is the total number of the involved observations;  $E(\hat{R}_{X' \cup X_k})$  and  $E(\hat{R}_{X'})$  are the expected values of  $\hat{R}_{X' \cup X_k}$  and  $\hat{R}_{X'}$ ;  $r(\hat{R}_{X' \cup X_k})$  and  $r(\hat{R}_{X'})$  are the rank scores.

### 3 Application to cyber risk data

In this section we consider real cyber loss data, organised in terms of different cyber attack severity levels, to evaluate the performance of our cyber risk measurement, based on the combination between rank regression models and the Shapley-Lorenz decomposition approach.

The available data are provided by the Italian Association for Cybersecurity (e.g., [1]), and consist of 6,865 worldwide observations on serious cyber attacks, in the years 2011-2017. Here we focus on a sample data, consisting of 808 cyber attacks observed in 2017, and as potential factors impacting on the severity levels, we consider the type of attacker, technique of attacks, victims and the corresponding continent of origin.

In order to detect the factors, among attacker, attack technique, victim type and location (continent), which most affect the severity levels, we have applied our proposed rank regression model, provided by equation (2), and we have calculated the Shapley-Lorenz marginal contributions associated with the variables attacker, victim type, attack technique and continent, using formula (3). For the sake of clarity, here we report only the extended computation of the Shapley-Lorenz values when measuring the marginal contribution associated with the type of attacker (Att) variable which is included in all the possible model configurations containing the remaining predictors, i.e. victim type (Vic), attack technique (Tec) and continent (Con):

The Shapley-Lorenz decomposition approach to mitigate cyber risks

$$\begin{aligned}
LZ^{Att}(\widehat{Severity}) &= (1/4)(LZ(\hat{R}_{Att,Vic,Tec,Con}) - LZ(\hat{R}_{Vic,Tec,Con})) \\
&+ (1/12)(LZ(\hat{R}_{Att,Vic,Tec}) - LZ(\hat{R}_{Vic,Tec})) \\
&+ (1/12)(LZ(\hat{R}_{Att,Vic,Con}) - LZ(\hat{R}_{Vic,Con})) + (1/12)(LZ(\hat{R}_{Att,Tec,Con}) - LZ(\hat{R}_{Tec,Con})) \\
&+ (1/12)(LZ(\hat{R}_{Att,Vic}) - LZ(\hat{R}_{Vic})) + (1/12)(LZ(\hat{R}_{Att,Tec}) - LZ(\hat{R}_{Tec})) \\
&+ (1/12)(LZ(\hat{R}_{Att,Con}) - LZ(\hat{R}_{Con})) + (1/4)(LZ(\hat{R}_{Att})).
\end{aligned}$$

The results, both in terms of Shapley-Lorenz values and global Shapley values, are displayed in Table 1. The local Shapley values provide a measure of the marginal contribution related to a given predictor in explaining the predictions at single unit level. If the local Shapley values are summed across units, a ‘‘global’’ variable importance measure is derived. By comparing the global Shapley values with the Shapley-Lorenz values, the main issue that arises is that, contrary to the global Shapley values the Shapley-Lorenz values are normalised and thus easier to be interpreted.

Table 1: Marginal contribution of each predictor in terms of the Shapley-Lorenz values and the global Shapley values

Additional covariate ( $X_k$ )	$LZ_{d=1}^{X_k}(\widehat{Severity})$	Global Shapley
Type of attacker	0.072	-748.96
Type of victim	0.115	5.15
Technique of attack	0.058	-34.36
Continent	0.032	-25.67

From Table 1, it results that according to the Shapley-Lorenz values, the variable describing the type of victim provides the highest marginal contribution in the prediction of cyber severity, across all the possible model configurations. A further relevant variable is represented by the type of attacker. If resorting to the global Shapley values we note that they are characterised by a sign that is positive only in the case of the victim type, indicating that victim type increases the mean loss severity, differently from the others.

## 4 Conclusions

In this paper, a new methodology to assess cyber risks in the case of loss data expressed on an ordinal scale, is illustrated. The proposed approach combines rank regression models with a novel XAI method based on the combination between the Shapley value-based formula and the Lorenz Zonoid tools. The application of our methodology leads to the detection of the main factors impacting on the cyber attack severity degree, allowing to adopt the best practices to mitigate cyber risks.

## References

1. Clusit 2018: Report on ICT security in Italy (2018) (in Italian), [https://web.uniroma1.it/infosapienza/sites/default/files/rapporto\\_clusit\\_2018.pdf](https://web.uniroma1.it/infosapienza/sites/default/files/rapporto_clusit_2018.pdf)
2. Cox Jr, L.T.A.: Evaluating and improving risk formulas for allocating limited budgets to expensive risk-reduction opportunities. *Risk Anal.*, **32(7)**, 1244–1252 (2012)
3. Giudici, P., Raffinetti, E.: Lorenz Model Selection. *J. Classif.*, **37**, 754–768 (2020a)
4. Giudici, P., Raffinetti, E.: Shapley-Lorenz eXplainable Artificial Intelligence. *Expert Syst. Appl.* (2020b)
5. Giudici, P., Raffinetti, E.: Cyber risk ordering with rank-based statistical models. *AStA Adv. Stat. Anal.* (2020c).
6. MacKenzie, CA.: Summarizing Risk Using Risk Measures and Risk Indices. *Risk Anal.*, **34(12)**, 2143–2162 (2014)

# A spatially adaptive estimator for the function-on-function linear regression model with application to the Swedish Mortality dataset

*Uno stimatore spazialmente adattivo per il modello di regressione lineare con regressore e risposta funzionale con un'applicazione al dataset Swedish Mortality*

Fabio Centofanti, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, Simone Vantini

**Abstract** In this work, we consider a spatially adaptive smoothing spline estimator for the function-on-function linear regression model where each value of the response, at any domain point, depends on the full trajectory of the predictor. The considered estimator adapts more easily to the true coefficient function over regions of large curvature and does not undersmooth over the remaining part of the domain because of two spatially adaptive penalties. The latter are based on initial estimates of the partial derivatives of the regression coefficient function. The performance of the proposed estimator is analysed by means of the well-known Swedish Mortality dataset.

**Abstract** *In questo lavoro, viene presentato uno stimatore spazialmente adattivo per il modello di regressione lineare funzionale, in cui ogni valore della risposta, in un qualsiasi punto del dominio, dipende dalla traiettoria completa del predittore. Lo stimatore considerato è capace di adattarsi con maggiore flessibilità ai cambiamenti di curvatura del coefficiente di regressione funzionale grazie all'introduzione di due penalità spazialmente adattive. Tali penalità vengono determinate a partire da stime iniziali delle derivate parziali del coefficiente di regressione. La bontà dello stimatore viene analizzata tramite il dataset Swedish Mortality.*

**Key words:** functional data analysis, function-on-function linear regression, adaptive smoothing, functional regression

---

Fabio Centofanti, Antonio Lepore, Biagio Palumbo  
Dept. of Industrial Engineering, University of Naples Federico II, P.le V. Tecchio 80, 80125, Naples, Italy  
e-mail: [fabio.centofanti@unina.it](mailto:fabio.centofanti@unina.it), [antonio.lepore@unina.it](mailto:antonio.lepore@unina.it), [biagio.palumbo@unina.it](mailto:biagio.palumbo@unina.it)

Alessandra Menafoglio, Simone Vantini  
MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy  
e-mail: [alessandra.menafoglio@polimi.it](mailto:alessandra.menafoglio@polimi.it), [simone.vantini@polimi.it](mailto:simone.vantini@polimi.it)

## 1 Methods

Due to advances in technologies and computing capacity, complex datasets are increasingly available and have prompted important methodological developments. In this regard, functional data analysis (FDA) tackles the problem of working with data that can be modelled as functions defined on a compact domain [10, 7, 6, 8, 5]. In particular, the generalization of the classical multivariate regression analysis to the case where the predictor and/or the response have a functional form is referred to as functional regression [9, 10]. In this work, we analyse the function-on-function (FoF) linear regression models, where the response function depends linearly on the complete trajectory of the predictor at any domain point. That is,

$$Y_i(t) = \int_{\mathcal{S}} X_i(s) \beta(s, t) ds + \varepsilon_i(t) \quad t \in \mathcal{T}, \quad (1)$$

for  $i = 1, \dots, n$ . The pairs  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , are independent realizations of the predictor  $X$  and the response  $Y$ , which are smooth random processes with realizations in  $L^2(\mathcal{S})$  and  $L^2(\mathcal{T})$ , i.e., the Hilbert spaces of square integrable functions defined on the compact sets  $\mathcal{S}$  and  $\mathcal{T}$ , respectively, and are assumed, without loss of generality, to have zero mean. The functions  $\varepsilon_i$  are i.i.d. zero-mean random errors, independent of  $X_i$ . The function  $\beta$  is the *coefficient function*.

In this work, we consider the *spatially adaptive* estimator of  $\beta$  that has been proposed in [3] and referred to as *adaptive smoothing spline* (AdaSS) estimator. It is obtained as the minimum of an objective function, composed by the usual sum of squared errors plus two adaptive smoothness penalties. The latter rely on two roughness parameters, which are functions defined on  $\mathcal{S} \times \mathcal{T}$ , that produce different amount of penalty over the domain, and, thus, allow the estimator to spatially adapt, i.e., to take into account varying degree of roughness. For example, the AdaSS estimator may accommodate the local behavior of  $\beta$  by imposing a heavier penalty in regions of lower smoothness. The two roughness parameters are chosen based on an initial estimate of the partial derivatives of  $\beta$ . The reasoning behind this choice is to allow the penalization contribution to be small (large) over region where the initial estimated curvature (i.e. partial derivatives) is large (small). This is an extension to the FoF linear regression model of the idea of Storlie et al. [12] and Abramovich and Steinberg [1].

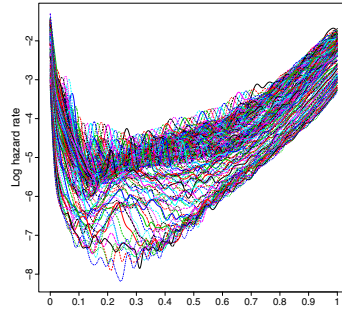
From a computational point of view, the unknown parameters, which are needed to compute the AdaSS estimator, are chosen by means of an evolutionary algorithm that aims to reduce the computational burden of the popular grid-search method [2].

## 2 Real-Data Example: Swedish Mortality dataset

In this section, we apply the AdaSS estimator to the *Swedish Mortality dataset* (<http://mortality.org>), which is very well known in the functional liter-



ature as benchmark dataset [4, 11]. In this analysis, we consider the log-hazard rate functions of the Swedish females mortality data for year-of-birth cohorts that refer to females born in 1751-1935 with age 0-80. The value of a log-hazard rate function at a given age is the natural logarithm of the ratio of females died and the number of females alive with that age. The 184 considered log-hazard rate functions [4] are shown in Figure 1. Without loss of generality, they have been normalized to the domain  $[0, 1]$ .



**Fig. 1** Log-hazard rate functions for Swedish female cohorts from 1751 to 1935.

The functions from 1751 (1752) to 1934 (1935) are considered as observations  $X_i$  ( $Y_i$ ) of the predictor (response) in (1),  $i = 1, \dots, 184$ . The relation between two consecutive log-hazard rate functions is the subject of the study.

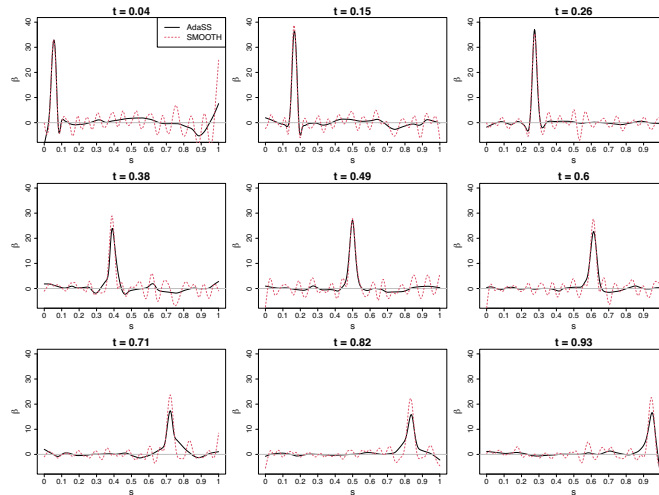
The AdaSS estimator is compared with the estimator proposed in [10], referred to as SMOOTH, with regularization obtained by introducing two roughness penalties. We aim to demonstrate that the AdaSS estimator has advantages, in terms of both prediction accuracy and interpretability, over the SMOOTH estimator.

To assess the predictive performance of both methods, for 100 times, 166 observations out of 184 are randomly chosen, as training set, to fit the model. The 18 remaining ones are used as test set to calculate the prediction mean squared error (PMSE). The averages and standard deviations of PMSEs are shown in Table 1. The AdaSS estimator turns out to outperform the SMOOTH estimator in terms of predictive performance.

**Table 1** The prediction mean squared error (PMSE) for the SMOOTH and AdaSS estimators. The numbers outside the parentheses are the averages of the PMSE over 100 replications, and the numbers inside parentheses are the corresponding standard errors.

	PMSE
SMOOTH	0.005938 (0.0000)
AdaSS	0.005706 (0.0000)

Figure 2 shows the AdaSS estimates along with the SMOOTH estimates for different values of  $t$ . The proposed estimator proves to be more interpretable than the competitor, being smooth where the coefficient function is likely flat, but still able to capture peaky patterns. On the contrary, the SMOOTH estimate is particularly rough over low-curvature regions.



**Fig. 2** AdaSS (solid line) and SMOOTH (dashed line) estimates of the coefficient functions for different values of  $t$  in the Swedish Mortality dataset.

In face of these results, the AdaSS estimator demonstrates to outperform the competitor in terms of both prediction accuracy and interpretability for the problem of estimating the relation between two consecutive log-hazard rate functions in the Swedish Mortality dataset.

## References

1. Abramovich, F., Steinberg, D.M.: Improved inference in nonparametric regression using lk-smoothing splines. *Journal of Statistical Planning and Inference* **49**(3), 327–341 (1996)
2. Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: *Advances in neural information processing systems*, pp. 2546–2554 (2011)
3. Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Adaptive smoothing spline estimator for the function-on-function linear regression model. *arXiv preprint arXiv:2011.12036* (2020)
4. Chiou, J.M., Müller, H.G.: Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association* **104**(486), 572–585 (2009)
5. Ferraty, F., Vieu, P.: *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media (2006)

6. Horváth, L., Kokoszka, P.: Inference for functional data with applications, vol. 200. Springer Science & Business Media (2012)
7. Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons (2015)
8. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. CRC Press (2017)
9. Morris, J.S.: Functional regression. *Annual Review of Statistics and Its Application* **2**, 321–359 (2015)
10. Ramsay, J., Silverman, B.: *Functional Data Analysis*. Springer Series in Statistics. Springer (2005)
11. Ramsay, J.O., Hooker, G., Graves, S.: *Functional data analysis with R and MATLAB*. Springer Science & Business Media (2009)
12. Storlie, C.B., Bondell, H.D., Reich, B.J.: A locally adaptive penalty for estimation of functions with varying roughness. *Journal of Computational and Graphical Statistics* **19**(3), 569–589 (2010)

# POSetR: a new computationally efficient R package for partially ordered data

*POSetR: un nuovo pacchetto R, ad alta efficienza computazionale, per dati parzialmente ordinati*

Alberto Arcagni, Alessandro Avellone, Marco Fattore

**Abstract** In this paper, we introduce `POSetR`, a new R package providing highly efficient routines for the treatment of partially ordered data. After motivating the need for a new package on posets, we describe the main functionalities of `POSetR` and give hints on its possible uses.

**Abstract** *Obiettivo di questo articolo è presentare il nuovo pacchetto `POSetR`, per il trattamento di dati parzialmente ordinati, in ambiente R. Dopo averne motivato la necessità, l'articolo descrive brevemente le principali funzionalità del pacchetto e ne indica i possibili utilizzi.*

**Key words:** Linear extensions, Mutual ranking probabilities, Partially ordered set, R.

## 1 Introduction

In this paper, we introduce `POSetR`, a new and efficient R [17] package for the analysis of partially ordered data. The package combines high level R instructions, with low-level core routines implemented in C++, so as to preserve user-friendliness, still assuring for high computational performances. In the following, we (i) briefly discuss the statistical relevance of partially ordered data, (ii) introduce existing R

---

Alberto Arcagni

Dipartimento Metodi e Modelli per l'Economia, il Territorio e la Finanza - Università di Roma La Sapienza, Via Del Castro Laurenziano 9 - 00161, ROMA. e-mail: alberto.arcagni@uniroma1.it

Alessandro Avellone

Dipartimento di Statistica e Metodi Quantitativi - Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8 - 20126 MILANO. e-mail: alessandro.avellone@unimib.it

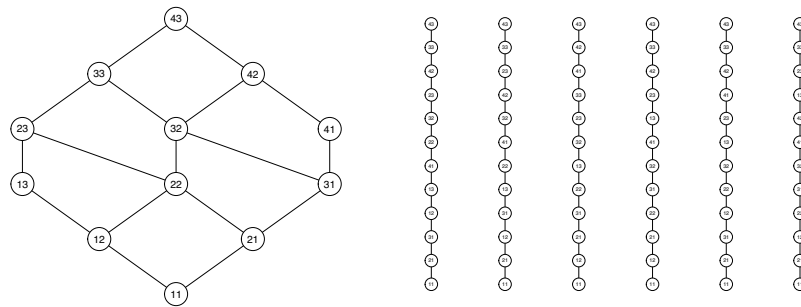
Marco Fattore

Dipartimento di Statistica e Metodi Quantitativi - Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8 - 20126 MILANO. e-mail: marco.fattore@unimib.it

resources for poset analysis, (iii) provide an overview of the new package and (iv) apply it to data of financial knowledge in Italy.

## 2 Why partially ordered sets?

Many problems in data analysis involve the treatment of multidimensional systems of ordinal indicators., e.g. for the construction of rankings and synthetic indicators, in contexts like the evaluation of multidimensional poverty, quality-of-life or customer satisfaction. Each ordinal indicator provides a *linear order* (possibly with ties) of the statistical units; in general, however, units are ordered differently by different indicators, preventing them to be “globally” ordered. However, they might be *partially* ordered. Indeed, if unit  $a$  gets “better” scores than unit  $b$  on all of the indicators, then  $a$  “dominates”  $b$ . Thus, some pairs of units can be ordered, some others cannot, producing a partially ordered set, or a *poset* (see [4] and Figure 1). Posets are the natural mathematical structure associated to ordinal multi-indicator systems and, in general, to data described in terms of comparabilities and incomparabilities (e.g. data on preferences); as such, they are also the natural setting to develop a sound “multidimensional ordinal data analysis”, as investigated in recent methodological advances [2, 11, 12, 13, 14, 15].



**Fig. 1** Hasse diagram of a poset built on two ordinal indicators with 4 and 3 degrees and examples of linear extensions.

### 3 Computational issues

Posets are combinatoric objects and often generate computationally non-trivial problems. In particular, many posetic data analysis procedures are essentially based on the construction of *linear extensions* (LEs) and on the computation of so-called *mutual ranking probabilities* (MRPs), two related concepts which in general involve heavy computations. The notion of LE of a poset is illustrated in Figure 1. Given a poset  $\pi$  built on a set  $X$ , a LE  $\ell$  of  $\pi$  is a linear order of the elements of  $X$  such that if  $a < b$  in  $\pi$ , then  $a < b$  in  $\ell$ ; more expressively,  $\ell$  is obtained by resolving all the incomparabilities of  $\pi$ , without affecting the dominances in it. It can be proven that any finite poset can be reconstructed from the set of its LEs [18], which then make it possible to decompose statistical problems on complex partial order structures, into “subproblems” on simple linear orders. For example, ranking extraction, scoring and evaluation over multidimensional ordinal indicator systems [2, 11, 15] are based on the computation of MRPs which in turn involves LE generation, MRPs being the fraction of LEs where an element, say  $a$ , dominates another element, say  $b$  [14, 15]. Since posets usually encountered in real applications have an extremely high number of LEs, these must be sampled [3]; sampled LEs can nevertheless be of the order of billions, making it crucial to have efficient implementations of sampling algorithms.

### 4 R resources for posetic analysis

Currently, there is just one R package devoted to posets, namely `parsec` [10], developed as a first software implementation of the procedures described in [11, 12, 13, 14, 15]. It implements a quite wide set of tools for basic mathematical analysis of partially ordered sets and for the statistical analysis of partially ordered data, but it is mainly designed to deal with posets built out of multidimensional indicator systems and has its major focus on multi-criteria evaluation. Although it provides quite efficient LE generation algorithms, imported from package `netrankr` [5] (which is not devoted to posets, but implements such algorithms for other purposes), `parsec` is not flexible enough, to effectively adapt to the incoming methodological advances and to the increasing range of statistical applications of posetic tools.

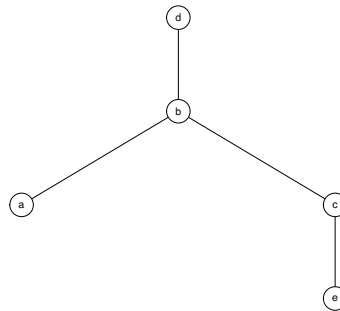
### 5 Overview of POSetR

POSetR is a new posetic package, internally written in C++ and integrated with R via the `Rcpp` package ([7], [8]), developed to provide a well-designed, efficient and flexible “engine” for LE generation. To introduce it, we show some of its main functionalities in action. To create a poset, the set of dominances between objects must be set, as a two-column matrix where elements in the first column are dominated

by corresponding elements in the second one, and passed to the poset constructor; function `summary` then provides synthetic infos on the poset:

```
> library(POSetR)
> dom <- matrix(c( "a", "b", "c", "b", "b", "d","e", "c" ),
+             ncol = 2, byrow = TRUE)
> p <- poset(dom, elements = c("a", "b", "c", "d", "e") )
> summary(p)
  5 elements
  8 strict comparabilities
  2 incomparabilities
```

A function `print` is also available, to get the list of dominances of the input poset. The Hasse diagram of the poset is provided by function `plot` (see Figure 2). There are also special functions to generate specific types of posets: `chain` for chains (i.e. linear orders) and `productOrder` for the product order of two posets (typically used to turn a multi-indicator system into a poset). The key function of the package is `LEapply`. It generates the LEs of the input poset, employing a high performance C++ code, *at the same time* evaluating an argument function on each extension and taking the average over the set of LEs. The function passed as an argument to `LEapply` can be any user-defined R function or a function implemented in the C++ library of the package (the fundamental function `MutualRankingProbability` is one of them):



**Fig. 2** Hasse diagram of poset `p` produced by the `plot` command.

```
> LEapply(x = p, FUN = "MutualRankingProbability")
  a    b    c    d    e
a 1.00 1.00 0.66 1.00 0.33
b 0.00 1.00 0.00 1.00 0.00
c 0.33 1.00 1.00 1.00 0.00
d 0.00 0.00 0.00 1.00 0.00
e 0.66 1.00 1.00 1.00 1.00
```

POSetR: a new computationally efficient R package...

LEapply makes POSetR much more flexible than existing posetic packages, which do not allow the easy implementation of user-defined functions over LEs. This is quite limiting, since many posetic applications to data analysis involve the computation of various statistics over LE, something that in POSetR can be efficiently done with a single call to LEapply, by properly choosing the argument function (see also the example in Section 6). As mentioned, real posets usually have an extremely high number of LEs, therefore LEapply implements both a state-of-the-art algorithm for full LE generation [16] and an MCMC algorithm for LE quasi-uniform sampling (see Bublely and Dyer [3]) and allows users to choose between the two. Interestingly, based on simulations, the computational performances of POSetR are of the same order of magnitude as those of netrankr, if not better.

## 6 Example: financial knowledge in Italy

We consider the data on financial knowledge provided by Bank of Italy, for year 2020 [1] which comprise the binary scores (0 - No knowledge; 1 - Knowledge) of 2036 individuals, aged 18-79, on 7 financial knowledge dimensions, namely:  $k_1$  - Purchasing power;  $k_2$  - Interest paid on a mortgage;  $k_3$  - Simple interest calculation;  $k_4$  - Compound interest calculation,  $k_5$  - Risk and reward;  $k_6$  - Inflation definition;  $k_7$  - Diversification. The  $2^7 = 128$  possible binary score patterns (knowledge profiles) are partially ordered *componentwise* and to each of them the corresponding relative frequency is associated. To summarize the data, we want to identify the *median* knowledge pattern of the population. The partial order structure of the data, however, makes the notion of the median a fuzzy one [9]; the “degree of membership to the median” of a profile is computed as the fraction of LEs of the financial knowledge poset in which it represents the median (indeed, linear extensions are completely ordered and on the median can be defined in the classical way). Knowledge profiles are first enumerated from 1 to 128, then LEapply(p, FUN = isMedian, generator = "BublelyDyer", bublelydyer.nit = n^3) is called, where isMedian is a function returning, for each LE, a binary vector of length 128, whose  $i$ -component is 1 if and only if the  $i$ -th profile is the median in that LE (given the number of profiles, the Bublely-Dyer MCMC algorithm has been used). More than 40 profiles happen to represent the median pattern in some LEs, but most of them have very small membership degrees (*md*, for short). The three profiles with the highest degrees are 0110110 (*md*: 0.23), 1100110 (*md*: 0.11) and 0100111 (*md*: 0.11). Such profiles can be considered as representative of the main features of financial knowledge in Italy. Interestingly they share the lack of knowledge about *compound interest* (4th component) and *diversification* (7th component), two dimensions of key importance for financial awareness.



## 7 Conclusion

We have presented the new R package `POSetR` and its basic functionalities, particularly the highly efficient C++ implementation of the “engine”, for generating linear extensions and flexibly computing user-defined functions on them. Future software implementations will move along two lines: (i) improving `POSetR` adding new functionalities, mainly oriented to the mathematical treatment of partial order structures and (ii) developing and integrating procedures for the statistical analysis of partially ordered data, to provide an ecosystem for applied statisticians.

## References

1. D’Alessio G., De Bonis R., Neri A. Rampazzi C. (2020) *Financial literacy in Italy: the results of the Bank of Italy’s 2020 survey*, Bank of Itali . Occasional Papers 588.
2. Bruggemann R., Patil G. P. (2011). *Ranking and Prioritization for Multi-indicator Systems*, Springer.
3. Bublely R., Dyer M. (1999). *Faster random generation of linear extensions*, Discrete Math., 201, 81-88.
4. Davey B. A., Priestley B. H. (2002). *Introduction to Lattices and Order*, CUP.
5. David Schoch (2017). *netrankr: An R package to analyze partial rankings in networks*.
6. De Loof K. (2010). *Efficient computation of rank probabilities in posets*, Ph.D dissertation, available on <https://biblio.ugent.be/publication/874495>.
7. Dirk Eddelbuettel and Romain Francois (2011). Rcpp: *Seamless R and C++ Integration*, Journal of Statistical Software, 40(8), 1-18. URL <https://www.jstatsoft.org/v40/i08/>.
8. Dirk Eddelbuettel and James Joseph Balamuta (2018). *Extending R with C++: A Brief Introduction to Rcpp*, The American Statistician. 72(1). URL <https://doi.org/10.1080/00031305.2017.1375990>.
9. Fattore M. (2008). *Hasse diagrams, poset theory and fuzzy poverty measures*, Rivista internazionale di scienze sociali 1, 63-75..
10. Fattore M., Arcagni A. (2014). *PARSEC: an R package for poset-based evaluation of multidimensional poverty*, in: Bruggemann R., Carlsen L., Wittmann J. (Eds.) *Multi-Indicator Systems and Modelling in Partial Order*, Springer, Berlin.
11. Fattore M. (2016). *Partially ordered sets and the measurement of multidimensional ordinal deprivation*, Social Indicators Research. 128(2) pp. 835-858 2, DOI: 10.1007/s11205-015-1059-6.
12. Fattore M., Arcagni A. (2018). *F-FOD: Fuzzy First Order Dominance analysis and populations ranking over ordinal multi-indicator systems*, Social Indicators Research DOI: 10.1007/s11205-018-2049-2.
13. Fattore M., Arcagni A. (2018). *A reduced posetic approach to the measurement of multidimensional ordinal deprivation*, Social Indicators Research. 136(3), 1053-1070.
14. Fattore M., Arcagni A., Maggino F. (2019). *Optimal scoring of partially ordered data, with an application to the ranking of smart cities*, Smart Statistics for Smart Applications Book of Short Papers SIS 2019 - ISBN 9788891915108.
15. Fattore M., Arcagni A. (2020). *Ranking extraction in ordinal multi-indicator systems*, Book of Short Papers - SIS 2020 - ISBN 9788891910776 - Pearson.
16. Habib, M., Medina, R., Nourine, L., and Steiner, G. (2001). *Efficient algorithms on distributive lattices*, Discrete Applied Mathematics, 110(2-3), 169-187.
17. R Core Team (2020). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
18. Schröder. (2002). *Ordered set. An introduction*, Birkäuser.

# Multi Split Conformal Prediction

## *Intervalli di previsione non parametrici basati su una procedura di ricampionamento*

Aldo Solari and Vera Djordjilović

**Abstract** Split conformal prediction is a computationally efficient method for performing distribution-free predictive inference in regression. It involves, however, a one-time random split of the data, and the result depends on the particular split. To address this problem, we propose multi split conformal prediction, a simple method based on Markov's inequality to aggregate split conformal prediction intervals across multiple splits.

**Abstract** *Split conformal prediction è un metodo computazionalmente efficiente per ottenere intervalli di previsione non parametrici per problemi di regressione. Questo metodo, tuttavia, richiede una suddivisione casuale dei dati, e quindi il risultato dipende da questo evento aleatorio. Per risolvere questo problema, proponiamo un semplice metodo basato sulla disuguaglianza di Markov per aggregare diversi intervalli di previsione.*

**Key words:** aggregated conformal prediction, split conformal prediction, Markov's inequality, multiple data splitting.

## 1 Introduction

*Conformal prediction* is a general framework for constructing marginally valid prediction sets. The main reference is the 2005 book *Algorithmic Learning in a Random World*, although the subject was pioneered by Vladimir Vovk and colleagues already in the 90s. Recently, it has attracted increasing attention in the statistical community. In spite of its elegance and theoretical appeal, the computational cost of the

---

Aldo Solari

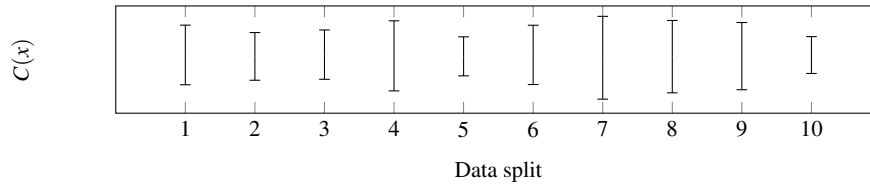
Department of Economics, Management and Statistics, University of Milano-Bicocca, e-mail: aldo.solari@unimib.it

Vera Djordjilović

Department of Economics, Ca' Foscari University of Venice, e-mail: vera.djordjilovic@unive.it

original method, termed *full conformal prediction*, proved to be rather prohibitive in practical applications.

To address this issue, several Authors have proposed *split conformal prediction* as a computationally efficient version of conformal prediction. It involves, however, a one-time random split of the data, and the result can strongly depend on the particular split. This kind of randomness of the prediction interval parallels the “*p*-value lottery” discussed in [4]. An illustration of the potential impact of a single random split is shown in Figure 1 featuring 10 prediction intervals obtained from 10 different data splits.



**Fig. 1** Realizations of 10 split conformal prediction intervals  $C(x)$  for the same test point  $x$ .

A straightforward strategy for overcoming this problem is to aggregate results obtained from multiple data splits. Methods of *aggregated conformal prediction* with proven coverage guarantees include *K*-fold cross-conformal prediction, jackknife+/*K*-fold CV+, and *K*-subsample conformal. See Table 1 for an overview.

<i>Method</i>	<i>Coverage</i>	<i>Reference</i>
Cross-conformal	$\geq 1 - 2\alpha - a(n, K)$	[8]
Jackknife+/CV+	$\geq 1 - 2\alpha - \min\{a(n, K), b(n, K)\}$	[1]
Subsampling conformal	$\geq \min\{2, K\}\alpha$	[2]

**Table 1** Aggregated conformal prediction methods with proven coverage guarantees, where  $a(n, K) = (2 - 2/K)/(n/K + 1)$  and  $b(n, K) = (1 - K/n)/(K + 1)$ .

The coverage guarantees of all methods listed in Table 1 exploit the fact that double of the average *p*-value is a valid *p*-value, a result established by [6]. Only the factor  $b(n, K)$  derived in Theorem 4 of [1] is based on a different argument that makes use of Landau’s theorem for tournaments.

In this short contribution, we propose *multi split conformal prediction*, a simple method based on Markov’s inequality to aggregate split conformal prediction intervals across multiple splits [7]. The proposed method is similar in spirit to *p*-value aggregation and stability selection. In particular, the multi split prediction set includes those points that are included in single split prediction intervals with frequency greater than a user defined threshold. Notably, the Bonferroni-intersection method of [3] and the jackknife+/CV+ of [1] can be seen as special cases of the proposed approach, for details, see [7].

## 2 Conformal prediction

Assume that  $(X_1, Y_1), \dots, (X_n, Y_n), (X, Y)$  are  $n+1$  independent identically distributed random vectors from  $P_{X,Y}$  on the sample space  $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ . Suppose that the realizations  $(x_1, y_1), \dots, (x_n, y_n)$  and  $x$  are available, and we want to predict  $Y$  based on  $x$ . More specifically, we aim to construct a prediction set  $C_\alpha \subseteq \mathbb{R}$  such that its marginal coverage is at least  $1 - \alpha$ , i.e.

$$\mathbb{P}(Y \in C_\alpha(X)) \geq 1 - \alpha, \quad (1)$$

where the probability is taken over  $(X_1, Y_1), \dots, (X_n, Y_n), (X, Y)$ .

Let  $\phi_\alpha = \phi(Z_1, \dots, Z_n, Z) \in \{0, 1\}$  be a Bernoulli random variable, where  $Z_i = (X_i, Y_i)$  for  $i = 1, \dots, n$  and  $Z = (X, Y)$ . Denote by  $\phi_\alpha^y = \phi(Z_1, \dots, Z_n, Z_y)$  with  $Z_y = (X, y)$ .

**Theorem 1.** Assume that  $\phi_\alpha$  is a Bernoulli random variable such that  $\mathbb{E}(\phi_\alpha) \leq \alpha$ . The prediction set  $C_\alpha(x) = \{y \in \mathbb{R} : \phi_\alpha^y = 0\}$  satisfies (1).

Informally,  $\phi_\alpha^y$  can be thought of as a test for the null hypothesis that  $Y$  assumes the value of  $y$ , that is  $H_y : Y = y$ . Theorem 1 then states that a valid prediction set can be obtained by inverting a collection of such tests. For the proof of this and the remaining results we refer to [7].

## 3 Single split conformal prediction

Consider a partition of  $(1, \dots, n)$  into a learning set  $L$  of size  $w$  and an inference set  $I$  of size  $m = n - w$ , independently of the observed data values. Define a statistic  $R = R((Z_l)_{l \in L}, Z)$ . Statistic  $R$ , also referred to as a conformity score in conformal inference, serves as a measure of “plausibility” of the value  $y$  as a realization of  $Y$  for the observed value of  $X$ . Examples include

$$R = |Y - \hat{\mu}_L(X)|, \quad R = \max\{\hat{q}_L^\gamma(X) - Y, Y - \hat{q}_L^{1-\gamma}(X)\}, \quad (2)$$

where  $\hat{\mu}_L$  is an estimator of  $\mathbb{E}(Y|X)$  based on  $(Z_l)_{l \in L}$ , and  $\hat{q}_L^\gamma$  is an estimator of the  $\gamma$ -quantile of  $Y|X$  based on  $(Z_l)_{l \in L}$ .

Denote the inference set by  $I = (i_1, \dots, i_m)$ . Let  $R_j = R((Z_l)_{l \in L}, Z_{i_j})$ ,  $j = 1, \dots, m$ . For  $\alpha \in (0, 1)$ , let  $R_\alpha = R_{(\lceil (1-\alpha)(m+1) \rceil)}$ , where  $R_{(1)} \leq \dots \leq R_{(m)}$  are the ordered statistics obtained by sorting the values of  $R_1, \dots, R_m$  in increasing order, with ties broken arbitrarily.

**Lemma 1.** For any  $\alpha \in (0, 1)$ ,  $\phi_\alpha = \mathbb{1}\{R > R_\alpha\}$  is such that  $\mathbb{E}(\phi_\alpha) \leq \alpha$ , with equality if and only if  $R_1, \dots, R_m, R$  are almost surely distinct and  $\alpha \in \Delta = \{i/(m+1)\}_{i=1}^m$ .

Equality  $\mathbb{E}(\phi_\alpha) = \alpha$  implies that the prediction set  $C_\alpha$  has exact coverage, i.e.  $\mathbb{P}(Y \in C_\alpha(X)) = 1 - \alpha$ . The proof of Lemma 1 is based on the *permutation testing principle* [5, 7].

Algorithm 1 describes how to compute the split conformal prediction set.

---

**Algorithm 1** Split Conformal Prediction

---

**Require:** data  $(x_1, y_1), \dots, (x_n, y_n), x$ , inference sample size  $m$ , statistic  $R$ , level  $\alpha \in (0, 1)$

1: split  $\{1, \dots, n\}$  into  $L$  of size  $w$  and  $I$  of size  $m = n - w$

2: compute  $R_\alpha = R_{(\lceil (1-\alpha)(m+1) \rceil)}$

3: compute  $C_\alpha(x) = \{y \in \mathbb{R} : R \leq R_\alpha\}$

**return** split conformal prediction set  $C_\alpha(x)$

---

## 4 Multi split conformal prediction

Choose the number of splits  $B \in \mathbb{N}$ . Partition  $(1, \dots, n)$  into  $L^{[b]}$  of size  $w^{[b]}$  and  $I^{[b]}$  of size  $m^{[b]} = n - w^{[b]}$ , and choose a statistic  $R^{[b]}$ , for  $b = 1, \dots, B$ . For  $\beta \in (0, 1)$ ,  $\phi_\beta^{[b]} = \mathbb{1}\{R^{[b]} > R_\beta^{[b]}\}$  has  $\mathbb{E}(\phi_\beta^{[b]})$  by Lemma 1. Let

$$V_\beta = \sum_{b=1}^B \phi_\beta^{[b]}$$

be the number of successes (1s), with  $\mathbb{E}(V_\beta) \leq B\beta$ . The following Theorem provides an upper bound for the probability of at least  $k$  successes out of  $B$  trials, i.e. for  $\mathbb{P}(V_\beta \geq k)$ .

**Theorem 2.** *Let  $\lambda$  be a non-negative integer such that, for a given integer  $1 \leq k \leq B$  and  $\beta \in (0, 1)$ , the following holds:*

$$\sum_{u=0}^{k-1} \mathbb{P}(V_\beta \in [k-u, k]) \geq \sum_{u=0}^{\lambda} \mathbb{P}(V_\beta \in [k, k+u]) \quad (3)$$

Then

$$\mathbb{P}(V_\beta \geq k) \leq \frac{B\beta}{k + \lambda}.$$

The parameter  $\lambda$  can be regarded as a smoothing parameter. The value  $\lambda = 0$  reduces (4) to Markov's bound. However, Markov's inequality can be sharpened under constraints on the shape of the distribution of  $V_\beta$ .

Let  $\Pi_\beta = 1 - V_\beta/B = B^{-1} \sum_{b=1}^B \mathbb{1}\{Y \in C_\beta^{[b]}(X)\}$  be the proportion of prediction sets  $C_\beta^{[b]}(X)$  that include  $Y$ . For  $\alpha \in (0, 1)$  and a threshold  $\tau = 1 - k/B$ , the multi split conformal prediction set

$$C_\alpha^\tau(x) = \{y \in \mathbb{R} : \Pi_\beta^y > \tau\} \quad (4)$$

has coverage at least  $1 - \alpha$  by Theorem 1 with  $\phi_\alpha = \mathbb{1}(\Pi_\beta \leq \tau)$ , where  $\beta = \alpha(1 - \tau)$  with no assumptions or  $\beta = \alpha(1 - \tau + \lambda/B)$  under the assumption (3) of Theorem 2. Algorithm 2 describes how to compute the multi split conformal prediction set.

---

**Algorithm 2** Multi Split Conformal
 

---

**Require:** data  $(x_1, y_1), \dots, (x_n, y_n)$ ,  $x$ , number of splits  $B \in \mathbb{N}$ , inference sample sizes  $(m^{[b]})_{b=1}^B$ , statistics  $(R^{[b]})_{b=1}^B$ , threshold  $\tau \in [0, (B-1)/B]$ , level  $\alpha \in (0, 1)$ , smoothing parameter  $\lambda \in \mathbb{N}_0$ .

- 1: **for**  $b \leftarrow 1$  to  $B$  **do**
- 2:   compute  $C_\beta^{[b]}(x)$  using Algorithm 1 with  $m^{[b]}$ ,  $R^{[b]}$  and level  $\beta = \alpha(1 - \tau + \lambda/B)$
- 3: **end for**
- 4: compute  $C_\alpha^\tau(x) = \{y \in \mathbb{R} : \Pi_\beta^y > \tau\}$

**return** multi split conformal prediction set  $C_\alpha^\tau(x)$

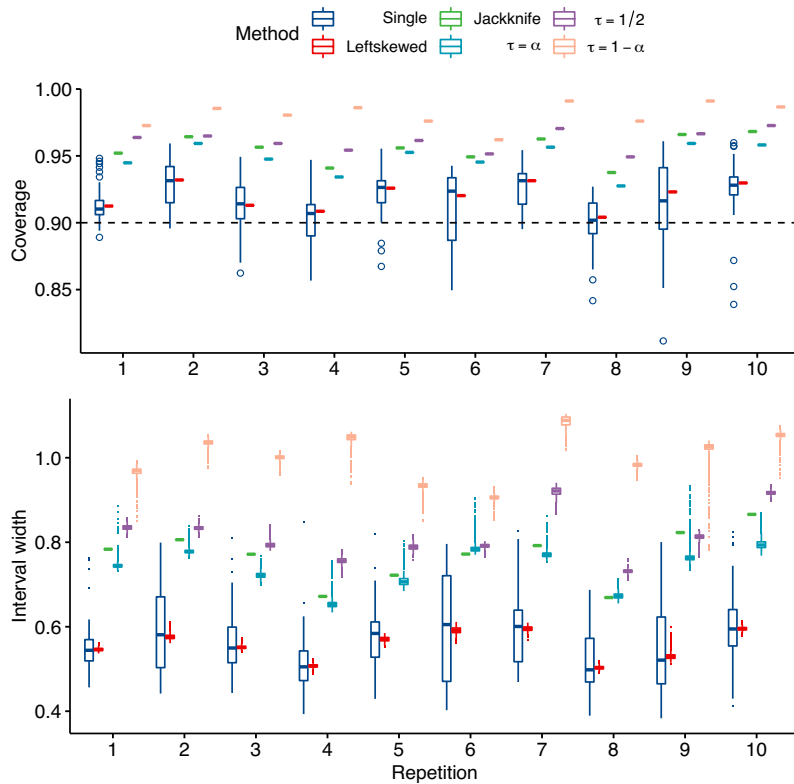
---

## 5 Example

We now apply multi split conformal prediction on the Communities and Crime data set (<http://archive.ics.uci.edu/ml/datasets/communities+and+crime>, accessed: March, 2020). The data set contains information on 1994 communities, with covariates such as median income, family size, etc., and the goal is to predict a response variable defined as the per capita violent crime rate. After removing categorical variables and variables with missing data,  $d = 99$  covariates remain.

We replicate the experiment in [1]. We randomly sample  $n = 200$  data points from the full data set, to use as the training data. The remaining 1794 points form the test set. We use  $R = |Y - \hat{\mu}_L(X)|$  where  $\hat{\mu}_L$  is estimated by the ridge regression algorithm with penalty parameter chosen as  $0.001c^2$ , where  $c$  is the largest singular value of the training data matrix. We set the coverage level to  $1 - \alpha = 90\%$ , the number of random split to  $B = 51$  and size for the inference set to  $m = 99$ . We construct  $B$  single split intervals and the multi split interval by using  $\tau = \alpha$ ,  $\tau = 1/2$ ,  $\tau = 1 - \alpha$  and Jackknife+ with no assumptions and  $(\tau, \lambda) = ((B-1)/2B, (B-1)/2)$  with assumption (3), which we refer to as “Leftskewed”. For each method, we calculate its empirical coverage and interval width on the test set. We then repeat this procedure 10 times, with the train/test split formed randomly each time.

Figure 2 displays the results. Intervals obtained by the “Leftskewed” method exhibit coverage and width comparable to single split intervals, but with substantially reduced variability, as expected. Assumption free methods reflect the conservativeness of Markov’s inequality and can not compete with the exact coverage single split method.



**Fig. 2** Coverage and interval width for single split and multi split prediction sets on the Communities and Crime data set.

## References

1. Barber, Rina Foygel and Candes, Emmanuel J and Ramdas, Aaditya and Tibshirani, Ryan J: Predictive inference with the jackknife+. *Annals of Statistics* **49**, 486–507 (2021)
2. Gupta, Chirag and Kuchibhotla, Arun K and Ramdas, Aaditya K: Nested conformal prediction and quantile out-of-bag ensemble methods. *arXiv preprint arXiv:1910.10562* (2019)
3. Lei, Jing and G' Sell, Max and Rinaldo, Alessandro and Tibshirani, Ryan J and Wasserman, Larry: Distribution-free predictive inference for regression. *JASA* **113**, 1094–1111 (2018)
4. Meinshausen, Nicolai and Meier, Lukas and Bühlmann, Peter: P-values for high-dimensional regression. *JASA* **104**, 1671–1681 (2009)
5. Pesarin, Fortunato: *Multivariate permutation tests: with applications in biostatistics*. Wiley Chichester (2001)
6. Rüschendorf, Ludger: Random variables with maximum sums. *Advances in Applied Probability* 623–632 (1982)
7. Solari A, Djordjilović V: Multi Split Conformal Prediction. *arXiv:submit/3626397* (2021)
8. Vovk, Vladimir: Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence* **74**, 9–28 (2015)

# Changes in the consumption of fruits and vegetables among university students during master courses: an analysis of data automatically collected from cashier transactions

*Cambiamenti del consumo di frutta e verdura tra gli studenti universitari nei tre anni successivi l'immatricolazione sulla base dei dati raccolti dalle transazioni effettuate per l'accesso alle mense universitarie*

Valentina Lorenzoni, Giuseppe Turchetti, Lucio Masserini

## Abstract

Previous evidence showed a degradation of eating habits among university students with high prevalence of unhealthy eating behaviours among freshmen that may change in further years when students acquire consciousness about their behaviours. As the frequency of consumption of both fruits and vegetables are well recognised indicators of healthy habits, the present study assessed changes in the consumption of both fruits and vegetables among university students during master courses using data automatically recorded by cashier transactions and related to meals consumed at the canteens of a large University in central Italy. Results highlighted that the frequency of choice of both fruits and vegetables was low in the first year of enrolment and significantly increased in subsequent years.

## Abstract

*Le evidenze disponibili sottolineano un peggioramento delle abitudini alimentari tra gli studenti universitari con un'alta prevalenza di comportamenti non salutari tra le matricole, che possono cambiare negli anni successivi con l'acquisizione di consapevolezza verso il cibo. Poiché la frequenza di consumo di frutta e verdura sono considerati indicatori di abitudini salutari, il presente studio ha valutato i cambiamenti nel consumo di frutta e*

---

<sup>1</sup> Valentina Lorenzoni, Institute of Management Scuola Superiore Sant'Anna , Pisa, Italy; email: [valentina.lorenzoni@santannapisa.it](mailto:valentina.lorenzoni@santannapisa.it)

<sup>2</sup> Giuseppe Turchetti, Institute of Management Scuola Superiore Sant'Anna , Pisa, Italy; email: [Giuseppe.Turchetti@santannapisa.it](mailto:Giuseppe.Turchetti@santannapisa.it)

<sup>3</sup> Lucio Masserini, Department of Economics and Management, University of Pisa, Pisa, Italy; email: [luccio.masserini@unipi.it](mailto:luccio.masserini@unipi.it)



*verdura tra gli studenti universitari durante il corso di laurea utilizzando dati acquisiti automaticamente dalle transazioni di cassa e relativi ai pasti consumati presso le mense di un' università del centro Italia. I risultati evidenziano un aumento significativo della frequenza di scelta di frutta e verdura negli anni successivi al primo.*

**Key words:** Diet, Fruit, Vegetables, University Students, Beta Regression

## 1 Introduction

Despite literature about eating habits among university students is limited, particularly in Europe, available evidence tend to equal student food with 'bad' food (Kapinos et al., (2014); Nelson et al., (2008)). Reasons for poor eating habits in this population have been mainly ascribed to the transition period university students have to face, the reduced influence of parents as compared to the greater importance of peers' effect and the high degree of duties and responsibilities they are required to absolve (Doherty et al., (2011); Sharma et al., (2009)).

In details, while the frequent consumption of vegetables and fruits are among well recognized indicators for healthy eating, with potential beneficial effect on both acute and chronic diseases (Slavin and Lloyd, (2012)), several studies suggested that the assumption of both fruits and vegetable is limited among university students and particularly among freshmen (García-Meseguer et al., (2014); Huang et al., (1994); Small et al., (2013); Teleman et al., (2015)) with potential effect not only on their present and future health status but also on their overall well-being and academic achievement.

Using data automatically recorded by cashier' transactions and related to meals consumed at the canteens of a large University in central Italy,, the present study aims at evaluating changes in the frequency of selection of both fruits and vegetables among students during their master courses.

## 2 Methods

### 2.1 *Study population and data*

The population under study was composed of students enrolled in courses in a large University in central Italy who accessed the canteens in the academic years from 2010-11 to 2013-14. In order to observe the behaviour over a sufficient period of time, only students enrolled in first-degree courses and accessing the canteen at least 40 times all over the three years of analysis were considered.

Analyses were performed merging two different sources of data: the administrative archive of the University and the database of Azienda Regionale per il Diritto allo Studio Universitario (DSU). The first source contained demographic characteristics of students and data about their career progression, while the second one contained all the records of consumed meals.

The two source of data were merged by the anonymous student ID and analysis about food choices was allowed by the fact that one of the canteens serving the University is

Trends in the consumption of fruits and vegetables among university students during master courses: an analysis using data automatically collected from cashier transactions equipped with an automatic system that allow for the recording of all consumed meals. In detail, for each meal consumed, in addition to the student ID a set of variables were available, such as date and time, number and type of courses chosen, dishes selected and the price charged. On the basis of these data, the frequency of fruits and vegetables selected was measured in terms of the number of times these items were selected over the total number of accesses. Calculating this measure for each student generates a continuous doubly bounded random variable defined in the unit interval (0,1). In details two responses variables were used in the present study: a) the proportion of vegetables choice over the total of meals consumed; b) the proportion of fruits choice over the total of meals consumed. For both variables a multiple model considering also age, gender, scholarship and geographical origin was adapted.

## 2.2 Statistical analysis

To model the proportion of meals including vegetables and fruits all over the period of observation, we used a beta regression model with robust standard errors. The class of beta regression models (Ferrari and Cribari-Neto, (2004)), are well-suited to model continuous variables constrained in the unit interval (0,1) and assumes that the response variable ( $y$ ) is beta distributed. Following the parameterization proposed by Ferrari and Cribari-Neto, the beta density function can be written as follows:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}$$

where  $0 < \mu < 1$ ,  $\phi > 0$  and  $\Gamma(\cdot)$  denotes the gamma function. Moreover,  $E(Y) = \mu$  and  $\text{Var}(Y) = \mu(1-\mu)/(\phi+1)$ . The parameter  $\phi$  is known as a “precision” parameter since, for fixed  $\mu$ , the larger  $\phi$ , the smaller the variance of the response variable  $y$ . Since the beta regression is heteroscedastic, both location and precision are modelled by specifying the following two submodels, one for the location parameter  $\mu$  and another for the precision parameter  $\phi$ . Given  $\mathbf{x}_i$  and  $\mathbf{w}_i$  the vector of fixed and known covariates, and let  $\boldsymbol{\beta}$  and  $\boldsymbol{\delta}$  be the vector of regression coefficients, the location submodel is:

$$g(\mu_i) = \mathbf{x}_i \boldsymbol{\beta},$$

where  $g(\cdot)$  is the logit function and the precision submodel is:

$$h(\phi_i) = \mathbf{w}_i \boldsymbol{\delta},$$

where  $h(\cdot)$  is the log function. Finally, parameter estimation is performed by maximum likelihood (ML).

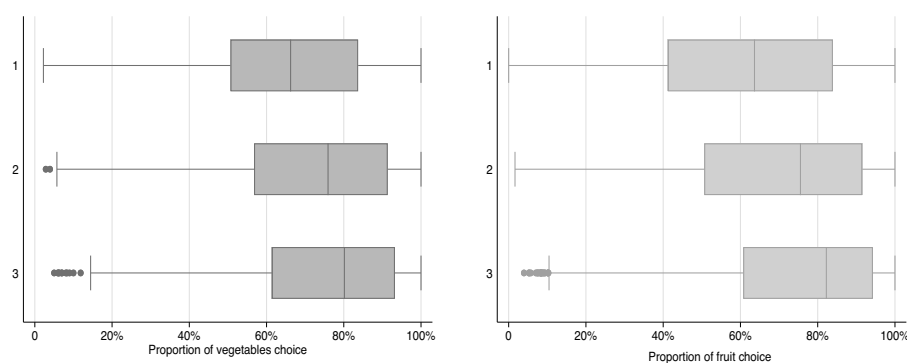
## 3 Results

A total of 2,825 students enrolled in master courses at University of Pisa and having observations over a 3-year period (starting from matriculation) were included in the analysis. The main characteristics of the study population at enrolment are shown in Table 1.

**Table 1:** Main Characteristics of the study population

Age	20.5±1.8
Gender	
Female	1,177 (41.7%)
Male	1,648 (58.34)
Scholarship older	
No	1,536 (54.4%)
Yes	1,289 (45.6%)
Geographical Origin	
North Italy	233 (8.3%)
Central Italy	1,024 (36.3%)
South Italy	730 (25.8%)
Italian islands	567 (20.1%)
Foreign	271 (9.6%)

The selection of both vegetables and fruits during meals consumed at the university canteen significantly increased over years ( $p < 0.001$ ). In details, while freshmen included vegetables in their meals about 65% of times, values increased to 72% and 76% in the second and third year respectively; for fruits values increased from about 61% in the first year to 75% in the third year, see Figure 1.

**Figure 1:** Box-plot showing the proportion of vegetables and fruit (choice) over years.

Results from the beta regression model suggested that in the second and third year of registration the frequency of selection of vegetables significantly increased by 0.29 (SE 0.02) and 0.44 (SE 0.02) respectively (Table 2).

Trends in the consumption of fruits and vegetables among university students during master courses: an analysis using data automatically collected from cashier transactions

**Table 2:** Results from the multiple beta regression models for the response variable “Proportion of vegetables choice”

	<i>Coef.</i>	<i>SE</i>	<i>95%CI</i>	<i>P-value</i>
<b>Response variable: Proportion of vegetables choice</b>				
Registration year 1	<i>(ref)</i>			
Registration year 2	0.290	0.016	0.258-0.322	<0.001
Registration year 3	0.445	0.024	0.399-0.491	<0.001
Age	0.032	0.009	0.015-0.050	<0.001
Male gender	-0.038	0.032	-0.101-0.024	0.227
Scholarship holders	0.227	0.030	0.168-0.286	<0.001
Geographical Origin				
North Italy	<i>(ref)</i>			
Central Italy	0.211	0.062	0.089-0.332	0.001
South Italy	-0.039	0.062	-0.161-0.084	0.536
Italian islands	0.062	0.064	-0.062-0.187	0.325
Foreign	0.569	0.072	0.429-0.710	<0.001
Pseudo R <sup>2</sup>	0.11			

Similarly results emerged when modelling the proportion of fruits choice, with coefficients being 0.32 (SE 0.02) and 0.57 (SE 0.03) for the second and first year compared with the enrolment year (Table 3).

**Table 3:** Results from the multiple beta regression models for the response variable “Proportion of fruits choice”

	<i>Coef.</i>	<i>SE</i>	<i>95%CI</i>	<i>P-value</i>
<b>Response variable: Proportion of fruits choice</b>				
Registration year 1	<i>(ref)</i>			
Registration year 2	0.318	0.018	0.282-0.352	<0.001
Registration year 3	0.566	0.025	0.517-0.615	<0.001
Age	0.027	0.010	0.008-0.046	0.005
Male gender	-0.052	0.033	-0.117-0.013	0.118
Scholarship holders	0.292	0.032	0.229-0.355	<0.001
Geographical Origin				
North Italy	<i>(ref)</i>			
Central Italy	0.227	0.069	0.092-0.362	0.001
South Italy	0.041	0.069	-0.094-0.175	0.553
Italian islands	0.155	0.070	0.018-0.291	0.026
Foreign	0.618	0.075	0.470-0.765	<0.001
Pseudo R <sup>2</sup>	0.12			

Moreover these models also suggested that the proportion of both fruits and vegetables selection were positively related with age, scholarship and being from central Italy or being foreign instead of being from the North of Italy.

## 4 Conclusions

Results from the present study confirm that unhealthy eating habits (in this case represented by a low proportion of meals that include fruits and vegetables) occur more frequently in freshmen thus potentially increasing their risk of developing overweight and obesity as well as impacting on both physical and mental health (Blichfeldt and Gram, (2013); Vella-Zarb and Elgar, (2010)).

With the increasing availability of data routinely collected even for purpose other than research there is nowadays the possibility to directly monitor behaviour, such as eating habits, with the opportunity to timely encourage the development of policies and new strategies to promoting healthy habits among university students, with the possibility of also targeting interventions to specific subgroups.

## References

- 1 Blichfeldt, B.S., Gram, M.: Lost in Transition? Student food consumption. *High. Educ.* (2013) <https://doi.org/10.1007/s10734-012-9543-2>
- 2 Doherty, S., Cawood, J., Dooris, M.: Applying the whole-system settings approach to food within universities. *Perspect. Public Health.* (2011) <https://doi.org/10.1177/1757913911413344>
- 3 Ferrari, S.L.P., Cribari-Neto, F.: Beta regression for modelling rates and proportions. *J. Appl. Stat.* (2004) <https://doi.org/10.1080/0266476042000214501>
- 4 García-Meseguer, M.J., Burriel, F.C., García, C.V., Serrano-Urrea, R.: Adherence to Mediterranean diet in a Spanish university population. *Appetite.* (2014); 78, 156–164. <https://doi.org/10.1016/j.appet.2014.03.020>
- 5 Huang, Y.L., Song, W.O., Schemmel, R.A., Hoerr, S.M.: What do college students eat? Food selection and meal pattern. *Nutr. Res.* (1994); 14, 1143–1153. [https://doi.org/10.1016/S0271-5317\(05\)80242-8](https://doi.org/10.1016/S0271-5317(05)80242-8)
- 6 Kapinos, K.A., Yakusheva, O., Eisenberg, D.: Obesogenic environmental influences on young adults: Evidence from college dormitory assignments. *Econ. Hum. Biol.* (2014);12, 98–109. <https://doi.org/10.1016/j.ehb.2013.05.003>
- 7 Nelson, M.C., Story, M., Larson, N.I., Neumark-Sztainer, D., Lytle, L.A.: Emerging adulthood and college-aged youth: An overlooked age for weight-related behavior change. *Obesity.* (2008) <https://doi.org/10.1038/oby.2008.365>
- 8 Sharma, B., Harker, M., Harker, D., Reinhard, K.: Living independently and the impact on young adult eating behaviour in Germany. *Br. Food J.* (2009) <https://doi.org/10.1108/00070700910957285>
- 9 Slavin, J.L., Lloyd, B., 2012. Health benefits of fruits and vegetables. *Adv. Nutr.* <https://doi.org/10.3945/an.112.002154>
- 10 Small, M., Bailey-Davis, L., Morgan, N., Maggs, J.: Changes in Eating and Physical Activity Behaviors Across Seven Semesters of College: Living On or Off Campus Matters. *Heal. Educ. Behav.* (2013); 40, 435–441. <https://doi.org/10.1177/1090198112467801>
- 11 Telemán, A.A., De Waure, C., Soffiani, V., Poscia, A., Di Pietro, M.L.: Nutritional habits in Italian university students. *Ann. Ist. Super. Sanita* (2015);51, 99–105. <https://doi.org/10.4415/ANN-15-02-05>
- 12 Vella-Zarb, R.A., Elgar, F.J. Predicting the freshman 15: Environmental and psychological predictors of weight gain in first-year university students. *Health Educ. J.* (2015) <https://doi.org/10.1177/0017896910369416>

## 4.16 New challenges in clustering and classification techniques

# A Dynamic Stochastic Block Model with infinite communities

## *Un modello dinamico con blocchi aleatori e numero infinito di comunità*

Roberto Casarin and Ovielt Baltodano López

**Abstract** This contribution proposes the use of bayesian non-parametric techniques to make inference on the number of communities in a Dynamic Stochastic Block Model which is then applied to real network data on international financial flows.

**Abstract** *Questo contributo si propone l'uso di metodi bayesiani non-parametrici per fare inferenze sul numero di comunità in un modello dinamico con blocchi aleatori, il quale dopo viene applicato alla rete di flussi finanziari internazionali.*

**Key words:** Stochastic block models, bayesian non-parametric methods.

## 1 Introduction

The increase of network data, e.g. online social networks, has shown the importance of clustering and community structures. In this sense, a Dynamic Stochastic Block Model (DSBM) allows to capture heterogeneous relationships between nodes and potential role changes in their interaction. [9, 6] proposed the use of Hidden Markov chains in order to extend the mixture distribution used in a static setting. However, there is no inference and therefore no measure of uncertainty on the number of communities.

On the other hand, in the field of time series analysis, the use of Hidden Markov chains with infinite states has a long tradition. An important extension was proposed by [4]. In their contribution, they introduce state persistence in a Hierarchical Dirichlet process framework used in a hidden Markov chain model. In a nonlinear context, [2] applies the same strategy to a Generalized Auto-Regressive Conditional heteroskedasticity (GARCH) model. In this paper, we combine this persistent

---

Roberto Casarin  
University Ca' Foscari of Venice e-mail: r.casarin@unive.it

Ovielt Baltodano López  
University Ca' Foscari of Venice e-mail: ovielt.baltodano@unive.it

Hierarchical Dirichlet process and hidden Markov chain with infinite states to the DSBM. This is in line with [3], but their contribution was centered on a mixed-membership setting, that is, each node can play different roles at the same time, while here we assume each node can be in only one community at each point in time. The description of our model is presented in Section 2. Moreover, in Section 3, we use empirical data on the bilateral financial flows between countries given its relevance for financial stability and interdependence, and to exemplify the use of the DSBM with infinite communities.

## 2 A DSBM with infinite communities

A weighted graph can be defined as the ordered triplet  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, Y)$ , where  $\mathcal{V} = \{1, \dots, N\}$  is the set of nodes,  $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$  and  $Y$  is a weight matrix,  $Y \in \mathbb{R}^N \times \mathbb{R}^N$ . The  $(i, j)$ -th element of  $Y$  is  $Y_{ij} = 0$  if  $(i, j) \notin \mathcal{E}$  and  $Y_{ij} = a \in \mathbb{R} \setminus \{0\}$  if  $(i, j) \in \mathcal{E}$ . We define the sequence of sets  $\mathfrak{V} = \mathcal{V}_1, \dots, \mathcal{V}_Q$ , with  $Q \in \{1, 2, \dots\}$  a partition of  $\mathcal{V}$ , if each element  $\mathcal{V}_j \subset \mathcal{V}$  (called block or community in what follows) satisfies:  $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$  and  $\mathcal{V}_1 \cup \dots \cup \mathcal{V}_Q = \mathcal{V}$ .

In this paper we assume a sequence of graphs  $\mathcal{G}_{1:T} = \{\mathcal{G}_t, t = 1, \dots, T\}$  is available and a latent sequence of partitions  $\mathfrak{V}_{1:T} = \{\mathfrak{V}_t, t = 1, \dots, T\}$  drives the topology of the graph. Following [9] and [6], the partition sequence  $\mathfrak{V}_{1:T}$  is induced by a set of  $N$  hidden Markov chain processes. The membership of  $i \in \mathcal{V}$  is captured by  $Z_i = \{Z_{it}, t = 1, \dots, T\}$ , which evolves following a Markov Chain process with transition matrix  $P$ , where entry  $q, r \in \mathcal{Q}$  is given by  $P_{qr} \in (0, 1)$  and each row  $P_q$  sum up to one. At time  $t$ , the node  $i$  belongs to the block  $\mathcal{V}_q$  if  $Z_{it} = q$ . Although the chains are independent, they share the same transition matrix, thus  $P$  gives information on the level of persistence of the communities as a whole.

The node partition induces edge clusters with different existence probabilities and weights. Further, we assume that the contemporaneous network  $Y_t$  given  $Z_{1:T}$  and  $Y_{1:T}$  only depends on  $Z_t = \{Z_{1t}, \dots, Z_{Nt}\}$  and each entry of the adjacency matrix is distributed as

$$Y_{ijt} | Z_{it} = q, Z_{jt} = r, \theta_{ijt} \sim (1 - v_{qr})\delta(y) + v_{qr}f(y|\lambda_{qr}) \quad (1)$$

which is a zero-inflated distribution family, where  $\delta(\cdot)$  denotes the Dirac function at zero and  $f(\cdot|\lambda_{qr})$  is a probability density function with parameter  $\lambda_{qr}$  and support set  $\mathbb{R} \setminus \{0\}$ . The community structure is used to allow for partial parameter pooling, that is the edge parameters  $\theta_{ijt} = (v_{ijt}, \lambda_{ijt})' = \theta_{qr}^*$  if  $i \in \mathcal{V}_q$  and  $j \in \mathcal{V}_r$  at time  $t$ .

Usually, the number of communities is given and the choice depends on some specific criteria. For instance, [6] chooses the model with the highest integrated classification likelihood criterion, after fitting models with different  $\mathcal{Q}$  cardinality. In order to infer the number of communities, a Bayesian non-parametric framework can be applied, which to allow for infinite states Markov chains. Since the number of state is infinite, i.e.  $\mathcal{Q} = \{1, 2, \dots\}$ , the transition matrix  $P$  becomes infinite di-



mensional and a parsimonious model is needed for  $P$ , which preserves the labelling of the communities in the different rows. [7] proposed a hierarchical Dirichlet Process (DP) to tie the different rows of  $P$  by providing the same centering measure for each row, that is

$$\begin{aligned} Z_{it} | Z_{it-1} = q &\sim G_q, q \in \mathcal{Q} \\ G_q | \omega, G_0 &\sim \text{DP}(\omega, G_0) \\ G_0 | \eta, H &\sim \text{DP}(\eta, H), \end{aligned} \tag{2}$$

where  $\text{DP}(\alpha, H)$  denotes a Dirichlet process with precision parameter  $\alpha$  and centering (or base) measure  $H$ . Nevertheless, [4] underline the fact that (2) does not differentiate between the main diagonal of  $P$  and the transition across different groups, essentially affecting the state persistence. Therefore, using the extension proposed by [4] for the analysis of time-series, (2) can be extended in line with the Chinese restaurant franchise with loyal customer,

$$\begin{aligned} Z_{it} | P_q, Z_{it-1} = q &\sim \text{Ca}(P_q) \\ P_q | \omega, \pi &\sim \text{DP}\left(\omega + \kappa, \frac{\omega\pi + \delta(r - q)}{\omega + \kappa}\right) \\ \pi &\sim \text{Stick}(\eta) \\ \theta_{qr}^* &\sim H \end{aligned} \tag{3}$$

where  $\text{Ca}(p)$  denotes a categorical (or multinoulli) distribution with probability parameter  $p$ . The parameters  $P_q, \pi$  are the weights of the stick-breaking representation of  $G_q$  and  $G_0$ , and  $\kappa$  is a parameter increasing the self-transition probability  $P_{qq}, q \in \mathcal{Q}$ . The  $q$ -th element of the infinite vector  $\pi$  is given by  $\pi_q = \xi_q \prod_{l=1}^{q-1} (1 - \xi_l)$  and  $\xi_l \sim \text{Beta}(1, \eta)$ . The Fig. 1 summarizes the structure of the DSBM with infinite communities, at each point time the allocation variables  $Z_{it}$  and  $Z_{jt}$  of the corresponding pair  $(i, j)$  determines which parameters  $\theta_{qr}^*$  applies. Their membership changes on the basis of the infinite dimension  $P$ , whose rows have a specific Dirichlet process under the same centering measure.

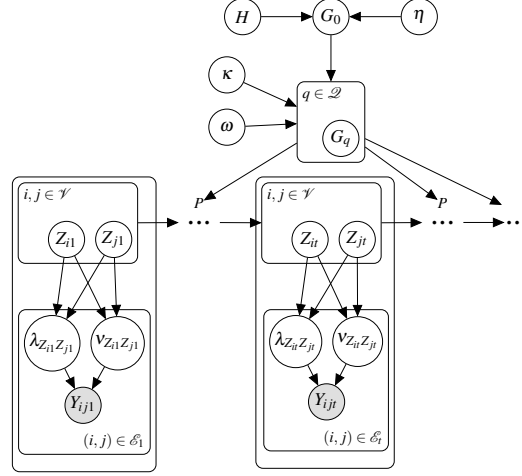
In the case of a weighted network whose active edges have a support  $\mathbb{R} \setminus \{0\}$ , a zero-inflated normal can be used in (1) with parameters  $\lambda_{qr} = (\beta_{qr}, \sigma_{qr}^2)'$ . Furthermore, (1) can be rewritten as

$$Y_{ijt} | D_{ijt}, Z_{it} = q, Z_{jt} = r, \theta_{ijt} \sim \begin{cases} \delta(y) & \text{if } D_{ijt} = 0 \\ f(y | \lambda_{qr}) & \text{if } D_{ijt} = 1 \end{cases} \tag{4}$$

where  $D_{ijt}$  is an observable indicator variable such that  $D_{ijt} = 1$  if  $(i, j) \in \mathcal{E}_t$  and  $D_{ijt} = 0$  if  $(i, j) \notin \mathcal{E}_t$ ,

$$D_{ijt} | Z_{it} = q, Z_{jt} = r, \theta_{ijt} \sim \text{Bern}(v_{qr}) \tag{5}$$

Under this representation, a full Gibbs sampling procedure can be derived after using a set of slice sampling auxiliary variables  $u_{it}$  applied to the stick-breaking

**Fig. 1** Directed Acyclic Graph of DSBM with infinite communities


representation in (3). The main full conditional posteriors are presented in Table 1 where the inference also covers the hyperparameters  $\eta, \kappa, \omega$ .<sup>1</sup> The closed form of the parameters of the full conditional posteriors and further details, such as the auxiliary variables  $\bar{m}_{\cdot q}$  and  $g$ , are standard in the literature [e.g. 4, 2]. Additionally, the allocation variables  $Z$  are sampled from Forward filtering backward sampling [5].

**Table 1** Gibbs sampling

Prior	Full Conditional Posterior
$\beta_{qr} \sim N(\underline{\beta}_{qr}, \underline{\Sigma}_{qr})$	$\beta_{qr} Y, \dots \sim N(\bar{\beta}_{qr}, \bar{\Sigma}_{qr})$
$\sigma_{qr}^2 \sim \text{IG}(d_{qr}/2, e_{qr}/2)$	$\sigma_{qr}^2 Y, \dots \sim \text{IG}(\bar{d}_{qr}/2, \bar{e}_{qr}/2)$
$v_{qr} \sim \text{Beta}(\underline{b}_{qr}, \underline{c}_{qr})$	$v_{qr} Y, \dots \sim \text{Beta}(\bar{b}_{qr}, \bar{c}_{qr})$
$P_q \sim \text{DP}(\omega + \kappa, \frac{\omega\pi + \delta(r-q)}{\omega + \kappa})$	$P_q Y, \dots \sim \text{Dir}(\omega\pi_1 + n_{q\cdot}, \dots, \omega\pi_q + \kappa + n_{qq}, \dots, \omega\pi_{Q+1})$
$\pi \sim \text{Stick}(\eta)$	$\pi Y, \dots \sim \text{Dir}(\bar{m}_{\cdot 1}, \dots, \bar{m}_{\cdot Q}, \eta)$
$u_{it} \sim \text{Uni}(0, 1)$	$u_{it} Y, \dots \sim \text{Uni}(0, k^{\mathbb{1}(Z_{it-1}=Z_{it})} P_{Z_{it-1}Z_{it}})$
$\omega + \kappa \sim G(\zeta_1, \zeta_2)$	$\omega + \kappa Y, \dots \sim G(\zeta_1 + m - s, (1/\zeta_2 - \sum_{q=1}^Q \log k_q)^{-1})$
$\rho \sim \text{Beta}(\chi_1, \chi_2)$	$\rho Y, \dots \sim \text{Beta}(\chi_1 + g, \chi_2 + m - g)$
$\eta \sim G(\psi_1, \psi_2)$	$\eta Y, \dots \sim G(\psi_1 + \bar{Q} - s, (1/\psi_2 - \log k)^{-1})$

<sup>1</sup> In the case of  $\kappa$  and  $\omega$ , the prior is set on  $\kappa + \omega$  and  $\rho = \kappa / (\kappa + \omega)$ .

### 3 Application

The financial flows at international level have experienced significant changes in the last decades including the 2008 crisis [e.g., 8]. The DSBM with infinite communities can identify specific network structures and its evolution, which can have potential consequences in terms of contagion and financial stability. In this sense, the following is an application of the model described in Section 2 to the data collected by Bank for international Settlements (BIS) on bilateral cross-border claims (and liabilities). As in [1], given that the data presented by the BIS is in a bank-country format, that is a banking system reports its position with respect to a country, we transform the data to a country-country format for most of the cases using data triangulation, with this the flows comprise other sectors and the missing data are minimized. The resulting network includes 31 countries for the period 2001–2019.<sup>2</sup>

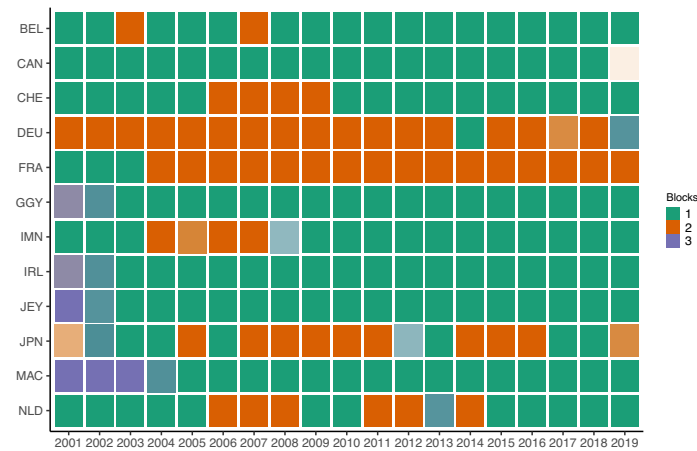
The main results are showed in Fig. 2 and Table 2. Regarding the number of communities, the 53% of draws result in three communities, but still there is some uncertainty given the relative frequency of four communities. Using the former number, 38% of countries have experience at least one change of membership in the period 2001-2019. These countries are presented in Fig.2. Although state persistence is high, there are sudden changes in JPN, NDL, DEU and BEL. Other countries, such as US and GBR are not in the figure because they remain in the same community. In the case of CAN, it seems stable in terms of posterior mode, but in 2019 it starts a transition to another community represented with lower posterior probability.

**Table 2** Relative frequency of the number of communities in the network of financial flows

Q	2	3	4	5	6	7	8	9
	0.12	52.70	32.97	9.70	3.28	1.00	0.20	0.01

<sup>2</sup> This sample covers only reporting countries, no destinations, a subset of the countries available. The countries (dependencies or relevant regions) are: Austria, Australia, Belgium, Brazil, Canada, Switzerland, Chile, Germany, Denmark, Spain, Finland, France, United Kingdom, Guernsey, Greece, Hong Kong SAR China, Ireland, Isle of Man, Italy, Jersey, Japan, South Korea, Luxembourg, Macao SAR China, Mexico, Netherlands, Philippines, Sweden, Taiwan, United States and South Africa.

**Fig. 2** Countries' membership by year, only the countries which experience a change of membership are included (color intensity is proportional to the posterior probability of the posterior mode)



## References

- [1] Brei, M., von Peter, G.: The distance effect in banking and trade. *Journal of International Money and Finance* **81**, 116–137 (2018)
- [2] Dufays, A.: Infinite-state Markov-switching for dynamic volatility. *Journal of Financial Econometrics* **14**(2), 418–460 (2016)
- [3] Fan, X., Cao, L., Da Xu, R.Y.: Dynamic infinite mixed-membership stochastic blockmodel. *IEEE Transactions on Neural Networks and learning systems* **26**(9), 2072–2085 (2014)
- [4] Fox, E.B., Sudderth, E.B., Jordan, M.I., Willsky, A.S.: A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics* pp. 1020–1056 (2011)
- [5] Kim, C., Nelson, C.R.: State-space models with regime switching: classical and Gibbs-sampling approaches with applications, vol. 1. The MIT press (1999)
- [6] Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(4), 1119–1141 (2017)
- [7] Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**(476), 1566–1581 (2006)
- [8] Tonzer, L.: Cross-border interbank networks, banking risk and contagion. *Journal of Financial Stability* **18**, 19–32 (2015)
- [9] Yang, T., Chi, Y., Zhu, S., Gong, Y., Jin, R.: Detecting communities and their evolutions in dynamic social networks—a bayesian approach. *Machine learning* **82**(2), 157–189 (2011)

# Cross-Subject EEG Channel Selection for the Detection of Predisposition to Alcoholism

## *Selezione Generalizzabile tra Soggetti di Canali EEG per l'Identificazione di predisposizione all'alcolismo*

Michela Carlotta Massi, Francesca Ieva

**Abstract** Electroencephalogram (EEG) is a powerful technology for the early detection, among others, of alcoholism. However, multiple electrodes placed on the scalp to record brain signals may introduce noisy and redundant information, hinder performance and increase computational times in the task of automated decoding of EEG signals. In this work we propose a novel end-to-end Representation Learning-based algorithm to select the most relevant EEG channels to perform detection of predisposition to alcoholism, in a subject-agnostic way. Indeed, EEG signals are characterized by strong subject-specific variance potentially affecting the generalizability of the selection. Results are promising, especially compared to the very limited literature on cross-subject EEG channel selection.

**Abstract** *L'Elettroencefalogramma è una tecnologia potente per la diagnosi precoce di alcolismo. Tuttavia, la presenza di molteplici elettrodi sullo scalpo per la registrazione dei segnali cerebrali può introdurre informazione ridondante e rumorosa, penalizzare la resa ed aumentare i tempi computazionali di approcci automatici alla classificazione di segnali EEG. In questo lavoro proponiamo un nuovo algoritmo end-to-end basato su Representation Learning, per selezionare i canali più rilevanti per il riconoscimento della predisposizione all'alcolismo in modo agnostico al soggetto. Infatti i segnali EEG sono caratterizzati da forte variabilità tra soggetti, che aumenta la complessità di ottenere una selezione generalizzabile. I risultati ottenuti sono promettenti, specialmente comparandosi alla letteratura molto limitata nel campo della selezione canali EEG indipendentemente dal soggetto.*

**Key words:** Representation Learning, Signal Processing, Feature Selection, EEG Channel Selection

---

Michela Carlotta Massi

MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano

CADS - Center for Analysis Decisions and Society, Human Technopole

e-mail: michelacarlotta.massi@polimi.it

Francesca Ieva

MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano

CADS - Center for Analysis Decisions and Society, Human Technopole

CHRP - Center for Healthcare Research and Pharmacoepidemiology, Bicocca University

## 1 Introduction

One of the main difficulties in early detection of alcoholism is the unreliability of the information presented by patients with addiction [10]; this hampers diagnosis and reduces the effectiveness of treatment. However, alcohol affects the Central Nervous System (CNS) directly, causing changes in brain functions. One way to check the changes caused by alcohol is through an EEG exam which can identify different types of brain activities through electrodes placed on the scalp. Automatically decoding EEG signals call for novel Statistical and Machine Learning approaches [4], to reduce time and efforts on the clinicians side. A typical EEG exam foresees the recording of signals from multiple sites of the head. However, applying a large number of EEG channels may present several drawbacks: it could (i) include noisy and redundant signals; (ii) induce longer preparation times and (iii) lead to higher computational time and lower performance in the automated processing of signal data for early detection of alcoholism. The development of effective channel selection algorithms is one of the most relevant strategies to overcome all the aforementioned issues at once [6]. However, EEG data is known to be highly subject variant. To the best of our knowledge, only a very limited number of studies devoted to subject-independent channel selection can be found in literature. One recent example can be found in [5], with very poor results ( $\sim 61\%$  average accuracy).

In this work we propose an algorithm to diagnose a predisposition to alcoholism by exploiting only a subset of the available channels, selected in a subject-agnostic fashion. Moreover, the algorithm here proposed is end-to-end, meaning that it does not require preprocessing of raw signals, which is oftentimes a cumbersome and knowledge-intensive procedure.

## 2 Materials: EEG Database

The dataset used in this work is a large public EEG database, available through UCI Machine Learning Repository<sup>1</sup>. This dataset was developed to examine genetic predisposition, through EEG signals, to alcoholism. To elicit the Event-Related Potential (ERP), a modified delayed Visually Evoked Potential (VEP) matching-to-sample task was used, in which two picture stimuli (i.e. objects chosen from the 1980 Snodgrass and Vanderwart picture set [11]) appeared in succession: a first picture stimulus (S1) was followed by a second stimulus (S2) either matching or non-matching the first picture. The database includes 122 subjects, equally splitted between two classes, alcoholic and controls. Each subject completed 120 trials. The signal acquisition is performed according to the 10–20 International System with 64 electrodes placed on the scalps of the subjects and recordings were sampled at 256 Hz (3.9-msec epoch) for 1 second. In this experiment we aimed at classifying trials (the class was determined by the subject associated to each trial) on the basis of the brain signals produced as response to the first visual stimulus (S1).

---

<sup>1</sup> <https://archive.ics.uci.edu/ml/datasets/EEG+Database>

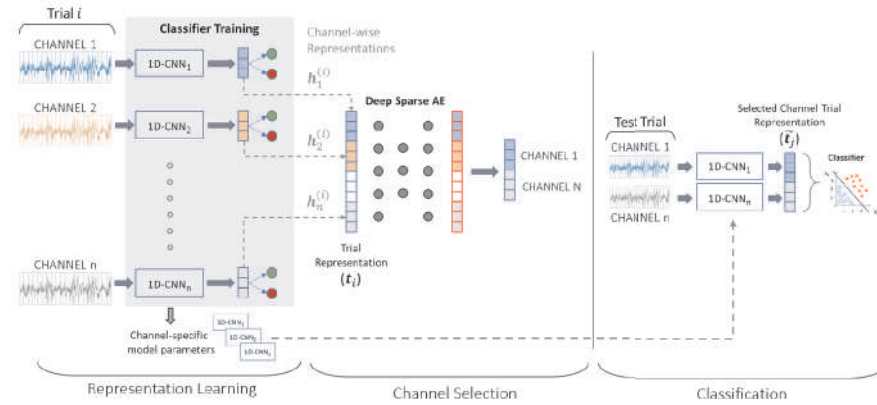


Fig. 1 Algorithm process flow

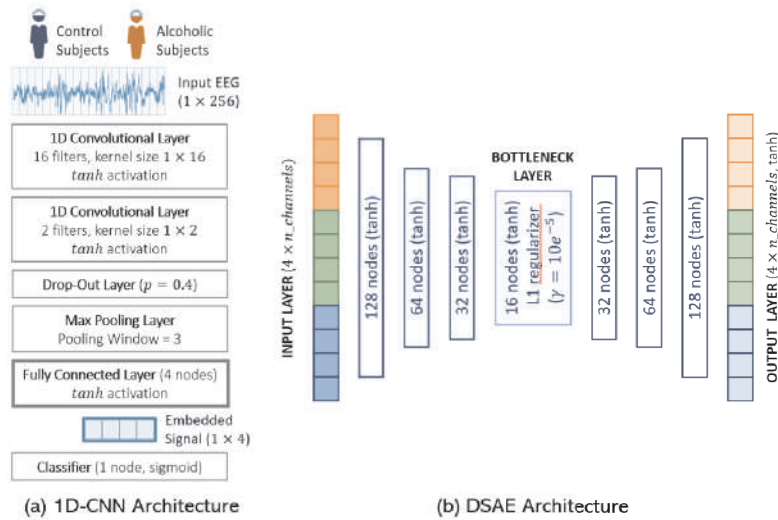
### 3 Proposed Methodology

To detect predisposition to alcoholism by exploiting the least number of EEG channels we propose an End-to-End Representation Learning (RL)-based algorithm that reduces signal dimensionality in a channel-wise fashion and selects the most relevant electrodes across subjects. The algorithm is composed of modules tailored to address different parts of the process and provide a variety of advantages for the task at hand. Its multi-step process is depicted in Figure 1.

The algorithm is designed to learn channel-specific 1-Dimensional Convolutional Neural Networks (1D-CNN) to embed signals grouped by electrode in a latent space of small dimensionality that maximizes intra-class separability. To do that, we consider each 1DCNN as composed of an *encoder* and a subsequent *classifier*. The encoder maps the signals from the  $J$ -dimensional input space, into an  $M$ -dimensional embedding space, where  $M < J$ . The whole model is then parametrized with supervised training to classify the signals as originating from alcoholics or controls. After training, the embedded  $M$ -dimensional vectors from each of the  $C$  channels are extracted from the *encoder*, and the algorithm builds a unique representation of each trial by concatenating the  $C$  embeddings into a trial vector  $\mathbf{t} \in \mathbb{R}^{1 \times (M \times C)}$ .

After that, the Channel Selection (CS) module relies on a Feature Selection method developed in [8]. This method exploits Deep Sparse AutoEncoders (DSAE) in an ensemble-like fashion to select the most relevant features to discriminate between minority and majority class. In particular, it analyzes the feature-wise average *difference* in Reconstruction Error ( $\Delta RE$ ) between classes after training each component on majority class only. In this context of application, the algorithm is adapted to select channels instead of single features. Indeed, Channels are ranked in terms of average channel  $\Delta RE$ , and top  $K$  channels are selected.

At test time, our algorithm transfers the parametrized subgroup of selected channel-specific 1D-CNNs to embed new signals, obtaining new trial vectors  $\hat{\mathbf{t}} \in \mathbb{R}^{1 \times (M \times K)}$  of small dimensionality and high predictive power that can be fed to any classifier.



**Fig. 2** (a) Architectural details of the 1D-CNNs employed for the described experiment. (b) DSAE components’ architectural details. DSAEs are exploited for channel selection from embedded trial vectors.

### 4 Experimental and Implementation Details

To test the cross-subjectivity of our channel selection procedure, we splitted the 122 subjects in training (102 subjects, equally subdivided between alcoholics and controls) and test group (20 new subjects, 50% alcoholics and 50% controls). The former group with all its trials was exploited for channel-wise 1DCNNs training and channel selection, while the latter was supplied to the algorithm to test the classification performance.

To perform the channel-wise embedding of the EEG recordings we had to train several channel-specific models. We opted for a shallow 1D-CNN and the specific architectural details are reported in Figure 2.(a). Hyperparameters were chosen by randomly sampling 10,000 training signals irrespectively of the channel - to make the tuning generalizable across electrodes - equally splitted between classes, and performing random search of the best combination. After setting hyperparameters, each channel-specific 1D-CNN was trained for 200 epochs with a batch size of 1,000 signals. For what concerns the channel selection module, details are reported in Figure 2.(b). Each DSAE model in the ensemble (30 components in total) was trained for 300 epochs with a batch size of 500 training trials. The whole algorithm was implemented in Python 3.7, exploiting Keras framework with Tensorflow backend and scikit-learn. To evaluate the performance on test set we adopted several classifiers, but we report here results for the best performer only, i.e. Support Vector Machines (SVM). We evaluated whether the channel reduction would impact the performance of the classifier by first trying to classify trials using all 61 channels (after their embedding via 1DCNNs and transformation into trial vectors) and then with smaller



subsets of 30, 20, 15, 10 and 5 most relevant channels. The performance of the classification was measured using the Area Under the ROC curve (AUROC) and Accuracy metrics by cross-validating 10 times.

#### 4.1 Results

Results for this experiment are reported in Table 4.1. This experimental setting is lightly comparable to [9, 10] and benchmark algorithms therein, even though performance measurements and data splitting criteria are not always clear from the original papers. Our algorithm obtains a satisfactory accuracy, and a very high AUROC performance, indicating a great precision in identifying the positive class (i.e. *alcoholics*). Our best classifier (SVM) with only 5 electrodes surpasses the performance in [9] with 4 channels (75.13% average accuracy). However, in this work the authors exploit PCA for channel selection applied to the whole dataset, and the lack of information on splitting criteria or performance standard deviations suggest that they are reporting training accuracy measures, which are overestimated compared to our test values. The average accuracy reported more recently in [10] with 11 channels ( $\sim 93\% \pm 3.3$  with the best proposed approach and SVM classifier) is therein defined as the state-of-the-art on this data. Their performance is higher compared to ours with a similar number of channels. However, in their work they perform channel selection evaluating the mean-variance of each channel for all subjects in the dataset before proceeding with feature extraction and classification, therefore their selection is not comparable to our subject-agnostic approach.

N Ch.	AUROC		Accuracy	
	Mean	Std	Mean	Std
61	0.905	0.013	0.807	0.015
30	0.895	0.018	0.816	0.017
20	0.879	0.020	0.798	0.023
15	0.862	0.016	0.793	0.024
10	0.872	0.021	0.786	0.025
5	0.858	0.018	0.762	0.015

**Table 1** Trial Classification Results with SVM classifier in terms of AUROC and Accuracy

## 5 Discussion and Conclusions

In this work we proposed an algorithm to perform cross-subject EEG Channel selection for the task of detecting predisposition to alcoholism. The reduction of EEG channels have several statistical and practical advantages as mentioned in Introduction. Moreover, the algorithm here presented was applied to the very specific clinical task of early diagnosis of predisposition to alcoholism. This clinical application could greatly benefit by more efficient and effective decoding of brain signal recordings, but EEG technology finds application in several other medical fields, s.a. clinical and neurological diagnosis of diseases and disorders like Alzheimer’s disease [7], depression [2], traumatic brain injuries [1] and in the recently spotlighted field of Brain-Computer Interfaces (BCI) [3], all heavily relying on EEG technology because of its high portability, relative low cost, high temporal resolution and few

risk to users. All these fields share the same needs and complexities of the context discussed in this work, therefore would take advantage of an effective cross-subject selection of the most relevant electrodes to aid EEG decoding tasks. Clinical trials could turn more efficient by cutting set-up times, and novel portable BCI technologies could benefit by a generalizable reduction of electrodes that effectively satisfy the needed tasks. In addition, the end-to-end approach of our algorithm makes it more easily transferrable to different tasks, as it avoids long and seldom highly task-specific signal preprocessing procedures. For this reason, this first application and its promising results can open up a stream of further developments to apply our methodology to several other medical EEG decoding domains.

## References

1. Albert, B., Zhang, J., Noyvirt, A., Setchi, R., Sjaaheim, H., Velikova, S., Strisland, F.: Automatic eeg processing for the early diagnosis of traumatic brain injury. In: 2016 World Automation Congress (WAC), pp. 1–6. IEEE (2016)
2. Cai, H., Sha, X., Han, X., Wei, S., Hu, B.: Pervasive eeg diagnosis of depression using deep belief network with three-electrodes eeg collector. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1239–1246. IEEE (2016)
3. Fathima, S., Kore, S.K.: Enhanced differential evolution-based eeg channel selection. In: Symposium on Machine Learning and Metaheuristics Algorithms, and Applications, pp. 162–174. Springer (2019)
4. Gemein, L.A., Schirrmeister, R.T., Chrabaszcz, P., Wilson, D., Boedecker, J., Schulze-Bonhage, A., Hutter, F., Ball, T.: Machine-learning-based diagnostics of eeg pathology. *NeuroImage* p. 117021 (2020)
5. Handiru, V.S., Prasad, V.A.: Optimized bi-objective eeg channel selection and cross-subject generalization with brain–computer interfaces. *IEEE Transactions on Human-Machine Systems* **46**(6), 777–786 (2016)
6. Lan, T., Erdogmus, D., Adami, A., Pavel, M., Mathan, S.: Salient eeg channel selection in brain computer interfaces by mutual information maximization. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 7064–7067. IEEE (2006)
7. Lehmann, C., Koenig, T., Jelic, V., Prichep, L., John, R.E., Wahlund, L.O., Dodge, Y., Dierks, T.: Application and comparison of classification algorithms for recognition of alzheimer’s disease in electrical brain activity (eeg). *Journal of neuroscience methods* **161**(2), 342–350 (2007)
8. Massi, M.C., Ieva, F., Gasperoni, F., Paganoni, A.M.: Feature selection for imbalanced data with deep sparse autoencoders ensemble. *arXiv:2103.11678* (2021)
9. Ong, K.M., Thung, K.H., Wee, C.Y., Paramesran, R.: Selection of a subset of eeg channels using pca to classify alcoholics and non-alcoholics. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 4195–4198. IEEE (2006)
10. Silva, F.H., Medeiros, A.G., Ohata, E.F., Reboucas Filho, P.P.: Classification of electroencephalogram signals for detecting predisposition to alcoholism using computer vision and transfer learning. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 126–131. IEEE (2020)
11. Snodgrass, J.G., Vanderwart, M.: A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory* **6**(2), 174 (1980)

# Some Issues on the Parameter Selection in the Spectral Methods for Clustering

## *Alcune Note sulla Scelta dei Parametri nei Metodi Spettrali di Clustering*

Cinzia Di Nuzzo and Salvatore Ingrassia

**Abstract** Spectral methods for clustering have emerged as effective approaches for finding non-convex clusters in the data; moreover, such methods do not require assumptions on the data because they are based on a matrix of pairwise similarities between the observations depending on some kernel function. The main underlying idea is to cluster the data in a suitable feature space depending on a spectral-based mapping rather than in the original space of the units. Two main issues concern the choice of the kernel function and the estimation of the number of groups. In this paper, we analyze some different proposals presented in literature and provide an explorative approach for the selection of both the number of groups and the proximity measure between the observations.

**Abstract** *I metodi spettrali di clustering costituiscono approcci efficaci per il raggruppamento di dati in accordo a cluster di forma non convessa; inoltre i metodi spettrali non richiedono assunzioni sulla distribuzione dei dati essendo basati sulla matrice di similarità in accordo ad una prefissata funzione kernel. L'idea dei metodi di clustering spettrale è quella di raggruppare i dati in accordo ad una specifica trasformazione spettrale anziché nello spazio originale. In questo contesto, due problemi da affrontare riguardano la scelta della funzione kernel e la stima del numero di gruppi. In questo articolo, vengono analizzati alcuni approcci presentati in letteratura e viene proposto un approccio di tipo esplorativo per la scelta del numero di gruppi e della misura della prossimità fra le osservazioni.*

**Keywords** Spectral clustering, kernel functions, mixture models

---

C. Di Nuzzo  
Department of Economics and Business, University of Catania, Italy. e-mail:  
cinzia.dinuzzo@phd.unict.it

S. Ingrassia  
Department of Economics and Business, University of Catania, Italy. e-mail:  
s.ingrassia@unict.it

## 1 Introduction

Spectral methods for clustering have emerged as effective approaches for finding non-convex clusters in the continuous data, see e.g. [7], [5]. Let  $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  be a set of points in  $\mathcal{X} \subseteq \mathbb{R}^p$ . In order to group the data  $\mathcal{D}$  in  $K$  cluster, the first step concerns the definition of a symmetric and continuous function  $\kappa: \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$  called the *kernel function*. Afterwards, a *weighed matrix*  $W = (w_{ij})$  can be assigned by setting  $w_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ , for  $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ ; in particular, in spectral clustering algorithms, a quite popular choice is the *Gaussian kernel* given by

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\varepsilon) \quad (1)$$

for some fixed parameter  $\varepsilon > 0$ . We introduce the *normalized graph Laplacian* as the matrix  $L_{\text{sym}} \in \mathbb{R}^{n \times n}$  given by  $L_{\text{sym}} = I - D^{-1/2}WD^{-1/2}$ , where  $I$  denotes the  $n \times n$  identity matrix and  $D = \text{diag}(d_1, d_2, \dots, d_n)$  is the *degree matrix* with  $d_i$  being the *degree* of the vertex  $v_i$  defined as  $d_i = \sum_{j \neq i} w_{ij}$ .

The Laplacian matrix  $L_{\text{sym}}$  is positive semi-definite with  $n$  non-negative eigenvalues. For a fixed  $K \ll n$ , let  $\{\gamma_1, \dots, \gamma_K\}$  be the eigenvectors corresponding to the smallest  $K$  eigenvalues of  $L_{\text{sym}}$ . Then the *normalized Laplacian embedding in the  $K$  principal subspace* is defined as the map  $\Phi_\Gamma: \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \rightarrow \mathbb{R}^K$  given by  $\Phi_\Gamma(\mathbf{x}_i) = (\gamma_{1i}, \dots, \gamma_{Ki})$ , for  $i = 1, \dots, n$ , where  $\gamma_{1i}, \dots, \gamma_{Ki}$  are the  $i$ -th components of  $\gamma_1, \dots, \gamma_K$ , respectively. In other words, the function  $\Phi_\Gamma(\cdot)$  maps the data from the input space  $\mathcal{X}$  to a *feature space* defined by the  $K$  principal subspace of  $L_{\text{sym}}$ . Finally, let  $Y = (\mathbf{y}'_1, \dots, \mathbf{y}'_n)$  be the  $n \times K$  matrix, the embedded data in the feature space, where  $\mathbf{y}_i = \Phi_\Gamma(\mathbf{x}_i)$  for  $i = 1, \dots, n$ . The embedded data  $Y$  are clustered according to some clustering procedure, usually the  $k$ -means algorithm is taken into account.

## 2 Practical issues for the Spectral Clustering algorithm

The spectral clustering approach does not require any assumption on the distribution of the data because it is based on the spectral analysis of the similarity matrix. However, we need to select the number of clusters  $K$  and the functional form of the kernel function  $\kappa$  and its parameter(s) as well, for example, the radius  $\varepsilon$  in (1). To this end, in literature many approaches have been proposed and rules of thumb for the choice of the parameters, but there is no unique criterion that can be adopted in general. In the following, we summarize the main ideas proposed in literature and afterward, rather than suggesting a new recipe, we propose an explorative approach for the choice of the number of clusters  $K$  and the parameter in the kernel functions.

**Estimating the kernel function and local scaling parameter.** The choice of the kernel function in spectral clustering algorithms is crucial because it affects the entire data structure in the graph, and consequently, the structure of the Laplacian

and its eigenvectors. An optimal kernel function should lead to a weighted matrix  $W$  having (as much as possible) diagonal blocks: in this case, we get well-separated groups and we are also able to understand the number of groups in that data set by counting the number of blocks. We remark that the choice of the kernel function will also affect the shape of the eigenvectors, in fact with a correct choice of  $\kappa$  the eigenvectors are approximately distributed according to cone structure or aligned to the indicator vectors.

In the Gaussian kernel (1) the main problem concerns the choice of the scale parameter  $\varepsilon$ . In this framework, [9] suggested to calculate a local scaling parameter  $\varepsilon_i$  for each  $\mathbf{x}_i$  and proposed the following kernel function

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\varepsilon_i \varepsilon_j}\right) \quad (2)$$

with  $\varepsilon_i = \|\mathbf{x}_i - \mathbf{x}_h\|$ , where  $\mathbf{x}_h$  is the  $h$ -th neighbor of point  $\mathbf{x}_i$  (similarly for  $\varepsilon_j$ ). This approach has been called *self tuning*. The advantage of this choice is that we get a similarity matrix that does not depend on any parameter so that the algorithm of spectral clustering will be based on the pairwise proximity of the points. The kernel (2) leads to a similarity matrix that depends on the pairwise proximity between the points. However, despite the name *self-tuning* the approach is not completely automatic, we have to select the number  $h$  of neighbors of the point  $\mathbf{x}_i$ . [9] suggested to select  $h = 7$ , but this choice cannot be adopted in general, as we will see later on in Section 3.

In [10], the local density is considered to amplify the intra-cluster similarity taking into account the common neighbors of points  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . The kernel function proposed is

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\varepsilon^2 (\text{CNN}(\mathbf{x}_i, \mathbf{x}_j) + 1)}\right) \quad (3)$$

where  $\mathcal{B}(\mathbf{x}_i, \tau)$  is the sphere centered in  $\mathbf{x}_i$  with radius  $\tau$  (analogously for  $\mathcal{B}(\mathbf{x}_j, \tau)$ ) and  $\text{CNN}(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{B}(\mathbf{x}_i, \tau) \cap \mathcal{B}(\mathbf{x}_j, \tau)$  is the number of points in the join region between the spheres. This kernel is able to capture group membership in cases where point density is very important as in cases where the shape of clusters consists of spirals or U-shaped. In this case, both  $\tau$  and  $\varepsilon$  must be tuned for each data set.

Quite recently, [4] combined the ideas of the kernels (2) and (3), to obtain a kernel function that could take advantages of both kernels functions. Then, once set the radius of the sphere  $\tau$  and the  $h$ -th nearest neighbor, the proposed kernel has the following expression

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\varepsilon_i \varepsilon_j (\text{CNN}(\mathbf{x}_i, \mathbf{x}_j) + 1)}\right). \quad (4)$$

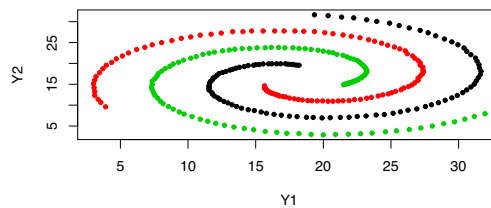
Another kernel function that takes into account the local geometry of each point has been introduced by [2]. This approach automatically determines an adaptive scale control for each point and considers also the point density.

**Estimating the number of clusters  $K$ .** The analysis of the eigenvalues of the Laplacian matrix  $L_{\text{sym}}$  provides a heuristic method to select the number of clusters  $K$  by selecting the number of groups that maximizes the eigengap between two consecutive eigenvalues. Nevertheless, if the groups are not well separated, the first eigenvalues are not exactly 0 and for this reason, the eigengap method is not a good approach in general for real data sets. In such cases, we can consider the number of “almost-connected components” of the graph. Moreover, we remember that if the adjacency matrix is a block matrix (or roughly a block matrix),  $K$  is the number of its blocks.

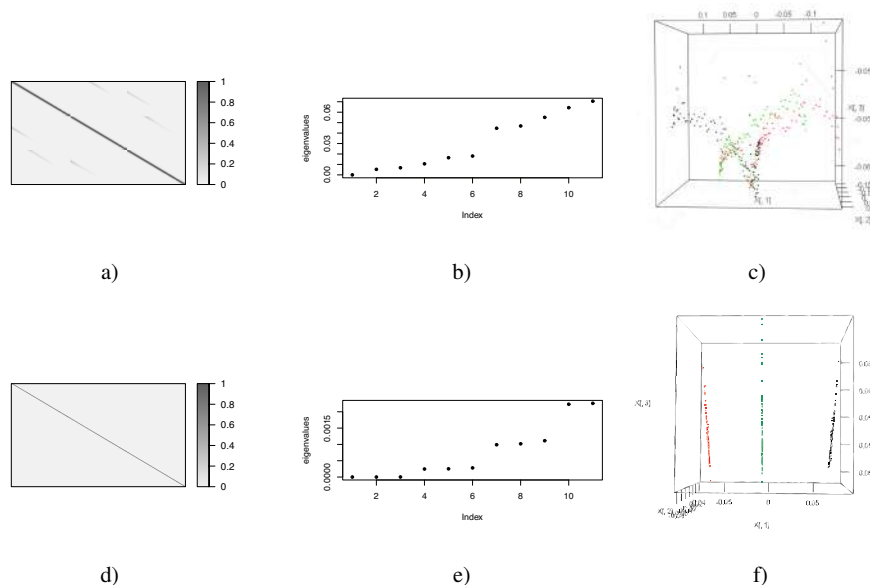
Another way to select the number  $K$  is to consider the results in [3] and [8], so if the eigenvectors assume a cones structure, then  $K$  is the number of the cones in the feature space.

In order to obtain a good result of the spectral clustering, the eigenvectors must be as much aligned as possible to the indicator vectors. This observation has been used by [9] to find the optimal number of groups, the authors try to align as much as possible the eigenvectors to the indicator vectors minimizing a cost function to find the optimal number of groups. Another approach is due to [1], where the authors propose a method for adapting Bartlett’s test for equal variances to the spectral clustering case. Finally, the most recent results are presented in [4], where the authors find  $K$  involving an eigenvector distribution analysis, in particular this method examines the multimodality of the eigenvectors thanks to Dip test statistic.

**A model-based approach.** Usually, the embedded data are clustered according to a  $k$ -means algorithm. The cluster structure of the embedded data has been investigated in [8], [3]. According to these results, if the data are well-separated the normalized Laplacian embedding has an orthogonal cone structure, for this reason, a model-based approach appears to be more effective. Both the  $k$ -means and the mixtures yield convex clusters, but the latter ones allow for more flexible shapes. A related approach for the spectral clustering and the Gaussian mixture models have been previously considered in [6], where the authors propose to use a post-processing step by Spectral Clustering so that the clusters of any shape are discovered. In this way, can be obtained a significant improvement of the clustering performance.



**Fig. 1** Spirals data.



**Fig. 2** Spirals data. Results for the spectral clustering with the self-tuning approach: Similarity matrix  $W$  (a and d); Eigenvalues (b and e); Eigenvectors (c and f). Results concern  $h = 7$  (a, b and c) and  $h = 2$  (d, e and f).

### 3 Main results and conclusions

The performances of many different functions are compared based on simulated data sets. From our experiments, we can state that there is no automatic way to select the kernel function. However, among the proposals, the best kernel functions are (2) and (4), as long as their free parameters are correctly selected. We propose to select the parameters of the kernel functions in such a way that the first eigenvectors are as aligned as possible, providing well-separated directions. As for the number of groups  $K$ , the best way to select it is to evaluate the number of blocks in the similarity matrix and the number of directions in the feature space, trying to reach an agreement between the number of blocks, the number of eigenvector directions and the eigengap candidates. In literature it is usual to consider the number of groups  $K$  that maximizes the eigengap, however, often the number of groups is provided by the first eigengap. We propose to select some candidates for  $K$  in such a way that this quantity emerges from the analysis of three main features: the eigengap, the number of the blocks of the similarity matrix and the number of the directions given out by the first eigenvectors.

Here we present only one example considering the spirals data shown in Figure 1 based on the kernel function (2). [9] propose to select in general  $h = 7$  and this choice leads to the results shown in Figures 2a), 2b) and 2c). As we can see, in this case, we are not able to select the number of groups by looking at the eigenvalues and eigenvectors. Moreover, Figure 2c) shows that the eigenvectors are not aligned (at least, approximately). Therefore, we have to decrease  $h$  to get aligned eigenvectors as possible and a good solution has been found for  $h = 2$ . The results are shown in Figures 2d), 2e) and 2f). In this case, we will obtain the right clustering, and the number of groups can be easily selected both from the number of directions in the eigenvectors space and from the first eigengap between the eigenvalues. We remark that, in this case, the analysis of the similarity matrix in Figure 2d) does not provide useful information. Moreover, we point out that, in other cases, the similarity matrix often provides very good information on the number of groups even as long as the parameters of the kernel function are properly selected.

Finally, several examples have provided significant evidence about the use of the Gaussian mixtures in the last step clustering of the spectral clustering algorithm rather than considering the  $k$ means.

## References

1. Bruneau, P., Parisot, O., and Otjacques, B.: A heuristic for the automatic parametrization of the spectral clustering algorithm. In 22nd International Conference on Pattern Recognition, pages 1313–1318 (2014)
2. David, G. and Averbuch, A.: Spectralcat: Categorical spectral clustering of numerical and nominal data. *Pattern Recognition*, **45**(1), 416 – 433 (2012)
3. Garcia Trillos N., Hoffman F., Hosseini B.: Geometric structure of graph laplacian embeddings. arXiv preprint arXiv:1901.10651 (2019)
4. John, C. R., Watson, D., Barnes, M. R., Pitzalis, C., and Lewis, M. J.: Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, **36**(4), 1159–1166 (2019)
5. Meila M.: Spectral clustering: a tutorial for the 2010's. In: Hennig C., Meila M., Murtagh F., Rocci R. (eds) *Handbook of Cluster Analysis*, Chapman and Hall/CRC (2015)
6. Muzeau J., Oliver-Parera M., Ladret P., Pascal B.: Combining mixture models and spectral clustering for data partitioning. In: Campilho A WZ Karray F (ed) *Image Analysis and Recognition. ICIAR 2020. Lecture Notes in Computer Science*, vol 12132, pp 63–75 (2020)
7. von Luxburg U.: A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416 (2007)
8. Schiebinger G., Wainwright MJ., Yu B.: The geometry of kernelized spectral clustering. *The Annals of Statistics* 43(2):819–846 (2015)
9. Zelnik-Manor, L. and Perona, P.: Self-tuning spectral clustering. *Adv. Neural Inf. Process. Syst.*, **17** (2004)
10. Zhang, X., Li, J., and Yu, H.: Local density adaptive similarity measurement for spectral clustering. *Pattern Recognition Letters*, **32**(2), 352 – 358 (2011)



# The link-match tale: new microdata from unit level association

## *Linkage e matching per generare nuovi database integrati a livello individuale*

Riccardo D'Alberto, Meri Raggi and Daniela Cocchi

**Abstract** Connections among Big Data, administrative registers, general censuses and smart surveying are receiving increasing attention in several domains where statistics is more and more important. When there are different data sets at hand, Record Linkage and Statistical Matching are usually applied to integrate the information which they separately collect. Sometimes, naming and methodological features of the two methods have been shallowly used, contributing to the confusion about their potential applications and scopes. This work aims to spread light on the specific purposes they are meant, clarify to what extent they are similar and how much they differ when they are used interchangeably, discussing a toy example employing the *Collection Faure* data.

**Abstract** *Le potenzialità di un uso congiunto di diverse fonti di dati, integrando Big Data, registri amministrativi, censimenti generali e smart survey, suscitano crescente attenzione. Due metodi frequentemente utilizzati per integrare database differenti sono il Record Linkage e lo Statistical Matching. Le molteplici denominazioni attribuite a questi metodi e la loro apparente somiglianza contribuiscono alla confusione su rispettivi scopi e possibili impieghi. Questo lavoro si propone di far luce sugli obiettivi specifici che Record Linkage e Statistical Matching perseguono, sulle loro caratteristiche fondamentali, analizzando somiglianze, differenze, peculiarità di utilizzo ed eventuali possibilità congiunte di applicazione, tramite un esempio che utilizza i dati della "Collection Faure".*

**Key words:** record linkage, statistical matching, imputation, data integration, individual data

---

Riccardo D'Alberto e-mail: [riccardo.dalberto@unibo.it](mailto:riccardo.dalberto@unibo.it) \*

Meri Raggi e-mail: [meri.raggi@unibo.it](mailto:meri.raggi@unibo.it)

Daniela Cocchi e-mail: [daniela.cocchi@unibo.it](mailto:daniela.cocchi@unibo.it)

Dept. of Statistical Sciences "P. Fortunati", Via delle Belle Arti 41, 40126 – Bologna (BO)

\* *corresponding author*

## 1 Introduction

Connections among Big Data, administrative registers, general censuses and smart surveying are receiving increasing attention in several domains of statistics. The Official Statistics promotes innovative paths for the future development of production and dissemination, through the construction of multi-source databases by integrating different sources. Data integration is promoted as a suitable strategy to address the future provision of information, compensating the progressive reduction of the classical, massive collections of primary data which suffer of high costs and inefficient timeliness. Non-exhaustive examples are [5–7] from EUROSTAT, [1] from the Australian Bureau of Statistics and [15] from the US Census Bureau.

Popular methods to integrate data are Record Linkage (RL) and Statistical Matching (SM). They offer different approaches to the common challenge of aggregating the information available in different sources.

RL dates back to [4] and consist of techniques for identifying the records which are present in different data sets but do refer to the same unit. Since unavailability of unique identifiers may occur or they can be misreported, RL uses the similarity in the values of the variables observed in the data sets at hand in order to declare if they identify the same unit or not. RL is also denoted in several other ways: entity identification, instance merge, data cleaning, etc. [10]. In spite of this variety, RL is straightforwardly identified by *i*) its main aim, i.e. the building of a longer data set from the ones at hand, *ii*) the use of computer science methods, *iii*) the satisfaction of informative needs for which the initial data sets were not meant to [8].

SM may be dated back to [12] and it is used either to integrate individual information coming from different data sets (micro approach) or to construct the joint distribution of variables that are not originally observed together (macro approach). Like RL, SM receives different names: data fusion, data matching, file concatenation, etc. [14]. Its main features are that *i*) there is (at least) one variable observed in both data sets and (at least) two variables that are separately observed, one for each data set [2], *ii*) the number of units that overlap among the data sets is negligible [14] and, *iii*) data sets at hand are identified as “donor(s)” and “recipient(s)” [3].

RL and SM differ since the former identifies the units to which records of different data sets refer to, while the latter deals with units that are just “similar”. Moreover, RL treats data sets at hand as symmetrical while in SM, hierarchical relations exist and/or have to be defined.

Sometimes, the name of the two methods has been used shallowly and there is still confusion about their potential applications and scopes. The present work, focusing on the joint use of RL and SM, aims to clarify 1) for which purposes they are meant, 2) to what extent they are similar/how much they differ, 3) if they can be used interchangeably and, 4) to deepen the grey region that the two methods share, with an application on the *Collection Faure* [11] real data.

## 2 Unit level association: linkage and matching

The simplified situation of two data sets, A and B, is considered. Information available on the a-th unit (with  $a = 1, \dots, n_A$ ) is in data set A, while the information on the b-th unit (with  $b = 1, \dots, n_B$ ) comes from data set B.

The values of a variable  $X$  consist of  $e_k$  elements (often typographic characters) with  $k = 1, \dots, K$  (for the sake of simplicity,  $K_{X_A} = K_{X_B}$ ), a binary agreement indicator for these elements for each a and b is definable as  $\gamma_{e_k^a e_k^b}$ . This equals 1 if  $e_k^a = e_k^b$ , and 0 if  $e_k^a \neq e_k^b, \forall a, b$ . Therefore,  $\Gamma_{ab} = K^{-1} \sum_{k=1}^K \gamma_{e_k^a e_k^b}$  can be defined as the ‘‘similarity weight’’ associated to the variable  $X$ , i.e. the frequency of similar characters that the variable  $X$  does present between the a-th and b-th units.

The a-th and b-th units are potentially identifiable as the same record based on  $X$ . They can be matched if  $\Gamma_{ab} \geq \Gamma_{ab}^*$ , where  $\Gamma_{ab}^*$  is the chosen threshold. Then, let be  $M \in (0, 1)$  the true (theoretical) matches index, such that  $M = 1$  indexes a matched pair of units,  $M = 0$  an unmatched pair. It follows that  $U = (1 - M)$  indexes the true (theoretical) unmatched. Being that the true match status is unknown, the conditional probability of a pair of units having a similarity pattern  $\Gamma_{ab}$  given that they are truly matched is:  $m_{ab} = P(\Gamma_{ab} | M_{ab} = 1) \equiv P(\Gamma_{ab} | M_{ab})$ . Hence, the conditional probability that a pair of units have a dissimilarity pattern  $\Gamma_{ab}$  given that they are truly unmatched can be defined as:  $u_{ab} = P(\Gamma_{ab} | U_{ab} = 1) \equiv P(\Gamma_{ab} | U_{ab})$ . These probabilities were defined at first by [9] and they are commonly known as  $m$ - and  $u$ -probabilities. They highlight a relevant feature: in RL, the units observed by data at hand are (implicitly) assumed as, at least partially, overlapping. The variables observed in both data sets can be either misreported or do change over time, while unique units identifiers are absent or misreported. In these cases, RL aims to match, by means of pseudo-identifiers, the units observed in different data sets which detect the same entity. The likelihood ratio ( $m_{ab}/u_{ab}$ ) is the matching weight to use for each pair.

In SM, it is unknown whether the set of units is overlapping or not, neither assumptions are made in this sense. Two data sets, A and B, collect information on (at least) one variable  $X$  observed in both and (at least) two variables that are observed in an exclusive way, one in A, e.g.,  $Y$  and the other in B, e.g.,  $Z$ . The donor data set is B, while A is the recipient. Following [2, 3], assumptions are: 1) A and B contain information on two representative samples of the same population; 2)  $A \cup B$  is a unique sample of the  $n_A + n_B$  iid observations from  $f(X, Y, Z)$ .

Non-parametric SM is applied to integrate what observed only in one data set with the information observed only in the other data set. Let  $Z$  be the variable to be transferred from B to A, i.e. the ‘‘imputation’’ variable. Imputation is based on the presence of the matching variable  $X$  in both A and B. The goal is to create a synthetic (complete) data set  $C_{n_A \times 3} = \{X^A, Y^A, Z^B\}$ . This data set is called ‘‘synthetic’’ because it does not come from the direct observation/collection of information; in other words, it is artificial [3]. It is ‘‘complete’’ in the sense that, at the end, all the variables collected either in one or in the other data source are aggregated [3].

Let  $d_{ab}$  be a distance function between units with respect to the matching variable  $X$ . If  $x_a = x_b$ ,  $d_{ab} = 0$ , while  $d_{ab} > 0$  if  $x_a \neq x_b$ . Then, let  $\omega_{ij} \in \{0, 1\}$  be the index

of the potential matching of units, with  $\omega_{ab} = 0$  if  $d_{ab} > t$ , while  $\omega_{ab} = 1$  if  $d_{ab} \leq t$ , where  $t$  is the threshold defining the maximum level of distance that is required in order that two units are considered a matched pair. The Manhattan distance [13] is a very popular one:  $d_{ab} = |x_a - x_b|$ , with  $t \leq \min_{1 \leq b \leq n_B} (d_{ab}), \forall a$ .

Similarly to blocking in RL [16], in SM the potential pairs of donor and recipient units can be restricted in homogeneous sub-groups constituting “donation classes” as follows. Let  $n_B^{NA}$  be the size of the donor and recipient pairs set. Let  $T$  be a discretized variable observed, in addition to  $X$ , both in  $A$  and in  $B$ , whose categories  $T_r$ , with  $r = 1, \dots, R$ , identify the donation classes. In this case the aforementioned size set is restricted to  $(n_{B,T_r})^{NA,T_r}$ .

In SM, there are not  $m$ - and  $u$ -probabilities assessing the uncertainty of true or false matches. The units’ association at individual level is defined by their dissimilarity with respect to the observed  $X$  and the SM identification problem is connected to the (unobserved) joint distribution  $f(X, Y, Z)$  generated by imputation in the synthetic (complete) data set. Following [14], four levels of matching validity are stressed: 1) individual values preservation, 2) preservation of the joint distribution  $f(X, Y, Z)$ , 3) correlation structure preservation and, 4) preservation of the bivariate marginal distribution, usually considered the minimum requirement.

### 3 Collection Faure (real) data application

The private collection of Jean Faure (1830-1914) is endowed of different inventories dated 1937, 1939, 1948. From these inventories, three data sets are extracted for demonstrative purposes and enriched with information on the dimensions of the art pieces. The variables (in Table 1) are the artist name, the art piece title and its *medium* (painting, drawing, watercolour, sculpture, etc.), height and length. The art piece quotation is present only in 1939, while the dimensions are included only in 1937 and 1939. No unique identifier integrates the archives: the ids from 1937 and 1939 are not unique, while the art pieces’ titles may change.

Table 1: Variables extracted from the archives of the *Collection Faure*

1937 (n=102)	1939 (n=120)	1948 (n=168)
<i>id-37</i>	<i>id-39</i>	
<i>artist</i>		<i>artist</i>
<i>title-37</i>		<i>title-48</i>
<i>medium</i>	<i>medium</i>	<i>medium</i>
<i>height*</i>	<i>height*</i>	
<i>length*</i>	<i>length*</i>	
	<i>quotation</i>	

\* These variables have been added by internet retrieval

The goal of this work is to build one data set that integrates all information from the three archives, regardless their time sequence. This is achieved with two different

The link-match tale: new microdata from unit level association

approaches, by two steps which involve both RL and SM, in order to investigate possible differences in the final result. One approach is illustrated in Figure 1 (first SM, second RL), while the other is illustrated in Figure 2 (first RL, second SM). The choice is not neutral: results are affected by the chosen sequence.

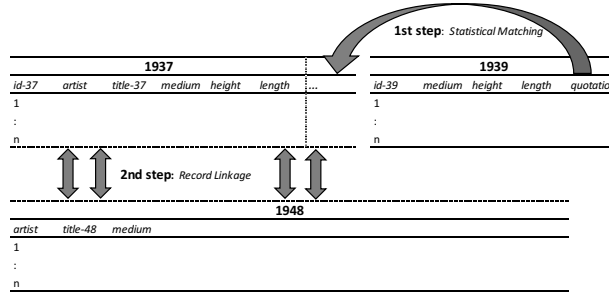


Fig. 1: Match-link approach

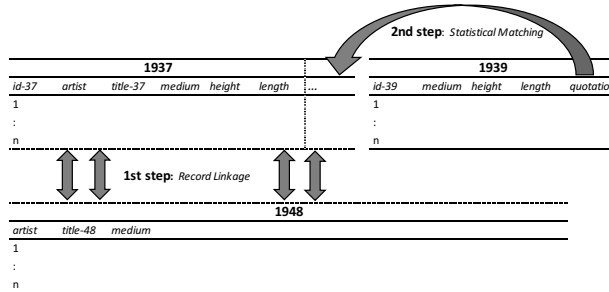


Fig. 2: Link-match approach

Following Figure 1, a synthetic (complete) data set is created by SM imputation, transferring the variable *quotation* from the 1939 donor into the 1937 recipient data set. This imputation is based on the matching variables *height* and *length* (common variables in 1937 and 1939). Donation classes are defined by the variable *medium*. The mean number of donors at the minimum distance is 1.505 (median=1), while the mean recipient-donor distance is 0.205 (median=0). The synthetic data set has 93 rows since in nine cases the value of *medium* is missing. Finally, this data set is completed by probabilistic RL with the 1948 inventory, based on the *artist* and *title* variables and by blocking on *medium*. The  $m$ - and  $u$ -probabilities estimated by the EM algorithm are 0.958 and 0.075 for the variable *artist*, while they are 0.639 and 0.008 for the variable *title*. The final linked data set has 173 rows.

In the approach illustrated in Figure 2, the linked data set is created by probabilistic RL between the 1937 and 1948 inventories. The  $m$ - and  $u$ -probabilities for the same variables are 0.956 and 0.073 (*artist*) and 0.639 and 0.008 (*title*). Finally, the linked data set is imputed with the *quotation* from the 1939 inventory at the aforementioned matching conditions. In relation to the former approach, these

conditions lead to analogous matching results (which show different matched pairs with the same imputed distribution), while the final data set has 175 rows.

Following the two different approaches, the resulting synthetic data sets contain the same variables but they differ in the rows number and in the distribution of the imputed variable due to the differential presence of the missing values. This highlights the importance of the order choice in the approaches' sequence.

## References

1. Australian Bureau of Statistics – ABS – Multi-Agency Data Integration Project (MADIP), <https://www.abs.gov.au/websitedbs/d3310114.nsf/home/statistical+data+integration+-+madip+research+projects> [accessed: Feb, 2021]
2. Conti PL, Marella D, Scanu M, Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators, *Comput Stat Data An*, 53, 354–365, (2008)
3. D'Orazio M, Di Zio M, Scanu M, *Statistical matching: Theory and practice*, 256 pp. John Wiley & Sons, Hoboken (2006)
4. Dunn HL, Record Linkage, *Am J Public Health*, 36, 1412–1416, (1946)
5. European Commission – EUROSTAT – ESSnet AdminData Project, <https://ec.europa.eu/eurostat/cros/content/use-administrative-and-accounts-data-business-statistics.en> [accessed: Feb, 2021]
6. European Commission – EUROSTAT – Data Warehouse Project <https://ec.europa.eu/eurostat/cros/content/data-warehouse.en> [accessed: Feb, 2021]
7. European Commission – EUROSTAT – ESSLait Project <https://ec.europa.eu/eurostat/cros/content/esslait.en> [accessed: Feb, 2021]
8. Fellegi IP, Record linkage and public policy – A dynamic evolution, In: Alvey W, Jamerson B, (eds.), *Record Linkage Techniques – Proceedings of an International Record Linkage Workshop and Exposition*, 3–12, (1997)
9. Fellegi IP, Sunter AB, A theory for record linkage, *J Am Stat Assoc*, 64, 1183–1210, (1969)
10. Gu L, Baxter R, Vickers D, Rainsford C, Record linkage: Current practice and future directions, 32 pp. CSIRO Mathematical and Information Sciences Technical Report, (2003)
11. Harvard Dataverse, La collection du docteur Jean Faure (1862-1942): des inventaires au musée d'Aix-les-Bains <https://doi.org/10.7910/DVN/YOOFYE> [accessed: Feb, 2021]
12. Okner BA, Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File, *Ann Econ Soc Meas*, 1, 325–342, (1972)
13. Rodgers WL, An evaluation of statistical matching, *J Bus Econ Stat*, 2, 91–102 (1984)
14. Rässler S, *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*, 238 pp. Springer, New York (2002)
15. Schachter J, Data Integration at the U.S. Census Bureau: Recent Efforts to Improve Estimates of International Migration, 10 pp. United Nations Statistical Commission High-level Panel Technical Report, (2019)
16. Steorts RC, Ventura SL, Sadinle M, Fienberg SE, A Comparison of Blocking Methods for Record Linkage, In: Domingo-Ferrer J (ed.), *Privacy in Statistical Databases*, 253–268, (2014)

## 4.17 New developments in Bayesian methods

# Spatio-temporal analysis of the Covid-19 spread in Italy by Bayesian hierarchical models

## *Analisi spazio-temporale della diffusione del Covid-19 in Italia attraverso modelli gerarchici bayesiani*

Nicoletta D'Angelo, Giada Adelfio and Antonino Abbruzzo

**Abstract** In this paper, we investigate the spatio-temporal spread pattern of the virus Covid-19 in Italy, during the first wave of infections, from February to October 2020. We provide a disease mapping of the virus infections, by using the Besag-York-Mollié model and its spatio-temporal extensions. Our results confirm the effectiveness of the lockdown action, and show that, during the first wave, the virus spread by an inhomogeneous spatial trend and each province was characterised by a specific temporal trend, independent of the temporal evolution of the observed cases in the other provinces.

**Abstract** *In questo articolo, analizziamo la diffusione spazio-temporale del virus Covid-19 in Italia, durante la prima ondata di infezioni, da febbraio a ottobre 2020. Forniamo, inoltre, una mappatura della malattia usando il modello Besag-York-Mollié e le sue estensioni spazio-temporali. I nostri risultati confermano l'efficacia del lockdown e mostrano che, durante la prima ondata, il virus si è diffuso in maniera non omogenea nello spazio, e ogni provincia è stata caratterizzata da uno specifico trend temporale, indipendentemente dall'evoluzione temporale dei casi rilevati nelle altre province.*

**Key words:** Italian Covid-19; Besag-York-Mollié model; Spatio-temporal models; Disease Mapping.

## 1 Introduction

This paper provides a model proposal for describing the Covid-19 diffusion in Italy from February to October 2020, interpreting the spatio-temporal evolution of the infections caused by the most recently discovered coronavirus during the first wave.

---

Nicoletta D'Angelo, Giada Adelfio and Antonino Abbruzzo  
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy e-mail:  
nicoletta.dangelo@unipa.it, giada.adelfio@unipa.it, antonino.abbruzzo@unipa.it



The Covid-19 Italian data are collected at an aggregate level, reporting the daily counts of the infected people in the regions and provinces. Therefore, they represent an example of areal data which can be described by the Besag–York–Mollié (BYM) model ([2]). A spatio-temporal model is proposed, interpreting the infection spread along all the Italian provinces, accounting also for the temporal evolution of the government restricting actions. Section 2 introduces the hierarchical spatio-temporal model. Section 3 presents the application to the Italian Covid-19 spread pattern. Finally, conclusions are in Section 4.

## 2 Spatio-temporal models for disease mapping

The BYM model is an extension of the ICAR (Intrinsic Conditional Autoregressive) model, obtained by adding spatially unstructured random effects to the spatially structured random effect. The latter is a realisation of a GMRF with zero mean and a sparse precision matrix capturing strong spatial dependence. Let

$$Y_i \sim \text{Poisson}(E_i \lambda_i)$$

be the random variable number of cases in the region  $D_i, i = 1, \dots, n$ , and  $E_i$  the corresponding expected cases count for the  $i$ -th spatial unit, computed externally through available population's information. The BYM model, in its general form, is defined with the linear predictor for the relative risk specified on the logarithmic scale:

$$\log(\lambda_i) = \alpha + u_i + v_i, \tag{1}$$

where  $\alpha$  is the intercept quantifying the average of the counts in all the  $n$  regions. The random spatial process is the sum of two area-specific effects: an independent Gaussian process  $u_i$ , with variance  $\sigma_u^2$  and a GMRF  $v_i$ , with variance  $\sigma_v^2$ . For each region, the value of the GMRF component depends on the average from the neighbouring regions

$$v_i | v_{-i}, \sigma_v^2 \sim \text{GMRF} \left( \frac{\sum_{j \in d_i} v_j}{d_i}, \frac{\sigma_v^2}{d_i} \right),$$

where  $d_i$  is the number of areas which share boundaries with the  $i$ -th one. Two regions are usually defined as neighbours if they share a common border. More in detail, the parameter  $u_i$  represents the unstructured residual, modelled as

$$u_i | \sigma_u^2 \sim N(0, \sigma_u^2).$$

When the ratio  $\sigma_u^2 / \sigma_v^2$  increases, the random process' spatial dependence increases as well, providing a smoother surface for the intensity.

The spatio-temporal disease mapping extension is widely used in disease surveillance studies. In practice, the standard disease mapping model (1) is modified for accounting for a temporal component, which is indexed by  $t = 1, \dots, T$ . For each time  $t$ , a parametric structure, as well as a non parametric trend can be specified for

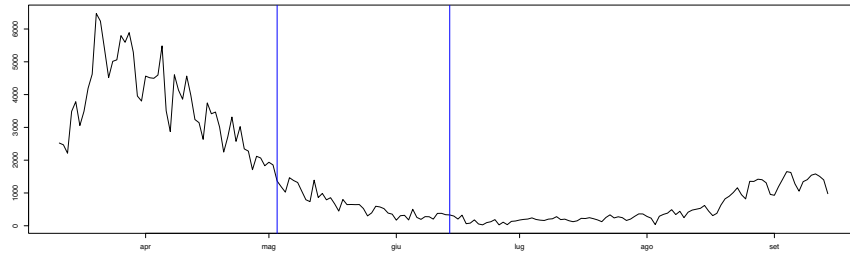


Fig. 1: Daily new cases in Italy during the first wave of infections.

the log-intensity. The spatio-temporal model can be further extended to allow for interactions between space and time components, for studying the evolution of the temporal trend of the studied phenomenon among different areas, with the following specification:

$$\log(\lambda_i) = \alpha + u_i + v_i + \gamma_t + \phi_t + \delta_{it}. \quad (2)$$

where the term  $\gamma_t$  represents the temporally structured effect, modelled dynamically using a random walk of order 1 (or of order 2), defined as  $\gamma_t | \gamma_{t-1} \sim \text{Normal}(\gamma_{t-1}, \sigma^2)$ , and where  $\phi_t$  is a Gaussian exchangeable prior  $\phi_t \sim \text{Normal}(0, 1/\tau_\phi)$ . The parameter vector  $\delta$  has a Gaussian distribution with a precision matrix given by  $\tau_\delta \mathbf{R}_\delta$  where  $\tau_\delta$  an unknown scalar, while  $\mathbf{R}_\delta$  is the structure matrix, identifying the type of temporal and/or spatial dependence between the elements of  $\delta$ . In [4] and [3] four different definitions for the structure matrix are provided.

### 3 Application to Italian Covid-19 data

In this section, we analyse the number of people infected by the Covid-19 in all the 107 Italian provinces, from February 24th to October 7th, 2020. On the 7th of October, due to the new rise of the detected coronavirus cases, the Italian Government postponed the end of the state of emergency to 31 January 2021. Stricter rules were reintroduced to limit the spread of Covid-19, such as imposing the use of protection mask outdoors and forbidding demonstrations and gatherings of people. For this reason, we chose the 7th of October as the end of the first wave of infections in Italy and as the beginning of the second wave, that we do not include in our analysis.

Therefore, we propose a spatio-temporal analysis, accounting also for the temporal domain information, splitting the time frame into three windows, referring to the time stamps identified by the Italian Government. Therefore, we considered the following temporal time frames:

- Phase 1: lockdown (March, 9th - May, 3rd);
- Phase 2: easing of containment measures (May, 4th - June, 14th);

- Phase 3: coexistence with Covid-19 (June, 15th - October, 7th).

These are represented in Figure 1, individuated by the vertical blue straight lines, together with the overall number of new cases recorded in each day of the first wave of the Italian infections.

First, a purely spatial BYM model without external covariates is fitted, with the log-linear predictor specified as in (1), fitting a unique spatial model referring to the whole time period [March, 9th - October, 7th]. Then, we fit spatio-temporal models as in (2), including the time component defined by the three phases and accounting for the four different kind of interactions among the random effects, as specified in [4].

The chosen model with the lowest DIC is the one with the interaction of type II, and it suggests that the structured temporal main effect  $\gamma_t$  and the unstructured spatial effect  $u_i$  interact. In other words, for each  $i$ -th area, the temporal component is described by the parameter vector  $\delta_{i1}, \dots, \delta_{iT}$  and has an autoregressive structure, independent of the ones of the other areas. This result means that the temporal evolution of the cases in each considered province is independent of the temporal evolution of the other provinces' cases.

The posterior means of the district-specific relative risk of detecting cases  $\exp(v + u)$ , not reported here for brevity, can be mapped providing useful information. The overall risk is higher in the Northern regions and provinces, decreasing from the North to the South of Italy, for the given temporal frames. Conditionally to the areas, the infection risk is overall lower in the first phase, that is the period starting with the lockdown action taken by the Italian government. Later, in the second phase, it increases in districts and further increases in the third phase. In particular, in the Northern macro-area, the most risk-exposed districts during the phases are those that have experienced the highest number of cases during the first phase, namely Milan, Bergamo, Brescia and Turin. In the central macro-area, the most affected province is the capital Rome. In the Southern regions, the most affected city is Naples. The specific estimated risk of the most affected provinces, that are also the most populated ones in the three macro-regions, are reported in Figure 2.

## 4 Results and conclusions

In this paper, we provide a model proposal for analysing the spatio-temporal spread pattern of the Covid-19 in Italy. According to the applied spatio-temporal model, both the spatial and temporal random effects are significant and, in particular, the interaction between the structured temporal component and the unstructured spatial one is also significant. This suggests that the temporal evolution of the cases in each considered province is independent of the temporal evolution of the other provinces' cases. Therefore, the contagions, and also their temporal trend, may be caused by some province specific aspects, rather than by the subjects' spatial movements. For a better comprehension of this spread, point models accounting for further information would be of paramount importance.

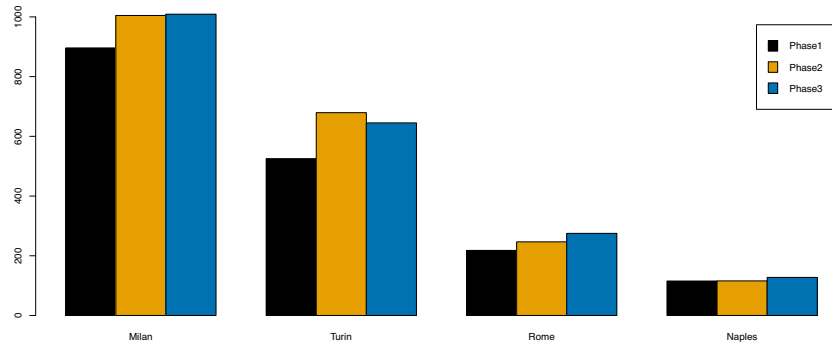


Fig. 2: Top province-specific estimated risk for the three macro-regions.

As a future hint, the availability of more complete datasets, as well as geocoded health data, would be crucial to refer to methods and models accounting for the self-exting behaviour of the epidemic phenomenon, as well as for external factors ([5, 1]).

## References

1. Adelfio, G. and Chiodi, M. (2020). Including covariates in a space-time point process with application to seismicity. *Statistical Methods and Applications*.
2. Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.
3. Blangiardo, M., Cameletti, M., Baio, G. and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*, 4:33–49.
4. Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, 19(17-18):2555–2567.
5. Meyer, S., Held, L., and Höhle, M. (2017). Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. *Journal of Statistical Software*, 77(11):1–55.

# Modelling of accumulation curves through Weibull survival functions

## *Modellazione di curve di accumulazione tramite funzioni di sopravvivenza Weibull*

Alessandro Zito, Tommaso Rigon and David B. Dunson

**Abstract** Recently, Zito et al. (2020) introduced a Bayesian nonparametric framework to model the sequential discovery of distinct entities in a sequence of labelled objects, such as biological species or words in a text. These discoveries are summarized through accumulation curves, which count the number of new species observed over time. The authors directly specified the probability of a new discovery through survival functions of independent and identically distributed random variables, and provided extensive theoretical support. In this paper, we extend the contribution of Zito et al. (2020) by considering the case when these variables follow a Weibull distribution. As an illustration of this methodology, we analyze the Barro Colorado Island tree counts dataset available in the R package `vegan`.

**Abstract** Recentemente, Zito et al. (2020) hanno introdotto una classe di modelli bayesiani nonparametrici per la modellazione di scoperte sequenziali in una collezione di oggetti, come ad esempio specie biologiche o parole in un testo. Tali scoperte vengono tipicamente rappresentate tramite curve di accumulazione, che riportano il conteggio del numero di nuove specie osservate nel tempo. La probabilità di una nuova scoperta è modellata attraverso funzioni di sopravvivenza di variabili aleatorie indipendenti ed identicamente distribuite. In questo manoscritto, estendiamo il contributo di Zito et al. (2020) considerando il caso in cui tali variabili obbediscano una distribuzione Weibull. Come illustrazione di tale metodologia, analizziamo un dataset presente sul pacchetto R `vegan`.

**Keywords:** Accumulation curves, Poisson-binomial distribution, Species sampling models, Weibull distribution

---

Alessandro Zito  
Duke University, Durham, NC, USA e-mail: [alessandro.zito@duke.edu](mailto:alessandro.zito@duke.edu)

Tommaso Rigon  
Università degli Studi Milano-Bicocca, Milano, Italy e-mail: [tommaso.rigon@unimib.it](mailto:tommaso.rigon@unimib.it)

David B. Dunson  
Duke University, Durham, NC, USA e-mail: [dunson@duke.edu](mailto:dunson@duke.edu)

## 1 Introduction

Let  $(X_n)_{n \geq 1}$  be a sequence of labelled entities taking values in  $\mathbb{X}$ , which is the space of biological species or words in a corpus. As  $n$  increases and more objects are observed, some labels may appear more than once. In particular, suppose that among the first  $n$  entities,  $X_1, \dots, X_n$ , a total of  $K_n \leq n$  distinct labels are recorded. The trajectory  $(K_n)_{n \geq 1}$  of the cumulative number of new entities, or species, detected over time is known as the *accumulation curve* [2]. Our goal is to describe a flexible framework for accumulation curves that allows one to i) obtain in- and out-of-sample predictions for the number of distinct species, and ii) retrieve asymptotic estimates for the *species richness*, defined as  $\lim_{n \rightarrow \infty} K_n = K_\infty$ .

In the same spirit of [8], let  $(D_n)_{n \geq 1}$  be a sequence of independent Bernoulli random variables, each indicating whether the entity observed at the  $(n+1)$ th step is new or already encountered, namely

$$\mathbb{P}(D_{n+1} = 1) = \mathbb{P}(X_{n+1} = \text{“new”} \mid X_1, \dots, X_n), \quad n \geq 1.$$

Thus, the associated accumulation curve can be obtained by summing these discovery indicators, so that  $K_n = \sum_{i=1}^n D_i$ , for any  $n \geq 1$ , provided that natural assumptions over the discovery probabilities  $\pi_n = \mathbb{P}(D_n = 1)$  are satisfied. Specifically, it is natural to require that  $\pi_1 = \mathbb{P}(X_1 = \text{“new”}) = 1$ , as the first entity ever observed is necessarily new. Additionally, we impose that  $\pi_n > \pi_{n+1}$  for every  $n \geq 1$ , meaning that the probability of detecting a new label decreases over time, and that  $\lim_{n \rightarrow \infty} \pi_n = 0$ , which implies that new discoveries eventually stop occurring. These three requirements are automatically satisfied within the following general framework.

**Definition 1 (Zito et al. (2020)).** Let  $T$  be a random variable on  $(0, \infty)$  with continuous cumulative distribution function  $F(t; \theta)$  indexed by  $\theta \in \mathbb{R}^p$  and let  $S(t; \theta) = 1 - F(t; \theta)$  be its survival function. The set of probabilities  $(\pi_n)_{n \geq 1}$  are said to be *directed by*  $S(t; \theta)$  if

$$\pi_n = \mathbb{P}(T_n > n - 1) = S(n - 1; \theta), \quad \text{for any } n \geq 1,$$

where  $(T_n)_{n \geq 1}$  are independent and identically distributed random variables according to  $F(t; \theta)$ .

A key advantage of the framework introduced in Definition 1 is its high flexibility in the choice of the survival function  $S(t; \theta)$ . In particular, [8] show that a particular family of distributions, called the three-parameter log-logistic, leads to strong relationships with the Bayesian nonparametric literature on species sampling models [6] and recovers the Dirichlet process of [1] as a special case. However, several other survival functions can be considered. In this manuscript, we focus on the case  $T \sim \text{Weibull}(\phi, \lambda)$ , so that we have

$$\pi_{n+1} = S(n; \phi, \lambda) = \phi^{n^\lambda}, \quad \phi \in (0, 1), \lambda > 0. \quad (1)$$

The likelihood associated to this random mechanism is

$$\mathcal{L}(\phi, \lambda \mid D_1, \dots, D_n) \propto \prod_{i=2}^n \left\{ \phi^{(i-1)\lambda} \right\}^{D_i} \left\{ 1 - \phi^{(i-1)\lambda} \right\}^{1-D_i}. \quad (2)$$

This form leads to substantial flexibility in prediction of trajectories  $(K_n)_{n \geq 1}$  with parameters  $\phi$  and  $\lambda$  controlling the shape, and  $K_\infty$  finite regardless of the parameter values. Properties are described in detail in the following sections.

## 2 Theoretical properties

In the following we describe some important theoretical properties associated with  $T \sim \text{Weibull}(\phi, \lambda)$ . For a more extensive discussion beyond the Weibull case, refer to [8]. If  $(\pi_i)_{i=1}^n$  are directed by  $S(t; \phi, \lambda)$  as in equation (1),  $K_n \sim \text{PB}\{1, \phi, \phi^{2\lambda}, \dots, \phi^{(n-1)\lambda}\}$ , where PB denotes the so-called Poisson-binomial distribution. This is a direct result of  $K_n$  being the sum of independent Bernoulli random variables. In particular, its expected value is

$$\mathbb{E}(K_n) = \sum_{i=1}^n \phi^{(i-1)\lambda}. \quad (3)$$

The expectation in equation (3) is a natural candidate for the in-sample estimate of the species accumulation curve. On the other hand, the out-of-sample prediction can be directly derived from the convenient conjugacy property of the Poisson-binomial distribution. Let  $K_m^{(n)} = K_{n+m} - K_n = \sum_{j=n+1}^{m+n} D_j$  for any  $n, m \geq 1$ . It is easy to see that  $K_m^{(n)} \sim \text{PB}\{\phi^{n\lambda}, \dots, \phi^{(n+m-1)\lambda}\}$ . Hence, if we observe  $K_n = k$  distinct species within the first  $n$  samples, we conclude that

$$\mathbb{E}(K_{n+m} \mid D_1, \dots, D_n) = k + \sum_{j=n+1}^{m+n} \phi^{(j-1)\lambda}, \quad (4)$$

since  $\mathbb{E}(K_n \mid D_1, \dots, D_n) = k$  and  $\mathbb{E}(K_m^{(n)} \mid D_1, \dots, D_n) = \mathbb{E}(K_m^{(n)}) = \sum_{j=n+1}^{m+n} \phi^{(j-1)\lambda}$ . Equation (4) is suitable for out-of-sample predictions: it represents the expected number of species that would be detected if  $m$  additional samples were observed. Such a quantity is of key interest in the species sampling model literature [4].

The infinite limits of equations (3) and (4), when  $n \rightarrow \infty$  and  $m \rightarrow \infty$ , respectively, lead to natural estimators for the species richness. Moreover, the convergence of the prior species richness  $K_\infty$  and the posterior species richness  $(K_\infty \mid D_1, \dots, D_n)$  can be characterized by the following proposition.

**Proposition 1 (Zito et al. (2020)).** *Let  $K_n \sim \text{PB}\{1, S(1; \theta), \dots, S(n-1; \theta)\}$ . Then, there exists a possibly infinite random variable  $K_\infty$  such that  $\lim_{n \rightarrow \infty} K_n \rightarrow K_\infty$ , almost surely, with  $\mathbb{E}(K_\infty) = \sum_{i=0}^{\infty} S(i)$ . Moreover,*

$$\mathbb{E}(T) \leq \mathbb{E}(K_\infty) \leq \mathbb{E}(T) + 1. \quad (5)$$

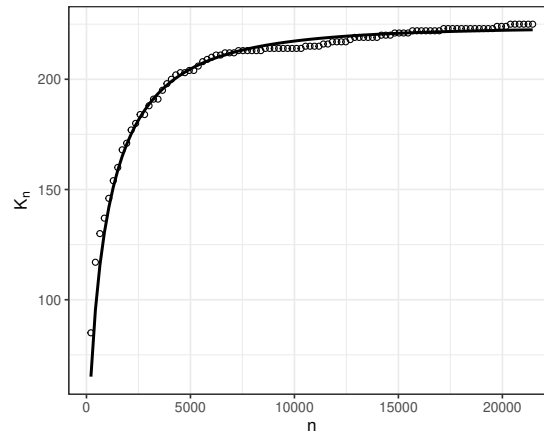
In other words, Proposition 1 implies that the latent variable  $T$  plays a major role in determining the asymptotic behavior of the species richness. This is clear from equation (5), which highlights that an infinite number of distinct species will occur if and only if  $T$  has infinite expectation. When  $T \sim \text{Weibull}(\phi, \lambda)$ , we have that

$$\mathbb{E}(T) = \Gamma\left(1 + \frac{1}{\lambda}\right) \left(-\frac{1}{\log \phi}\right)^{\frac{1}{\lambda}} < \infty, \quad (6)$$

for every  $\lambda > 0$  and  $\phi \in (0, 1)$  and with  $\Gamma(\cdot)$  denoting the gamma function. Hence, under a Weibull survival function, equation (3) always converges to a finite constant when  $n \rightarrow \infty$ . As for the convergence of equation (4), notice that  $\lim_{m \rightarrow \infty} \mathbb{E}(K_m^{(n)}) \leq \lim_{m \rightarrow \infty} \mathbb{E}(K_{n+m}) < \infty$  by Proposition 1.

### 3 Application

We analyse the Barro Colorado Island tree count dataset [3, 7] of the R package `vegan` [5]. The data contain the frequency of appearance of distinct species of trees in 50 different plot samples of 1 hectare square of forest in the island. All trees have a diameter of at least 10 centimeters at breast height. For a detailed description of the geographical coordinates and the environmental characteristics, see [3, 7]. At an aggregate level, the data comprise  $n = 21,457$  trees grouped into  $K_n = 225$  distinct species. Our goal is to determine whether the resulting accumulation curve derived from the tree frequencies is close to convergence and, if so, how many distinct new species should be expected if more trees were inspected in similar areas.



**Fig. 1** Observed discoveries (dots) and estimated accumulation curve  $\mathbb{E}(K_n)$  (solid line) for the tree species counts of the Barro Colorado Island data.



Figure 1 displays the observed data and the in-sample estimate  $\mathbb{E}(K_n)$  for the Barro Colorado Island data, whose frequencies are computed by summing the tree counts across sampling locations. We construct the accumulation curve displayed with a Monte Carlo procedure. In particular, we randomly simulate 1000 different orderings of species obtainable from the raw frequencies, and estimate the parameters  $\lambda$  and  $\phi$  via empirical Bayes by numerically maximizing the likelihood in equation (2) on each associated discoveries sequence. Then, we retain the ordering and the parameters for which the model reaches the highest likelihood value among the resamples. This approach deals with the fact that accumulation curves are inherently order dependent [2]. For a more thorough discussion of order dependence and estimation, refer to [8].

In the Barro Colorado Island data, the observed curve quickly reaches convergence and becomes nearly flat towards the end. The estimate for  $\mathbb{E}(K_n)$  accurately captures this behavior, with estimates for  $\phi$  and  $\lambda$  being approximately 0.777 and 0.357, respectively. Moreover, we obtain that the estimated posterior species richness equals  $\mathbb{E}(K_\infty | D_1, \dots, D_n) = 227$  species, i.e. two more than the observed value  $K_n = 225$ . This suggests that if the sampling procedure were performed in similar nearby areas to the one considered, few new species are likely to be detected.

## References

1. Ferguson, T. S.: A Bayesian analysis of some nonparametric problems. *Ann. of Stat.* **1**(2), 209–230, (1973)
2. Gotelli, N. J. and R. K. Colwell: Quantifying biodiversity: procedures and pitfalls in themeasurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391, (2001)
3. Harms K.E., Condit R., Hubbell S.P. and Foster R.B.: Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. *J. Ecol.* **89**, 947–959, (2001)
4. Lijoi, A., R. H. Mena, and I. Prünster . Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**(4), 769–786, (2007)
5. Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlenn, D., Minchin, P.R., O'Hara, R. B., Simpson, G. L., Solymos, P. M., Stevens, H. H., Szoecs, E., and Wagner, H.: *vegan: Community Ecology Package*. R package version 2.5-7. (2020) <https://CRAN.R-project.org/package=vegan>
6. Pitman, J.: Some developments of the Blackwell-MacQueen urn scheme. In T. S. Ferguson, L. S. Shapley, and J. B. MacQueen (Eds.), *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, Volume 30 of IMS Lecture notes, Monograph Series, pp. 245–267. Hayward: Institute of Mathematical Statistics, (1996)
7. Pyke, C. R., Condit, R., Aguilar, S., and Lao, S.: Floristic composition across a climatic gradient in a neotropical lowland forest. *J. Veg. Sci.* **12**, 553–566, (2001)
8. Zito, A., Rigon T., Ovaskainen O., Dunson, D.B.: Bayesian nonparametric modelling of sequential discoveries. *arXiv:2011.06629*, (2020)

# Model fitting and Bayesian inference via power expectation propagation

## *Stima ed inferenza Bayesiana tramite power expectation propagation*

Emanuele Degani, Luca Maestrini and Mauro Bernardi

**Abstract** We study a message passing approach to power expectation propagation for Bayesian model fitting and inference. Power expectation propagation is a class of variational approximations based on the notion of  $\alpha$ -divergence that extends two notable approximations, namely mean field variational Bayes and expectation propagation. An illustration on a simple model allows to grasp benefits and complexities of this methodology and sets the basis for applications on more complex models.

**Abstract** *Studiamo l'approccio message passing al power expectation propagation per la stima e l'inferenza Bayesiana. Power expectation propagation è una classe di approssimazioni variazionali basata sulla nozione di divergenza  $\alpha$  che estende due approssimazioni notevoli, mean field variational Bayes ed expectation propagation. Un'illustrazione su un semplice modello consente di cogliere benefici e complessità di questa metodologia, ponendo le basi per applicazioni su modelli più complessi.*

**Key words:**  $\alpha$ -divergence, approximate Bayesian inference, factor graph, message passing, variational approximation.

## 1 Introduction

Bayesian inference deals with updating a prior distribution  $p(\theta)$  on a parameter vector  $\theta$  through the model likelihood  $p(\mathbf{y}|\theta)$  for the observed data  $\mathbf{y}$  to obtain the posterior distribution  $p(\theta|\mathbf{y}) = p(\mathbf{y}|\theta)p(\theta)/p(\mathbf{y})$ . Typically the marginal likelihood  $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$  cannot be evaluated explicitly and Markov chain Monte Carlo (MCMC) methods have been the main toolkit to sample from the posterior

---

Emanuele Degani, Mauro Bernardi  
Department of Statistical Sciences, University of Padua, Italy  
e-mail: degani@stat.unipd.it, e-mail: mauro.bernardi@unipd.it,

Luca Maestrini  
School of Mathematical and Physical Sciences, University of Technology Sydney, Australia  
e-mail: luca.maestrini@uts.edu.au

density for decades. Nevertheless, MCMC algorithms may suffer of slow convergence and poor mixing behaviors that can compromise inferential conclusions [4].

Variational inference methods [3, 11] take a different perspective on the problem. Instead of sampling from  $p(\theta|\mathbf{y})$ , variational approaches are used to approximate the posterior density with an approximating density  $q(\theta)$  chosen from a suitable family  $\mathcal{Q}$  of distributions. The most common Bayesian variational methods find the optimal approximating density by solving

$$q^*(\theta) = \arg \min_{q(\theta) \in \mathcal{Q}} \text{KL}(q(\theta) \| p(\theta|\mathbf{y})), \quad (1)$$

with  $\text{KL}(q(\theta) \| p(\theta|\mathbf{y}))$  denoting the Kullback–Leibler divergence between  $q$  and  $p(\cdot|\mathbf{y})$ . Practical solutions arise imposing a convenient partition  $\{\theta_1, \dots, \theta_M\}$  of  $\theta$  such that  $q(\theta) = \prod_{i=1}^M q(\theta_i)$  and employing a convex optimization scheme (see e.g. Section 10.1.1. of [2]) known as *mean field variational Bayes (MFVB)*.

Another variational inference technique, proposed in [8] and named *expectation propagation (EP)*, is built upon the optimization problem

$$q^*(\theta) = \arg \min_{q(\theta) \in \mathcal{Q}} \text{KL}(p(\theta|\mathbf{y}) \| q(\theta)), \quad (2)$$

where the arguments of the Kullback–Leibler divergence in (1) are reversed. This leads to a different class of iterative optimization schemes that [10] recasts into a *message passing* on a factor graph framework. The message passing paradigm allows for distributed and scalable fitting of variational approximations. [5] exploit the results of [10] and provide an explicit algorithm for performing EP on a simple statistical model, studying issues and challenges related to its implementation.

In this article we study a generalization of both MFVB and EP known as *power expectation propagation (Power-EP)* that was proposed by [9] to make (2) more tractable. This method yields a class of appealing message passing algorithms and we explore their use for statistical model fitting. Section 2 describes Power-EP and introduces a message passing technique to solve the optimization problem on models with factor graph representations. Section 3 provides explicit illustration on a simple model and Section 4 investigates the quality of the variational approximation via a simulation study. Final considerations and further developments are described in Section 5.

## 2 Power-EP and message passing

Power-EP solves the following optimization problem:

$$q_\alpha^*(\theta) = \arg \min_{q_\alpha(\theta) \in \mathcal{Q}} D_\alpha(p(\theta|\mathbf{y}) \| q_\alpha(\theta)), \quad \alpha \in (-\infty, \infty) \setminus \{0\},$$

where  $D_\alpha(p(\theta|\mathbf{y}) \| q_\alpha(\theta)) \equiv (\alpha(1-\alpha))^{-1} \{1 - \int_{\Theta} p(\theta|\mathbf{y})^\alpha q_\alpha(\theta)^{1-\alpha} d\theta\}$  is the  $\alpha$ -divergence of Amari [1]. It possesses two notable limiting cases:

$$D_\alpha(p(\theta|\mathbf{y}) \| q(\theta)) \xrightarrow{\alpha \rightarrow 0} \text{KL}(q(\theta) \| p(\theta|\mathbf{y})) \quad \text{and} \quad D_1(p(\theta|\mathbf{y}) \| q(\theta)) = \text{KL}(p(\theta|\mathbf{y}) \| q(\theta)),$$

meaning that Power-EP reduces to MFVB and EP for  $\alpha \rightarrow 0$  and  $\alpha = 1$ , respectively. Hence, the quality of Power-EP approximations varies with  $\alpha$ , and for certain  $\alpha$  values the approximations may outperform those obtained with MFVB and EP. We restrict our attention to approximations arising from  $\alpha \in (0, 1]$ , that is to the class of approximations that has MFVB and EP as extreme and opposite cases.

[10] provides an approximate solution to the minimization in (2) based on message passing on factor graphs, for a given  $\alpha$ . We employ this strategy and describe a message passing procedure for fitting models having a factor graph representation via Power-EP (see e.g. [6, §2.3] for a primer on factor graphs).

Consider a model whose joint density function can be factorized into  $N$  different factors  $p(\boldsymbol{\theta}, \mathbf{y}) = \prod_{j=1}^N f_j(\boldsymbol{\theta}_{\text{neigh}(j)})$ , with  $\text{neigh}(j) \equiv \{1 \leq i \leq M : \theta_i \text{ is a neighbor of } f_j\}$ . Introduce an approximating density to the posterior distribution  $q_\alpha(\boldsymbol{\theta})$  that can be written as  $q_\alpha(\boldsymbol{\theta}) = \prod_{i=1}^M q_\alpha(\theta_i)$ . Using a Power-EP approach, each density  $q_\alpha(\theta_i)$  can be obtained as the product of *messages* reaching  $\theta_i$  from the neighboring factors. For each  $1 \leq i \leq M$  and  $1 \leq j \leq N$ , the Power-EP *factor to stochastic node message* updates are given by

$$m_{f_j \rightarrow \theta_i}^{(\alpha)}(\theta_i) \leftarrow \text{proj} \left\{ Z^{-1} [m_{f_j \rightarrow \theta_i}^{(\alpha)}(\theta_i)]^{1-\alpha} m_{\theta_i \rightarrow f_j}^{(\alpha)}(\theta_i) \int [f_j(\boldsymbol{\theta}_{\text{neigh}(j)})]^\alpha \right. \\ \left. \times \prod_{i' \in \text{neigh}(j)/\{i\}} [m_{f_j \rightarrow \theta_{i'}}^{(\alpha)}(\theta_{i'})]^{1-\alpha} m_{\theta_{i'} \rightarrow f_j}^{(\alpha)}(\theta_{i'}) d\boldsymbol{\theta}_{\text{neigh}(j)/\{i\}} \right\} / m_{\theta_i \rightarrow f_j}^{(\alpha)}(\theta_i), \quad (3)$$

where the  $\leftarrow$  symbol means that the function of  $\theta_i$  on the left-hand side is updated according to the expression on the right-hand side,  $\text{proj}\{p\}$  is the operator that projects the density function  $p$  onto an appropriate exponential family (see [5, §2.3]) and  $Z$  is the normalizing constant of  $p$ . For each  $1 \leq i \leq M$  and  $1 \leq j \leq N$ , the Power-EP *stochastic node to factor message* updates have form

$$m_{\theta_i \rightarrow f_j}^{(\alpha)}(\theta_i) \leftarrow \prod_{j' \neq j : i \in \text{neigh}(j')} m_{f_{j'} \rightarrow \theta_i}^{(\alpha)}(\theta_i). \quad (4)$$

Optimization can be performed by iteratively updating the factor graph messages via (3) and (4) upon convergence. Convergence can be assessed by monitoring the  $\alpha$ -approximate marginal log-likelihood defined as

$$\log \tilde{p}(\mathbf{y}; q_\alpha) \equiv \sum_{i=1}^M \log s_{\theta_i}^{(\alpha)} + \frac{1}{\alpha} \sum_{j=1}^N \log s_{f_j}^{(\alpha)}, \quad \text{with } s_{\theta_i}^{(\alpha)} \equiv \int \prod_{j: i \in \text{neigh}(j)} m_{f_j \rightarrow \theta_i}^{(\alpha)}(\theta_i) d\theta_i \\ \text{and } s_{f_j}^{(\alpha)} \equiv \frac{\int (f_j(\boldsymbol{\theta}_{\text{neigh}(j)}))^\alpha \prod_{i \in \text{neigh}(j)} m_{\theta_i \rightarrow f_j}^{(\alpha)}(\theta_i) (m_{f_j \rightarrow \theta_i}^{(\alpha)}(\theta_i))^{1-\alpha} d\boldsymbol{\theta}_{\text{neigh}(j)}}{\int \prod_{i \in \text{neigh}(j)} m_{\theta_i \rightarrow f_j}^{(\alpha)}(\theta_i) m_{f_j \rightarrow \theta_i}^{(\alpha)}(\theta_i) d\boldsymbol{\theta}_{\text{neigh}(j)}}. \quad (5)$$

At convergence, the optimal approximating densities can be obtained from

$$q_\alpha^*(\theta_i) \propto \prod_{j: i \in \text{neigh}(j)} m_{f_j \rightarrow \theta_i}^{(\alpha)}(\theta_i) = m_{\theta_i \rightarrow f_j}^{(\alpha)}(\theta_i) m_{f_j \rightarrow \theta_i}^{(\alpha)}(\theta_i). \quad (6)$$

It is worth noting that when  $\alpha = 1$ , the resulting  $q_1^*(\theta)$  approximation matches the one from EP. Consequently, expressions (3.5)–(3.11) of [5], and results of [6] can be immediately retrieved fixing  $\alpha = 1$  in expressions (3)–(6).

### 3 Simple illustrative example

The general expressions of Section 2 providing a message passing solution to Power-EP are anything but intuitive and the computational steps behind (3)–(6) are difficult to glean. Therefore, we make explicit illustration on the simple Bayesian Normal random sample model studied in [5]. The model we consider is:

$$y_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \sim \mathcal{N}(\mu_\mu, \sigma_\mu^2), \sigma^2 | a \sim \text{Inv-Gamma}\left(\frac{1}{2}, \frac{1}{a}\right), a \sim \text{Inv-Gamma}\left(\frac{1}{2}, \frac{1}{A^2}\right), \quad (7)$$

for  $1 \leq i \leq n$ , where  $\mu_\mu \in \mathbb{R}$ ,  $\sigma_\mu > 0$  and  $A > 0$  are fixed hyperparameters, and the hierarchical specification on  $\sigma^2$  is such that  $\sigma \sim \text{Half-Cauchy}(A)$ . The joint density function then factorizes as  $p(\mathbf{y}, \mu, \sigma^2, a) = p(\mathbf{y} | \mu, \sigma^2) p(\mu) p(\sigma^2 | a) p(a)$ .

Consider the approximation  $q_\alpha(\mu, \sigma^2, a) = q_\alpha(\mu) q_\alpha(\sigma^2) q_\alpha(a)$  to the posterior density. Application of (3) and enforcement of conjugacy constraints give rise to the following expressions for the Power-EP factor to stochastic node messages:

$$\begin{aligned} m_{p(\mathbf{y} | \mu, \sigma^2) \rightarrow \mu}^{(\alpha)}(\mu) &\propto \exp\left(\left[\begin{array}{c} \mu \\ \mu^2 \end{array}\right]^T \eta_{p(\mathbf{y} | \mu, \sigma^2) \rightarrow \mu}^{(\alpha)}\right), \quad m_{p(\sigma^2 | a) \rightarrow a}^{(\alpha)}(a) \propto \exp\left(\left[\begin{array}{c} \log a \\ 1/a \end{array}\right]^T \eta_{p(\sigma^2 | a) \rightarrow a}^{(\alpha)}\right), \\ m_{p(\mathbf{y} | \mu, \sigma^2) \rightarrow \sigma^2}^{(\alpha)}(\sigma^2) &\propto \exp\left(\left[\begin{array}{c} \log \sigma^2 \\ 1/\sigma^2 \end{array}\right]^T \eta_{p(\mathbf{y} | \mu, \sigma^2) \rightarrow \sigma^2}^{(\alpha)}\right), \quad m_{p(\mu) \rightarrow \mu}^{(\alpha)}(\mu) \propto \exp\left(\left[\begin{array}{c} \mu \\ \mu^2 \end{array}\right]^T \eta_{p(\mu) \rightarrow \mu}^{(\alpha)}\right), \\ m_{p(\sigma^2 | a) \rightarrow \sigma^2}^{(\alpha)}(\sigma^2) &\propto \exp\left(\left[\begin{array}{c} \log \sigma^2 \\ 1/\sigma^2 \end{array}\right]^T \eta_{p(\sigma^2 | a) \rightarrow \sigma^2}^{(\alpha)}\right), \quad m_{p(a) \rightarrow a}^{(\alpha)}(a) \propto \exp\left(\left[\begin{array}{c} \log a \\ 1/a \end{array}\right]^T \eta_{p(a) \rightarrow a}^{(\alpha)}\right). \end{aligned}$$

Here the symbol  $\eta$  denotes natural parameter vectors of exponential families. Straightforward application of (4) leads to similar and conjugate expressions for the Power-EP stochastic node to factor messages. Application of (6) leads to the optimal approximating densities for the parameters of interest  $q_\alpha^*(\mu)$  and  $q_\alpha^*(\sigma^2)$ :

$$q_\alpha^*(\mu) \propto \exp\left(\left[\begin{array}{c} \mu \\ \mu^2 \end{array}\right]^T \eta_{q_\alpha^*(\mu)}\right) \quad \text{and} \quad q_\alpha^*(\sigma^2) \propto \exp\left(\left[\begin{array}{c} \log \sigma^2 \\ 1/\sigma^2 \end{array}\right]^T \eta_{q_\alpha^*(\sigma^2)}\right), \quad (8)$$

which correspond to a  $\mathcal{N}(-[\eta_{q_\alpha^*(\mu)}]_1 / (2[\eta_{q_\alpha^*(\mu)}]_2), -1 / (2[\eta_{q_\alpha^*(\mu)}]_2))$  density function for  $\mu$  and an  $\text{Inv-Gamma}(-[\eta_{q_\alpha^*(\sigma^2)}]_1 - 1, -[\eta_{q_\alpha^*(\sigma^2)}]_2)$  density function for  $\sigma^2$ , respectively, with  $\eta_{q_\alpha^*(\mu)}$  and  $\eta_{q_\alpha^*(\sigma^2)}$  vectors of length 2.

Given that the resulting Power-EP messages belong to exponential families, their updates can be performed just by updating their  $\eta$  natural parameter vectors. Derivations of these updates and the explicit expression of  $\log \tilde{p}(\mathbf{y}; q_\alpha)$  from (5) are not provided here for brevity, but follow steps similar to those in [5, §A.5].

Algorithm 1 lists the iterative natural parameter updates for fitting the Bayesian random sample model via Power-EP message passing. Within the expressions of the natural parameter updates,  $\eta_{f_j \leftrightarrow \theta_i}^{(\alpha)} \equiv (1 - \alpha)\eta_{f_j \rightarrow \theta_i}^{(\alpha)} + \eta_{\theta_i \rightarrow f_j}^{(\alpha)}$  for meaningful com-

binations of  $f_j \in \{p(\boldsymbol{\mu}), p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2), p(\boldsymbol{\sigma}^2|a), p(a)\}$  and  $\boldsymbol{\theta}_i \in \{\boldsymbol{\mu}, \boldsymbol{\sigma}^2, a\}$ . Functions  $G^N(\cdot)$ ,  $G^{IG1}(\cdot)$  and  $G^{IG2}(\cdot)$  are defined in [5, §A.4] and involve quadrature methods for evaluating non-analytic functions that are described in [5, §2.1].

---

**Algorithm 1** *Power-Expectation Propagation message passing algorithm for determining the parameter of the optimal density functions  $q_\alpha^*(\boldsymbol{\mu})$  and  $q_\alpha^*(\boldsymbol{\sigma}^2)$  of interest for approximate Bayesian inference on the Normal random sample model (7).*

---

Input:  $\mathbf{y} = (y_1, \dots, y_n)^T$ ,  $\boldsymbol{\mu}_\mu$ ,  $\boldsymbol{\sigma}_\mu > 0$  and  $A > 0$ . Create:  $\mathbf{c} = (n, \sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2)^T$ .

Select: Power-EP factor  $\alpha \in (0, 1]$ .

$$\text{Initialize: } \eta_{p(\boldsymbol{\mu}) \rightarrow \boldsymbol{\mu}}^{(\alpha)} \leftarrow \begin{bmatrix} \boldsymbol{\mu}_\mu / \boldsymbol{\sigma}_\mu^2 \\ -1 / (2\boldsymbol{\sigma}_\mu^2) \end{bmatrix}, \quad \eta_{p(a) \rightarrow a}^{(\alpha)} \leftarrow \begin{bmatrix} -3/2 \\ -1/A^2 \end{bmatrix}, \quad \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\mu}}^{(\alpha)} \leftarrow \begin{bmatrix} 0 \\ -1/2 \end{bmatrix},$$

$$\eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \quad \eta_{p(\boldsymbol{\sigma}^2|a) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \quad \eta_{p(\boldsymbol{\sigma}^2|a) \rightarrow a}^{(\alpha)} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix},$$

$$\eta_{\boldsymbol{\mu} \rightarrow p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)}^{(\alpha)} \leftarrow \eta_{p(\boldsymbol{\mu}) \rightarrow \boldsymbol{\mu}}^{(\alpha)}, \quad \eta_{a \rightarrow p(\boldsymbol{\sigma}^2|a)}^{(\alpha)} \leftarrow \eta_{p(a) \rightarrow a}^{(\alpha)}.$$

Cycle until the relative change in  $\log \tilde{p}(\mathbf{y}; q_\alpha)$  is negligible:

$$\eta_{\boldsymbol{\sigma}^2 \rightarrow p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)}^{(\alpha)} \leftarrow \eta_{p(\boldsymbol{\sigma}^2|a) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)},$$

$$\eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\mu}}^{(\alpha)} \leftarrow G^N \left( \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \leftrightarrow \boldsymbol{\mu}}^{(\alpha)}, \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \leftrightarrow \boldsymbol{\sigma}^2}^{(\alpha)}; \boldsymbol{\alpha} \mathbf{c} \right) + (1 - \alpha) \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\mu}}^{(\alpha)},$$

$$\eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)} \leftarrow G^{IG1} \left( \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \leftrightarrow \boldsymbol{\sigma}^2}^{(\alpha)}, \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \leftrightarrow \boldsymbol{\mu}}^{(\alpha)}; \boldsymbol{\alpha} \mathbf{c} \right) + (1 - \alpha) \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)},$$

$$\eta_{\boldsymbol{\sigma}^2 \rightarrow p(\boldsymbol{\sigma}^2|a)}^{(\alpha)} \leftarrow \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)},$$

$$\eta_{p(\boldsymbol{\sigma}^2|a) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)} \leftarrow G^{IG2} \left( \eta_{p(\boldsymbol{\sigma}^2|a) \leftrightarrow \boldsymbol{\sigma}^2}^{(\alpha)}, \begin{bmatrix} [\eta_{p(\boldsymbol{\sigma}^2|a) \leftrightarrow a}^{(\alpha)}]_1 + 2(1 - \alpha) \\ [\eta_{p(\boldsymbol{\sigma}^2|a) \leftrightarrow a}^{(\alpha)}]_2 / \alpha \end{bmatrix}; 3\boldsymbol{\alpha} \right) + (1 - \alpha) \eta_{p(\boldsymbol{\sigma}^2|a) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)},$$

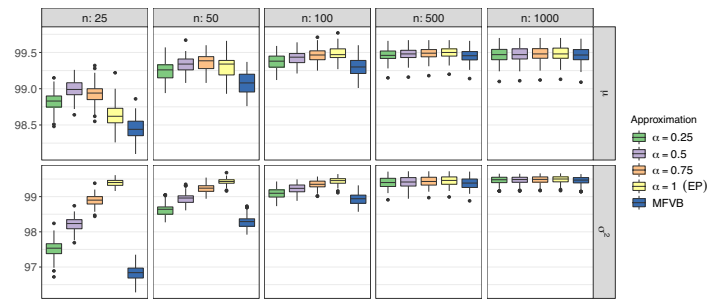
$$\eta_{p(\boldsymbol{\sigma}^2|a) \rightarrow a}^{(\alpha)} \leftarrow G^{IG2} \left( \eta_{p(\boldsymbol{\sigma}^2|a) \leftrightarrow a}^{(\alpha)}, \begin{bmatrix} [\eta_{p(\boldsymbol{\sigma}^2|a) \leftrightarrow \boldsymbol{\sigma}^2}^{(\alpha)}]_1 + 2(1 - \alpha) \\ [\eta_{p(\boldsymbol{\sigma}^2|a) \leftrightarrow \boldsymbol{\sigma}^2}^{(\alpha)}]_2 / \alpha \end{bmatrix}; \boldsymbol{\alpha} \right) + (1 - \alpha) \eta_{p(\boldsymbol{\sigma}^2|a) \rightarrow a}^{(\alpha)}.$$

Output for (8):  $\eta_{q_\alpha^*(\boldsymbol{\mu})}^{(\alpha)} = \eta_{p(\boldsymbol{\mu}) \rightarrow \boldsymbol{\mu}}^{(\alpha)} + \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\mu}}^{(\alpha)}$ ,  $\eta_{q_\alpha^*(\boldsymbol{\sigma}^2)}^{(\alpha)} = \eta_{p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)} + \eta_{p(\boldsymbol{\sigma}^2|a) \rightarrow \boldsymbol{\sigma}^2}^{(\alpha)}$ .

---

## 4 Simulation study

We assess the performances of Power-EP for fitting model (7) through a simulation study. For each sample size  $n \in \{25, 50, 100, 500, 1000\}$ , we generate 100 random samples from the  $N(0, 1)$  distribution and obtain the optimal Power-EP approximating densities of interest  $q_\alpha^*(\boldsymbol{\mu})$  and  $q_\alpha^*(\boldsymbol{\sigma}^2)$  for  $\alpha \in \{0.25, 0.5, 0.75, 1\}$  via Algorithm 1, and MFVB approximations using Algorithm 1 of [7]. We set diffuse priors with hyperparameters  $\boldsymbol{\mu}_\mu = 0$  and  $\boldsymbol{\sigma}_\mu = A = 10^5$ . For each replicate, we evaluate the quality of the approximation computing, for  $\boldsymbol{\theta} = \boldsymbol{\mu}, \boldsymbol{\sigma}^2$ ,  $\text{accuracy}\{q_\alpha^*(\boldsymbol{\theta})\} \equiv 100(1 - 0.5 \int |q_\alpha^*(\boldsymbol{\theta}) - p(\boldsymbol{\theta}|\mathbf{y})| d\boldsymbol{\theta})$ . The ‘true’ marginal posterior densities are obtained via kernel density estimation applied to MCMC samples obtained with the `rstan` library [12], after excluding an appropriate burn-in sample. Figure 1 summarizes the results and compares the approximations. For small sample sizes, Power-EP approximations with  $\alpha = 0.25, 0.5, 0.75$  overperform both EP and MFVB in terms of accuracy for  $\boldsymbol{\mu}$ , whereas EP provides a better approxi-



**Fig. 1** Accuracy values of the approximating  $q^*$ 's for  $\mu$  and  $\sigma^2$ , at different sample sizes.

mation for  $\sigma^2$ . As  $n$  increases, the accuracy of the approximations becomes more uniform for both  $\mu$  and  $\sigma^2$ .

## 5 Conclusions and further developments

We studied Power-EP as a message passing approach for fitting models that have a factor graph representation through the minimization of the  $\alpha$ -divergence between the posterior and an approximating density. Power-EP includes the more common MFVB and EP approximations, which can be outperformed by approximations based on appropriate choice of  $\alpha$ , especially when the number of observations is limited. Implementation of Power-EP for a wide set of  $\alpha$  values comes with a higher computational cost, that could be reduced applying optimization strategies based on automatic differentiation. Further directions include the exploration of methods for automatic selection of  $\alpha$  values that produce better approximations and application to more complex statistical models.

## References

1. Amari, S.: Differential-Geometrical Methods in Statistics. Springer, New York (1985)
2. Bishop, C. M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
3. Blei, M. D., Kucukelbir, A., McAuliffe, J. D.: Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017)
4. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B.: Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC, (2013)
5. Kim, A. S. I., Wand, M. P.: The Explicit Form of Expectation Propagation for a Simple Statistical Model. *Electron. J. Stat.* **10**, 550–581 (2016)
6. Kim, A. S. I., Wand, M. P.: On Expectation Propagation for Generalised, Linear and Mixed Models. *Aust. N. Z. J. Stat.* **60**, 75–102 (2018)
7. Luts, J., Broderick, T., Wand, M. P.: Real-time semiparametric regression. *J. Comput. Graph. Stat.* **23**, 589–615 (2014)
8. Minka, T. P.: Expectation propagation for approximate Bayesian inference. *Proc. of the XVII Conf. on Uncert. in Art. Intel.*, Morgan Kaufmann Publishers Inc., 362–239 (2001)
9. Minka, T. P.: Power EP. *Micr. Res. Tech. Rep.* **149** (2004)
10. Minka, T. P.: Divergence measures and message passing. *Micr. Res. Tech. Rep.* **173** (2005)
11. Ormerod, J. T., Wand, M. P.: Explaining variational approximations. *Am. Stat.* **64**, 140–153 (2010)
12. Stan Development Team.: *rstan* 2.21.2: the R interface to Stan. <http://mc-stan.org/> (2020)

# Bayesian quantile estimation in deconvolution

## *Stima bayesiana di quantili in problemi di deconvoluzione*

Catia Scricciolo

**Abstract** Estimating quantiles of a population is a fundamental problem in non-parametric statistics, with high practical relevance. This note deals, from a Bayesian point of view, with quantile estimation in deconvolution problems with known error distribution. We pursue the analysis for error distributions whose characteristic functions decay polynomially fast, the so-called ordinary smooth errors that lead to mildly ill-posed inverse problems. Our method is based on Fourier inversion techniques for density deconvolution and the estimation procedure for single quantiles is minimax-optimal under minimal conditions.

**Abstract** La stima dei quantili di una popolazione è un problema fondamentale nella statistica non parametrica, di grande rilevanza pratica. Questa nota si occupa, da un punto di vista bayesiano, della stima di quantili in problemi di deconvoluzione con distribuzione nota dell'errore. L'analisi viene condotta per distribuzioni dell'errore aventi funzione caratteristica che decade in modo polinomiale. Il metodo adottato si basa su tecniche d'inversione di Fourier per la deconvoluzione di densità e la procedura di stima per singoli quantili risulta ottimale in senso minimax sotto condizioni minimali.

**Key words:** Bayesian Quantile Estimation, Deconvolution, Ordinary Smooth Error

## 1 Introduction

Quantile estimation is a fundamental problem in nonparametric statistics from both the methodological and practical point of view. Estimated quantiles are relevant in applications, however, since quantiles depend nonlinearly on the underlying distribution, it is not always clear how to estimate them, even more in deconvolution

---

Catia Scricciolo  
Università degli Studi di Verona, Via Cantarane, 24, 37129 Verona,  
e-mail: [catia.scricciolo@univr.it](mailto:catia.scricciolo@univr.it)



problems, see, *e.g.*, § 1.1.2 in [9], pp. 13–14, and the monograph by [12], where observations are affected by additive measurement errors that should be taken into account, otherwise quantile estimates based on the observed measurements would be biased. For example, since high blood pressure can cause cardiovascular diseases, it is important to determine reference values, in particular, percentiles of systolic and diastolic blood pressure by features like age, sex etc. The observer is aware that blood pressure is measured with some error due to the lack of precision of the measurement device, which forces to consider indirect observations with implicit measurement errors instead of outcomes of the quantity of interest.

We describe the problem more formally. Suppose that we observe independent and identically distributed (iid) random variables  $Y_1, \dots, Y_n$  such that

$$Y_i = X_i + \varepsilon_i, \quad i = 1, \dots, n,$$

that is,  $Y_i$  is the signal  $X_i$  additively corrupted by the measurement error  $\varepsilon_i$ , which is independent of  $X_i$  and has density  $f_\varepsilon$ . If also  $X$  has density, say  $f_X$ , then  $f_Y = f_X * f_\varepsilon$ . We assume that the error density is completely known and its Fourier transform  $\hat{f}_\varepsilon(t) := \int_{\mathbb{R}} e^{it} f_\varepsilon(u) du, t \in \mathbb{R}$ , verifies the following condition: for some  $\beta > 0$  there exists a constant  $R > 0$  such that

$$|\hat{f}_\varepsilon(t)|^{-1} \leq R(1 + |t|)^\beta \quad \text{and} \quad |\hat{f}_\varepsilon^{(1)}(t)| \leq R(1 + |t|)^{-(\beta+1)}, \quad t \in \mathbb{R}. \quad (1)$$

Condition (1), which implies that  $\hat{f}_\varepsilon$  decays polynomially fast, characterizes the so-called ordinary smooth errors. For  $\tau \in (0, 1)$ , let

$$q^\tau = Q(\tau) \equiv F_X^{-1}(\tau) := \inf\{x : F_X(x) \geq \tau\}$$

be the  $\tau$ -quantile of the population  $X$ . The problem is to estimate  $q^\tau$  from indirect observations  $Y_1, \dots, Y_n$ . Quantile estimation in deconvolution with measurement error distribution satisfying condition (1) leads to nonlinear functional estimation in a mildly ill-posed inverse problem.

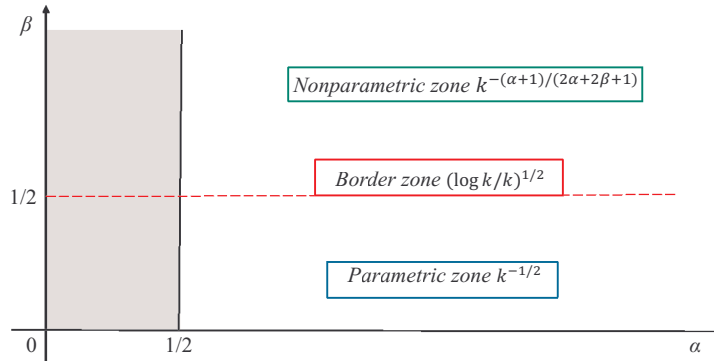
The problem of quantile estimation in deconvolution for the case of known error distribution has been studied by [10], while the more realistic situation where also the error distribution is unknown and has to be estimated from a sample  $\varepsilon_1^*, \dots, \varepsilon_m^*$  has only recently been investigated by [5]. The former authors proposed a quantile estimator obtained by inverting a distribution function estimator constructed using a direct inversion formula instead of integrating the canonical density deconvolution estimator as in [6], which resulted in a non-optimal (in the minimax sense) analysis of the method. Dattner *et al.* [5], instead, used a plug-in method for distribution function estimation based on a deconvolution density estimator which leads to a minimax optimal procedure under a local  $\alpha$ -Hölder condition on  $f_X$  for  $\alpha \geq 1/2$ , with rates

$$\psi_k(\alpha, \beta) := \begin{cases} k^{-1/2}, & \text{for } \beta < 1/2, \\ (\log k/k)^{1/2}, & \text{for } \beta = 1/2, \\ k^{-(\alpha+1)/(2\alpha+2\beta+1)}, & \text{for } \beta > 1/2, \end{cases} \quad \text{where } k := (n \wedge m),$$

that differ, for  $\beta = 1/2$ , only by a logarithmic factor from the lower bound

$$k^{-(\alpha+1)/[2\alpha+(2\beta\vee 1)+1]}. \tag{2}$$

The existence of different rate *régimes* for  $\beta < 1/2$ ,  $\beta = 1/2$  and  $\beta > 1/2$  was already pointed out in Theorem 3.2 by [10] and in Theorem 2.1 by [4] for estimating the cumulative distribution function  $F_X$ , see Fig. 1. The same distinction also holds when the error distribution is known, with  $\psi_n(\alpha, \beta)$  for all  $\alpha, \beta > 0$ .



**Fig. 1** Partition of the *Regular zone*  $\mathcal{L} := \{(\alpha, \beta) : \alpha \geq 1/2, \beta > 0\}$ : *Parametric zone*  $\{(\alpha, \beta) \in \mathcal{L} : \beta < 1/2\}$ , *Border zone*  $\{(\alpha, \beta) \in \mathcal{L} : \beta = 1/2\}$ , *Nonparametric zone*  $\{(\alpha, \beta) \in \mathcal{L} : \beta > 1/2\}$ .

In this note, we consider the inverse problem of estimating quantiles of the distribution of  $X$  from the Bayesian point of view. Some results on Bayesian nonparametric quantile estimation in the direct problem, based on a Dirichlet process prior law for the population distribution, were already present in Ferguson’s seminal paper [7]. The limiting distribution of the posterior quantile process has been derived by [3], who showed that the posterior law of the rescaled and recentered quantile function converges weakly to a Gaussian process, as the sample size increases. Confidence bands for the quantile function are constructed based on bootstrap approximations of the posterior quantile process. Also the paper by [11] develops and discusses methods for carrying out nonparametric Bayesian inference on the quantile function based on a Dirichlet process prior. The limiting distribution of the quantile process corresponding to a normalized inverse-Gaussian process has been given in [2], see also [1] for the study of the limiting distribution of the quantile process based on prior laws belonging to a general class of popular Bayesian nonparametric priors.

We are aware of no results on Bayesian nonparametric inference for the quantile function in deconvolution problems with known or unknown error distribution. In Section 2, we give sufficient conditions on the prior law and the true data generating process so that the posterior leads to an optimal (in the minimax sense) quantile estimation procedure when the error distribution is known and ordinary smooth.

## 2 Posterior rates for quantile estimation in deconvolution

Let  $Y^{(n)} := (Y_1, \dots, Y_n)$  be a sample of  $n$  iid observations drawn from the true data generating process  $P_0 \equiv P_{0Y}$ , with density  $f_{0Y} = f_{0X} * f_\varepsilon$ . Given  $\tau \in (0, 1)$ , let  $q_{0X}^\tau$  be the  $\tau$ -quantile of  $F_{0X}$ . We want to estimate  $q_{0X}^\tau$  taking a Bayesian nonparametric approach. Let  $\Pi_n$  be a prior law on the set  $\mathcal{F}$  of Lebesgue absolutely continuous probability measures on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  and let  $\Pi_n(\cdot | Y^{(n)})$  be the resulting posterior

$$\Pi_n(B | Y^{(n)}) = \frac{\int_B \prod_{i=1}^n f_Y(Y_i) \Pi(d f_X)}{\int_{\mathcal{F}} \prod_{i=1}^n f_Y(Y_i) \Pi(d f_X)}, \quad B \in \mathcal{B}(\mathbb{R}),$$

where  $f_Y = f_X * f_\varepsilon$ . Our goal is to assess the posterior contraction rate  $\gamma_n = o(1)$  such that

$$\Pi_n(|q_X^\tau - q_{0X}^\tau| \leq \gamma_n | Y^{(n)}) = 1 + o_{P_0}(1),$$

where  $q_X^\tau$  is the  $\tau$ -quantile of  $F_X$ . We give sufficient conditions on  $\Pi_n$  and  $f_{0X}$  so that the posterior distribution leads to a minimax-optimal estimation procedure.

We introduce some notation. Let  $\langle \alpha \rangle$  be the largest integer strictly smaller than  $\alpha > 0$ . For any interval  $I \subseteq \mathbb{R}$  and function  $g$  on  $I$ , the Hölder norm is

$$\|g\|_{C^\alpha(I)} := \sum_{k=0}^{\langle \alpha \rangle} \|g^{(k)}\|_{L^\infty(I)} + \sup_{x, y \in I: x \neq y} \frac{|g^{(\alpha)}(x) - g^{(\alpha)}(y)|}{|x - y|^{\alpha - \langle \alpha \rangle}}.$$

Let  $C_B(I)$  be the set of continuous and bounded functions on  $I$  and  $C^\alpha(I, R) = \{g \in C_B(I) : \|g\|_{C^\alpha(I)} \leq R\}$  the set of continuous and bounded functions on  $I$  with Hölder norm uniformly bounded by  $R > 0$ . For  $\delta > 0$ , let

$$B_{\text{KL}}(P_0; \delta^2) := \left\{ P \in \mathcal{F} : P_0 \left( \log \frac{P_0}{P} \right) \leq \delta^2, P_0 \left( \log \frac{P_0}{P} \right)^2 \leq \delta^2 \right\}$$

be a *Kullback-Leibler* type neighborhood of  $P_0$  of radius  $\delta^2$ , where  $P_0 f$  stands for the expected value  $\int f dP_0$ .

**Proposition 1.** *Let  $\mathbb{E}[|\varepsilon|] < \infty$  and  $\hat{f}_\varepsilon$  verify condition (1) for  $\beta \geq 1$ . For  $\alpha, R, r, \zeta > 0$ , let  $f_{0X}(\cdot - q_{0X}^\tau) \in C^\alpha([-\zeta, \zeta], R)$  and*

$$\inf_{x \in [-\zeta, \zeta]} f_{0X}(x - q_{0X}^\tau) \geq r. \tag{3}$$

*Let the prior law  $\Pi_n$  assign probability one to the set of  $f_X(\cdot - q_{0X}^\tau) \in C^\alpha([-\zeta, \zeta], R)$ . If for  $\eta_n = o(1)$  such that  $n\eta_n^2 \rightarrow \infty$  and constants  $C_1, C_2 > 0$ , we have*

$$\Pi_n(B_{\text{KL}}(P_0; \eta_n^2)) \geq \exp(-C_1 n \eta_n^2) \tag{4}$$

and

$$\Pi_n(\|f_X - f_{0X}\|_\infty \leq C_2 \eta_n | Y^{(n)}) = 1 + o_{P_0}(1), \tag{5}$$

then, for  $M$  large enough,

$$\Pi_n(|q_X^\tau - q_{0X}^\tau| > M(\eta_n \log n)^{(\alpha+1)/(\alpha+\beta)} | Y^{(n)}) = o_{P_0}(1).$$

*Proof.* We give a sketchy proof. Note that, while  $q_{0X}^\tau$  is fixed, the  $\tau$ -quantile  $q_X^\tau$  of  $F_X$  is *random*. Without loss of generality, we suppose that  $q_X^\tau > q_{0X}^\tau$ . Since  $F_X$  has density  $f_X$ , there exists a (random) point  $q_*^\tau$  between  $q_{0X}^\tau$  and  $q_X^\tau$  such that  $F_X(q_X^\tau) - F_X(q_{0X}^\tau) = (q_X^\tau - q_{0X}^\tau)f_X(q_*^\tau)$ . Then,

$$\begin{aligned} 0 = \tau - \tau &= F_X(q_X^\tau) - F_{0X}(q_{0X}^\tau) = \int_{q_{0X}^\tau}^{q_X^\tau} f_X(x) \, dx + \int_{-\infty}^{q_{0X}^\tau} [f_X(x) - f_{0X}(x)] \, dx \\ &= (q_X^\tau - q_{0X}^\tau)f_X(q_*^\tau) + \underbrace{[F_X(q_{0X}^\tau) - F_{0X}(q_{0X}^\tau)]}_{=:\Delta}. \end{aligned}$$

If  $f_X(q_*^\tau)$  is bounded away from zero, then

$$|q_X^\tau - q_{0X}^\tau| = \frac{|F_{0X}(q_{0X}^\tau) - F_X(q_{0X}^\tau)|}{f_X(q_*^\tau)} = \frac{|\Delta|}{f_X(q_*^\tau)}.$$

For a constant  $0 < \eta < r$  not depending on  $f_X$  nor  $q_*^\tau$  and sufficiently large  $n$  so that  $\eta_n < \eta$ , by convergence in (5), we have

$$\eta \geq \|f_X - f_{0X}\|_\infty \geq |f_X(x) - f_{0X}(x)| \quad \text{for every } x \in [q_{0X}^\tau - \zeta, q_{0X}^\tau + \zeta].$$

Since the interval  $[q_{0X}^\tau - \zeta, q_{0X}^\tau + \zeta]$  eventually includes both points  $q_X^\tau$  and  $q_*^\tau$ , it follows that  $f_X(q_*^\tau) > f_{0X}(q_*^\tau) - \eta \geq \inf_{x \in [q_{0X}^\tau - \zeta, q_{0X}^\tau + \zeta]} f_{0X}(x) - \eta \geq r - \eta > 0$ .

Let  $K$  be an  $(\langle \alpha \rangle + 1)$ -order kernel as in Assumption A of [5], p. 3, with bandwidth  $b > 0$ . We have that

$$\begin{aligned} |\Delta| &= \left| \int_{-\infty}^{q_{0X}^\tau} [f_X - f_{0X}](x) \, dx \right| \\ &\leq \left| \int_{-\infty}^{q_{0X}^\tau} [K_b * f_{0X} - f_{0X}](x) \, dx \right| + \left| \int_{-\infty}^{q_{0X}^\tau} [K_b * (f_X - f_{0X})](x) \, dx \right| \quad (6) \\ &\quad + \left| \int_{-\infty}^{q_{0X}^\tau} [K_b * f_X - f_X](x) \, dx \right| \\ &=: I + II + III. \end{aligned}$$

In order to bound term *III*, recall that, by assumption,  $f_X(\cdot - q_{0X}^\tau) \in C^\alpha([- \zeta, \zeta], \mathbb{R})$ . Then, by Lemma 5.2 in [5], both terms *I* and *III* are  $O(b^{\alpha+1})$ . It can be shown that, for  $b$  small enough,  $II = O(\|F_X - F_{0X}\|_\infty) + O(b^{-(\beta-1)}\|f_X - f_{0X}\|_\infty \log(1/b))$ . Combining previous bounds on *I*, *II* and *III*, we have

$$|\Delta| = O(b^{\alpha+1}) + O(\|F_X - F_{0X}\|_\infty) + O(b^{-(\beta-1)}\|f_X - f_{0X}\|_\infty \log(1/b)). \quad (7)$$

For  $\beta \geq 1$ , under assumption (4), a modification of Lemma 2 in [13] yields that  $\Pi_n(\|F_X - F_{0X}\|_\infty \leq K\eta_n \mid Y^{(n)}) = 1 + o_{P_0}(1)$ . The assertion follows by combining this fact with assumption (5).  $\square$

### *Discussion and final remarks*

The lower bound on  $\hat{f}_\varepsilon$  and the upper bound on  $\hat{f}_\varepsilon^{(1)}$  in condition (1) are standard assumptions in deconvolution problems and are used to derive an upper bound on the quantile estimation error, see relationship (7). The assumption on local Hölder smoothness of  $f_{0X}$  at  $q_{0X}^\tau$  allows to control the bias term  $I$  in (6). Since the quantile function is estimated pointwise, the smoothness of  $f_{0X}$  is described locally in a Hölder scale and not globally by a decay condition on the Fourier transform of  $f_{0X}$ . Under these conditions on  $\hat{f}_\varepsilon$  and  $f_{0X}$ , the sup-norm convergence rate  $\eta_n$  in the direct density estimation problem is, up to a logarithmic factor, of the order  $n^{-(\alpha+\beta)/(2\alpha+2\beta+1)}$ . Consequently, quantiles are estimated, up to a log-factor, at the minimax-optimal rate  $n^{-(\alpha+1)/(2\alpha+2\beta+1)}$ , see the rate in (2). General sufficient conditions on the prior law and the data generating process to derive posterior sup-norm contraction rates are given in [8], along with examples of priors attaining minimax bounds. Derivation of the upper bound on the estimation error in (7) is based on Fourier inversion techniques that seem promising for extending results from single quantiles to the entire quantile function estimation in the  $L^1$ -norm.

### References

1. Al Labadi, L., Abdelrazeq, I.: On functional central limit theorems of Bayesian nonparametric priors. *Stat. Methods Appl.* **26**(2), 215–229 (2017)
2. Al Labadi, L., Zarepour, M.: On asymptotic properties and almost sure approximation of the normalized inverse-Gaussian process. *Bayesian Anal.* **8**(3), 553–568 (2013)
3. Conti, P.L.: Approximated inference for the quantile function via Dirichlet processes. *Metron* **62**(2), 201–222 (2004)
4. Dattner, I., Goldenshluger, A., Juditsky, A.: On deconvolution of distribution functions. *Ann. Statist.* **39**(5), 2477–2501 (2011)
5. Dattner, I., Reiß, M., Trabs, M.: Adaptive quantile estimation in deconvolution with unknown error distribution. *Bernoulli* **22**(1), 143–192 (2016)
6. Fan, J.: On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19**(3), 1257–1272 (1991)
7. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230 (1973)
8. Giné, E., Nickl, R.: Rates of contraction for posterior distributions in  $L^r$ -metrics,  $1 \leq r \leq \infty$ . *Ann. Statist.* **39**(6), 2883–2911 (2011)
9. Giné, E., Nickl, R.: Mathematical foundations of infinite-dimensional statistical models. *Cambridge Series in Statistical and Probabilistic Mathematics*, [40]. Cambridge University Press, New York (2016)
10. Hall, P., Lahiri, S.N.: Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.* **36**(5), 2110–2134 (2008)
11. Hjort, N.L., Petrone, S.: Nonparametric quantile inference using Dirichlet processes. In: *Advances in statistical modeling and inference, Ser. Biostat.*, vol. 3, pp. 463–492. World Sci. Publ., Hackensack, NJ (2007)
12. Meister, A.: Deconvolution problems in nonparametric statistics, *Lecture Notes in Statistics*, vol. 193. Springer-Verlag, Berlin (2009)
13. Scricciolo, C.: Bayesian Kantorovich deconvolution in finite mixture models. In: A. Petrucci, F. Racioppi, R. Verde (eds.) *New Statistical Developments in Data Science*, pp. 119–134. Springer International Publishing, Cham (2019)

# Bayesian inference for discretely observed non-homogeneous Markov processes

## *Inferenza Bayesiana per processi markoviani non-omogenei discretamente osservati*

Rosario Barone and Andrea Tancredi

**Abstract** Inference for continuous time non homogeneous multi-state Markov models may present considerable computational difficulties when the process is only observed at discrete time points without additional information about the state transitions. In fact, the likelihood can be obtained numerically only by solving the Chapman-Kolmogorov equations satisfied by the model transition probabilities. In this paper we propose to make Bayesian inference bypassing the likelihood calculation by simulating the whole continuous trajectories conditionally on the observed points via a Metropolis-Hastings step based on a piecewise homogeneous Markov process. A benchmark data set in the multi-state model literature is used to illustrate the resulting inference.

**Abstract** *L'inferenza per processi multi-stato markoviani non-omogenei a tempo continuo può presentare problemi computazionali quando il processo viene osservato solamente in determinati istanti. Infatti, la funzione di verosimiglianza può essere ottenuta esclusivamente risolvendo numericamente le equazioni di Chapman-Kolmogorov relative alle probabilità di transizione del processo. In questo lavoro mostriamo come, da un punto di vista bayesiano, il calcolo della verosimiglianza può essere evitato simulando le traiettorie continue condizionatamente ai punti osservati tramite un passo Metropolis-Hastings, utilizzando come modello generatore un processo markoviano omogeneo a tratti. L'approccio proposto sarà utilizzato per stimare i parametri del modello in un data set di riferimento nella letteratura dei modelli multi stato.*

**Key words:** Metropolis-Hastings, Multi-State models, Panel data, Uniformization

---

Rosario Barone  
Sapienza University of Rome, e-mail: rosario.barone@uniroma1.it

Andrea Tancredi  
Sapienza University of Rome, e-mail: andrea.tancredi@uniroma1.it

### 1 Introduction

Let  $\{X(t), t \geq 0\}$  be a continuous time multi-state process with state space  $\mathcal{S} = \{1, 2, \dots, S\}$ . Models for continuous time multi-state process  $X(t)$  can be defined via the transition intensity functions

$$q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \rightarrow 0} \frac{P\{X(t + \delta t) = s | X(t) = r, \mathcal{F}_t\}}{\delta t}$$

representing the instantaneous probability of a transition from state  $r$  to state  $s$  at time  $t$  when  $\mathcal{F}_t$  is the past history up to time  $t$ . Considering

$$P\{X(t + \delta t) = s | X(t) = r, \mathcal{F}_t\} = \begin{cases} \gamma_{rs} \delta t + o(\delta t) & s \neq r \\ 1 + \gamma_{rr} \delta t + o(\delta t) & s = r \end{cases} \tag{1}$$

where  $\gamma_{rs} \geq 0$  and  $\gamma_{rr} = -\sum_{s \neq r} \gamma_{rs} = -\gamma_r$ , we have a homogeneous Markov continuous time model, which is governed by the rate matrix  $\mathbf{G} = (\gamma_{rs} : r, s \in \mathcal{S})$ . Then  $P\{X(u+t) = s | X(u) = r\}$  is the  $(r, s)$  element of the exponential matrix

$$\exp(t\mathbf{G}) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mathbf{G}^r.$$

We may represent  $\{X(t), t \geq 0\}$  as a sequence of the visited states  $S_0, S_1, \dots$  and jump times  $Z_1, Z_2, \dots$ . In particular, we have a homogeneous Markov process if the sequence of states is a Markov chain with transition probabilities  $p_{rs} = \gamma_{rs} / \gamma_r$  and the sojourn times  $W_j = Z_j - Z_{j-1}$  are independent exponential random variables with rate  $\gamma_r$  depending on the visited state. The density of the complete sample path between 0 and  $T$  with state sequence  $s_0, s_1, \dots, s_n$  and jump times  $z_1, z_2, \dots, z_n$  is

$$p_M(s, z) = \left( \prod_{rs} p_{rs}^{n_{rs}} \right) \left( \prod_r \gamma_r^{n_r} e^{-\gamma_r d_r} \right) \times \frac{1}{\gamma_{s_\ell}} \tag{2}$$

where  $n_{rs} = \sum_{j=1}^n I(s_{j-1} = r, s_j = s)$  is the number of  $r \rightarrow s$  transition,  $n_r = \sum_{s \neq r} n_{rs}$  is the number of departures from the state  $r$  and  $d_r = \sum_{j=1}^{n+1} (z_j - z_{j-1}) I(s_{j-1} = r)$  is the total amount of time spent in the state  $r$ , with  $z_0 = 0$  and  $z_{n+1} = T$ . Finally note that  $s_\ell$  denotes the last visited state whose complete sojourn time is truncated. Note also that if the trajectory assumes an absorbing state the factor  $1/\gamma_{s_\ell}$  does not have to be included.

When the transition intensity of the process varies over time, that is the rate matrix  $G(t) = (\gamma_{rs}(t) : r, s \in \mathcal{S})$  depends on  $t$ , the Markov continuous time process is non-homogeneous and the transition probabilities  $P\{X(u+t) = s | X(u) = r\}$  will depend on both  $u$  and  $t$ . Also in the non-homogeneous case  $\{X(t), t \geq 0\}$  may be represented as the sequence of the visited states  $S_0, S_1, \dots$ , and jump times  $Z_1, Z_2, \dots$ . The sojourn time  $W_j = Z_j - Z_{j-1}$  for the state  $r$  given the entry time  $Z_{j-1} = z_{j-1}$  has density

$$f(w_j) = \gamma_r(z_j) e^{-\int_{z_{j-1}}^{z_j} \gamma_r(t) dt}$$

for  $i = 1, 2, \dots$  and the transition probabilities for the state sequences are  $P(S_j = r | S_{j-1} = s, Z_j = z_j) = \gamma_{rs}(z_j) / \gamma_r(z_j)$ . In this work we assume that  $\gamma_{rs}(z) = \gamma_r(z) \cdot p_{rs}$ , for  $s \neq r$  and  $r, s = 1, \dots, S$  hence the density of the complete sample path between 0 and  $T$  with state sequence  $s_0, s_1, \dots, s_n$  and jump times  $z_1, z_2, \dots, z_n$  is

$$p_{IM}(s, z) = \left( \prod_{rs} p_{rs}^{n_{rs}} \right) \left( \prod_{j=1}^{n+1} \prod_r \left[ \gamma_r(z_j) e^{-\int_{z_{j-1}}^{z_j} \gamma_r(\tau) d\tau} \right]^{I(s_{j-1}=r)} \right) \times \frac{1}{\gamma_{s_\ell}(z_n + 1)}. \quad (3)$$

## 2 Inference for non-homogeneous Markov models

Now suppose that the process is observed only at fixed points so that the state sequence  $s$  and the transition times  $z$  are not available. Let  $y = (y_0, y_1, \dots, y_m)$  be the observed states at times  $0 = t_0 < t_1 < \dots < t_m < t_{m+1} = T$ . In this framework, inference for the rate parameters of the homogeneous Markov model does not present particular issues beyond the numerical calculation of the transition probabilities  $P(X(t_j) = y_j | X(t_{j-1}) = y_{j-1})$  via the the exponential matrix. Instead the non homogeneous case is generally handled by assuming piecewise homogeneous Markov models. A wider class of non homogeneous models was proposed in [4] with spline rates functions  $\gamma_{rs}(t)$  depending on unknown parameters. Likelihood inference was obtained numerically by solving the set of Chapman-Kolmogorov differential equations satisfied by the transition probabilities  $P(X(t_j) = y_j | X(t_{j-1}) = y_{j-1})$  for each set of parameters defing the rates.

To make Bayesian inference for the non homogeneous case with general rate functions  $\gamma_r(t; \beta_r)$  for  $r = 1, \dots, S$ , we propose a Metropolis within Gibbs algorithm where at each step of the algorithm we generate the whole trajectory  $x(t)$  for  $t \in [0, t]$  given the observations at times  $0 = t_0 < t_1 < \dots < t_m < t_{m+1} = T$ . That is we simulate the sequences  $s$  and  $z$  conditionally on the observed points of the process. In particular we propose a trajectory generated by a piecewise homogeneous Markov process with constant rate matrix on the intervals  $[t_j, t_{j+1}]$  for  $j = 0, \dots, m$  accepting the proposed trajectory via the corresponding Metropolis-Hastings acceptance probability ratio. Note that the density of a piecewise Markov trajectory can be written as

$$p_{PW}(s, z) = \prod_{j=0}^m \left[ \left( \prod_{rs} p_{rs}(t_j)^{n_{rs}(t_j)} \right) \left( \prod_r \gamma_r(t_j)^{n_r(t_j)} e^{-\gamma_r(t_j) d_r(t_j)} \right) \times \frac{1}{\gamma_{s_\ell}(t_j)} \right]$$

where the quantities  $n_{rs}(t_j)$ ,  $n_r(t_j)$ ,  $p_{rs}(t_j)$ ,  $\gamma_r(t_j)$  and  $s_\ell(t_j)$  are analogous to those defining the density (2) but now refer to the interval  $[t_j, t_{j+1}]$ . Moreover the simulation of the non-homogeneous process given the observation  $y$  is straightforward since in every interval  $[t_j, t_{j+1}]$  we need only to simulate the homogeneous Markov process conditionally on  $[y_j, y_{j+1}]$ . This can be easily obtained by the uniformization



algorithm proposed in [1] to simulate a homogeneous Markov process conditioned on the endpoints. Then assuming to have a new simulated trajectory  $(s', z')$  from the piecewise homogeneous process conditioned on the observed points  $y$  we may accept it via a Metropolis-Hastings step with acceptance probability given by

$$\min \left\{ 1, \frac{p_{IM}(s', z'|y)p_{PW}(s, z|y)}{p_{IM}(s, z|y)p_{PW}(s', z'|y)} \right\} = \min \left\{ 1, \frac{p_{IM}(s', z')p_{PW}(s, z)}{p_{IM}(s, z)p_{PW}(s', z')} \right\}$$

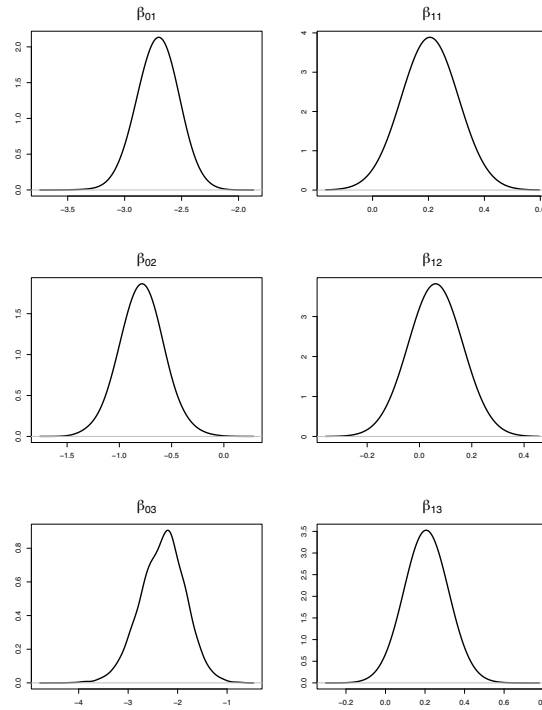
Note that for the parameters  $\gamma_r(t_j)$  of the generating piecewise Markov process we take the values  $\gamma_r(t_j; \beta_r)$  corresponding to the rates of the non homogeneous process at the beginning of each interval  $[t_j, t_{j+1}]$ . The updating of the parameters  $\beta_r$  at each step of the MCMC algorithm can be obtained via standard Metropolis steps depending also on the analytical function used to model the time dependence which in the proposed application is

$$\gamma_r(t) = e^{\beta_{0r} + \beta_{1r}t} \quad r = 1, \dots, S.$$

Note also that the proposed algorithm can be easily generalized to handle models with absorbing states and panel data configurations where a set of observed states  $y_i = (y_{i0}, y_{i1}, \dots, y_{im_i})$  at the follow-up times  $(t_{i0}, t_{i1}, \dots, t_{im_i})$  are available for  $i = 1, \dots, n$ , i.e. for each sample unit.

### 3 Application

We consider a data set describing the progression of the coronary allograft vasculopathy (CAV), a disease leading to the deterioration of arterial walls which is a common cause of death after heart transplantation. The data report the disease status (CAV-free (1), mild CAV (2) and moderate or severe CAV (3)) observed approximately each year after transplant for a set of 622 subjects followed up until their most recent visit if alive at the end of the observation period or until death (state (4)). This data set have been extensively analyzed by fitting different multi-state models, see for example [2,3,4, 5]. We fitted the non-homogeneous Markov model permitting transitions only to the adjacent states or to the death state. The model parameters are  $\theta = (p_{12}, p_{14}, p_{21}, p_{23}, p_{24}, p_{32}, p_{34}, \beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \beta_{03}, \beta_{13})$ . We assumed as non-informative prior for  $\beta_{r0}$  and  $\beta_{r1}$ ,  $r = 1, 2, 3$ , a Normal distribution with 0 mean and standard deviation 1000 and a Dirichlet prior on the transition probabilities  $p_r$ . We ran the MCMC algorithm for 10000 iterations with a burnin of 2000. The posterior summaries of all the parameters are reported in Table 1 while Figure 1 shows the posterior distributions for the intercept and the slope of the log rate parameters in the transient states.



**Fig. 1** CAV data set. Posterior densities for the intercept and the slope of the log rate parameters in the transient states

**Table 1** CAV data set. Posterior means, standard deviations and 0.025, 0.975 quantiles for the time non-homogeneous Markov model with  $p_{13} = p_{31} = 0$ .

	$\beta_{01}$	$\beta_{11}$	$\beta_{02}$	$\beta_{12}$	$\beta_{03}$	$\beta_{13}$
$E(\cdot x)$	-2.71	0.21	-0.79	0.06	-2.30	0.21
$SD(\cdot x)$	0.16	0.02	0.20	0.03	0.45	0.05
$q_{0.025}(\cdot x)$	-3.02	0.16	-1.18	0.00	-3.21	0.11
$q_{0.975}(\cdot x)$	-2.41	0.25	-0.40	0.12	-1.49	0.31

	$p_{12}$	$p_{14}$	$p_{21}$	$p_{23}$	$p_{24}$	$p_{32}$	$p_{34}$
$E(\cdot x)$	0.71	0.29	0.36	0.49	0.15	0.31	0.69
$SD(\cdot x)$	0.03	0.03	0.04	0.05	0.04	0.06	0.06
$q_{0.025}(\cdot x)$	0.65	0.25	0.29	0.40	0.09	0.20	0.57
$q_{0.975}(\cdot x)$	0.76	0.35	0.45	0.58	0.23	0.43	0.80

## References

1. Hobolth, A. and Stone, E. A. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *The Annals of Applied Statistics* **3**, 1204–1231. (2009)

2. Sharples, L. D., Jackson, C. H., Parameshwar, J., Wallwork, J., and Large, S. R. Diagnostic accuracy of coronary angiography and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation* **76**, 679–682 (2003)
3. Tancredi, A. Approximate Bayesian inference for discretely observed continuous-time multi-state models. *Biometrics* **75** 966-977 (2019)
4. Titman, A.C. Flexible nonhomogeneous Markov models for panel observed data. *Biometrics* **63** 780–787
5. Titman, A. C. , Sharples, L. D.. Semi-Markov models with phase-type sojourn distributions. *Biometrics* **66**, 742–752. (2010)

## 4.18 New developments in composite indicators applications

# Building composite indicators in the functional domain: a suggestion for an evolutionary HDI

## *Indicatori composti nel dominio funzionale: una proposta per un HDI evolutivo*

Francesca Fortuna, Alessia Naccarato and Silvia Terzi

**Abstract** Aim of this paper is to integrate composite indicator building with functional data analysis. For different geographical areas, we will exploit the time evolution of a composite indicator of well-being to highlight new interpretative issues and provide new frameworks for data interpretation. Specifically, an evolutive composite indicator is proposed, obtained by weighing the original indicator (the Human Development Index in our example application) with the first derivative of the function that approximates its temporal dynamics. We compute an evolutionary Human Development Index for the Asian Least Developed Countries.

**Abstract** *Lo scopo di questo articolo è di integrare la costruzione di indicatori composti con l'analisi funzionale dei dati. Per diverse aree geografiche, terremo conto dell'evoluzione temporale di un indicatore composto di benessere per fornire nuovi e più ampi contesti interpretativi. Nello specifico, viene proposto un indicatore composto "evolutivo", ricavato ponderando l'indicatore originale (Human Development Index nel nostro esempio applicativo) con la derivata prima della funzione che approssima la sua dinamica temporale. Come esempio applicativo, calcoliamo l'Human Development Index evolutivo per i paesi meno sviluppati dell'Asia.*

**Key words:** Composite indicators, Human Development Index, Functional data analysis

---

Francesca Fortuna  
"Roma Tre" University, Rome (Italy) e-mail: francesca.fortuna@uniroma3.it

Alessia Naccarato  
"Roma Tre" University, Rome (Italy) e-mail: alessia.naccarato@uniroma3.it

Silvia Terzi  
"Roma Tre" University, Rome (Italy) e-mail: silvia.terzi@uniroma3.it

## 1 Introduction

This paper aims to provide an original methodological approach for the analysis of composite indicators (CIs) using functional data analysis (FDA). The latter refers to the analysis of curves or functions in a continuous domain and assumes the existence of unknown smooth functions, which generate and underlie the data (see [5] and [2] for a detailed introduction to FDA). Within this framework, CIs should be considered as functions rather than scalar vectors. Specifically, we extend the concept of temporal sequences to CIs as functional data. The basic idea is that, in most cases, latent phenomena measured by CIs live in a continuous domain. Thus, contrary to the usual context, which considers CIs from a static point of view [4, 6], we exploit their evolutionary aspect highlighting new interpretative issues and providing new frameworks for data interpretation. The development of a functional approach within the context of CIs provides several advantages. First of all, it enriches interpretation by evaluating the CI's evolution (in time or space). Secondly, the functional approach can tackle cases where data are not sampled at equally spaced time points. Finally, it is possible to introduce new analytical tools, such as the derivatives, which may sometimes complement the original data with useful information [3, 1]. The latter aspect is particularly relevant to CIs because derivatives are potential quantifications of the function's behaviour in an evolutionary perspective. In our suggestion, the information provided by the CI will be integrated with the information regarding its temporal evolution, so as to define the evolutionary CI, say *ECI*. Each CI will be integrated in order to discount (reward) for a decreasing (increasing) evolution. Specifically, we propose an evolutionary CI, which is defined as  $ECI = CI(1 \pm \alpha)$ , where  $\alpha$  is a weight determined by the value of the first derivative.

Perhaps the most widely-used and well-known composite indicator of well-being is the Human Development Index (HDI) [7]. It is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indexes (life expectancy, education, Gross National Income) for each of the three dimensions.<sup>1</sup> As we can read in the UNDP site <http://hdr.undp.org/en/humandev>, the human development approach is about expanding the richness of human life, improving the lives people lead rather than assuming that economic growth will lead, automatically, to greater wellbeing for all. It means developing people's abilities and giving them a chance to use them. It is about providing people with opportunities, and improving people's well-being. Thus, however we choose to measure human development, it would be useful to complement it with information on the evolution of its components.

The remainder of the paper is organized as follows. Sect. 2 illustrates the analysis of CIs in a functional framework and introduces the evolutionary CI. Sect. 3 shows the main results obtained by applying the proposed approach to the time series of the HDI index for Asian Least Developed Countries from 1990 to 2019. Finally, Sect. 4 presents the conclusions of this study and further suggestions.

---

<sup>1</sup> Before 2011 HDI was computed as a weighted average of the three dimensions.

## 2 Composite indicators in a functional framework

Despite the continuous nature of functional data, in real applications, sample curves are observed with error in a discrete set of sampling points,  $t_1 < t_2 < \dots < t_L$  of  $\mathcal{T}$ . Thus, the observed data evaluated on  $t_1, t_2, \dots, t_L$ , can be expressed as follows:

$$y_{il} = y_i(t_{il}) + \varepsilon_{il}, \quad l = 1, \dots, L; \quad i = 1, \dots, n, \quad (1)$$

where  $y_{il}$  is the observed value for the  $i$ -th unit at the sampling point  $t_l$ ,  $y_i(t_{il})$  is a smooth function and  $\varepsilon_{il}$  is a measurement error [5]. One usual solution to reconstruct the functional form starting from the discrete and noisy observations, is to assume that sample paths belong to a finite-dimension space spanned by a suitable basis  $\{\phi_1(t), \phi_2(t), \dots, \phi_K(t)\}$ , so that the reconstructed smooth function for the  $i$ -th unit can be expressed as follows:

$$y_i(t) = \sum_{k=1}^K a_{ik} \phi_k(t), \quad i = 1, \dots, n, \quad (2)$$

where  $y_i(t)$ ,  $a_{ik}$  and  $\phi_k(t)$  represent the reconstructed smooth function, the basis coefficients and the basis functions for the  $i$ -th unit, respectively. Since we aim to evaluate the CI's evolution, our attention is focused not directly on functions but on the first derivative of the splines approximation, which permit to highlight a growth, a deceleration or a constant trend of the CI, depending on whether the first derivative is positive, negative or flat. Whenever the evolution of the CI is of interest, it could also be useful to embed this information into the CI. Our suggestion is to increase (or decrease) the CI by multiplying it by a coefficient that reflects this trend. Thus, the evolutionary CI, can be computed as follows:

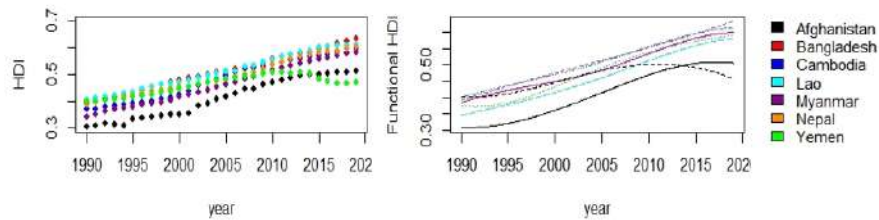
$$ECI = CI(1 + d_1) \quad (3)$$

Of course, we don't want this additional information, the evolution-integration, to prevail respect to the CI's level, thus, in some applications, it might be necessary to establish a criterion to choose the maximum weight ascribable to  $d_1$ .

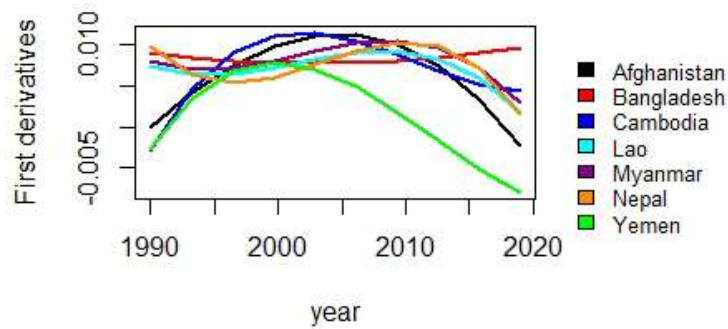
## 3 Application

In this section, we applied the suggested integration to the time series of the HDI indexes for the Asian Least Developed Countries. The list of Least Developed Countries (LDC) includes developing countries that, according to the United Nations, have the lowest socio-economic development indicators, with the lowest HDI of all countries in the world. Nine Asian countries are classified as LDC, however, we considered only seven of them, the ones that have a complete annual HDI index series from 1990 to 2019 (see Table 1 for the list of the countries). The HDI time

series of each country can be considered as a function observed in a discrete set of sampling points. The raw data were converted into a sample of functions adopting a B-splines basis expansion as in Equation (2). Specifically, the basis coefficients were obtained by least square approximation with  $K = 5$  cubic B-splines basis, chosen by cross validation. Fig. 1 shows both the raw HDI time series (left side) and the reconstructed functional HDI (right side) for the Asian LD countries. Using



**Fig. 1** Raw (left side) and functional (right side) HDI data for the Asian LD countries.



**Fig. 2** First and second derivatives for the Asian LD countries

the information provided by the first derivatives (see Fig. 2), the evolutionary HDI can be developed as in Equation (3). In the comparison between the rankings based on HDI and on the evolutionary HDI we found a strong concordance. The cases of mismatch between the HDI and the *EHD*I rankings are 6 out of 30 and are reported in Table 1. For these cases, the joint analysis of HDI and  $d_1$  values (which we don't report for reasons of space) revealed that *EHD*I is able to reflect the evolutionary dynamic of the indicator. Indeed, for HDIs that are identical or differ only slightly,



**Table 1** Ranking of the Asian LD counties obtained with HDI and *EHDI* for the cases of mismatching.

Country	HDI(1992)	<i>EHDI</i> (1992)	HDI(2005)	<i>EHDI</i> (2005)	HDI(2008)	<i>EHDI</i> (2008)
Afghanistan	7	7	7	7	7	7
Bangladesh	2	2	1	1	2	2
Cambodia	5	5	3	3	3	3
Lao	1	1	2	2	1	1
Myanmar	6	6	6	6	6	5
Nepal	4	3	5	4	4	4
Yemen	3	4	4	5	5	6
Country	HDI(2011)	<i>EHDI</i> (2011)	HDI(2012)	<i>EHDI</i> (2012)	HDI(2014)	<i>EHDI</i> (2014)
Afghanistan	7	7	7	7	7	6
Bangladesh	1	1	1	2	2	2
Cambodia	3	4	4	4	4	4
Lao	2	2	2	1	1	1
Myanmar	5	5	5	5	5	5
Nepal	4	3	3	3	3	3
Yemen	6	6	6	6	6	7

*EHDI* rewards the country that presents a higher value (even slightly higher) of  $d_1$ , thus recognising a growth of the indicator over time.

#### 4 Conclusions and further suggestions

By construction, the suggested *EHDI* accounts for the evolution (growth or decrease) of HDI and provides useful insight in people’s well-being improvement or worsening. In our application, the rankings based on the *EHDI* show moderate changes with respect to HDI rankings: some countries move one position up or down in the ranking (see Table 1). This happens most often to Yemen (4 times out of 6) and to Nepal (3 times out of 6). Not surprisingly, if we look at the derivatives (Fig. 2), since the derivatives of these two countries show values and trends that are quite different from those of the derivatives of the other countries: Yemen has decreasing HDI (a negative derivative) during most of the time span; while Nepal has alternating phases of growth, decline, growth, decline.

Suggestions for deeper investigation include further in-depth analysis to choose an optimal weight for the first order derivative  $d_1$ , and the inclusion of the second derivative  $d_2$  in the evolutionary component to highlight the rate of growth or decrease. Let us write the evolutive composite indicator in a general form as  $ECI = CI(1 + \alpha)$ . For what concerns the first issue, instead of setting  $\alpha = d_1$ , we could assign a weight to the first order derivative  $d_1$ , in order for  $\alpha$  to be bounded, and the evolutive component to be a relevant but not a too heavy component of human development the CI. For the second issue, we could set  $\alpha = d_1 + d_1d_2$ , to

include in the evolutionary component the rate of growth or decrease of the CI's trend.

## References

1. Di Battista, T., Fortuna, F., Maturo, F.: Environmental monitoring through functional biodiversity tools. *Ecological Indicators* **60**, 237–247 (2016)
2. Ferraty, F., Vieu, P.: *Nonparametric Functional Data Analysis*. Springer-Verlag, New York (2006)
3. Fortuna, F., Maturo, F., Di Battista, T.: Clustering functional data streams: Unsupervised classification of soccer top players based on Google trends. *Quality and Reliability Engineering International* **34**, (7), 1448–1460 (2018)
4. OECD: *Handbook on constructing composite indicators. Methodology and user guide*. OECD Publications, Paris (2008)
5. Ramsay, J.O., Silverman, B.W. *Functional Data Analysis*, 2nd edition. Springer, New York (2005)
6. Saisana, M., Tarantola, S.: *State-of-the-art report on current methodologies and practices for composite indicator development*. European Commission-JRC, EUR 20408 EN, Ispra (2002)
7. UNDP: *Human Development Report 2020*. Palgrave Macmillan, New York (2020).

# Small Area Estimation of Inequality Measures via Simplex Regression

## *Stima per piccole aree di misure di disuguaglianza per mezzo della regressione Simplex*

Silvia De Nicolò, Maria Rosaria Ferrante, and Silvia Pacci

**Abstract** This paper intends to propose a small area estimation strategy for Atkinson inequality measures. Classical proposals in area-level context show limitations when dealing with asymmetric heavy-tailed data and complex variance function. We consider alternative distributional assumption and propose a Simplex regression area level model. An application to EU-SILC income data is provided.

**Abstract** *Lo scopo di questo lavoro è proporre una strategia di stima per piccole aree per la stima dell'indice di disuguaglianza di Atkinson. I modelli area level proposti in letteratura appaiono limitati in caso di dati asimmetrici a code alte e variance function complesse. Specificando assunzioni alternative, proponiamo un modello area level con regressione Simplex e ne presentiamo un'applicazione ai dati dell'indagine EU-SILC.*

**Key words:** Atkinson Index, Inequality Measures, Simplex Regression, Small Area Estimation.

## 1 Introduction

The demand for reliable estimates of economic inequality measures for small areas is growing due to its importance in better planning public regional policies. Their estimation in small domains by using data taken from income surveys, such as EU-SILC, implies that the number of units sampled at area

---

Silvia De Nicolò  
Department of Statistical Sciences, University of Padova  
e-mail: [silvia.denicolo@phd.unipd.it](mailto:silvia.denicolo@phd.unipd.it)

Maria Rosaria Ferrante and Silvia Pacci  
Department of Statistical Sciences "P. Fortunati", University of Bologna  
e-mail: [maria.ferrante@unibo.it](mailto:maria.ferrante@unibo.it), [silvia.pacci@unibo.it](mailto:silvia.pacci@unibo.it)

level is generally too small to obtain reliable estimates. Thus, small area estimation techniques are required, allowing estimators to borrow strength across areas through auxiliary information, see [12] for a comprehensive review. The body of literature concerning estimation of inequality measures in small domains is very poor, including [5] for Gini Index at area level and [15] for Gini Index and Quintile Share Ratio via M-quantile-based models at unit level. Moreover, the estimation of alternative measures, as opposed to the widely used Gini index, may enable a more meaningful assessments of different aspects of economic inequality. Gini constitutes itself as a stochastic dominance measure, based on partial ordering of probability distributions [10]. Two very different distributions - one having more inequality amongst the poor, the other amongst the rich can have exactly the same Gini Index value. When the distributional dominance fails, welfare-based measures such as Atkinson Indexes, may provide for a *complete* ranking among alternative distributions. Furthermore, these measures satisfy the additive decomposability axiom, as opposed to Gini, allowing to decompose inequality into within areas and between areas components. Atkinson index is defined as:

$$A_{x_1, \dots, x_N}(\varepsilon) = \begin{cases} 1 - \frac{1}{\mu} \left( \frac{1}{N} \sum_{i=1}^N x_i^{1-\varepsilon} \right)^{1/(1-\varepsilon)} & \text{for } 0 \leq \varepsilon \neq 1 \\ 1 - \frac{1}{\mu} \left( \prod_{i=1}^N x_i \right)^{1/N} & \text{for } \varepsilon = 1, \end{cases}$$

where  $x$  is the variable of interest and  $\mu$  expected value. The parameter  $\varepsilon$  expresses the level of inequality aversion, as  $\varepsilon$  increases, the index becomes more sensitive to changes at the lower end of the income distribution. In our estimation strategy, we considered area level models [12], to be estimated by adopting a Hierarchical Bayesian approach via Hamiltonian MCMC.

## 2 Our model proposal

In the context of small area estimation of measures in  $(0, 1)$ , a huge body of literature is dedicated to proportions, implementing Fay-Herriot [12] and Beta regression models [7] with non-linear linking model. Small area models generally assume sampling variance or dispersion as known due to identifiability issues. Its separated estimation from the data involves rigorous steps requiring an in-depth analysis of the variance function. When dealing with proportions, variance under s.r.s. is  $V(y) = \mu(1 - \mu)/n$  (with  $\mu$  expected value), under binomial generating process assumption. However when dealing with non-proportions measures with a different and perhaps more complex variance function, the Beta distribution parametrization appears limited and too restrictive leading to (a) re-parametrizations that dramatically complicate the density geometry leading to several problems in the estimation step; (b) over-simplification of the variance function itself that may limit measures support or constrain expected values in such a way that shrinking process

is undermined. As a first conclusion, Beta regression does not seem to provide enough flexibility to model complex variance functions. On the other hand, the Simplex distribution [?], pertaining to the dispersion models family, appears more suitable. This two parameters-distribution is known to be an alternative to Beta distribution in terms of robustness [3], over-dispersion control [8] and large left and right skewness modeling [1]. Moreover, simplex distribution seems to work better than Beta with parameters close to the boundaries of the support [3]. Simplex regression model is used in [11], [13], an R package [17] is available. A proposal of covariate measurement error model involving simplex distribution is found in [1], residual and influence analysis to a general class of simplex regression is provided by [3].

The generic Atkinson index estimator is negatively biased in small samples, direct estimators have been bias-corrected following [2]. Our small area model proposal with Simplex distribution for  $y_d$ , the bias-corrected direct estimator of a generic Atkinson index and  $x_d$  a set of  $p$  generic covariates for  $m$  small areas is as follows:

$$\begin{cases} y_d|\theta_d \sim S^-(\theta_d, \sigma_d^2) & \forall d = 1, \dots, m \\ \text{logit}(\theta_d) = x_d^T \beta + v_d \end{cases} \quad (1)$$

where (a)  $v_d \sim N(0, \sigma_v^2)$  or alternatively (b)  $v_d \sim t(0, \nu, \sigma_v^2)$ . Consider that, by parametrization of the Simplex,  $\theta_d = E(y_d|\theta_d)$  and

$$\sigma_d^2 = E[d(y_d; \theta_d)|\theta_d] = E\left[\frac{(y_d - \theta_d)^2}{y_d(1 - y_d)\theta_d^2(1 - \theta_d)^2}\right]. \quad (2)$$

We estimated via MCMC:  $\beta$ , a vector  $(p + 1) \times 1$  of regression coefficients, random effects dispersion  $\sigma_v^2$  and d.o.f.  $\nu$  in case of  $t$  distributed random effects. Dispersion coefficients  $\sigma_d^2$  are supposed known, as usual, and estimated from data by mean of a two steps procedure as follows. Initially, it is estimated by a proper bootstrap procedure developed taking into account the complex sampling design [4]. The area specific bootstrap estimates from  $B = 1000$  repeated samples are obtained by using bias-corrected method of moments estimators of the dispersion parameter as in [11]. Secondly, those estimates are smoothed via a Generalized Variance Function approach as in [4] in order to reduce bootstrap sampling error. To do so, we derived the variance function of Atkinson indexes as follows.

**Theorem 1.** *Under the assumption of log-normality of income variable i.e.  $\log(x_{jd}) \sim N(\mu_d, \phi_d^2)$ , with  $j = 1, \dots, N_d$  the individuals and  $d = 1, \dots, m$  the areas, the s.r.s. estimator of Atkinson ( $\varepsilon$ ) measure has variance function*

$$V(\hat{\theta}_d) \cong \frac{2\theta_d^2}{n_d} \exp\{-2\theta_d\}. \quad (3)$$

*Proof.* Under the underlined assumptions, let us consider  $\theta_d = \text{Atk}(\epsilon) \forall$  real and complex  $\epsilon \geq 0$  and  $\neq 1$ , but this result can be easily generalized to the case  $\epsilon = 1$ , is equal to

$$\theta_d = 1 - \exp\{-\epsilon\phi_d^2/2\}, \tag{4}$$

with  $\phi_d^2$  estimated by  $s_d^2 = \frac{1}{n_d-1} \sum_{j=1}^{n_d} (\log(x_{jd}) - \hat{\mu}_d)^2$ . By applying the normal distribution theory  $V(s_d) \cong \frac{\phi_d^2}{2n_d}$ , and using delta method:

$$V(\hat{\theta}_d) = V(1 - \exp\{-\frac{\epsilon s_d^2}{2}\}) \cong \frac{\epsilon^2 \phi_d^4}{2n_d} \exp\{-\epsilon\phi_d^2\} \cong \frac{2\theta_d^2}{n_d} \exp\{-2\theta_d\} \tag{5}$$

where the last right hand side equation is obtained by expanding (4), so that  $\phi_d^2 \cong 2\theta_d/\epsilon$ .  $\square$

Let us consider that  $\sigma_d^2 \cong V(y_d|\theta_d)/[(1-\theta_d)^3\theta_d^3]$  by second-order expansions of (2) around  $\theta_d$ , by combining it with (3), let easily assume that  $\sigma_d^2 = f(\theta_d) \times IF/n_d$  with  $IF$  denoting a design-effect variance inflation factor induced by the complex sampling, assumed not to vary across areas, and  $n_d$  area sample size under complex sampling. Therefore, considering  $\psi = 1/IF$ , we introduce the following smoothing model:

$$2 \frac{\exp\{-2\theta_d\}}{\theta_d(1-\theta_d)^3 \hat{\sigma}_{dboot}^2} = n_d\psi + \epsilon_d \tag{6}$$

where  $\epsilon_d$  are zero-mean and heteroskedastic residuals, estimated via generalized least squares. The smoothed estimator follows from (3) by replacing  $\theta_d$  with  $y_d$  and  $n_d$  with  $n_d\hat{\psi}$ .

### 3 Application and Results

An application to evaluate equalized income inequality in Italian NUTS-3 regions is provided by using 2017 EU-SILC data. The Atkinson index considered has parameter  $\epsilon = 2$ , having an high degree of inequality aversion and asymmetry [2], which leads our model to be particularly suitable. In order to stem heavy non-robustness and bias issues, income data have been treated by semi-parametric pareto and inverse pareto tail modeling procedure following [6] and [9]. As auxiliary variables we considered both fiscal and registry office data related to each of the 107 provinces. In particular: population density, aged dependency ratios, % of foreigners residents, people in higher education ratio, average taxable income, % of residents filling tax forms, % of residents filling tax forms with income lower than/greater than double national median and lastly, Atkinson measure calculated on income classes declared by tax forms.

We proceeded estimating model in Section (2) by using different priors and by comparing it to two baseline models, (1) a Logit Fay-Herriot, i.e. with sampling model  $y_d|\theta_d \sim N(\theta_d, D_d)$  and (2) a Beta sampling model  $y_d|\theta_d \sim Beta(\theta_d(1-\phi_d), (1-\theta_d)(1-\phi_d))$ , both with linking model  $logit(\theta_d) \sim N(x_d^T\beta, \sigma_v^2)$ , with  $D_d$  and  $\phi_d$  known and estimated separately similarly to what explained in Section (2). Priors for regression parameters  $\beta$  were set out to be  $\beta \sim N_p(\bar{0}, \Sigma)$  with  $\Sigma$  diagonal matrix with diagonal  $10 \times \bar{1}_p$ , then we considered 4 different settings for the Simplex model: (1) with Normal random effect and  $\sigma_v^2 \sim InvGamma(0.01, 0.01)$  prior, (2) Student-t random effect with  $\nu \sim Exp(0.5)$  and  $\sigma_v^2 \sim InvGamma(0.01, 0.01)$  prior (3) Student-t random effect with  $\nu \sim Exp(0.5)$  and  $\sigma_v \sim Half\ t_3(0, 100)$  priors, (4) Student-t random effect with  $\nu \sim Exp(0.1)$  and  $\sigma_v \sim Half\ t_3(0, 100)$ .

Some diagnostic measures have been used for comparison, as regards goodness-of-fit, DIC [14] and looic, based on leave-one-out cross-validation [16], whereas a precision improvement measure, the Standard Deviation Reduction measure:  $SDR(\theta_d) = 1 - \sqrt{V(\theta_d|data)/\hat{V}(y_d|\theta_d)}$  has been used to evaluate model-based estimators performances. Variances for each small area are calculated using the exact formula.

As clear from results set out in Table (1), our model shows better performance both in terms of fitting and variability reductions. Simplex regression better captures skewness and the slowly-decaying upper tail of the distribution, thus avoiding to underestimate economic inequality. In fact model based estimates of simplex model appears to be slightly higher than the ones estimated by Fay-Herriot model, due to the conditional mean  $\mu_d = E(y_d|X)$  sensibility

	looic	dic	asdr
logit F-H	-294.83	-319.06	0.29
Beta	-296.33	-318.16	0.30
1	-298.47	-320.05	0.31
2	-306.95	-322.47	0.34
3	-304.65	-321.34	0.34
4	-302.58	-321.59	0.33

**Table 1** Looic, DIC and Average SDR of the models estimated.

to heavy tails. Moreover even further robustness-aware tools are required as showed by the implementation of *Student's t* random effect which results in better performances in term of goodness of fit and shrinkage. By considering model with setting (2), we first check the adequacy by using a posterior predictive approach via a discrepancy measure  $P(y_d < \theta_d^{rep})$ , where  $\theta_d^{rep}$  is generated from the posterior predictive. Values near 0 and 1 would show evidence of under or over estimation. In our application, the average over areas was 0.506, with quartiles equal to 0.29, 0.52, 0.68, which means a good fit with only two values outside (0.1, 0.9). The sd reduction ranges from -8% to 69% with quartiles 23%, 35% and 45%. Moreover, model estimates shows design consistency, i.e. convergence on direct estimators in large samples.

## 4 Concluding Remarks

We proposed a Simplex regression small area model for Atkinson Index, which provides a more flexible framework with respect to Beta regression, allowing to incorporate complex variance function, and which appears to work better than classical gaussian area models avoiding to underestimate inequality. Further directions of research involve expanding it to other measures and developing a multivariate context.

## References

1. Carrasco, J.M., Reid, N.: Simplex regression models with measurement error. *Communications in Statistics: Simulation and Computation* **0**(0), 1–16 (2019)
2. De Nicolò, S., Ferrante, M.R., Pacei, S.: Mind the Income Gap: Behaviour of Inequality Estimators from Complex Survey Small Samples. Working Paper (2021)
3. Espinheira, P.L., de Oliveira Silva, A.: Residual and influence analysis to a general class of simplex regression. *Test* **29**(2), 523–552 (2020)
4. Fabrizi, E., Ferrante, M.R., Pacei, S., Trivisano, C.: Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Computational Statistics and Data Analysis* **55**(4), 1736–1747 (2011)
5. Fabrizi, E., Trivisano, C.: Small area estimation of the Gini concentration coefficient. *Computational Statistics and Data Analysis* **99**, 223–234 (2016)
6. Finkelstein, M., Tucker, H.G., Alan Veeh, J.: Pareto Tail Index Estimation Revisited. *North American Actuarial Journal* **10**(1), 1–10 (2006)
7. Janicki, R.: Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Communications in Statistics - Theory and Methods* **49**(9), 2264–2284 (2020)
8. Jørgensen, B.: *The Theory of Dispersion Models*, vol. 41 (1997)
9. Masseran, N., Yee, L.H., Safari, M.A.M., Ibrahim, K.: Power law behavior and tail modeling on low income distribution. *Mathematics and Statistics* **7**(3), 70–77 (2019)
10. Muliere, P., Scarsini, M.: A note on stochastic dominance and inequality measures. *Journal of Economic Theory* **49**(2), 314–323 (1989)
11. Qiu, Z., Song, P.X., Tan, M.: Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics* **35**(4), 577–596 (2008)
12. Rao, J.N., Molina, I.: *Small-area estimation*. Wiley Series in Survey Methodology (2015)
13. Song, P.X., Qiu, Z., Tan, M.: Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal* **46**(5), 540–553 (2004)
14. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A.: The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 485–493 (2014)
15. Tzavidis, N., Marchetti, S.: Robust domain estimation of income-based inequality indicators. *Analysis of Poverty Data by Small Area Estimation* pp. 171–186 (2016)
16. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**(5), 1413–1432 (2017)
17. Zhang, P., Qiu, Z., Shi, C.: simplexreg: An R package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Software* **71**(11) (2016)



# Relational Well-Being and Poverty in Italy

## *Benessere relazionale e povertà in Italia*

Elena Dalla Chiara and Federico Perali

**Abstract** The study measures the many dimensions of poverty taking into account both the material and non-material dimensions of well-being. The monetary dimension alone is not sufficient to fully describe also the non-material dimensions of need. We estimate poverty for different types of income by adding wealth (current income) and the value of domestic production (extended income) to disposable income. The study shows that the ability to produce domestic and relational goods, after taking into account of the monetary dimensions, are very important factors that redraw the map of poverty in Italy.

**Abstract** *Lo studio misura le molte dimensioni della povertà tenendo conto sia del benessere materiale sia non materiale. La sola dimensione monetaria non è sufficiente a descrivere in modo completo le dimensioni non materiali del bisogno. Abbiamo stimato la povertà per diverse tipologie di reddito aggiungendo al reddito disponibile la ricchezza (reddito corrente) e il valore della produzione domestica (reddito esteso). Lo studio mostra che la capacità di produrre beni di cura in famiglia e beni di relazione, dopo aver considerato le dimensioni monetarie di mercato, sono fattori importanti che ridisegnano la mappa della povertà in Italia.*

**Key words:** current income, extended income, multidimensional poverty, well-being.

## 1 Introduction

Poverty in general is a state of deprivation that limits the ability to achieve a minimum standard of living of material well-being associated with a level of income that allows to accomplish and consume market and non-market goods. The monetary dimension alone is not sufficient to fully describe the non-material dimensions of

---

<sup>1</sup> Elena Dalla Chiara, Interdepartmental Center of Economic Documentation (CIDE), University of Verona, email: elena.dallachiara@univr.it  
Federico Perali, University of Verona (Italy), Department of Economics and CHILD, email: federico.perali@univr.it

need (Sumner, 2004). Not all goods and services that are important to people are obtained from the market, such as child or elder care, a peaceful atmosphere and good relationships in the family and in the community we live. The prices describing the value of these goods and services are not defined by the market but are implicit and differ from person to person. These non-market goods influence a person's well-being and affect their ability to be and do what they most desire (Sen, 1987). These relational dimensions are generally important, but dramatically so in times of health and economic emergencies. The primary objective of this study is to quantify the relevance of both the domestic and relational goods produced by households (Donati, 2019; Matteazzi et al., 2020) and to explain how the traditional poverty map based on monetary metrics changes if deprivation is also defined in relation to a set of non-monetary attributes that influence the level of well-being.

It is also important to include in the analysis of the monetary dimension relevant information, often overlooked in the literature, such as regional differences in the cost of living and quality of life, and the value of wealth. This paper devotes particular attention to these aspects that, if neglected, would invalidate the estimates of primary interest in this study concerning the importance of relations for the well-being of the person.

## 2 Methodology. How to Measure Well-Being: Not just income

A family maximizes its well-being by taking employment opportunities in non-business activities remunerated at market values, in business activities, and in household and care activities into account. The level of employment also depends on the availability of income, assets and inherent household capabilities. This information allows to estimate disposable, current and extended income (Lustig, 2018). Therefore, we can also describe monetary poverty in a multidimensional way. In addition, when we make comparisons across individuals or households, it is important to take into account differences in purchasing power between households living in different regions, the quality of the services offered and the household equivalent scale. For these reasons, this study uses the quality adjusted true cost of living index estimated by Menon et al. (2020) as an income deflator and the OECD-modified scale as equivalence scale (Istat, 2019) to derive the quality adjusted real equivalized (henceforth *QARE*) income declined in terms of disposable, current and extended income.

We extend the analysis to include deprivation dimensions about family relationship and well-being by following a multidimensional aggregation technique (Alkire, 2008; Alkire and Foster, 2011). This approach implies a model of poverty that moves beyond the traditional definition of poverty based on the income dimension alone. The Multidimensional Poverty Index (MPI) adopted in our study is based on a socio-economic and a relational subset of 6 and 4 equally weighted dimensions, respectively. Table 1 reports these dimensions and their associated

Relational Well-Being and Poverty in Italy

poverty thresholds. A household is defined as multidimensional poor if it is simultaneously deprived in half of the dimensions considered.

**Table 1:** Socio-economic and relational dimensions of multidimensional poverty

<i><b>Poverty Dimensions</b></i>	<i><b>Description and Threshold (in parenthesis)</b></i>
<i>Socio-economic dimensions</i>	
Poverty of income	<i>QARE</i> disposable income (half median)
Poverty of wealth	Movable and immovable family assets (half median)
Poverty of education	Education level - household head (middle school)
Poverty of parents	Number of parents in the household (single parent)
Poverty of jobs	Presence of unemployed in the family (unemployed members)
Poverty of time	Time devoted to child and home care (half median)
<i>Relational dimensions</i>	
Poverty of bonding capital	Trust in family members (7 on scale 0-10)
Poverty of bridging capital	Trust in friends (7 on scale 0-10)
Poverty of relationship 1	Relationship satisfaction with children (7 on scale 0-10)
Poverty of relationship 2	Satisfaction with the time spent together (7 on scale 0-10)

### **3 Results. The Many Dimensions of Poverty**

To implement this multidimensional analysis, we use the integrated database to measure the standard of living in Italy created by Dalla Chiara et al. (2019). The results are presented in two parts. The first part deals with monetary poverty alone and also including assessments of household production. The second part reports measures of non-monetary multidimensional poverty introducing relational aspects.

#### **3.1 Monetary Poverty Adjusting for Quality of Services**

The difference in purchasing power explains the reversal of the North-South gradient in the incidence of poverty at the macro-region and national levels analyzing household income, that is observed when incomes are transformed into real (Table 2). What is important to observe is the narrowing of the North-South gap when equivalent scale and regional differences in the cost of living are taken into account. The shift from per-capita to equivalent adults leads to an overall reduction in poverty levels due to the fact that the income of the single individual is generally lower than family income, while in terms of equivalent adult the income of one-person household becomes relatively higher and is placed in higher quintiles of the distribution.

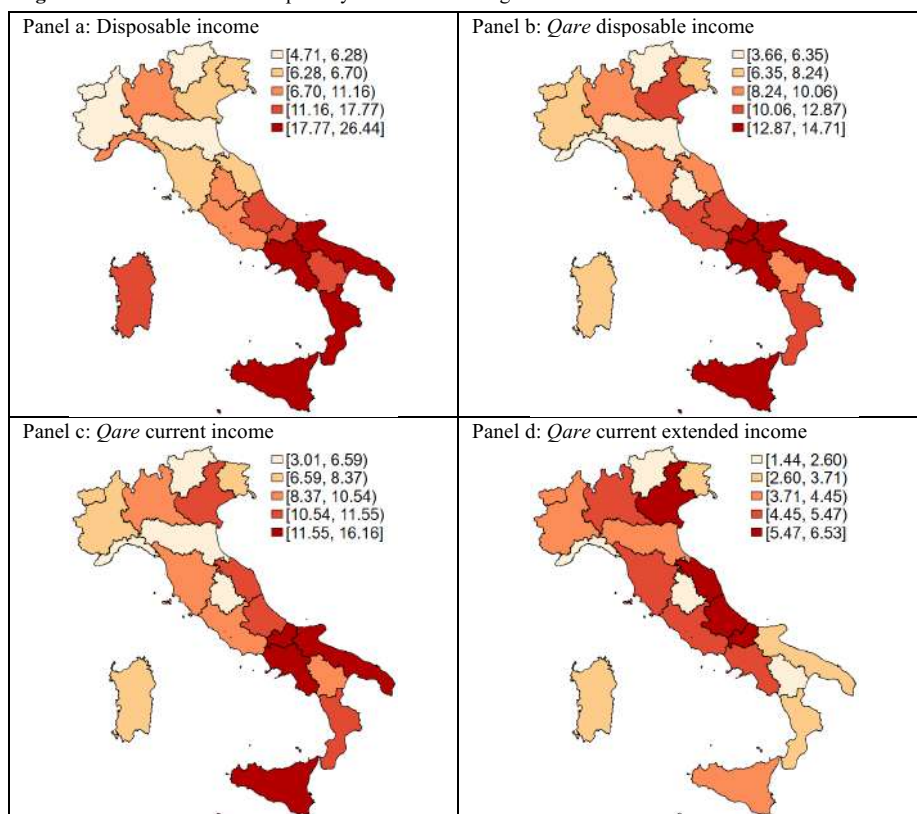
The representation of the poverty map is completed with other monetary dimensions as the value of assets (current income) and household production (extended income) and the non-monetary dimensions described in the second part.

Figure 1 shows how the incidence of relative poverty changes significantly according to the type of income taken into consideration. The relative poverty of Trentino Alto Adige is in all cases in the lower quintile of the distribution of relative poverty, while Sicily moves from the highest to the third quintile if extended income is considered.

**Table 2:** Incidence of relative poverty: comparison of different definition of disposable income (y)

<i>Macro-Region</i>	<i>Disposable income (y)</i>	<i>Real y</i>	<i>Equivalent y</i>	<i>Equivalent real y</i>	<i>QARE y</i>
North	12.69	16.17	6.16	8.71	7.77
Centre	12.95	13.81	7.83	8.34	8.84
South	20.58	13.76	19.38	10.07	12.29
Italy	15.10	14.91	10.47	9.02	9.36

**Figure 1:** Incidence of relative poverty for different categories of income

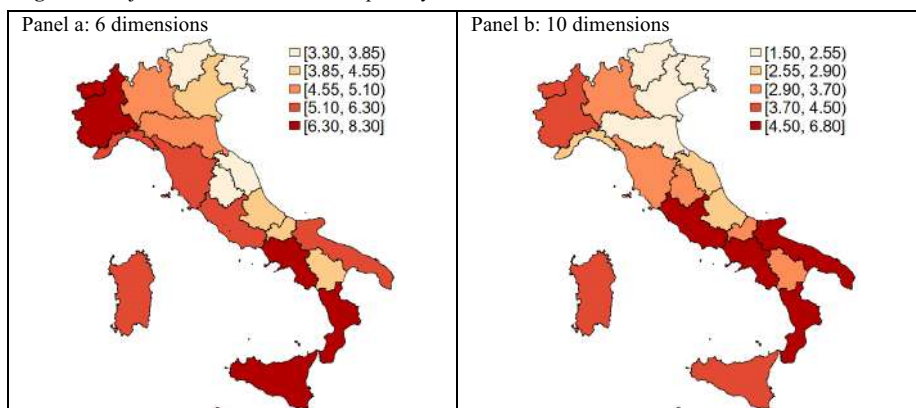


### 3.2 *Multidimensional Socio-economic and Relational Poverty*

The multidimensional poverty analysis is limited to families with children because the relational dimensions (Table 1) are less meaningful for individuals living alone. Figure 2 presents the adjusted measure of multidimensional poverty incidence (M0) for the 6 socio-economic dimensions (panel a) and for the 10 socio-economic and relation dimensions (panel b) that weighs poverty incidence (H0) with the average of the subset of dimensions in which households are jointly poor. Analyzing socio-economic dimensions, regions in the East of Italy are relatively less poor than regions in the West. Adding the relational dimensions reduces the overall poverty level but also changes the map. Northern regions improve their relative position while the situation worsens in Umbria, Molise, Basilicata and Apuglia.

It is very informative to analyze the relative contribution of each dimension to multidimensional poverty. Wealth explains about a quarter of multidimensional poverty, about ten percent more than labor income. Homeownership, which is more common among the elderly, is a strong protective factor against poverty risk. While unemployment, as expected, is a factor more important in the South, the presence of a single parent in the household and less investment of time in caring for children and the elderly are relatively more important risk factors in the North. The education level of the household head is the relatively least relevant dimension, especially in the North. This information is critical for designing effective interventions to mitigate and prevent exposure to poverty risk. Much attention should be paid to the consequences for the well-being of family members associated with single-parent situations or where there are situations in which it is difficult to reconcile work and family time.

**Figure 2:** Adjusted incidence of relative poverty



## 4 Conclusions

This study provides an extended measure of poverty for Italy that also includes non-monetary dimensions. The purpose is to highlight the importance of the multiple components contributing to well-being, that in general are not made available through markets, placing special emphasis on the relevance of the production of household goods and relational aspects in the household. The study shows that in the multidimensional poverty estimation, time spent caring for children contributes to poverty as much as income. It is a risk factor especially in Northern Italy. Among the relational well-being dimensions, trust in friends and satisfaction about the time spent with family members contribute to poverty significantly more than income.

This evidence illustrates the role that the family plays in Italy as a factor in preventing and treating the risk of poverty, especially in times of health, economic, or environmental emergencies. It also shows the relevance of relational dimensions to improve the efficiency and cost-effectiveness of public action to fight poverty and improve accuracy of targeting aid to fragile families.

## References

1. Alkire, S.: Choosing Dimensions: The Capability Approach and Multidimensional Poverty. In: Kakwani, N., Silber, J. (eds.) *The Many Dimensions of Poverty*. Palgrave Macmillan, New York (2008)
2. Alkire, S., Foster, J.: Counting and Multidimensional Poverty Measurement. *Journal of Public Economics* (2011) doi: 10.1016/j.jpubeco.2010.11.006
3. Dalla Chiara, E., Menon, M., Perali, F.: An Integrated Database to Measure Living Standards. *Journal of Official Statistics* (2019) doi: 10.2478/jos-2019-0023
4. Donati, P.P.: *Discovering Relational Assets. To Generate a New Sociality*. Rubbettino Editore (2019)
5. Istat: *BES Report: Equitable and Sustainable Well-Being in Italy* (2019)
6. Lustig N.: *Commitment to Equity Handbook - Estimating the Impact of Fiscal Policy on Inequality and Poverty*. Brookings Institution Press, Washington, D.C (2018)
7. Matteazzi, E., Menon, M., Perali, F.: Family: economic subject at risk and protective factor from material and relational poverty risk. *Family and Relational Poverty, Family International Monitor Report* (2020)
8. Menon, M., Perali, F., Ray, R., Tommasi, N.: *Regional Price Parities Accounting for Differences in the Quality of Services: The Tale of the Two Italies*. Department of Economic Sciences, University of Verona, Working Paper Series (2020)
9. Sen, A.: *The Standard of Living: Tanner Lectures in Human Values*. Cambridge University Press, Cambridge (1987)
10. Sumner, A.: *Economic Well-being and Non-economic Well-being*. United Nation University, World Institute for Development Economics Research WIDER Research Paper No. 30 (2004)

# A composite indicator to assess sustainability of agriculture in European Union countries

## *Un indicatore composito per valutare la sostenibilità dell'agricoltura nei paesi dell'Unione Europea*

Alessandro Magrini and Francesca Giambona

**Abstract** In this paper, we propose a composite indicator to measure sustainability of agriculture in European Union (EU) countries, based on the geometric aggregation of twelve indicators through weights endogenously determined according to the Benefit of Doubt (BoD) approach. We considered a broad set of indicators covering the economic, social and environmental dimensions, a comprehensive set of countries (25 in total), and data over a long period (15 years) to explore not only the level of sustainability, but also its evolution in time. The results show that Finland, Germany, Austria and Sweden have the highest average score, but also an average annual growth below the median. Instead, Czechia, Slovakia, Italy, Romania, Portugal and Spain have the lowest average score but, excepting Portugal, they show a higher average annual growth than the countries with the highest average score.

**Abstract** *In questo articolo proponiamo un indicatore composito per la misura della sostenibilità nei paesi dell'Unione Europea (EU), basato sull'aggregazione geometrica di dodici indicatori con pesi determinati endogeneamente secondo l'approccio BoD (Benefit of Doubt). Abbiamo considerato un insieme ampio di indicatori che copre la dimensione economica, sociale e ambientale, un insieme esaustivo di paesi (25 in totale) e dati su un periodo lungo (15 anni) per esplorare sia il livello della sostenibilità che la sua tendenza. I risultati mostrano che Finlandia, Germania, Austria e Svezia hanno il punteggio medio di sostenibilità più alto, ma anche una crescita media annua sotto la mediana. Invece, Cechia, Slovacchia, Italia, Romania, Portogallo e Spagna hanno il punteggio medio di sostenibilità più basso ma, ad eccezione del Portogallo, hanno una crescita media annua più elevata rispetto ai paesi col punteggio medio di sostenibilità più alto.*

**Key words:** benefit of doubt, EU agriculture, sustainable development, sustainable dimensions, weighted product approach.

---

Alessandro Magrini, Department of Statistics, Computer Science, Applications – University of Florence, Italy, e-mail: [alessandro.magrini@unifi.it](mailto:alessandro.magrini@unifi.it)  
Francesca Giambona, Department of Statistics, Computer Science, Applications – University of Florence, Italy, e-mail: [francesca.giambona@unifi.it](mailto:francesca.giambona@unifi.it)

## 1 Introduction

Nowadays, the agricultural sector is called to face in the front row the challenge of satisfying food demand of the rapidly increasing world population. For this reason, sustainability of agriculture has become a widely spread theme among international decision makers, and it has found large space in the 2030 agenda for the Sustainable Development Goals (SDGs) of the United Nations. Agricultural sustainability is a multidimensional concept involving the efficient production of goods and services (economic dimension), the management of natural resources (environmental dimension), and the improvement of conditions in rural areas (social dimension) [8, 5, 4, 6]. As such, the major critical issue in sustainability assessment concerns the modality of synthesizing the considered indicators. Composite indicators represent a commonly adopted methodology, but they are subjected to several arbitrary choices, like the determination of the weights and the degree of compensability [10].

Existing empirical studies on agricultural sustainability in the EU have at least one of the following three limitations: (i) some sustainable dimensions are disregarded or accounted only partially, thus agricultural sustainability cannot be understood in all of its relevant aspects; (ii) the focus is at farm level or on a small set of countries, thus it is not possible to draw an exhaustive picture of the sustainability level of EU agriculture; (iii) cross-sectional data are considered, thus the evolution of sustainability in time cannot be assessed. The most valuable studies in the literature include [2, 3], where longitudinal data on a broad set of countries are considered but the social dimension of sustainability is disregarded, and [9, 7, 1], where all the three sustainable dimensions are taken into account on a broad set of countries but the assessment relies on cross-sectional data.

This paper contributes to the literature by proposing a composite indicator for the measurement of agricultural sustainability in EU countries. Our composite indicator is based on twelve basic indicators covering the economic, environmental and social dimensions. We adopt the weighted product method (geometric aggregation) with weights endogenously determined according to the Benefit of Doubt (BoD [11]) approach, and consider yearly data on 25 EU countries in the period 2004–2018 (15 years) to assess not only the level of sustainability in EU countries, but also its trend. In Section 2, the data are described and the methodology is presented. In Section 3, the results are reported and discussed. Section 4 contains concluding remarks.

## 2 Data and methodology

The selection of the indicators was based on theory and guidelines in the literature [8, 5, 4, 6], and data collection relied on publicly available statistics from Eurostat, FAO and OECD. We selected a set of indicators and a temporal window as large as possible balancing representativeness of the three sustainable dimensions (economic, social and environmental) and availability of time series data. The resulting dataset included twelve indicators: five for the economic, three for the social and



four for the environmental dimension, measured yearly on 25 EU countries (Croatia, Cyprus and Malta excluded due to data unavailability) in the period 2004–2018 (15 years).

The selected indicators for the economic dimension cover the objectives of productivity (total factor productivity index of agriculture, from Eurostat), capital investments (net capital stocks to gross value added in agriculture, from FAO), manager turnover (ratio young/elderly for farm managers, from Eurostat) and profitability (real income of agricultural factors per paid annual work unit, and net entrepreneurial income per unpaid annual work unit, both from Eurostat).

The selected indicators for the social dimension refer to the ability of agriculture to deal with inequality and abandonment in rural areas, and include: median equivalised net income, at-risk-of-poverty rate and unemployment rate, all measured in rural areas and sourced from Eurostat.

The selected indicators for the environmental dimension cover two objectives: increase of practice favouring a positive development of the natural environment (share of production of renewable energy from agriculture, and area under organic cultivation to total agricultural area, both from Eurostat), and reduction of the negative pressure to the natural environment (greenhouse gas emissions per hectare due to agriculture, from FAO, and gross nitrogen balance per hectare, from OECD).

To construct our composite indicator for agricultural sustainability in EU countries, we adopted the weighted product method (geometric aggregation of the basic indicators), with weights determined endogenously according to the Benefit of Doubt (BoD) approach [11], i.e., maximizing the score of each country.

Let  $i = 1, \dots, n$  denote the country,  $I_{ij}$  the measurement of the  $j$ -th indicator ( $j = 1, \dots, p$ ) on country  $i$ , and  $w_{ij}$  the weight of indicator  $j$  for country  $i$ . Our composite indicator is defined as:

$$CI_i = \prod_{j=1}^p I_{ij}^{w_{ij}} \quad i = 1, \dots, n \quad (1)$$

where the weights  $w_{ij}$  are determined by solving, for each  $i$ :

$$\max \prod_{j=1}^p I_{ij}^{w_{ij}} \quad \text{subjected to:} \quad \prod_{j=1}^p I_{kj}^{w_{kj}} \leq e, \quad k = 1, \dots, n; \quad w_{ij} \geq 0, \quad j = 1, \dots, p \quad (2)$$

To avoid excessively low weights, the proportion constraint  $I_{ij}^{w_{ij}} \geq (\prod_{j=1}^p I_{ij}^{w_{ij}})^{0.05}$  was added, so that each basic indicator contributes at least 5% to the composite.

The indicators have different scales, thus they were preliminarily normalized as  $I_{ij} = 1 + \left( \frac{I_{ij} - m_j}{M_j - m_j} \right)$ , where  $m_j$  and  $M_j$  are empirical minimum and maximum of the  $j$ -th indicator, excepting those with negative polarity (unemployment, poverty, emissions and nitrogen balance) which were normalized as  $I_{ij} = 2 - \left( \frac{I_{ij} - m_j}{M_j - m_j} \right)$ .

Since the maximum score resulting from our composite indicator is  $e$  but the minimum can vary between 1 and  $e$ , the scores were rescaled as  $CI_i^* = \frac{CI_i - \min(CI)}{e - \min(CI)}$ , so that their range becomes  $[0, 1]$  and the results are easier to interpret.

### 3 Results and discussion

Figure 1 shows the trajectories of the scores resulted from our composite indicator. We see that the considered EU countries have an approximately increasing or stable trend of sustainability in the period 2004–2018, with the exception of Denmark, Greece, Ireland and Portugal, which show a definitely decreasing trend.

**Fig. 1** Trajectories of sustainability scores by country in the period 2004–2018.

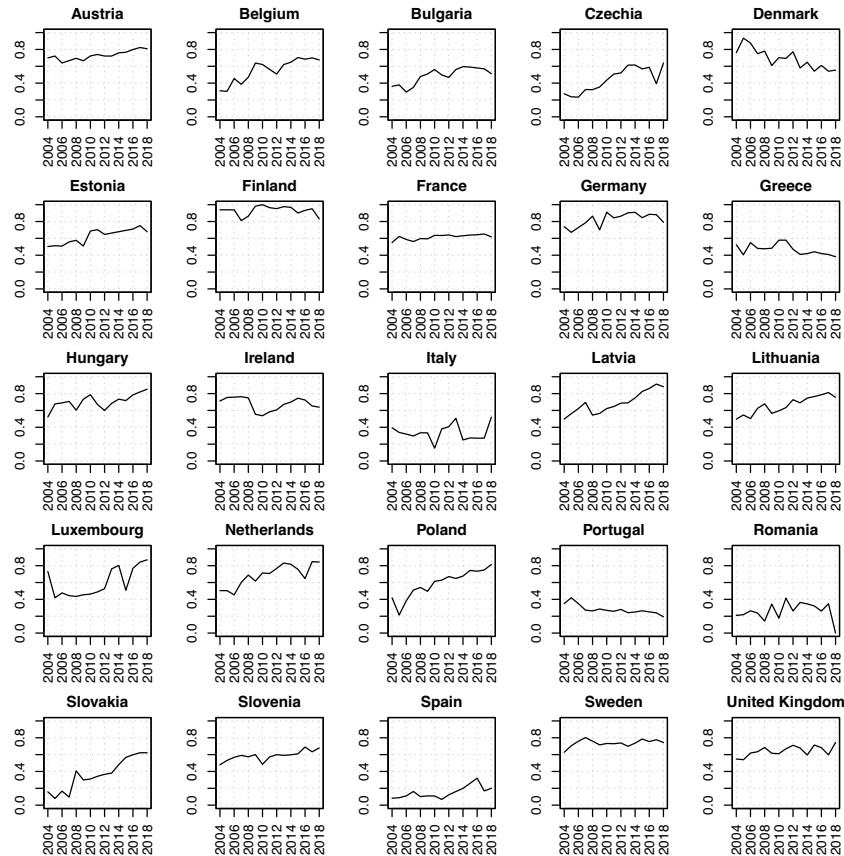
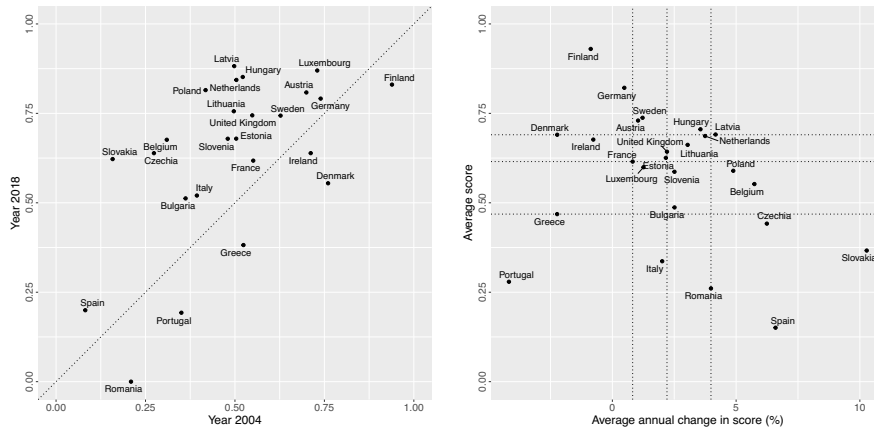


Figure 2, left panel, displays the change in score between 2004 and 2018, where it can be noted that Denmark has the highest rate of decrease, while Slovakia, Poland, Latvia, Belgium and Czechia have the highest rate of increase.

Figure 2, right panel, compares the average score of each country across the period 2004–2018 with the corresponding average annual change. We see that countries with average score above the third quartile are Finland, Germany, Austria, Sweden, Hungary and Denmark. However, excepting Hungary, these countries have an

A composite indicator to assess sustainability of agriculture in EU countries

**Fig. 2** Comparison of sustainability scores in 2004 and in 2018 (left panel), and of the average sustainability score of each country across the period 2004–2018 with its corresponding average annual change (right panel). Dotted lines in the right panel indicate quartiles.



average annual change below the median, with Denmark showing even a negative change. Instead, countries with average score below the first quartile are Czechia, Slovakia, Italy, Romania, Portugal and Spain. Interestingly, excepting Portugal, these countries have higher average annual growth than the ones with the highest average score, with Slovakia and Spain showing the highest average annual growth across all the considered countries.

#### 4 Concluding remarks

We have proposed a composite indicator to measure agricultural sustainability in EU countries, based on the geometric aggregation of twelve indicators through weights endogenously determined according to the BoD approach. Our proposal is innovative with respect to existing studies because we considered: (i) a broad set of indicators to cover the economic, social and environmental dimensions; (ii) a comprehensive set of countries (25 in total) to provide an exhaustive picture of agricultural sustainability in the EU; (iii) data over a long period (15 years) to explore not only the level of sustainability, but also its evolution in time.

The main limitation of our work relies in quality and availability of data, an issue affecting all the multidimensional assessments due to the practical difficulty of collecting reliable measurements on a large number of indicators. The longitudinal nature of our analysis entails a further complication, because publicly available time series are typically short and may present a number of missing values.

Future work will be directed towards the validation of our composite indicator. On one hand, a sensitivity analysis will be carried out to assess the impact of

different weighting and aggregation methods, in particular uniform weights (equal importance of the indicators) and arithmetic aggregation (full compensability). On the other hand, the decomposition into the contributions of the three sustainable dimensions will be addressed, with the purpose of determining the dimension-specific weights explicitly. Accounting for the hierarchical structure implied by the three dimensions is expected to improve interpretability and to limit the tendency of the BoD approach to attribute excessively low and high weights to some indicators.

## References

- [1] R. Cataldo, C. Crocetta, M. G. Grassia, N. C. Lauro, M. Marino and V. Voytsekhovska (2020). Methodological PLS-PM framework for SDGs system. *Social Indicators Research*, published: 20 January 2020.
- [2] S. E. Cristache, M. Vută, E. Marin, S. I. Cioacă and M. Vută (2018). Organic versus conventional farming: A paradigm for the sustainable development of the European countries. *Sustainability*, **10**: 4279.
- [3] B. Czyzewski, A. Matuszczak, A. Grzelak, M. Guth and A. Majchrzak (2020). Environmental sustainable value in agriculture revisited: How does Common Agricultural Policy contribute to eco-efficiency? *Sustainability Science*, **356**(43):1 44-165.
- [4] Food and Agriculture Organization (2013). Sustainability Assessment of Food and Agriculture Systems. Indicators. FAO, Rome, IT.
- [5] D. Hayati, Z. Ranjbar and E. Karami (2010). Measuring agricultural sustainability. In: Lichtfouse E. (ed.), *Biodiversity, Biofuels, Agroforestry and Conservation Agriculture*, pages 73-100. Springer, Cham, CH.
- [6] L. Latruffe, A. Diazabakana, C. Bockstaller, Y. Desjeux, J. Finn, E. Kelly, M. Ryan and S. Uthes (2016). Measurement of sustainability in agriculture: A review of indicators. *Studies in Agricultural Economics*, **118**, 123-130.
- [7] A. Nowak, A. Krukowski and M. Różańska-Boczula (2019). Assessment of sustainability in agriculture of the EU countries. *Agronomy*, **9**(12): 890.
- [8] J. Pretty (2008). Agricultural sustainability: Concepts, principles and evidence. *Philosophical Transactions of The Royal Society B*, **363**: 447-465.
- [9] M. Radovanovic and N. Lior (2017). Sustainable economic environmental planning in Southeast Europe: Beyond-GDP and climate change emphases. *Sustainable Development*, **25**(6): 580-594.
- [10] S. Terzi, A. Otoi, E. Grimaccia, M. Mazziotta and A. Pareto (2021). Open issues in composite indicators. A starting point and a reference on some state-of-the-art issues. Roma TrE-Press, Rome, IT.
- [11] P. Zhou, B. W. Ang and D. Q. Zhou (2010). Weighting and aggregation in composite indicator construction: a multiplicative optimization approach. *Social Indicator Research*, **96**: 169-181.

# Interval-Based Composite Indicators with a Triplex Representation: A Measure of the Potential Demand for the "Ristori" Decree in Italy.

*Indicatori compositi basati su dati ad intervallo facendo uso di una rappresentazione Triplex: la misurazione della potenziale domanda del decreto "Ristori" in Italia*

Carlo Drago<sup>1</sup>

**Abstract** In this work, we propose a new approach to constructing interval-based composite indicators based on the triplex representation. So, we measure the principal value of the indicator and simultaneously the value's uncertainty due to the different assumptions as different weightings associated how the indicator and their ranks can vary considering different assumptions or weights. The approach is useful not only on the construction of the composite indicators but also for a reliable interpretation of the results. The application shows the usefulness of the approach in detecting the regions higher the potential demand for economic support due to the Covid-19 emergency.

**Abstract** *In questo lavoro proponiamo un nuovo approccio alla costruzione di indicatori compositi basati su intervalli, basati sulla rappresentazione triplex. In questo caso siamo in grado di misurare non solo il valore principale dell'indicatore e l'incertezza del valore dovuta alla diversa assunzione come differenti ponderazioni associate ma anche come l'indicatore e il loro rango possono variare considerando differenti assunzioni o pesi. Quindi l'approccio è utile non solo sulla costruzione degli indicatori compositi, ma anche su un'interpretazione affidabile dei risultati.*

---

<sup>1</sup> Carlo Drago, University "Niccolò Cusano", Via Don Carlo Gnocchi 3, 00166, Rome and NCI University, London. Northern & Shell Tower, 4 Selsdon Way, Isle of Dogs, London E14 9GL, United Kingdom. We thank an anonymous referee for their helpful suggestions.

*L'applicazione mostra l'utilità dell'approccio nell'analisi della domanda di sostegno economico ("ristori") dovuto all'emergenza Covid-19 in Italia.*

**Key words:** interval data, symbolic data, composite indicators

## 1 The Statistical Problem

There are many situations where it is challenging to evaluate many complex phenomenon measures using a single value. In this case, we lose relevant information because we are aggregating the data in a unique measure. The use of interval-valued data can be an essential approach for retaining all the knowledge about the phenomenon under investigation and avoiding information loss (Billard and Diday 2003 Gioia 2009).

In this case, the complex information is related to the problematic measurement of a composite indicator. In this way, there can be a relevant issue on constructing the composite indicator, which contains some subjective choices. For instance, the weighting of the composite indicators can be differently identified. Following Greco et al. (2019) it is necessary to transform the subjective aspect of weighting the metrics more manageable and, most transparent, the most important thing to consider. After the composite indicator construction, it is necessary to evaluate the robustness and the sensitivity of the choices done on the final results of the composite indicator (Saisana et al. 2005).

In this respect, before considering any statistical analysis, it is crucial to passing through this stage of symbolic data analysis to define new formats for the data (Billard Diday 2003). These data should maintain all the complex information of the data. In this context, the use of composite indicators using interval data can be a possibility to represent the uncertainty accordingly on the data.

## 2 Interval-Based Composite Indicators Using a Triplex representation

To better represent a composite indicator based on a single value, it is possible simultaneously to consider different values obtained considering different initial weights or assumptions. The approach proposed by Drago (2017 and 2019 see also Gatto and Drago 2020) is to consider the possible different weighting schemes randomly as factors in the construction of the composite indicator. Using a Monte-Carlo simulation, obtain the different composite indicators associated with each assumption on the weighting scheme. Finally, the interval-based composite indicator can be constructed considering the different values obtained by the Monte-Carlo simulation and the center, the lower and the upper bound, representing extreme

Interval-Based Composite Indicators using the Triplex Representation scenarios from the construction of the possible composite indicators. Following Bertrand and Goupil (2000) and Billard (2010) given:

$$Y_u = [a_u, b_u], u = 1, \dots, n \quad (1)$$

Where  $a_u$  is the lower bound of the interval and  $b_u$  is the upper bound. We have the mean of the intervals:

$$\bar{Y} = \frac{1}{2n} \sum_{u \in E} (b_u + a_u) \quad (2)$$

And also variance:

$$S^2 = \frac{1}{3n} \sum_{u \in E} (b_u^2 + b_u a_u + a_u^2) - \frac{1}{4n^2} [\sum_{u \in E} (b_u + a_u)]^2, 1, \dots, n \quad (3)$$

The use of the interval data is crucial because it retains the original data's relevant information, considering all the different weighting schemes.

In this respect, the data's triplex representation (Apostolatos et al. 1968, Nickel 1969) can be useful to consider the most relevant assumption of the different composite indicators constructed. The intervals and the triplex representation both have their arithmetic, allowing the statistical analysis of the different sets of composite indicators obtained (Williamson 1989 Gioia and Lauro 2005).

So we are explicitly considering in our interval-based composite indicator a value of the interval obtained using the classical approach in constructing the composite indicator. The advantage is twofold: to retain the information of the original composite indicator, and also to allow to evaluate the two radii, as sub-intervals constructed by considering the difference between the upper bound and the original composite indicator value, but also the difference between the most likely result of the composite indicator and the lower bound. These measures can be essential in interpreting the results because it allows measurement of the variability of the measures considering extreme scenarios.

### 3 Application Results

To analyze the proposed approach, we consider an application based on the analysis of the interest and potential demand for "Decreto Ristori" as a policy measure to

sustain the Italian Economy from the Covid-19 Emergency. In this sense, the policy measure was the DL 28 October 2020, n. 137, containing "Additional urgent measures in the field of health protection, support to workers and businesses, justice and safety, connected to the epidemiological emergency from Covid-19" (V.A. 2021 MEF 2021). This decree was very relevant because it was significant for many businesses in Italy, so the potential demand or interest measured as google queries on the period 2020-10-25- 2021-2-28 (essential to consider also the additional renewals of the decree, which identify different measure and different policies). The interest measured in this sense is absorbing because it allows identifying relevant Italian zones special in need of economic sustain.

More specifically, we collected the queries by using Google Trends (see Google Trends 2021). The different queries are: "decreto ristori", "decreto ristori bis", "decreto ristori ter", "decreto ristori quater", "decreto ristori quinquies", "fondo perduto decreto ristori", "decreto ristori codici ateco" and "finally bonus decreto ristori". The queries are performed for the Italian territory. Obtained the different values for the regions, we consider the Monte-Carlo simulation using different weights schemes. So we obtain the visual representation of the simulation as a heat map in figure 1. We can then compute the different intervals for the region, considering the lower bound, the equal weights scenario, and the upper bound. Using the triplex representation, we can compute the two radii. The two radii (lower radius L.R. and upper radius U.R.) are useful for obtaining a measure of uncertainty related to the composite indicator's computation. Finally, we compute the prototype as the "interval average," considering all the different intervals computed for each region. The mean computation is obtained using the triplex arithmetic (Williamson 1989 and Gioia and Lauro 2005). The final results are shown in Table 1. The result clearly shows some regions in Italy in which it is higher the potential demand and interest for the decree "Ristori." In particular, these regions are Calabria, Campania, and Apulia (Puglia). In these regions, it is higher the measured level of poverty (see Drago 2020). The remarkable difference is the Sicily which does not show a higher the potential demand and interest for the decree "Ristori." This result can suggest that the zones in more severe economic difficulty tend to demand high for economic sustain. It is possible using this indicator to identify them. The results can also be interpreted considering the radii of the difference between the main scenario and the upper and the radii. These differences can have a relevant applicative result because it shows in the apposite ranking how different from the main scenario, the potential demand in some query (related to a single topic). It could be essential to consider the main scenario (or the center of the interval in a different context considering the data as a genuine interval representation) and how can be different the lower and the upper bounds.

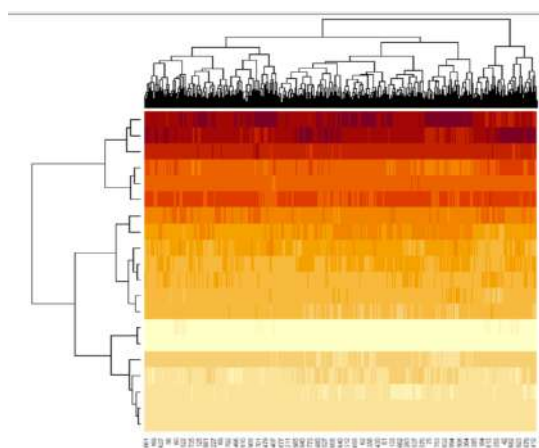
**Table 1:** Interest and the Potential Demand for the Decree "Ristori" over the period 2020-10-25 and the period 2021-2-28

Region	LB	EW	UB	L.R. (1)	U.R. (2)	rank (1)	rank (2)
Calabria	0.85	1.21	1.52	0.36	0.31	1	1



Interval-Based Composite Indicators using the Triplex Representation

Campania	0.92	1.15	1.44	0.23	0.29	6	4
Apulia	0.71	0.92	1.18	0.21	0.26	8	6
Tuscany	0.48	0.65	0.86	0.17	0.21	11	9
Lombardy	0.25	0.49	0.77	0.24	0.28	5	5
Sicily	0.37	0.49	0.61	0.12	0.12	15	15
Piedmont	0.09	0.31	0.55	0.22	0.24	7	7
Umbria	-0.12	0.14	0.38	0.25	0.24	3	8
Basilicata	-0.24	0.01	0.31	0.25	0.30	4	3
Abruzzo	-0.24	-0.06	0.11	0.19	0.16	9	12
Lazio	-0.20	-0.09	0.02	0.12	0.10	16	18
Marche	-0.33	-0.15	0.03	0.18	0.19	10	11
Emilia-Romagna	-0.34	-0.20	-0.01	0.15	0.19	12	10
Veneto	-0.56	-0.44	-0.31	0.12	0.13	14	13
Molise	-0.85	-0.55	-0.24	0.30	0.31	2	2
Liguria	-0.70	-0.58	-0.47	0.11	0.11	17	17
Sardinia	-0.69	-0.60	-0.53	0.08	0.08	20	20
Friuli-Venezia Giulia	-0.76	-0.66	-0.56	0.10	0.10	19	19
Trentino-Alto Adige/South Tyrol	-1.11	-1.00	-0.88	0.11	0.12	18	16
Aosta	-1.19	-1.05	-0.93	0.15	0.12	13	14
<b>Prototype computed</b>	-0.18	0	0.19	0.18	0.19		



**Figure 1:** Heat map of the different simulations performed (in the columns) for each ranked region (in the row)

## 4 Conclusions

The results in this work show the possibility of using the triplex representation in the interval to interpret the different intervals in a useful way to economic policy. The main scenario (or the center of the interval) can analyze the potential demand or the interest in the topic. The obtained relevant result is the identification of the relevant zones with a higher potential demand for economic support. There are relevant advantages to using the triplex representation to interpret the result of the analysis. First of all, we are explicitly taking into account the main scenario (in this sense, the equal weight scenario includes the main result of a classical composite indicator outcome).

Secondarily, this representation allows the computing of the lower and the upper radii to represent the limitations of the classical approach identifying where there is stronger the request and interest for some queries or in the application for different economic supports.

## References

1. Apostolatos, N., Kulisch, U. Krawczyk, R. Lortz, B. Nickel K. -Wipperman, W. (1968) The Algorithmic Language Triplex-Algol 60. *Numerische Mathematik* 11 (1968)
2. Billard L. (2010) *Symbolic Data Analysis: Basic Statistics*. Compstat August 2010.
3. Billard, L., & Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, 98(462), 470-487.
4. Drago, C. (2017) Interval Based Composite Indicators. FEEM Working Paper No. 42.2017
5. Drago, C. (2019) Decomposition of the Interval Based Composite Indicators by Means of Biclustering. CLADAG 2019 12-th Scientific Meeting Classification and Data Analysis Group Cassino, September 11 – 13, 2019 Book of Short Papers/Abstracts
6. Drago, C. (2020) The Analysis and the Measurement of Poverty: An Interval Based Composite Indicator Approach. Preprints 2020, 2020120695 (doi: 10.20944/preprints202012.0695.v1).
7. Gatto, A., & Drago, C. (2020). Measuring and modeling energy resilience. *Ecological Economics*, 172, 106527.
8. Gioia F. (2009) Portfolio Selection Models with Interval Data. *Statistica Applicata* Vol. 21, n. 1, 2009 43
9. Gioia, F., & Lauro, C. N. (2005). Basic statistical methods for interval data. *Statistica applicata*, 17(1), 75-104.
10. Google Trends (2021) Data source: Google Trends (<https://www.google.com/trends>) page accessed the 28/2/2021
11. Greco, S., Ishizaka, A., Tasiou, M., & Torrioni, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, 141(1), 61-94.
12. MEF (2021) Decreto ristori: le misure a favore di chi è in difficoltà <https://www.mef.gov.it/covid-19/Decreti-ristori-le-misure-a-favore-di-chi-e-in-difficolta/> Page accessed the 28/2/2021
13. Nickel K. (1969) Triplex-Algol and its Applications, in *Topics in Interval Analysis*, E. Hansen, ed., Oxford University Press, Oxford.
14. Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2), 307-323.
15. V. A. (2021) Decreto Ristori <https://www.pmi.it/tag/decreto-ristori#:~:text=Il%20cosiddetto%20Decreto%20Ristori%2C%20ovvero,ottobre%20e%20pubblicat%20in%20Gazzetta> page accessed the 28/2/2021
16. Williamson, R. C. (1989). Probabilistic arithmetic (Doctoral dissertation, University of Queensland).

## 4.19 New developments in GLM theory and applications

# Variational inference for the smoothing distribution in dynamic probit models

## *Inferenza variazionale per la distribuzione di lisciamiento nell'ambito della regressione di modelli probit dinamici*

Augusto Fasano and Giovanni Rebaudo

**Abstract** Recently, [1] provided closed-form expressions for the filtering, predictive and smoothing distributions of multivariate dynamic probit models, leveraging on unified skew-normal distribution properties. This allows to develop algorithms to draw independent and identically distributed samples from such distributions, as well as sequential Monte Carlo procedures for the filtering and predictive distributions, allowing to overcome computational bottlenecks that may arise for large sample sizes. In this paper, we briefly review the above-mentioned closed-form expressions, mainly focusing on the smoothing distribution of the univariate dynamic probit. We develop a variational Bayes approach, extending the partially factorized mean-field variational approximation introduced by [2] for the static binary probit model to the dynamic setting. Results are shown for a financial application.

**Abstract** Recentemente, [1] hanno derivato le espressioni esatte delle distribuzioni di filtraggio, predittive e di lisciamiento nell'ambito del modello probit dinamico multivariato, sfruttando le proprietà delle distribuzioni normali asimmetriche. Questo ha permesso di sviluppare algoritmi per ottenere campioni indipendenti ed identicamente distribuiti da tali distribuzioni oltre a schemi di campionamento sequenziale per le distribuzioni di filtraggio e predittive, permettendo di superare i problemi computazionali che possono sorgere per dimensioni campionarie elevate. Nel presente articolo, riassumiamo tali risultati con particolare riferimento alla distribuzione di lisciamiento nell'ambito del modello probit univariato dinamico e sviluppiamo un approccio variazionale, estendendo al caso dinamico il metodo di inferenza variazionale introdotto da [2] per il modello probit statico. I risultati sono presentati in un'applicazione finanziaria.

**Key words:** Dynamic Probit Model, Hidden Markov Model, Variational Inference, Unified Skew-Normal Distribution

---

Augusto Fasano  
Collegio Carlo Alberto and ESOMAS Department, C.so Unione Sovietica 218/bis, Turin  
e-mail: [augusto.fasano@unito.it](mailto:augusto.fasano@unito.it)

Giovanni Rebaudo  
Department of Statistics and Data Sciences, the University of Texas at Austin, TX 78712-1823  
e-mail: [giovanni.rebaudo@austin.utexas.edu](mailto:giovanni.rebaudo@austin.utexas.edu)

### 1 Introduction

Let us consider a hidden Markov model with binary observations  $y_t \in \{0; 1\}^m$ ,  $t = 1, \dots, n$  and state variables  $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_{1t}, \dots, \boldsymbol{\theta}_{pt})^\top \in \mathbb{R}^p$ . Adapting the notation proposed in, e.g., [3] to our setting, we aim to develop a novel variational approximation for the joint smoothing distribution in the following dynamic probit model

$$p(y_t | \boldsymbol{\theta}_t) = \Phi((2y_t - 1)\mathbf{x}_t^\top \boldsymbol{\theta}_t), \tag{1}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N_p(\mathbf{0}, \mathbf{W}_t), \quad t = 1 \dots, n, \tag{2}$$

with  $\boldsymbol{\theta}_0 \sim N_p(\mathbf{a}_0, \mathbf{P}_0)$ ,  $\{\boldsymbol{\varepsilon}_t\}_{t \geq 1} \perp \{\boldsymbol{\theta}_t\}_{t \geq 0}$  and  $\boldsymbol{\varepsilon}_{t_1} \perp \boldsymbol{\varepsilon}_{t_2}$  for any  $t_1 \neq t_2$ . In (1),  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution, while  $\mathbf{x}_t$  represents a known covariate vector. In the following, we set  $\mathbf{a}_0 = \mathbf{0}$  to ease notation.

Representation (1)–(2) can be alternatively obtained via the dichotomization of an underlying state-space model for the univariate Gaussian time series  $z_t \in \mathbb{R}$ ,  $t = 1, \dots, n$ , which is regarded, in econometric applications, as a set of time-varying utilities. Indeed, adapting classical results from static probit regression [4], model (1)–(2) is equivalent to

$$y_t = \mathbb{1}(z_t > 0) \tag{3}$$

$$z_t = \mathbf{x}_t^\top \boldsymbol{\theta}_t + \eta_t, \quad \eta_t \sim N(0, 1), \tag{4}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N_p(\mathbf{0}, \mathbf{W}_t), \quad t = 1 \dots, n, \tag{5}$$

having  $\boldsymbol{\theta}_0 \sim N_p(\mathbf{0}, \mathbf{P}_0)$ ,  $\{\eta_t\}_{t \geq 1} \perp \{\boldsymbol{\varepsilon}_t\}_{t \geq 1}$  and  $\eta_{t_1} \perp \eta_{t_2}$  for any  $t_1 \neq t_2$ .

As is clear from model (4)–(5), if  $\mathbf{z}_{1:n} = (z_1, \dots, z_n)^\top$  were observed, then, calling  $\boldsymbol{\theta}_{1:n} = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_n^\top)^\top$ , the joint smoothing density  $p(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n})$  and its marginals  $p(\boldsymbol{\theta}_t | \mathbf{z}_{1:n})$ ,  $t \leq n$ , could be obtained in closed-form by Gaussian-Gaussian conjugacy [3]. However, in (3)–(5) only a dichotomized version  $y_t$  of  $z_t$  is available. Thus the smoothing density is  $p(\boldsymbol{\theta}_{1:n} | \mathbf{y}_{1:n})$ , which is not Gaussian.

### 2 Literature review

In the context of static probit regression, [5] recently proved that the posterior distribution for the probit coefficients, under either Gaussian or unified skew-normal (SUN) [6] priors, is itself a SUN with parameters that can be derived in closed-form. Leveraging these findings, [1] showed that also in the more challenging multivariate dynamic probit setting, the filtering, predictive and smoothing densities of the state variables have SUN kernels. We recall that a random vector  $\boldsymbol{\theta} \in \mathbb{R}^q$  has SUN distribution,  $\boldsymbol{\theta} \sim \text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ , if its density function  $p(\boldsymbol{\theta})$  can be expressed as

$$\phi_q(\boldsymbol{\theta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_h[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\theta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta}]}{\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})},$$

where the covariance matrix  $\boldsymbol{\Omega}$  of the Gaussian density  $\phi_q(\boldsymbol{\theta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$  can be decomposed as  $\boldsymbol{\Omega} = \boldsymbol{\omega} \bar{\boldsymbol{\Omega}} \boldsymbol{\omega}$ , i.e. by rescaling the correlation matrix  $\bar{\boldsymbol{\Omega}}$  via the diagonal

---

**Algorithm 1:** Independent and identically distributed sampling from  $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ 


---

- [I] Sample  $\mathbf{U}_{0:1:n}^{(1)}, \dots, \mathbf{U}_{0:1:n}^{(R)}$  independently from a  $N_{p \cdot n}(\mathbf{0}, \tilde{\boldsymbol{\Omega}}_{1:n} - \boldsymbol{\Delta}_{1:n} \boldsymbol{\Gamma}_{1:n}^{-1} \boldsymbol{\Delta}_{1:n}^\top)$ .
  - [II] Sample  $\mathbf{U}_{1:1:n}^{(1)}, \dots, \mathbf{U}_{1:1:n}^{(R)}$  independently from a  $N_n(\mathbf{0}, \boldsymbol{\Gamma}_{1:n})$ , truncated below  $\mathbf{0}$ .
  - [III] Compute  $\boldsymbol{\theta}_{1:n}^{(1)}, \dots, \boldsymbol{\theta}_{1:n}^{(R)}$  via  $\boldsymbol{\theta}_{1:n}^{(r)} = \boldsymbol{\omega}_{1:n}(\mathbf{U}_{0:1:n}^{(r)} + \boldsymbol{\Delta}_{1:n} \boldsymbol{\Gamma}_{1:n}^{-1} \mathbf{U}_{1:1:n}^{(r)})$  for each  $r$ .
- 

scale matrix  $\boldsymbol{\omega} = (\boldsymbol{\Omega} \odot \mathbf{I}_q)^{1/2}$ , with  $\odot$  denoting the element-wise Hadamard product. See [6] for additional details on the SUN distribution.

From now on,  $\boldsymbol{\Omega}$  will actually denote the covariance matrix of the zero-mean normally distributed vector  $\boldsymbol{\theta}_{1:n}$ . Even though this might seem an abuse of notation with respect to the SUN density above, we show that this matrix actually coincides with the second parameter of the SUN joint smoothing density reported in Theorem 1 below. By the recursive formulation (2), we have that  $\boldsymbol{\theta}_{1:n}$  is normally distributed thanks to closure properties of Gaussian random variables with respect to linear transformations, while  $\boldsymbol{\Omega}$  shows the following block structure. Calling  $\mathbf{G}_l^t = \mathbf{G}_t \cdots \mathbf{G}_l$ ,  $l \leq t-1$ ,  $\boldsymbol{\Omega}$  is formed by  $(p \times p)$ -dimensional blocks  $\boldsymbol{\Omega}_{[tl]} = \text{var}(\boldsymbol{\theta}_t) = \mathbf{G}_1^t \mathbf{P}_0 \mathbf{G}_1^{t\top} + \sum_{l=2}^t \mathbf{G}_l^t \mathbf{W}_{l-1} \mathbf{G}_l^{t\top} + \mathbf{W}_t$ , for  $t = 1, \dots, n$ , and  $\boldsymbol{\Omega}_{[tl]}^\top = \text{cov}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_l) = \mathbf{G}_{l+1}^t \boldsymbol{\Omega}_{[ll]}$ , for  $t > l$ . As a direct consequence of Theorem 2 in [1] adapted to the simpler model (1)-(2), the following theorem holds.

**Theorem 1.** *Under model (1)–(2), the joint smoothing distribution has the form*

$$(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \sim \text{SUN}_{p \cdot n, n}(\mathbf{0}, \boldsymbol{\Omega}_{1:n}, \boldsymbol{\Delta}_{1:n}, \mathbf{0}, \boldsymbol{\Gamma}_{1:n}),$$

with  $\boldsymbol{\Omega}_{1:n} = \boldsymbol{\Omega}$ ,  $\boldsymbol{\Delta}_{1:n} = \tilde{\boldsymbol{\Omega}} \boldsymbol{\omega} \mathbf{D}^\top \mathbf{s}^{-1}$ ,  $\boldsymbol{\Gamma}_{1:n} = \mathbf{s}^{-1} (\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^\top + \mathbf{I}_n) \mathbf{s}^{-1}$ , where  $\mathbf{D}$  is an  $n \times (p \cdot n)$  block-diagonal matrix having block entries  $\mathbf{D}_{[tl]} = (2y_t - 1) \mathbf{x}_t^\top$ ,  $t = 1, \dots, n$ ,  $\mathbf{s} = [(\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^\top + \mathbf{I}_n) \odot \mathbf{I}_n]^{1/2}$  and  $\mathbf{I}_n$  defines the  $n$ -dimensional identity matrix.

By Theorem 1 and the additive representation of the SUN [6], we can get the following probabilistic characterization, which can be used to draw i.i.d. samples from the smoothing distribution as in Algorithm 1:

$$(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \stackrel{d}{=} \boldsymbol{\omega}_{1:n}(\mathbf{U}_{0:1:n} + \boldsymbol{\Delta}_{1:n} \boldsymbol{\Gamma}_{1:n}^{-1} \mathbf{U}_{1:1:n}),$$

with  $\mathbf{U}_{0:1:n} \sim N_{p \cdot n}(\mathbf{0}, \tilde{\boldsymbol{\Omega}}_{1:n} - \boldsymbol{\Delta}_{1:n} \boldsymbol{\Gamma}_{1:n}^{-1} \boldsymbol{\Delta}_{1:n}^\top)$ , while  $\mathbf{U}_{1:1:n}$  is distributed according to a  $N_n(\mathbf{0}, \boldsymbol{\Gamma}_{1:n})$  truncated below  $\mathbf{0}$ . From this representation, we see that the most computationally demanding part of drawing i.i.d. samples is sampling from an  $n$ -variate truncated Gaussian (point [II] of Algorithm 1). Although recent results [7] allow efficient simulation in small-to-moderate time series, this i.i.d. sampler might become computationally impractical for longer time series. In this paper, we propose a variational approximation for the smoothing distribution to overcome such computational issues. This approximation is based on methods developed by [2], which we extend here to the dynamic setting.

### 3 Variational approximation for the smoothing distribution

[2] recently introduced a partially factorized mean-field variational Bayes (PFM-VB) approximation for static probit models, which allows to perform approximate posterior inference without incurring in computational issues arising from the i.i.d. sampling. See [5] for details. Such a procedure has also been extended to categorical observations [8], providing notable approximation accuracy, especially in high dimensional settings. In this section, we adapt such results to develop a variational procedure for approximate inference on the smoothing distribution in dynamic probit models. Adapting [2], our PFM-VB procedure aims at providing a tractable approximation for the joint posterior density  $p(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n} | \mathbf{y}_{1:n})$  of the states vector  $\boldsymbol{\theta}_{1:n}$  and the partially observed variables  $\mathbf{z}_{1:n} = (z_1, \dots, z_n)^\top$ , within the PFM class of partially factorized densities  $\mathcal{Q}_{\text{PFM}} = \{q_{\text{PFM}}(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n}) : q_{\text{PFM}}(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n}) = q_{\text{PFM}}(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n}) \prod_{i=1}^n q_{\text{PFM}}(z_i)\}$ . Differently from classic mean-field (MF) approximations, this enlarged class does not assume independence among  $\boldsymbol{\theta}_{1:n}$  and  $\mathbf{z}_{1:n}$ , thus providing a more flexible family of approximating densities. This form of factorization allows to remove the main computationally demanding part of  $p(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n} | \mathbf{y}_{1:n})$ , while retaining part of its dependence structure. Indeed, adapting [9] and letting  $\mathbf{V} = (\boldsymbol{\Omega}^{-1} + \mathbf{X}^\top \mathbf{X})^{-1}$  and  $\mathbf{X}$  a  $n \times (p \cdot n)$  block-diagonal matrix with block-diagonal entries  $\mathbf{X}_{[t]} = \mathbf{x}_t^\top$ ,  $t = 1, \dots, n$ , the joint density  $p(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n} | \mathbf{y}_{1:n})$  under the augmented model (3)-(5) can be factorized as  $p(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n} | \mathbf{y}_{1:n}) = p(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n}) p(\mathbf{z}_{1:n} | \mathbf{y}_{1:n})$ , where  $p(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n}) = \phi_{p \cdot n}(\boldsymbol{\theta}_{1:n} - \mathbf{V} \mathbf{X}^\top \mathbf{z}_{1:n}; \mathbf{V})$  and  $p(\mathbf{z}_{1:n} | \mathbf{y}_{1:n}) \propto \phi_n(\mathbf{z}_{1:n}; \mathbf{I}_n + \mathbf{X} \boldsymbol{\Omega} \mathbf{X}^\top) \prod_{i=1}^n \mathbb{1}[(2y_i - 1)z_i > 0]$  denote the densities of a  $p \cdot n$ -variate Gaussian and an  $n$ -variate truncated normal, respectively. From this, we can note that the main source of intractability comes from the truncated normal density.

The optimal PFM-VB solution  $q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n}) \prod_{i=1}^n q_{\text{PFM}}^*(z_i)$  within  $\mathcal{Q}_{\text{PFM}}$  is the minimizer of the Kullback–Leibler (KL) divergence [10]

$$\begin{aligned} \text{KL}[q_{\text{PFM}}(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n}) || p(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n} | \mathbf{y}_{1:n})] &= \mathbb{E}_{q_{\text{PFM}}(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n})} [\log q_{\text{PFM}}(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n})] \\ &\quad - \mathbb{E}_{q_{\text{PFM}}(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n})} [\log p(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n} | \mathbf{y}_{1:n})]. \end{aligned}$$

Alternatively, it is possible to obtain  $q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n}) \prod_{i=1}^n q_{\text{PFM}}^*(z_i)$  by maximizing the evidence lower bound  $\text{ELBO}[q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n})]$ . See [11] and [2] for details. Adapting Theorem 2 in [2], it is immediate to obtain the following theorem.

**Theorem 2.** *Under model (1)-(2), the KL divergence between  $q_{\text{PFM}}(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n}) \in \mathcal{Q}_{\text{PFM}}$  and  $p(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n} | \mathbf{y})$  is minimized at  $q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n}) \prod_{i=1}^n q_{\text{PFM}}^*(z_i)$  with*

$$\begin{aligned} q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n}) &= p(\boldsymbol{\theta}_{1:n} | \mathbf{z}_{1:n}) = \phi_{p \cdot n}(\boldsymbol{\theta}_{1:n} - \mathbf{V} \mathbf{X}^\top \mathbf{z}_{1:n}; \mathbf{V}), \\ q_{\text{PFM}}^*(z_i) &= \frac{\phi(z_i - \mu_i^*; \sigma_i^{*2})}{\Phi[(2y_i - 1)\mu_i^* / \sigma_i^*]} \mathbb{1}[(2y_i - 1)z_i > 0], \quad \sigma_i^{*2} = (1 - \mathbf{X}_{[i]} \mathbf{V} \mathbf{X}_{[i]}^\top)^{-1}, \end{aligned} \quad (6)$$

where  $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_n^*)^\top$  solves the system  $\mu_i^* - \sigma_i^{*2} \mathbf{X}_{[i]} \mathbf{V} \mathbf{X}_{[-i]}^\top \bar{\mathbf{z}}_{-i}^* = 0$ ,  $i = 1, \dots, n$ , with  $\mathbf{X}_{[-i]}$  denoting the matrix  $\mathbf{X}$  with the  $i$ th row  $\mathbf{X}_{[i]}$  removed, while  $\bar{\mathbf{z}}_{-i}^*$  is an  $n - 1$

---

**Algorithm 2:** CAVI algorithm for  $q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n}) = q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n} \mid \mathbf{z}_{1:n}) \prod_{i=1}^n q_{\text{PFM}}^*(z_i)$

---

[I] For each  $i = 1, \dots, n$ , set  $\sigma_i^{*2} = (1 - \mathbf{X}_{[i]} \mathbf{V} \mathbf{X}_{[i]}^\top)^{-1}$  and initialize  $\bar{z}_i^{(0)} \in \mathbb{R}$ .

[II] **for**  $t$  from 1 until convergence of  $\text{ELBO}[q_{\text{PFM}}^{(t)}(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n})]$  **do**

**for**  $i$  from 1 to  $n$  **do**

        [II.1] Set  $\mu_i^{(t)} = \sigma_i^{*2} \mathbf{X}_{[i]} \mathbf{V} \mathbf{X}_{[-i]}^\top (\bar{z}_1^{(t)}, \dots, \bar{z}_{i-1}^{(t)}, \bar{z}_{i+1}^{(t)}, \dots, \bar{z}_n^{(t-1)})^\top$ .

        [II.2] Set  $\bar{z}_i^{(t)} = \mu_i^{(t)} + (2y_i - 1) \sigma_i^* \phi(\mu_i^{(t)} / \sigma_i^*) \Phi[(2y_i - 1) \mu_i^{(t)} / \sigma_i^*]^{-1}$ .

**Output:**  $q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n})$  as in Theorem 2.

---

vector obtained by removing the  $i$ th element  $\bar{z}_i^* = \mu_i^* + (2y_i - 1) \sigma_i^* \phi(\mu_i^* / \sigma_i^*) \Phi[(2y_i - 1) \mu_i^* / \sigma_i^*]^{-1}$ ,  $i = 1, \dots, n$ , from the vector  $\bar{\mathbf{z}}^* = (\bar{z}_1^*, \dots, \bar{z}_n^*)^\top$ .

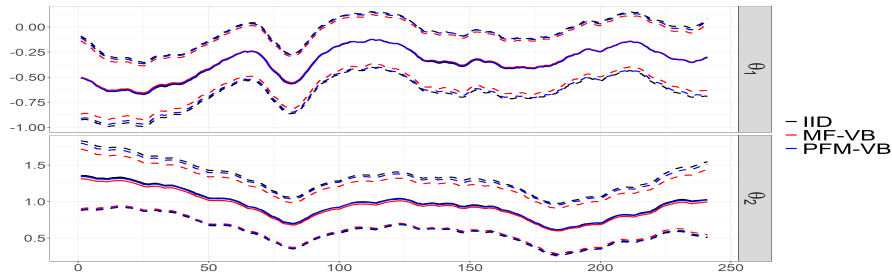
Algorithm 2 shows how to obtain  $q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n} \mid \mathbf{z}_{1:n}) \prod_{i=1}^n q_{\text{PFM}}^*(z_i)$  via a coordinate ascent variational inference (CAVI) algorithm that iteratively optimizes each  $\mu_i^*$ , keeping the rest fixed [11]. In addition to retaining part of the dependence structure of the true posterior, the PFM-VB solution also admits closed-form moments, as shown in Corollary 1 below, whose proof can be found in [2]. If more complex functionals are desired, they can be easily computed via Monte Carlo integration, since, exploiting (6), in order to get i.i.d. samples from  $q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n}, \mathbf{z}_{1:n})$  it is sufficient to draw values from  $p \cdot n$ -variate Gaussians and univariate truncated normals, avoiding the computational issues of the truncated multivariate normals in Algorithm 1.

**Corollary 1.** Let  $q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n}) = \mathbb{E}_{q_{\text{PFM}}^*(\mathbf{z}_{1:n})}[q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n} \mid \mathbf{z}_{1:n})]$ , then  $\mathbb{E}_{q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n})}(\boldsymbol{\theta}_{1:n}) = \mathbf{V} \mathbf{X}^\top \bar{\mathbf{z}}^*$  and  $\text{var}_{q_{\text{PFM}}^*(\boldsymbol{\theta}_{1:n})}(\boldsymbol{\theta}_{1:n}) = \mathbf{V} + \mathbf{V} \mathbf{X}^\top \text{diag}[\sigma_1^{*2} - (\bar{z}_1^* - \mu_1^*) \bar{z}_1^*, \dots, \sigma_n^{*2} - (\bar{z}_n^* - \mu_n^*) \bar{z}_n^*] \mathbf{X} \mathbf{V}$ , where  $\bar{z}_i^*$ ,  $\mu_i^*$  and  $\sigma_i^*$ ,  $i = 1, \dots, n$  are defined as in Theorem 2.

## 4 Financial application

We illustrate the performance of the variational approximation derived in Section 3 on a financial application considering a dynamic probit regression for the daily opening directions of the French CAC40 stock market index from January 4th, 2018 to December 28th, 2018, for a total of  $n = 241$  observations. In this study,  $y_t = 1$  if the opening value of the CAC40 on day  $t$  is greater than the corresponding closing value in the previous day, and  $y_t = 0$  otherwise. We consider two covariates: the intercept and the opening direction of the NIKKEI225, regarded as binary covariates  $\xi_t$ . Since the Japanese market opens before the French one,  $\xi_t$  is available before  $y_t$  and, hence, provides a valid predictor for each day  $t$ . Thus, with reference to model (1)-(2),  $p = 2$  and  $\mathbf{x}_t = (1, \xi_t)^\top$ . Moreover, we take  $\mathbf{W}_t = \text{diag}(0.01, 0.01)$  for every  $t$  and  $\mathbf{P}_0 = \text{diag}(3, 3)$ . See [1] for details on the hyperparameters' setting. The extent of the quality of the PFM-VB approximation is displayed in Figure 1. There, we plot  $\mathbb{E}[\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}]$  and  $\mathbb{E}[\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}] \pm \sqrt{\text{var}[\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}]}$ , estimated with  $10^4$  samples from the i.i.d. sampler, with the PFM-VB solution, exploiting Corollary 1, and with a mean-field variational Bayes (MF-VB) approximation, where independence among  $\boldsymbol{\theta}_{1:n}$  and  $\mathbf{z}_{1:n}$  is enforced, by adapting [12] to the current setting. We observe that the





**Fig. 1**  $\mathbb{E}[\theta_{1:n} | \mathbf{y}_{1:n}]$  (—) and  $\mathbb{E}[\theta_{1:n} | \mathbf{y}_{1:n}] \pm \sqrt{\text{var}[\theta_{1:n} | \mathbf{y}_{1:n}]}$  (- - -) for the i.i.d. sampler, the MF-VB algorithm and the PFM-VB solution.

PFM-VB approximation—differently from the MF-VB—almost perfectly matches the quantities of interest of the smoothing distribution. To better understand the improvements of PFM-VB over MF-VB, the average absolute difference in the estimated means of  $\theta_{1t}$  and  $\theta_{2t}$ ,  $t = 1, \dots, 241$ , with respect to the ones obtained with the i.i.d. sampler are 0.003 and 0.008 for the PFM-VB and 0.009 and 0.031 for the MF-VB, respectively. Considering the average difference of the log-standard-deviations, we obtain 0.04 and 0.05 for the PFM-VB, while these values equal 0.14 and 0.16 for the MF-VB, showing a much higher overshrinkage towards 0. Finally, the PFM-VB solution allows to compute the desired moments in only 1.1 seconds, similar to MF-VB, showing a much lower computational time than the i.i.d. sampler, which requires 115.4 seconds. Code to produce Figure 1 and additional outputs are available at the following link: <https://github.com/augustofasano/Dynamic-Probit-PFMVB>.

**Acknowledgements** The authors wish to thank Daniele Durante for carefully reading a preliminary version of this manuscript and providing insightful comments.

## References

1. Fasano, A., Rebaudo, G., Durante, D., and Petrone, S.: A closed-form filter for binary time series. arXiv:1902.06994, (2019)
2. Fasano, A., Durante, D., and Zanella, G.: Scalable and accurate variational Bayes for high-dimensional binary regression models. ArXiv:1911.06743, (2019)
3. Petris, G. Petrone, S., and Campagnoli, P.: Dynamic Linear Models with R. Springer, (2009)
4. Albert, J. H. and Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc., **88**, 669–679 (1993)
5. Durante, D.: Conjugate Bayes for probit regression via unified skew-normal distributions. Biometrika, **106**, 765–779 (2019)
6. Arellano-Valle, R. B. and Azzalini, A.: On the unification of families of skew-normal distributions. Scand. J. Stat., **33**, 561–574 (2006)
7. Botev, Z. I.: The normal law under linear restrictions: simulation and estimation via minimax tilting. J. R. Stat. Soc. Series B Stat. Methodol., **79**, 125–14 (2017)
8. Fasano, A. and Durante, D.: A class of conjugate priors for multinomial probit models which includes the multivariate normal one. arXiv:2007.06944, (2020)
9. Holmes, C. C. and Held, L.: Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Anal., **1**, 145–168 (2006)
10. Kullback, S. and Leibler, R. A.: On information and sufficiency. Ann. Stat., **22**, 79–86 (1951)
11. Blei, D. M., Kucukelbir, A., and McAuliffe, J. D.: Variational inference: A review for statisticians. J. Am. Stat. Assoc., **112**, 859–877 (2017)
12. Consonni, G. and Marin, J.M.: Mean-field variational approximate Bayesian inference for latent variable models. Computational Statistics & Data Analysis, **52**, 790–798 (2007)

# Interpretability and interaction learning for logistic regression models

## *Interpretabilità ed interaction learning per modelli di regressione logistica*

Nicola Rares Franco<sup>1</sup>, Michela Carlotta Massi<sup>1,2</sup>, Francesca Ieva<sup>1,2,3</sup>, Anna Maria Paganoni<sup>1,2,3</sup>

**Abstract** Logistic Regression (LR) is a renowned statistical model that finds use in many contexts, from medical fields to social sciences. Contrary to black-box models, one of the key features of LR is interpretability. However, it can be difficult to preserve this latter property when the covariates affect the response variable through high-order interactions. Here, we address this problem in the case of categorical predictors, and we propose a few approaches for identifying and including interaction terms in LR models, with possible benefits both in performance and interpretability. We then test the methodology on simulated data and discuss future developments.

**Abstract** *La Regressione Logistica (LR) è un noto modello statistico che trova applicazioni in svariati contesti, dall'ambito medico a quello delle scienze sociali. A differenza dei modelli black-box, una delle caratteristiche della LR è l'interpretabilità. Tuttavia, può essere difficile preservare quest'ultima proprietà quando l'influenza delle covariate sulla variabile risposta è caratterizzata da interazioni di ordine elevato. Nel presente lavoro, affrontiamo questa problematica nel caso di predittori categorici, proponendo alcuni approcci per l'identificazione e l'inclusione delle interazioni nei modelli LR, con possibili benefici nelle prestazioni e nell'interpretabilità. Testiamo quindi la metodologia attraverso uno studio di simulazione e discutiamo dei possibili sviluppi futuri.*

**Key words:** interaction learning, interpretability, logistic regression, categorical data

---

<sup>1</sup>MOX Laboratory, Math Department, Politecnico di Milano, Milan, Italy

<sup>2</sup>CADS-Center for Analysis, Decisions and Society, Human Technopole, Milan, Italy

<sup>3</sup>CHRP-National Center for Healthcare Research and Pharmacoepidemiology, University of Milano-Bicocca, Milan, Italy

## 1 Introduction

Model interpretability can be an essential requirement, especially in contexts such as decision making and healthcare [1]. For this reason, when it comes to modelling the probability of an event in terms of other variables (or *covariates*), Logistic Regression (LR) remains a popular choice that is often preferred over other alternatives, such as decision trees, support vector machines and neural networks [3]. However, things quickly get complicated in the presence of high-order interaction effects. In fact, high-order terms are notoriously harder to interpret. Also, their inclusion typically results in very large models that can suffer from reduced reliability [10].

Here, we focus on the case in which the covariates are categorical and their influence on the response variable involves complex interactions. We note that, despite not being the most general setting, this situation is frequently encountered in genomics and healthcare, see e.g. [6, 7].

## 2 Methodology

We first introduce some notation. Let  $Y$  be a binary random variable (r.v.), and let  $X_1, \dots, X_p$  be  $p$  categorical r.v.s, each admitting  $l_1, \dots, l_p$  levels respectively. We address the problem of modelling the function

$$f(x_1, \dots, x_p) := \mathbb{P}(Y = 1 \mid X_1 = x_1, \dots, X_p = x_p), \quad (1)$$

using LR. In particular, we focus on the case in which  $f$  is best described by interaction terms of its arguments. Without loss of generality, we assume each  $X_i$  takes values in  $\{1, \dots, l_i\} \subset \mathbb{N}$ . Let us formally introduce the collection of all interactions,

$$\mathcal{I}(X_1, \dots, X_p) := \left\{ \prod_{j \in J} \mathbf{1}_{\{x_j\}}(X_j) \mid J \subseteq \{1, \dots, p\}, 1 \leq x_j \leq l_j \right\}.$$

Note that, as the  $X_i$  are categorical, the interactions are actually prescribed in terms of the so-called dummy variables. In particular, each interaction describes a different combination of the covariates' levels. We further note that, with little abuse of notation,  $\mathcal{I}(X_1, \dots, X_p)$  also includes the primary effects (no interaction), as those are obtained when  $J$  is a singleton. Since  $|\mathcal{I}(X_1, \dots, X_p)| = (l_1 + 1) \cdot \dots \cdot (l_p + 1) \geq 2^p$ , the straight addition of all interaction effects to the model is often unsuited.

In order to overcome this drawback and recover the model interpretability, we proceed as follows. First, we propose a way of filtering the interaction terms so that a small but relevant pool is extracted. Then, we discuss how to include the interactions in the LR model so that the interpretability is preserved.

## 2.1 Interaction learning

We are given a dataset  $\{(x_{i,1}, \dots, x_{i,p}, y_i)\}_{i=1}^N$  consisting of  $N$  i.i.d. realizations of  $X_1, \dots, X_p$  and  $Y$ . We wish to explore the data in order to point out the elements in  $\mathcal{S}(X_1, \dots, X_p)$  that are actually useful for modelling (1). The latter is an open problem [4] that is sometimes referred to as *interaction learning*.

For each interaction  $T$ , let  $\{t_i\}_{i=1}^N$  be the corresponding observed values. Since all r.v.s in  $\mathcal{S}(X_1, \dots, X_p)$  are dichotomous, the empirical frequency of the event  $\{T = 1\}$  is equivalent to the empirical mean  $\bar{t} := \frac{1}{N} \sum_{i=1}^N t_i$ . To start, we discard all those interactions for which  $\bar{t} = 0, 1$ , as we cannot make any inference on them. Then, we may apply several criteria to further filter the list  $\mathcal{S}(X_1, \dots, X_p)$ . One way is to fix a threshold  $\delta > 0$  and remove all  $T$  for which  $\bar{t} < \delta$ . This corresponds to keeping only the frequent interactions, procedure that can be optimized by taking advantage from the flourishing literature of *frequent itemsets* in data mining [2, 5]. A further possibility is to limit this preliminary search to the collection of data for which  $y_i = 1$  (equiv.  $y_i = 0$ ). This is the approach adopted in [5], which is better suited in the presence of class imbalance, i.e. when  $\mathbb{P}(Y = 1) \ll \mathbb{P}(Y = 0)$ . Other methods such as Random Intersection Trees [9] may be used as well.

Whichever is the chosen approach, the first screening results in a list of candidate interactions  $\mathcal{S}_0 \subseteq \mathcal{S}(X_1, \dots, X_p)$ . Typically,  $L_0 := |\mathcal{S}_0|$  is very large and a further filtering is needed. For this purpose, one may rank the interactions in  $\mathcal{S}_0$  according to a suitable metric, such as: the effect size on the response variable (logworth); the odds-ratio or some quantity derived from it (e.g. the absolute value of the logarithm, as in [5]); importance measures (e.g. using random forests), and so on. The candidates list can be then sorted in descending order,  $\mathcal{S}_0 = \{T_k\}_{k=1}^{L_0}$ , with  $T_1$  being the most relevant. The idea is to exploit the ranking in order to extract a sublist  $\mathcal{S}_* \subset \mathcal{S}_0$  with  $L_* := |\mathcal{S}_*| \ll L_0$ . The sublist should be chosen in a way that avoids redundancy and favors the model generalizability. For this reason, we desist from letting  $\mathcal{S}_* = \{T_k\}_{k=1}^{L_*}$ , as it is likely that the interactions ranked at the top are very *similar* one another. Instead, we propose the use of a *dissimilarity measure*. The latter is a map  $d : \mathcal{S}(X_1, \dots, X_p) \times \mathcal{S}(X_1, \dots, X_p) \rightarrow \mathbb{R}$  that quantifies the "difference" between two given interactions. An example can be found below.

$$d(T, S) := \max\{|T|, |S|\} - \text{ReLU}(|T \cdot S| - |T| - |S|), \quad (2)$$

where  $|\cdot|$  is the interaction *length*, i.e. the number of  $X_i$ 's that are involved in its definition, while  $\text{ReLU}(x) := \max\{x, 0\}$  is the so-called linear unit rectifier. By design, the dissimilarity measure in (2) returns larger values for interactions that: (i) have different order; (ii) refer to different levels of some common  $X_i$  (in fact, in this latter case  $|T \cdot S| = 0$ ). We exploit  $d$  in order to define  $\mathcal{S}_* := \{T_{k_i}\}_{i=1}^{L_*}$  as follows,

$$k_1 := 1, \quad k_{i+1} := \arg \max_k \min_{j=1, \dots, i} d(T_k, T_{k_j}),$$

where, in case more maximizers exist, smaller indexes are preferred (as they correspond to interactions with a higher ranking). The idea is that such construction

of  $\mathcal{S}_*$  should favor both relevance and diversity. Also, as  $L_*$  is user-specified, the sublist can be made as small as desired. The final LR model would then be,

$$\text{logit } \mathbb{P}(Y = 1 \mid X_1, \dots, X_p) = \beta_0 + \sum_{i=1}^{L_*} \beta_i T_{k_i}, \quad (3)$$

where the  $\beta_i$  are the model parameters. However, it is also possible to include the interactions in alternative ways that foster the model interpretability. We discuss those in the next subsection.

### 2.2 Increasing interpretability

In the case of categorical data, interactions are easier to interpret as they encode possible level combinations. Still, (3) becomes harder to visualize as soon as  $L_*$  is large. In fact, even though  $L_*$  is user-specified, it can happen that mild to large values are needed to ensure a suitable performance. As a remedy, we propose to further compress the information by clustering the interactions in  $\mathcal{S}_*$ . We shall detail two approaches. The first one, is to roughly compress the interactions using only two indexes, a risk index and a protection index. More precisely, for each  $T \in \mathcal{S}_*$ , let  $\text{OR}_T$  be the empirical odds-ratio associated to the pair  $(T, Y)$ . Let  $\mathcal{R} := \{T \in \mathcal{S}_* \mid \text{OR}_T > 1\}$  and  $\mathcal{P} := \{T \in \mathcal{S}_* \mid \text{OR}_T < 1\}$  be respectively the risk and protection interactions. Then, we define the risk index  $R$  and the protection index  $P$  as

$$R := \sum_{T \in \mathcal{R}} T, \quad P := \sum_{T \in \mathcal{P}} T.$$

The corresponding LR model now reads,

$$\text{logit } \mathbb{P}(Y = 1 \mid X_1, \dots, X_p) = \beta_0 + \beta_1 R + \beta_2 P. \quad (4)$$

Compared to (3), the above is far easier to understand: the model considers two opposite contributes, each synthesizing the effect of multiple level combinations.

As second approach we propose a less coarse reduction, where we keep separated all those interactions that refer to incompatible situations (i.e. they involve different levels of some common  $X_i$ ). To do so, let  $\mathcal{R}$  and  $\mathcal{P}$  be as before. We define  $\mathcal{R}_1 \subseteq \mathcal{R}$  as the largest collection of interactions that are mutually compatible (hence  $T \cdot S \neq 0$  for all  $T, S \in \mathcal{R}_1$ ). Iteratively, we let  $\mathcal{R}_{i+1} \subseteq \mathcal{R} \setminus \bigcup_{j=1}^i \mathcal{R}_j$  be the largest sublist with the same property. We define similarly the lists  $\mathcal{P}_i$ , so that  $\mathcal{R} = \mathcal{R}_1 \cup \dots \cup \mathcal{R}_r$  and  $\mathcal{P} = \mathcal{P}_1 \cup \dots \cup \mathcal{P}_p$ . Then, we construct the risk and protection indexes as

$$R_i := \sum_{T \in \mathcal{R}_i} T, \quad P_j := \sum_{T \in \mathcal{P}_j} T,$$

where  $i = 1, \dots, r$  and  $j = 1, \dots, p$ . The corresponding LR model is

**Fig. 1** Reference picture for the tic-tac-toe example. Cells are numbered from 1 to 9 starting from the bottom left. In the reported example  $X_1 = 5, O_1 = 4, X_2 = 7, O_2 = 3$ . The value of the target variable  $Y$  is only known at the end of the game.

7	X		
4	O	X	
1			O

$$\text{logit } \mathbb{P}(Y = 1 | X_1, \dots, X_p) = \gamma_0 + \sum_{i=1}^r \alpha_i R_i + \sum_{j=1}^p \beta_j P_j. \tag{5}$$

We interpret the above as a variant of (4), where the two contributes have been splitted to let incompatible situations weight differently.

### 3 Simulation study and results

We apply the approach in Section 2 to a dataset of  $N = 1000$  simulated tic-tac-toe games. The data concerns five variables, namely  $X_1, O_1, X_2, O_2$  and  $Y$ . Following the convention in Figure 1, the first four variables report the cells chosen by the two players during the first two turns; the target variable, instead, indicates whether player 2 won the game ( $Y = 1$ ) or not ( $Y = 0$ ). Within the observed data,  $x_{i,1} = 5$  for all  $i = 1, \dots, N$ , meaning that all games started with a cross at the center of the table. For this reason,  $X_1$  was ignored. In order to properly test the methodology, we split the data into a training and a test set (with 50:50 ratio). This means that all the procedures in Section 2, as well as the fitting of the models, are limited to the training set. The resulting LR model is then tested on the remaining data.

To identify the candidate set  $\mathcal{S}_0$ , we apply a frequency threshold of  $\delta = 0.1$  and only scan the minority class  $\{y_i = 1\}$ . For the filtering step, we operate separately on the lists  $\mathcal{R}_0 := \{T \in \mathcal{S}_0 | OR_T > 1\}$ ,  $\mathcal{P}_0 := \{T \in \mathcal{S}_0 | OR_T < 1\}$ , which we rank respectively by descending and ascending odds-ratio. As dissimilarity measure we use (2), while we set  $L_* = 15$  (to be counted twice, as we filter the two lists separately). We implement both (4) and (5) and compare them via Area Under the Curve (AUC). As benchmark, we also fit a LR model with no interactions terms.

Results are reported in Table 1. Both the proposed approaches outperform the benchmark model, reflecting the fact that interactions among the covariates strongly impact the target variable. In particular, model (4) appears to be the most appealing thanks to its higher interpretability (3 vs 17 fitted parameters). For comparison, we note that the LR model without interactions has to deal with 25 unknowns: one intercept and eight coefficients for each of the three predictors (recall that each  $X_i$  gives rise to 8 dummie variables, as  $X_1 = 5$  in every game and thus the level "5" is never attained). In this sense, models (4) and (5) provide an improvement both in terms of AUC and model interpretability.

**Table 1** Results for the tic-tac-toe example. No interactions = LR without interaction terms, Two-index = model in (4), Multi-index = model in (5). The three are compared in terms of Area Under the ROC Curve. Each model requires the fitting of a different number of parameters.

Model	Number of fitted parameters	AUC
No interactions	25	0.582
Two-index	3	0.776
Multi-index	17	0.803

## 4 Conclusions

We have presented a way of including high-order interactions in LR models so that interpretability is preserved. The numerical results are promising, and we believe that these out-of-the-box approaches can be of particular interest in real-life applications, such as healthcare. In fact, once fitted, these models are particularly easy to use and understand, even for non-experts such as medical doctors, clinicians, etc. A first application to genomic data can be found in [8], where the authors used model (4) to build polygenic risk scores that account for interactions among genetic loci. As future works, we would like to investigate the case of non-categorical covariates as well as other grouping criteria.

## Acknowledgements

This project has received funding under the ERA PerMed Cofund program (grant agreement No ERAPERMED2018-244, RADprecise - Personalized radiotherapy).

## References

1. Ahmad, M.A., Eckert, C., Teredesai, et al.: Interpretable Machine Learning in Healthcare. *IEEE Intelligent Informatics Bulletin*, **19**(1) (2018).
2. Ceddia, G., Martino, L.N., Parodi, A., et al.: Association rule mining to identify transcription factor interactions in genomic regions. *Bioinformatics*, **36**(4), 1007–1013 (2020).
3. Dreiseitl, S., Ohno-Machado, L.: Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.*, **35**(5-6), 352–359 (2002).
4. Lim, M., Hastie, T.: Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graph. Stat.*, **24**(3), 627–654 (2015).
5. Massi, M.C., Franco, N.R., Ieva, F., et al.: Learning High-Order Interactions via Targeted Pattern Search. *arXiv preprint*, arXiv:2102.12974 [cs.LG] (2020).
6. Massi, M.C., Gasperoni, F., Ieva, F., et al.: A Deep Learning approach for validating the effect of SNPs on late RT toxicity for prostate cancer patients. *Front. Oncol.*, **10**, 2033 (2020).
7. Moore, J.H.: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, **56**(1-3), 73–82 (2003).
8. Franco, N.R., Massi, M.C., Ieva, F., et al.: Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity. *Radiotherapy and Oncology*, In Press, DOI: 10.1016/j.radonc.2021.03.024 (2021).
9. Shah, R.D., Meinshausen, N.: Random Intersection Trees. *J. Mach. Learn. Res.*, **15**, 629–654 (2014).
10. Sur, P., Candès, E.J.: A modern maximum-likelihood theory for highdimensional logistic regression. *Proceedings of the National Academy of Sciences*, **116**(29), 14516–14525 (2019).

# Entropy estimation for binary data with dependence structures

## *Stima dell'entropia per dati binari con strutture di dipendenza*

Linda Altieri and Daniela Cocchi

**Abstract** We propose a new method for entropy estimation: instead of intervening with corrections on the global estimator as in previous approaches, we focus on improving the estimation of its components, i.e. the probabilities of the study variable. Our method allows to account for covariates, temporal and spatial dependence. Following a Bayesian approach, we estimate the posterior distribution of probabilities in a wide variety of situations. A posterior distribution for entropy is then derived, which may be synthesized as wished. When covariates are available, the entropy point estimator varies according to the covariate values; with temporal dependence, the estimated entropy is a curve; for spatial data, it can be displayed as an entropy smooth surface for the area under study.

**Abstract** *Questo lavoro propone un nuovo approccio alla stima dell'entropia: invece che correggere lo stimatore globale, come fatto in letteratura, miglioriamo la stima delle componenti dell'entropia, cioè le probabilità della variabile oggetto di studio. Il nostro metodo può tenere conto di covariate, dipendenza temporale e spaziale. Con un approccio Bayesiano, stimiamo le distribuzioni a posteriori delle probabilità su diversi scenari. Di conseguenza, si ricava una distribuzione a posteriori dell'entropia che si può sintetizzare a piacere. Quando ci sono covariate, la stima puntuale dell'entropia prende un valore differente per ogni valore delle covariate; con dipendenza temporale, lo stimatore per l'entropia è una curva; con dipendenza spaziale, lo stimatore è una superficie liscia sull'area di studio.*

**Key words:** Entropy estimation, CAR models, Bayesian logistic regression, binary variables, temporal dependence, spatial entropy, correlated data, covariates

---

Linda Altieri

Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy,  
e-mail: linda.altieri@unibo.it

Daniela Cocchi

Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy  
e-mail: daniela.cocchi@unibo.it



## 1 Introduction

Shannon's entropy is a successful measure in many fields, as it is able to synthesize several concepts in a single number: entropy, information, heterogeneity, surprise, contagion. The entropy of a categorical variable  $X$  is (Cover and Thomas, 2006)

$$H(X) = \sum_{i=1}^I \log p(x_i) \log \frac{1}{p(x_i)}. \quad (1)$$

In several case studies, the interest lies in describing and synthesizing data. This requires advanced computational tools and spatial correlation, if present, should be accounted for (see, e.g., Batty, 1976, Leibovici, 2009). When it comes to measuring the entropy of spatial data, an approach proposed and validated by recent papers (Altieri et al., 2018, 2019) allows to decompose entropy into a term quantifying the spatial information and a second term quantifying the residual heterogeneity. In other case studies, the aim lies in estimating the entropy of a phenomenon, i.e. in making inference rather than description. The standard approach (Paninski, 2003) relies on the so-called 'plug-in' estimator, which substitutes probabilities with observed relative frequencies in the computation of entropy. It is the non-parametric as well as the maximum likelihood estimator; it is biased, but performs well when  $I < \infty$  is known and independence is an acceptable assumption; otherwise, the estimator needs correction. The most popular proposals at this regard are in Miller (1955); Efron and Stein (1981); in more recent years, Zhang (2012) proposed a non-parametric solution with fast decaying bias. Under a Bayesian framework, the NSB estimator (Nemenman et al., 2002) has been proposed. In all proposals, independence among realizations is assumed and no auxiliary information is considered.

Two main limits concern entropy estimation. Firstly, the above mentioned works focus on improving the performance of the plugin estimator, instead of proposing alternatives. Secondly, no study faces the task of estimating entropy for variables presenting dependence on available covariates, spatial/temporal association or other types of dependence. Spatial entropy studies never consider the aspect of inference and simply use the relative frequencies as probabilities.

In this paper, we take a new perspective to entropy estimation that enriches both the area of entropy estimator proposals and the one of spatial entropy approaches. The focus is moved from the index  $H(X)$  itself to its components: since entropy is a deterministic function of the probability mass function (pmf) of the variable of interest, such pmf should be properly estimated. A Bayesian logistic regression approach allows to derive the variable pmf for binary data; then, the posterior distribution of entropy is straightforward. Point estimates, credibility intervals and other syntheses may be obtained via standard Bayesian inference. This approach can be extended to spatial and/or temporal data, with or without auxiliary information: in such cases, entropy is allowed to vary according to the model, and can be represented by a curve for continuous covariates or temporal dependence, or by a surface for spatial data.

## 2 Modelling probabilities

Let  $X$  be a binary response variable and consider a series of  $n$  realizations, which are independent given the distribution parameters, indexed by  $u = 1, \dots, n$ , each presenting a value  $x_u \in \{0, 1\}$ . For each  $X_u \sim \text{Ber}(p_u)$ , a logistic model for the parameter in a Bayesian framework can be defined as (Cressie, 1993)

$$\begin{aligned} \text{logit}(p_u) &= z_u' \beta + \phi_u \\ \phi &\sim \text{IGMRF}(0, \tau_\phi K) \end{aligned}$$

with parameters  $\beta \sim N(0, 10^{-6})$  and  $\tau_\phi \sim \text{Gamma}(a_\phi, b_\phi)$ . The vector  $z_u'$  contains the covariates associated to the  $u$ -th unit, while  $\phi_u$  is the random effect. The Intrinsic Gaussian Markov Random Field structure matrix  $K$  defines the type of dependence between the random effects, such as temporal or spatial correlation. In the simplest case of independence,  $K = I$ . In the present work, versions of the above model are estimated using the software R and INLA. Once parameter estimates of the logistic model are returned, we derive posterior estimates for the probabilities  $p_u$  and finally for entropy (1), where we obtain one entropy value linked to each realization  $u$ .

## 3 Entropy estimation across scenarios

In order to assess the validity of our methods both in absolute terms and in comparison to a selection of existing methods, we build a series of increasingly complex scenarios. For each scenario, we generate  $S = 1000$  replicated sequences of  $N = 2500$  observations from a binary  $X$ , where the probability of success  $p_u$  may vary across observations according to covariates or data dependence structures. As competitors, we choose the plug-in estimator (ML), the Miller-Madow correction (MM), Zhang's non-parametric estimator (Zh), the NSB Bayesian estimator with, respectively, a Laplace prior (BLapl) and with Jeffrey's prior (BJeff). Our estimator is abbreviated as BMB, i.e. Bayesian Model-Based. The comparison is run in terms of mean square error (MSE).

The independence scenario was tested with different probabilities of success from 0.05 to 0.50 (with a 0.05 step). We do not report results, as in such basic situation the BMB estimator is identical to the ML estimator. All methods return good estimates in terms of MSEs.

### 3.1 Covariates

We start with a binary covariate  $Z_1$  with values 0.1 and 2. Under the model  $\text{logit}(p_u) = z_1(u)\beta$ , with  $\beta = 1$ , two possible success probabilities are returned:

$p_u|_{z_1=0.1} = 0.88$  and  $p_u|_{z_1=2} = 0.52$ ; thus, they lead to two entropy values:  $H_{0.1} = 0.365$  and  $H_2 = 0.692$ . All existing methods ignore the presence of the covariate and return a similar single entropy value. Our estimator is able to include the influence the covariate and returns two entropy values. The resulting estimates and MSEs are in Table 1: as can be seen, our estimator has by far the best performance.

The same idea underlies a continuous covariate  $Z_2$  that takes a different value in

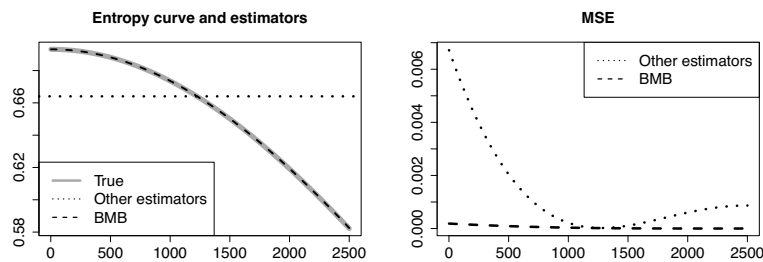
**Table 1** Data with binary covariate  $Z_1$  - Estimates and MSE

True	Estimates		MSE	
	Other estimators	BMB	Other estimators	BMB
$H = 0.365$	0.609	0.367	0.0597	0.0003
$H = 0.692$	0.609	0.692	0.0068	<0.0001

$[0, 1]$  for each observation. The success probability depends on the covariate (with  $\beta = 1$ ) and entropy now becomes a curve as in the left panel of Figure 1, where we see that the existing estimators do not grasp the variability of entropy and return a single value (flat dotted line), while our approach has a very good performance both in absolute terms and in comparison to the other ones (see the MSE curves in the right panel).

### 3.2 Temporal dependence

Temporal data are generated under a ARIMA(1,1,0) process, with the constraint that probabilities must be between 0 and 1. The model for the BMB estimator includes an intercept and an AR(1) temporal effect (determined by the matrix  $K$ ). Results are in Figure 2 and again show the superiority and flexibility of our approach wrt previous proposals, even if the model chosen for estimation is simpler wrt the generating process. Should a single number for entropy estimation be desired, our approach is still preferable: one could choose the posterior mean or median as point estimate for entropy, and the resulting MSE is still smaller compared to the existing methods.



**Fig. 1** Data with continuous covariate  $Z_2$  - Estimates and MSE

### 3.3 Spatial dependence

In order to account for spatial dependence, two spatial configurations are considered with different autocorrelation strengths: a compact and a multicluster scenario are built with Thomas processes (a Poisson cluster process). The chosen model for BMB includes an intercept and a standard RW2d spatial effect. Results are shown in Figure 3. Our proposal captures the behaviour of entropy in these last scenarios too, and returns a smaller MSE wrt other estimator, which all have similar results.

## 4 Concluding remarks

The present work proposes a novel approach to entropy estimation that is a substantial step forward in all cases where auxiliary information may affect entropy and independence is an unrealistic assumption: it is the first entropy estimator that allows to exploit all the available information. Our model-based approach is very flexible and may be complicated as wished. In the simple case of independence, our estimator is a well-performing alternative to the available methods. When data have a more complex structure, we produce a more suitable output: in the case of covariates, en-

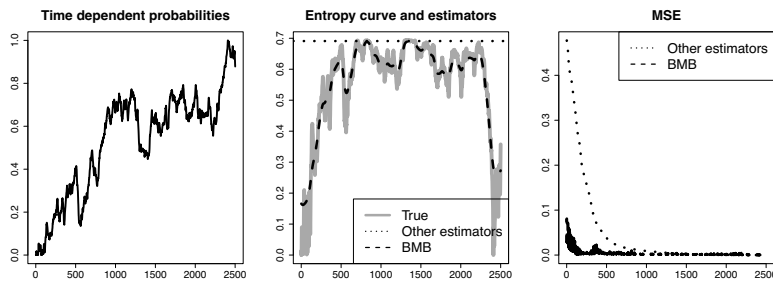


Fig. 2 Temporal data - Temporal  $p$  series, entropy estimates and MSE

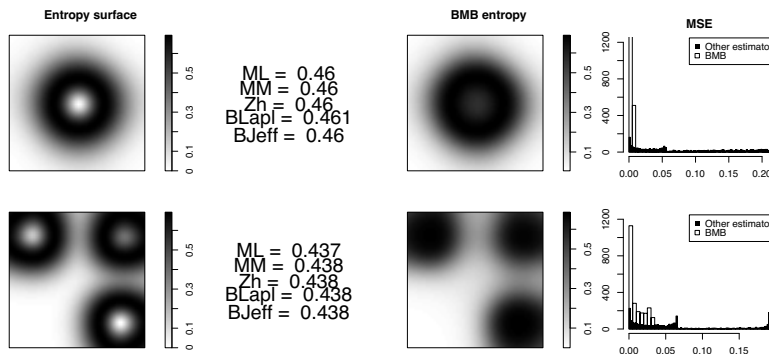


Fig. 3 Spatial data - generating entropy, literature estimators, BMB entropy and MSE; top panels: compact scenario, bottom panels: multicluster scenario

trophy varies according to the covariate values; in the case of temporal dependence, entropy is estimated as a smooth curve; for spatial data, entropy is a two-dimensional surface over the observation area. The results of the study show that in all deviations from independence our approach has a very good performance, grasping the entropy structure even when we choose to estimate with a simpler model than the generating process. Evaluation of the MSE wrt a selection of standard approaches identifies our proposal as the best performing one.

Our proposal is suitable for ecological and environmental data: an example comes from biodiversity studies, where the presence and the amount of species may depend on several factors, such as environmental covariates, spatial location, temporal structures. Traditional Shannon's entropy is only based on the relative frequencies of the observed species, while recent ecological studies attempt at finding a relationship between entropy and environmental factors or temporal/spatial effects, which can be explored with our approach. Further examples regard studies of the heterogeneity of land use and urban expansion, or of natural phenomena such as earthquakes and wildfires, whose entropy may be studied in relation to complex underlying spatial structures. An extension to multinomial data for such applications is currently under development.

## References

- Altieri, L., D. Cocchi, and G. Roli (2018). A new approach to spatial entropy measures. *Environmental and Ecological Statistics* 25(1), 95–110.
- Altieri, L., D. Cocchi, and G. Roli (2019). Advances in spatial entropy measures. *Stochastic Environmental Research and Risk Assessment* 33.
- Batty, M. (1976). Entropy in spatial aggregation. *Geographical Analysis* 8, 1–21.
- Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory. Second Edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Cressie, N. A. C. (1993). *Statistics for spatial data (rev. ed.)*. New York, Wiley.
- Efron, B. and C. Stein (1981). The jackknife estimate of variance. *Annals of statistics* 9, 586–596.
- Leibovici, D. G. (2009). *Defining spatial entropy from multivariate distributions of co-occurrences*. Berlin, Springer: In K. S. Hornsby et al. (eds.): COSIT 2009, Lecture Notes in Computer Science 5756, pp 392-404.
- Miller, G. (1955). *Note on the bias of information estimates*. Glencoe, IL free press: In H. Quastler (ed.) Information Theory in psychology II-B, pp. 95-100.
- Nemenman, I., F. Shafee, and W. Bialek (2002). *Entropy and inference, revisited*. Cambridge, MA: MIT Press: In T. G. Dietterich, S. Becker, and Z. Ghahramani (eds.) Advances in neural information processing, 14.
- Paninski, L. (2003). Estimation of entropy and mutual information. *Journal of Neural Computation* 15, 1191–1253.
- Zhang, Z. (2012). Entropy estimation in Turing's perspective. *Journal of Neural Computation* 24, 1368–1389.

# A Comparison of Some Estimation Methods for the Three-Parameter Logistic Model

## *Un confronto di Alcuni Metodi di Stima per il Modello Logistico con Tre Parametri*

Michela Battauz and Ruggero Bellio

**Abstract** The three-parameter logistic model is frequently used to account for guessing in multiple-choice items. However, this model is affected by estimation difficulties due to the weak identifiability of its parameters. In this note, a Bayesian approach is compared to two different shrinkage estimation approaches, given by a bias-reduction method based on the adjustment of the score function and by penalized likelihood estimation, respectively. The results of some experiments suggest that the regularization performed by the Bayesian estimation is similar to that of the penalized likelihood method, at least for the typical sample sizes for which the model is employed in applications.

**Abstract** *Il modello logistico con tre parametri è frequentemente usato per tener conto della possibilità di indovinare la risposta giusta nelle domande a risposta multipla. Tuttavia, questo modello incontra difficoltà di stima a causa della debole identificabilità dei parametri. In questa nota si confronta un approccio Bayesiano con un approccio di regolarizzazione frequentista. Più specificatamente, si considerano il metodo di riduzione della distorsione basato su un aggiustamento della funzione punteggio e una stima basata su di una funzione di verosimiglianza penalizzata. I risultati suggeriscono che la regolarizzazione operata dall'approccio Bayesiano è per molti versi simile a quella prodotta dalla stima di massima verosimiglianza penalizzata, soprattutto per le tipiche dimensioni campionarie per cui il modello viene utilizzato nelle applicazioni.*

**Key words:** Bayesian estimation, Bias Reduction, Item Response Theory, Penalized Likelihood, Regularization, Shrinkage Estimation.

---

Michela Battauz

University of Udine - Department of Economics and Statistics, via Tomadini 30/A - Udine (Italy),  
e-mail: michela.battauz@uniud.it

Ruggero Bellio

University of Udine - Department of Economics and Statistics, via Tomadini 30/A - Udine (Italy)  
e-mail: ruggero.bellio@uniud.it

## 1 Introduction

Item Response Theory models are used to analyze the data collected through tests or questionnaires. An area where these models have found large application is that of educational assessment. When the students' abilities are measured using multiple choice items, the probability of giving a correct response is positive even for examinees with extremely low ability levels, due to the possibility of guessing the right answer. The Three-Parameter Logistic (3PL) model is particularly suited to take this aspect into account. However, the weak identifiability of the parameters of the model results in large variability of the estimates and in convergence problems in the numerical maximization of the log-likelihood function [8]. Some proposals to overcome this issue have been made in [1, 2], where it is argued that some sort of regularization technique is called for, and some likelihood-based approaches for shrinkage estimation are proposed. In particular, here we focus on the proposals made in [2], consisting in the application of the bias reduction methodology proposed by [3], and of a penalized maximum likelihood estimation similar in nature to that made in [6] for a different IRT model.

The aim of this note is to compare these proposals to a Bayesian approach, consisting in a hierarchical model with weakly informative prior specification, as endorsed in [4]. In the next section the 3PL model is introduced and the estimation methods of interest are briefly reviewed. The performances of the various approaches are then compared using a real data set, supplemented with a small scale simulation study. The note ends with some concluding remarks.

## 2 Model and Methods

Let  $Y_{ij}$  be the binary response of person  $i$  to item  $j$ . In the 3PL model the probability of a correct response depends on the subject ability and some item parameters, and a convenient parameterization for it is given by

$$P(Y_{ij} = 1 | \beta_{1j}, \beta_{2j}, \beta_{3j}, \theta_i) = \pi_{ij} = F(\beta_{3j}) + \{1 - F(\beta_{3j})\} F\{\beta_{1j} + \beta_{2j} \theta_i\}, \quad (1)$$

where  $F$  is the logistic function

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}. \quad (2)$$

The parameter  $F(\beta_{3j})$  is called *guessing*, as it represents the probability of a correct response for examinees at extremely low levels of ability  $\theta_i$ .

Assuming that the responses are independent given the latent variable  $\theta_i$ , for which a standard normal distribution is assumed, by integrating out the random abilities it is possible to obtain the log-likelihood function for the item parameters

$$\ell(\beta) = \sum_{i=1}^n \log \int_{\mathbb{R}} \prod_{j=1}^J \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \phi(\theta_i) d\theta_i, \tag{3}$$

which is commonly used to obtain the parameter estimates in a frequentist approach. As already mentioned, the resulting Maximum Likelihood Estimator (MLE) is affected by weak identifiability of the item parameters, with occasional outlying values of the estimated  $\beta_3$  even for large sample sizes.

The frequentist proposals made in [2] are given in the following. The bias reduction method of [3] adds a term to the score function  $\mathbf{S}(\beta) = \partial \ell(\beta) / \partial \beta$  in order to obtain an estimator that has asymptotically smaller bias than that of the maximum likelihood estimator, achieving some regularization as a side effect of the procedure [7]. The method employs the estimating equation

$$\mathbf{S}^*(\beta) = \mathbf{S}(\beta) + \mathbf{A}(\beta) = 0, \tag{4}$$

where  $\mathbf{A}(\beta)$  is an adjustment term.

A different approach consists in adding a penalty term to the log-likelihood function. As proposed in [2], the ridge-type penalized log-likelihood function is defined as follows

$$\ell_p(\beta) = \ell(\beta) - \lambda \sum_{j < k} (\beta_{3j} - \beta_{3k})^2, \tag{5}$$

where the penalty forces the guessing parameters toward a common value. The selection of the tuning parameter  $\lambda$  can be performed using cross-validation, information criteria or following an empirical Bayes approach.

For implementing a full Bayesian approach, we need to supplement the model with some prior distributions for the item parameters. Here we assume a trivariate normal distribution for the three parameters of a given item

$$(\beta_{1j}, \log(\beta_{2j}), \beta_{3j})^\top \sim N(\mu, \Sigma), \tag{6}$$

followed by hyperpriors assumed for  $\mu$  and  $\Sigma$ . Similarly to [5] we adopt conjugate priors for the multivariate normal (see also [4], p. 73), namely

$$\Sigma \sim \text{Inv-Wishart}_{v_0}(I), \tag{7}$$

$$\mu | \Sigma \sim N(\mu_0, \Sigma/k_0), \tag{8}$$

where  $I$  is the identity matrix, and with the hyperprior parameters chosen to achieve weakly informative priors. Note that a natural choice for  $\mu_0$  consists in a three-dimensional vector with the first two elements equal to zero, and the third one equal to  $F^{-1}(1/m)$ , being  $m$  the number of alternatives in multiple choice items.



### 3 A Real-Data Example

The methods presented in this paper were applied to the same dataset employed by [2]. This is a test composed of 14 multiple-choice items in mathematics, which was part of the final exam at the third year of vocational high school in Italy, and all the 3,843 students who was administered this test were included in the sample. Though the sample is not small, the regularization carried out by shrinkage estimation is apparent, hence it is suitable for an illustration. The analysis has been made using the R software, and actually the dataset is included in the R package `S3PL` associated to [2] and available at <https://github.com/micbtz/S3PL>.

The `Stan` probabilistic programming language [9] was employed for Bayesian estimation in R, resulting in a rather simple implementation; see also the online *case study* about the 2PL model available at the software website (<https://mc-stan.org>). For the normal-inverse-Wishart distribution we set  $k_0 = 5$  and  $\nu_0 = 5$ , as in [5]. The two likelihood-based estimates reported here correspond to (4) and (5), with the tuning parameter for the latter obtained by repeated 10-fold cross validation.

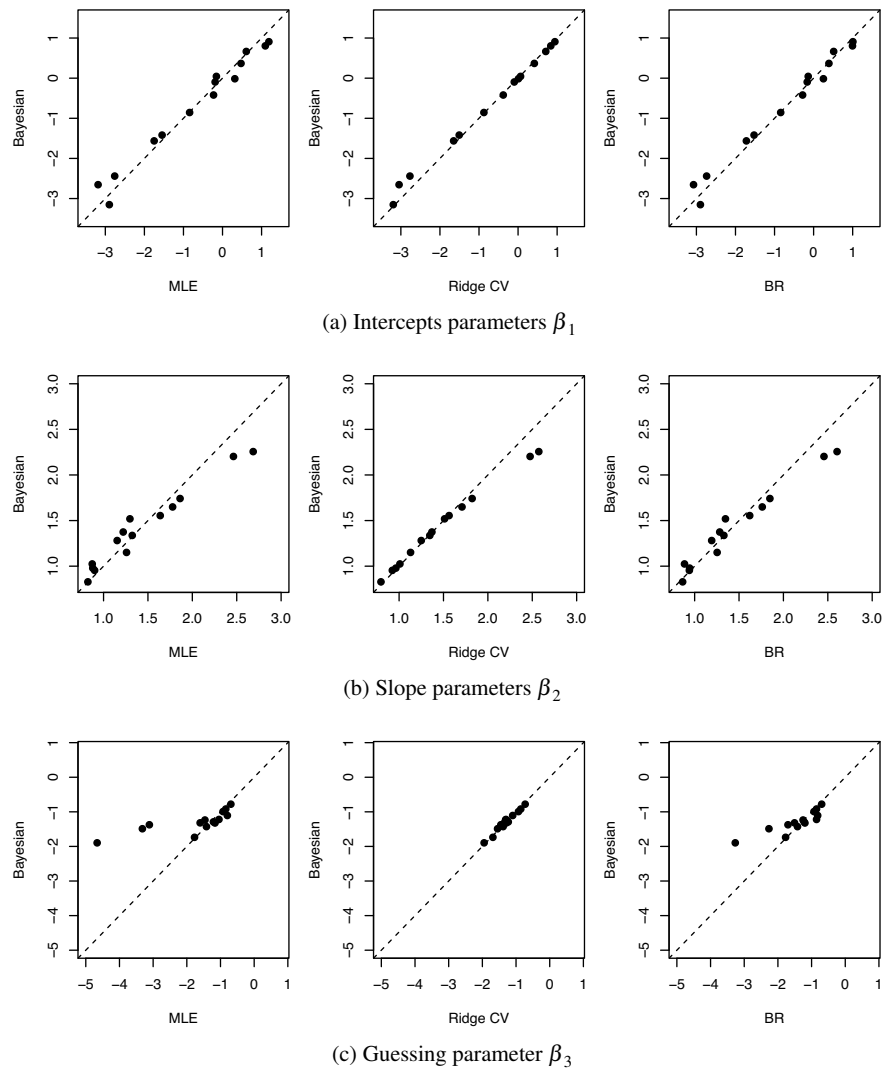
Figure 1 compares the three sets of frequentist parameter estimates to the Bayesian ones, given by the posterior means. The shrinkage of the Bayesian estimates of the guessing parameters compared to the MLE is noticeable, whereas the intercepts and the slopes do not exhibit such an effect. The penalized estimates (Ridge CV) are very similar to the Bayesian estimates, while the estimates obtained with the bias reduction (BR) method are closer to the MLE ones.

### 4 A Small Scale Simulation Study

A simulation study has been carried out for a more thorough comparison of the regularization properties of the various methods. For the likelihood-based methods extensive simulation results are reported in [2], and actually these estimation methods scale well for large sample sizes, since the individual abilities are integrated out. On the other hand, although the `Stan` implementation is rather smooth, the MCMC sampling is quite time consuming, since the sampling of the  $n$ -dimensional vector of abilities has to be carried out along with the model parameters. For such reason, the investigation was limited to just 100 simulated data sets for each sample size. The same simulation design of [2] was implemented, limiting the study to three sample sizes, namely  $n = 200$ ,  $n = 500$  and  $n = 1000$ , skipping instead the case with  $n = 5000$ . The root mean square error (RMSE), absolute bias (B) and mean absolute error (MAE) are reported in Table 1. These values are the average over all the items, and they can be compared with the results reported in table S3 in [2].

Even from such limited study, some useful suggestions emerge. The most striking fact is that for  $n = 1000$  the Bayesian regularization is very similar to that based on the penalized likelihood (5), and both methods provide a clear improvement over MLE, in line with the results of the previous section. For smaller sample sizes the Bayesian posterior means are not worse than any of the frequentist methods, and

A Comparison of Some Estimation Methods for the Three-Parameter Logistic Model



**Fig. 1** Comparison of the estimates of in the real-data example.

they actually seem preferable for  $n = 200$ . The additional gain seems to come from the fact that the Bayesian approach applies shrinkage to all the item parameters, differently from penalized maximum likelihood estimation, where the shrinkage on  $\beta_1$  and  $\beta_2$  derives from the direct shrinkage enforced on  $\beta_3$  coupled by the strong correlation existing among the three parameter estimates of the same item.

**Table 1** Simulation results

$n$	$\beta_1$			$\beta_2$			$\beta_3$		
	RMSE	B	MAE	RMSE	B	MAE	RMSE	B	MAE
200	0.31	0.29	0.26	0.26	0.17	0.18	0.56	0.68	0.52
500	0.27	0.26	0.24	0.21	0.15	0.15	0.53	0.64	0.49
1000	0.24	0.23	0.20	0.17	0.13	0.13	0.49	0.59	0.44

## 5 Conclusion and Ongoing Work

For the estimation of 3PL models, we strongly recommend practitioners to supplement maximum likelihood estimation with some form of regularization. This work suggests that Bayesian estimation represents another possible route for the task, that may be appealing to many. For routine applications, a further option could be given by a MAP approach, with abilities integrated out from the likelihood function and some shrinkage-inducing priors applied to the item parameters. The application of the resulting estimation method could actually be rather fast, and the aforementioned `S3PL` package offers some R and C++ routines for a straightforward implementation of this approach. Some further investigation may be of some interest, especially for possible implications for large samples of examinees. More specifically, it would be interesting to observe which methods perform better for sample sizes of thousands of examinees in terms of bias and root mean square error. Other extensions of some interest may concern the application of the same methodologies to other IRT models, such as models for ordinal responses.

## References

1. Battauz, M., Bellio, R.: A regularized estimation approach for the three-parameter logistic model. In: Abbruzzo, A., Brentari, E., Chiodi, M., Piacentino, D. (eds.) *Book of Short Papers SIS 2018*, pp. 224–232. Palermo, IT: 49th Scientific Meeting of the Italian Statistical Society (2018)
2. Battauz, M., Bellio, R.: Shrinkage estimation of the three-parameter logistic model. *Br. J. Math. Stat. Psychol.*, online first (2021)
3. Firth, D.: Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38 (1993)
4. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: *Bayesian Data Analysis* (Third Edition). Chapman and Hall/CRC, Boca Raton (2014)
5. Glas C.A.W., van der Linden W.J.: Computerized adaptive testing with item cloning. *Appl. Psychol. Meas.* **27**, 247–261 (2003)
6. Houseman, E.A., Marsit, C., Karagas, M., Ryan, L.M.: Penalized item response theory models: application to epigenetic alterations in bladder cancer. *Biometrics*, **63**, 1269–1277 (2007)
7. Kosmidis, I.: Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdiscip. Rev. Comput. Stat.*, **6**, 185–196 (2014)
8. Patz, R. J., Junker, B. W.: Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *J. Educ. Behav. Stat.* **24**, 342–366 (1999)
9. R Package `rstan`: the R interface to Stan. Stan Development Team, R package version 2.21.1 (2020).

# A statistical model to identify the price determinations: the case of Airbnb.

## *Un modello statistico per identificare le determinanti del prezzo: il caso Airbnb.*

Giulia Contu, Luca Frigau, Gian Paolo Zammarchi, Francesco Mola

**Abstract** We propose an indicator to estimate the effects of *transports, culture, crowd, managerial, accommodation* dimensions on price determination for Airbnb accommodations. The indicator is defined using a proportional odds model in two steps. In the first phase, we estimate for each accommodation the probability to belong to a specific category of the price moving from *very-low* to *very-high*. Then, we estimate the average concentration index to identify which is the price class more likely for each observation and which dimension can better explain the price. Then, we assign the price class to each observation using the median of the probabilities of the model's fitted values and identify the most significant dimension. Afterward, we aggregate the concentration index calculated for each observation for the district of *Trastevere* in Rome, with the aim to identify its most relevant dimensions. The results highlight a significant impact of the *managerial* and *accommodation* dimensions.

**Abstract** *Si propone un indicatore per stimare l'effetto delle dimensioni del trasporto, cultura, affollamento, management e caratteristiche dell'alloggio sulla determinazione del prezzo per gli alloggi Airbnb. L'indicatore è definito utilizzando un proportional odds model come spiegato nel Vector Generalized Additive Model. L'indicatore viene costruito in due fasi. Nella prima fase si stima la probabilità*

---

Giulia Contu

University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: giulia.contu@unica.it

Luca Frigau

University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: frigau@unica.it

Gian Paolo Zammarchi

University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: gp.zammarchi@unica.it

Francesco Mola

University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: mola@unica.it

*di ogni alloggio di appartenere ad una delle cinque categorie di prezzo che vanno da molto basso a molto alto. Successivamente, si stima l'indice di concentrazione medio per identificare la fascia di prezzo in cui ricade l'osservazione e quale dimensione possa spiegare meglio il prezzo. Poi, la fascia di prezzo viene assegnata a ciascuna osservazione utilizzando la mediana del vettore di probabilità e si identificano le dimensioni più significative. Successivamente, l'indice di concentrazione calcolato per ciascuna osservazione viene aggregato per il quartiere di Trastevere di Roma, con l'obiettivo di individuare la dimensione più rilevanti per il quartiere. I risultati hanno evidenziato un impatto significativo della dimensione management e caratteristiche dell'alloggio.*

**Key words:** proportional odds model, price determination, price dimensions, Airbnb accommodation, price indicator

## 1 Introduction

Airbnb is one of the most famous platforms where it is possible to book apartments, private and shared rooms. It has been founded in 2008 in San Francisco and has grown significantly over the years. It operates in more than 65,000 cities and 191 countries and sells millions of room nights to tourists and travelers all around the globe. Tourists choose Airbnb accommodation for different aspects, such as a wider range of listings, a favorable price-quality ratio, a lower price, the possibility of choosing between a private or a shared environment, the possibility of meeting new people, or living a more authentic experience [3].

Different researchers have previously investigated which aspects can impact on the Airbnb prices and they have proposed different models. Generally, these models include the price as the response variable, and features grouped in categories referable to site characteristics, reputation, convenience, personal, and amenities attributes as independent variables [2]. The methodologies included in these models are ordinary least squares (see for instance [4]), panel data analysis (see for instance [5]), quantile regressions (see for instance [6]), hedonic price models (see for instance [2]), price equations with spatial effects [7], pricing strategy model [8], machine learning (see for instance [9]).

In this work, we propose a statistical composite indicator of the price that estimates the impact of five different dimensions on Airbnb accommodations' prices. The indicator is defined using *proportional odds model* as explained in Vector Generalized Additive Model by [10]. In our analysis we considered the Airbnb accommodation located in the district of Rome called *Trastevere*.

Five sections, besides the introduction, complete this study. The second and third sections are related to the research design: data and methodology are described. The results are explained in the fourth section. Finally, the fifth section focuses on concluding remarks, limitations, and future developments.

A statistical model to identify the price determinations: the case of Airbnb.

## 2 Data

The study has been realized using a dataset composed of six groups of variables: i.e. *transports*, *culture*, *crowd*, *managerial*, *accommodation* (Table 1). Specifically, *transports* data consider the minimum distance from the accommodations to the closest subway stop, the number of bus stops in a range of 200 meters, and the distance from the accommodations to the city center (i.e. Pantheon). The *crowd* dimension measures the number of other Airbnb accommodations and hotels in a range of 50 and 500 meters. The data about the *culture* include the number of monuments in a range of 500 and 2000 meters. The *managerial* dimension identifies the aspects that can be directly chosen by the hosts as the provided services and the additional fees to be applied. The *accommodation* data include the listing types offered on the Airbnb platform, the number of bedrooms and bathrooms. Finally, the accommodation published nightly rate (*price*) has been categorized into five classes according to their quantiles and labeled as *very-low*, *low*, *medium*, *high*, *very-high*. The data have different origins: those related to *transport*, *crowd*, and *culture* have been downloaded from the website *Open Data Roma Capitale* [1]; the others have been provided by the company *Airdna*.

In this study, we take into account only the year 2016 because the data we had available covered only that year.

**Table 1** Variables

Label	Group	Description
price	instrumental	Classes of Published Nightly Rate (euros)
distCenter	transports	Distance from city center (meters)
minDistSubway	transports	Minimum distance from the closest subway stop (meters)
distBus200	transports	Number of bus stops in a range of 200 meters
airbnb_close50	crowd	Number of airbnb in a range of 50 meters
airbnb_close500	crowd	Number of airbnb in a range of 500 meters
hotels_close50	crowd	Number of hotels in a range of 50 meters
hotels_close500	crowd	Number of hotels in a range of 500 meters
monuments_close500	culture	Number of monuments in a range of 500 meters
monuments_close2000	culture	Number of monuments in a range of 2000 meters
Bedrooms	accommodation	Number of bedrooms in a vacation rental listing
Bathrooms	accommodation	Number of bathrooms in a vacation rental listing
Privateroom	accommodation	Dummy variable: private room type (binary variable)
Entirehome	accommodation	Dummy variable: entire home type (binary variable)
Cancellationpolicy	managerial	Cancellation policy for the vacation rental listing (binary variable)
ResponseTimemin	managerial	Average time in minutes a host responds to (minutes)
MinimumStay	managerial	The default minimum night stay required by host
BusinessReady	managerial	Host who provides business facilities (binary variable)
Superhost	managerial	High quality experienced host (binary variable)
NumberofPhotos	managerial	Number of photos in a vacation rental listing

### 3 Methodology

To define the indicator, we have used the *proportional odds model*, also called *cumulative logit model*, as explained in the Vector Generalized Additive Model by [10]. The model is used when the  $Y$  is ordinal and it is defined through:

$$\text{logit } P(Y \leq j|\mathbf{x}) = \eta_j(\mathbf{x}), \quad (1)$$

subject to the following constraint

$$\eta_j(\mathbf{x}) = \beta_{(j)1}^* + \mathbf{x}_{[-1]}^T \beta_{[-(1:M)]}^*, \quad j = 1, \dots, M \quad (2)$$

where  $j$  identifies the level of  $Y$  and moves from 1 to  $M$ ;  $\mathbf{x}_{[-1]}$  is the  $\mathbf{x}$  with the first element deleted;  $*$  denotes the regression coefficient that are to be estimated [10, p. 11]. We fit a model for each of the five dimensions, in which we use the corresponding features to estimate the price class. In other words, the model allows estimating the probability that a specific accommodation belongs to one of the five classes of the price by using the feature of a single dimension. We assume that the higher concentration of probabilities of the model's fitted values in a single price class the better capability of that dimension in explaining the price variable. In order to measure the probability concentration for each observation  $i$  we used the complementary of the normalized Gini index

$$\rho = 1 - \frac{1 - \sum_{j=1}^5 f_{ij}^2}{4/5} \quad (3)$$

where  $j$  identified the five price classes.

We define two tools to discover which group of variables is more important in explaining the price. The first tool computes the average of the  $\rho_i$  as

$$\bar{\rho} = n^{-1} \sum_{i=1}^n \rho_i \quad (4)$$

that assumes a value in the range  $[0, 1]$  and provides directly the information about the importance of the single group in the definition of the price class. The second tool assigns a price class to each observation by using the median of the probabilities of the model's fitted values, and then operates an aggregation of the  $\rho_i$  by the price classes estimated by the model, where the median corresponds to the estimation of the price class for the whole zone.

A statistical model to identify the price determinations: the case of Airbnb.

## 4 Results

We have applied the model on the data related to the 1802 Airbnb accommodations located in the district of *Trastevere* in Rome. Firstly, we have fitted the models and then estimated the probabilities for each observation to belong to the five price classes. We have estimated that the number of Airbnb accommodation in the class *very-low* is equal to 98, in the class *low* is equal to 257, in the class *median* is equal to 424, in the class *high* is equal to 540 and, finally, in the class *very-high* is equal to 483. The main results are illustrated in Table 2. It emerges that the price is mainly determined by *accommodation* (58.76%) and *managerial* (40.92%), whilst the other three dimensions are less significant in explaining the price in this district. Then we can define the most representative price class for the district of *Trastevere*. The results show that the accommodations are rented at a *high* price.

**Table 2** Price classes and dimensions relevance ( $\rho$ ) in Trastevere.

Price class	<i>transports</i>	<i>crowd</i>	<i>culture</i>	<i>accommodation</i>	<i>managerial</i>	Tot
<i>very-low</i>	0.00	0.00	0.00	0.41	220.86	221.27
<i>low</i>	0.00	0.00	0.00	309.23	414.55	723.78
<i>median</i>	3.50	3.88	1.44	176.01	87.92	272.75
<i>high</i>	0.00	0.00	0.00	287.40	388.84	676.24
<i>very-high</i>	0.00	0.00	0.00	876.58	36.72	913.30
Tot	3.50	3.88	1.44	1649.63	1148.89	
Tot in %	0.12	0.14	0.05	58.76	40.92	

## 5 Conclusions

We have defined a price indicator composed by five dimensions using a proportional odds model. The indicator allows evaluating the relevance of each dimension on the price and the dominant price class range in the district. We focused on *Trastevere*, one of the most famous and tourist neighborhoods of Rome. The results have highlighted the relevance of dimensions *managerial* and *accommodation*. Less impact has been recorded for the other groups of variables.

We recognize some relevant aspects in the definition of the indicator. Firstly, we use the average concentration index to define the relevance of each dimension. The index allows attributing weights able to measure the impact of each group of variables has on the price. The second innovative aspect is related to the estimation level. We offer a double point of view of the price determination. To one side, we define a model able to evaluate the indicator for each accommodation offering the possibility to support the single host in the price determination. On the other side, we aggregate the results in terms of the geographical area to better explain the price and its determinants taking into account the geographical proximity.



We identify some limitations in this study. Firstly, we have analyzed the impact of the determinants' price only for one district. Taking into account more districts it can be interesting to comprehend if weights attribute at the dimensions can assume different values concerning the characteristics of the specific area of the city. Secondly, the price indicator has been estimated for a geographical area. However, we believe that inserting the time variation can be interesting to evaluate possible differences in terms of dimensions' impact. We are not sure for instance if the proximity of the subway and bus stops have the same impact in the different annual seasons.

## References

1. <https://dati.comune.roma.it/>
2. Faye B (2021) Methodological discussion of airbnb's hedonic study: A review of the problems and some proposals tested on bordeaux city data. *Annals of Tourism Research* 86:103079
3. Skalska T (2017) Sharing economy in the tourism market: Opportunities and threats. *KNUV* (4 (54)):248–260
4. Voltes-Dorta A, Sánchez-Medina A (2020) Drivers of airbnb prices according to property/room type, season and location: A regression approach. *Journal of Hospitality and Tourism Management* 45:266–275
5. Falk M, Larpin B, Scaglione M (2019) The role of specific attributes in determining prices of airbnb listings in rural and urban locations. *International Journal of Hospitality Management* 83:132–140
6. Wang D, Nicolau JL (2017) Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com. *International Journal of Hospitality Management* 62:120–131
7. Tang LR, Kim J, Wang X (2019) Estimating spatial effects on peer-to-peer accommodation prices: Towards an innovative hedonic model approach. *International Journal of Hospitality Management* 81:43–53
8. Ye P, Qian J, Chen J, Wu Ch, Zhou Y, De Mars S, Yang F, Zhang L (2018) Customized regression model for airbnb dynamic pricing. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp 932–940
9. Kalebhasti PR, Nikolenko L, Rezaei H (2019) Airbnb price prediction using machine learning and sentiment analysis. *arXiv preprint arXiv:190712665*
10. Yee TW (2015) *Vector generalized linear and additive models: with an implementation in R*. Springer

## 4.20 New developments in social statistics analysis

# Data-based Evaluation of Political Agents Against Goals Scheduling

## Valutare agenti politici su linee di governo

Giulio D'Epifanio<sup>1</sup>

**Abstract** An approach is outlined for indexing social-agents (eg governors of health-care-districts) in a social system with respect to their government-worthiness (in increasing welfare-fitness of citizens) against a planned guideline. Such a guideline is provided by the policy-maker-mandator through scheduled goals toward achievement of the ideal social goal. The value-increments on the scheduled goals chain, which will enter the proposed index, are interpreted herein in a probabilistic setting as "worthiness increases", standardized on the reference-agent chosen to represent a "best practice". Agents performance-data are back-forward interpreted on the goals chain, "as if" they were, virtually, the final effect due to the agent government activities, along the preimposed goal chain over the which, for any agent, value increases are integrated by the index, starting from a reference base-line situation.

**Abstract** *Si delinea un approccio per valutare, attraverso un indice, agenti-governatori rispetto alla loro capacità di far avanzare il benessere sociale dei propri cittadini, secondo una linea guida politica preimpostata con una catena di obiettivi che avanzano verso il raggiungimento dell'obiettivo ideale finale. Il significato di "incremento di merito" nell'avanzare obiettivi è interpretato, connesso al rischio di fallire un obiettivo, con riferimento ad un'agente di riferimento, magari virtuale. Un metodo operativo è delineato per assegnare tali valori in un quadro probabilistico. I dati degli agenti sono interpretati, virtualmente, "come se" essi fossero esiti imputabili, per ogni agente, alla sua capacità di governo sulla linea guida preimpostata, lungo la quale l'indice proposto (nella classe di funzionali tipo Yaari-Quiggin) integra "incrementi di merito", partendo da una situazione iniziale di riferimento.*

**Key words:** policy government evaluation, performance worthiness, social agents indexing

---

University of the Study of Perugia. Department of Political Science. ggiuliodd@gmail.com

## 1 Introduction

From the view of the policy-maker-mandator(PMM), agents<sup>1</sup>  $\{A_1, A_2, \dots, A_q\}$  in a social system (eg health-care-district governors as in Table 1) have to be benchmarked with respect to their “*ability in governing*” the own citizens (possibly in a multi-objective setting) against a pre-figured political-guideline, virtually scheduled by the PMM through a goals-path<sup>2</sup> progressing towards the wished ideal goal. Of course<sup>3</sup>, even the more sophisticated data-analysis technical may be lacking of substantial relevancy to the PMM purpose (eg [1],[5] for a review on statistical approaches), whenever the political aims in decision-making were neglected in designing the PMM’s evaluation framework. For instance, adhering to an economic principle that demands “*the highest health, at the lowest cost*”, the PMM aims to improve “*social fitness*”, by interpreting it on “*welfare-sustainable health-care*”. But conflicting problematic (“improving health quality” and “contrasting the social expenditure”) arises on. Thus, the ideal goal  $O_{Full}$  (“*the best health level, at the minimum public health-care cost*”), which is wished by the PMM for each citizen, could be figuratively associated to a preimposed planned goals-chain

$$O_0 \preceq O_1 \preceq O_2 \preceq \dots \preceq O_l \preceq \dots \preceq \dots \preceq O_{L-1} \preceq O_L := O_{Full}, \quad (1)$$

which advances<sup>4</sup> (coherent with the PMM’s criteria) toward the achievement of  $O_{Full}$ , starting from the “tautological” (dummy) base-line goal  $O_0$  (“*at least the lowest health-level, regardless of public health-care cost*”) which is automatically achieved by anyone. For instance herein, on the space  ${}_A Y \otimes_B Y$  of the potential-effects-axes<sup>5</sup> related to government actions, interposed goals (among the “lesser stringent” dummy-goal  $O_0 \equiv \{{}_A Y \geq 0\} \wedge \{{}_B Y \geq 0\}$  and the ideal pursued one  $O_{Full} \equiv \{{}_A Y \geq 4\} \wedge \{{}_B Y \geq 4\}$  wished for anyone) could be sketched, by the PMM’s mind, using a Pert-like diagram as that in Fig. 1.

A data example. Related to the indicators  $({}_A Y, {}_B Y)$  above, agents data (having sampled citizens in typed domain  $\mathcal{D}$ , eg “elderly fem. aged over 75”) are reported in Table (1a). Then, after having processed citizens against detec-

<sup>1</sup> typed in a certain class. specified by social mission. environment and reference conditions

<sup>2</sup> it implicitly specifies the “performance trait” to be evaluated

<sup>3</sup> the data-analyst may recommend the most efficient means to an end. but not the end to be pursued which is inside the PMM’s mind (see [6]. [7]. for general framing and some review)

<sup>4</sup> Guttman order:  $O_l \preceq O_{l+1} \Leftrightarrow$  “whenever goal  $O_{l+1}$  is achieved. also  $O_l$  has been achieved”

<sup>5</sup> they are operationalized through a pair of indicators: the “*individual health status*” level  ${}_A Y \in \{0, 1, 2, 3, 4\}$  (the vertical axis) and the “*individual expenditure*” level  ${}_B Y \in \{0, 1, 2, 3, 4\}$ (the horizontal axis); here  ${}_A Y$  provides SF-36-like health-scores on levels {“*very bad*”(0), “*bad*”(1), “*moderate*”(2), “*good*”(3), “*excellent*”(4)};  ${}_B Y$  classifies health-care expenditure on levels {“*very high*”(0), “*high*”(1), “*moderate*”(2), “*small*”(3), “*very small*”(4)}

tors of the goals in chain (1), the associated outcome-classifier<sup>6</sup>  $Y$  produced the agents performance-data reported in Table (1b).

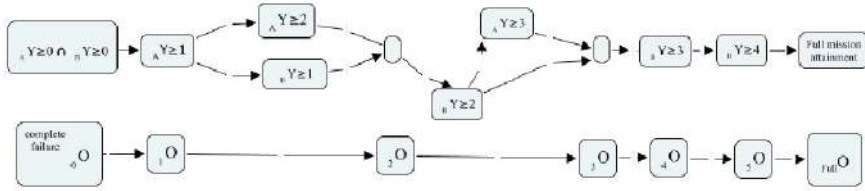


Fig. 1: Example. Pert-like diagram which specifies goals chain

	Expenditure						Expenditure					
Health status	very high	high	mode rate	small	very small		Health status	very high	high	mode rate	small	very small
very bad	0	0	0	0	0		very bad	2	2	0	0	0
bad	2	9	5	0	0		bad	3	13	9	1	0
moderate	0	3	18	0	0		moderate	1	20	31	4	1
good	0	3	17	3	0		good	1	3	59	8	1
excellent	0	0	14	9	4		excellent	0	6	48	18	3
Agent A1							Agent A2					
	Expenditure						Expenditure					
Health status	very high	high	mode rate	small	very small		Health status	very high	high	mode rate	small	very small
very bad	0	3	1	0	0		very bad	3	8	1	0	0
bad	2	24	16	0	0		bad	23	83	48	1	0
moderate	0	20	48	1	0		moderate	1	79	156	9	2
good	0	2	53	3	0		good	1	20	236	24	1
excellent	0	0	49	30	2		excellent	0	6	198	100	13
Agent A3							Reference Agent					
	Expenditure						Expenditure					
Health status	very high	high	mode rate	small	very small		Health status	very high	high	mode rate	small	very small
very bad	1	3	0	0	0		very bad	1	3	0	0	0
bad	16	37	18	0	0		bad	16	37	18	0	0
moderate	0	36	59	4	1		moderate	0	36	59	4	1
good	0	12	107	10	0		good	0	12	107	10	0
excellent	0	0	87	43	4		excellent	0	0	87	43	4
Agent A4							Agent A4					

(a) social agents data (incl. ref. stand. agent  $A^*$ )

perf. lev. $Y$ :	0	1	2	3	4	5	6
	$Pr\{Y \geq 0\}$	$Pr\{Y \geq 1\}$	$Pr\{Y \geq 2\}$	$Pr\{Y \geq 3\}$	$Pr\{Y \geq 4\}$	$Pr\{Y \geq 5\}$	$Pr\{Y \geq 6\}$
agent A1	1	1	0,816	0,540	0,183	0,045	0,045
agent A2	1	0,982	0,863	0,385	0,128	0,017	0,012
agent A3	1	0,984	0,818	0,539	0,137	0,008	0,0078
agent A4	1	0,990	0,828	0,373	0,130	0,009	0,009
ref. st. ag. $A^*$	1	0,921	0,739	0,484	0,125	0,0239	0,010

(b) social agent performance-data. processed by Table (1a) on goals chain (1)

Table 1: Example, health-care social-agents data (incl. ref. st. agent  $A^*$ )

The PMM would need a basic index (and its proper versions), fully normalized on the PMM's reference framework (including goals-chain (1) and the choice of reference-standard agent  $A^*$ ), such that, whenever applied to any agent  $A$ ,

<sup>6</sup> letting  $O_l \leftrightarrow (Y \geq l)$ .  $l := 0, \dots, L$ . an ordinal scale  $Y \in \{0, 1, \dots, l, \dots, L\}$  remains identified with goals chain (1) from the which it will inherit semantics. and vice versa

it takes into input performance-data of  $A$  (eg see Table 1b) to provide the  $A^*$ -standardized value for the *worthiness* of  $A$  in governing own citizens.

## 2 Methodological lines

Evaluating. For any agent  $A \in \{A_1, \dots, A_q\}$ , each of its citizens (in specified social-domain-type  $\mathcal{D}$ ) is sequentially checked against goal-achievement detectors associated to goals-chain (1); it will advance along such a chain until a goal  $O_{l+1}$  is failed, thus finishing ranked at the reached level  $l \in Y$ . Of course, for each level  $l \in Y$ , “the larger the percentage of citizens which social agent  $A$  would have guided beyond over that level; the greater the government capacity of that agent”, which is represented<sup>7</sup> by parameter  $Pr\{Y > l\}[A] \stackrel{est.}{\cong} Rel\ Freq\{Y > l\}[A] = (1 - F_Y[p[A]](l))$ . Thus, for each level  $l \in Y$ , the “social-value increment” due to  $A$  (in moving, perhaps figuratively, a certain percentage mass of its citizens beyond over level  $l$ ) would be (at least roughly) provided by the product:  $(Val(O_l) - Val(O_{l-1})) \cdot (1 - F_Y[p[A]](l))$ , where (crossing goals-chain (1)) pro-capita standardized-value-increases  $\omega_l := \Delta_{l-1}Val := Val(O_l) - Val(O_{l-1}) \geq 0$  need to be specified. Such a specification should be fully independent on by any actual agent to be evaluated, but coherent on the criteria adopted by the PM’s even implicit in the choice of reference-agent  $A^*$ . Herein, the “criterion of intrinsic worthiness” (eg see [3], [4]) is recalled to specify and also practically to determine meaningful worthiness-increases  $\omega_l^*$ , standardized on the government capacity of chosen “reference agent”  $A^*$ . Such a criterion yields (conventionally setting  $\omega_0^* := 0$ )

$$\omega_l^* := \varphi_l(Pr\{Y = l - 1 | Y \geq l - 1; \mathcal{P}^*[A^*]\}), \quad l := 1, \dots, L \quad (2)$$

provided some cont. monot. transf.<sup>8</sup>  $\varphi_l(\cdot)$ , eg the identity function, of the  $A^*$ -standardized risks<sup>9</sup>  $Pr\{Y = l - 1 | Y \geq l - 1; \mathcal{P}^*[A^*]\}$  of failing goal  $O_l \leftrightarrow (Y \geq l)$  in advancing goals-chain (1); here  $\mathcal{P}^*[A^*]$  denote the (actual or figurative) population of the citizens governed by  $A^*$ . Now, provided value-increases  $\omega_l^*$ , accumulating “social-value increments” above, it yields the following index<sup>10</sup>

$$A \mapsto W[A; \omega^*] := \sum_{l:=1}^L \varphi_l\left(\frac{Pr\{Y=l-1; \mathcal{P}^*\}}{Pr\{Y \geq l-1; \mathcal{P}^*\}}\right) \cdot (1 - F_Y[p[A]](l)) \quad (3)$$

<sup>7</sup> here  $p[A] := (p_0, p_1, \dots, p_L)[A]$  denotes the relative(%) distribution which describes the behavior of agent  $A$ ; so that  $F_Y[p(A)](l) \stackrel{def}{=} p_0(A) + p_1(A) + \dots + p_{l-1}(A)$

<sup>8</sup>  $\varphi_l(\cdot)$  could be chosen to specifies types of design-requirements (eg. scale additivity)

<sup>9</sup> given that (at certain conditions referred to  $A^*$ ) a certain goal  $O_{l-1}$  has been achieved, the higher the risk of failing the next goal  $O_l$ , the greater the government-worthiness of that agent which would take masses of its citizens to improve over beyond it

<sup>10</sup> it interprets a version of the Yaari-Quigging functional beyond usual utility-based meaning (eg see [2] pp. 559); it is a Choquet integral from the which it inherits formal properties

Of course, it could be meaningfully re-ranged<sup>11</sup> on interval  $[0, 100\%]$

$$A \in \{A_1, \dots, A_p\} \mapsto W^*[A; \omega^*] := \frac{W[A; \omega^*] - W[A_{worst}; \omega^*]}{W[A_{best}; \omega^*] - W[A_{worst}; \omega^*]} \quad (4)$$

### 3 Interpretations and lines of advancing

Back-forward interpreting. For any agent<sup>12</sup>  $A \in \{A_1, \dots, A_q\}$ , let  $A[T]$  denote its status as described by data-snapshot (formatted as in Table 1b) at date  $T$ . A forward dynamic process could be now figuratively considered, “as if” governor  $A$  would have actually implemented political-acts, to advance its own citizens on preimposed goals-chain (1), starting from an initial situations  $A[T_0]$  dated  $T_0$  ( $T_0$  being an eventually hypothetical initial time,  $T_0 < T$ ). By default herein, such an hypothetical initial situation is associated to the worst social-situation, ie  $A[T_0] \stackrel{def}{=} A_{worst} := (100\%, 0, \dots, 0)$ . Thus, data-snapshot of  $A$  at current date  $T$  (as reported in Table 1b) could be interpreted as the final result due to the political acts of  $A$ , in advancing welfare-fitness of its citizens on the improvement guideline, represented by chain (1), starting from the initial worst social situation. Meanwhile on such a guideline, index (3) integrates  $A^*$ -standardized worthiness-increases  $\omega^*$  to provide the evaluation  $W[A[T]; \omega^*]$  of performance of  $A$  at current date  $T$ . Therefore,  $W[A[T]; \omega^*]$  represents<sup>13</sup> the gross effect due to the worthiness of  $A$  in improving conditions of its citizens on period  $[T_0, T]$ , hypothetically starting from the initial worst social situation.

Let  $A_p$  denote an agent in  $\{A_1, \dots, A_q\}$ , actually competing with  $A$ , chosen to establish a base-line in benchmarking. Then, the evaluation of the government capacity of  $A$ , against that of  $A_p$  on the same (hypothetical) period  $[T_0, T]$  is provided by the following  $A^*$ -standardized variational index:  $\Delta W_{[T_0, T]}[A; A_p] := W[A[T]; \omega^*] - W[A_p[T]; \omega^*]$ . It is discounted for by the common exposition, on the same period  $[T_0, T]$ , to the same environment influence.

Let suppose now that a further data-snapshots (similar to that in Table 1b) has been taken but at a date  $T_1$ , previous the current one, so that  $T_0 < T_1 < T$ . Then, the worthiness-gain/deficit of  $A$  over  $A_p$  on the same (real) period  $[T_1, T]$ , is provided by the following  $A^*$ -standardized variational index:

<sup>11</sup> the agent performance is graduated on the ‘percentage of gained worthiness’ in advancing chain (1). from the complete social-failure. represented by virtual-agent  $A_{worst}$  (such that  $p[A_{worst}] := (100\%, 0, \dots, 0)$ ) toward the full achievement of the wished final social-goal represented by virtual-agent  $A_{best}$  (such that  $p[A_{best}] := (0, 0, \dots, 0, 100\%)$ )

<sup>12</sup> established a social-domain-type  $\mathcal{D}$  for citizens

<sup>13</sup> it could be interpreted by variations:  $\Delta W_{[T_0, T]}[A; A_{worst}] := (W[A[T]; \omega^*] - W[A[T_0]; \omega^*])$ . where  $W^*[A[T_0]; \omega^*] = W^*[A_{worst}; \omega^*] = 0$

$\Delta W_{[T_1, T]}[A; A_p] := (W[A[T]; \omega^*] - W[A[T_1]; \omega^*]) - (W[A_p[T]; \omega^*] - W[A_p[T_1]; \omega^*])$ . Such an index would evaluate the variation in the value of  $A$  with respect to that of  $A_p$ , over the same period  $[T_1, T]$  (thus both the agents subjected to the same environmental exposure), but each agent starting from a different initial situation at date  $T_1$ .

Model based indexing. To take into account various “reference domains”  $\mathcal{D}_{|x_r}$ , related to specific conditions  $x := X \in \{x_1, \dots, x_r, \dots, x_R\}$  of citizens (eg by income levels), a model based index<sup>14</sup> could be used as the following:

$$A \mapsto \sum_{r=1}^R q_r \cdot \left\{ \sum_{l=1}^L \varphi_l \left( \frac{\exp(\hat{a}_l + \hat{b}_l \cdot x_r)}{1 + \exp(\hat{a}_l + \hat{b}_l \cdot x_r)} \right) \cdot (1 - F_{Y|x_r}[p_{|x_r}[A]](l)) \right\}$$

Here,  $q_r \geq 0$  weights<sup>15</sup> ( $\sum_{i=1}^R q_r = 1$ ) the reference domains. A more complex example in modeling is presented in [4].

Index sensitivity checking. It could be also of interest to consider effects on index (3), due to various choices about reference-standard agent  $A^*$ , perhaps interpreting different political stakeholder-views.

## References

1. Bird S.M., Cox D., Farewell V.T., Goldstein H., Holt T., Smith P.C. (2005). Performance indicators: good, bad, and ugly. *JRSS-A*, 168: 1-27
2. Chateaufneuf A., Cohen M., Meilijson I. (2004). Four Notions of Mean-preserving Increase in Risk Attitudes and Applications to the Rank-dependent Expected Utility Model. *Journal of Mathematical Economics* 40, 547-571
3. D'Epifanio G. (2011) Sviluppo di un Indice Multi-attributo per la Valutazione del Merito. In “*Criteri e indicatori per misurare l'efficacia delle attività universitarie*”, vol I, pp. 279. CLEUP, Padova (prepr.: <https://1drv.ms/b/s!AsAje28BQ7S-qAiDBcCkRsd3QIU?e=1bnkFf>)
4. D'Epifanio G. (2018). Indexing the Normalized Worthiness of Social Agents. in *Springer Proceedings in Mathematics & Statistics* 227. C. Perna et al. (eds.), pp 263-274. Studies in Theoretical and Applied Statistics. Springer International Publishing AG, Cham, Switzerland (pre-print)
5. Fisher, N.I. (2019). A comprehensive approach to problems of performance measurement. *JRSS-A*, 182, 3, 755-803
6. Franco-Santos M., Lucianetti L., Bourne M (2012) Contemporary Performance Measurement System: a Review of their Consequences and a Framework for Research. *Management Accounting Research*, 23(2), 79-119
7. Ross, P.H.; Lipsey, M.W.; Henry, G.T.. *Evaluation: A systematic approach*. Sage publications, 2018

<sup>14</sup> it uses a sequence of logistic models, crossing levels  $l := 1, \dots, L$ , to specify  $\mathcal{D}_{|x_r}$ -domain specific value-increases  $\omega_{l|x_r}^*$  through determining parameters  $\hat{a}_l$  and  $\hat{b}_l$  using data from reference agent  $A^*$

<sup>15</sup> the weights should represent the political relevancy of the “social reference domains” to the overall aim, as perceived by the PMM



# **Local heterogeneities in population growth and decline. A spatial analysis for Italian municipalities**

## ***Eterogeneità locali nella crescita e nel declino della popolazione. Un'analisi spaziale nei comuni italiani***

Federico Benassi, Annalisa Busetta, Gerardo Gallo, Manuela Stranges

**Abstract** The unequal demographic dynamics at municipality level lead to a progressive fragility of the territory and its social system. The municipalities in demographic malaise tend to be increasingly dusty in size and peripheral in localization, and their local spatial aggregation increased over time. To try to identify the signs of the population variation, we estimated a spatial regression model. It results that the variation of population is significant affected by the value it assumes in the neighbouring municipalities, thus confirming the spatial nature of the phenomenon and indicating the existence of a spatial diffusion process. The presence of schools in the municipality emerges as a crucial factor for the increase/decrease of the population.

**Abstract** *Le dinamiche demografiche disuguali a livello comunale portano ad una progressiva fragilità del territorio e del suo sistema sociale. I comuni in malessere demografico tendono ad essere sempre più piccoli e periferici, e la loro l'aggregazione spaziale è aumentata nel tempo. Per cercare di identificare i prodromi della diminuzione della popolazione abbiamo stimato un modello di regressione spaziale che ha mostrato come la variazione della popolazione sia influenzata in modo significativo dal comportamento dei comuni limitrofi. La presenza di scuole nel comune emerge come fattore determinante per lo spopolamento/popolamento.*

**Key words:** demographic malaise, Italy, spatial lag models, spatial demography, local analysis

---

<sup>1</sup> Federico Benassi, Istat; Annalisa Busetta, Università di Palermo; Gerardo Gallo, Istat; Manuela Stranges, Università della Calabria.

## 1. Introduction

In Italy, the population trend is strongly territorially differentiated with some municipalities that show a systematic loss of population and others with an equally continuous and significant demographic increase. These unequal demographic dynamics potentially lead to a progressive fragility of the territory and its social system, with consequences in terms of sustainability, socio-economic development and well-being. This strong territorial heterogeneity is also a European phenomenon. Indeed territories subject to depopulation are realities that are gradually becoming weaker and unsafe [1], whereas other territories that grow very quickly - typically large urban and metropolitan areas - clash with other problems that arise from the processes of concentration [2]. To overcome the consequences of heterogeneous demographic dynamics at European level it is recognized the need to support the development of "polycentric territorial systems", or rather territorial areas characterized by well-interconnected medium-sized cities [3]. European Commission affirms that a territorial redistribution of the population and a balanced growth of the territories are necessary conditions for a significant, lasting and sustainable development of the various local realities [3,4]. In this perspective strategies and policies to deal with the problems associated with shrinking regions - i.e. isolated and/or peripheral realities that are in systematic demographic decline - both at European [5] and at Italian level have been developed<sup>2</sup>

Previous studies on population at municipality level [6] showed that from the demographic point of view there are some "more dynamic" situations contrasted by others characterized by demographic malaise that tend to be increasingly dusty in size and become peripheral in localization. Indeed if we look it with a spatial lens the situation seems even more complex. Spatially informed demographic research on the Spanish context seeks in fact to examine whether and how geographical environments directly affect outcomes such as demographic growth and decline [7]. In the Italian context the increasing level of spatial clustering of the variable 'average annual growth rate', is showed globally by the value of the Moran's I [8] global spatial autocorrelation index: the index increases from 0.401 in the period 1981-1991, reach 0.536 in 2001-2011 and then decline to 0.436 in 2011-2019.

Figure 1 maps the local univariate Moran's index [9]. Based on this index's coefficient each municipality was classified as: (i) High-High (HH) hot spots (high growth rates with similar values among neighbouring municipalities), (ii) Low-Low (LL) cold spots (low growth rates with similar values among neighbouring municipalities), (iii) High-Low (HL) potential spatial outliers (high fertility values with low fertility values among neighbouring municipalities), (iv) Low-High (LH) potential spatial outliers (low values with high growth rates among neighbouring municipalities) or finally, (v) units spatially uncorrelated with neighbours.<sup>3</sup>

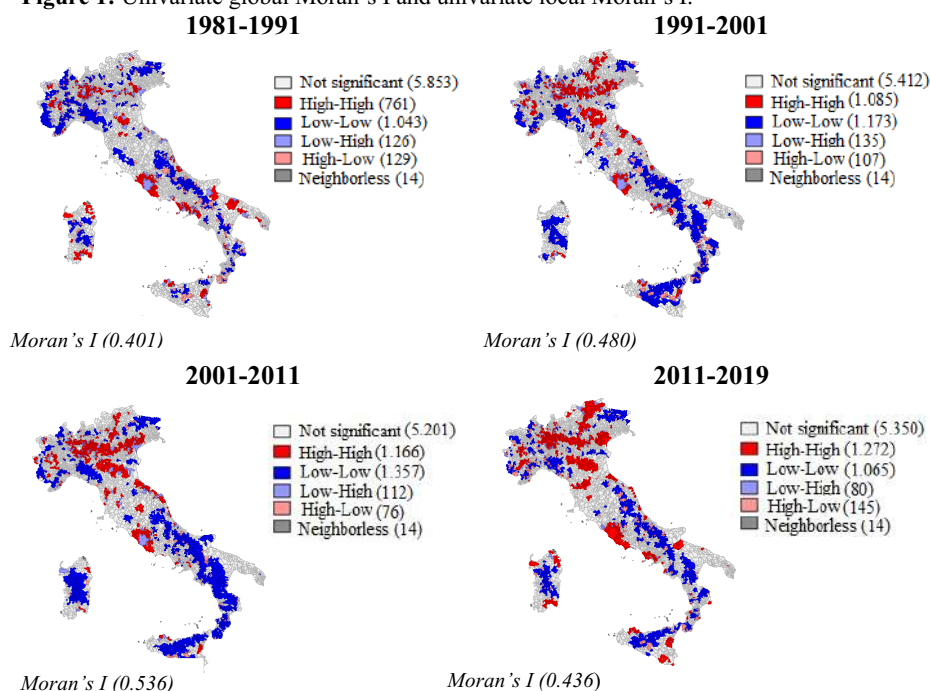
From the analysis of the maps we note both the transition phases of Italian metropolitan area - mainly characterized by the fundamental role of foreigners in the last period [10] - and the consolidation of some demographically 'weak' areas (low -low) in the inland areas of the South and in the peripheral areas of the North and the Centre. The result is a broken up territory in which growing and spatially compact territories, are opposed to large territories in population decrease. The general framework in which some local realities win and some others lose represents a sort of dual space that is detrimental to social cohesion and sustainable development.

---

<sup>2</sup> "Strategia Nazionale Aree Interne" <https://www.agenziacoesione.gov.it/strategia-nazionale-aree-interne/>

<sup>3</sup> It is important to keep in mind that the reference to high and low is relative to the mean of the variable, and should not be interpreted in an absolute sense.

**Figure 1:** Univariate global Moran's I and univariate local Moran's I.



Note: spatial matrix of the "queen" contiguity type of order 1. Statistically significant results  $p \leq 0.05$   
 Source: our elaboration on Istat data. For the years 1951-2011 the data are from census sources. For 2019 the data are from a demographic (pre census) source revised by Istat (<http://demo.istat.it/>).  
 Moran's global univariate index and its local version were calculated using GeoDa by Luc Anselin version 1.14.0, 24 August 2019.

## 2. The description of the approach and/or models used

To try to identify the forerunners of the population increase or decrease of the municipalities, a spatial regression model was used. Specifically, we use a spatial lag model in which the dependent variable (the annual average growth rate of each municipality in the last period) is placed in relation both to a set of covariates measured at the 2011 Census and to the same 'spatially lagged' variable  $y$  [11, 12]. This type of model is the most compatible with the concept of (spatial) diffusion processes because it implies an influence of the neighbourhood structure that is not simply artificial but actually estimated. It is important to clarify that these spatial lag models are designed to produce indirect evidence of diffusion in cross-sectional data (as in our case).

The explanatory variables used in the model analyse whether the signs of future population growth/decrease can be identified among the characteristics detected for each municipality at the 2011 Census. The covariates included in the model concern seven dimensions<sup>4</sup>: the *geo-orographic* one (altimetry zone and coast line of the municipality); the

<sup>4</sup> For geographic, demographic, social, living conditions, housing degradation, employment status

*demographic* dimension (amount of the population in classes, percentage of preschool children and elderly over 75, percentage of foreign population); the *social* dimension and mobility (percentage of young people living alone, mobility for study and work reasons and long-distance mobility); *living conditions* (families in a condition of material and social vulnerability); *housing degradation, employment status* (unemployment rate, male and female activity rates, and employment rate of young people aged 15 to 29); the *school infrastructure* (presence/absence of primary, lower and upper secondary school) and the *economic-productive environment* (share of employees in the agricultural and industrial sector, and number of entrepreneurs or self-employed or freelance professionals working in the tourism sector).

### 3. Preliminary results and conclusions

The results of the model (Table 1) show how the variation of  $y$  is significantly influenced by the value it assumes in the spatial neighbourhood of each municipality, thus confirming the spatial nature of the phenomenon and indicating the existence of a spatial diffusion process which manifests itself net of the influence of the other explanatory variables included in the model. Furthermore, it can be appreciated how the geo-orography of the territory is one of the dimensions, together with the demographic one, with a stronger association with the growth / decrease of the population. The location of the municipality in the plain or in the hills, as well as its coastline, are among the factors that most portend a growth in the municipal population. The demographic dimension is significantly and consistently associated with population variation. In particular, the age structure is among the main forerunners of population growth / decrease. The percentage of children up to 6 years is positively associated with the 2011-2019 change in the population, suggesting a demographic increase, while the high percentage of elderly over 75 years shows a negative association. The percentage of foreigners, although statistically significant, makes a much smaller contribution to the growth / decrease of the population.

The social dimension also plays an important role. The share of young people living alone is significantly associated with the increase in the population of the last 10 years. Leaving the family of origin is therefore confirmed as an important phase of transition that constitutes the necessary premise for the formation of the family (and the other stages of transition to the adult state) and an important driving force for stimulating population growth through fertility. The ability to reach places of study and work in a short time is certainly one of the factors that slows down the depopulation of the territories. At the municipal level, the percentage of the population that moves to reach the places of study and work registered in 2011 is negatively associated with the variation in the following years. The share of the population that takes more than 60 minutes a day for the one-way journey (home-work) is not significant.

The variables related to the living conditions of families have significant but limited effects. The share of households with potential economic and material difficulties (measured through the social and material vulnerability index and the share of housing degradation) is

---

dimensions the variables come from Istat ('Principali statistiche geografiche sui comuni' <https://www.istat.it/it/archivio/156224> and '8mila census' <http://ottomilacensus.istat.it/>). The variables related to economic-productive environment dimension come from the 2011 Population Census, whereas those linked to the school infrastructure are provided by the Ministry of Education and Scientific Research. Finally geographical information (shape files) on municipalities are available at Istat ('Confini delle unità amministrative a fini statistici' (<https://www.istat.it/it/archivio/222527#:~:text=I%20confini%20delle%20unit%C3%A0%20amministrative,in%20contestazione%20e%20isole%20amministrative>)).

negatively associated with the population change between 2011 and 2019. The economic-productive structure in 2011 has a significant relationship with the population change; in particular, the share of employed in agriculture and industry slightly negatively associated with the increase in population. On the other hand, the tourist vocation of the territories in 2011 proves to be a useful factor to avoid, or at least to slow down, depopulation.

**Table 1:** Results of a spatial lag model<sup>(a)</sup> on the average annual growth rate in the period 2011-2019

Explanatory variables	Coefficients	Significance
Y spatially lagged	0.249	0.000
Coastal municipality (ref. Non-coastal municipality)	1.684	0.000
Coastal mountain (ref. Inner mountain)	-0.799	0.172
Inner hill	0.585	0.001
Coastal hill	0.717	0.019
Flat land	0.332	0.123
From 2.500 to 4.999 (ref. <2.500)	0.644	0.004
From 5.000 to 19.999	1.408	0.000
From 20.000 to 49.999	1.181	0.021
From 50.000 to 149.999	0.865	0.285
150.000 and over	-1.429	0.376
% less than 6 years	0.670	0.000
% over 75 years	-0.607	0.000
% foreigners	0.006	0.003
% youths living alone	0.130	0.000
Housing degradation	-0.095	0.005
Unemployment rate	-0.026	0.153
Male activity rate	0.075	0.000
Female activity rate	0.063	0.000
Employment rate 15-29 years	0.024	0.027
% workers in agriculture	-0.068	0.000
% workers in industry	-0.094	0.000
Study work mobility	0.038	0.000
Long-distance mobility	0.016	0.383
Material and social vulnerability	-0.026	0.067
Primary school	0.601	0.015
Lower secondary school	-0.010	0.963
Higher secondary school	0.140	0.502
Entrepreneurs in Tourism	0.007	0.000
R-squared	0.534	
Lag coeff. (Rho)	0.249	
Akaike info criterion	50675,5	
N° municipalities	7,926	

Note: (a) contiguity matrix of the 'Queen' type of order 1; Spatial lag model was calculated using GeoDa by Luc Anselin version 1.14.0, 24 August 2019.

Source: our elaboration on census source data (see footnote 4)

The school emerges as a crucial factor related to the depopulation of the municipality or vice versa its population, being naturally both cause and consequence. All other things being equal, the presence of a primary school in the municipality in 2011 is positively linked to the increase in the population in subsequent years. On the other hand, given the ministerial requirements for the creation of classes, it is the same lack or the reduced number of children residing in the municipality that leads to the closure of primary school and the movement of families with children to neighbouring municipalities where the school is present. The vicious circle feeds itself, confirming, however, that the possibility of sending children to school is an essential element to stem the depopulation of small municipalities, especially in the internal areas of the country.

Italy's delay in using the internet is also confirmed by the share of non-users (i.e. people who have never used the internet or used it more than a year ago) between 16 and 74 years equal to 20% against 11% of the European average [13]. The need for investments to improve the internet throughout the country (especially in the smaller and peripheral municipalities), as well as to increase computer literacy and the provision of technological equipment by families (there are still too many families who do not have adequate IT tools and skills) has been evident in recent months. The recent experience of Covid-19 has shown both the limits and opportunities of the distance learning for schools of different levels. Starting from the elements that emerged in this health crisis, it could be possible to rethink alternative and/or supplementary forms of distance learning, if not for primary school, at least for the secondary and university education system as a policy to stem the depopulation of the most remote and isolated. To do this, however, it is essential to invest in reducing the digital divide and increasing access to ICT (Information Communication Technology) services, especially in small municipalities, in internal areas and in the South of the country. In fact, the latest Istat Annual Report [13] highlighted how in 2019 the internet was used regularly by only 74% of individuals between 16 and 74 years of age, compared to 85% in the EU28 [13], with more disadvantaged situations for less educated families (in which the highest qualification within the household is the middle school certificate), in the South and in municipalities with up to 2,000 inhabitants.

## References

1. Lasanta, T., Arnáez, J., Pascual, N., Ruiz-Flaño, P., Errea, M.P. & Lana-Renault, N. [2017], *Space-time process and drivers of land abandonment in Europe*, «Catena», vol. 149, pp. 810-823.
2. Kempen, R.V. e Marcuse, P. [1997], *A new spatial order in cities?*, «American Behavioral Scientist», vol. 41, n. 3, pp. 285-298.
3. European Commission [1999], *European Spatial Development Perspective. Towards balanced and sustainable development of the territory of EU*. Luxemburg, Office for Official Publications of the European Communities.
4. Vanolo, A. [2003], *Per uno sviluppo policentrico dello spazio europeo. Sistemi innovativi territoriali nell'Europa sud-occidentale*, Milano, Franco Angeli.
5. Espon [2017] *Shrinking rural regions in Europe Towards smart and innovative approaches to regional development challenges in depopulation rural regions*, Policy Brief, Espon Egtc.
6. Golini, A., Mussino, A. e Savioli, M. [2000], *Il malessere demografico in Italia: una ricerca sui comuni italiani*, Bologna, Il Mulino.
7. Burillo, P., Salvati, L., Matthews, S.A., Benassi, F. [2020], *Local-Scale Fertility Variations in a Low-Fertility Country: Evidence from Spain (2002–2017)*, «Canadian Studies in Population» vol. 47(4), 279-295.
8. Moran, P.A.P. [1948], *The interpretation of Statistical Maps*, «Biometrika», 35, pp. 255-260.
9. Anselin, L. [1995]. *Local Indicators of Spatial Association – Lisa*, «Geographical Analysis», 27(2), pp. 93-115.
10. Strozza, S., Benassi, F., Gallo, G., Ferrara, R. [2016], *Recent demographic trends in the major Italian urban agglomerations: the role of foreigners*, «Spatial Demography», vol. 4(1), pp. 39-70.
11. Anselin, L. [2001]. *Spatial econometrics*. In Baltagi B.H. (2001). *A companion to Theoretical Econometrics*: pp.310-33 Blackwell Publishing.
12. Matthews, S.A., Parker D. M. [2013], *Progress in Spatial Demography*. «Demographic Research», vol. 28 (10): 271-232.
13. Istat (2020c), *Rapporto annuale 2020*. La situazione del Paese. Istat.

# The assessment of environmental and income inequalities

## *La valutazione delle disuguaglianze ambientale ed economica*

Michele Costa

**Abstract** We analyze the income and environmental inequalities and their interplay by means of both the multidimensional poverty measurement and the Gini index decomposition. We stress how overlap between the two dimensions plays a relevant role and allows powerful insights on the contribution of the environmental dimensions to poverty. Our finding underlines that bad environmental conditions are more likely among low-income units and represent a relevant inequality factor.

**Abstract** *Le disuguaglianze economica ed ambientale vengono analizzate, insieme alla loro interazione, facendo riferimento alla misura multidimensionale della povertà e alla scomposizione dell'indice di Gini. Si sottolinea come la sovrapposizione tra le due dimensioni abbia un ruolo centrale e consenta un quadro approfondito del contributo della dimensione ambientale alla povertà. I nostri risultati sottolineano come cattive condizioni ambientali siano più probabili per le unità a basso reddito e rappresentino un importante fattore di disuguaglianza.*

**Key words:** Environmental inequality, income inequality, overlap

## 1 Introduction

Environmental inequality refers to unequal distribution of opportunities related to environment and has a strong impact on the economic and social system for a wide range of reasons, both ethical, normative and economic.

There is a natural correlation between income and environmental inequalities [6], two of the main dimensions of poverty. In this paper we aim to tackle the interplay between these two dimensions, which is a key issue in inequality analyses and can be considered as a major threat to economic resilience [1].

---

Michele Costa

Department of Economics, University of Bologna, piazza Scaravilli 2, Bologna, Italy, e-mail: michele.costa@unibo.it

We tackle the analysis of the interplay between income and environmental dimensions by building on both the multidimensional measurement of poverty and the decomposition of the Gini index. In particular, in the analysis of the income and environmental inequalities, we underline the role of overlap existing between the two dimensions.

There is a wide range of environmental risks and, depending on the definition which we adopt for the environmental dimension [5], results and policy implications change, while always confirming that bad environmental conditions are more likely among low-income units [3]. The methods we are going to present are extremely flexible and can be adapted to different definitions and the analysis of multiple risks.

## 2 Methodology

The most widespread measure of inequality, the Gini index, has already been successfully used in the study of environmental inequality, both in its traditional expression and in an environmental Paglin-Gini extension proposed by [7] or a spatialized Gini index applied to environmental segregation [8].

A relevant source of information about interplay between income and environmental dimensions is represented by overlap existing between income distributions related to low and high quality of environmental conditions. In the absence of overlap, environmental dimension fully explains income inequality, while, on the contrary, a perfect overlap suggests that environmental and economic dimensions are independent.

In order to evaluate the degree of the overlap we have two measures at our disposal, both introduced by Gini: the probability of transvariation, which measures the frequency of overlapping occurrences, and the intensity of transvariation, which evaluates the extent of the overlap.

Beside probability and intensity of transvariation, we can also take into account the overlap within the Gini index, through its decomposition. In particular, we refer to the decomposition proposed by Dagum [4], which has among its strengths the role attributed to overlap.

Given a population of  $n$  units, the traditional unidimensional poverty representation is based on a vector of incomes,  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  and on a poverty line  $z_y$  on the basis of which we define a zero-one vector  $\mathbf{g}_y$  where  $g_{yi} = 1$  if  $y_i < z_y$  and  $g_{yi} = 0$  otherwise. The number of poor units with respect to income,  $q_y$ , is the sum of the vector  $\mathbf{g}_y$ .

By adding the environmental dimension, a vector  $\mathbf{e} = (e_1, e_2, \dots, e_n)$  representative of the environmental conditions is introduced, together with a poverty line  $z_e$  on the basis of which we define a zero-one vector  $\mathbf{g}_e$  where  $g_{ei} = 1$  if  $e_i < z_e$  and  $g_{ei} = 0$  otherwise. The number of poor units with respect to the environmental dimension,  $q_e$ , is the sum of the vector  $\mathbf{g}_e$ . The use of a  $z_e$  threshold allows for the treatment of many different environmental risks and is robust with respect to various definitions of the environmental dimension.



The assessment of environmental and income inequalities

In the following we refer to the case of only two groups, poor and non-poor, for each of the two dimensions, but it is possible to generalize to the case of more than two groups, a situation that we would like to address in a future work.

The joint analysis of the economic and environmental dimensions leads to classify the  $n$  units of the population in four classes which can be reported in the 2x2 Table 1, where we find the  $q$  poor units in both dimensions, the  $q_e - q$  units which are poor with respect environmental dimension but not according to income, the  $q_y - q$  units poor only with respect to income, and the  $n - q_e - q_y - q$  units that are not poor for both dimensions.

**Table 1** Poor and non-poor units by income and environmental conditions

		Income		
		poor	non poor	
Environment	poor	$q$	$q_e - q$	$q_e$
	non poor	$q_y - q$	$n - q_e - q_y + q$	$n - q_e$
		$q_y$	$n - q_y$	$n$

In absence of overlapping, we have  $q_e = q_y = q$ , which implies both  $q_e - q = 0$  and  $q_y - q = 0$ , and also  $n - q_e - q_y + q = n - q$ : poor (non poor) units according to income are also poor (non poor) units for the environment. On the other side, when the overlap is perfect, the conditional distributions are the same, that is  $q_y/n = q/q_e = (q_y - q)/(n - q_e)$ .

The simplest indicators for the two dimensions are the head count ratios

$$H_y = \sum_{i=1}^n g_{yi}/n = q_y/n \text{ and } H_e = \sum_{i=1}^n g_{ei}/n = q_e/n$$

from which a two-dimensional indicator can be derived as a weighted average:

$$H_{ye} = (H_y w_y + H_e w_e)/(w_y + w_e).$$

Among the possible weighting structures we refer to  $w_y = \log(n/q_y)$  and  $w_e = \log(n/q_e)$ , following the proposal by Cerioli and Zani [2] which aim to measure the intensity of deprivation and social exclusion related to each dimension.

In order to analyze the interplay between environmental and income dimension the three indices  $H_y$ ,  $H_e$  and  $H_{ye}$  are extremely helpful since, on their basis, it is possible to compare the three sets of poor units identified by them.

The joint analysis of the two dimensions can also be obtained by dividing the population in 2 subgroups, the first with the  $q_e$  poor units and the second with  $(n - q_e)$  non poor units according to the environmental conditions, and deriving the Gini index for income as

$$G = G_1 p_1 s_1 + G_2 p_2 s_2 + G_{12} p_1 s_2 + G_{21} p_2 s_1$$

where  $p_1 = q_e/n$  and  $p_2 = (n - q_e)/n$  indicate the population shares,  $s_1 = p_1\bar{y}_1/\bar{y}$  and  $s_2 = p_2\bar{y}_2/\bar{y}$  the income shares,  $G_1$  and  $G_2$  the Gini indices of the two subgroups, and  $G_{12} = G_{21}$  with

$$G_{12} = \frac{1}{n_1 n_2 (\bar{y}_1 + \bar{y}_2)} \sum_{i=1}^{n_1} \sum_{r=1}^{n_2} |y_{1i} - y_{2r}|$$

The Dagum's decomposition of the Gini index, alongside the two traditional components of inequality within and inequality between the subgroups, also introduces a component related to overlap. The inequality within component  $G_w$  is simply obtained as a weighted average of the Gini indices of the subgroups:

$$G_w = G_1 p_1 s_1 + G_2 p_2 s_2$$

Given  $\bar{y}_1 < \bar{y}_2$ , the components of inequality between subgroups  $G_b$  and of inequality related to overlap  $G_o$  are derived by  $G_{12}$ , attributing to  $G_b$  the differences  $|y_{1i} - y_{2r}|$  if  $y_{1i} < y_{2r}$  and to  $G_o$  the differences  $|y_{1i} - y_{2r}|$  if  $y_{1i} > y_{2r}$ .

A further relevant poverty and inequality indicator is the Sen index (1976) [9], which can be expressed as

$$S = H_y(I_p + (1 - I_p)G_p)$$

where  $I_p$  is the mean over the poor of the normalized poverty gap,

$$I_p = \frac{1}{q_y} \sum_{i=1}^{q_y} \left( \frac{z_y - y_i}{z_y} \right)$$

and  $G_p$  is the Gini index of the poor.

Sens's proposal is based on the three  $I$ s, i.e., the three key elements of poverty: its size,  $H_y$ , its depth,  $I_p$ , and its distribution among the poor,  $G_p$ . In order to take into account the effects of the environmental dimension on the Sen index, it is possible to decompose  $G_p$ , which also allows to investigate the role of environmental conditions on the income distribution among the poor.

### 3 A case study on Italian data

In order to illustrate the previous methods, we develop a case study on the data from the Bank of Italy's Survey on Households Income and Wealth for 2006, which is, unfortunately, the last year in which information on environmental conditions is available. The variable of interest classifies the location of dwellings into four groups: degraded areas, neither prestigious nor degraded areas, prestigious areas, other. We set the 60% of the median of the equivalent income as  $z_y$ , the poverty line for equivalent income, and we consider households living in degraded areas to be environmentally poor. On the basis of  $z_y$ , we detect 1314 poor units, while, on

the basis of  $z_e$ , we have 350 poor units. Table 2 shows how only half of the poor units according to the environmental dimension are also poor according to income, while 85% of the non-poor according to the environmental dimension are also not poor according to income: a good match is highlighted between the groups of the non-poor, while there are significant differences between the groups of the poor.

**Table 2** Italian households 2006, poor and non-poor by income and environmental conditions

		Income		
		poor	non poor	
Environment	poor	176	174	350
	non poor	1138	6280	7418
		1314	6454	7768

The first results which can be derived from Table 2 are the unidimensional head count ratios for the two inequality dimensions which are analyzed here. As for income,  $H_y = 0.181$  indicates the presence of 18.1% of poor families, against 5.8% suggested by  $H_e$  in reference to the environmental dimension. Moving to the multidimensional indicator  $H_{ze}$ , which takes into account the two dimensions jointly, we have a percentage of poor families equal to 9.6%. The sets of poor units identified by  $H_y$ ,  $H_e$  and  $H_{ze}$  coincide only for the 176 units that are poor for both dimensions, while the multidimensional index also classifies as poor some units which are poor according only to one of the two dimensions.

As intuitive from the first row of Table 2, the probability of transvariation, which evaluates the frequency of overlap between the two dimensions, is 50.3%; the extent of overlap is measured by the intensity of transvariation, which is equal to 39%, thus indicating a non-negligible, albeit not high, value.

The overall income inequality indicator provided by the Gini index is equal to 0.318, while the poverty Sen index is equal 0.071.

To include the environmental dimension in the analysis, it is possible to decompose the Gini index on the basis of the two groups of 350 and 7418 units which are, respectively, poor and non-poor with respect to the environmental conditions. From Table 3, it is possible to observe how the inequality within represents the most relevant component, with a weight of  $G_w/G = 91.19\%$ , while the inequality between and the overlap component weighs  $G_b/G = 7.23\%$  and  $G_o/G = 1.57\%$ , respectively.

This result should not be read as an indication that the environmental dimension is of little relevance, but as an effect of the weight (94.2%) of the non-poor group on the total. If we look at the decomposition of  $G_p$ , the Gini index related only to the  $q_y$  poor units, the results are quite different: the inequality within weighs  $G_{pw}/G_p = 75.16\%$ , the inequality between  $G_{pb}/G_p = 15.53\%$  and the overlap component  $G_{po}/G_p = 9.32\%$ .

Overall, we identify significantly different patterns for the poor and the non-poor with respect to the environmental dimension, with a strong overlap with the economic dimension only for the non-poor group.

**Table 3** Results for income and environmental inequalities, Italian households 2006

unidimensional and multidimensional indicators					
$H_y = 0.181$	$H_e = 0.048$		$H_{ye} = 0.096$		
Gini index decomposition					
$G = 0.318$	$G_w = 0.290$	$G_b = 0.023$	$G_o = 0.005$		
Sen index components					
$S = 0.071$	$I_p = 0.279$	$G_p = 0.161$	$G_{pw} = 0.121$	$G_{pb} = 0.025$	$G_{po} = 0.015$

## 4 Conclusions

Multidimensional poverty indicators and Gini index decomposition allow powerful insights into the interaction between environmental and economic dimensions, especially in reference to the overlap between the two dimensions. These methods are extremely flexible with respect to different definitions of the environmental dimension and can be implemented using a wide range of variables of any type.

As we also find in a case study on Italian data, the poor and non-poor groups show significantly different patterns in the interaction between environmental and economic inequalities.

Environmental inequality is typically stable over time, it is likely linked to persistent poverty, and can therefore be extremely useful in correctly identifying poor units. Our finding confirms and underlines that income alone provides only partial information on poverty conditions, while adding the environmental dimension leads to a relevant improvement both for the inequality evaluation and for the implementation of policy actions.

## References

1. Adger, W.N., de Campos, R.S., Siddiqui, T., Szaboova, L.: Commentary: Inequality, precarity and sustainable ecosystems as elements of urban resilience. *Urb. Stu.*, **57**, 1588-1595 (2020)
2. Cerioli, A., Zani, S.: A fuzzy approach to the measurement of poverty. In: Dagum, C., Zenga, M. (eds.) *Income and wealth distribution, inequality and poverty*, pp. 272-284. Springer, Berlin (1990)
3. Choi, P., Min, I.: Measuring environmental inequality from air pollution and health conditions. *Appl. Econ. Lett.*, **27**, 615-619 (2020)
4. Dagum, C.: A new approach to the decomposition of the Gini income inequality ratio. *Empirical Economics*, **22**, 515-531 (1997)
5. Downey, L.: Assessing environmental inequality: how the conclusions we draw vary according to the definitions we employ. *Sociol. Spectr.* (2005) doi: 10.1080/027321790518870
6. Meya, J.N.: Environmental inequality and economic valuation. *Environ. and Res. Eco.* (2020) doi: 10.1007/s10640-020-00423-2
7. Millimet, D.L., Slottje D.: An environmental Paglin-Gini. *Appl. Econ. Lett.* **9**, 271-274 (2002)
8. Schaeffer, Y., Tivadar, M.: Measuring environmental inequalities: insights from the residential segregation literature. *Ecol. Econom.*, **164**, 1-14 (2019)
9. Sen, A.K.: Poverty: an ordinal approach to measurement. *Econometrica*, **44**, 219-231 (1976)

# Household financial fragility across Europe

## *La fragilità finanziaria delle famiglie in Europa*

Marianna Brunetti and Elena Giarda and Costanza Torricelli

**Abstract** We investigate households' financial fragility in 12 European countries using the first wave of the Household Finance and Consumption Survey. Financial fragility is defined by accounting for both income constraints and portfolio composition and modelled by means of a bivariate probit. We find that in all countries holding an illiquid portfolio increases the likelihood of being financially fragile, while having a mortgage generally reduces it, albeit there are relevant differences across countries. Decomposing the observed differences wrt Germany (reference country), we prove that household characteristics drive all countries towards higher financial fragility, while in the Netherlands and Luxemburg the institutional set-up is able to counterbalance this effect.

**Abstract** *La fragilità finanziaria delle famiglie è analizzata utilizzando dati della prima wave della Household Finance and Consumption Survey relativi a 12 paesi europei. La condizione di fragilità finanziaria è definita tenendo conto sia della condizione reddituale che della composizione del portafoglio finanziario della famiglia, ed è modellata tramite un probit bivariato. I risultati mostrano forti differenze tra i paesi, sebbene in tutti un portafoglio poco liquido è associato ad una maggiore probabilità di fragilità finanziaria mentre un indebitamento ipotecario generalmente la riduce. Scomponendo le differenze osservate rispetto al paese di riferimento (Germania), si rileva come il setting economico ed istituzionale dei paesi sia in alcuni casi (Olanda e Lussemburgo) in grado di controbilanciare i fattori idiosincratici delle famiglie.*

**Key words:** household financial fragility, portfolio composition, counterfactual analysis, HFCS.

---

<sup>1</sup> Marianna Brunetti, University of Rome Tor Vergata, CEFIN & CEIS; email: [marianna.brunetti@uniroma2.it](mailto:marianna.brunetti@uniroma2.it)

Elena Giarda, Prometeia & CEFIN; email: [elena.giarda@prometeia.com](mailto:elena.giarda@prometeia.com)

Costanza Torricelli, University of Modena and Reggio Emilia & CEFIN; email: [costanza.torricelli@unimore.it](mailto:costanza.torricelli@unimore.it)

## 1 Introduction

The issue of short-term households' financial fragility was brought to the forefront by the 2007-08 financial crisis and has become even more relevant with the recent outbreak of the Covid-19 pandemic, which has led to an unexpected and unprecedented economic shock at global level.

The existing literature focusing on household financial fragility relies on a variety of indicators, most of them related to indebtedness. For instance, Brown and Taylor (2008), Faruqui (2008), Keese (2009) and Jappelli et al. (2013) focus on objective indicators such as the debt-to-income ratio, the debt-service ratio, and the mortgage income gearing. Others, such as May and Tudela (2005), Duygan-Bump and Grant (2009), Magri and Pico (2009), Beck et al. (2010) and Georgarakos et al. (2010), rely on questions concerning the financial burden due to housing costs or being in arrears on mortgages or other debt payments.

Others use metrics not necessarily linked to indebtedness, based e.g. on net wealth, saving and consumption (e.g. Brown and Taylor, 2008; Fuenzalida and Ruiz-Tagle, 2009; Giarda, 2013), or on subjective self-reported indicators such as having difficulties making ends meet (Christelis et al., 2009; McCarthy, 2011), poor living standards (Worthington, 2006) or questions over the ability to cope with unexpected expenses (Lusardi et al., 2011).

Moreover, most existing contributions analyse financial fragility focusing on a single country only. Exceptions include Jappelli et al. (2013), and ECB (2005), which looks only at indebted households, as well as Ampudia et al. (2016), Bankowska et al. (2015) and Gambacorta et al. (2020). However, Ampudia et al. (2016) rely on an indicator based on the financial margin being negative, so that the proposed metric of distress seems to capture mostly an income effect. The same applies to Gambacorta et al. (2020), who provide cross-country descriptive evidence on vulnerable households, i.e. those holding liquid assets (deposits, bonds, and listed equities) which are insufficient to keep the household above the national at-risk-of-poverty line for three months. In the same spirit, Bankowska et al. (2015) consider financially vulnerable those households with unsustainable debt both with respect to income (debt service-to-income higher than 40%) and with respect to assets (debt-to-assets higher than 100%), but again their definition is limited by construction to indebted households.

Against this backdrop, our contribution is a cross-country comparative analysis based on an indicator of fragility, which is free of the subjectivity-bias and can be applied to all households regardless of their debt position. Moreover, we are the first decomposing the observed international differences in financial fragility into two components, one which arises from differences in household characteristics and one from the economic environment in which comparable households live, we also deal with the literature using counterfactual analysis. Whilst this methodology has been largely used in the labour literature (see, among others, Arulampalam et al., 2007; Autor et al., 2008), to the best of our knowledge only Christelis et al. (2013) apply it to household finance. However, their analysis is restricted to the over-50s and more

Household financial fragility across Europe

importantly it is applied to households' assets and mortgage holdings not to the assessment of financial fragility.

## 2 Methodology

For each household  $i$  in country  $j$ , we generate the following two dichotomous variables:

$$y_{1ij} = I(\text{Income} \geq \text{regular expenses})_{ij} \quad (1)$$

$$y_{2ij} = I(\text{Liquid assets} \geq 1500\text{€})_{ij} \quad (2)$$

For the variable  $y_{1ij}$  we rely on a question asking the household whether last year's regular expenses were higher, about the same or lower than its income, and set the variable to 1 for each household  $i$  in country  $j$  declaring its regular expenses to be lower than or equal to income, and 0 otherwise. Variable  $y_{2ij}$  is set to 1 when household  $i$  in country  $j$  holds a liquidity buffer, which is given by the sum of sight and saving accounts and certificates of deposits, worth at least 1500€.

We then model the two outcomes above by means of a bivariate probit, in which the vector of explanatory variables contains – besides countries fixed effects – demographic (households' size, gender, age, marital status and education of the household's financial head), socio economic (occupational status and gross yearly income and net worth, in quartiles) as well as portfolio controls, the latter being a dummy for having a mortgage, and an indicator of the portfolio illiquidity due to housing, defined as the ratio between the residential home value and household total assets.

In order to assess international differences in the determinants of financial fragility, we first estimate the bivariate probit model for each country separately. This allows us to compute, for each country, the average probability of being financially fragile  $p_j = \Pr(y_{1ij} = 1; y_{2ij} = 0 \mid \mathbf{x}_{ij}) = \Phi_2(\mathbf{x}'_{ij}\boldsymbol{\beta}_1; -\mathbf{x}'_{ij}\boldsymbol{\beta}_2; -\rho)$ . Then, we choose Germany as a reference country and we label its probability phase. The difference in the average observed financial fragility between the base country and each country can be decomposed as:

$$(p_{base} - p_j) = (p_{base} - \hat{p}_j) + (\hat{p}_j - p_j) \quad (6)$$

where  $\hat{p}_j$  is the average estimated counterfactual probability of financial fragility that households living in country  $j$  would exhibit if they lived in the reference country. It is obtained by applying the estimated coefficients of the base country to the households living in each country  $j$ .

The first term of the right-hand side of equation (6) represents the contribution of the characteristics of the households (or, more in general, the population) living in country  $j$  to the observed difference in the average probability of financial fragility between the base country and country  $j$ . The second term is the difference between the counterfactual average probability of financial fragility that households living in country  $j$  would exhibit if they lived in the base country and the actual average probability of financial fragility in their country  $j$ . Thus, it captures the contribution

of each country's economic environment. In order to assess their statistical significance, we compute bootstrap standard errors by drawing (with replacement) observations from the full sample of all countries and by repeating the estimation and the decomposition 250 times.

### 3 Dataset and Main Results

We employ the first wave of the Household Finance and Consumption Survey (HFCS), which gathers harmonised data on households' finances and consumption of fifteen euro-area countries (Austria, Belgium, Cyprus, Germany, Spain, Finland, France, Greece, Italy, Luxemburg, Malta, the Netherlands, Portugal, Slovenia, and Slovakia) and refers to 2008-2010. The final estimation sample consists of 12 countries and 34,699 observations.

Table 1 reports the average estimated probabilities of being financially fragile by country (column A), the corresponding differences with the estimated probability in Germany (column B) and the decomposition of these differences into household or population effects (column C) and economic-environment effects (column E). We first detect a great variability across countries in the probability of financial fragility, ranging from around 11% in the Netherlands to more than 50% in Slovenia. Second, the differences with Germany vary in sign and especially in magnitude. Only two countries (the Netherlands and Luxemburg) have an average probability lower than that of Germany, while the remaining exhibit a higher one (slightly so for Austria). Third, as for the decomposition of these differences into their two components, we observe that the household effect is negative in all countries, thus contributing to increase the level of each country's financial fragility compared to Germany.

However, it more than compensates the household effect only in Luxemburg and the Netherlands, which thus result as the only countries with an incidence of financial fragility lower than Germany. This means that protective effect of the economic-institutional setting is strong enough to offset the households' characteristics effect. On the contrary, in Austria and Belgium, the economic-institutional setting is not relevant enough to entirely counterbalance the differences stemming from the household characteristics, but it helps these countries getting closer to the level of financial fragility of Germany. In three Mediterranean countries (Spain, Italy and Portugal), the economic-institutional environment does not contribute significantly to the observed differences with the base country, therefore the difference with Germany is entirely driven by the configuration of the characteristics of the households.

Finally, among the countries where both effects are negative, the institutional effect is generally stronger – in absolute value – than households' characteristics (i.e., Cyprus, Greece, Slovakia and Slovenia). This points to a weak protection offered by the institutional set-up against financial fragility in these countries, especially in Slovakia and Slovenia.

To sum up, there is a great variability in the incidence of financial fragility across countries, with only the Netherlands and Luxemburg exhibiting a lower level



Household financial fragility across Europe

than Germany. In all countries, household characteristics drive financial fragility to a higher level than Germany and only in Luxemburg the economic-institutional setting seems to be able to counterbalance this effect. Finally, Slovakia and Slovenia are the countries where the economic-institutional environment shields the least from financial fragility.

**Table 1:** Probability of financial fragility by country: average level, difference with Germany and its decomposition

	Average probability of Financial Fragility (A)	Difference with the probability of FF in Germany (B) = 0.142 - (A)	Household effect (C)	Bootstrap st. error (D)	Institutional effect (E)	Bootstrap st. error (F)
Netherlands	0.1134	0.0285	0.000	(0.012)	0.029*	(0.016)
Luxembourg	0.1137	0.0282	-0.063***	(0.017)	0.092***	(0.018)
Austria	0.1428	-0.0009	-0.060***	(0.009)	0.059***	(0.013)
Belgium	0.1542	-0.0122	-0.037***	(0.011)	0.025*	(0.014)
Spain	0.1923	-0.0504	-0.049*	(0.027)	-0.002	(0.028)
Italy	0.2377	-0.0958	-0.090***	(0.026)	-0.006	(0.029)
Cyprus	0.2943	-0.1524	-0.047***	(0.016)	-0.106***	(0.024)
Portugal	0.3317	-0.1898	-0.119**	(0.053)	-0.071	(0.053)
Greece	0.4091	-0.2671	-0.119***	(0.025)	-0.148***	(0.028)
Slovakia	0.4561	-0.3142	-0.155***	(0.022)	-0.159***	(0.025)
Slovenia	0.5047	-0.3628	-0.079***	(0.023)	-0.283***	(0.039)

Note: in column (B), a positive sign signals a country's probability of financial fragility lower than in Germany, while a negative sign a higher one. Standard errors of columns (D) and (F) are bootstrapped by drawing with replacement observations from the full sample, by country, and by repeating the estimation and the decomposition 250 times.

## References

1. Ampudia, M., H. van Vlokhoven and D. Zochowski, "Financial fragility of Euro area households," *Journal of Financial Stability*, 27, 250-262, 2016.
2. Arulampalam, W., A. L. Booth, and M. L. Bryan, "Is There a Glass Ceiling over Europe? Exploring the Gender Pay Gap across the Wage Distribution," *Industrial and Labor Relations Review*, 60:2, 163-186, 2007.
3. Autor, D. H., L. F. Katz, and M. S. Kearney, "Trends in U.S. Wage Inequality: Revising the Revisionists," *The Review of Economics and Statistics*, 90, 300-323, 2008.
4. Bankowska, K., Lamarche, P., Osier, G., and Pérez-Duarte, S. "Measuring household debt vulnerability in the euro area" in: Bank for International Settlements (ed.), *Indicators to support monetary and financial stability analysis: data sources and statistical methodologies*, vol. 39, 2015.
5. Beck, T., K. Kibuuka and E. Tiongson, "Mortgage Finance in Central and Eastern Europe Opportunity or Burden?," *The World Bank, Policy Research Working Paper* 5202, 2010.
6. Brown, S. and K. Taylor, "Household Debt and Financial Assets: Evidence from Germany, Great Britain and the USA," *Journal of the Royal Statistical Society, Series A*, 171(3), 615-643, 2008.
7. Brunetti, M., E. Giarda and C. Torricelli, "Is Financial Fragility a Matter of Illiquidity? An Appraisal for Italian Households," *Review of Income and Wealth*, 62(4), 628-649, 2016.
8. Christelis, D., D. Georgarakos and M. Haliassos, "Differences in Portfolios Across Countries: Economic Environment Versus Household Characteristics," *Review of Economics and Statistics*, 95(1), 220-236, 2013.
9. Duygan-Bump, B. and C. Grant, "Household Debt Repayment Behaviour: What Role do Institutions Play?," *Economic Policy*, 24(57), 107-140, 2009.

Marianna Brunetti and Elena Giarda and Costanza Torricelli

10. European Central Bank, "Assessing the Financial Vulnerability of Mortgage Indebted Euro Area Households Using Micro-level Data," *Financial Stability Review*, 150-158, 2005.
11. Faruqui, U., "Indebtedness and the Household Financial Health: An Examination of the Canadian Debt Service Ratio Distribution," Bank of Canada, Working Paper 46, 2008.
12. Fuenzalida, M. and J. Ruiz-Tagle, "Households' Financial Vulnerability," Central Bank of Chile, Working Paper 540, 2009.
13. Gambacorta, R., A. Rosolia, and F. Zanichelli, "All in it together, but with differences: The finances of European households through the pandemic", 2020, available at: <https://voxeu.org/article/finances-european-households-through-pandemic>
14. Georgarakos, D., A. Lojschova and M. Ward-Warmedinger, "Mortgage Indebtedness and Household Financial Distress," European Central Bank, Working Paper 1156, 2010.
15. Giarda, E., "Persistence of financial distress amongst Italian households: Evidence from dynamic models for binary panel data," *Journal of Banking and Finance*, 37(9), 3425-3434, 2013.
16. Jappelli, T., M. Pagano and M. Di Maggio, "Households' Indebtedness and Financial Fragility," *Journal of Financial Management, Markets and Institutions*, 1, 26-35, 2013.
17. Keese, M., "Triggers and Determinants of Severe Household Indebtedness in Germany," SOEP paper 239, 2009.
18. Lusardi, A., D. Schneider and P. Tufano, "Financially fragile households: Evidence and implications," *Brookings Papers on Economic Activity*, Spring, 83-134, 2011.
19. Magri, S. and R. Pico, "Arrears on Mortgages: Differences Across Countries and Their Effect on the Pricing of the Loan," 2nd Australasian Finance and Banking Conference (Available from <http://ssrn.com/abstract=1460700>), 2009.
20. May, O. and M. Tudela, "When is mortgage indebtedness a financial burden to British households? A dynamic probit approach," Bank of England, Working Paper N. 277, 2005.
21. McCarthy, Y., "Behavioural characteristics and financial distress," European Central Bank, Working Paper 1303, 2011.
22. Worthington, A.C., "Debt as a source of financial stress in Australian households," *International Journal of Consumer Studies*, 30(1), 2-15, 2006.

# Refugees' perception of their new life in Germany

## *La Nuova Vita dei Rifugiati in Germania*

Daria Mendola and Anna Maria Parroco

**Abstract** Since 2015, Germany has been hosting noticeable incoming flows of refugees and asylum seekers and despite the quality of life of refugees is expected to be improved in the aftermath of their arrival to Germany, refugees are still facing several problems of integration and economic deprivation. Using a sample of individuals from the first wave of the German IAB-BAMF-SOEP Survey of Refugees, we present some preliminary analyses on their life satisfaction (LS). A gamma glm was estimated to focus on the association among levels of LS and main socio-demographic characteristics as well as post-migration factors. Greater stability (both in the legal and personal sphere) in refugees' lives is positively associated to LS; while education increases consciousness, hence decreasing LS. Interesting insights come out for policy design.

**Abstract** Dal 2015 la Germania accoglie flussi considerevoli di rifugiati e richiedenti asilo. Sebbene la qualità della loro vita sia oggettivamente migliorata essi fronteggiano ancora problemi di integrazione e deprivazione economica. Ricorrendo al campione tratto dalla prima onda della indagine tedesca sui rifugiati IAB-BAMF-SOEP, vengono qui presentate alcune analisi sulla soddisfazione generale per la nuova vita (LS). L'associazione tra i livelli di LS e le principali caratteristiche socio-demografiche nonché i fattori post migratori è stimata attraverso un glm con link Gamma. Gli elementi di stabilità nelle vite dei rifugiati (legati sia alla sfera legale che personale) sono associati a più alti livelli di LS, mentre l'istruzione elevata aumenta l'insoddisfazione. Spunti per politiche di integrazione emergono dalle analisi condotte.

**Key words:** life satisfaction, subjective well-being, asylum seekers, Gamma generalized linear model

## 1 New Life in a High-Income Country

Germany has a leading role in Europe regarding the hosting of refugees, mainly Syrians, Afghans, and Iraqis. There has been a huge increase in the incoming flow since 2015, and in 2019 (the last data available) Germany was the fourth country in the world for the number of accommodated refugees [12]. These numbers, accumulated year by year, have been undermining the endurance of the German welfare system.

Once in Germany, there are no more risks to their lives, and the government provides accommodation and extensive access to health services. However, most of the refugees struggle to fully reach integration in the new society: finding a job and becoming economically independent, understanding the German cultural norms and rebuilding a network of relationships.

Most of the studies about socio-economic characteristics of refugees focus on education, employment, and language skills as the drivers for their integration. Amongst the main predictors of refugees' well-being, we find mental and general health, family ties, and housing conditions ([10], [1], [6], [11]).

Less is known about refugees' subjective well-being and their own perception of the new living conditions in the host country. Hence, life satisfaction (LS) of refugees is still an under-explored theme.

Issues of mental health (such as depression, anxiety, or post-traumatic distress) are reported in a recent and increasing strand of literature for refugees hosted also in highly developed countries (see, e.g., [9] on Sweden; [11] and [7] on Germany).

The role of family support in shaping refugees' well-being is widely acknowledged in the scientific literature. Fleeing wars and persecution, most refugees have left one or more relatives in their homeland or lost them during the conflicts or along migratory routes ([6] and [3]). On arrival in safe countries of destination, material living conditions are usually much better, while the psychological health is often affected by the trauma suffered. Mental health concerns are often reported for refugees even years after their arrival in safe countries ([9] and [11]). Consequences on their life satisfaction are indeed still debated and literature has often provided contradicting results [8].

The aim of this contribution is to add to studies concerning life satisfaction among refugees hosted in highly developed countries. This paper is divided into the following three sections: The subsequent section introduces data from the first wave of the IAB-BAMF-SOEP survey of refugees in Germany and presents the statistical model adopted to explain determinants of LS. The following Section 3 sets out our statistical analyses, and Section 4 concludes discussing the main results from this study.

## 2 Data and Methods

The IAB-BAMF-SOEP Survey of Refugees [2], the first wave of which was carried out in 2016, gathered socio-demographic information about those refugees and asylum seekers who entered Germany between 2013 and 2016. The general structure is that of the main European household surveys (namely the SOEP), with information recorded for both individuals and households.

Our sample is made up of 3,408 individuals: most of them are men (62%); Afghans, Eritreans, Iraqis, and Syrians account for about 83% of the sample. They are also quite young (mean age of 33.5 years)

In this survey, life satisfaction is understood to be a subjective aspect of the quality of life (see [5]); the main variable consists of people's self-assessment of their overall life satisfaction ("How satisfied are you currently with your life in general?" arranged on an 11-point scale). LS answers show the usual negatively skewed distribution with a generally high mean (mean = 7.28, standard deviation = 2.31, skewness = -0.88).

We estimated a Gamma regression model to focus on the association among levels of LS and main individual and household level characteristics. This model is a member of the class of the generalized linear models, where the link function  $g(\cdot)$ , which transforms the expectation of the response variable,  $\mu_i = E(Y_i)$ , to the linear predictor is a Gamma function, accommodating skewed distributions of the Y. Since LS has a negative skewness, we transformed the variable to be consistent with positive skewness of Gamma distribution.

Explanatory variables were organized into four thematic areas:

a) *sociodemographic variables*: sex of the respondents, their education level (arranged in three ordinal levels, according to ISCED standards), geographical area of origin (Syria, Afghanistan, Iraq, former USSR; Africa; Balkan region, other countries), and quantiles of age;

b) *post-migration personal factors*: time spent in Germany (as the number of years passed between arrival in Germany and the time of the interview); legal residence permit (dummy variable in which we combined refugees, entitled to asylum and holders of subsidiary/humanitarian and other forms of international protection into one category, and placing into the other one those awaiting the response to asylum application and those whose application was dismissed) and concerns about their own economic situation (a lot, somewhat, not at all);

c) *post-migration family related factors*: family arrangements (household size and presence of a partner/spouse possibly cohabitant) and accommodation (shared with others or private);

d) *post-migration subjective well-being factors*: satisfaction with specific life domains (current living arrangements, the quality of the food, the privacy that they have, the safety of their neighbourhood and their own current health) on a 11-point scale.

### 3 Results

Satisfaction with life was generally rated lower by men (average score is 7.15; IC<sub>95%</sub>: 7.05-7.25) than by women (7.50; IC<sub>95%</sub>: 7.38-7.61), by people without or with a pending legal status (7.01; IC<sub>95%</sub>: 6.88-7.15) than by refugees and holders of international protection (7.45; IC<sub>95%</sub>: 7.36-7.55).

A Gamma generalized linear model was estimated to provide possible explanations of LS through the sets of covariates presented above. As already mentioned, we reversed the LS scores, so that  $LS^* = \max(LS) - LS$ . This means that positive coefficients for X variables express a decrease in LS. Table 1 displays the model estimates and their significance levels.

*Socio-demographic factors:* a weak significance emerges for the gender, with women more satisfied than men, other things being equal. People in the third quartile of age (33-40 years) show lower satisfaction with life when compared to the youngest. Education influences LS too: highly educated respondents are less satisfied than those with lowest levels of education, other things being equal; instead, respondents with low and medium level of education report the same LS. The country/area of origin is significantly associated with life satisfaction only for Afghans when compared to Syrians: the former have a higher level of life satisfaction. No statistically significant differences emerge between Syrians and people coming from other geographical areas.

*Post-migration personal factors:* As expected, even controlling for main socio-demographic characteristics, respondents' LS is higher among those who obtained any kind of legal protection than among those who had not (yet) received their residence permit. LS increases with the duration of the permanence in Germany (although the significance is only at 10%). LS decreases with the extent of financial concerns. Particularly, people partially concerned or not concerned at all with financial issues show higher level of LS than those very concerned.

*Post-migration family related factors:* the two covariates accounting for family arrangements are associated significantly with LS. Indeed, according to international studies (see, e.g., [3]), higher household size and having a cohabiting partner/spouse -which are both proxies of social support and, more in general, of social capital- increase refugees' LS. Particularly, not having a partner or living separated from him/her (that is not in the same house nor in the same city) lowers the life satisfaction, even controlling for other personal and family characteristics. Unexpectedly, respondents who live in private houses have a lower level of satisfaction than those who live in shared ones, other things being equal. These last results could be related to the feeling of loneliness and heavier organizational tasks even if this hypothesis would need a further in-depth analysis.

*Post-migration subjective well-being:* As accounted for in many studies, also perceived well-being measures, related to specific life domains, are highly significantly associated with overall life satisfaction. Increasing levels of satisfaction with health, living arrangements, privacy in the current living arrangements and feeling safe with neighbourhood positively affect LS.

**Table 1:** Generalized linear model (Gamma family, link log) for reversed Life Satisfaction

Variable	Beta	Variable	Beta
<b>Socio-demographic factors</b>		<b>Economic concerns</b> (ref. a lot)	
Female	-0.063*	Somewhat concerned	-0.141***
<b>Quartiles of age</b> (ref. Q <sub>1</sub> : 18-26 yrs)		Not concerned at all	-0.392***
Q <sub>2</sub> (27-32 yrs)	0.057	<b>Post-migration family factors</b>	
Q <sub>3</sub> (33-40 yrs)	0.092**	<b>Family arrangements</b>	
Q <sub>4</sub> of age (>40 yrs)	0.024	Household size	-0.031***
<b>Nation group</b> (ref. Syria)		Partner or Spouse (ref. none)	
Afghan	-0.158***	... cohabiting partner/spouse	-0.138***
Africa	-0.063	... not cohab. partner/spouse	0.027
Balkans	-0.118	<b>Accommodation</b>	
Former USSR	-0.144	Private apartment	0.081**
Iraq	-0.061	<b>Post-migration subjective well-being</b>	
Other nations	-0.020	Satisfaction with privacy	-0.023***
<b>Education</b> (ref. low)		Satisfaction with health	-0.068***
Middle school	0.029	Satisf. w. living arrangements	-0.088***
High school	0.148***	Satisfaction with safety	-0.013*
<b>Post-migration personal factors</b>		Satisfaction with food	-0.017**
Legal permit	-0.071**		
Years since arrival in Germany	-0.036*	Constant	2.776***
AIC	12939.68	N=3,408	* p<0.10 ** p<0.05 *** p<0.01
BIC	13099.16		

#### 4 Discussion and conclusions

Studies on living conditions of refugees usually focus on objective characteristics, such as education, language skills, or performances in the labour market. On the contrary, this study focuses on a subjective trait, indirectly related to the refugees' integration in the new hosting countries: their satisfaction with current (new) life.

Using the results from the first wave of the German survey of refugees, we provide preliminary analyses of the determinant of refugees and asylum seekers life satisfaction. Our estimates pointed out how higher life satisfaction is associated with the condition of being woman, Afghans, in cohabitation with a partner or with a large family, poorly educated, with having a legal permit to stay in Germany, and with the duration of their permanence in Germany.

Interestingly, those factors addressing a greater stability in people lives (e.g., the status of refugee or the international protection, having had enough time to satisfactorily settle in the new country, as well as living as a couple and without financial concerns) appear to be correlated with greater life satisfaction (consistently with [4]). Hence, to foster social integration and increase LS of refugees and asylum seekers, it stands out as crucial to shorten the process for the issue of the status of refugees or of the international and humanitarian permits (which are also related to the possibility of family reunification) and promote opportunities for economic independence (pre-requisite for the formation of new family unions). We believe that LS would deserve to be more extensively studied, not only since it is an indirect measure of the success of integration policies adopted by hosting states, but *per se*, as a measure of individual well-being.

Among the limitations of this contribution, we acknowledge the lack of a deeper analysis of the migratory history. Indeed, since immigrants, and refugees in particular, are a heterogeneous group with a great variety of immigration-related experiences, their past experiences can affect current evaluation of life satisfaction both in terms of *inertia* of negative feelings accumulated during the travel phase of their migration, and in terms of resilience.

## References

- 1 Belau, M.: The impact of the housing situation on the health-related quality of life of refugees located in North Rhine-Westfalia, Germany. *European Journal of Public Health*, **29**(Supplement\_4), ckg185-175 (2019).
- 2 Brückner, H., Rother, N., & Schupp, J.: IAB-BAMF-SOEP-Befragung von Geflüchteten: Überblick und erste Ergebnisse **29**, 77. DEU (2016).
- 3 Busetta, A., Mendola, D.: The effect of family networks on refugees health conditions. In: *Giornate di Studio sulla Popolazione 2019, January 24-26, 2019 – Milan, Italy* (2018).
- 4 Colic-Peisker, V.: Visibility, settlement success and life satisfaction in three refugee communities in Australia. *Ethnicities*, **9**(2), 175--199 (2009).
- 5 Cummins, R.A.: Objective and Subjective Quality of Life: An Interactive Model. *Social Indicators Research*, **52**, 55--72 (2000).
- 6 Gambaro, L., Kreyenfeld, M., Schacht, D., Spieß, C. K.: Refugees in Germany with children still living abroad have lowest life satisfaction. *DIW Weekly Report*, **8**(42), 415--425 (2018).
- 7 Georgiadou, E., Zbidat, A., Schmitt, G. M., Erim, Y.: Prevalence of mental distress among Syrian refugees with residence permission in Germany: a registry-based study. *Frontiers in Psychiatry*, **9**, 393, 1--12 (2018).
- 8 Hendriks, M.: The happiness of international migrants: A review of research findings. *Migration Studies*, **3**(3), 343--369 (2015).
- 9 Leiler, A., Bjärtå, A., Ekdahl, J., Wasteson, E.: Mental health and quality of life among asylum seekers and refugees living in refugee housing facilities in Sweden. *Social psychiatry and psychiatric epidemiology*, **54**(5), 543--551 (2019).
- 10 Phillips, D.: Moving towards integration: the housing of asylum seekers and refugees in Britain. *Housing Studies*, **21**(4), 539--553 (2006).
- 11 Walther, L., Fuchs, L. M., Schupp, J., & von Scheve, C.: Living Conditions and the Mental Health and Well-being of Refugees: Evidence from a Large-Scale German Survey. *Journal of Immigrant and Minority Health*, **22**, 903--913(2020).
- 12 UNHCR (2021). Refugee data finder. Available at: <https://www.unhcr.org/refugee-statistics/download/?url=vvO2> (Accessed 5<sup>th</sup> April 2021).



## 4.21 New perspectives in clinical trials

# Improved maximum likelihood estimator in relative risk regression

## *Alcuni miglioramenti alla massima verosimiglianza per la regressione sul rischio relativo*

Euloge C. Kenne Pagui, Francesco Pozza and Alessandra Salvan

**Abstract** Relative risk allows to parametrize in a simple and interpretable way the effect of a group of explanatory variables on a binary response of interest. Despite its ease of interpretation, statistical inference on this parameter is often challenging and limited to specific experimental designs. Starting from a previous work in the field, we propose the use of two estimating procedures for the cases in which there are many explanatory variables, but the interest is mainly focused on the effect of a binary risk factor. These methods, aiming at mean and median bias reduction of the maximum likelihood estimator, rely on a systematic correction of the likelihood estimating equation. We show through a simulation study and an application that the proposed methods perform better than ordinary maximum likelihood.

**Abstract Sommario** Il rischio relativo permette di parametrizzare in modo semplice ed interpretabile l'effetto di un insieme di variabili esplicative su una risposta di tipo binario. Tuttavia, l'inferenza statistica su questa quantità è spesso difficile da implementare in pratica. Utilizzando come punto di partenza un modello recentemente sviluppato in letteratura, in questo articolo vengono proposti due nuovi metodi per l'inferenza sul rischio relativo quando l'interesse è focalizzato sull'effetto di un fattore di esposizione binario. Tali metodologie sono derivate da alcuni metodi per la riduzione della distorsione in media e mediana dello stimatore di massima verosimiglianza. Attraverso uno studio di simulazione e un'applicazione su dati reali, si mostra come gli stimatori proposti abbiano migliori proprietà distributive rispetto allo stimatore originale.

**Key words:** Bias reduction, likelihood, relative risk

---

Euloge C. Kenne Pagui  
University of Padova, Department of Statistical Sciences, e-mail: [kenne@stat.unipd.it](mailto:kenne@stat.unipd.it)

Francesco Pozza  
University of Padova, Department of Statistical Sciences, e-mail: [francesco.pozza.2@phd.unipd.it](mailto:francesco.pozza.2@phd.unipd.it)

Alessandra Salvan  
University of Padova, Department of Statistical Sciences, e-mail: [salvan@stat.unipd.it](mailto:salvan@stat.unipd.it)

## 1 Introduction

Many studies are focused on the effect of a set of observed variables on a binary response of interest. In this context, logistic regression is by far the most popular statistical model. Under this model, the effects of the explanatory variables are parameterized in terms of log-odds ratio. This choice is, in some way, obligated in experimental designs, such as case-control studies, where a direct estimation of relative risk is not feasible. On the other hand, there are many contexts in which relative risk can be directly estimated and gives more interpretable results (see for example McNutt et al., 2003). Unfortunately, log-binomial regression, whose coefficients can be interpreted in term of the logarithm of relative risk, is often impossible to use in practice because of problems related to numerical instability (Barros and Hirakata, 2003). For this reason, in the last two decades many different methods for the estimation of relative risk have been proposed. A particularly attractive one, for the case when the interest is focused on the effect of a binary exposure and its interactions with the other explanatory variables, was proposed by Richardson et al. (2017). In particular, the novelty of their approach relies on a different parameter specification, where the parameter of interest and the nuisance part are variation independent. This makes it easier to obtain the maximum likelihood estimates and their standard errors.

Even though solving many difficulties regarding the inference on relative risk, likelihood inference under the model developed by Richardson et al. (2017) may be subject to a significant bias, in particular with small sample sizes or a high number of explanatory variables. For this reason, in this article, we implement the mean and median bias reduction methods proposed by Firth (1993) and Kenne Pagui et al. (2017) respectively. Indeed, these techniques allow to improve the sample properties of the maximum likelihood estimator mainly by reducing its mean and median bias.

The rest of the article is organized as follows. In Section 2 we briefly introduce the approach proposed by Richardson et al. (2017) to inference on relative risk while Section 3 describes the mean and median reduction methods developed by Firth (1993) and Kenne Pagui et al. (2017). Finally, in Section 4 and Section 5 the performance of the bias reduced estimators is compared with the maximum likelihood through a simulation study and a real-data application.

## 2 Modelling relative risk

Let  $y_1, \dots, y_n$  be realizations of  $n$  independent binary random variables  $Y_1, \dots, Y_n$ . For each observation there are  $p$  explanatory variables  $(x_i, t_i) = (x_{i0}, x_{i1}, \dots, x_{ip-2}, t_i)$ , where  $x_{i0} = 1, i = 1, \dots, n$  and  $t_i$  is a binary exposure. Moreover, we assume that the expected value of  $Y_i, i = 1, \dots, n$ , depends on the row vector  $(x_i, t_i)$ , and we write  $E(Y_i; x_i, t_i) = P(Y_i = 1; x_i, t_i) = \pi(x_i, t_i) = \pi_i$ . In the following, the parameter of interest is the logarithm of the relative risk associated to  $t_i$  and defined as  $\log(RR_i) = \log(\pi_{i1}/\pi_{i0})$ , where  $\pi_{i1} = \pi(x_i, t_i = 1)$  and  $\pi_{i0} = \pi(x_i, t_i = 0)$ . In this

context Richardson et al. (2017) developed a new method for inference on  $\log(\text{RR}_i)$ . In particular they propose the parameterization

$$\eta_{i1} = \log(\text{RR}_i),$$

and

$$\eta_{i2} = \log(\pi_{i0}\pi_{i1} / [(1 - \pi_{i0})(1 - \pi_{i1})]),$$

where  $\eta_{i1}$  and the nuisance parameter  $\eta_{i2}$  are variation independent. Moreover,  $\eta_{i1}$  and  $\eta_{i2}$  are linked to the explanatory variables using two linear predictors  $\eta_{i1} = \gamma x_i$  and  $\eta_{i2} = \beta x_i$ , where  $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_{p-2})$  and  $\beta = (\beta_0, \beta_1, \dots, \beta_{p-2})$  with  $(\gamma, \beta) = \theta \in \mathbb{R}^{2(p-1)}$ . The maximum likelihood estimator  $\hat{\theta} = (\hat{\gamma}, \hat{\beta})$  is obtained by maximizing the log-likelihood without any constraint on the parameter space. A major drawback of this proposal is that, compared to models which have only a linear predictor, the number of parameters tends to be higher. In a situation with many explanatory variables, or when the number of observations is low, this could heavily affect the properties of the maximum likelihood estimator.

### 3 Modified score equations

Let  $\ell(\theta)$  be log-likelihood function for a generic parametric model with parameter  $\theta \in \Theta \subseteq \mathbb{R}^d$ . Let  $U_r = U_r(\theta) = \partial \ell(\theta) / \partial \theta_r$  be the  $r$ -th element of the score function,  $U(\theta)$ , for  $r = 1, \dots, d$ . The observed and the expected information are denoted with  $j(\theta) = -\partial^2 \ell(\theta) / (\partial \theta \partial \theta^T)$  and  $i(\theta) = E_\theta \{j(\theta)\}$ , respectively. Firth (1993) and Kenne Pagui et al. (2017) show that it is possible to obtain estimators with smaller mean and median bias, compared to the maximum likelihood one, through a suitable modification of the maximum likelihood estimating equation. The subsequent papers by Kosmidis and Firth (2010) and Kenne Pagui et al. (2020) give an alternative and more computational attractive matrix expression of these adjustment terms. In particular, the mean bias reduced estimator  $\hat{\theta}^*$  proposed by Firth (1993) can be obtained by solving the equation

$$U^*(\theta) = U(\theta) + A^*(\theta) = 0.$$

The vector  $A^*(\theta)$  has elements  $A^*(\theta)_r = \text{Tr}[i(\theta)^{-1} \{P_r(\theta) + Q_r(\theta)\}] / 2$ , where  $P_r(\theta) = E_\theta \{U(\theta)U(\theta)^T U_r(\theta)\}$ ,  $Q_r(\theta) = -E_\theta \{j(\theta)U_r(\theta)\}$  for  $r = 1, \dots, d$ , and  $\text{Tr}(\cdot)$  is the trace operator. Similarly, the median bias reduced estimator,  $\tilde{\theta}$ , is obtained by solving

$$\tilde{U}(\theta) = U(\theta) + \tilde{A}(\theta) = 0,$$

with  $\tilde{A}(\theta) = A^*(\theta) - i(\theta)\tilde{F}_2$ . The vector  $\tilde{F}_2$  has entries  $\tilde{F}_{2,r} = [i(\theta)^{-1}]_r^T F_{2r}$  for  $r = 1, \dots, d$ , where  $[i(\theta)^{-1}]_r$  represents the  $r$ -th column of  $i(\theta)^{-1}$  and  $F_{2r}$  is a vector with elements  $F_{2s,r} = \text{Tr}[h_r(\theta) \{P_s(\theta)/3 + Q_s(\theta)/2\}]$  for  $s = 1, \dots, d$ . Above,  $h_r(\theta) = [i(\theta)^{-1}]_r [i(\theta)^{-1}]_r^T / i^{rr}(\theta)$  is a  $d \times d$  matrix where  $i^{rr}(\theta)$  is the  $(r, r)$  el-

ement of  $i(\theta)^{-1}$ . Under mild regularity conditions,  $\hat{\theta}^*$  is unbiased with an error of order  $O(n^{-2})$ , i.e.  $E_{\theta}(\hat{\theta}^*) = \theta + O(n^{-2})$ . Since the bias is strictly tied to the parametrization,  $\hat{\theta}^*$  is equivariant only under linear transformations. The median bias reduced estimator is component-wise median unbiased with an error of order  $O(n^{-3/2})$  in the continuous case, i.e.  $P_{\theta}(\tilde{\theta}_r \leq \theta_r) = 1/2 + O(n^{-3/2})$ . Moreover,  $\tilde{\theta}$  is invariant under monotone component-wise transformations of  $\theta$ . Asymptotically, both  $\hat{\theta}^*$  and  $\tilde{\theta}$  have the same multivariate normal distribution as the maximum likelihood estimator,  $\mathcal{N}_d(\theta, i(\theta)^{-1})$ .

## 4 Simulation study

In this section, we compare through a simulation study the maximum likelihood estimator with the mean and median bias reduced ones for the model proposed by Richardson et al. (2017). In particular, the three estimators are evaluated in terms of estimated relative bias (RBIAS), empirical probability of underestimation (PU), root mean square error (RMSE) and empirical coverage of the 95% Wald-type confidence intervals (WALD). Except for the RMSE, all the other measures are reported in percentages.

We consider the model with

$$\eta_{i1} = \gamma_0 x_{i0} + \gamma_1 x_{i1} + \gamma_2 x_{i2}, \quad i = 1, \dots, n,$$

and

$$\eta_{i2} = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2}, \quad i = 1, \dots, n,$$

where  $x_{i0} = 1$ ,  $x_{i1}$  is a realization of a uniform random variable between -1 and 1 and  $x_{i2}$  comes from a Poisson with mean 2.5. For half of the units in the sample the binary exposure factor  $t_i$  is equal to 1. The parameters of the model have been randomly chosen from a uniform random variable with values between -1 and 2. In particular, they are fixed to  $\gamma = (-0.55, 0.75, -0.24)$  and  $\beta = (1.65, 1.82, -0.86)$ .

Using this setting, we consider three different sample sizes  $n = 30, 50, 100$  and for each of them we generate 10 000 samples of the response variable holding the design matrix fixed. From Table 1, it is possible to see how the median bias reduced estimator  $\tilde{\gamma}$  has values of PU that are systematically closer to 50%, while  $\hat{\gamma}^*$  has smaller RBIAS. In addition,  $\tilde{\gamma}$  performs better than  $\hat{\gamma}^*$  in terms of RBIAS and the two bias reduced estimators have smaller values of RMSE. The coverage of Wald-type confidence intervals is comparable for all the estimators. The differences among the three estimators tend to be smaller with increasing  $n$ , but they are not negligible also for  $n = 100$ . In Table 1 we only report the results for the estimators of  $\gamma$ . Similar conclusions are drawn for the estimators of  $\beta$ .

**Table 1** Simulation results for  $\hat{\gamma}$ ,  $\hat{\gamma}^*$  and  $\tilde{\gamma}$

$n$		PU	RBIAS	RMSE	WALD
30	$\hat{\gamma}_0$	45.0	3.6	1.28	97.4
	$\hat{\gamma}_0^*$	48.6	-2.7	0.94	99.3
	$\tilde{\gamma}_0$	48.6	-5.9	1.10	98.9
	$\hat{\gamma}_1$	47.5	26.7	1.25	97.6
	$\hat{\gamma}_1^*$	55.0	-5.8	0.86	98.2
	$\tilde{\gamma}_1$	51.4	6.4	1.04	98.0
	$\hat{\gamma}_2$	58.9	-60.9	0.62	97.2
	$\hat{\gamma}_2^*$	46.4	5.7	0.43	97.9
	$\tilde{\gamma}_2$	50.7	-17.8	0.53	97.7
50	$\hat{\gamma}_0$	50.8	18.5	0.87	95.7
	$\hat{\gamma}_0^*$	46.3	-1.8	0.71	97.0
	$\tilde{\gamma}_0$	49.1	-10.0	0.78	96.6
	$\hat{\gamma}_1$	49.2	15.7	1.03	94.2
	$\hat{\gamma}_1^*$	55.5	-2.5	0.90	94.8
	$\tilde{\gamma}_1$	50.9	9.1	0.96	94.9
	$\hat{\gamma}_2$	52.9	-20.2	0.39	97.3
	$\hat{\gamma}_2^*$	48.0	-0.8	0.30	98.0
	$\tilde{\gamma}_2$	50.3	-10.5	0.35	97.8
100	$\hat{\gamma}_0$	50.7	-9.7	0.51	92.6
	$\hat{\gamma}_0^*$	46.2	0.0	0.46	94.2
	$\tilde{\gamma}_0$	49.6	-6.2	0.49	93.8
	$\hat{\gamma}_1$	49.8	3.3	0.58	95.6
	$\hat{\gamma}_1^*$	52.0	-1.5	0.52	96.5
	$\tilde{\gamma}_1$	50.4	1.5	0.55	96.1
	$\hat{\gamma}_2$	51.4	-6.4	0.22	94.4
	$\hat{\gamma}_2^*$	48.8	0.0	0.20	95.6
	$\tilde{\gamma}_2$	50.4	-3.5	0.21	95.3

## 5 Application

We examine now the respiratory dataset analysed in Everitt and Hothorn (2009, Section 13.1). These data are obtained from a multicentre clinical trial developed to test the effect of a new treatment on a respiratory disease. Each one of the 111 participants in the study was randomly assigned to the treatment or to a placebo and the respiratory status was certified in five subsequent monthly visits. Along with the respiratory status, three other variables were collected: Centre, categorical variable with two levels which identifies the centre in which the data were gathered, Gender, the sex of the patient, and Age, the age in years of the patient. In the following, we consider only the respiratory status during the 5-th visit.

The model assumed for the data is

$$\eta_{i1} = \gamma_0 + \gamma_1 \text{Centre}_i + \gamma_2 \text{Gender}_i + \gamma_3 \text{Age}_i, \quad i = 1, \dots, 111,$$

and

$$\eta_{i2} = \beta_0 + \beta_1 \text{Centre}_i + \beta_2 \text{Gender}_i + \beta_3 \text{Age}_i, \quad i = 1, \dots, 111.$$

Table 2 shows the estimates obtained from the methods described in the previous sections, along with their standard errors and their 95% Wald-type confidence intervals. The mean and median bias reduced estimates tend to be slightly different from the maximum likelihood estimate. This is particularly clear for  $\gamma_0$  which parametrizes the main effect of the treatment. These differences also affects the confidence intervals, leading for  $\gamma_3$  to different conclusions on the significance of the parameter at level 0.05.

**Table 2** Respiratory data. Estimates for the parameter  $\gamma$ , estimated standard errors and 95% Wald interval for  $\hat{\gamma}$ ,  $\hat{\gamma}^*$  and  $\tilde{\gamma}$

$\gamma$	Estimate	Standard error	Wald 95%
$\hat{\gamma}_0$	-0.526	0.500	-1.506 - 0.453
$\hat{\gamma}_0^*$	-0.479	0.503	-1.465 - 0.508
$\tilde{\gamma}_0$	-0.493	0.504	-1.481 - 0.495
$\hat{\gamma}_1$	-0.127	0.386	-0.884 - 0.630
$\hat{\gamma}_1^*$	-0.110	0.388	-0.871 - 0.652
$\tilde{\gamma}_1$	-0.116	0.388	-0.877 - 0.645
$\hat{\gamma}_2$	0.083	0.418	-0.737 - 0.902
$\hat{\gamma}_2^*$	0.109	0.439	-0.753 - 0.970
$\tilde{\gamma}_2$	0.094	0.435	-0.759 - 0.948
$\hat{\gamma}_3$	0.030	0.015	0.001 - 0.059
$\hat{\gamma}_3^*$	0.027	0.015	-0.002 - 0.057
$\tilde{\gamma}_3$	0.029	0.015	-0.001 - 0.058

## References

1. Barros, A.J. & Hirakata, V.N (2003). Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology* **3**, (21).
2. Everitt, B. S. and Hothorn, T. (2009), *A Handbook of Statistical Analyses Using R*, Boca Raton, Florida: Chapman & Hall/CRC Press, 2nd edition
3. Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27-38.
4. Kenne Pagui, E. C., Salvan, A. & Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika* **104**, 923-38.
5. Kenne Pagui, E. C., Salvan, A. & Sartori, N. (2020). Efficient implementation of median bias reduction with applications to general regression models. URL: <https://arxiv.org/pdf/2004.08630.pdf>
6. Kosmidis, I. & Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics* **4**, 1097-1112.
7. McNutt, L.A., Wu, C., Xue, X. & Hafner, J.P (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology* **157**, 940-43.
8. Richardson, T. S., Robins, J. M. & Wang, L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association* **112**, 1121-30.

# Development and validation of a clinical risk score to predict the risk of SARS-CoV-2 infection

## *Sviluppo e validazione di un indice prognostico in grado di predire il rischio di infezione da SARS-CoV-2*

Laura Savaré, Valentina Orlando and Giovanni Corrao

**Sommario** To date, there is a lack of studies describing the clinical characteristics of patients most at risk of SARS-CoV-2 infection. We aimed (i) to identify clinical predictors of SARS-CoV-2 infection risk, (ii) to develop and validate a score predicting SARS-CoV-2 infection risk, and (iii) to compare it with unspecific scores. A retrospective case-control study was carried. Odds ratios for associations between candidate predictors and risk of infection were estimated by means of conditional logistic regression. SARS-CoV-2 Infection Score (SIS) was developed by generating a total aggregate score obtained from assignment of a weight at each selected covariate using coefficients estimated from the model. Conditions and diseases that make people more vulnerable to SARS-CoV-2 infection were identified.

**Sommario** *Ad oggi mancano studi che descrivono le caratteristiche cliniche dei soggetti più a rischio di infezione da SARS-CoV-2. Abbiamo quindi eseguito un'ampia indagine volta a identificare i predittori clinici dell'infezione per sviluppare e validare un indice predittivo di tale rischio confrontandolo con indici di fragilità aspecifici. È stato condotto uno studio retrospettivo caso-controllo e, mediante modelli di regressione logistica condizionata, è stata valutata l'associazione tra i predittori individuati e il rischio di infezione. Tramite l'assegnazione di un peso per ciascuna condizione, proporzionale ai coefficienti stimati dal modello, è stato sviluppato il SARS-CoV-2 Infection Score (SIS). Questo studio ha identificato condizioni e malattie che rendono gli individui più vulnerabili all'infezione da SARS-CoV-2.*

**Key words:** SARS-CoV-2 infection, risk score, population-based cohort study

---

Laura Savaré <sup>1,2,4</sup>, Valentina Orlando <sup>3</sup> and Giovanni Corrao <sup>4</sup>

<sup>1</sup>MOX - Department of Mathematics, Politecnico di Milano, Milan, Italy

<sup>2</sup>CADS, Center for Analysis Decisions and Society, Human Technopole, Milan, Italy

<sup>3</sup>CIRFF, Center of Drug Utilization and Pharmacoeconomics, University of Naples Federico II, Naples, Italy

<sup>4</sup>National Centre for Healthcare Research & Pharmacoepidemiology, at the University of Milano-Bicocca, Milan, Italy

e-mail: laura.savare@polimi.it valentina.orlando@unina.it giovanni.corrao@unimib.it



## 1 Introduction

Since December 2019, the novel coronavirus (SARS-CoV-2) pandemic spread rapidly from the Hubei province in China to all the rest of the world causing at the current time (February 2021) over 100,000,000 cases [2]. The epidemic increased exponentially in Italy, earlier than in any other Western Country. SARS-COV-2 causes a Coronavirus disease 2019 (Covid-19), for which minor symptoms are anosmia, ageusia, gastrointestinal symptoms, headache, and cutaneous manifestations and major symptoms are fever, cough, dyspnoea [4]. Due to these major symptoms it may be considered necessary to hospitalize patients for respiratory complications.

Several hospital-based studies [3, 5], including a systematic review of literature and meta-analysis [6], focused on the attempt for predicting the progression of the disease towards developing critical manifestations or death. These studies are important from the clinical practice point of view for identifying patients at whom early treatment must be guaranteed. However, as most infections are not life-threatening [5], it becomes increasingly important to stratify population for identifying people at higher risk of infection. Despite this, to our best knowledge, no studies on this topic have been still published.

We therefore performed a large investigation based upon healthcare utilization database from the Italian Region of Campania aimed (1) to identify clinical predictors of the risk of SARS-CoV-2 infection, (2) to develop and validate a score overall predicting the risk of SARS-CoV-2 infection, and (3) to compare discriminant power of such a score with that from unspecific scores of clinical profile.

## 2 Dataset

Residents in Campania who were beneficiaries of the Regional Health Service (RHS) aged 30 years or older formed the target population (almost 3.9 million people, around 9% of the Italian population of that age group). Italian citizens have equal access to essential healthcare services provided by the National Health Service (NHS). An automated system of healthcare utilization (HCU) databases allows managing NHS within each Italian region, including Campania. HCU data report a variety of information drawn from services provided fully or in part free of charge from NHS to beneficiaries of NHS (e.g. the ICD-CM-9 codes of inpatient diagnoses and services supplied from public or private hospitals, the ATC codes of outpatient drugs dispensed from pharmacies). This allowed to Campania Region of designing, building and routinely managing the so-called Campania Region Database (CaReDB) which formed the data source for the current study.

From the beginning of the Covid-19 epidemic, a surveillance system was implemented to detect all cases identified by reverse transcription-polymerase chain reaction (RT-PCR) testing for SARS-CoV-2. A diagnostic algorithm was based on the protocol released by the World Health Organization (WHO), i.e., on nasopha-

SARS-CoV-2 Infection Score (SIS) to predict patients most at risk of Covid-19

ryngeal swab specimens tested with at least two real-time RT PCT assays targeting different genes (E, RdRp and M) of SARS-CoV-2.

These various types of data (i.e., CaReDB and Covid-19 registry) can be interconnected, since a single individual identification code is used by all databases for each citizen enrolled. To preserve privacy, each identification code was automatically deidentified, the inverse process being allowed only to the Regional Health Authority on request from judicial authorities.

### 3 Methods

The date of SARS-CoV-2 infection diagnosis was considered as the index date and patients were extracted from the Covid-19 registry until June 10, 2020. A total of 4,629 subjects positive to SARS-CoV-2 were identified. Among these, we excluded i) patients with missing demographic information (N=469) and ii) patients younger than 30 years at the index date (N=663). Finally, 3,497 patients were included into the study as cases. Among them, 453 patients died during the observational period.

For each case, up to five controls were randomly selected from the target population to be matched for gender, age at index date and municipality of residence. The density incidence approach was used for selecting controls since patients who had a confirmed diagnosis of SARS-CoV-2 infection were eligible as potential controls until they became cases, and all matches had to be at risk of SARS-CoV-2 infection.

A list of 47 diseases and conditions potentially predicting the risk of SARS-CoV-2 infection was developed starting from the lists included in several comorbidities scores and in some systematic reviews on Covid-19 risk factors [6].

Seven out of ten of the 3,497 1:5 case-control sets were randomly selected to form the so-called training set. Conditional logistic regression was used to estimate odds ratios (ORs), with 90% confidence intervals (CIs), for the association between candidate predictors and the odds of SARS-CoV-2 infection. The least absolute shrinkage and selection operator (LASSO) method was applied for selecting the diseases / conditions able to independently predict the SARS-CoV-2 infection. The coefficients estimated from the model were used for assigning a weight at each selected covariate. A weight was assigned to each coefficient by multiplying it by 10 and rounding it to the nearest whole number. The weights thus obtained were then summed to generate a total aggregate score. To simplify the system, i.e., with the aim of accounting for excessive heterogeneity of the total aggregate score, the latter was categorized by assigning increasing values of 1, 2, 3 and 4 to the categories of the aggregate score of 0, 1-2, 3-4,  $\geq 5$ , respectively. The so obtained index was denoted SARS-CoV-2 Infection Score (SIS).

Performance of SIS was explored by applying the corresponding weights to the so-called validation set consisting of the 1,048 1:5 case-control sets who did not enter the training set. To evaluate the clinical utility of SIS for predicting infection, we considered the receiver operating characteristic (ROC) curve analysis and used area

under the ROC curve (AUC) as a global summary of the discriminatory capacity of the scores.

Some unspecific scores surrogating general clinical profile of each case and control included into the study were considered. In particular, the number of drugs with different 3rd level ATC dispensed to, and comorbidities with different ICD-9-CM experienced by each case and control within two-years prior baseline (2018-2019) were recorded. Categorization was made by assigning increasing values of 1, 2, 3 and 4 to 0, 1-4, 5-9 and  $\geq 10$  drugs (comedication score) and 1, 2, 3 and 4 to 0, 1-2 and  $\geq 3$  comorbidities (comorbidity score). In addition, cases and controls were categorized according to the Multisource Comorbidity Score (MCS), a new index of patients' clinical status derived from inpatients diagnostic information and outpatient drug prescriptions provided by the regional Italian data and validated for outcome prediction [1]. With the aim of comparing discriminatory ability of specific (SIS) and unspecific (comedications, comorbidities and MCS) predictors of SARS-CoV-2 infection, ROC curves and corresponding AUCs were again used.

## 4 Results

Patients suffering from diabetes, anaemias, mental disorders (dementia / Alzheimer's disease, psychosis and anxiety), Parkinson's disease, glaucoma, diseases of the circulatory system (heart failure and hypertension), chronic respiratory, inflammatory bowel, and rheumatologic conditions showed statistical evidence of increased risk of infection with respect to patients who did not suffer from them. Likely because of low power, only 7 conditions resulted significantly associated with the risk of fatal Covid-19 disease, but there was no relevant difference in the estimates with respect to the risk of SARS-CoV-2 infection.

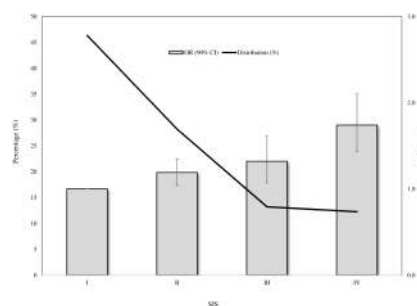
Fifteen conditions significantly contributed to the SIS. Factors which most contributed to the total aggregate score were dementia / Alzheimer's disease, kidney disease, psychosis, inflammatory bowel disease and rheumatologic conditions, while diabetes, anaemias, anxiety, Parkinson's disease, glaucoma, heart failure, hypertension, arrhythmia, thyroid disorders and chronic respiratory disease provided small, although significant, contributions. Figure 1 shows that, as the SIS value increases, the OR progressively increases, being the odds of SARS-CoV-2 infection among people with the highest SIS value (SIS = IV), 1.74 times higher than those unaffected by any SIS contributing conditions (SIS = I). The prevalence of controls stratified according to the SIS score gradually decreases from 50% (SIS = I) to 12% (SIS = IV).

Generic/unspecific scores surrogating clinical profile showed to be associated with the risk of SARS-CoV-2 infection, showing patients with  $\geq 10$  drug treatments, those with  $\geq 3$  comorbidities, and those with MCS value  $\geq 4$ , increased risk of 65%, 36% and 45% with respect to patients cotreatments, comorbidities and MCS value = I, respectively.

AUC (90% CI) of SIS, cotreatment and comorbidity scores and MCS respectively

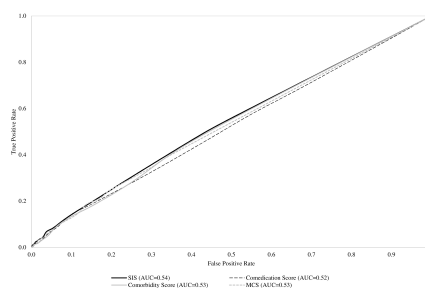
## SARS-CoV-2 Infection Score (SIS) to predict patients most at risk of Covid-19

**Figure 1** SARS-CoV-2 Infection Score (SIS) distribution among controls, and corresponding trend in odds ratios (and 90% confidence intervals) along categories of SIS. SARS-CoV-2 Infection Score: I, II, III and IV to 0, 1-2, 3-4 and  $\geq 5$ .



had values of 0.54 (0.52 to 0.56), 0.52 (0.50 to 0.54), 0.53 (0.51 to 0.55), and 0.53 (0.51 to 0.55) (Fig 2). There were no evidence that specific and unspecific scores had different discriminatory ability.

**Figure 2** Receiver Operating Characteristics (ROC) curves comparing discriminant power of SARS-CoV-2 Infection Score (SIS), and selected unspecific score surrogating clinical profile (cotreatments, comorbidities and Multisource Comorbidity Score).



## 5 Conclusion

Despite our results confirm that a wide range of diseases and conditions likely increase vulnerability to SARS-CoV-2 infection, and probably its more severe clinical manifestations, we have not been able to develop a score that accurately may predict the risk of infection. In addition, we found that predictive ability of the score obtained by weighting risk factors of SARS-CoV-2 infection did not overcome that of some generic scores of comorbidities and comedication.

This can be explained by several limitations of our approach, which generate estimates biased towards the null. First, exposure misclassification regards our inability to carefully capture conditions and diseases through algorithms based on healthcare utilization databases. Second, it is well known that outcome misclassification can bias epidemiologic results. For Covid-19, suboptimal test sensitivity, despite excellent specificity, results in an overestimation of cases in the early stages of an outbreak, and substantial underestimation of cases as prevalence increases.

It should be noticed, however, that both, exposure and outcome misclassification likely drew estimates towards the null (i.e., underestimate the strength of the association between their presence and the outcome risk) so generating uncertainty for the weighting approach of score developing. Third, our choice of accepting a 0.10 first type error, and of consequently reporting 90% confidence intervals, is justified by the exploratory nature of our study, but at the same time likely generate false positive signals, so limiting discriminant power of the score. Forth, the lack of specific data regarding the clinical outcome for the stratification of Covid-19 positive patients in terms of home isolation, hospitalization and admission in intensive care. Furthermore, because data on stays in long-term facilities are not recorded in our database, we cannot exclude that the higher risks associated with mental disorders observed in our study could be explained by confounding, i.e., patients who suffered from these conditions are often hospitalized in these structures where the risk of infection can be particularly high. Finally, the lack of information on biologic markers potentially able to predict infection, and severity of its clinical manifestations, is another limitation of our study, as for example, according to the current literature, some laboratory hallmarks have been shown to predict infection, particularly in more severe cases.

In conclusion, taking the limitations we discussed into account, we identified conditions and diseases that make people more vulnerable to SARS-CoV-2 infection. These findings contribute to inform public health, and clinical decisions regarding risk stratifying. However, further research is need for developing a score reliably predicting the risk, possibly by integrating healthcare utilization with clinical and biological data or by considering different constructions of the score, e.g by categorizing it according to the predicted probability classes of the risk of SARS-CoV-2 infection, using a nomogram.

## Riferimenti bibliografici

1. Corrao, G.: Developing and validating a novel multisource comorbidity score from administrative data: a large population-based cohort study from Italy. *BMJ Open*. (2017) doi:10.1136/bmjopen-2017-019503.
2. Dong, E.: An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*. 20:533-4 (2020)
3. Liang, W.: Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. *JAMA Intern Med*. (2020) doi:10.1001/jamainternmed.2020.2033
4. Vaira, LA.: Olfactory and gustatory function impairment in COVID-19 patients: Italian objective multicenter-study. *Head Neck*. (2020) doi: 10.1002/hed.26269
5. Xie, J.: Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. medRxiv preprint doi:10.1101/2020.03.28.20045997
6. Zheng, Z.: Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *J Infect*. (2020) doi:10.1016/j.jinf.2020.04.021

# Functional representation of potassium trajectories for dynamic monitoring of Heart Failure patients

## *Rappresentazione funzionale delle traiettorie di potassio per il monitoraggio dinamico dei pazienti affetti da Scompenso Cardiaco*

Caterina Gregorio<sup>1,2</sup>, Giulia Barbati<sup>1</sup> e Francesca Ieva<sup>2</sup>

**Abstract** Monitoring potassium is crucial in Heart Failure patients since pharmacological therapy can cause alterations which have been found to increase the risk of hospitalisations and death. It would be essential for the clinical practice to have a tool for the dynamic surveillance of potassium taking into account measurement error and the history of the longitudinal process. Specifically, a functional representation of potassium in a dynamic framework represents a first step towards this direction. The data comes from the Trieste Observatory of Cardiovascular Diseases integrating clinical and administrative regional health data.

**Abstract** *Monitorare il potassio nei soggetti affetti da Scompenso Cardiaco è fondamentale per un efficace trattamento di questi pazienti. La terapia farmacologica in alcuni casi infatti causa delle alterazioni nel potassio che possono portare ad un aumento del rischio di ospedalizzazione e morte. Nella pratica clinica, sarebbe fondamentale avere a disposizione uno strumento per la sorveglianza del potassio che tenga conto degli errori di misurazione e della natura longitudinale del processo. Una rappresentazione funzionale del potassio in un'ottica dinamica rappresenta un primo passo in questa direzione. I dati provengono dall'Osservatorio delle Malattie Cardiovascolari di Trieste che integra le informazioni provenienti dai registri clinici e dai dati amministrativi sanitari regionali.*

**Key words:** Functional Data Analysis, Dynamic Models , Heart Failure, Potassium

## 1 Introduction

Heart failure is a consequence of many cardiovascular diseases and, despite improvements in treatments, mortality and hospitalisation rates remain high. One of

---

<sup>1</sup> Biostatistics Unit - Department of Medical Science, University of Trieste, Trieste 34100, Italy

<sup>2</sup> MOX – Department of Mathematics, Politecnico di Milano, Milan 20133, Italy  
e-mail: caterina.gregorio@polimi.it gbarbati@units.it francesca.ieva@polimi.it

the challenges that clinical research faces in this field is an individualised optimisation of the treatments.

Potassium has been found to be an important biomarker to monitor since it plays a fundamental role in the heart functioning. Heart Failure itself together with the pharmacological treatment are likely to cause potassium alterations. Both low levels (hypokalemia) and high levels (hyperkalemia) of potassium can lead to life-threatening conditions. The normal range of serum potassium is typically cited as 3.5–5.0 mmol/L. However, recent studies have raised serious concerns about its validity [1, 2, 6]. The main issue is that, in clinical practice, hyperkalemia is defined by only one measurement of potassium over 5 or 5.5 mmol/L and this often leads the medical doctor to decide to suspend life-saving therapy. However, it may be the case that dropping from the therapy is even more dangerous than potassium itself. Therefore, it would be very important to go beyond the "cut-off" paradigm and exploiting the (1) functional, (2) longitudinal and (3) dynamic nature of the variable representing the potassium trajectory over time while developing statistical tools to dynamically monitor potassium. The aim of this work is to provide a mathematical representation of potassium trajectories to dynamically monitor potassium and propose a procedure to dynamically reconstruct its functional form together with its first derivative.

## 2 Data

The integration of administrative and Electronic Health Recording Systems offers new opportunities to elaborate prognostic models in a real-world context. Nowadays, these data play an important role in extracting Real World Evidence that help in manage highly complex conditions such as patients in the cardiovascular setting [4, 3]. The data source used in this work is an example of such resources. Data was obtained by the interrogation of the administrative regional health data of Friuli Venezia Giulia Region, integrated with data derived from the Outpatient and Inpatient Clinic E-chart (Cardionet®). This integrated database constitutes the Trieste Observatory of Cardiovascular Diseases. It contains longitudinal information regarding diagnostic codes, laboratory tests, procedures, and drugs prescriptions collected by cardiologists during routine clinical practice as well as diagnosis at discharge from hospitals.

## 3 Methods

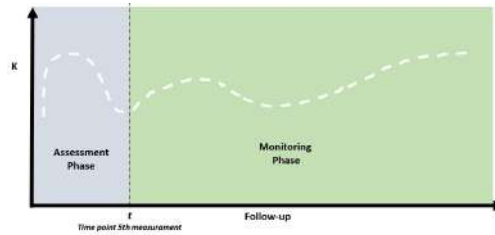
Subjects diagnosed with Heart Failure entered the study from the date of the first potassium measurement for which a cardiological visit could be found in a interval of 1 year before and 3 months after the blood test. They were observed either until the time of death or the administrative end of the study. The administrative censor-

ing date was June 2019.

Repeated potassium measurements can be seen as an individual data streams generated by an underlying longitudinal trajectory for each subject. This time-varying covariate is characterised by being *partially* observed, since patients have been observed for different observation periods. It is also *gradually* observed since it is measured contemporary to the follow-up.

In this work we used a statistical procedure which allows the longitudinal trajectory to be first reconstructed then dynamically updated as new observations arrives. The procedure is based on the observation period being divided in two phases (Fig.1): *assessment* and *monitoring*. From a statistical point of view, these coincides with the estimation and the updating of the individual longitudinal potassium trajectories. The *assessment* period is defined by the time window in which the first five measurements are taken from each subject. These are used to estimate a first preliminary trajectory. Using Functional Data Analysis (FDA) techniques, potassium can be represented by a function of time and it can be estimated by smoothing the data. In the *monitoring phase*, the Sherman-Morrison Formula [7] was used to recursively estimate the functional form of potassium trajectories according to the the subsequent measures.

**Fig. 1** The two phases of the potassium trajectory reconstruction for an individual subject. The *assessment* period is used for a preliminary estimation of the functional form, then the *monitoring* phase begins as new measurements are collected.

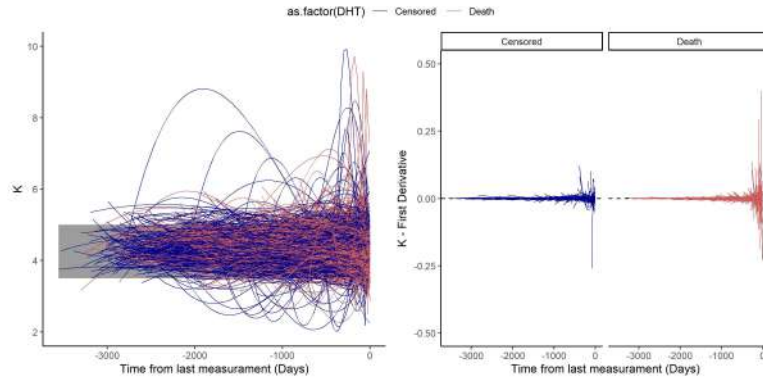


## 4 Results

The dataset included 1500 patients affected by Heart Failure who were observed for a median time of 48 months (IQR: 28-70). Over this period, the median number of potassium measurements per subject was 19 (IQR: 11-33). The mean value of potassium in the first 5 measurements was 4.32 (SD:0.4) and 79% of patients had at least one measurement outside of the normal range (79% were hyperkalemic and 51% were hypokalemic) and the majority of them had repeated episodes of dyskalemia. Moreover, the number of deaths observed in the cohort was 717 (48%), while the total number hospitalisations (including repeated events) for Heart Failure was 1274. The cohort was selected among those having at least 6 potassium measurements.



For each subjects  $i$ ,  $m_i$  measurements were taken during the observation period:  $y_i = \{y_{i1}, \dots, y_{im}\}$ . Let  $\{f(t)_i\}, i = 1, \dots, 1500$  to be the underlying unknown individual function representing the potassium trajectory. We can assume some functional form by using basis expansion to help us reconstruct its functional from the data points:  $f(t)_i = \sum_{k=1}^p \beta_{ik} h_{ik}(t)$ . We used cubic bs-spline bases.



**Fig. 2** Left Panel: Functional representation of potassium trajectories coloured according to outcome (red:death; blue: censored). Right Panel: First derivatives obtained from the functional representation.

In Fig.2 the potassium trajectories estimated according to all available measurements until the end of the observation period and the first derivatives are shown. The derivatives highlight that trajectories of patients who died show a higher variability than the ones from censored patients. When monitoring patients, however, new measurements arrive over time and ideally the cardiologist would need the trajectory to be updated according to the latest measurement.

Therefore, we introduced a dynamic functional representation suitable for this ongoing data collection.

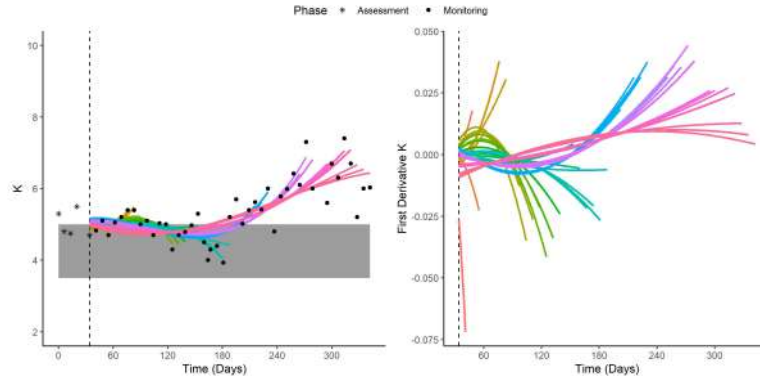
A preliminary estimate of  $f(t)_i$  was obtained using the Least Square Criterion by using the subject's first 5 measurements. Once we have obtained,  $\hat{\beta}_i$  for each of the subject, the **Sherman-Morrison Formula** was used to sequentially update the smoothing according to the 6th measurement:

$$\hat{\beta}_i = \hat{\beta}_i + \lambda W_{i(5)} h_{i6} (y_{i6} - h_{i6} \hat{\beta}_i), \tag{1}$$

where  $\lambda = \frac{1}{1+h_{i(6)}^T W_5 h_{i(6)}}$ ;  $W_{i(5)}$  is the inverse of  $(H_{i(5)}^T H_{i(5)})$ ;  $H_{i(5)}$  is the  $5 \times p$  matrix containing the bases functions evaluated on the first 5 measurements;  $h_{i6}$  and  $y_{i6}$  are the basis expansion of the time of measurement and measurement itself corresponding to the 6th measurement respectively.

Functional representation of potassium in HF

The formula allows to obtain the updated version of  $\hat{\beta}_i$  by using the prediction error:  $e_{i6} = y_{i6} - h_{i6}\hat{\beta}_i$  made by estimating the new observation with the old coefficients. After the new vector  $\hat{\beta}_i$  is obtained, also  $W_{i(6)}$  can be easily updated only using  $\lambda$ ,  $W_{i(5)}$  and  $h_{i(6)}$ . The formula can be used-recursively for any number of new measurements with no computational effort.



**Fig. 3** Estimated potassium trajectories for one subject (left panel) and its first derivative (right panel). Each line represents the functional representation obtained by a new measurement.

In order to adapt the procedure to the problem of estimation potassium trajectories, we did the following :

1. for the *assessment phase*, 4 bases were used for all subjects;
2. in the *monitoring phase*, a new basis was added when it resulted in a gain in terms of GCV.

In Fig.3, the result of the procedure for one subject is shown. In the *assessment phase*, no trajectory is drawn since it is used only as preliminary estimation. Around 6 months of follow-up, the subject develops hyperkalemia which persists from that time forward. Moreover, this dynamic functional representation allows also to obtain the derivative which can be useful to capture the change in the biomarker up to a specified point of time.

## 5 Conclusions

Monitoring potassium in Heart Failure with quantitative tools can assist cardiologists making informed decisions with regards to patients' treatment. The approach proposed allows to obtain a dynamic representation of potassium trajectories through Functional Data Analysis. It addresses measurement error and it is able

to integrate information of new measurements as they arrive. Further methodological research should be devoted in studying the interference of external covariates such as changes in the pharmacological therapy and the development of alert criteria based on the dynamic functional representation of potassium trajectories.

## References

1. Cooper, L.B., Savarese, G., Carrero, J.J., Szabo, B., Jernberg, T., Jonsson, Å., Dahlbom, C., Dahlström, U., Larson, A. and Lund, L.H.: Clinical and research implications of serum versus plasma potassium measurements. *European journal of heart failure*, **21(4)**, 536–537. (2019)
2. Ferreira, J.P., Butler, J., Rossignol, P., Pitt, B., Anker, S.D., Kosiborod, M., Lund, L.H., Bakris, G.L., Weir, M.R. and Zannad, F.: Abnormalities of potassium in heart failure. JACC state-of-the-art review. *Journal of the American College of Cardiology*, **75(22)**, 2836–2850. (2020)
3. Gasperoni, F., Ieva, F., Barbati, G., Scagnetto, A., Iorio, A., Sinagra, G., Di Lenarda, A.: Multi-state modelling of heart failure care path: A population-based investigation from Italy. *PloS one*, **12(6)**, e0179176. (2017)
4. Iorio A., Sinagra G., Di Lenarda A.: Administrative database, observational research and the Tower of Babel. *Int J Cardiol*. **284**,118–119. (2019)
5. Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis*. Springer Series in Statistics. Springer New York, 2005.
6. Savarese, G., Xu, H., Trevisan, M., Dahlström, U., Rossignol, P., Pitt, B., Lund, L.H. and Carrero, J.J.: Incidence, predictors, and outcome associations of dyskalemia in heart failure with preserved, mid-range, and reduced ejection fraction. *JACC: Heart Failure*, **7(1)**, 65–76. (2019)
7. Sherman, J., Morrison, W. J.: Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *Annals of Mathematical Statistics*. **21 (1)**, 124–127. (1950)

# Effect of lung transplantation on the survival of patients with cystic fibrosis: IMaCh contribution to registry data

## *Trapianto di polmone e sopravvivenza in fibrosi cistica: il contributo di IMaCh all'analisi di dati di registro*

Cristina Giudici, Nicolas Brouard and Gil Bellis

**Abstract** Using the Interpolated Markov Chain (IMaCh) software, we analyse transitions from different degrees of pulmonary function and mortality with and without transplantation, and compute life expectancy for cystic fibrosis patients, starting from different level of their respiratory health. We use data from the French Cystic Fibrosis Registry. The period of the study is 2008–2013 (7112 patients). Individuals enter the analysis in different years and their health status is regularly monitored. Health refers to the *volume* of air that can forcibly be blown out in first 1 second, after full inspiration (FEV<sub>1</sub>). Globally, we found higher mortality for those who have been transplanted but also lower probability of transition towards critical respiratory functions.

**Abstract** Lo studio analizza i dati del registro francese della fibrosi cistica utilizzando il software IMaCh - Interpolated Markov Chain. Vengono stimate le transizioni di salute e mortalità in presenza/assenza di trapianto di polmone e calcolata la speranza di vita a partire da diversi livelli di funzionalità polmonare. Quest'ultima è misurata attraverso il volume di aria espirata nel corso del primo secondo di una espirazione massima forzata (FEV<sub>1</sub>). Lo studio è svolto su 7112 pazienti presenti nel Registro tra il 2008 e il 2013. I pazienti che hanno subito un trapianto di polmone mostrano una mortalità più alta, ma anche una minore probabilità di transizione verso livelli critici di funzionalità polmonare.

**Key words:** Interpolated Markov Chain (IMaCh), Registry data, Cystic fibrosis

---

<sup>1</sup> Cristina Giudici, Sapienza University of Rome; email: [cristina.giudici@uniroma1.it](mailto:cristina.giudici@uniroma1.it)  
Nicolas Brouard, Institut National d'Etudes Démographiques (INED); email: [brouard@ined.fr](mailto:brouard@ined.fr)  
Gil Bellis, Institut National d'Etudes Démographiques (INED); email: [bellis@ined.fr](mailto:bellis@ined.fr)

## 1 Background

Cystic fibrosis (CF) is a multiorgan genetic disease that affects primarily the lungs and often leads to progressive respiratory insufficiency and premature death. CF is the most common hereditary disease among children in Europe (12,3 per 100.000). Median age of survival with CF has increased in several countries, and nowadays a steadily growing number of patients are adults.

Actually, among CF patients, morbidity and mortality is mostly caused by bronchiectasis, small airways obstruction, and progressive respiratory impairment (Nkam (2017)).

In the last decades, remarkable improvements in quality of life and clinical outcomes in patients with cystic fibrosis have been achieved thanks to innovative therapies. Although, for those patients who failed to respond to standard therapy lung transplantation remains the only treatment option with the potential to ameliorate symptoms, preserve quality of life, and extend life (Thabut et al. (2013)). In particular, bilateral lung transplantation has been shown to be an important therapeutic option for end-stage CF pulmonary disease (Hirche (2014)).

Several statistical models have been developed to identify prognostic factors in CF patients. One of the most significant predictors for survival is the Forced Expiratory Volume in 1 second (FEV1). Patients with low FEV1 may be referred for lung transplantation with the aim of improving their life expectancy and their quality of life (Nkam (2017)). Although, predicting life and health expectancy with or without transplantation is still a major issue.

In France, a national cystic fibrosis Registry (Registre français de la mucoviscidose) was created in 1992, and managed by the Institut National d'Etudes Démographiques (INED) since 1998. In October 2001 the Ministry of Health introduced systematic neonatal screening for cystic fibrosis on a national scale, and universal screening of newborn babies was introduced in 2003.

A number of studies have been carried out on this kind of data using classical statistical model (Nkam (2017)). Using the latest version of Interpolated Markov Chain (IMaCh) approach we analyse transitions from different degrees of pulmonary function and mortality and we compute life expectancy at different ages with and without transplantation.

## 2 Data: the French Cystic Fibrosis Register

Data are collected via questionnaires sent once a year to the healthcare centres cooperating with the Registry, in mainland France and Réunion Island. Individuals enter the Registry in different years and patient health status is regularly monitored. For each annual survey, each participating center reports on the patients seen at least once in the year. In particular, patients may visit their health center either a few times each year, or only once, or not every year according to their state of health. Globally,

the number of patients treated in the healthcare centres cooperating with the Registry increased from 2,168 in 1994 to 6,408 in 2013. Actually the Registry contains longitudinal data on more than 8,000 patients, which represents approximately 90% of all CF patients in France (Bellis et al. (2015)).

Among other analysis, the Forced Expiratory Volume in 1 second, after full inspiration (FEV<sub>1</sub>) is measured in patients aged 6 years old or over. In case of multiple measures during the same year, only the best value of FEV<sub>1</sub> is recorded. This measure is considered good when it is greater than 80% of the predicted value. On the contrary, we consider that patients having a FEV<sub>1</sub> lower than 40% of the predicted value may be referred for lung transplantation, with the aim of improving their life expectancy and their quality of life.

The data examined in this article concern the deaths of patients included in the Registry database from 2008 to 2013. The database provides up to 6 measures for each individual, one for each year. A total of 7,112 patients were registered between 2008 and 2013. As 1,308 patients do not have any measure of the FEV<sub>1</sub> (they are mostly children under the age of 6), only 5,804 patients have been considered for the analysis. During the period, the median age of patients increased, the transplantation rate increased as well and the death rate decreased (see table 1). In 2013 a total of 668 lung transplanted patients was still alive.

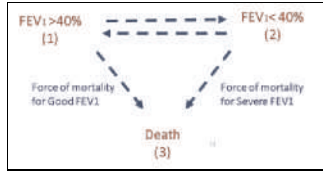
**Table 1:** *Characteristic of the Register Population*

	2008	2009	2010	2011	2012	2013
Patients (N)	5,419	5,700	5,685	6,077	6,277	6,408
Median age of patients (years)	17.2	18.2	18.7	19.2	19.7	20.3
Max age of patients (years)	75	77	80	87	88	87
Adults >=18 (%)	44.8	46.2	47.9	48.7	49.8	51.0
Mean FEV <sub>1</sub>	71.0	71.0	71.8	75.9	77.0	77.7
std FEV <sub>1</sub>	26.9	26.7	26.3	25.8	25.5	25.1
Death (N)	56	64	56	67	51	45
Mortality rate (x1000)	10.3	11.2	9.9	11.0	8.1	7.0
Median age at death (years)	29.7	25.2	29.1	26.5	32.3	34.6
Transplantations (N)	61	70	72	96	94	94
Transplantations (x1000)	11.3	12.3	12.7	15.8	15.0	15.0
Age at transplantation (year)	26.2	26.1	27.6	27.3	29.4	27.6

### 3 Methods

We estimated the age-specific flows of entry into and exit from critical respiratory functions, and the matrix of the transition probabilities between good (coded 1) and

severe (coded 2) FEV<sub>1</sub> and death (coded 3), using version 0.99 (r19) of the software IMaCh (Interpolation of Markov Chains)<sup>1</sup>.



**Figure 1:** States and possible transitions among states

The probability for an individual aged  $x$ , observed in the health state  $i$  to find him/herself in state  $j$  after one year is indicated by  $p_{ij}^x$  and the transition probabilities are estimated based on a series of  $3 \times 3$  matrices:

$$P_{ij}^x = \begin{pmatrix} p_{11}^x & p_{12}^x & p_{13}^x \\ p_{21}^x & p_{22}^x & p_{23}^x \\ 0 & 0 & 1 \end{pmatrix}$$

The first and the second rows represent transitions for individuals who begin the interval respectively in good and severe FEV<sub>1</sub>. The third row represents the absorbing state of death. The probabilities of transition are then parameterized using the following multinomial logit model:

$$\ln \frac{p_{ij}^x}{p_{ii}^x} = \alpha_{ij} + \beta_{ij}x + \gamma_{ij}z$$

Where  $z$  is a time varying covariate which could influence the transition from one state to another or to death. This covariate corresponds to the answer to the question: “Was the patient already transplanted (1) at the time of each FEV<sub>1</sub> measure or not (0)?”.

A complicating factor was that, if patients have had the opportunity to be transplanted in year  $t$ , the FEV<sub>1</sub>( $t$ ) might be measured before or after transplantation, but this information was not recorded. Moreover, some patients might have no measure at all during the year of transplantation.

<sup>1</sup> IMaCh is a publicly available computer program introduced by Lièvre, Brouard and Heatcote (2003) and mostly used for the estimation of Health Expectancy from longitudinal surveys. It allows to estimate transition probabilities by the method of maximum likelihood, using a discrete time embedded Markov chain approach. Transitions are supposed to occur at any time and death is always an additional competing risk. See Brouard (2019) for theory and applications. Several applications can also be found in literature, see for example Molla and Madans (2008) and Giudici et al. (2013).

To overcome this problem, we set up a decision rule based on the expected increase in the ventilation capacity just after the surgery.

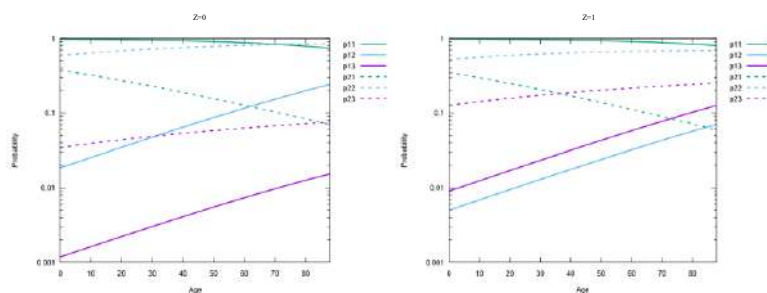
## 4 Results

Figure 2 shows the transition probabilities from different initial state of respiratory function for non-lung transplanted ( $z = 0$ ) and lung transplanted ( $z = 1$ ) patients.

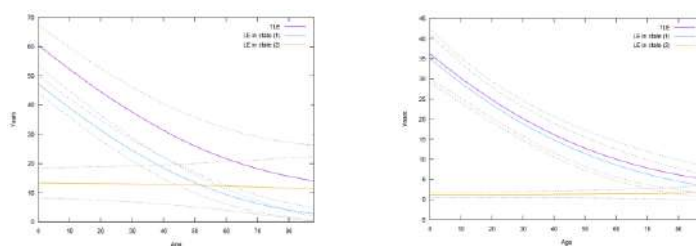
As expected, the probability of dying is always higher among those with severe FEV<sub>1</sub> ( $p_{23}$  vs  $p_{13}$ ) and for transplanted patients ( $z=1$  vs  $z=0$ ), and the probability of recovering ( $p_{21}$ ) is decreasing with age.

The transplantation modifies mainly the transition rate towards severe FEV<sub>1</sub> ( $p_{12}$ ) which is lower for those who have been transplanted.

On the basis of transition probabilities estimates, IMaCh computes life expectancy for patients in state 1 and 2 by age, given that they were in that state initially. Globally, total life expectancy at birth for the analysed population is 54 years and, at any age, patients may expect to live 11 years in critical respiratory health. Life expectancy for lung transplanted patients is almost totally free from severe respiratory distress. At the age of 27 (median age of transplantation) they may expect to live 21 years (Figure 3).



**Figure 2:** Conditional probabilities to be observed in state  $j$  being in state  $i$ , 12 months before ( $z = 0$ ) and after ( $z = 1$ ) lung transplantation



**Figure 3:** Life expectancies according to the state of health for non-transplanted (on the left) and transplanted (on the right) patients



## 5 Conclusions

The aim of our analysis was to compute life and health expectancy for CF patients with and without lung transplantation, starting from Registry data and using the latest version of IMACh software.

For a given year, registry data do not necessarily provide the exact date at which respiratory health is measured and neither the date of the surgery in case of transplantation. To overcome this lack of information we set up a decision rule based on the expected increase in the ventilation capacity just after the surgery, and create a time dependent covariate which indicates, for each year, if the patient was already transplanted at the time of the respiratory measure or not.

Globally, we found higher mortality for those who have been transplanted but also lower probability of transition towards critical respiratory functions.

Our approach allows a more complex life expectancy analysis than classical approach on median age at death and contribute to exploit CF registry data. Nonetheless, the study is not without limitation. Indeed, CF prognostic factor are very complex. In particular, female gender is recognized in several studies as having a negative impact on survival. Moreover, in addition to compromised respiratory function, mortality in CF may be also related to liver complications or compromised nutritional status (Buzzetti et al. (2008)). Finally, we did not distinguish single lung and double or heart-lung transplantation.

To conclude, the French Registry of Cystic Fibrosis provides an excellent basis for measuring and monitoring mortality and survival of CF patient in France and our study suggests a useful approach to the registry data analysis.

## References

1. Bellis, G., Dehillotte C., Lemonnier L.: *Registre Français De La Mucoviscidose. Bilan Des Données 2014. Vaincre la Mucoviscidose et Institut national d'études démographiques, INED, Paris (2016)*
2. Brouard N.: *Theory and Applications of Backward Probabilities and Prevalences in Cross-Longitudinal Surveys*. In Rao, A.S.R.S., Rao, C.R. (eds). *Handbook of Statistics*, Elsevier, Vol. 40, pp. 435-486 (2019) doi 10.1016/bs.host.2018.11.009
3. Buzzetti R. et al.: *An overview of international literature from cystic fibrosis registries:1. Mortality and survival studies in cystic fibrosis*. *J. Cyst. Fibros.* 8, 229-237 (2009)
4. Giudici, C., Arezzo, M. F., & Brouard, N.: *Estimating health expectancy in presence of missing data: an application using HID survey*. *Stat. Methods Appt.*, 22(4), 517-534 (2013)
5. Hirche, T.O. et al.: *Practical guidelines: lung transplantation in patients with cystic fibrosis*. *Pulm. Med.* (2014)
6. Lievre, A., Brouard, N., & Heathcote, C.: *The estimation of health expectancies from cross-longitudinal surveys*. *Math. Popul. Stud.* 10(4), 211-248 (2003)
7. Molla, M. T., & Madans, J. H.: *Estimating healthy life expectancies using longitudinal survey data: methods and techniques in population health measures*. *Vital Health Stat.* 2 (146), 1-24 (2008)
8. Nkam, L., Lambert, J., Latouche A., Bellis G., Burgel P.R., Hocine M.N.: *A 3-year prognostic score for adults with cystic fibrosis*. *J. Cyst. Fibros.* 16 (2017)
9. Thabut, G., Christie J.D., Mal H., Fournier M., Brugiere O., Leseche G., et al.: *Survival benefit of lung transplant for cystic fibrosis since lung allocation score implementation*. *Am J Respir Crit Care Med.* 187 (2013)

# Categories and Clusters to investigate Similarities in Diabetic Kidney Disease Patients

## *Categorie e clusters per investigare la similarità fra i pazienti affetti da nefropatia diabetica*

Veronica Distefano, Maria Mannone, Claudio Silvestri, and Irene Poli

**Abstract** Heterogeneous responses to therapeutical treatments across patients and over time is a common and serious problem for several diseases. Precision medicine research focuses in developing procedures to take treatment decisions for the individual patient using all the information available for the patient, including demographic and clinical variables and the response to the followed treatment. In this paper we adopt category theory and the cluster analysis to achieve insight into specific disease pathways and patient subgroups. We analyze a longitudinal dataset of patients affected by diabetic kidney disease (highly prevalent in type 2 diabetes) and monitored at different time points in the response to various treatment regimes. This analysis, based on distances between patients in different time points and in time evolution, divides patients into clusters that show the relevant role of some variables in affecting the progress of the disease.

**Abstract** *L'eterogeneità nella risposta a trattamenti terapeutici tra pazienti e nella sua evoluzione temporale rappresenta un problema comune a molte malattie. La medicina di precisione si propone di sviluppare metodologie di supporto alle decisioni di trattamento per ogni singolo paziente usando tutte le informazioni disponibili sul paziente, includendo perciò le variabili demografiche, le variabili cliniche e la risposta ai trattamenti. In questo lavoro noi adottiamo la teoria delle categorie e la cluster analysis per ottenere elementi informativi su diversi sviluppi della*

---

Veronica Distefano

European Centre for Living Technology, Ca' Foscari University of Venice, Italy, e-mail: veronica.distefano@unive.it

Maria Mannone

European Centre for Living Technology, Ca' Foscari University of Venice, Italy, and Department of Mathematics and Computer Sciences, University of Palermo, Italy, e-mail: maria.mannone@unive.it, mariacaterina.mannone@unipa.it

Claudio Silvestri

Dipartimento di Scienze Ambientali, Informatica e Statistica and European Centre for Living Technology, Ca' Foscari University of Venice, Italy, e-mail: claudio.silvestri@unive.it

Irene Poli

European Centre for Living Technology, Ca' Foscari University of Venice, Italy, e-mail: irenpoli@unive.it

*malattia nel tempo e sull'esistenza di gruppi di pazienti con comportamenti diversi. Analizziamo un insieme di dati longitudinali relativi a pazienti con diabete di tipo 2 e complicazioni renali, osservati in differenti punti temporali con riferimento a un insieme di variabili e alla risposta a differenti trattamenti. Questa analisi, basata sulle distanze tra pazienti in successivi punti temporali, individua clusters di pazienti e il ruolo di certe variabili nell'influenzare la progressione della malattia.*

**Key words:** category theory, cluster analysis, DKD disease

## 1 Introduction

Precision medicine greatly profits from statistical and mathematical techniques to envisage differences and similarities between patients, their treatments, and outcomes. The increasing interest on these topics leads to a more extended development and use of methodological procedures. One of the research areas where precision medicine is currently considered is the treatment of diabetes of type 2, with kidney clinical complication (DKD). DKD patients show significative heterogeneity in the disease progress, and thus there is a clinical need for individualized treatments. Patient clustering [1, 6] appears as a useful explorative way to achieve some information about the evolution of the disease. We focus on a longitudinal dataset of DKD patients from the DC-ren project,<sup>1</sup> and we aim to highlight similarities between patients, in terms of initial conditions, therapeutical treatments, and responses to the treatments. The dataset consists of mixed data, which include quantitative and qualitative variables, concerning clinical-laboratorial data, socio-demographical aspects, and different treatment responses.

In this paper, we aim to derive clusters of patients observed in different time points, where the clustering approach will be based on distances between patients. The evaluation of distance between patients is an essential step to investigate patients' characteristic profiles. In order to derive an integrated approach to visualize and highlight dynamic patterns of the disease, we adopt the category theory, an abstract branch of mathematics, developed to formalize the concept of *transformations between transformations* in a flexible way [4]. This approach is used in a variety of areas of research, which include biology, physics, chemistry, and computer science [2], and it is useful to investigate problems in an abstract way and visualize connections and temporal dynamics. In this framework, a category is constituted by objects (points) and morphisms between them (arrows), and provides a clear way to model similarity and equivalence. One of the most powerful ideas of category theory is the notion of *functor*, which can be thought as a generalization of the concept of function. A functor maps objects and morphisms from a category to objects and morphisms of another category. A mapping between functors is a natural transformation, and leads to generate nested structures. The novelty of this work consists in deriving an integrative approach to visualize and explore patterns in longitudinal

---

<sup>1</sup> <https://dc-ren.eu/>

data. Longitudinal data on DKD are collected in visits at baseline and in subsequent follow-ups. We focus on the first three time points  $t_0, t_1, t_2$ , comparing the response to the therapy treatment of the set of patients. In order to evaluate the heterogeneity of patients at different time points and how this heterogeneity evolves in time, we evaluate matrices of distances between patients. We introduce  $D(t_0)$  as the matrix of distances between patients at time  $t_0$ ;  $D(t_1)$  and  $D(t_2)$  as the matrices of distances between patients at  $t_1$  and  $t_2$ , respectively. We also introduce  $D(t_0, t_1)$  as the matrix of distances between  $D(t_0), D(t_1)$ , and  $D(t_1, t_2)$  as the matrix of distances between  $D(t_1), D(t_2)$ . With this approach adopting category theory and cluster analysis, we achieve a small set of clusters, which represent the most similar patients in their behavior in time and response to the therapy treatments. These results will be very helpful in building a decision system which allows to derive the best treatment for each patient. The paper is organized as follows. In Section 2, we adopt elements of the category theory to envisage a strategy for deriving matrices of distance between patients, and in Section 3, we build clusters of patients with respect to the progression of the disease.

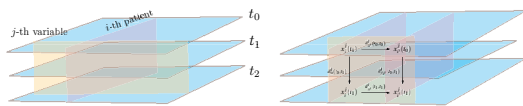
## 2 Method

In our analysis, we will consider a set of data concerning  $n$  patients,  $p$  variables, and three time points  $t_0, t_1, t_2$ . Each element is the observation  $x_i^j(t_k)$ , where:  $i$  indicates the individual (the patient),  $i = 1, \dots, n$ ;  $j$  indicates the variable ( $X^j$ ) on which we observe  $x_i^j$ ,  $j = 1, \dots, p$ ;  $k$  indicates the time point,  $k = 0, 1, 2$ , and thus there are three time points:  $t_0$  (also called *baseline*), and  $t_1, t_2$  (the first and second *follow-ups*, respectively). We define two kinds of distances: distance  $d_{i,i'}^j(t_k, t_k)$  between observations of variable  $j$  at the time  $k$  for different patients  $i, i' = 1, \dots, n$  (horizontal distance); distance  $d_{i,i}^j(t_k, t_{k'})$  between observations of variable  $j$  through different times  $k, k' = 0, 1, 2$  for patient  $i$  (vertical distance). Having built a set of distance values, we can achieve an *enriched category* with metrics in  $\mathbb{R}$  [4]. More precisely, we can describe observations and distances as an enriched *double category* whose objects are  $x_i^j(t_k)$  and whose morphisms are vertical and horizontal distances, as in the following diagram 1.

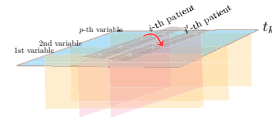
$$\begin{array}{ccc}
 x_i^j(t_0) & \xrightarrow{d_{i,i'}^j(t_0,t_0)} & x_{i'}^j(t_0) \\
 \downarrow d_{i,i}^j(t_0,t_1) & & \downarrow d_{i',i'}^j(t_0,t_1) \\
 x_i^j(t_1) & \xrightarrow{d_{i,i'}^j(t_1,t_1)} & x_{i'}^j(t_1)
 \end{array} \tag{1}$$

Observations for the same variable through time and patients constitute a lattice. Horizontal composition shows comparisons between multiple patients at the same time; vertical composition shows comparisons of the same patient through time. Mappings from variables to variables can be formalized as *functors*. There is a lat-

tice for each variable. These lattices are the vertical sections in the representation of Figure 1 (left). In fact, observations and distances for each patient are represented by transversal sections (Figure 1, right). Functors map a lattice into the other. The outcomes (success/unsucces) of the therapeutical treatment are evaluated through the variations of a response variable. We compute the dissimilarity matrices, to compare patients and their disease time evolution. We evaluate as a first step the information in the variable values of each patient at certain times. The  $i$ -th patient  $p_i(t)$  is described by the vector  $[x_i^1(t_k), x_i^2(t_k), \dots, x_i^p(t_k)]$ . Each element of the dissimilarity matrix is the distance between patients, as described in Figure 2.



**Fig. 1** Representation of the dataset.



**Fig. 2** Representation of matrix elements of  $D(t_k)$ .

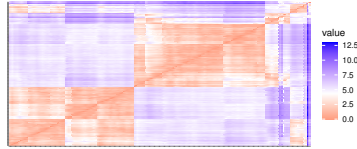
To measure the distance, we adopt the coefficient  $s(p_i, p_{i'})$  proposed in [5], which takes the following expression:  $s(p_i, p_{i'}) = \frac{\sum_{j=1}^p s^j(x_i^j, x_{i'}^j) \delta^j(x_i^j, x_{i'}^j)}{\sum_{j=1}^p \delta^j(x_i^j, x_{i'}^j)}$ , where:  $p_i$  is the  $i$ -th patient,  $p_{i'}$  is the  $i'$ -th patient,  $s^j$  is based on the Gower similarity [3],  $x_i^j$  is the value of the  $j$ -th variable for the  $i$ -th patient,  $x_{i'}^j$  is the value of the  $j$ -th variable for the  $i'$ -th patient;  $\delta^j(x_i^j, x_{i'}^j)$  is a coefficient to be 0 if  $p_i$  or  $p_{i'}$  have a missing value for the  $j$ -th variable, and 1 if they do not. The elements of the dissimilarity matrix, describing the *distance between patients*, are computed as  $d(p_i(t_k), p_{i'}(t_k)) = 1 - s(p_i(t_k), p_{i'}(t_k))$ . The dissimilarity matrices, computed at different time points, are then  $D(t_0), D(t_1), D(t_2)$ , and on these matrices we build matrices of distances of distances  $D(t_0, t_1), D(t_1, t_2)$ .

### 3 Results

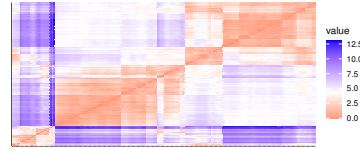
In order to evaluate the behaviors of patients with respect to variables and therapeutical treatments, we build clusters of patients according to the measures of distance described in Section 2. The dataset, used in the DC-ren project and based on the PROVALID study,<sup>2</sup> is about diabetic kidney disease (DKD), with  $n = 241$  patients observed in longitudinal way in three data points and  $p = 21$  variables, which include clinical and social-demographic variables and treatment responses. In order to find clusters of patients, we build matrices of distances between patients at the three data points,  $D(t_0), D(t_1), D(t_2)$ , and, based on them, matrices of distances between distances,  $D(t_0, t_1), D(t_1, t_2)$ . The data are collected at baseline ( $t_0$ ) and follow-ups

<sup>2</sup> SysKid project <https://cordis.europa.eu/project/id/241544/>

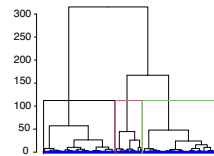
$(t_1, t_2)$ , and contain information on the therapy adopted. The patients received four different treatments,  $a_1, a_2, a_3, a_4$ . The outcome is a binary variable, indicating *success* or *unsuccess* of the therapeutical treatment. In this paper, we adopt the hierarchical clustering method without deciding *a priori* the number of clusters.



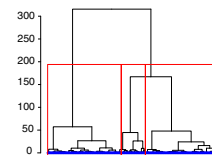
**Fig. 3** Matrix  $D(t_0, t_1)$ . (Darker blue indicates a greater dissimilarity).



**Fig. 4**  $D(t_1, t_2)$ . (Darker blue indicates a greater dissimilarity).



**Fig. 5** Dendrogram of  $D(t_0, t_1)$ .

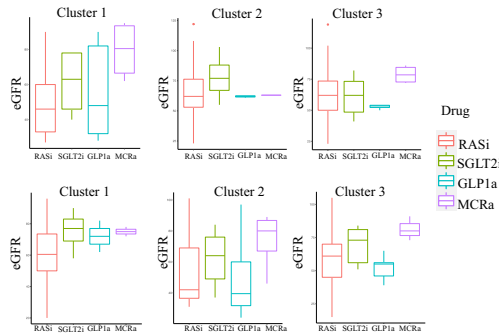


**Fig. 6** Dendrogram of  $D(t_1, t_2)$ .

To obtain  $D(t_0, t_1)$ , represented in Figure 3, we compute  $D(t_0)$  associated with the distances between patients at  $t_0$ , and matrix  $D(t_1)$  associated with the distances between patients at  $t_1$ , both evaluated with Gower distance. To obtain  $D(t_1, t_2)$ , represented in Figure 4, we follow the same procedure. The sequence of clusters obtained by using a hierarchical clustering is visualized through the Ward dendrogram by using the matrices  $D(t_0, t_1)$  and  $D(t_1, t_2)$ , and we achieve  $K = 3$  as the optimal number of clusters. The height of the dendrogram is the distance between the clusters, as shown in Figures 5 and 6. In our study, the comparison between four linkage methods (average, single, complete, ward), shows that the Ward-type linkage method identifies the better clustering structure. The Ward method finds the distance between two clusters as the minimum within-cluster variance. In Table 1, we can see that the mean values of the vast majority of the variables within clusters 2 and 3 in  $D(t_0, t_1)$  and the variables within clusters 1 and 3 in  $D(t_1, t_2)$  do not show significant differences. On the contrary, we notice relevant differences in cluster 1 in  $D(t_0, t_1)$  and cluster 2 in  $D(t_1, t_2)$ . In particular, the patients in these clusters present higher body mass index (BMI) and triglyceride values, while the levels of HbA1c, cholesterol HDL, and estimated Glomerular Filtration Rate (eGFR) are lower. In particular, the mean eGFR of these patients is  $<60 \text{ mL/min/1.73 m}^2$ , which increases the risk factor to cardiovascular disease. Figure 7 shows the boxplots of clusters of patients for different treatments and different eGFR levels. The 47.3% of patients in cluster 1 in  $D(t_0, t_1)$ , is present in cluster 2 in  $D(t_1, t_2)$ . These patients are mostly treated with  $a_1$  and  $a_3$ . The 66% of them had *unsuccessful* outcome in  $D(t_0, t_1)$  and  $D(t_1, t_2)$ . From this analysis, it is possible to find groups of patients with behaviors different between them, and with respect to the time evolution of the disease.

**Table 1** Description of the clusters of distances between patients:  $D(t_0, t_1)$  and  $D(t_1, t_2)$ , with variable mean values and standard deviation values within within brackets.

Variables	$D(t_0, t_1)$			$D(t_1, t_2)$		
	cluster1 (n= 38)	cluster2 (n= 100)	cluster3 (n= 103)	cluster 1 (n= 103)	cluster 2 (n= 34)	cluster 3 (n= 104)
eGFR	55.13 (± 21.59)	65.65 (± 18.34)	62.73 (± 17.71)	62.14 (± 17.88)	57.24 (± 21.90)	59.25 (± 17.27)
Age	66.55 (± 11.38)	69.55 (± 7.81)	66.10 (± 8.35)	68.03 (± 7.76)	65.82 (± 11.52)	65.38 (± 8.60)
BMI	34.74 (± 9.74)	30.87 (± 6.52)	30.27 (± 5.81)	31.24 (± 6.76)	33.90 (± 10.45)	30.10 (± 5.98)
Body weight	96.68 (± 22.35)	85.43 (± 16.35)	85.32 (± 13.97)	86.38 (± 17.00)	94.62 (± 24.09)	84.88 (± 14.05)
Systolic	137.29 (± 15.80)	137.08 (± 14.22)	135.22 (± 14.44)	136.99 (± 14.07)	135.29 (± 16.19)	133.73 (± 14.45)
Diastolic	76.87 (± 10.36)	77.57 (± 9.19)	75.72 (± 10.34)	75.97 (± 8.72)	76.88 (± 10.51)	76.20 (± 9.22)
Blood_glucose	164.63 (± 78.84)	148.84 (± 50.50)	146.02 (± 48.59)	150.81 (± 62.27)	162.97 (± 74.99)	146.05 (± 44.79)
HbA1c	7.65 (± 1.58)	7.16 (± 1.16)	7.28 (± 1.30)	7.35 (± 1.41)	7.69 (± 1.24)	7.12 (± 1.12)
Serum_creatinine	1.37 (± 0.46)	1.06 (± 0.33)	1.08 (± 0.31)	1.13 (± 0.41)	1.29 (± 0.46)	1.16 (± 0.39)
Serum_cholesterol	194.03 (± 53.31)	184.37 (± 51.58)	181.45 (± 43.61)	178.25 (± 43.29)	184.32 (± 42.98)	178.80 (± 43.64)
Serum_cholesterol LDL	99.58 (± 32.36)	99.09 (± 35.50)	98.68 (± 32.98)	94.09 (± 28.24)	92.11 (± 33.25)	95.27 (± 27.59)
Serum_cholesterol HDL	46.26 (± 14.85)	52.18 (± 15.84)	49.07 (± 12.77)	50.21 (± 15.96)	46.94 (± 22.31)	49.81 (± 13.70)
Serum_triglycerides	211.39 (± 124.06)	176.37 (± 103.07)	172.66 (± 119.14)	180.36 (± 168.67)	247.88 (± 152.54)	167.35 (± 81.73)
Serum_potassium	4.61 (± 0.60)	4.56 (± 0.48)	4.55 (± 0.46)	4.53 (± 0.50)	4.62 (± 0.60)	4.52 (± 0.55)
Hemoglobin	13.64 (± 1.81)	13.48 (± 1.45)	13.74 (± 1.45)	13.39 (± 1.63)	13.93 (± 1.53)	13.33 (± 1.68)
Serum_albumin	4.38 (± 0.48)	4.51 (± 0.43)	4.53 (± 0.54)	4.44 (± 0.50)	4.57 (± 0.63)	4.48 (± 0.47)
Crp	0.51 (± 0.53)	0.61 (± 1.23)	0.51 (± 1.08)	1.01 (± 3.01)	0.48 (± 0.47)	0.51 (± 1.04)
Mean uacr	173.37 (± 379.24)	48.63 (± 124.42)	23.58 (± 33.45)	69.52 (± 180.33)	168.19 (± 364.34)	31.06 (± 66.02)



**Fig. 7** Box plots of clusters of patients for the different treatments and different eGFR levels. Top:  $D(t_0, t_1)$ . Bottom:  $D(t_1, t_2)$ . Drug  $a_1$  is RASi,  $a_2$  is SGLT2i,  $a_3$  is GLP1a, and  $a_4$  is MCRa.

**Acknowledgements** This research activity is part of the project DC-ren that has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 848011. We are grateful to the researchers of the European Centre for Living Technology (ECLT) for very helpful discussions and suggestions.

**References**

1. Amiri, S., Clarke, B. S., Clarke, J. L.: Clustering categorical data via ensembling dissimilarity matrices. *J. Comput. Graph. Statist.* **27** (1): 195–208 (2017)
2. Carlsson G., Mémoli, F.: Classifying Clustering Schemes. *Foundations of Computational Mathematics* **13**, 221–252 (2013)
3. Gower J.: A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–871 (1971)
4. Grandis, M.: *Higher Category Theory*. World Scientific, Singapore (2020)
5. Hummel, M., Edelmann, D., and Kopp-Schneider, A.: Clustering of samples and variables with mixed-type data. *Plos One* **12** (11), e0188274 (2017)
6. Park, S., Xu, H., Zhao, H.: Integrating Multidimensional Data for Clustering Analysis With Applications to Cancer Patient Data. *Journal of the American Statistical Association* **116** (533), 14–26 (2021)

## 4.22 New perspectives in models for multivariate dependency



# Parsimonious modelling of spectroscopy data via a Bayesian latent variables approach

## *Modellazione parsimoniosa di dati spettroscopici mediante un approccio Bayesiano a variabili latenti*

Alessandro Casa, Tom F. O'Callaghan and Thomas Brendan Murphy

**Abstract** Recent years have seen increased attention in the dairy sector towards cattle feeding regimens with grass-based one leading to healthier and more expensive products, thus more susceptible to adulteration. Hence, statistical tools guaranteeing milk authenticity and discriminating samples from different diets are needed. Spectroscopy data are routinely used in this context, nonetheless they introduce challenges, such as high-dimensionality and the peculiar wavelengths relationships, that have to be tackled. In this work a modification of the standard Factor Analysis is proposed. The data are mapped into a low-dimensional latent space while clustering the observed variables thus highlighting redundancies and providing more parsimonious summaries of the data and insights on diet induced differences in the milk.

**Abstract** *Nel settore lattiero-caseario si sta assistendo ad un crescente interesse verso l'alimentazione degli animali. L'allevamento al pascolo è considerato sinonimo di prodotti più salutari, venduti ad un prezzo più elevato e suscettibili a sofisticazione. Sono quindi necessari strumenti statistici che garantiscano l'autenticità del latte e che discriminino campioni associati a diete diverse. I dati spettroscopici, utilizzati spesso in questo contesto, introducono alcune difficoltà da affrontare quali l'elevata dimensionalità e le particolari relazioni tra diverse lunghezze d'onda. In questo lavoro proponiamo una modifica dell'analisi fattoriale standard; si riduce la dimensionalità dei dati e si ottiene una partizione delle variabili capace di evidenziare ridondanze e di fornire informazioni sull'impatto sul latte di diete diverse.*

**Key words:** dairy science, chemometrics, factor analysis, clustering

---

Alessandro Casa, Thomas Brendan Murphy  
Vistamilk Research Centre, School of Mathematics & Statistics, University College Dublin,  
Belfield, Dublin 4, Ireland e-mail: alessandro.casa@ucd.ie, brendan.murphy@ucd.ie

Tom F. O'Callaghan  
Vistamilk Research Centre, School of Food & Nutritional Sciences, University College Cork,  
College Rd, Cork, Ireland e-mail: tom\_ocallaghan@ucc.ie

## 1 Introduction

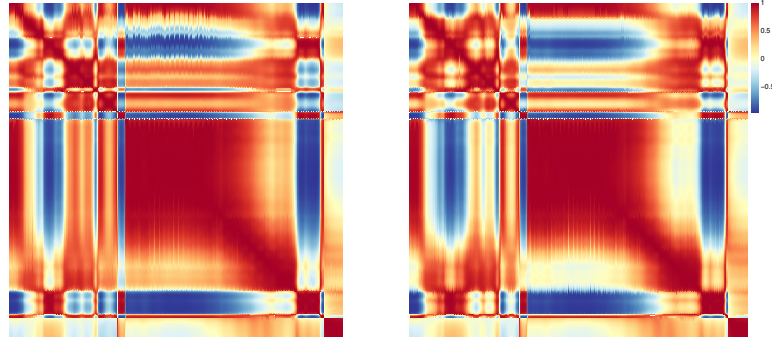
Recently increased consumer awareness has led to radical changes in those industry sectors producing foodstuffs of animal origin. Dairy farming has been especially involved in this transition with cattle feeding regimen attracting particular attention. In fact pasture based feeding has been demonstrated to lead to improvements in the products quality and it is regarded as more respectful of the animal well-being. As such these products are demanding higher prices, thus being more susceptible to food adulteration. Therefore proper methods being able to distinguish between milk from pasture and non-pasture diets are needed.

In this framework, infrared spectroscopy techniques represent a cheap, rapid and non-disruptive way to collect large amounts of data that have been already fruitfully used to determine different milk characteristics. Nonetheless, a thorough exploration of the usefulness of spectroscopy data to authenticate cow feeding regimens is still missing. From a statistical standpoint these data introduce some challenges, such as high-dimensionality and the peculiar correlations among the wavelengths, that have to be addressed. *Factor analysis* (FA, [2]) is particularly useful for tackling some of these challenges given its ability to map the observed data into a lower-dimensional latent space while simultaneously aiming to explain the correlations among the features. Nonetheless, even if it effectively reduces the dimensionality of the data, standard FA does not account for possible redundancies in the features, often witnessed in spectral data as confirmed by the blocky structures in Figure 1.

For this reason, in this work, we propose a modification of the standard FA model which, by producing a partition of the wavelengths, allows to detect their intrinsic redundancies. The clustering of the variables can subsequently be used to gain useful insights about similarly behaving spectral regions. The rest of the paper is structured as follows. In Section 2 we describe the data which motivates our proposal which is in turn introduced in Section 3 along with model estimation and model selection strategies. Empirical results are finally reported in Section 4.

## 2 Dairy diet data

The data we consider in this work consists of 4320 mid-infrared spectra of milk samples collected weekly over a three year time span (from 2015 to 2017) produced by Holstein-Freisian cows on three dietary treatments. The treatments included grass (GRS) and clover (CLV) based outdoor feeding strategies and total mixed ration (TMR) based one where, on the contrary, cows are maintained indoors and where the nutrients are combined in a single mix of concentrates, grass and maize silage. Given their strong compositional similarities, in our work we merged together the first two classes into a pasture-based diet group. The total number of cows involved is equal to 120 with multiple measurements for each single animal. The samples considered are the ones collected mainly in the summer months, representing the period of milk production with the highest prevalence of grass growth. Finally note



**Fig. 1** Sample correlation matrices computed on the milk samples produced by pasture fed cows (on the left) and total mixed ration fed cows (on the right).

that, for each sample, 1060 reflectance measurements have been collected in the region spanning from  $925\text{cm}^{-1}$  to  $5010\text{cm}^{-1}$ .

### 3 Parsimonious Bayesian Factor Analysis

Standard FA parsimoniously summarizes dependence structures among high dimensional observations. Let  $X = \{x_1, \dots, x_n\}$ , with  $x_i \in \mathbb{R}^p$ , the set of observed data assumed to be centered. Factor analysis models each observation  $x_i$  as follows

$$x_i = \Lambda u_i + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

where  $\Lambda \in \mathbb{R}^{p \times K}$  is the loading matrix,  $u_i \in \mathbb{R}^K$  are the factor scores with  $u_i \sim N_K(0, \mathbb{I})$  and  $K$  the number of factors while  $\varepsilon_i \sim N_p(0, \Psi)$  with  $\Psi$  diagonal. Therefore marginally  $x_i \sim N(0, \Sigma = \Lambda \Lambda^T + \Psi)$  implying that the correlation between the original variables is modelled through  $\Lambda$ . Moreover, since in practical applications  $p > K$ , the model entails a parsimonious decomposition of  $\Sigma$ .

The detection of uninformative features has often been tackled in the FA framework but, to the best of our knowledge, possible redundancy has not been addressed yet. A variable is defined as redundant when it carries similar information with respect to the one provided by other variables, usually because of their strong correlations. In order to account for it we introduce a model where some of the variables are mapped into the factor latent space by means of the same loading coefficients, providing information about possible grouping structures in the features. The proposed model is defined as

$$x_i = Z \Lambda_c u_i + \varepsilon_i, \quad i = 1, \dots, n \quad (2)$$

with  $x_i, u_i$  and  $\varepsilon_i$  defined above while  $Z \in \mathbb{R}^{p \times G}$  is a latent allocation matrix where  $G$  is the number of variables clusters and  $z_{jg} = 1$  if the  $j$ -th variables belongs to the  $g$ -th group and 0 otherwise. Finally  $\Lambda_c \in \mathbb{R}^{G \times K}$  is the matrix whose  $g$ -th row contains

the unique and representative loadings for the  $g$ -th variable cluster. Note that, as a consequence,  $\tilde{\Lambda} = Z\Lambda_c$  has duplicate row values; we believe that this represents a sensible way to account for redundancy by constraining the variables carrying the same information to share the same relations with the latent factors.

The distributional results hold as before with  $(x_i|Z) \sim N(0, \tilde{\Sigma} = \tilde{\Lambda}\tilde{\Lambda}^T + \Psi)$  entailing an even greater reduction in the number of covariance parameters to estimate. Moreover the estimation of  $Z$  allows us to obtain a clustering of the variables which, from an interpretative standpoint, might give relevant information about the phenomenon under study.

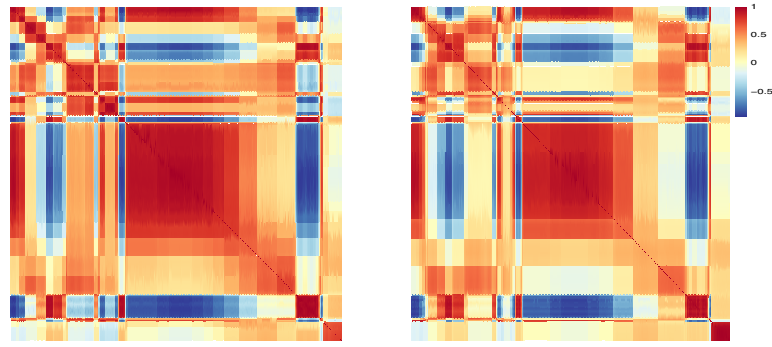
Different strategies might be adopted to estimate the parameters in (2). In this work, we take a Bayesian approach by assuming standard independent prior distributions for  $\Lambda_c$ ,  $u_i$  and  $\Psi$ . The corresponding hyperparameters are chosen in order to induce uninformative and to avoid the *Heywood problem*. Some words of caution are required for the allocation matrix  $Z$ . Here we consider a product partition model (PPM, see [1]) which assumes that the prior probability is expressed as

$$\pi(\mathbf{c}) \propto \prod_{g=1}^G \rho(C_g)$$

where  $\mathbf{c} = \{C_1, \dots, C_G\}$  is a clustering of the indices  $\{1, \dots, p\}$  with  $C_g$  containing the ones belonging to the  $g$ -th cluster. More specifically, we consider  $\pi(\mathbf{c}) \propto \alpha_Z^G \prod_{g=1}^G (|C_g - 1|)!$  representing a common choice in a Bayesian clustering framework as it shares strong connections with the widely used Dirichlet process prior. With a slight abuse of notation and considering the correspondence between the representation of a partition  $\mathbf{c}$  as a collection of disjoint subsets  $\{C_1, \dots, C_G\}$  and the one via the allocation matrix  $Z$ , we write  $Z \sim \text{PPM}(\alpha_Z)$ .

Due to the conditionally conjugate nature of the prior distributions considered we adopt a Metropolis withing Gibbs algorithm in order to sample from the posterior distribution. In the Metropolis step, in order to sample the allocation matrix  $Z$ , we modify one of the moves of the *allocation sampler* proposed by [4] in order to enhance a faster exploration of the partition space.

In the proposed framework both  $K$  and  $G$  have to be chosen. Several approaches have been studied, often relying on model comparisons using information criteria. Here, in order to avoid computationally infeasible exhaustive global searches over wide ranges of values for  $K$  and  $G$ , we propose an ad hoc initialization strategy providing a configuration  $(K_{init}, G_{init})$  to be used as the starting point of a local search. The procedure resorts to model-based clustering techniques (see [3] for a recent review) to group the rows of  $\Lambda$  in (1) in order to mimick the repeated rows structures of  $\tilde{\Lambda}$ ; the different estimated models are then compared via BIC. Note that some unreported sensitivity analyses on synthetic data showed that this strategy produces reasonable configurations for the number of factors and clusters and that a global search is not strictly needed when covariance reconstruction is the aim.



**Fig. 2** Estimated correlation matrices computed on the milk samples produced by pasture fed cows (on the left) and TMR fed cows (on the right).

## 4 Empirical results

In this section, we show the performances of the proposed method when applied to the data introduced in Section 2, separately for pasture and TMR samples. The hyperparameters involved, as well as  $(K, G)$ , have been selected according to the considerations in the previous section. Prior to running the methodology some wavelengths, considered to be highly noisy, have been removed; consequently the final dataset consists of  $n = 4320$  milk samples and  $p = 533$  wavelengths.

In Figure 2 the estimated correlation matrices are reported. For the pasture samples the initialization strategy selects  $K_{pasture} = 4$  and  $G_{pasture} = 25$  while  $K_{TMR} = 4$  and  $G_{TMR} = 25$  for the TMR ones. The reconstructions are coherent with the patterns seen in Figure 1 as it is confirmed by the values of the *Mean Squared Errors* measuring the discrepancies between the sample and the estimated correlation matrices ( $MSE_{pasture} = 0.021$  and  $MSE_{TMR} = 0.035$ ). The graphical inspection shows that our approach, by clustering the variables, favours the appearance of blocky structures that simplifies the interpretation of wavelengths relationships and highlights even more the differences in the correlations among different spectral regions between milk samples from pasture fed and TMR fed cows. Note that this may serve as a starting point to study how diet regimens impact the chemical processes underlying the spectral behaviour.

Other insights are provided by the investigation of the obtained variable groupings. The partitions obtained from pasture and TMR samples are quite similar with an *Adjusted Rand Index* equal to 0.65; if, on one hand, strong similarities are expected since we are examining milk samples differing only because of the diet, on the other hand this behaviour may be seen as a signal about the existence of real wavelength clustering structures and as a confirmation of the presence of a traceable redundancy. Moreover these results may be used to build new variables defined as summaries of the groups, possibly helpful for prediction purposes.

Finally note that the indications obtained, when paired with previously conducted studies, can lead to other interesting insights. As an example some works



# Bias reduction in the equicorrelated multivariate normal

## *Riduzione della distorsione nel modello normale multivariato equicorrelato*

Elena Bortolato, Euloge Clovis Kenne Pagui

**Abstract** In the multivariate normal model, the maximum likelihood estimates can be highly inaccurate with small sample size, or in presence of many covariates. The variance and correlation may result in substantial bias and therefore compromise the inferential conclusions. The paper focuses on the equicorrelated normal model and uses the mean and median bias reduction methods to improve the accuracy of inference. The properties of the resulting estimators are assessed through extensive simulation studies and one application.

**Abstract** *Nel modello normale multivariato, le stime di massima verosimiglianza possono essere altamente imprecise nel caso in cui la numerosità campionaria non sia particolarmente elevata, o in presenza di molte covariate. Gli stimatori dei parametri di varianza e correlazione risultano distorti e possono compromettere l'attendibilità delle conclusioni inferenziali. Questo lavoro pone l'attenzione sul modello normale multivariato equicorrelato e applica i metodi di riduzione della distorsione in media e in mediana per migliorare l'accuratezza dell'inferenza. Le proprietà degli stimatori risultanti sono verificate mediante ampi studi di simulazione e si prende inoltre in considerazione un'applicazione ad un dataset reale.*

**Key words:** bias reduction, confidence intervals, likelihood, multivariate normal

## 1 Introduction

The equicorrelated multivariate model was intensively studied in the decades, both for theoretical properties of estimates (Basu, 1972; De and Mukhopadhyay, 2019), and for building flexible extensions and applications (Engle and Kelly, 2012). One of

---

Elena Bortolato  
University of Padova, Department of Statistical Sciences, e-mail: elena.bortolato.1@phd.unipd.it

Euloge C. Kenne Pagui  
University of Padova, Department of Statistical Sciences, e-mail: kenne@stat.unipd.it

the problematic aspects is related to the bias arising from the estimators of the variance and the correlation parameters. The standard approach to inference based on maximum likelihood (ML) might not be accurate when the sample size  $n$  is small, or in presence of many covariates. Eventhough the ML estimators of the regression parameters are unbiased, the variance and correlation parameters may result in substantial bias and therefore misleading the inferential conclusion. This affects not only the covariance and correlation parameters, but especially the standard errors of regression coefficients.

As a result, confidence intervals provided by Wald's construction, might be unreliable. In this paper, we show that applying adjustment to the score function according to the procedures derived by Firth (1993) and Kenne Pagui et al. (2017) aiming at mean and median bias reduction (BR) respectively, improves the accuracy of the inference. The performance of the ML, mean and median BR estimators are assessed through Monte Carlo simulations under different settings. An application to a real dataset is considered. Both mean and median bias reduction estimators show better coverages than that obtained with ML estimators.

## 2 Model specification

Consider  $n$  independent observations from a  $q$ -variate normal,  $Y_i \sim N_q(\mu_i, V)$ ,  $i = 1, \dots, n$ , with  $\mu_i = X_i\beta$ , where  $X_i$  is a  $q \times p$  design matrix and  $\beta = (\beta_1, \dots, \beta_p)$ . Let  $N = n \times q$ , and  $Y = (Y_1, \dots, Y_n)^T$ , then  $Y \sim N_N(\mu, \mathcal{V})$ , with  $\mu = (\mu_1, \dots, \mu_n)^T \in \mathbb{R}^N$ . In the above, the  $N \times N$  block diagonal matrix  $\mathcal{V}$  has form

$$\mathcal{V} = \begin{pmatrix} V & 0 & \dots & 0 \\ 0 & V & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V \end{pmatrix}, \quad \text{with } V = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}$$

The model has a total of  $p + 2$  parameters. Denoting by  $\Omega$  the inverse of  $\mathcal{V}$ , the log-likelihood is

$$\ell(\theta; y) = -\frac{n}{2}[(q - 1)\log(1 - \rho) + q\log\sigma^2 + \log(q\rho - \rho + 1)] - \frac{1}{2}(y - X\beta)^T \Omega (y - X\beta),$$

where  $\theta = (\beta_1, \dots, \beta_p, \sigma^2, \rho)^T$ . The ML estimator  $\hat{\theta} \sim N_{p+2}(\theta, i^{-1}(\theta))$ .

## 3 Bias reduction

Let  $U(\theta) = \partial\ell(\theta)/\partial\theta$ ,  $j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^T$  and  $i(\theta) = E[i(\theta)]$  be the score vector, the observed information and the Fisher information.



The bias expansion of the ML estimator ( $\hat{\theta}$ ) has form  $E_{\theta}[\hat{\theta} - \theta] = b(\theta) + O(n^{-2})$ , where  $b(\theta) = i(\theta)^{-1}A^*(\theta)$  with  $A^*(\theta)$  having components  $A_r^*(\theta) = \frac{1}{2}\text{tr}\{i(\theta)^{-1}[P_r(\theta) + Q_r(\theta)]\}$ . In the latter,  $P_r(\theta)$  and  $Q_r(\theta)$  are  $p + 2 \times p + 2$  matrices defined as  $P_r(\theta) = E[U(\theta)U(\theta)^T U_r(\theta)]$ ,  $Q_r(\theta) = E[-j(\theta)U_r(\theta)]$ ,  $r = 1, \dots, p + 2$ . Firth (1993) proposed an adjusted score of form

$$U^*(\theta) = U(\theta) + A^*(\theta),$$

where the adjustment term  $A^*(\theta)$  of order  $O(1)$ , is built in such a way that  $b(\theta)$  is implicitly removed. The resulting estimator,  $\theta^*$  (mean BR estimator), solution of the  $U^*(\theta) = 0$ , has smaller bias than that of ML, that is  $E_{\theta}[\theta^*] = \theta + O(n^{-2})$ . Kenne Pagui et al. (2017) in a similar way develop an adjusted score of form

$$\tilde{U}(\theta) = U(\theta) + \tilde{A}(\theta),$$

built in such a way that the resulting estimator,  $\tilde{\theta}$  (median BR estimator), is componentwise third-order median unbiased, that is  $Pr_{\theta}(\tilde{\theta}_r < \theta_r) = 1/2 + O(n^{-3/2})$ . The adjustment term is  $\tilde{A}(\theta) = A^*(\theta) - i(\theta)F(\theta)$ , where  $F(\theta)$  is a vector of components  $F_r = [i(\theta)^{-1}]_r^T \tilde{F}_r$ ,  $r = 1, \dots, p + 2$ . The vector  $\tilde{F}_r$  has elements  $\tilde{F}_{r,t} = \text{tr}\{h_r[(1/3)P_t + (1/2)Q_t]\}$ ,  $t = 1, \dots, p + 2$  and the matrix  $h_r$  is defined as  $h_r = \{[i(\theta)^{-1}]_r [i(\theta)^{-1}]_r^T\} / i^{rr}(\theta)$ , where  $[i(\theta)^{-1}]_r$  is the  $r$ -th column of  $i(\theta)^{-1}$  and  $i^{rr}(\theta)$  its  $r$ -th element. The estimators  $\tilde{\theta}$  and  $\theta^*$  have the same asymptotic distribution of the ML estimator and this can be used to construct confidence intervals.

#### 4 Simulation studies

We present two simulation studies, in which we compare ML estimator with the mean and median BR estimators. The former focuses on independent and identical distribution case while the latter involves covariates. We draw 10000 samples from  $Y \sim \mathcal{N}(\mu, \mathcal{V})$ , with  $n = 10$  and considering  $q = 5, 15$ . The true parameter values are  $\mu = 10, \sigma^2 = 5, \rho = 0.9$ . The performance of the estimators are evaluated in terms of percentage of underestimation,  $PU = R^{-1} \sum_{r=1}^R I_{\{\hat{\theta}_r \leq \theta\}}$ , with  $I$  denoting the indicator function, the relative bias,  $RB = R^{-1} \sum_{r=1}^R (\hat{\theta}_r - \theta) / \theta$ , empirical 95% Wald confidence interval (WALD) and influence of bias on mean square error,  $IBMSE = B^2 / SD^2$ , which indicates the relative increase due to bias on the mean square error from its absolute minimum. Here,  $B$  and  $SD$  denote the bias and standard deviation, respectively. Then we repeat the experiment increasing the sample size to  $n = 20$ . Results are summarized in table 1. The ML estimator tends to underestimate the variance and correlation parameters and this is more evident for smaller  $n$  and larger  $q$ , the bias has also an high impact on the IBMSE index. From the IBMSE, we note that the effect of the bias on the standard error of  $\hat{\rho}$  is more pronounced. The mean and median BR estimators succeed in achieving their own desirable goals, respectively, and the results are preferable than the ML estimator.

Bias reduction methods produce the empirical coverage of confidence intervals which is closer to the nominal 95% level compared to those obtained with the ordinary ML. To assess the properties of bias reduction methods in a regression frame-

		$q = 5$			$q = 15$			
		ML	mean BR	median BR	MLE	mean BR	median BR	
$n = 10$	PU	$\mu$	50.00	50.00	50.00	50.01	50.01	50.01
		$\sigma^2$	64.88	55.90	49.78	65.92	57.20	50.72
		$\rho$	62.68	41.92	50.20	64.98	42.75	50.56
	RB	$\mu$	0.03	0.03	0.03	0.02	0.02	0.02
		$\sigma^2$	-8.97	0.38	7.59	-9.66	-0.52	6.52
		$\rho$	-3.68	-0.57	-1.70	-3.64	-0.57	-1.56
	WALD	$\mu$	90.00	91.85	92.60	90.42	92.03	92.58
		$\sigma^2$	80.60	85.74	88.00	80.11	85.15	87.54
		$\rho$	94.37	87.15	91.16	94.30	88.38	91.38
IBMSE	$\mu$	0.00	0.00	0.00	0.00	0.00	0.00	
	$\sigma^2$	5.20	0.01	2.59	6.33	0.01	2.01	
	$\rho$	21.71	0.72	5.82	25.84	0.87	6.02	
$n = 20$	PU	$\mu$	49.08	49.08	49.08	50.25	50.25	50.25
		$\sigma^2$	59.66	53.72	49.25	60.35	54.51	49.74
		$\rho$	57.93	43.04	49.47	60.15	43.83	50.31
	RB	$\mu$	0.09	0.09	0.09	-0.02	-0.02	-0.02
		$\sigma^2$	-4.22	0.44	3.75	-4.53	0.04	3.29
		$\rho$	-1.56	-0.08	-0.69	-1.61	-0.14	-0.69
	WALD	$\mu$	92.63	93.42	93.78	92.86	93.54	93.89
		$\sigma^2$	87.13	89.80	91.26	86.78	89.66	91.06
		$\rho$	94.58	90.14	92.57	94.66	90.81	92.77
IBMSE	$\mu$	0.04	0.04	0.04	0.00	0.00	0.00	
	$\sigma^2$	2.13	0.02	1.42	2.57	0.00	1.14	
	$\rho$	12.20	0.04	2.72	15.09	0.15	3.16	

Table 1: First simulation study under independent and identical distribution. Estimation of parameter  $\theta=(\mu, \sigma^2, \rho)$ :  $\mu = 10, \sigma^2 = 5, \rho = 0.9$ .

work, we run a simulation study considering 10000 samples of size 20 from the model

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4},$$

where  $x_{i1}$  is drawn from a Uniform in  $(-10,10)$ ;  $x_{i2}$  from an exponential distribution of rate  $\frac{1}{2}$ ;  $x_{i3}$  is generated from a Bernoulli  $B(1,0.5)$  and  $x_{i4}$  from a  $B(1,0.2)$ . The true values for the parameter is set to  $\beta = (2,0.3,-1,3,-0.5)$ , with  $\sigma^2 = 5$  and  $\rho = 0.9$ . We first consider  $q = 2$ . From table 2, the estimators of  $\sigma^2$  and  $\rho$  obtained with the adjusted score fulfill the expected properties. Both mean and median BR estimators perform better than the ML one with respect to the four performance measures. In particular, with the mean BR the RB is significantly reduced with respect to ML while median BR has PU closer to 50% . Similar results are obtained with  $q = 5$  . In this case, the estimator of  $\beta$  is unbiased and identical for the three methods. As a result, PU, RB and IBMSE are equal as shown in table 3. Under the

Bias reduction in the equicorrelated multivariate normal

	$q = 2$				$q = 5$			
	PU	RB	WALD	IBMSE	PU	RB	WALD	IBMSE
$\hat{\sigma}^2$	83.60	-24.38	64.35	87.75	82.60	-23.10	64.48	83.39
$\hat{\rho}$	73.41	-5.32	95.52	45.47	79.89	-4.93	91.89	67.17
$\sigma^{2*}$	56.16	-0.71	86.40	0.04	55.17	0.07	87.02	0.00
$\rho^*$	44.87	-0.53	87.08	0.82	45.40	-0.38	88.50	0.71
$\hat{\sigma}^2$	50.97	3.46	88.02	0.91	51.04	4.11	88.71	1.36
$\hat{\rho}$	50.71	-1.38	90.07	4.96	50.42	-0.93	90.63	3.87

Table 2: Simulations with covariates: estimation of  $\sigma^2$  and  $\rho$ .

two scenarios ( $q = 2$  and  $q = 5$ ), it is remarkable the good performance of the bias reduced estimators in terms of the coverages of confidence intervals.

	$q = 2$					$q = 5$				
	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$
PU	50.19	50.01	48.90	49.16	50.82	49.83	50.67	50.44	50.24	48.96
RB	-0.27	-0.23	-0.44	0.45	6.15	0.04	-0.30	0.29	0.05	-2.90
ML	89.58	88.69	88.83	89.52	89.34	88.76	89.12	88.97	89.28	88.92
WALD mean BR	93.55	93.04	93.16	93.69	93.28	93.16	93.23	93.08	93.62	93.16
WALD median BR	93.88	93.59	93.68	94.02	93.70	93.63	93.69	93.59	93.94	93.48
IBMSE	0.00	0.01	0.02	0.02	0.04	0.00	0.01	0.01	0.00	0.01

Table 3: Simulations with covariates: estimation of regression coefficients.

## 5 Application

We consider the `Stroke` dataset (Dobson e Barnett, 2008), available in the R package `MLGdata` on CRAN. This was collected with the aim of study post-heart attack rehabilitation therapies. Patients were assigned to three experimental groups: A, treated with the innovative therapy; B, treated with traditional therapy in the same hospital as the patients of group A; C, treated with traditional therapy in a different hospital. For each of the 24 patients, 8 measures of functional ability were obtained in consecutive weeks. The study aimed to verify whether treatment A was more effective than the others. The model considered is

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5},$$

where  $x_{i1} = 1$  or  $x_{i2} = 1$  if the subject belongs to the B or C group therapy,  $x_{i3}$  refers to the week, while  $x_{i4}$ ,  $x_{i5}$  represent the interaction terms between the group and the week. Results in table 4 show that the standard errors of the regression coefficients are different for the three approaches.

	ML	mean BR	median BR
$\beta_0$	29.82 (7.05)	29.82 (8.07)	29.82 (7.60)
$\beta_1$	3.35 (9.97)	3.35 (11.41)	3.35 (10.75)
$\beta_2$	-0.02 (9.97)	-0.02 (11.41)	-0.02 (10.75)
$\beta_3$	6.32 (0.46)	6.32 (0.45)	6.32 (0.46)
$\beta_4$	-1.99 (0.66)	-1.99 (0.63)	-1.99 (0.66)
$\beta_5$	-2.69 (0.65)	-2.69 (0.63)	-2.69 (0.66)
$\sigma^2$	425.57 (104.88)	547.69 (141.17)	490.86 (123.63)
$\rho$	0.83 (0.04)	0.88 (0.03)	0.85 (0.03)

Table 4: Stroke data: estimates and standard errors in parenthesis.

## References

1. Basu, J. P. (1972). Statistical analysis of equicorrelated samples from multivariate population (Doctoral dissertation, Texas Tech University).
2. Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27 – 38.
3. Engle, R. and Kelly, B. (2012) Dynamic equicorrelation *Journal of Business & Economic Statistics*, 30(2), 212-228 Taylor & Francis
4. Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika*, 104, 923 – 938.
5. Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2020). Efficient implementation of median bias reduction. arXiv preprint arXiv:2004.08630.
6. De, S. K., and Mukhopadhyay, N. (2019). Two-stage fixed-width and bounded-width confidence interval estimation methodologies for the common correlation in an equi-correlated multivariate normal distribution. *Sequential Analysis*, 38(2), 214-258.
7. Sartori, N., Salvan, A. and Pace, L. (2020). Package ‘MLGdata’. <https://CRAN.R-project.org/package=MLGdata>

# Some results on identifiable parameters that cannot be identified from data

*Alcuni risultati su parametri identificabili che non possono essere identificati dai dati*

Christian Hennig

**Abstract** It can be shown that some theoretically identifiable parameters cannot be identified from data, meaning that no consistent estimator of them can exist. Examples are a constant correlation between Gaussian observations (in presence of such correlation not even the mean can be identified from data), cluster memberships in a fixed classification model underlying  $k$ -means clustering. I will define non-identifiability from data and indistinguishability from data. Two different constant correlations between Gaussian observations cannot even be distinguished from data.

**Abstract** È possibile dimostrare che alcuni parametri teoricamente identificabili non possono essere identificati a partire dai dati. Per esempio questo avviene nel caso di correlazione costante tra osservazioni con distribuzione normale (in tal caso nemmeno la media è identificabile dai dati), oppure nel caso dei cluster-labels nel modello di partizionamento sottostante il metodo del  $k$ -means. In questo lavoro si propongono le nozioni di “non identificabilità dai dati” e di “indistinguibilità dai dati”. Due correlazioni costanti e distinte tra osservazioni Gaussiane non sono nemmeno “distinguibili dai dati”.

**Key words:** identifiability from data, distinguishability, correlation, dependence,  $k$ -means clustering

## 1 Introduction

The starting point of this work is the realisation that it is impossible for marginally Gaussian distributed random variables  $X_1, \dots, X_n$  with  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ , to diagnose from the data whether the observations are i.i.d, or whether there is a correlation  $\rho > 0$  between them, see Section 2 and Theorem 1. This is a big problem

---

Christian Hennig  
Dipartimento di Scienze Statistiche “Paolo Fortunati”, Università di Bologna,  
Via delle Belle Arti, 41, 40126 Bologna, e-mail: christian.hennig@unibo.it

for practical data analysis, because it means that it is not possible to distinguish in any way (including misspecification testing) between i.i.d. Gaussian data and Gaussian data with a constant positive correlation. But in the latter case the mean, the standard estimator for  $\mu$  on which all standard inference is based, is inconsistent. In fact, in that case,  $\mu$  is not identifiable from data either.

In order to formalise this, in Section 3 I will define a concept of parameters not being identifiable from data, meaning that a consistent estimator of them cannot exist. This is essentially different from the classical definition of identifiability in statistics (see Definition 1.5.2 in [2]). Identifiability is necessary but not sufficient for the existence of consistent estimators. In the above situation both  $\rho$  and  $\mu$  are identifiable according to the classical definition; I will discuss how it is generally possible that identifiable parameters cannot be identified from data.

It is possible to further differentiate between non-identifiability from data and indistinguishability, meaning that not even observable events exist that have different probabilities under the different parameters. In the above situation, two different means are distinguishable, but two different constant correlations are not. A further example for these concepts involves the cluster membership parameters in a fixed classification model for  $k$ -means clustering, see Section 4.

## 2 Constant correlation between Gaussian observations

A model assumption for much standard statistical inference is to assume i.i.d. Gaussian  $X_1, \dots, X_n$ ,  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ . Now consider Gaussian  $X_1, \dots, X_n$  with correlation  $\text{Cor}(X_i, X_j) = \rho > 0$  constant for any  $i \neq j$  (model M1).

This is a problem for inference about  $\mu$ , because in the latter situation, for the arithmetic mean  $\bar{X}_n$ :

$$\mathcal{L}(\bar{X}_n) = \mathcal{N}\left(\mu, \frac{(1-\rho)\sigma^2}{n} + \rho\sigma^2\right) \xrightarrow{n \rightarrow \infty} \mathcal{N}(\mu, \rho\sigma^2).$$

This means that the mean is inconsistent for  $\mu$  as long as  $\rho\sigma^2 > 0$ . In fact, the model can equivalently be written as a model with a single realisation of a random effect  $Y$ ,  $i = 1, \dots, n$ :

$$X_i = \mu + Y + E_i, \quad Y \sim \mathcal{N}(0, \tau_1^2), \quad E_i \sim \mathcal{N}(0, \tau_2^2), \quad \sigma^2 = \tau_1^2 + \tau_2^2, \quad \rho = \frac{\tau_1^2}{\tau_1^2 + \tau_2^2},$$

and it can be seen that  $\mu$  is confounded with the unobservable  $Y$  and can therefore not be consistently identified in any way.

It is therefore of interest to detect such correlations when trying to do inference based on the mean, but this is impossible, because data from such a model look exactly like i.i.d. data, just with mean  $\mu + Y$  with unobservable  $Y$  rather than  $\mu$ .

This is formalised using the concept of “(non-)identifiability from data”.

### 3 Identifiability and distinguishability from data

Let  $X_1, X_2, \dots, X_n, \dots$  random variables on a space  $\mathcal{X}$ , for  $n \in \mathbb{N}$ :  $\mathcal{L}(X_1, \dots, X_n) = P_{n;\theta}$  with parameter  $\theta \in \Theta$ . The spaces  $\mathcal{X}$  and  $\Theta$  can be very general. Results may not concern all information in the parameter, which could be more than one-dimensional. The parameter part of interest for identifiability considerations is called  $\lambda \in \Lambda$  with  $\theta = \lambda$  or  $\theta = (\lambda, \psi)$ ,  $\psi \in \Psi$ .

**Definition 1.**  $\lambda$  is called **identifiable from data** if it is possible to find a consistent sequence of estimators  $(T_n)_{n \in \mathbb{N}}$  with  $T_n = \mathcal{X}^n \mapsto \Lambda$ ,  $\forall \theta \in \Theta$ :  $T_n(X_1, \dots, X_n) \rightarrow \lambda$  in probability.

Traditionally, statistical identifiability of a parametric model  $(P_\theta)_{\theta \in \Theta}$  means that  $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$  [5, 2]; for parameter parts,  $\lambda_1 \neq \lambda_2 \Rightarrow P_{(\lambda_1, \psi_1)} \neq P_{(\lambda_2, \psi_2)}$  for any  $\psi_1, \psi_2$  is often referred to as partial identifiability [5]. If parameters are not (partially) identifiable, they can obviously not be identifiable from data, because no consistent estimator can tell equal distributions apart:

**Corollary 1.** *Parameters and parameter parts that are identifiable from data are also identifiable.*

Here, data generating mechanisms are treated that do not allow to identify parameters from data that are in fact identifiable in the traditional sense. Model M1 is an example. Obviously, models with different correlation parameters  $\rho_1 \neq \rho_2$  are different from each other, and  $\rho$  can be estimated consistently a the whole sequence of  $n$  observations is repeated *independently*. Generally, as opposed to traditional identifiability, the concept of identifiability from data involves potential constraints on what is observable, on top of the model definition itself. In Section 2, independent repetition of observations from model M1 does not happen; all available observations are dependent on all other observations. This makes consistent estimation of  $\rho$  impossible:

**Theorem 1.** *In model M1, neither  $\rho$  nor  $\mu$  are identifiable from data.*

Proofs are omitted in the short paper due to space limitations.

In fact there is a difference between trying to estimate  $\rho$  on one hand and  $\mu$  on the other hand. While  $\mu$  cannot be estimated consistently, in case that  $\rho$  is small, the data can give fairly precise information about its location, whereas there is no information in the data about  $\rho$  at all.

**Definition 2.** Two parameter values  $\lambda_1$  and  $\lambda_2$  are called **distinguishable from data** if  $\exists n, \alpha \in (0, 1]$ , and an observable set  $A$  so that

$$\forall \psi \in \Psi : P_{n;(\lambda_1, \psi)}(A) < \alpha, P_{n;(\lambda_2, \psi)}(A) \geq \alpha. \quad (1)$$

If  $\lambda$  is the only parameter,  $\psi$  can be chosen as a constant not influencing  $P$ .

Note that the choice of  $A$  can involve any information that the data hold about  $\psi$ , therefore indistinguishability according to the definition does not leave open the possibility to distinguish  $\lambda_1$  and  $\lambda_2$  by finding a set  $A_\psi$  dependent on  $\psi$  and estimating the required  $\psi$ , despite the requirement that (1) needs to hold for all  $\psi$ .

**Theorem 2.** *In model M1, any two  $\rho_1 \neq \rho_2 \geq 0$  are indistinguishable from data, whereas any two  $\mu_1 \neq \mu_2$  are distinguishable from data.*

A more general result can be shown regarding the non-identifiability from data of a parameter formalising independence and the strength of dependence for a general class of models in which the observation ordering is not informative for dependence.

#### 4 Cluster memberships in $k$ -means clustering

$k$ -means clustering can be interpreted as maximum likelihood (ML) estimator of a “fixed classification model” [1]: Let  $\mathbf{X}_1, \dots, \mathbf{X}_n$ ,  $\mathbf{X}_i \in \mathbb{R}^p$ ,  $i = 1, \dots, n$ , be independently distributed with

$$\mathcal{L}(\mathbf{X}_i) = \mathcal{N}_p(\boldsymbol{\mu}_{\gamma_i}, \sigma^2 \mathbf{I}_p), \quad \gamma_i \in \{1, \dots, k\}, \quad k > 1, \quad \sigma^2 \geq 0. \quad (2)$$

This model can be interpreted as generating  $k$  different Gaussian distributed clusters characterised by cluster means  $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k \in \mathbb{R}^p$ , all with the same spherical covariance matrix, and  $\gamma_i$  indicates the true cluster membership of  $\mathbf{X}_i$ . The  $\gamma_i$  take discrete values, and their number converges to  $\infty$  with  $n$ , so these are nonstandard parameters, but in many applications they are of practical interest.

$k$ -means clustering of data  $\tilde{\mathbf{X}}_n = (\mathbf{X}_1, \dots, \mathbf{X}_n)$  is defined as

$$\begin{aligned} T_n(\tilde{\mathbf{X}}_n) &= (\mathbf{m}_{1n}, \dots, \mathbf{m}_{kn}, g_{1n}, \dots, g_{nn}) \\ &= \underset{\mathbf{m}_1, \dots, \mathbf{m}_k, g_1, \dots, g_k}{\operatorname{arg\,min}} \quad W(\mathbf{m}_{1n}, \dots, \mathbf{m}_{kn}, g_{1n}, \dots, g_{nn}), \\ W(\mathbf{m}_{1n}, \dots, \mathbf{m}_{kn}, g_{1n}, \dots, g_{nn}) &= \sum_{i=1}^n \|\mathbf{X}_{in} - \mathbf{m}_{g_{in}}\|^2. \end{aligned}$$

For given  $\mathbf{m}_1, \dots, \mathbf{m}_k$ , the  $g_1, \dots, g_n$  minimising  $W$  are given by

$$g_i = \underset{j \in \{1, \dots, k\}}{\operatorname{arg\,min}} \|\mathbf{X}_i - \mathbf{m}_j\|, \quad i = 1, \dots, n.$$

Assume that the  $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$  are pairwise different and lexicographically ordered. This makes the model identifiable according to the traditional definition. The cluster membership parameters are another example for parameters that are identifiable according to the classical definition (because  $\gamma_i$  uniquely defines the distribution of  $\mathbf{X}_i$ ), but cannot be identified from data.

**Theorem 3.** *The parameters  $\gamma_i$ ,  $i \in \mathbb{N}$  in the model defined in (2) are not identifiable from data, although any two values of  $\gamma_i$  are distinguishable from data.*

It may be suspected that this is a consequence of the fact that for  $i = 1, \dots, n$ , only  $\mathbf{X}_i$  holds information about the parameter  $\gamma_i$ , and the number of these parameters goes to  $\infty$  with  $n \rightarrow \infty$ . But this is not quite true. More observations add information about



Parameters not identifiable from data

the clusters that can in turn be used to classify individual observations better. The problem here is rather the Gaussian distribution assumption, which implies that the marginal density of  $\mathbf{X}_i$  is everywhere nonzero, so that the single observation made of  $\mathbf{X}_i$  is not enough to determine with probability 1 to what cluster the observation belongs.

In fact, there is a different model setup in which the  $\gamma_i$  are identifiable from data, which requires that, where densities exist, the marginal density  $f_{\eta^*,n}(\mathbf{X}_i = \mathbf{x})$  is zero wherever  $f_{\eta,n}(\mathbf{X}_i = \mathbf{x}) > 0$ .

Defining

$$W(P) = (\boldsymbol{\mu}_1^*, \dots, \boldsymbol{\mu}_k^*) = \arg \min_{(\mathbf{m}_1, \dots, \mathbf{m}_k) \in (\mathbb{R}^p)^k} \int \min_{\mathbf{m} \in \{\mathbf{m}_1, \dots, \mathbf{m}_k\}} \|\mathbf{x} - \mathbf{m}\|^2 dP(\mathbf{x}),$$

[4] showed that for a distribution  $P$  satisfying

$$E_P \|\mathbf{X}\|^2 < \infty, W(P) \text{ is unique up to the numbering of the means,} \quad (3)$$

the  $k$ -means estimator  $(T_n^m)_{n \in \mathbb{N}}$ , where  $T_n^m(\tilde{\mathbf{X}}_n) = (\mathbf{m}_{1n}, \dots, \mathbf{m}_{kn})$ , is strongly consistent for  $W(P)$ . Assume further that

$$\forall j \neq l \in \{1, \dots, k\} : P\{\|\mathbf{X} - \boldsymbol{\mu}_j^*\|^2 = \|\mathbf{X} - \boldsymbol{\mu}_l^*\|^2\} = 0. \quad (4)$$

For  $\mathcal{L}(\mathbf{X}) = P$ ,  $j = 1, \dots, k$ , define

$$A_j = \left\{ \mathbf{X} : j = \arg \min_l \|\mathbf{X} - \boldsymbol{\mu}_l^*\|^2 \right\}, P_j = \mathcal{L}(\mathbf{X} | \mathbf{X} \in A_j), \pi_j = P(A_j).$$

$P_j$  is  $P$  constrained to the set  $A_j$  of points that are closest to the mean  $\boldsymbol{\mu}_j$  ( $A_1, \dots, A_k$  form a so-called Voronoi tessellation of  $\mathbb{R}^p$ ), and

$$P = \sum_{j=1}^k \pi_j P_j \quad (5)$$

(every distribution can be written as a mixture in this form). Now consider  $\mathcal{L}(\tilde{\mathbf{X}}_n) = P^*$  so that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  are independently distributed with

$$\mathcal{L}(\mathbf{X}_i) = P_{\gamma_i} \quad \gamma_i \in \{1, \dots, k\}, k > 1, i = 1, \dots, n. \quad (6)$$

This defines a fixed classification model associated to the mixture  $P$ . Let  $Q_P$  be an infinite i.i.d. product of categorical distributions on  $\{1, \dots, k\}$  with probabilities  $(\pi_1, \dots, \pi_k)$ . Assume for given  $P$  that  $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots)$  fulfill

$$\tilde{P} \left\{ \lim_{n \rightarrow \infty} T_n^m(\tilde{\mathbf{X}}_n) = W(P) | \mathbf{G} = \boldsymbol{\gamma} \right\} = 1. \quad (7)$$

According to [4], this happens in model (5) with probability 1.

**Theorem 4.** *Assuming (3), (4), and (7), the parameters  $\gamma_i$ ,  $i \in \mathbb{N}$ , in the fixed classification model defined by (6) are identifiable from data.*

This can be interpreted as implying that  $k$ -means does not actually estimate the centres of the spherical Gaussians in (2) for which it is ML, but rather the Voronoi tessellation resulting from  $P$ , and the resulting clusters are not necessarily spherical.

## 5 Conclusion

I provide a framework for describing situations in which certain parameters that are identifiable in a classical sense cannot actually be identified from data. Practical implications of the results shown here are that a constant correlation between any two observations in a simple Gaussian sample cannot be detected from the data, and if it exists, the mean cannot be identified either, although it is at least possible to have a weak indication of where the mean is, because two different means are distinguishable.

Cluster memberships in  $k$ -means clustering are not identifiable from data under the fixed classification model with spherical Gaussian clusters for which  $k$ -means is ML, but they can be detected if the true clusters are interpreted as defined by the optimal Voronoi tessellation of the underlying true distribution, which does not require spherical Gaussian components.

Another example of the introduced concepts is the indistinguishability from data of “missing at random” and “missing not at random”, which follows directly from the main result in [3]. In such a situation the source of indistinguishability despite traditional identifiability is that components that are essential parts of the model are unobserved (missing).

## References

1. Bock, H. H.: Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis* 23, 5–28 (1996).
2. Lehmann, E. L., Casella, G.: *Theory of Point Estimation* (2nd ed.), Springer, New York (1998)
3. Molenberghs, G., Beunckens, C., Sotito, C. and Kenward, M.G.: Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 371–388 (2008).
4. Pollard, D.: Strong consistency of  $k$ -means clustering. *Annals of Statistics* 9, 135–140 (1981).
5. Prakasa Rao, B. L. S.: *Identifiability in Stochastic Models*, Academic Press, Boston (1992).

## 4.23 Novel approaches for official statistics

# Web data collection: profiles of respondents to the Italian Population Census

## *Raccolta dati via web: profili dei rispondenti al Censimento italiano della popolazione*

Elena Grimaccia, Gerardo Gallo, Alessia Naccarato, Novella Cecconi, Alessandro Fratoni

**Abstract** Identifying the profile of the "web respondent" can help the survey designers to promote the participation in web-based surveys, with the aim of enhancing timeliness and reducing costs of surveys. Based on the 2019 Italian population census data, we estimated the set of familial and geographical characteristics corresponding to a greater probability that the interviewed will choose to respond online, and we identify the web respondents' profile. Results show that the households with lower probability of answering via web correspond to segments of the population generally affected by economic and social fragility.

**Abstract** *L'individuazione del profilo del "rispondente via web" può aiutare gli statistici a promuovere la partecipazione ad indagini basate sulla raccolta dati via web, con l'obiettivo di aumentare la tempestività e ridurre i costi delle indagini. Sulla base dei dati del censimento della popolazione italiana del 2019, sono state individuate le caratteristiche familiari e geografiche corrispondenti a una maggiore probabilità che l'intervistato scelga di rispondere online ed è stato identificato il profilo degli intervistati CAWI. I risultati mostrano che le famiglie con minore probabilità di risposta via web corrispondono a segmenti della popolazione generalmente più fragili economicamente e socialmente.*

**Key words:** Italian permanent population census, online respondent, CAWI profile

---

<sup>1</sup> Elena Grimaccia, Istat; email: elgrimac@istat.it

Alessia Naccarato, Department of Economics, Roma Tre University; email: alessia.naccarato@uniroma3.it

Gerardo Gallo, Istat; email: gegallo@istat.it

Novella Cecconi Istat; email: nceconni@istat.it

Alessandro Fratoni, Istat; email: fratoni@istat.it

## 1 Introduction

For statistical surveys, the on line data collection implies considerable advantages, such as the reduction of costs, the containment of interviewer effect (Bethlehem et al., 2011), and shorter survey time (Cobben and Bethlehem, 2013). These advantages are even more valuable for official statistical surveys in the case of large sample sizes, very large number of variables to be collected as well as the high quality standards they must ensure. This takes on particular importance in the case of the population census which, due to the extent of the information it collects on the usual resident population, represents one of the most important surveys carried out by National Statistical Institutes (NSIs). Moreover, the emergency due to the coronavirus pandemic and the impossibility of carrying out the field operations have further boosted the necessity for web-based interviews. However, even if the use of the internet is quite widespread in Europe, the respondents' attitudes toward online data collection is still not granted. In Italy, the Permanent Population and Housing Census (PPHC), started in 2018 by the Italian National Institute of Statistics (Istat), currently provides a mixed mode data collection in which respondents may choose to fill in the questionnaire via the web or in the traditional way. The availability of such data allows to study the different profiles of CAWI and not CAWI respondents. In general, the knowledge of the interviewees' profile is useful to define the data collection methodologies. During the survey process, decisions have to be made both to contact the eligible respondents and to solicit the compilation of the web questionnaire (Biffignandi and Pratesi, 2002). The interaction of the characteristics of the respondent with the decisions of the survey designer influences the response rate and the success of the survey (Biffignandi and Pratesi, 2002; Durrant and Steele, 2009). In this study, we analyse the determinants that influence the cooperation of the respondents to the Italian population census in computer-assisted web interview (CAWI) data collection in order to point out the specific characteristics of the population which could enhance the efficacy of the actions of survey designers to improve CAWI participation (on differences between those who participate online and offline in mixed-mode surveys (Blom et al. 2015)). The profiles of the CAWI respondents are finally analysed according to their geographical distribution, since in Italy geographical differences are particularly significant, being related with a number of factors of economic and social development (Benassi and Naccarato, 2017) and, therefore, it is necessary to adapt the data collection strategy to the geographical imbalances.

## 2 The Italian Population Census

In recent decades, developments in information technology and the ever-increasing availability of administrative data have led several European countries to develop innovative methods for the population census, with the aim of providing official data for the usual resident population. Statistics are developed by using registers and

Web data collection: profiles of respondents to the Italian Population Census

other administrative sources, together with information from either sample field data or full field enumeration for selected variables. Under this approach, called a combined census, the field data collection can cover the whole population or just a sample. Starting from 2018, the design of the PPHC is based on two different yearly sample surveys: an Areal survey and a List survey (Istat, 2020). The first is a “door-to-door” enumeration, which is conducted by the enumerators at the house number or enumeration area extracted from the address archive. The List survey sample HHs receive a letter inviting them to fill in the online questionnaire using the CAWI technique. If the HHs in the List sample are not available to fill in the online questionnaire, they can choose the “face-to-face” interview with the support of an interviewer. This different type of data collection in the List sample allows us to distinguish the HHs which choose to fill in the questionnaire via the web from those who prefer the traditional interview. Thanks to this information, it is possible to estimate the probability that they would prefer to respond online, to define the profiles of each type of HH (in terms of homogeneity with respect to the HH variables detected), and to study their geographical distribution. Therefore, the statistical units used in the study are the HHs that belong to the List sample of the 2019 PPHC survey, that represents about 68% of the total PPHC sample.

The distributions of characteristic variables vary according to data collection mode, CAWI and not-CAWI (Table 1).

**Table 1:** Distribution of Cawi and not cawi responses by geographical area - 2019

<i>Geographical area</i>	<i>Cawi</i>	<i>Not Cawi</i>	<i>Total</i>
North	58.91	41.09	100.00
Center	53.77	46.23	100.00
Mezzogiorno	36.55	63.45	100.00
Italy...	49.99	50.01	100.00

*Pearson chi2 = 3.4e+04 Pr = 0.000*  
*Kendall's tau-b= -0.1849 ASE = 0.001*

The overall CAWI response rate for the 2019 Italian PPHC is 49.9 per cent. The share of HHs that answer to a survey conducted with a CAWI methodology varies according to the residence of the HH: in the Northern area of Italy the share of CAWI-respondents is close to 59%, while in the “Mezzogiorno” areas (Southern area of Italy plus the two major Islands) the share of CAWI-respondents is stuck at 36%.

### 3 Analysis of CAWI-respondents’ profile

In order to identify the “CAWI profile”, that is to say the characteristics of the observed variables associated to the HHs with the highest probability of responding via the web, the probability that a HH responds in CAWI mode we estimated a Logit model (Biffignandi and Pratesi, 2002; Pratesi et al., 2004; Maslovskaya et al., 2019) in which the endogenous variable is the dichotomous variable that assumes the value

1 if the family responded CAWI and zero otherwise. The exogenous variables are the HH characteristics (namely: HH size, HH citizenship, younger HH member age, HH higher educational level). Among the exogenous variable of the model is also considered the geographical macro region variable that represents the effect of the economic and social differentiations that historically distinguish Northern Italy from the rest of the country. Differences in HH profiles could also be found in residence areas smaller than macro regions, due to different administrative systems and management of community services. Therefore, in the model regional fixed effects were also considered, as this allows to study the family characteristics that define the CAWI profile, net of the effects due to the region of residence of the HH.

Table 2 reports the results (odds ratios) of the logit model that identifies the profile of the HHs answering via web data collection.

Table 2 – Logit model estimate parameters (Odds ratios)

<i>Variable</i>	<i>Italy</i>	<i>North</i>	<i>Centre</i>	<i>Mezzogiorno</i>
Household highest educational level (base = Primary level education)				
Secondary	.73636***	.68698***	.67004***	.84439***
Tertiary	1.3552***	1.3055***	1.2388***	1.4879***
Household citizenship (base = All foreigners)				
All Italians	1.3829***	1.5434***	1.1888***	1.1798***
Mixed citizenship	.71418***	.72435***	.64223***	.71066***
Household size (base = 1 component)				
2	0.01075	-0.00269	0.01595	.0385***
3	.01706*	.02546*	-7.10E-05	.03673**
4	.10958***	.16946***	.08758***	.10517***
5 or more	-.23677***	-.21706***	-.23157***	-.21302***
Youngest household member age (base = 18–34 years old)				
35–64	.05153***	-.02121*	0.02619	.13425***
65+	-.04364***	-.249***	-.07834***	.26399***
Regional Fixed Effects				
Trentino Alto Adige	0.01767	0.02395		
Lombardy	.28144***	.2776***		
Piedmont	.06195***	.06068***		
Friuli Venezia Giulia	.03786*	.04156**		
Veneto	.11585***	.11006***		
Liguria	-0.01384	-0.00811		
Valle d'Aosta	-.16726***	-.17235***		
Tuscany	-.06017***		(base)	
Marche	-.22437***		-.15925***	
Lazio	-.22109***		-.15245***	
Umbria	-.20907***		-.14427***	
Abruzzo	-.52699***			(base)
Campania	-1.0489***			-.49973***
Sardinia	-.4825***			.06413***
Molise	-.92447***			-.39183***
Puglia	-.65984***			-.11443***
Basilicata	-.95069***			-.41128***
Sicily	-1.0459***			-.50054***
Calabria	-1.3369***			-.79589***
Constant	-1.6673***	-1.6947***	-1.4662***	-2.2103***

Legend: \* p<.05; \*\* p<.01; \*\*\* p<.001

The empirical evidence suggests that among HHs with a higher level of education the odds ratios (OR) of answering via web are significantly higher than for those who have a lower level degree: respectively 0.736 for Secondary education and

Web data collection: profiles of respondents to the Italian Population Census

1.355 for Tertiary. A higher level of education is a strong determinant of the availability to respond via web in every area of the Country.

Households composed by all foreigner members present a lower probability of using the CAWI option than those with at least an Italian member. In particular, the HHs with all Italians members present an OR of 1.383. The same result is obtained in the other Italian macro-regions: foreigner HHs should be made target of specific actions aiming at helping them responding via web in every area of Italy. The analysis of the size of the HH in term of number of components offers less univocal results. The probability of CAWI answering is significantly lower in HHs with 5 or more members; the same probability, instead, is significantly higher for HHs with four components, in all the areas of Italy. However, while in the Northern part of the peninsula other HH dimensions do not make such a difference, in the Regions of Mezzogiorno also two and three members HHs present a higher probability of CAWI answering. Moreover, empirical evidence suggests that in the HH where the age of the youngest members is between 35 and 64 years old the CAWI response rate is significantly higher than in HH where the rank age of youngest members is between 18 and 24 years old. The HHs composed by all elderly people present the lowest probability of CAWI response. However, the Mezzogiorno geographical area is the only one where older people accept to answer via web more than HH with younger members. It is worth mentioning that HH with at least a member with higher education present a higher probability of accepting the CAWI data collection regardless of age or citizenship.



**Figure 1:** Regional odds ratios of CAWI response

Lombardy and Veneto present the higher probability of a CAWI response (Figure 1). The other Northern regions, with the exception of the Valle D'Aosta, are found with a slightly lower odds of a response via web. The Mezzogiorno Regions' - except Abruzzo, Puglia, and Sardinia which are aligned with the Central geographical area – present instead the lower probability of a CAWI data collection mode.



## 4 Conclusions

The online data collection still faces—at least in Italy—widespread resistance from the population. In order to foster a positive attitude towards web interviewing, an awareness campaign is necessary, and this would be more effective if it is targeted at a specific population. Hence, there is a need to identify the HH profiles from which a web response is most unlikely. A further consideration should be made with regard to families living in southern Italy, since in this case the geographical factor plays a very important role. Indeed, in the Mezzogiorno geographical area, only 36% of the HHs choose to answer via the web.

This contribution has shown how some structural characteristics of HHs allow us to classify them on the basis of their preference to fill in the population census questionnaire online. The HHs with a higher level of education, composed of Italian citizens, living in the North and partly in Central Italy are those for whom the probability of answering via the web is higher. Since these characteristics are also those referring to HHs that historically live in better economic and social conditions, the results indicate that the attitude toward web data collection is favoured by those characteristics that other studies identify as less fragile conditions. In conclusion, HHs without any member who has a high level of education, HH's made of all foreign members, or that comprise only elderly people, should be made target of specific campaign and support for the web-based questionnaire.

## References

1. Benassi F. and Naccarato A., Households in potential economic distress. A geographically weighted regression model for Italy, 2001–2011, *Spatial Statistics*, August 2017, Vol.21, 362-376 (2017).
2. Bethlehem I., Cobben F. and Schouten B., *Handbook of Nonresponse in Household Surveys*, Wiley, New Jersey (US), (2011).
3. Biffignandi S. and Pratesi M., Modelling the Respondents' Profile in a web survey on Firms in Italy, In: A. Ferligoj and A. Mrvar (Editors), *Development in Social Science Methodology*, Metodoloski zvezki, Ljubljana: FDV (2002).
4. Blom A. G., Bosnjak M., Das S., Cornilleau A., Cousteaux A., Douhou S., and Krieger U., A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe, *Social Science Computer Review*, 1-18 (2015)
5. Cobben F. and Bethlehem J. G., *Web panels for official statistics*, Discussion Paper 201307, The Hague, The Netherlands: Statistics (2013).
6. Durrant G. B. and Steele F., Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys, *Journal of the Royal Statistical Society: Series A - Statistics in Society*, 361-381 (2009).
7. Istat, *Nota tecnica sulla produzione dei dati del Censimento Permanente: la stima della popolazione residente per sesso, età, cittadinanza, grado di istruzione e condizione professionale per gli anni 2018 e 2019*. <https://www.istat.it/it/files/2020/12/NOTA-TECNICA-CENSIPOP.pdf>, (2020).
8. Maslovskaya O., Durrant G. B., Smith P. W. F., Hanson T. and Villar A., What are the Characteristics of Respondents using Different Devices in Mixed-device Online Surveys? Evidence from Six UK Surveys, *International Statistical Review*, 87 (2) , 326-346 (2019).
9. Pratesi M., Lozar Manfreda K., Biffignandi S., and Vehovar V., List-based Web Surveys: Quality, Timeliness and Nonresponse in the Steps of the Participation Flow, *Journal of Official Statistics* (2004).

# Trusted Smart Surveys: architectural and methodological challenges at a glance

## *Un rapido sguardo ai problemi metodologici e architetturali nel contesto delle Trusted Smart Surveys*

Mauro Bruno, Francesca Inglese, Giuseppina Ruocco<sup>1</sup>

**Abstract** Smart statistics play an important role in the future of official statistics, in a world overwhelmed by smart technologies. This work illustrates the intermediate results of the ESSnet (European Statistical System network project) on smart surveys, which started its activities at the beginning of 2020. The main goals of the project are: (i) development and test of a (trusted) smart surveys framework within the European Statistical System (ESS), (ii) conceptualization of a new reference architecture for trusted smart surveys; iii) development of new methodological and technical skills within the ESS. This paper focuses on architectural and methodological challenges related to smart data sources, such as: i) combination of active and passive data collection; ii) errors due to measurement and to participant selectivity; iii) privacy concerns and ethical issues.

**Abstract** *Le statistiche 'smart' rivestono un ruolo significativo nell'ambito della statistica ufficiale, in un mondo caratterizzato da una tecnologia in continua evoluzione. Il lavoro illustra i risultati intermedi del progetto europeo ESSnet Smart Surveys, avviato all'inizio del 2020. I principali obiettivi del progetto sono: (i) sviluppare e testare un framework europeo di riferimento per le smart surveys; (ii) definire un nuovo modello architetturale per le smart surveys; (iii) sviluppare nuove professionalità metodologiche e tecniche. Il lavoro si focalizza sulle sfide metodologiche e architetturali legate alle fonti dati 'smart', ad esempio: i) combinazione di tecniche di rilevazione attive e passive; ii) errori di misurazione o dovuti alla selettività dei rispondenti; iii) privacy e valutazioni etiche.*

**Key words:** trusted smart surveys, sensor data, active/passive data collection

---

<sup>1</sup> Mauro Bruno, Istat; email: mbruno@istat.it  
Francesca Inglese, Istat; email: fringles@istat.it  
Giuseppina Ruocco, Istat; email: giruocco@istat.it

## 1 Introduction

Smart statistics play an important role in the future of official statistics, in a world overwhelmed by smart technologies. The term trusted smart surveys (TSS<sub>u</sub>) refers to an enhancement of the smart survey model using methodological and software solutions to increase the degree of trustworthiness and promote the public acceptance and participation. Constituent elements of a trusted smart survey are the strong protection of personal data based on privacy-preserving computation solutions, full transparency, and auditability of processing algorithms [1].

The development of smart statistics offers new possibilities to improve the quality of traditional social surveys. In smart statistics, respondents use smart devices (e.g., smartphones, activity trackers) to provide survey data. TSS<sub>u</sub> may include data collection processes in which respondents are asked to share existing data collected by trusted third parties, like government authorities and private stakeholders. Regardless of the type of smart data source, such new scenarios entail the revision of the statistical process to ensure accuracy and reliability, according to the principle of accountability and transparency of Official Statistics [2, 3].

This work illustrates the intermediate results of the ESSnet (European Statistical System network project) on Smart Surveys [4] which started its activities at the beginning of 2020. The main goals of the project are: (i) developing and testing (trusted) smart surveys framework within the European Statistical System (ESS), (ii) conceptualization of a new reference architecture for trusted smart surveys; (iii) development of new methodological and technical skills within the ESS. The ESSnet will deliver preparatory work to create a European wide platform to share and re-use smart survey solutions and components. The main benefits of implementing such a system are: i) reduction of respondent burden; ii) innovation of production processes to exploit new data sources; iii) a common framework to harmonise and enrich statistical surveys, as well as avoid data misuse.

## 2 Methodological challenges of a Trusted Smart Survey

The development of TSS<sub>u</sub> based on mobile devices offers new opportunities for social surveys to collect data, thus generating new and unfamiliar types of data that are not standardized in structure, format, or availability. The measurement capabilities of mobile devices can supplement or potentially even replace self-reports in surveys. In addition, internal sensors data collected passively, and respondents' activities on smartphones (e.g., taking pictures, scanning receipts) increase available data sources. Sensor data allow researchers to collect rich behavioural data, potentially with less measurement error and lower respondent burden than self-reports. However, there are multiple challenges to collecting these data: participant selectivity, (non) willingness to provide sensor data or perform additional tasks, privacy concerns and ethical issues, quality and usefulness of the

Trusted Smart Surveys: architectural and methodological challenges at a glance  
data and practical issues of implementation. These aspects have consequences in terms of both representation (selection) and measurement errors.

Representation errors are determined by the availability or not of a smartphone or other mobile devices by the individuals selected in the sample (coverage error) or by their willingness to participate (non-response). Participation is influenced by technological barriers, topic of the survey, duration of collection, respondent characteristics including privacy and security concerns and ability with smartphone and its tasks [4]. To maximize participation, it is necessary to define at the design phase the best contact strategies (app information, procedures to ensure data confidentiality, levels of incentives to counteract the privacy intrusiveness of passive data collection, etc.). However, a smart survey may not be able to reach parts of the population, making the sample of respondents not representative of the general population. Non-coverage remains a major problem, as this error can be the largest contributor to Total Survey Error [5]. To solve coverage issues, it may be necessary to consider other traditional data collection modes to intercept specific subpopulations not reachable through the tools of a smart survey.

Measurement errors depend on the nature of the smart survey data. For traditional data resulting from questionnaires or diaries, measurement errors are mainly due to the behaviour of the respondent and his familiarity with smartphone. For sensor data, measurement errors occur during the collect phase (errors in data generation, operating errors) and in the processing phase, such as in calibrating measured data, treatment of outliers, noise, etc., data manipulation to search for patterns or to explore the accuracy and precision of data.

The severity of the measurement errors needs to be analysed considering not only the type of sensor and analytic goals involved, but also the specific data sources. Indeed, a TSS<sub>u</sub> can employ more smart features, such as access to external sensors and personal and public online data, etc. In the TSS<sub>u</sub> design, many aspects must be considered, not least the integration of data from different sources and with different quality levels. New solutions for the collect phase and new methods for processing data, not yet explored in the traditional surveys, are relevant goals of the ESSnet on Smart Surveys.

### **3 Overview of TSS<sub>u</sub> architectural framework**

The design of an architectural framework for TSS<sub>u</sub> should be guided both by traditional survey requirements and by emerging issues related to new data sources (use of sensor data, use of machine learning techniques, use of data collected and held by private stakeholders, etc.).

The volume, dimension, and variety of such new data motivate the adoption of the Big Data paradigm. Therefore, the design activities have started from the Big Data REference Architecture and Layers (BREAL), developed in the context of the ESSnet Big Data II [5]. The main goal of the architectural analysis is to integrate BREAL concepts with a standard model used to describe traditional survey processes, i.e., GSBPM standard [6].

To this aim, the analysis has explored:

- *BREAL business functions* to identify the minimum set of BREAL business functions to describe the business layer of TSS<sub>u</sub> (e.g., Acquisition and Recording, Data Wrangling, Data Representation)
- *TSS<sub>u</sub> process steps* to highlight the relation between GSBPM sub-processes and BREAL business functions and specify for each GSBPM subprocess:
  - input and output data sources;
  - grouping subsets of related tasks (e.g., data ingestion, data transformation, data processing);
  - connections with overarching business functions (e.g., metadata, privacy, and quality).

The modelling of the TSS<sub>u</sub> architecture is an on-going activity, therefore the following sections briefly describe the preliminary result related to two GSBPM phases: *Run and Finalize Collection, Metadata Management*.

### **3.1 GSBPM Run and Finalize Collection process**

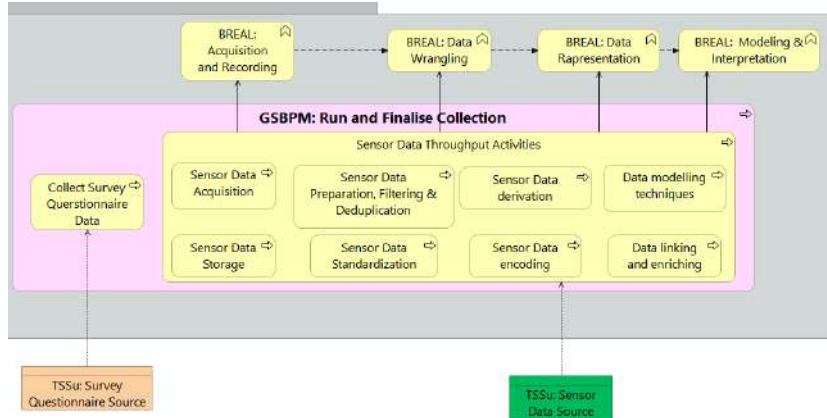
In the proposed model, the GSBPM phase *Run and Finalize collection* is split in two sub-processes: *Collect Survey Questionnaire Data* that includes traditional data processing techniques, and *Sensor Data Throughput Activities* that groups all tasks needed to transform collected sensor data in statistical data (Fig. 1). The combination of *Sensor Data Throughput Activities* sub-processes may vary in each survey, depending on the type of sensor data, agreements with the data provider, in-app, or in-house data processing. Once sensor data is stored, the processing tasks may be a combination of the following steps:

- *Data preparation & filtering*, a set of pre-processing steps to include only relevant information in the following tasks.
- *Data standardisation* for converting data to a target format.
- *Data derivation* to transform unstructured data to structured information.
- *Data encoding* for categorical data transformation in binary or numeric format.
- *Data modelling techniques*, grouping machine learning techniques and other methods to extract statistical information from sensor data.
- *Data linking and enriching* to integrate questionnaire and derived sensor data.

These activities allow to perform the following BREAL business functions: Acquisition and Recording, Data Wrangling, Data Representation and Modelling and Interpretation.

Trusted Smart Surveys: architectural and methodological challenges at a glance

**Figure 1:** TSS<sub>u</sub> Run and Finalize Collection process



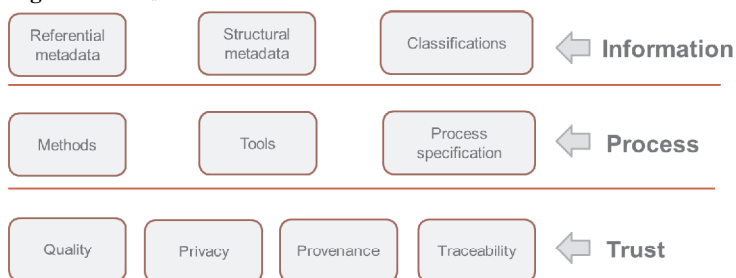
### 3.2 GSBPM Metadata management overarching process

The GSBPM business function *Metadata management* refers to several dimensions, such as variable description, classifications, data quality, provenance, etc. In order to reduce the complexity of such broad subject, the metadata involved in TSS<sub>u</sub> have been divided in three main subsets:

- *Information subset*, grouping referential and structural metadata, as well as classifications.
- *Process subset*, including the metadata for modelling data transformation in a statistical process chain.
- *Trust subset*, enabling the measurement of the trust level of a trusted smart survey, and monitoring data quality and provenance, privacy constraints and process traceability.

An analysis of existing frameworks and ontologies has been performed to identify and describe the items belonging to each element included in every subset. The following figure shows the metadata subsets, grouped according to the above criteria.

**Figure 2:** TSS<sub>u</sub> Metadata subsets



## 4 Conclusions

Further investigation is needed to design all the GSBPM phases involved in the statistical production chain. In general terms, the design of the business layer provides valuable insight that will guide future methodological and technological implementation activities. More precisely, the design of the business layer will allow to classify all the processes and sub-processes involved in the treatment of traditional and new data sources in TSS<sub>u</sub>. There are still many open challenges that will be addressed by several Proofs of Concepts during the second part of the ESSnet project. The exploration will concern both methodological and technical-architectural aspects: active and passive (sensor) data collection, machine learning algorithms for managing sensor data, incentives to reduce respondents' burden, in-app or in-house sensor data processing, metadata specification, technical infrastructure, and privacy techniques.

## References

1. Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., Skaliotis, M.: Trusted smart statistics: motivations and principles. *Statistical Journal of the IAOS*, 35 (2019). <https://ec.europa.eu/eurostat/cros/system/files/sji190584.pdf> 2)
2. United Nations. Fundamental principles of official statistics. Official Resolution adopted by the UN General Assembly on 29/1/2014. <https://unstats.un.org/unsd/dnss/gp/fp-new-e.pdf> G. Cited 15 Jan 1999. [http://www.rsc.org/dose/title of subordinate document](http://www.rsc.org/dose/title%20of%20subordinate%20document).
3. European statistics Code of Practice – revised edition (2017). <https://ec.europa.eu/eurostat/web/products-catalogues/KS-02-18-14>
4. Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P.; Kreuter, F.: Willingness to Participate in Passive Mobile Data Collection. In *Public Opinion Quarterly* 83, pp. 210–235 (2019).
5. Biemer, P. P., de Leeuw, E., Eckman S., Edwards B., Kreuter T., Lyberg L. E., Tucker N. C.: *Total Survey Error in Practice*. West B. T. (Editors). John Wiley & Sons, Inc., Hoboken, New Jersey (2017).
6. ESSnet on Smart Surveys (2020-2021). [https://ec.europa.eu/eurostat/cros/content/essnet-smart-surveys\\_en](https://ec.europa.eu/eurostat/cros/content/essnet-smart-surveys_en)
7. ESSnet on Big Data II, Work Package F, Deliverable F1. (2018-2021). [https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPF\\_Deliverable\\_F1\\_BREAL\\_Big\\_Data\\_Reference\\_Architecture\\_and\\_Layers\\_v.03012020.pdf](https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPF_Deliverable_F1_BREAL_Big_Data_Reference_Architecture_and_Layers_v.03012020.pdf)
8. Generic Statistical Business Process Model (GSBPM) v. 5.1. January (2019). <https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1>
9. Ricciato, F., Giannakouris, K., Wirthmann, A., Hahn, M.: Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics. *SIS* (2020). <https://it.pearson.com/content/dam/region-core/italy/pearson-italy/pdf/Docenti/Universit%C3%A0/Pearson-SIS-2020-atti-convegno.pdf>
10. Ricciato, F., Bujnowska, A., Wirthmann, A., Hahn, M., Barredo-Capelot, E.: A reflection on privacy and data confidentiality in official statistics. *ISI World Statistics Congress* (2019). [https://www.bis.org/ifc/events/isi\\_wsc\\_62/ips177\\_paper3.pdf](https://www.bis.org/ifc/events/isi_wsc_62/ips177_paper3.pdf)

# On bias correction in small area estimation: An M-quantile approach

## *Bias correction in stima per piccole aree: un approccio M-quantile*

Gaia Bertarelli, Francesco Schirripa Spagnolo, Raymond Chambers and David Haziza

**Abstract** In this paper we propose two bias correction approaches in order to reduce the prediction bias of the robust M-quantile predictors in small area estimation in the presence of representative outliers. A bootstrap procedure is considered for the estimation of the mean squared error. A Monte-Carlo simulation study is conducted. Results confirm that our approaches improve the efficiency and reduce the prediction bias of M-quantile predictors when the population contains units that may be influential if selected in the sample.

**Abstract** *L'obiettivo del lavoro è quello di proporre due approcci per ridurre l'errore di predizione degli stimatori basati modello di regressione M-quantile nella stima per piccole aree in presenza di outliers rappresentativi. Per valutare la variabilità degli stimatori è utilizzato un approccio bootstrap. Uno studio di simulazione è stato implementato ed i risultati hanno evidenziato che gli approcci proposti migliorano l'efficienza e riducono l'errore di predizione quando la popolazione contiene unità che possono essere influenti se selezionate nel campione.*

**Key words:** Robust methods; Small Area Estimation; M-quantile

---

Gaia Bertarelli

Istituto di Management Scuola Universitaria Superiore Sant'Anna, Pisa, Italy, e-mail: gaia.bertarelli@santannapisa.it

Francesco Schirripa Spagnolo

Dipartimento di Economia e Management, Università di Pisa, Pisa, Italy, e-mail: francesco.schirripa@ec.unipi.it

Raymond Chambers

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia, e-mail: ray@uow.edu.au

David Haziza

Département de mathématiques et de statistique, Université de Montréal, Canada, e-mail: haziza@dms.umontreal.ca



## 1 Introduction

Outliers occur frequently in sample surveys when the data distribution is highly skewed. Accordingly, to the terminology of [3] sample outliers can be classified into two categories. The first type is the ‘non-representative outliers’, which are sample elements whose data values are incorrect or they are unique. In this case, they can be identified and removed or corrected before estimation. However, in other cases, sample values associated with the outliers have been correctly recorded and they cannot be considered as unique. These are called ‘representative outliers’ because they are representative of the non-sampled part of the population; in other words, there is no reason to assume that there are no more similar outliers in the non-sampled part of the population. Such outliers values can seriously affect the survey estimates. Consequently, several methods have been developed in order to mitigate the effects of outliers on survey estimates.

Representative outliers are even more concerning in the small area estimation (SAE) context, where sample sizes are very small and the estimation is often model-based [5]. [4] addressed the issue of outlier robustness in SAE by proposing an M-quantile approach aiming at overcoming the issue of outliers by avoiding the normal assumption. [7] addressed the same issue from the perspective of linear mixed models. Both these approaches use plug-in robust prediction replacing parameter estimates in optimal but outlier-sensitive predictors by outlier robust versions. These predictors are efficient under the correct model but may be sensitive to the presence of outliers because they use plug-in robust prediction which usually leads to a low prediction variance and a considerable prediction bias. [6] and [5] proposed a bias correction method for models with continuous response variables. The main aim of this work is to propose new M-quantile predictors in SAE with correction terms for the bias. Two approaches are studied. The first estimator is a unified approach to M-quantile predictors based on a full bias correction and it could be viewed as a generalization of [3]. The second proposal is developed following the conditional bias approach by [1] and [6].

## 2 Bias corrected M-quantile-based estimator

Let  $\theta_i$  be a finite population parameter for area  $i$ . That is,  $\theta_i$  is a well-defined function of the values of a random variable  $Y$  associated with the  $N_i$  elements of such a small area finite population of interest. For ease of notation, we assume that both  $Y$  and  $\theta_i$  are scalar, and we denote

$$\theta_i = f(\mathbf{y}_{U_i}),$$

where  $\mathbf{y}_{U_i}$  denotes the vector of population values of  $Y$  for small area  $i$  and  $f$  is a known function. A basic sample survey inference problem is then one of predicting the value of  $\theta_i$  given a sample of  $n < N$  values from  $\mathbf{y}_U$ . Without loss of generality we

put  $\mathbf{y}_s$  equal to the population sub-vector defined by these values, where  $s$  denotes the set of sampled population units. We define (i)  $\mathbf{y}_{U_i}$  vector of population values of  $Y$  for area  $i$  with  $U = \bigcup_{i=1}^m U_i$  with  $m$  is the number of small areas; (ii)  $\mathbf{y}_{s_i}$  vector of sampled population values in small area  $i$  with  $s = \bigcup_{i=1}^m s_i$ . Suppose that, given  $\mathbf{y}_{s_i}$  we can impute the remaining values  $\hat{\mathbf{y}}_{U_i}$  denote this imputed vector. A popular method of predicting the unobserved value of  $\theta_i$  is via the Plug-In Predictor (PIP)

$$\hat{\theta}_i = f(\hat{\mathbf{y}}_{U_i}). \tag{1}$$

Adopting a model-based approach, the empirical PIP for  $\theta_i$  based on this plug-in approximation is

$$\hat{\theta}_i = f(\mathbf{y}_{s_i}, \{y_{ij}^{opt}; j \in U_i - s_i\}) \tag{2}$$

where the set  $U_i - s_i$  contains the  $N_i - n_i$  indices of the non-sampled units,  $y_{ij}^{opt} = E[y_{ij} | \mathbf{y}_s; \delta = \hat{\delta}]$  is the plug-in approximation of the minimum mean squared error predictor (MMSEP) of  $y_{ij}^{opt}$  for a non-sampled population unit  $j$  for area  $i$ , and  $\delta$  is a vector of unknown parameters. The above PIP (2) for small area can be also computed using the M-quantile approach. It can be obtained by using the estimated regression coefficients by M-quantile approach,  $\hat{\beta}_\tau$ , leading to

$$\hat{\theta}_i^{MQ} = f(\mathbf{y}_{s_i}, \{g^{-1}(\mathbf{x}_{ij}^T \hat{\beta}_\tau); j \in U_i - s_i\}), \tag{3}$$

where  $\tau_i$  represents the order of M-quantile for area  $i$ . Its computation varies depending on the type of the data.

We propose two small area estimators based on Generalised version of M-quantile regression models.

The first estimator is a unified approach to M-quantile predictors based on a full bias correction. Following [3], the first order approximation to the prediction bias of  $\hat{\theta}_i^{MQ}$  is

$$E[\hat{\theta}_i^{MQ} - \theta_i] \simeq \sum_{j \notin s_i} \left( \frac{\partial f}{\partial y_{ij}} \right)_{\mathbf{y}_{U_i} = \mathbf{m}_{U_i}} E[\hat{y}_{ij} - y_{ij}] \simeq \sum_{i \in r_j} \left( \frac{\partial f}{\partial y_{ij}} \right)_{\mathbf{y}_{U_i} = \hat{\mathbf{m}}_{U_{q_j}}} \left( \frac{\partial g^{-1}}{\partial \eta} \right)_{\eta = \mathbf{x}_{ij}^T \hat{\beta}_{q_j}} \mathbf{x}_{ij}^T E[\hat{\beta}_{q_j} - \beta_{q_j}],$$

The bias corrected robust predictor MQC for the population average of  $Y$  in the  $i$ th area will be:

$$\theta_i^{MQC} = N_i^{-1} \left( \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij} + \sum_{j \in r_i} \left( \frac{\partial f}{\partial y_{ij}} \right)_{\mathbf{y}_{U_i} = \hat{\mathbf{m}}_{U_{q_j}}} \left( \frac{\partial g^{-1}}{\partial \eta} \right)_{\eta = \mathbf{x}_{ij}^T \hat{\beta}_{q_j}} \mathbf{x}_{ij}^T \hat{\mathbf{B}}_i \right) \tag{4}$$

where  $d_{jh\bar{q}_j} = 2 \{ \bar{q}_j I(r_{hj} > 0) + (1 - \bar{q}_j) I(r_{hj} \leq 0) \}$  and  $\hat{\mathbf{B}}_i$  has to be computed depending of the type of the response variable. If  $y_{ij}$  is continuous

$$\hat{\mathbf{B}}_i = \left( \sum_{h=1}^m \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \mathbf{x}_{hj}^T \right)^{-1} \sum_{h=1}^m \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \hat{\sigma}_{hj} \phi \left\{ \frac{y_{hj} - \mathbf{x}_{hj}^T \hat{\beta}_{\tau_i}}{\hat{\sigma}_{hj}} \right\}. \tag{5}$$

The second proposal is developed following the conditional bias approach by [1] and [6]. In a model based approach, the conditional bias attached to unit  $ij$  is

$$B_{ij} = E[\hat{\theta} - \theta | s; Y_{ij} = y_{ij}].$$

The prediction error  $\hat{\theta}_i - \theta_i$  can be approximated as:

$$\hat{\theta}_i - \theta_i \simeq \sum_{j \in r_i} B_{ij}(I_{ij} = 0) + \sum_{j \in s_i} B_{ij}(I_{ij} = 1). \tag{6}$$

To determine the conditional bias, we need to distinguish two cases, whether the unit belongs to the sample or not. The main problem is that the conditional bias of a non-sampled unit can't be estimated since it depends on the  $Y$ -values on the non-sample units, which are not observed. A robust predictor of the mean in the  $i$ th area can be expressed as

$$N_i^{-1} \left( \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \boldsymbol{\beta}) - \sum_{j \in s_i} B_{ij}(I_{ij} = 1) + \phi \left\{ \sum_{j \in s_i} B_{ij}(I_{ij} = 1) \right\} \right)$$

where  $\phi$  is the Huber function. Translating the idea for MQ we have:

$$\theta_i^{MQD} = N_i^{-1} \left( \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\hat{q}_j}) - \sum_{h=1}^m \sum_{j \in s_h} \hat{B}_{jh}(I_{jh} = 1) + \phi \left\{ \sum_{h=1}^m \sum_{j \in s_h} \hat{B}_{jh}(I_{jh} = 1) \right\} \right). \tag{7}$$

The  $\phi$ -function in MQD depends on a tuning constant  $c$ . Using min-max method to compute the optimal tuning constant we obtain

$$\theta_i^{MQD} = N_i^{-1} \left( \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\hat{q}_j}) - \frac{1}{2} (\min \{B_{jh}(I_{jh} = 1)\} + \max \{B_{jh}(I_{jh} = 1)\}) \right) \tag{8}$$

where the conditional bias for unit  $j$  has to be computed depending of the type of the response variable. If  $y_{ij}$  is a continuous

$$\hat{B}_{hj}(I_{hj} = 1) = \sum_{i \notin s_i} \mathbf{x}_{ij}^T \left\{ \sum_{h=1}^m \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \mathbf{x}_{hj}^T \right\}^{-1} \hat{d}_{hj} \mathbf{x}_{hj} (y_{hj} - \mathbf{x}_{hj}^T \hat{\boldsymbol{\beta}}_{\epsilon_j}). \tag{9}$$

### 3 Model-based simulations

In this section, we provide results regarding model-based simulation scenarios for continuous variables. Following [5], population data are generated from  $m = 40$  small areas with samples selected by a simple random sampling without replacement within each area. The population and sample size are the same for all areas and are fixed at  $N_i = 100$  and  $n_i = 5$ . Values for  $x$  are generated as i.i.d. from a lognormal distribution with a mean of 1 and a standard deviation of 0.5 on the log scale. Values for  $Y$  are generated as  $y_{ij} = 100 + 5x_{ij} + u_i + \epsilon_{ij}$ , where  $i$  refers to

the areas and  $j$  to the population units. The random area and individual effects are independently generated according to the following scenarios:

- a) [0,0,0] - no outliers,  $u \sim N(0, 3)$  and  $e \sim N(0, 6)$ ;
- b) [e,0,0] - individual outliers only,  $u \sim N(0, 3)$  and  $e \sim \delta N(0, 6) + (1 - \delta)N(20, 150)$ ;  $\delta \sim Ber(0.03)$ ;
- c) [e,u,0] - outliers in both area (fixed) and individual effects,  $u \sim N(0, 3)$  for areas 1–36,  $u \sim N(9, 20)$  for areas 37–40 and  $e \sim \delta N(0, 6) + (1 - \delta)N(20, 150)$ .

Each scenario is independently simulated 1000 times. For each simulation the population values are generated according to the underlying scenario, a sample is selected in each area and the sample data are then used to compute estimates of each of the actual area means for  $y$ . Nine different estimators are used for this purpose: the M-quantile estimator MQ by [4] which serves as a reference for the MQ regression based estimators, the bias corrected M-quantile estimator MQBC by [5], the M-quantile estimator based on full bias correction MQC (see equation (4)), the M-quantile estimator based on conditional bias correction MQD (see equation (8)), the standard EBLUP which serves as a reference for all the considered estimators, the robust eblup REBLUP by [7] and its robust bias corrected version REBLUP-BC by [5], the CBEBLUP and CEBLUP predictorS by [6]. The influence function  $\phi$  that is used in MQBC, MQC, REBLUP BC, CBEBLUP and CEBLUP is a Huber proposal 2 type. For each estimator, we test three different tuning constant for the bias correction part equal to 3, 6 and 9. The performance of the proposed indicators is evaluated according to min-max plots (Figure 1). The values on the  $x$ -axis and  $y$ -axis on plots are:

$$AbsRBias = \frac{\text{Median}[AbsB(\theta_{ki})] - \min\{\text{Median}[AbsB(\Theta_i)]\}}{\max\{\text{Median}[AbsB(\Theta_i)]\} - \min\{\text{Median}[AbsB(\Theta_i)]\}}$$

and

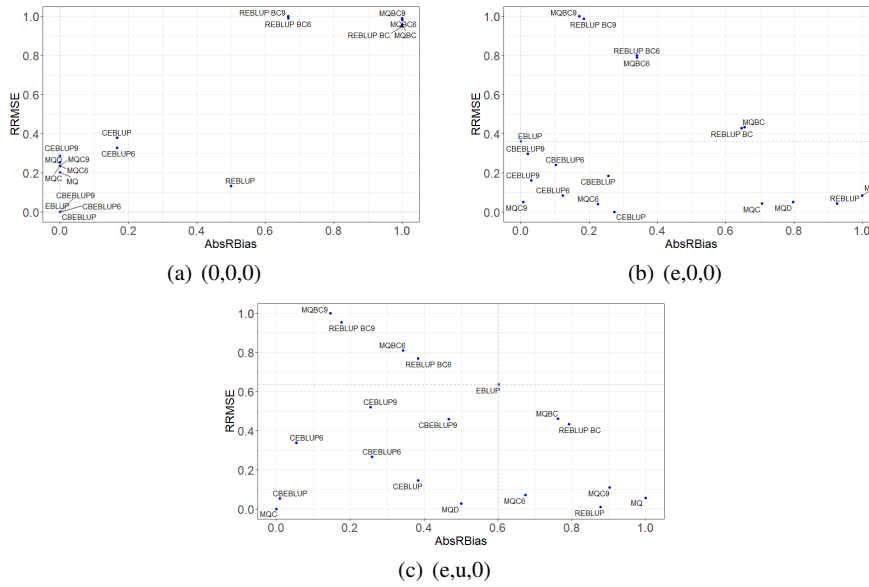
$$RRMSE = \frac{\text{Median}[RRMSE(\theta_{ki})] - \min\{\text{Median}[RRMSE(\Theta_i)]\}}{\max\{\text{Median}[RRMSE(\Theta_i)]\} - \min\{\text{Median}[RRMSE(\Theta_i)]\}},$$

where  $\theta_{ki}$  is the  $k$ th estimator in the  $i$ th area and  $\Theta_i$  is the vector all  $K$  predictors in area  $i$ .

Results confirm our expectations regarding the behaviour of the MQC and MQD estimators. With respect to MQ estimator, the new proposed estimators reduce the bias in the presence of outliers and the variance with only unit-outliers.

We now examine the performance of the MSE estimators. We use the bounded-block-bootstrap [2] for MQC and MQD, with a constant equal to 3 for scenarios a and b and equal to 1.345 for c. Results are reported in table 3.

Scenario	Estimator								
	MQ	MQBC	MQBC6	MQBC9	MQC	MQC6	MQC9	MQD	
[0,0,0]	-3.38	-6.85	-5.54	-5.38	2.16	2.12	2.14	2.24	
[e,0,0]	-18.01	-8.90	-5.16	-3.77	-2.32	-1.52	-2.26	-3.26	
[e,u,0]	-11.09	-8.96	-5.22	-3.83	4.15	-0.53	-2.66	3.51	



**Fig. 1** Min-Max plots for MQ, MQBC, MQC, MQD, EBLUP, REBLUP, REBLUP BC, CBE-LUP and CEBLUP under selected simulation scenarios.

## References

- [1] Beaumont J.F., Haziza D., Ruiz-Gazen A.: A unified approach to robust estimation in finite population sampling. *Biometrika* **100(3)**, 555–569 (2013)
- [2] Bertarelli G., Chambers, R., Salvati, N.: Outlier robust small domain estimation via bias correction and robust bootstrapping. *Stat. Methods Appl.* (2020) doi:10.1007/s10260-020-00514-w-
- [3] Chambers, R.: Outlier robust finite population estimation. *JASA* **81(396)**, 1063–1069 (1986)
- [4] Chambers, R., Tzavidis, N.: M-quantile models for small area estimation. *Biometrika* **93(2)**, 255–268 (2006)
- [5] Chambers, R., Chandra, H., Salvati, N., Tzavidis, N.: Outlier Robust Small Area Estimation. *J. Roy. Stat. Soc. B* **76(1)**, 47–69 (2014)
- [6] Dongmo-Jiongo, V., Haziza, D., Duchesne, P.: Controlling the bias of robust small area estimators. *Biometrika* **100(4)**, 843–858 (2013)
- [7] Sinha, S. K., Rao, J.: Robust small area estimation. *Canadian Journal of Statistics* **37(3)**, 381–399 (2009)
- [8] Chambers, R., Salvati, N. and Tzavidis, N.: Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the uk. *Journal of the Royal Statistical Society: Series A* **179 (2)**, 453-479 (2016)

# The address component of the Statistical Base Register of Territorial Entities

## *La componente indirizzi del Registro Statistico Base dei Luoghi (RSBL)*

D. Fardelli, E. Orsini, A. Pagano<sup>1</sup>

**Abstract** The new Statistical Base Register of Territorial and geographical entities of ISTAT (RSBL) is a multidimensional register integrating several components addresses with geographic coordinates, micro-zones and census blocks, buildings and housing units, administrative zones, statistical and functional zones. All components will be integrated with other components according hierarchical and geographical principles. The RSBL, with the other Registers of Institute, will provide a bridge between the statistical units, such as individuals, families and economic. In this paper, we present the component of addresses of RSBL, illustrating his structure, his process and some preliminary results on the data contained in it about the geocoding and georeferencing.

**Abstract** *Il Registro Statistico di Base dei Luoghi (RSBL) è il pilastro di tutte le attività che prevedono la georeferenziazione delle informazioni statistiche contenute negli altri registri o raccolte attraverso le indagini. La componente indirizzi di RSBL acquisisce le informazioni relative agli indirizzi derivanti dal progetto ANNCSU e attraverso il processo di integrazione, implementa al suo interno gli indirizzi di diversi archivi amministrativi (Lista Anagrafica, Anagrafe Tributaria, Catasto, Asia). Ogni indirizzo sarà corredato, oltre ad indicatori di qualità dedicati, di coordinata geografica e di sezione di censimento. Tale infrastruttura permetterà la georeferenziazione e la geocodifica alla sezione di censimento e alla griglia regolare delle unità statistiche (individui, famiglie, unità locali, etc.).*

**Key words:** Register, Integration, Georeferencing, Geocoding, Addresses, Geographic Coordinates, Harmonization, Geospatial integration.

---

<sup>1</sup> Davide Fardelli, ISTAT; fardelli@istat.it;  
Enrico Orsini, ISTAT; eorsini@istat.it;  
Andrea Pagano, ISTAT; andreapagano@istat.it.

## 1 Introduction

One of the pillars of NSIs modernization program is the system of integrated statistical registers as the basis for surveys and statistical production; this system has been denominated Integrated System of Statistical Registers (ISSR). The ISSR integrates information relating to: (i) individuals, families and cohabitation; (ii) economic units; (iii) places; and (iv) activities [4] [9]. In this system, every unit referred to places will be inside the Statistical Base Register of Territorial Entities (RSBL). The base registers are connected by codes and are maintained updated over time using mainly administrative sources [8]. According to Global Statistical Geospatial Framework (GSCF) [7] [3], the geospatial information is an important data sources for statistics. Therefore, RSBL will assume a dual role: (i) georeferencing and/or geocoding the statistical units (demographic/economic) [2] and (ii) spatial data production (e.g. surfaces, altitudes, distances, contiguities, statistics on buildings, population grid, etc.). The RSBL has been designed like a multidimensional register integrating several components with heterogeneous nature: addresses with geographic coordinates, micro-zones (old census block), buildings and dwellings units, administrative zones and statistical and functional zones. All components will be integrated with other components according hierarchical and geographical principles. The RSBL, with the other Registers, also will provide a bridge between the statistical units, such as individuals, families and economic. They will be geocoded at census block or regular grid. In this paper, we will illustrate the addresses component of RSBL. It has been released in a prototypal form in 2018 and a new release in 2020. The RSBL has been used like sample frame for the permanent census of population. The aim is to build the register only once and to keep it updated in time.

## 2 The structure of RSBL-Addresses

The addresses component of RSBL should include all the addresses existence on national territory. Every address will be admitted and identified in RSBL with a unified address code (CUI). The attribution of a code will simplify the integration with other registers, and the code will avoid errors of linkage, due at several form of strings of addresses. Every CUI will have a geographic coordinate and/or census block geocoded. The geographic information is always accompanied with quality indicators both of coordinate and of geocoding.

The innovation of RSBL is the integration of addresses from several administrative and geographical archives. The statistic unit in this component is the address interpreted like the direct or indirect access, from a street to a housing unit or other units like economic activities.

The National Archive of Addresses of Urban Streets (ANNCSU), born in 2012, is the first archive to populate RSBL. It is the primary source of register and it is considered like a benchmark, because is provided straight by municipality. It represents the administrative archive, which contains streets and addresses for the entire country

The address component of the Statistical Base Register of Territorial Entities (about 23.7 million of addresses). Each address is geocoded at the census block, and some municipality have started to insert the geographic coordinate too. In order, the second administrative archive is the municipal register of resident population (LAC). From LAC, we have extracted about 17.5 million of addresses. This source is very crucial, because the addresses are the key linkage straight with Population Register. The third source uploaded on RSBL is the register of the tax Agency (AT). We have treated about 43.5 million of addresses. In AT archive there are all the individuals registered at Internal Revenue Agency (IRS). The fourth administrative archive used is the Real Estate Registry of the Land Property, where there are about 17.8 million of addresses. This source is also important because each CUI will be associated with one or more unit of building and dwellings, so it will be determinant to integrate the component addresses and buildings/dwellings. The fifth, and the last – for the moment – source inserted in RSBL is the list of all the addresses used in the census population of 2011 related at family and buildings. This source is fitted of 11 million of addresses with associated the census block. Actually, inside RSBL have been processed about 110 million of addresses deriving from several administrative archives.

The register is built starting from the identification process of the address, based on an identification algorithm that performs a deterministic record linkage with a probabilistic threshold on its database acquired by commercial company. The software is specialized in automatic normalization of databases, eliminating errors, completing partial or missing information. Moreover, the software allows to verify and to certify that the information is univocal, through a complete and sophisticated deduplication process. Indeed, the purpose of normalization and deduplication is to ensure a high level of data quality in the Register. The algorithm, given the address string, performs three operations: (i) identification and certification of the quality of the address, (ii) normalization of the address string, and (iii) georeferencing of the address with latitudinal and longitudinal information. Normalization provides a form of the address string, because the official form remains in ANNCSU.

The integration of the sources is implemented in the register by loading the set of official and distinct addresses in *Address Registry*; instead, the set of different forms corresponding to the dictionary of synonyms are storage in the *Thesaurus*; and the set of unidentified addresses are temporarily discarded, to be revisited deeply. To obtain an efficiency integration of sources, RSBL has been projected according to an architecture to archive all the geographic and database information. The use of fundamental geospatial infrastructure and geocoding is one of the five principles of GSGF [7]. In particular, the availability of point coordinates will be of great importance for spatial analysis and the production of statistics on a regular grid, as required by the modernization of European statistics [1]. The set of official and distinct addresses are in the Address Register, including the geographical coordinate coherently placed in the administrative/statistical areas of the national territory. Each official address recognized in the system has a unique code (CUI) and a cluster of addresses from different sources, denominated forms of the official address. The set of different forms corresponds to the Thesaurus. The set of unrecognized addresses is stored in the register too. The next step involves the unrecognized addresses in deterministic and probabilistic parsing processes in order to increase the coverage of addresses recognized as official or forms of register.



The geographical coordinate constitutes the georeferencing component of the register and finally, through geocoding processes, the administrative/statistical component of the register is constructed. The architecture thus created connects all these components by associating them.

### 3 The geocoding process and harmonization

According the principle 1 of GSGF, the use of geospatial infrastructure and geocoding supports high quality, standardised physical address, property or building identifier, or other location description, in order to assign accurate coordinates and/or a small geographic area or standard grid reference to each statistical unit [7]. The Statistical Commission in 2018 encourage to assign to statistical unit record data with a location reference, and that ideally it should allow for geospatial coordinates with x- and y-values to be produced for each record [6].

Inside RSBL, an important step is the validation of the association between address, census block and municipality. The process of geocoding has a key role, because it has the role to allocate every address in a census block and in one municipality. A strong process with less error will implicate an output in municipal sub-level not distorted in all the statistical unis.

**Figure 1:** The representation of Coordinates, Census block and Municipality Borders



First of all, we calculate the census block for every address with the coordinate, according the several level of coordinates (punctual/interpolated/approximated) available in RSBL. This is possible with the use of GIS tools, specifically spatial join [5], as showed in figure 1. It is relevant to attribute an evaluation of quality for the deviation of address as real position and coordinate geographic (positional accuracy). Actually, we have three levels of quality: from Address Point (punctual), from

The address component of the Statistical Base Register of Territorial Entities vectorial road network database (interpolated) and, at least, the approximated when we take a coordinate from a close address (max 10 house numbers). We are implementing in RSBL, the coordinates of cadastral buildings too. In this way we are coding all the stages, so we have the information that indicate the accuracy of the location coordinate obtained for each address. The second step is the loading in the register of the census block declared in ANNCSU. The third step is the loading in the register of the census block declared during the Census of population in the 2011. The last phase is the integration of the three sources of information. Across the analysis of distribution of about 70 indicators on the convergence and divergence, it is possible attribute one census block for every address. The geocoding process will have three levels of quality (high, medium and low). In this process is verified the coherence between house number, census block and municipality. The figure 1 shows the house numbers as points, the census block borders as red lines and the municipality border as blue line. The use of the coordinates allows and strengthen the consistency of the association between each house number within the current census block within the effective municipal boundaries and will allow maintaining the correct association over time in harmony with the territorial variations.

#### **4 Results**

The RSBL-Addresses integrates five administrative sources, providing about 29 million addresses, with 27 million of addresses validate and certification of house numbers and 2 million with low quality referring only to the street. In total, over the 100 million addresses were processed, of which just about 29 million were univocally identified as a canonical and official form and assigned with a unique code (CUI). Just over 70 million addresses have been identified as synonyms for the canonical version and the remaining 4 million are the unidentified addresses for which the control and correction process will be starting. In RSBL has been focused a measurement of quality for addresses and geocoding process. Every address can be summarized according to indicator in high or low quality. About 27.8 million of addresses are with high quality and about 1 million with low quality. The addresses with low quality represent a big cluster because their identification is certificate until the street than house number. For the low quality addresses is need a strong operation to elevate them in high quality, because, actually is not possible geocoding to apply the process geocoding. It is also important to understand the under-coverage of the ANNCSU than of RSBL.

The table 1 describes the output of RSBL-Addresses observing the data in two dimensions: the georeferencing and the geocoding with quality indicators.

The high quality of georeferencing is assigned when there is a punctual coordinate, the medium quality is assigned when there is an interpolated coordinate, and a low quality when the coordinate is approximated to close address. It is important to increase the coverage of coordinates because all statistical unit record data should be collected or associated with a location reference, and that ideally it should allow for

geospatial coordinates with  $x$ - and  $y$ -values. When implementation of coordinates is not possible, we need an approximation at street level, and not at address level.

**Table 1:** Distribution of the quality of the geographic coordinates and addresses geocoded for the high quality CUI

		Geocoding					
		High	Medium	Low	No Geoc.	Total	(%)
Georeferencing	High	13.018303	1.388.058	235.751	3.651	14.645.763	52%
	Medium	2.475.985	1.750.426	129.168	3.263	4.358.842	15%
	Low	2.734.797	1.872.963	841.552	3.759	5.453.071	19%
	No Coord.	1.819.761	889.097	781.615	326.250	3.816.723	13%
	Total	<b>20.048.846</b>	<b>5.900.544</b>	<b>1.988.086</b>	<b>336.923</b>	<b>28.274.399</b>	100%
	(%)	71%	21%	7%	1%	100%	

## 5 Conclusions

The RSBL-addresses integrate several sources of administrative and geographic archives for the purpose to assign and validate a unique code for every single address (CUI). Every CUI should be georeferenced and geocoded to grid/census block. In this way will be possible to associate every statistic unit (economic/demographic) at the territory. In this paper, we presented some of the main preliminary results. The next step will be to update of the sources, increase the quality of the indicators of georeferencing and geocoding, increase the coverage of addresses georeferenced and reduce the discarded addresses.

## References

1. Commission Implementing Regulation (EU) 2018/1799 of 21 November 2018, in Official Journal of the European Union L296/19 22/11/2018.
2. Conference of European Statisticians Sixty-sixth plenary session Geneva, Guidelines on the use of registers and administrative data for population and housing censuses (2018).
3. Haldorson M., Mostrom J., Implementing the Statistical Geospatial Framework at Statistics Sweden, GEOSTAT 3- National Report (2018).
4. Radini R., Scannapieco M., Garofalo G., The Italian Integrated System of Statistical Registers: On the Design of an Ontology-based Data Integration Architecture, NTTS, Brussels (2017).
5. Rainer N., Kaminger I. and Katzlberger G., Adding Value to Statistical Information by Georeferencing, 98th DGINS Conference and 14th ESSC Meeting Prague (2012).
6. United Nations, Economic and Social Council, Statistical Commission - Report of the Expert Group on the Integration of Statistical and Geospatial Information, 6-9 March 2018, E/CN.3/2018/33.
7. United Nations, UN-GGIM, The Global Statistical Geospatial Framework, Department of Economic and Social Affairs-Statistic Division (2019).
8. UNECE, Register-based statistics in the Nordic countries: review of best practices with focus on population and social statics, United Nations Publication, ISBN 978-92-1-116963-8 (2007).
9. Wallgren B., Wallgren A., Register-based Statistics: Statistical Methods for Administrative Data, John Wiley & Sons, Ltd (2014).

# **A well-being municipal indicator using census data: first results**

## *Un indicatore di benessere comunale su dati di censimento: primi risultati*

Massimo Esposito

**Abstract** An increasing attention was recently devoted to the study of Sustainable Development Goals (SDGs) and then to the research of detailed territorial indicators able to measure them. This paper aims at reconstructing, for Italy, the evolution of a well-being municipal indicator, starting from the half of XX century and using census data. Here a first exploratory analysis, referred to the 1951 general population census, is presented. The results reflect, at least partially, the well-known territorial differentials among Northern and Southern regions.

**Abstract** Negli ultimi anni una crescente attenzione è stata rivolta allo studio degli obiettivi di sviluppo sostenibile (SDGs) e alla ricerca di indicatori territoriali dettagliati che lo misurano. Questo contributo si propone di ricostruire l'evoluzione in Italia, a partire dalla metà del secolo scorso, di un indicatore di benessere a livello comunale utilizzando dati di censimento. Si presentano gli esiti di una prima esplorazione dei dati, relativi al censimento generale della popolazione del 1951. I risultati ottenuti rispecchiano almeno in parte i ben noti differenziali territoriali fra Nord e Sud del paese.

**Key words:** SDGs, well-being indicators, population census

---

<sup>1</sup> Massimo Esposito, Department of Economics and Business, University of Sassari; email: mesposito@uniss.it

## 1 Introduction. Data and methods

In the last years the problem of finding a comprehensive measure of well-being, including not only economic aspects but also social and environmental ones, has involved public and private actors. In Italy, since 2010 the National Institute of Statistics has launched a wide research project (called “BES”, Benessere equo e sostenibile) in order to define such a measure. Every year it produces a report in which well-being is considered according to a multidimensional approach. Information at regional level are also provided. The literature on this topic is very wide: in the references section some of the most recent studies focussed on Italy are reported.

It would be interesting to extend such analysis in two directions: first, considering a more detailed territorial level. Second, investigating the past. This is the starting point of this work, aiming at reconstructing a well-being indicator for the Italian municipalities referred to the past decades.

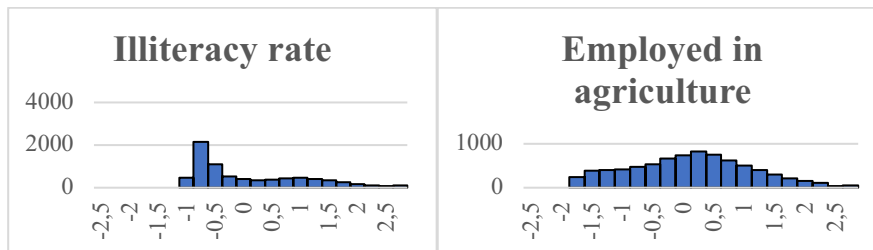
The first exploration of the problem takes into account census data, in particular those of the 9<sup>th</sup> general population census (1951). In that occasion, among the others, information about literacy, socio-economic status and houses, were collected at municipal level, and are still available.

Then, the indicator here proposed includes the following components:

- the illiteracy rate, calculated considering illiterate people with respect to the population aged more than 6;
- the percentage of active population (that is, aged more than 10) employed in agriculture;
- the mean number of people per house;
- the percentage of homeownershipper families;
- the rate of houses provided to toilet, water (aqueduct or well) and electrical wiring, considered as a proxy of quality houses

Obviously, the relationship between each component and the indicator is inverse for the first three and direct for the others. The preliminary step was the digitalization of data in a suitable format for the analysis. The following one consisted of a simple rank analysis carried out with respect the five components. To doing so, for each component the standardized distribution was considered. The pattern for all the Italian municipalities, shown in figure 1, is nearly symmetric only for the mean number of people per house.

**Figure 1:** Components of well-being indicator: distribution of Italian municipalities



A well-being municipal indicator using census data: first results



After standardization, to preserve the distances among municipalities, the ranking was normalized by attributing the score 1 to the first municipality and the score 1,000 to the last one. Finally, the indicator was obtained calculating the arithmetic mean of the five rankings.

The five components highlight a correlation generally positive, even if not marked (table 1). In particular, the highest figures concern house quality, whilst ownership seems to be weakly or negatively correlated with the others.

**Table 1:** Spearman's correlation coefficient for the components of the indicator

<i>Component</i>	<i>Illiteracy</i>	<i>Employed a.</i>	<i>People</i>	<i>Owners.</i>	<i>House q.</i>
Illiteracy	/	0.41	0.33	-0.02	0.70
Employed a.	0.41	/	0.15	-0.31	0.48
People p.h.	0.33	0.15	/	0.19	0.29
Ownership	-0.02	-0.31	0.19	/	-0.32
House q.	0.70	0.48	0.29	-0.32	/

## 2 First results

The results of this analysis are reported in table 2; the municipalities are grouped by quintiles and region to make easier the interpretation. As expected Lombard municipalities show high values of the indicator: about 40% of them belongs to the first quintile of national distribution. The other North-Western regions (Liguria, Piedmont and Aosta Valley) are also characterized by a considerable well-being level.

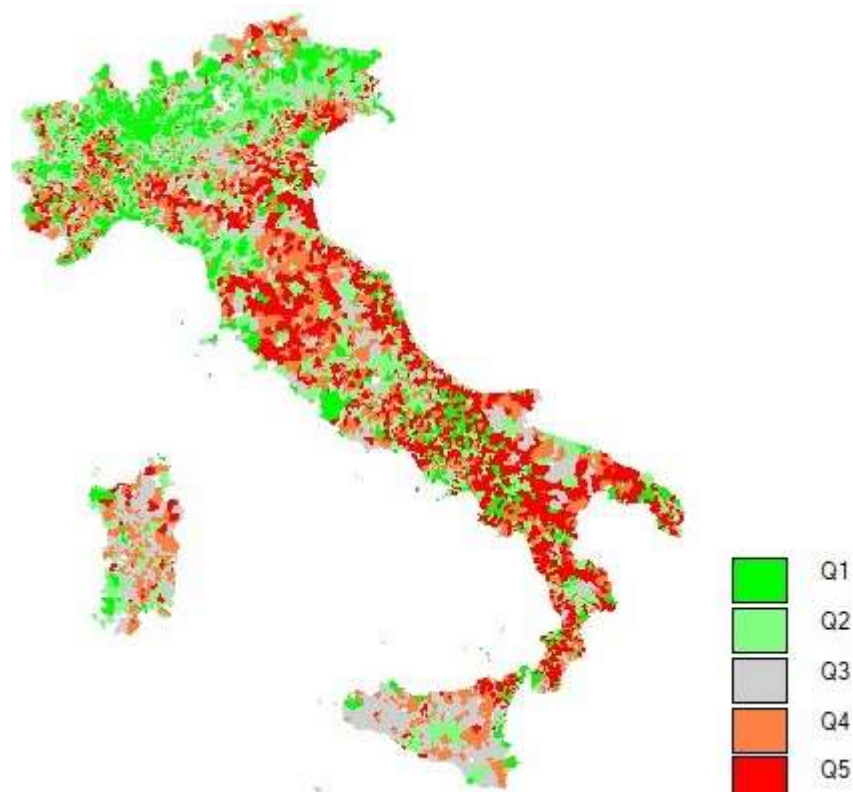
For the Southern regions, conversely, the indicator seems to trace a very different condition: the majority of their municipalities is positioned in the fourth and the fifth quintile (in Apulia, Basilicata and Calabria over 60%). Note that in some Central areas and in Emilia Romagna emerges also a low level of the indicator.

**Table 2:** Regional distribution (%) of the quintiles of the well-being indicator for the Italian municipalities

<i>Region</i>	<i>Q1</i>	<i>Q2.</i>	<i>Q3</i>	<i>Q4</i>	<i>Q5</i>
Aosta Valley	20.5	30.1	19.2	20.5	9.7
Piedmont	24.6	19.3	17.5	24.6	14.0
Lombardy	40.9	29.1	18.4	8.6	3.0
Bolzano	13.2	13.2	31.1	32.1	10.4
Trento	21.8	45.2	20.1	12.3	0.6
Venetum	11.8	21.1	23.9	23.2	20.0
Friuli V.G.	22.6	33.5	25.0	11.3	7.6
Liguria	36.0	19.9	14.7	19.5	9.9
Emilia Romagna	4.5	15.6	21.9	25.7	32.3
Tuscany	15.0	23.6	18.2	18.6	24.6
Marche	10.6	12.7	15.1	18.0	43.6
Umbria	2.2	11.0	20.9	34.0	31.9
Latium	8.7	12.3	18.3	29.8	30.9
Abruzzo-Molise	18.4	13.8	16.1	18.9	32.8
Campania	18.2	17.8	12.1	17.3	34.6
Apulia	10.9	12.1	15.8	19.1	42.1
Basilicata	11.9	13.5	4.8	11.1	58.7
Calabria	8.6	14.0	14.6	22.5	40.3
Sicily	5.1	12.2	43.2	25.9	13.6
Sardinia	3.3	11.4	38.6	37.4	9.3
ITALY	20.0	20.0	20.0	20.0	20.0

Finally, a quick graphical description is presented in figure 2, in which it is possible to confirm the previous considerations. The municipalities depicted in green belong to the first quintile, while those in red belong to the fifth quintile. The municipalities with high level are scattered in Northern Italy, as well as those with low level are located in the inner areas of Center and South or in particular rural areas (Polesine, Maremma, etc.).

**Figure 2:** Distribution of the well-being indicator for the Italian municipalities, 1951 census (quintiles)



Of course, some questions about the determinants of these differences are still unsolved, as well as some remarks should be kept in mind for further analyses. For example, municipalities are classified regardless of their demographic size and altitude. Well-being levels should include both subjective and objective determinants, but the latter cannot be easily detected especially for the past decades. Moreover, it is very likely that the role played by such determinants changes over time: in 1951 quality of household and literacy could affect well-being much more than they did in recent years.

Thus, the next step of this study will be devoted to improve the theoretical framework, to increase the number of variables to be included and to extend the analysis to the following censuses.

## References

1. Alaimo, L.S., Arcagni, A., Fattore, Maggino F., Quondamstefano, V.: Measuring Equitable and Sustainable Well-Being in Italian Regions: The Non-aggregative Approach, *Social Indicators Research* (2020), <https://doi.org/10.1007/s11205-020-02388-7>



2. Alaimo, L.S., Maggino, F.: Sustainable development goals indicators at territorial level: conceptual and methodological issues. *The Italian perspective*, *Social Indicators Research*, Vol. 147, n. 2 (2020): 383-419
3. Barrington-Leigh, C., Escande, A. *Measuring Progress and Well-Being: A Comparative Review of Indicators*. *Soc Indic Res* 135, 893–925 (2018). <https://doi.org/10.1007/s11205-016-1505-0>
4. Biggeri L., Laureti T., Secondi L.: Well-being and quality of life in Italy: Assessing and selecting indicators for local policy making. *Italian Journal of Applied Statistics* 24.2 (2014): 125-152.
5. Bonardo, D., Quondamstefano, V.: *Measuring Well-Being in Italian (Eco)regions*, *Social Indicators Research* (2020). <https://doi.org/10.1007/s11205-020-02315-w>
6. Burchi F., Gnesi C. (2016): A Review of the Literature on Well-Being in Italy: A Human Development Perspective, *Forum for Social Economics*, 45:2-3, 170-192, doi: 10.1080/07360932.2014.995197
7. Calcagnini, G., Perugini, F.: A Well-Being Indicator for the Italian Provinces. *Social Indicators Research* 142, 149–177 (2019). <https://doi.org/10.1007/s11205-018-1888-1>
8. Chelli F.M., Ciommi M., Emili A., Gigliarano C., Taralli S.: Assessing the Equitable and Sustainable Well-Being of the Italian Provinces, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. Special Issue on Aggregation in Welfare Economics, Vol. 24, Suppl. 1 (2016): 39-62
9. Ciommi M. et al.: A new class of composite indicators for measuring well-being at the local level: An application to the Equitable and Sustainable Well-being (BES) of the Italian Provinces. *Ecological indicators*, Vol. 76 (2017): 281-296.
10. Conigliaro, P.: Subjective Well-Being in Italian Regions, *Social Indicators Research* (2020). <https://doi.org/10.1007/s11205-020-02391-y>
11. Mazziotta M.: Socio-Economic Indicators for Measuring the Well-Being of Italian Municipalities, in *Scienze Regionali*, *Italian Journal of Regional Science Speciale/2019*, pp. 633-650, doi: 10.14650/94670

## 4.24 Prior distribution for Bayesian analysis

# On the dependence structure in Bayesian nonparametric priors

## *Sulla dipendenza nelle distribuzioni a priori Bayesiane e non parametriche*

Filippo Ascolani, Beatrice Franzolini, Antonio Lijoi, and Igor Prünster

**Abstract** Bayesian models for data grouped into distinct samples are typically defined within the framework of partial exchangeability. All currently known nonparametric priors for partially exchangeable data induce positive correlation both between observations coming from different samples as well as between the underlying random probability measures. However, such property is not implied by partial exchangeability and may not be appropriate in some applications. Using  $\sigma$ -stable completely random measures and Clayton-Lévy copulas, we propose a nonparametric prior that may induce either negative or positive correlation. The contents of these pages summarize some of the results derived in [1].

**Abstract** *La parziale scambiabilità è un'assunzione spesso utilizzata nei modelli Bayesiani per dati suddivisi in campioni. Tutte le distribuzioni non parametriche note per dati parzialmente scambiabili inducono correlazione positiva sia tra le osservazioni in diversi campioni, sia tra le misure di probabilità sottostanti. Tuttavia, la correlazione positiva non è implicata dalla parziale scambiabilità. In questo lavoro viene introdotta una distribuzione a priori nonparametrica che può indurre correlazione negativa o positiva e che fa uso delle misure completamente aleatorie  $\sigma$ -stabili e delle Clayton-Lévy copulas. Il contenuto di queste pagine riassume alcuni dei risultati derivati in [1].*

**Key words:** Bayesian nonparametrics, Completely random measure, Lévy copula, Negative correlation, Partial exchangeability

---

Filippo Ascolani  
Bocconi University and BIDSa, e-mail: [filippo.ascolani@phd.unibocconi.it](mailto:filippo.ascolani@phd.unibocconi.it)

Beatrice Franzolini  
Bocconi University and BIDSa, e-mail: [beatrice.franzolini@phd.unibocconi.it](mailto:beatrice.franzolini@phd.unibocconi.it)

Antonio Lijoi  
Bocconi University and BIDSa, e-mail: [antonio.lijoi@unibocconi.it](mailto:antonio.lijoi@unibocconi.it)

Igor Prünster  
Bocconi University and BIDSa, e-mail: [igor.pruenster@unibocconi.it](mailto:igor.pruenster@unibocconi.it)

## 1 Introduction

Traditional Bayesian models assume that data are exchangeable, which is a homogeneity condition implying the existence of a common underlying distribution from which observations have been sampled. More formally, a sequence of observations  $X = (X_i)_{i \geq 1}$  is said exchangeable if and only if  $\forall n \geq 1$ ,  $(X_1, \dots, X_n)$  is equal in distribution to  $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$  for any  $\sigma$  permutation of  $n$  elements. It should be clear that exchangeability is an appropriate assumption only when one would like to develop an inferential procedure which disregards any information that may be included in the order in which data were collected and stored.

However, this is not the case, for instance, when data are grouped into many samples corresponding to different experimental conditions or when discrete covariates information is available. In these situations a more plausible assumption is partial exchangeability. Two sequences of data  $X_1 = (X_{i,1})_{i \geq 1}$  and  $X_2 = (X_{i,2})_{i \geq 1}$ , where  $X_{j,i}$  is a random variable taking value in a Polish space  $(\mathbb{X}, \mathcal{X})$ , are said partially exchangeable if and only if for all  $n_1 \geq 1$  and  $n_2 \geq 1$ :

$$(X_{1,1}, \dots, X_{n_1,1}, X_{1,2}, \dots, X_{n_2,2}) \stackrel{d}{=} (X_{\sigma_1(1),1}, \dots, X_{\sigma_1(n_1),1}, X_{\sigma_2(1),2}, \dots, X_{\sigma_2(n_2),2})$$

for any  $\sigma_1$  and  $\sigma_2$  permutations of respectively  $n_1$  and  $n_2$  elements. Thanks to de Finetti's representation theorem for partial exchangeability [3], we know that  $X_1$  and  $X_2$  are partial exchangeable if and only if there exist two (possibly dependent) random probability measure  $\tilde{p}_1$  and  $\tilde{p}_2$  such that:

$$X_{i,j} \mid (\tilde{p}_1, \tilde{p}_2) \stackrel{ind}{\sim} \tilde{p}_j \quad \text{for } j = 1, 2 \quad (\tilde{p}_1, \tilde{p}_2) \sim Q$$

and  $Q$  plays the role of the prior.

In the last two decades there has been a growing interest in developing nonparametric priors for partially exchangeable data. See [5, 11] and references therein.

However, all existing and used nonparametric priors induce a non-negative correlation both between  $\tilde{p}_1(A)$  and  $\tilde{p}_2(A)$ , for every  $A \in \mathcal{X}$ , and between  $X_{i,1}$  and  $X_{i',2}$  for any  $i, i'$ . Such property is not implied by partial exchangeability and does not fit those applications where one has a priori information regarding negative correlation between observable in different groups.

In this work, after some preliminaries regarding completely random measures (Section 2), we introduce a novel nonparametric prior (Section 3) over  $(\tilde{p}_1, \tilde{p}_2)$  that may induce either negative or positive correlation between the observables. Lastly (Section 4), we develop an algorithm for sampling from the proposed prior and use it to show the conditional behaviour of  $\tilde{p}_2$  given  $\tilde{p}_1$ . The focus of this work is the prior law of  $(\tilde{p}_1, \tilde{p}_2)$ . For what concerns posterior inference, a comment can be found at the end of Section 3, while further details will be provided in forthcoming works.

## 2 Preliminaries on completely random measures

Consider a Polish space  $(\mathbb{X}, \mathcal{X})$  endowed with its Borel  $\sigma$ -algebra and  $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$  the space of boundedly finite measures on  $\mathbb{X}$ .

**Definition 1.** Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , a random element  $\tilde{\mu}$  from  $(\Omega, \mathcal{F}, \mathbb{P})$  into  $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$  is a completely random measure (CRM) on  $(\mathbb{X}, \mathcal{X})$  if, for every collection of pairwise disjoint sets  $(A_i)_{i \geq 1}^n$  in  $\mathcal{X}$ , the random variables  $\tilde{\mu}(A_1), \tilde{\mu}(A_2), \dots, \tilde{\mu}(A_n)$  are mutually independent.

If  $\tilde{\mu}$  is a CRM without deterministic component and fixed points of discontinuity, then  $\tilde{\mu}$  is almost surely discrete, i.e.

$$\tilde{\mu} \stackrel{a.s.}{=} \sum_{j=1}^{\infty} J_j \delta_{X_j}$$

and  $\tilde{\mu}$  is characterized by the following Laplace functional transform. For any measurable positive-valued function  $f$ ,

$$\mathbb{E} \left[ e^{-\int_{\mathbb{X}} f(x) \tilde{\mu}(dx)} \right] = \exp \left\{ - \int_{\mathbb{R}^+ \times \mathbb{X}} [1 - e^{-sf(x)}] \tilde{\nu}(ds, dx) \right\}$$

where  $\tilde{\nu}$  is called Lévy intensity and uniquely identifies the law of  $\tilde{\mu}$ . Finally, we assume that the jumps  $(J_j)_{j \geq 1}$  and the locations  $(X_j)_{j \geq 1}$  are independent, so that  $\nu(ds, dx) = \rho(s) ds \alpha(dx)$ . For more details on CRM, we refer to [8, 9]. CRMs have been proven a useful tool for prior specification. In particular, they may be normalized to obtain random probability measures, called normalized random measures with independent increments (NRMI), introduced in [12]. The notion of CRM can be extended to a vector of measures as follows:

**Definition 2.** Let  $\underline{\mu} = (\tilde{\mu}_1, \tilde{\mu}_2)$  be a vector of CRMs on  $\mathbb{X}$ . We say that  $\underline{\mu}$  is a completely random vector (CRV) on  $(\mathbb{X}, \mathcal{X})$  if, for every collection of pairwise disjoint sets  $(A_i)_{i \geq 1}^n$  in  $\mathcal{X}$ , the random vectors  $(\tilde{\mu}_1(A_1), \tilde{\mu}_2(A_1)), \dots, (\tilde{\mu}_1(A_n), \tilde{\mu}_2(A_n))$  are mutually independent.

The Laplace functional transform of  $\underline{\mu}$  is

$$\mathbb{E} \left[ e^{-\int_{\mathbb{X}} f_1(x) \tilde{\mu}_1(dx) - \int_{\mathbb{X}} f_2(x) \tilde{\mu}_2(dx)} \right] = \exp \left\{ - \int_{(\mathbb{R}^+)^2 \times \mathbb{X}} (1 - e^{-s_1 f_1(x) - s_2 f_2(x)}) \nu(ds_1, ds_2, dx) \right\}$$

for measurable  $f_1, f_2 : \mathbb{X} \rightarrow \mathbb{R}^+$ , where  $\nu$  is called joint Lévy intensity and uniquely identifies the law of  $(\tilde{\mu}_1, \tilde{\mu}_2)$ .

### 3 Atom-dependent $\sigma$ -stable normalized completely random measures

**Definition 3.** Let  $\xi = (\xi_1, \xi_2)$  be a CRV on  $(\mathbb{X} \times \mathbb{X}, \mathcal{X} \otimes \mathcal{X})$  with Lévy intensity  $\nu(ds_1, ds_2, dx_1, dx_2) = \rho(s_1, s_2) ds_1 ds_2 \alpha(dx_1, dx_2)$  such that

$$\int_0^{+\infty} \rho(s_1, s) ds_1 = \int_0^{+\infty} \rho(s, s_2) ds_2 = \frac{\sigma}{\Gamma(1-\sigma)} s^{-1-\sigma} ds, \quad 0 < \sigma < 1.$$

Then  $\tilde{\mu}_1(\cdot) = \xi_1(\cdot \times \mathbb{X})$  and  $\tilde{\mu}_2(\cdot) = \xi_2(\mathbb{X} \times \cdot)$  are called atom-dependent  $\sigma$ -stable CRMs with underlying joint Lévy intensity  $\nu$ .

**Proposition 1.** Consider  $\tilde{\mu}_1$  and  $\tilde{\mu}_2$  atom-dependent  $\sigma$ -stable CRMs, as defined in Definition 3, then  $\tilde{\mu}_j$  is a  $\sigma$ -stable CRM, for  $j = 1, 2$  and the a.s. discrete representation of  $\tilde{\mu}_1$  and  $\tilde{\mu}_2$  is:

$$\tilde{\mu}_1 \stackrel{a.s.}{=} \sum_{k \geq 1} W_{1,k} \delta_{(\theta_{1,k})} \quad \tilde{\mu}_2 \stackrel{a.s.}{=} \sum_{k \geq 1} W_{2,k} \delta_{(\theta_{2,k})}$$

where the two sequences of weights  $(W_{1,k})_{k \geq 1}$  and  $(W_{2,k})_{k \geq 1}$  are inherited from the underlying measures  $\xi_1$  and  $\xi_2$  and  $(\theta_{1,k}, \theta_{2,k}) \stackrel{iid}{\sim} G_0 \equiv \alpha/\alpha(\mathbb{X})$ .

**Definition 4.** The random probability measures  $\tilde{p}_1$  and  $\tilde{p}_2$  obtained normalizing two atom-dependent  $\sigma$ -stable CRMs  $\tilde{\mu}_1$  and  $\tilde{\mu}_2$  with underlying joint Lévy intensity  $\nu$ :

$$\tilde{p}_1(\cdot) = \frac{\tilde{\mu}_1(\cdot)}{\tilde{\mu}_1(\mathbb{X})} \quad \tilde{p}_2(\cdot) = \frac{\tilde{\mu}_2(\cdot)}{\tilde{\mu}_2(\mathbb{X})}$$

are called atom-dependent  $\sigma$ -stable NRMI.

In order to obtain a working model which makes use of atom-dependent  $\sigma$ -stable NRMI, the underlying joint Lévy intensity  $\nu$  has to be specified. A useful strategy to serve the purpose is to use Lévy copulas. See [2, 7, 10]. A popular Lévy copula is the Clayton’s one, which is given by the following expression:

$$C_\theta(x_1, x_2) = \{x_1^{-\theta} + x_2^{-\theta}\}^{-1/\theta}$$

The attractive feature of Clayton’s copula is that it depends only on one parameter,  $\theta$ , that fully characterizes the degree of dependence between the resulting CRMs  $\xi_1$  and  $\xi_2$ . As consequence, when Clayton’s copula is used to specify the law of two atom-dependent NRMI,  $\theta$  controls the portion of dependence between  $\tilde{p}_1$  and  $\tilde{p}_2$  induced by the joint distribution of the weights. In particular when  $\theta \rightarrow 0$  independence between  $\tilde{p}_1$  and  $\tilde{p}_2$  is approached, while the case of  $\theta \rightarrow +\infty$  corresponds to maximal dependence induced by the weights, i.e. the two sequences of weights are equal with probability 1. Applying Clayton’s Lévy copula to marginal Lévy  $\sigma$ -stables, one gets (see [4]):

On the dependence structure in Bayesian nonparametric priors

$$v(ds_1, ds_2, dx_1, dx_2; \theta) = \frac{(1 + \theta) \sigma (s_1 s_2)^{\sigma\theta - 1}}{\Gamma(1 - \sigma) (s_1^{\sigma\theta} + s_2^{\sigma\theta})^{\frac{1}{\theta} + 2}} \alpha(dx_1, dx_2) \quad (1)$$

**Theorem 1.** Consider the sampling model  $X_{i,j} \mid \tilde{p}_j \stackrel{\text{ind}}{\sim} \tilde{p}_j$  for  $j = 1, 2$  and  $i = 1, \dots, n_j$ , where  $\tilde{p}_1$  and  $\tilde{p}_2$  are atom-dependent  $\sigma$ -stable NRMIs with underlying joint Lévy intensity (1), then:

$$\text{Corr}(X_{i,1}, X_{i',2}) = g(\theta) \rho$$

where  $g: \mathbb{R}^+ \rightarrow (0, (1 - \sigma))$  and  $\rho$  is the correlation between two random variables jointly sampled from  $G_0$ .

Therefore, for appropriate choices of  $G_0$ , and in particular of  $\rho$ , the correlation between observations in different samples can be negative.

Lastly, concerning the possibility of deriving posterior inference, it is important to note that the representation of  $(\tilde{\mu}_1, \tilde{\mu}_2)$  in terms of the CRV  $(\xi_1, \xi_2)$  is crucial. Indeed, it allows to obtain posterior representation theorems generalizing the results provided in [6] for the exchangeable case.

## 4 Prior algorithm and simulations

We conclude this work with a simulation study, which shows the flexibility of the nonparametric prior introduced in the previous section when  $\alpha(dx_1, dx_2)$  is a multivariate Gaussian probability measure with zero means, unit variances and correlation  $\rho$ . To this end we need an algorithm to sample the infinite dimensional parameters  $\tilde{p}_1$  and  $\tilde{p}_2$  for different values of the hyperparameters  $\theta$  and  $\rho$ . Algorithm 1 serves the purpose and it has been obtained adapting the Algorithm 6.15 in [2] to the atom-dependent structure. We first sample a realization for  $\tilde{p}_1$  and then sim-

---

### Algorithm 1: Prior Sampler

---

```

for  $k \leftarrow 0$  to  $K$  do
  Sample  $T_k$  from an Exponential(1);
  Compute  $S_k^{(1)} = S_{k-1}^{(1)} + T_k$ ;
  Sample  $U_k$  from an Uniform(0, 1);
  Compute  $S_k^{(2)} = S_k^{(1)} \left( U_k^{-\theta/(1+\theta)} - 1 \right)^{-1/\theta}$ ;
  Compute  $W_{j,k} = (S_k^{(j)})^\sigma \Gamma(1 - \sigma)^{-1/\sigma}$  for  $j = 1, 2$ ;
  Sample  $(\theta_{1,k}, \theta_{2,k})$  from  $G_0$ ;

```

**end**

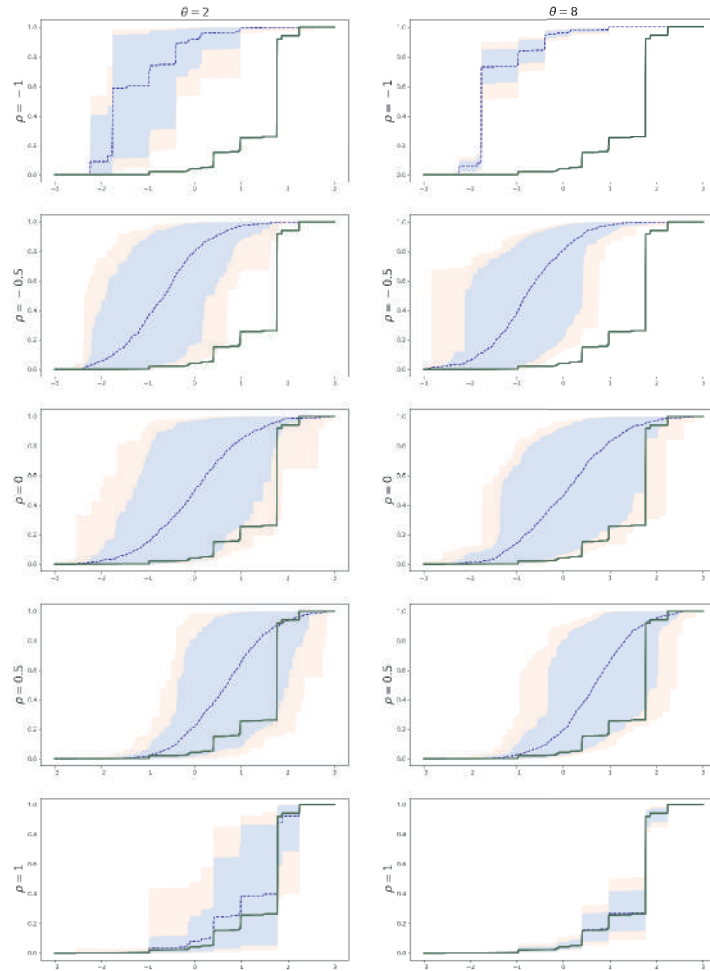
Compute  $\bar{W}_{j,k} = W_{j,k} / \sum_{k=1}^K W_{j,k}$  for  $j = 1, 2$  and  $k = 1, \dots, K$ ;

Obtain  $\tilde{p}_1 \approx \sum_{k=1}^K \bar{W}_{1,k} \delta_{\theta_{1,k}}$  and  $\tilde{p}_2 \approx \sum_{k=1}^K \bar{W}_{2,k} \delta_{\theta_{2,k}}$

---

ulate the conditional distribution of  $\tilde{p}_2$ , given  $\tilde{p}_1$ , under different hyperparameters choices. Figure 1 shows the results in terms of cumulative distribution functions. The plots in the first and second row ( $\rho = -1$  and  $\rho = -0.5$ ) show a strong and

mild negative correlation between the observables, represented by the opposite behaviour of  $\tilde{p}_2$  and  $\tilde{p}_1$ . While  $\tilde{p}_1$  associates high probabilities to positive values,  $\tilde{p}_2$  tends to associate high probabilities to negative values. While  $\rho$  increases, first the conditional distribution of  $\tilde{p}_2$  becomes independent from  $\tilde{p}_1$  ( $\rho = 0$ ) and then shows a behaviour similar to that of  $\tilde{p}_1$  ( $\rho = 0.5$  and  $\rho = 1$ ), corresponding to positive correlation of the observables.



**Fig. 1** Green solid line: a realization of the cumulative distribution function (cdf) corresponding to  $\tilde{p}_1$ , i.e.  $\int_{-\infty}^x \tilde{p}_1(dx)$ . Blue dashed lines: conditional expected value of the cdf corresponding to  $\tilde{p}_2$ , given the realization of  $\tilde{p}_1$ , i.e.  $\mathbb{E}[\int_{-\infty}^x \tilde{p}_2(dx) | \tilde{p}_1]$ . Light blue shaded area: 95% pointwise credible interval for the cdf corresponding to  $\tilde{p}_2$ . Pink shaded area: 99% pointwise credible interval for the cdf corresponding to  $\tilde{p}_2$ .



## References

1. Ascolani, F., Franzolini, B., Lijoi, A., Prünster, I.: Dependent nonparametric models inducing negative correlation. Technical report. (2021)
2. Cont, R., Tankov, P.: Financial modelling with jump processes. Chapman & Hall/CRC, Boca Raton, FL (2004)
3. de Finetti, B.: Sur la condition d'équivalence partielle. *Actual. Sci. Ind.* **739**, 5–18 (1938)
4. Epifani, I., Lijoi, A.: Nonparametric priors for vectors of survival functions. *Stat. Sin.* **20**, 1455–1484 (2010)
5. Foti, N.J., Williamson, S.A.: A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 359–371 (2013)
6. James, L.F., Lijoi, A., Prünster, I.: Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36**, 76–97 (2009)
7. Kallsen, J., Tankov, P.: Characterization of dependence of multidimensional Lévy processes using Lévy copulas. *J. Multivar. Anal.* **97**, 1551–1572 (2006)
8. Kingman, J.F.: Completely random measures. *Pac. J. Math.* **21**, 59–78 (1967)
9. Kingman, J.F.: Poisson Processes. Oxford Studies in Probability. Oxford University Press, Oxford (1993)
10. Palacio, A. R., Leisen, F.: Bayesian nonparametric estimation of survival functions with multiple-samples information. *Electron. J. Stat.*, **12**, 1330–1357 (2018)
11. Quintana, F. A., Mueller, P., Jara, A., MacEachern, S. N.: The dependent Dirichlet process and related models. arXiv preprint arXiv:2007.06129. (2020)
12. Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. *Ann. Stat.* **31**, 560–585 (2003)

# Anisotropic determinantal point processes and their application in Bayesian mixtures

## *Processi di punto anisotropici di tipo determinantal e loro applicazione nelle misture bayesiane*

Lorenzo Ghilotti, Mario Beraha and Alessandra Guglielmi

**Abstract** Repulsive mixture models have recently gained visibility in Bayesian statistics. In such models, a finite repulsive point process is assumed as prior distribution for the number of components and component-specific parameters. We assume a determinantal point process as such prior, proposing a simple construction of anisotropic determinantal point processes, that can better characterize repulsion when data have different scales along the axes. In turn, this produces better cluster estimates. We discuss the model on simulated data.

**Abstract** *I modelli mistura repulsivi hanno avuto di recente un incremento di visibilità in statistica bayesiana. In tali modelli, si assume un processo di punto finito repulsivo come prior sul numero di componenti e sui parametri specifici di ogni componente. In particolare, noi assumiamo un processo di punto di tipo determinantal, proponendo una semplice costruzione per processi di punto di tipo determinantal anisotropi, che possano meglio caratterizzare la repulsione quando i dati hanno dispersioni differenti lungo gli assi. Di conseguenza, modelli di questo tipo producono stime dei clusters migliori. Discutiamo il nostro modello su dati simulati.*

**Key words:** repulsive mixture models, determinantal point processes, anisotropic covariance function, spectral density

## 1 Introduction

Mixture models are a popular framework in Bayesian inference, providing useful tools for density estimation problems and cluster detection; see [4].

---

Lorenzo Ghilotti<sup>1</sup>, Mario Beraha<sup>1,2</sup> and Alessandra Guglielmi<sup>1</sup>

<sup>1</sup>Department of Mathematics, Politecnico di Milano, Milano, Italy

<sup>2</sup>Department of Computer Science, Università degli Studi di Bologna, Bologna, Italy

e-mail: lorenzo1.ghilotti@mail.polimi.it, {mario.beraha, alessandra.guglielmi}@polimi.it

Mixture models assume that data arise from one of  $M$  homogeneous populations, each suitably modelled by a density  $\{f_m\}_{m=1}^M$ , henceforth denoted as *component*. A set of nonnegative weights specifies the probability of each population to be selected. In the Bayesian setting, a prior is assumed on the weights, on the parameters governing the densities  $f_m$  and possibly on  $M$ . The most common formulation assumes that the parameters of the components are a priori independent and identically distributed, because of mathematical tractability, but, specifically for clustering purposes, it often reveals to be an oversimplification. As shown in [3], if the mixture model is misspecified, assuming component-specific parameters iid leads to overestimating the number of components, so that inference may produce redundant clusters of the data.

*Repulsive mixture models* use the notion of repulsion between cluster-specific parameters specifying prior that encourages well separated components, see for instance [2, 5] and the references therein. In particular, [1] proposes a general framework for this family of models, by assuming a *repulsive point process* as joint prior distribution for the location centers and  $M$ . Within the spectrum of repulsive point processes, *determinantal point processes* (DPPs) (see [6]) are rather appealing since they do not carry intractable normalizing constants and are defined through a covariance function. Often DPPs in the literature assume *stationary* and *isotropic* covariance functions, but this might be a modelling limitation.

In this work, we propose a simple construction for anisotropic DPPs, that preserves the analytical tractability of isotropic DPPs. The structure of the paper is as follows. Section 2 covers background material on (determinantal) point processes, while in Section 3 we introduce our anisotropic DPP. In Section 4 we assume this DPP as a joint prior for location parameters and the number of components in a Bayesian mixture model. We show the advantages of introducing anisotropism in a simulation study in Section 5.

## 2 Background on Determinantal Point Processes

Let  $R \subseteq \mathbb{R}^d$  be a compact set, a *finite point process*  $X$  on  $R$  is a finite random subset of  $R$ . Several choices are available to characterize  $X$ . For instance, we may assign the *product density functions*  $\rho^{(n)} : R^n \rightarrow [0, \infty)$ ,  $n = 1, 2, \dots$ ; see [7]. Intuitively, for any pairwise distinct points  $x_1, \dots, x_n$  in  $R$ ,  $\rho^{(n)}(x_1, \dots, x_n) dx_1 \cdots dx_n$  represents the probability that  $X$  has a point in an infinitesimal small region around  $x_i$  of volume  $dx_i$ , for each  $i = 1, \dots, n$ .

To define a DPP  $X$  on  $R$ , we consider a covariance function  $C : R \times R \rightarrow \mathbb{R}$  and define the product density functions  $\rho^{(n)}$  as

$$\rho^{(n)}(x_1, \dots, x_n) = \det\{[C](x_1, \dots, x_n)\}, \quad (x_1, \dots, x_n) \in R^n, \quad n = 1, 2, \dots$$

where  $[C](x_1, \dots, x_n)$  is the  $n \times n$  matrix with elements  $C(x_i, x_j)$  and *det* denotes the matrix determinant. Of course, some assumptions on  $C$  guarantee the DPP to exist.

Observe that  $\rho^{(n)}(x_1, \dots, x_n) = 0$  if  $x_i = x_j$ , for some  $i \neq j$ , since, in this case, the matrix  $[C](x_1, \dots, x_n)$  is not full rank. Consequently, if  $C$  is continuous,  $\rho^{(n)}(x_1, \dots, x_n) \rightarrow 0$  if  $x_i \rightarrow x_j$ , for some  $i \neq j$ . Hence, the probability of having two points in a given neighborhood tends to zero as the size of the neighborhood shrinks. Moreover,  $\rho^{(n)}(x_1, \dots, x_n) \leq \prod_{j=1}^n C(x_j, x_j)$ , and  $C(x, x)$  represents the *intensity function* of the process. Hence, DPPs are *repulsive* point processes: in fact, the joint probability of any points configuration is smaller than the case of independent point configurations. *Stationarity* is a common assumption for a point process, describing invariance under translations in  $\mathbb{R}^d$ . For DPPs, it is expressed by assuming  $C(x, y) = C_0(x - y)$ .

Under this assumption, conditions on the existence of the process are given as conditions on the spectral density  $\varphi = \mathcal{F}(C_0)$ , where  $\mathcal{F}$  indicates the Fourier transform. Additionally, if  $\varphi < 1$ , then the DPP has a density  $f$  with respect to the unit rate Poisson point process  $Y_1$  on  $R$ . That is, letting  $I(\cdot)$  denote the indicator function, it holds that

$$P(X \in F) = \mathbb{E}[I(Y_1 \in F)f(Y_1)]$$

for any collection of point patterns  $F$  contained in  $R$ .

Using the spectral density approach, [2] derived a Markov chain Monte Carlo (MCMC) sampling scheme based on split-merge reversible jump moves, while [1] proposed a Metropolis-within-Gibbs sampler based on spatial birth-death processes. In both these papers, the authors consider modeling directly  $\varphi$  and using the approximation of the density  $f$  proposed in [6] when  $R = [-1/2, 1/2]^d$ . In particular, [2, 1] work with an isotropic DPP on  $[-1/2, 1/2]^d$  and apply an affine transformation mapping the DPP onto the smallest rectangle containing all the data.

As shown in Section 4, isotropism might produce misleading results when such a process is adopted as a prior for Bayesian mixture modeling and a more complex (anisotropic) model should be preferred.

### 3 Anisotropic DPPs

Suppose to be modeling points  $x_1, \dots, x_M$  through a DPP. If we assume an isotropic DPP, this would result in  $C_0(x)$  of the form  $C_0(\|x\|)$  for  $\|\cdot\|$  the standard Euclidean norm. If the  $x_i$ 's represent spatial locations, then isotropy is likely to be a justifiable assumption. However, if the  $x_i$ 's represent more complex kinds of data, such as medical measurements on a patient, isotropy can be an oversimplification and more complex models could be more suitable. For instance, if  $x_i \in \mathbb{R}^2$ , one might want to model different scales along the two axes, i.e. having a DPP that considers close two points such as  $(x, 0)$ ,  $(x + d, 0)$  and distant two points such as  $(0, y)$ ,  $(0, y + d)$  for the same value of  $d$  (or viceversa). In this section, we show how this behavior can be achieved by constructing a stationary but anisotropic DPP.

Note that the kind of anisotropy we are interested in can be well represented by employing a different metric on  $\mathbb{R}^d$ . In particular, we consider a  $d \times d$  sym-

metric positive definite matrix  $\Lambda$  and define  $\|x\|_\Lambda^2 = x^T \Lambda x$ , for  $x \in \mathbb{R}^d$ . Through  $\Lambda$  it is possible to define several kind of anisotropic behaviors: for instance, if  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$ , we could well model data that have different scales along different axes, and, by considering a full-matrix  $\Lambda$  we could also have different scales along directions that are not parallel to the cardinal axes.

Hence, the problem is how to define a *valid* DPP with kernel  $C(x, y) = C_0(x - y)$  such that  $C_0(x) = C_0(\|x\|_\Lambda)$ . We were able to prove that, if  $B$  is a  $p \times d$  matrix with full rank, with  $p \geq d$ , and  $k > 0, \rho > 0$ , then the kernel

$$C_0(x) = \rho \exp\left(-\frac{\|Bx\|^2}{2k}\right), \quad x \in \mathbb{R}^d \tag{1}$$

defines a valid DPP. Moreover, the resulting DPP has a density with respect to the unit rate Poisson point process if  $\rho < \rho_{\max}$ , where  $\rho_{\max} = |B^T B|^{\frac{1}{2}} k^{-d/2} / (2\pi)^{d/2}$ . In this case, the Fourier transform  $\varphi = \mathcal{F}(C_0)$  has a closed form expression :

$$\varphi(x) = \rho \frac{(2\pi k)^{d/2}}{|B^T B|^{1/2}} \exp(-2\pi^2 k x^T (B^T B)^{-1} x), \quad x \in \mathbb{R}^d \tag{2}$$

Note that the expression of  $C_0$  in (1) recovers indeed the desired kind of anisotropy; since  $\|Bx\|^2 = x^T B^T B x$ , it is sufficient to let  $B = \Lambda^{1/2}$ . Moreover, since  $\rho$  controls the intensity of the DPP, one might want to fix the maximum admissible intensity  $\rho_{MAX}$  independently of  $B$ . By applying the change of variable  $c := |B^T B|^{1/2} k^{-d/2}$ , and substituting  $k = |B^T B|^{1/d} c^{-2/d}$ , we get a new parametrization of the covariance function

$$C_0(x) = \rho \cdot \exp\left(-c^{2/d} \|Bx\|^2 / (2|B^T B|^{1/d})\right) \quad \rho_{MAX} = c / (2\pi)^{d/2}. \tag{3}$$

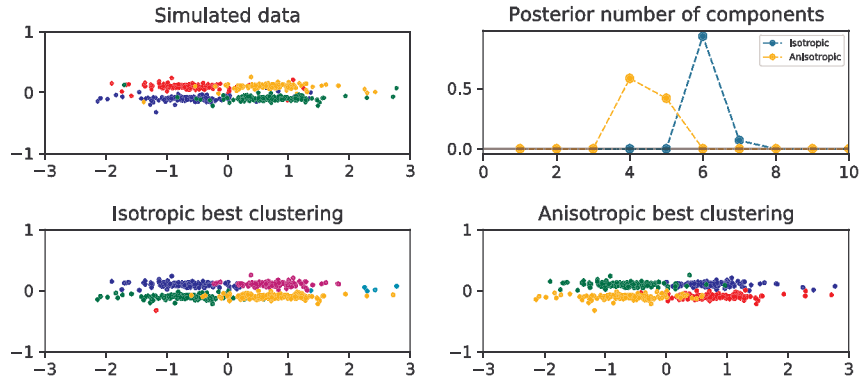
It is evident that, within this parametrization, parameter  $c$  tunes the maximum intensity allowed.

#### 4 Bayesian repulsive DPP mixture model

In this section, we introduce the Bayesian mixture model with an anisotropic DPP as a prior for the centers of the components. Let data  $y_1, \dots, y_n \in \mathbb{R}^d$ ; we assume

$$y_i | \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, M \stackrel{\text{iid}}{\sim} \sum_{h=1}^M w_h \mathcal{N}_d(\cdot | \mu_h, \boldsymbol{\Sigma}) \tag{4}$$

where  $\mathbf{w} = (w_1, \dots, w_M)$  are the weights,  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$  are the centers of the components,  $M$  denotes the total number of components, and  $\boldsymbol{\Sigma}$  is a covariance matrix that we assume to be known and fixed.



**Fig. 1** Simulated dataset (top left), posterior distribution of  $M$  (top right), cluster estimate under the isotropic (bottom left) and anisotropic (bottom right) prior.

**Prior assumptions.** The model is completed assuming the same hierarchical prior as in [1, 2]:

$$\{\mu_1, \dots, \mu_M; M\} \sim \text{DPP}(C_0; R), \quad w | M \sim \text{Dirichlet}_M(\alpha) \quad (5)$$

where  $\text{DPP}(C_0; R)$  denotes the distribution of a stationary determinantal point process on the compact set  $R \subset \mathbb{R}^q$  with kernel  $C_0$  and  $\text{Dirichlet}_M(\alpha)$  denotes the Dirichlet distribution on the  $M - 1$  dimensional simplex with parameters  $(\alpha, \dots, \alpha)$ . Assuming a DPP as a prior on the locations  $\mu_1, \dots, \mu_M$  induces repulsion between them, favoring well separated components, see [1]. Note that it also determines the distribution of the number of components  $M$ .

**Posterior inference.** We have designed a Metropolis-within-Gibbs MCMC algorithm to sample from the posterior distribution of  $(\mu, w, M)$  given  $y_1, \dots, y_n$ , as in [1]. The code has been implemented in C++. A central building block of the proposed MCMC scheme is the approximation of the DPP density as described in [6], by means of the Fourier transform in (2).

## 5 Simulation study

We present a simple simulated scenario to highlight the difference between the proposed anisotropic DPP prior and previously considered (isotropic) priors and the corresponding posterior inferences. We generated  $n = 600$  data from an equally weighted mixture of four bivariate Student-t distributions, with means  $m_1 = [-0.7, 0.1]$ ,  $m_2 = [-0.7, -0.1]$ ,  $m_3 = [0.7, 0.1]$ ,  $m_4 = [0.7, -0.1]$ , the same covariance matrix  $H = \text{diag}(0.1, 0.0005)$  and same degrees of freedom  $v = 3$ . Simulated data is shown in Figure 1, top left. Note that the dispersion is much more extreme along the horizontal axis than along the vertical one.

We consider two specifications for the DPP prior, fixing  $R = [-2, 2]^2$ . The first one (isotropic) assumes  $B$  in (3) to be the identity matrix, while the second one (anisotropic) assumes  $B = \text{diag}(1, 5)$ . We assume  $c = 6$ ,  $\rho = 0.9 \cdot \rho_{MAX}$  for both models and  $\alpha = 3$ ; see (5). Moreover, we assume the covariance  $\Sigma$  in (4) as  $\Sigma = \nu H$ . Note that the two models differ just on the *shape* of the repulsion: while the first assumes isotropism, the second induces a stronger repulsion along the horizontal axis and a weaker one along the vertical direction.

MCMC chains were run for 20,000 iterations, discarding the first 10,000 and keeping one iteration every 10, for a final sample size of 1,000. Figure 1(top right) shows the posterior distributions of  $M$  under the two priors. It is clear that the anisotropic DPP is more effective in recovering the true number of components. Moreover, Figure 1(bottom) shows cluster estimates obtained by minimizing the Binder loss function: the anisotropic DPP correctly recovers four clusters (bottom right), while the isotropic DPP (bottom left) estimates six of them.

## 6 Conclusion

In this paper, we have introduced a determinantal point process with anisotropism. Assuming this process as a prior in a Bayesian mixture model was shown to produce better cluster estimates in scenarios where data have different scales along different axes or directions.

The approach considered could be further extended to describe more complex models, such as an analogous of the Whittle-Matern DPP density in [6].

## References

1. Beraha, M., Argiento, R., Møller, J., Guglielmi, A.: MCMC computations for Bayesian mixture models using repulsive point processes. arXiv preprint arXiv:2011.06444 (2020)
2. Bianchini, I., Guglielmi, A., Quintana, F.A.: Determinantal point process mixtures via spectral density approach. *Bayesian Analysis* **15**, 187–214 (2020)
3. Cai, D., Campbell, T., Broderick, T.: Finite mixture models do not reliably learn the number of components (2020)
4. Frühwirth-Schnatter, S., Celeux, G., Robert, C.P.: Handbook of mixture analysis. CRC press (2019)
5. Fúquene, J., Steel, M., Rossell, D.: On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**(5), 809–837 (2019)
6. Lavancier, F., Møller, J., Rubak, E.: Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 853–877 (2015)
7. Møller, J., Waagepetersen, R.P.: Statistical Inference and Simulation for Spatial Point Processes. CRC press (2004)

# Bayesian Screening of Covariates in Linear Regression Models Using Correlation Thresholds

*Screening in ambito bayesiano delle covariate nei modelli di regressione lineare attraverso soglie di correlazione*

Ioannis Ntzoufras and Roberta Paroli

**Abstract** In this work, we propose a fast and simple Bayesian method based on simple and partial correlation coefficients to identify covariates which are not supported in terms of the Bayes Factors in normal linear regression models. By this way, when the number of the covariates is large, we can screen out the covariates with negligible effects and reduce the size of the model space in such a way that we can implement traditional Bayesian variable selection methods. We focus on the g-prior implementation where computations are exact but the approach is general and can be easily extended to any prior setup. The proposed method is illustrated using simulation studies.

**Abstract** *In questo lavoro si propone un metodo Bayesiano semplice e veloce, per identificare le covariate non significative, nei modelli lineare di regressione, basato sui coefficienti di correlazione semplice e parziale. Quando le covariate sono tante, esso consente di eliminare quelle poco importanti e ridurre lo spazio dei possibili modelli per poter quindi applicare i metodi tradizionali di selezione delle variabili restanti. Nel presente lavoro ci si focalizza sul caso delle g-prior, con le quali si possono ottenere risultati esatti, ma l'approccio é generale e può essere esteso a ogni altro tipo di distribuzione a priori. Il metodo viene illustrato tramite vari studi di simulazione.*

**Key words:** Bayes Factor, Bayesian Variable selection, g-prior, simple and partial correlation coefficient

---

Ioannis Ntzoufras

Department of Statistics, Athens University of Economics and Business, Greece, e-mail: ntzoufras@aueb.gr

Roberta Paroli

Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy, e-mail: roberta.paroli@unicatt.it



## 1 Introduction

Bayesian variable selection is an important problem which has been widely discussed both from a theoretical and a practical perspective (see [1]; [2]; [9] for literature review). Recent advances have been made in two directions, solving the theoretical properties of different choices of prior structure for the regression coefficients ([4] and [6]) and proposing algorithms that can explore huge models space consisting of all the possible subsets when a large number of covariates is considered, using either MCMC or other model search algorithms (see for example [5] and [3]).

In this work, we consider the Bayesian variable selection problem for normal linear regression models with Zellners  $g$ -priors [10]. Under our proposed approach, the relationships between the Bayes factor and the simple Pearson correlation coefficient (in the simple regression), or the partial correlation coefficient (in the multiple regression), are utilized in order to define proper correlation thresholds that the Bayes factor provides evidence against or in favour of the inclusion of a covariate in the final model. Hence, with this approach we can construct a fast screening algorithm based on simple and partial correlation measures based on purely Bayesian arguments.

## 2 Model and Motivation

### 2.1 Bayesian variable selection under the $g$ -prior

We consider the normal linear regression model with response vector  $\mathbf{y}$  and  $n \times p$  matrix  $\mathbf{X}$  of potential predictors. Let  $\gamma \in \{0, 1\}^p$  be the model indicator index of all  $2^p$  subsets of predictors in the full model  $\mathcal{M}$ ;  $\gamma_j$  be the index of whether covariate  $X_j$ , for  $j = 1, \dots, p$ , is included or not in model  $M_\gamma$  and  $p_\gamma = \sum_{j=1}^p \gamma_j$  denoting the number of active covariates in model  $M_\gamma$ .

Assuming that all models share the same error variance and an intercept is always included in the analysis, we are interested in comparisons between the reference null model  $M_0$  and all models  $M_\gamma \in \{\mathcal{M} \setminus M_0\}$ , i.e.

$$M_\gamma : \mathbf{y} \sim N(\beta_0 \mathbf{1} + \mathbf{X}_\gamma \beta_\gamma; \mathbf{I}_n \sigma^2) \quad \text{vs} \quad M_0 : \mathbf{y} \sim N(\beta_0; \mathbf{I}_n \sigma^2) \quad (1)$$

where  $\mathbf{X}_\gamma$  and  $\beta_\gamma$  are of dimension  $n \times p_\gamma$  and  $p_\gamma \times 1$ , respectively;  $\mathbf{1}$  is a  $n \times 1$  vector of ones and  $\mathbf{I}_n$  is the  $n \times n$  identity matrix.

The Bayesian model formulation is completed by specifying a prior distribution on the model parameters  $\pi(\beta_0, \beta_\gamma, \sigma^2 | M_\gamma)$ , and a prior mass  $Pr(M_\gamma)$  on model space.

One of the most popular choice of the prior specification for the regression coefficients is the Zellner's  $g$ -prior ([10] and [6]), due to its computational simplicity and its connection to the Bayes information criterion. In the original  $g$ -prior definition,

Zellner specified the following prior for the parameters of the normal regression model

$$\beta_\gamma^* | \gamma, \sigma^2 \sim \mathcal{N} \left( \mathbf{0}, g (\mathbf{X}_\gamma^{*T} \mathbf{X}_\gamma^*)^{-1} \sigma^2 \right) \text{ and } \pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad (2)$$

where  $\beta_\gamma^* = (\beta_0; \beta_\gamma)$  is the complete vector of regression coefficients, i.e. the vector including the intercept parameter;  $\mathbf{X}_\gamma^* = [\mathbf{1}; \mathbf{X}_\gamma]$  is the design matrix of dimension  $n \times (p_\gamma + 1)$  under model  $M_\gamma$ ; and  $\mathbf{0}$  is the vector of length  $n$  with all elements equal to zero.

Within the Bayesian framework, the formal approach for model or variable selection is based on the posterior model probabilities through the evaluation of Bayes factor ( $B_{\gamma,0}$ ) of model  $M_\gamma$  versus  $M_0$  which is defined as  $B_{\gamma,0}(\mathbf{y}) = \frac{m(\mathbf{y}|M_\gamma)}{m(\mathbf{y}|M_0)}$ , where  $m(\mathbf{y}|M)$  is the marginal likelihood of model  $M$ .

### 2.2 Importance of Covariate in Simple Regression

For the simple regression model, the expression of the Bayes factor can be derived (see [6]) as function of the Pearson correlation coefficient between  $Y$  and  $X_j$  (denoted by  $\rho_j$ ),  $g$  and  $n$ . If we examine the behaviour of the Bayes factor for varying values of  $g$ , different values of  $\rho_j$  and for fixed sample size  $n$ , we can easily conclude that, as expected, stronger evidence against the null model is obtained as the Pearson correlation increases for each given sample size  $n$  and value of  $g$ . So, for each given sample size, we can easily identify some thresholds for the simple correlation coefficient which correspond to covariates whose inclusion will be never supported by the Bayes Factor for all possible values of prior parameter value of  $g$ .

This analysis provides us the motivation to setup a fast screening covariate method which considerably reduces the model space.

The method is based on the concept of *non-important set of correlation coefficients* introduced by [7], which is defined as the set of correlations  $\rho_j$  which corresponds to covariates with Bayes factor less than  $\theta$ , for all possible values of  $g$ :

$$\{\rho_j : BF \leq \theta, \forall g \geq 0\} \quad j = 1, \dots, p.$$

From the explicit expression of  $BF$  (see [6]), we can obtain the expression of the threshold  $\rho_{j(\theta)}^2$  as a function of  $g$  and  $\theta$ :

$$\rho_{j(\theta)}^2 = \frac{g+1}{g} \left\{ 1 - [\theta^2(1+g)]^{-1/(n-1)} \right\} \quad j = 1, \dots, p. \quad (3)$$

So we can identify the non-important covariates by simply comparing the simple correlation coefficient of each covariate with the corresponding threshold.

### 2.3 Importance of Covariate in Multiple Regression

We can extend the previous formulas to the more realistic case of the multiple regression. Now we can express the Bayes factor in terms of the partial correlation coefficient and identify some thresholds of Bayesian significance for which the Bayes Factor takes a particular value. From the explicit expression of the multiple regression  $BF$  (see again [6]), we can obtain that the threshold  $\rho_{j,\gamma(\theta)}^2$  is given by:

$$\rho_{j,\gamma(\theta)}^2 = 1 - \frac{g(1 - R_\gamma^2)}{\theta^{2/(n-1)}(1+g)^{1/(n-1)}(1+g(1 - R_\gamma^2)) - 1} \quad (4)$$

with  $R_\gamma^2$  the usual coefficient of determination of the full model.

For a given value of  $\theta$ ,  $\rho_{j,\gamma(\theta)}^2$  are upper bounds for partial correlations that correspond to covariates with Bayes factor lower than  $\theta$ . So we can implement a fast Bayesian screening of the covariates by using only the information of their partial correlation coefficients.

Analogous thresholds can be derived by using the Bayes Factor under the unit-information approach ( $g = n$ ) or for the empirical Bayes approach, as described in [6].

Our proposed method is performed sequentially in three steps:

1. Calculate the  $p$  simple correlation coefficients and compare them with the simple regression thresholds (3), for all  $g$ . Remove the corresponding covariates from the model space, so the number of the active covariates becomes  $p^* < p$ ;
2. compute the  $p^*$  partial correlation coefficients for the model with the remaining  $p^*$  covariates. Compare them with the corresponding multiple regression thresholds (4). Decide which covariates should be screened out as non-important.
3. Remove from further analysis covariates that are always non-important for both 1) and 2). Implement any other Bayesian variable selection method in the remaining space.

This method is very efficient for  $p \leq n$ ; if  $p \geq n$  the procedure exhibits further complications and need special treatment which is not described here (work in progress by the authors).

### 3 An illustrative example

As an example we report here the result of a simulation where the simulation design of [8] is used. It consists of  $p = 15$  covariates and  $n = 50$  observations. The first 10 covariates are independent standard Normal while the last five are generated as a linear combination of the first five plus a standard normal error. Under this scheme the covariates are variously correlated: the last five are highly correlated, whereas they are moderately correlated with the first five covariates. The response variable  $Y$

is generated by the model  $Y = 2X_1 - X_5 + 1.5X_7 + X_{11} + 0.5X_{13} + \varepsilon$ , where  $\varepsilon$  is a Normal distribution with zero mean and variance equal to  $2.5^2$ . We ran 1000 simulations and repeat the two steps on 1000 datasets, counting the number of times where the simple or partial correlation coefficients are less than the corresponding thresholds. These results, in percentages, are shown in Table 1 for the unit information approach: it can be seen that the percentages of runs that in step 1 indicate that the covariates  $X_2, X_3, X_5, X_6$  and  $X_8, X_9, X_{10}$  belong to the non-importance area are very high, while in step 2 also the covariates  $X_{12}, X_{13}, X_{14}, X_{15}$  have percentages very high. So, after the two steps of our proposed algorithm, covariates  $X_2, X_3, X_5, X_6, X_8, X_9, X_{10}$  and  $X_{12}, X_{13}, X_{14}, X_{15}$  have a very high percentages to belong to the non-importance area and not to be included in the final model. For these 1000 simulated

**Table 1** Percentage of times that covariates belonging to the non-important set out of 1000 simulations by the Unit information approach; first line corresponds to step 1, second line to step 2.

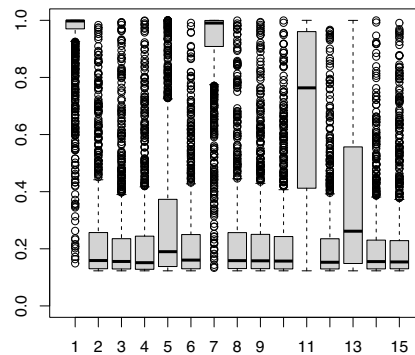
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$	
step 1		2	78	66	36	84	98	30	92	95	95	1	11	6	8	10
step 2		5	89	90	89	82	91	5	92	94	88	19	93	69	92	90
true effects	✓					✓		✓				✓			✓	

datasets we have computed also the posterior inclusion probabilities (PIP) for all the covariates. In Figure 1 the box-plots of these probabilities are illustrated: the covariates  $X_1, X_7, X_{11}$  are indicated as the important ones, as expected, and the values of their median PIPs are higher than 0.5. It can be observed that the variability of  $X_{11}$  is the highest. All covariates which indicate as non-important in our methods have median PIPs equal to zero. Covariates  $X_5$  and  $X_{13}$ , despite being in the true model, have small median PIP: in fact  $X_5$  is the covariate removed by our strategy while  $X_{13}$  have high correlation with  $X_{11}$ , so our method selects only the ones highly correlated with  $Y$ .

#### 4 Conclusion and Discussion

Our method allows to perform a first fast Bayesian screening of the non-important covariates in multiple regression model simply by comparing the partial correlation coefficients with some thresholds based on the Bayes factors. The Bayes factors are not have calculated so the computational cost is considerably reduced compared with other methods. In fact we have only to compare  $p$ , or less (in the second step), quantities while full Bayesian analysis will require the evaluation of  $2^p$  Bayes factors.

Finally, this method can be easily expanded to include more complicated terms such as polynomial terms or multiplicative (i.e. interaction) terms. The approach can be exactly the same, and such terms can be treated as extra covariates that can be evaluated by the corresponding correlations measures. Nevertheless, some



**Fig. 1** Boxplots of the PIP of the covariates over the 1000 generated datasets computed under the unit information BF.

deeper analysis might need to interpret the use of correlation or to adjust appropriately the prior distribution. Finally, even more difficult can be the problem with the use of dummy variables when we embody categorical factors in our analysis. In such approach, we might need to use alternative measures for fast screening since the Pearson or partial correlation is not appropriate in such case.

## References

1. Chipman, H., George, E.I., McCulloch, R.E., Clyde, M., Foster, D.P., Stine, R.A.: The Practical Implementation of Bayesian Model Selection. *Lecture Notes-Monograph Series of Institute of Mathematical Statistics* **38** 65–134 (2001)
2. Clyde, M., George, E.I.: Model uncertainty. *Statistical Science* **19** 81–94 (2004)
3. Dellaportas, P., Forster, J.J., Ntzoufras, I.: On Bayesian model and variable selection using MCMC. *Statistics and Computing* **12**, 27–36 (2002)
4. Fernandez, C., Ley, E., Steel, M.F.J.: Benchmark priors for Bayesian model averaging. *J. of Econometrics* **100**, 381–427 (2001)
5. Kohn, R., Smith, M., Chan, D.: Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* **11**, 313–322 (2001)
6. Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O.: Mixtures of g Priors for Bayesian Variable Selection. *J. of the Am. Stat. Ass.* **103**, 410–423 (2008)
7. Lykou, A., Ntzoufras, I.: On Bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing* **23**, 361–390 (2013)
8. Nott, D.J., Kohn, R.: Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–763 (2005)
9. O’Hara, R.B., Sillanpää, M.J.: A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis* **4**, 85–117 (2009)
10. Zellner, A.: On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision techniques* (1986)

## 4.25 Recent advances in clustering methods

# Biclustering longitudinal trajectories through a model-based approach

## *Classificazione a due vie di traiettorie longitudinali attraverso un approccio basato su modello*

Francesca Martella, Marco Alfó, Maria Francesca Marino

**Abstract** This work introduces a model-based biclustering approach for discrete multivariate longitudinal data. The proposed model considers a finite mixture of generalized linear models to cluster units and, within each mixture component, a flexible and parsimonious parameterization of the corresponding canonical parameter to cluster variables evolving in a similar manner across time. Model parameter estimates are obtained through an Expectation Maximization (EM) type algorithm and the performance of the proposed model are shown on both simulated and real dataset.

**Abstract** *In questo lavoro viene introdotto un approccio di classificazione a due vie per dati longitudinali multivariati discreti. Il modello proposto considera una mistura finita di modelli lineari generalizzati per classificare le unità e, all'interno di ogni componente della mistura, una parametrizzazione flessibile e parsimoniosa del parametro canonico corrispondente per classificare le variabili che evolvono in modo simile nel tempo. Le stime dei parametri del modello sono ottenute tramite un algoritmo di tipo Expectation Maximization (EM) e la performance del modello proposto è mostrata sia su dati simulati che reali.*

**Key words:** biclustering, discrete longitudinal data, finite mixture of generalized linear models

---

Francesca Martella, Marco Alfó  
Dipartimento di Scienze Statistiche, Sapienza Università di Roma  
e-mail: francesca.martella@uniroma1.it  
e-mail: marco.alfó@uniroma1.it

Maria Francesca Marino  
Dipartimento di Statistica, Informatica, Applicazioni G. Parenti, Università degli Studi Firenze  
e-mail: mariafrancesca.marino@unifi.it

## 1 Introduction

Biclustering technique dates back to the 1970s when the work of [3] appeared. It consists of an extension of the standard clustering approach aiming at jointly partitioning the set of units and the set of variables of a data matrix into homogeneous blocks, denominated biclusters. During the past decades biclustering approaches have been proposed in several scientific fields especially for analyzing large data matrix where the role of the two modes, which are usually units (rows) and variables (columns), can be exchangeable. Some of the most popular examples are: text mining, web-mining, bioinformatics, marketing, ecology, computer science, among others. Literature on biclustering is quite extensive, the interested reader is referred to [4] for a structured overview and to [5] for having a look at the available toolboxes introduced in the last few years. Specifically looking at mixture-based approaches, it is only in recent times that several biclustering methods have been proposed for discrete data ([9], [8], [5], [2]).

Here, we focus on multivariate discrete longitudinal data which represent a subclass of three-way data. A number of techniques have been developed for clustering three-way data ranging from sequential procedure (i.e. dimension reduction techniques applied to one of the way, and thereby conventional two-way data clustering technique applied to the reduced two-way data matrix) to simultaneous clustering and data reduction. On the other hand, only few examples of biclustering specific to three-way data are available in the literature up to our knowledge ([7], [6]). Specifically, we propose a mixture-based biclustering approach where, within each mixture component, the canonical parameter is suitably reparametrized to identify clusters of variables evolving in a similar manner across time by making use of adequate time functions. Notice that the proposed model can be seen as a longitudinal extension of the biclustering model proposed by [5]. Parameter estimates are derived within a maximum likelihood framework based on a EM type algorithm. The performance of the proposed approach is discussed on both simulated and real dataset.

## 2 Biclustering discrete data

The reference approach for biclustering discrete data is the model proposed in [5], where the key idea for introducing variable clustering was to modify finite mixtures of factor analyzers through a suitable reparameterization of the variable-specific parameters. In details, let  $n$  and  $p$  be the size of the observed units and the number of observed variables, respectively. For a given unit, the vector  $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})'$  is observed, where  $y_{ij}$  represents the value of the  $j$ -th variable for the  $i$ -th unit,  $i = 1, \dots, n$ ,  $j = 1, \dots, p$ . As it is usual in model-based clustering through finite mixtures, we assume that  $\mathbf{y}_i$  is drawn from a population  $\mathcal{P}$  formed by  $K$  subpopulations  $\mathcal{P}_k$ ,  $k = 1, \dots, K$ , where the prior probability that a generic unit comes from the subpopulation  $\mathcal{P}_k$  is indicated by  $\pi_k = \Pr(i \in \mathcal{P}_k)$ ,  $0 < \pi_k \leq 1$ ,  $k = 1, \dots, K$ , with  $\sum_{g=1}^k \pi_g = 1$ . Without loss of generality, let us assume that  $y_{ij}$  is a count and that



the  $p$  responses are conditionally independent Exponential Family (EF) response variables given the  $k$ -th component-specific effect. Thus, denoting by  $z_{ik}$  the unobserved component membership indicator, where  $z_{ik} = 1$  if the  $i$ -th unit belongs to the  $k$ -th component ( $i = 1, \dots, n, k = 1, \dots, K$ ), the conditional density of  $\mathbf{y}_i$  may be expressed as follows:

$$f(\mathbf{y}_i | z_{ik} = 1) = \prod_{j=1}^p f(y_{ij} | \theta_{j(k)}; \sigma_k) \quad (1)$$

$$= \prod_{j=1}^p \exp \left\{ \frac{y_{ij} \theta_{j(k)} - c(\theta_{j(k)})}{b(\sigma_k)} + d(y_{ij}; \sigma_k) \right\}$$

where  $f(y_{ij} | \theta_{j(k)})$  represents a generic density in the Exponential Family (EF) with canonical parameter  $\theta_{j(k)}$  corresponding to the  $j$ -th response for a unit in the  $k$ -th component;  $b(\cdot)$ ,  $c(\cdot)$ ,  $d(\cdot)$  are known functions and  $\sigma_k$  is the component-specific dispersion parameter.

To introduce variable partitioning within the  $k$ -th component (i.e. to identify segments), the parameter  $\theta_{j(k)}$  is modeled as

$$\theta_{j(k)} = \phi_k + \mathbf{a}'_{jk} \boldsymbol{\beta} \quad (2)$$

$j = 1, \dots, p, k = 1, \dots, K$ , where  $\phi_k$  is a component-specific latent effect,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q, \dots, \beta_Q)'$  where  $\beta_q$  represents a segment-specific latent effect and  $\mathbf{a}_{jk}$  is a  $Q$ -dimensional component-specific vector ( $Q \leq p$ ) whose elements “select” the membership of the  $j$ -th variable to one of the  $Q$  segments, with

$$a_{jkq} = \begin{cases} 1 & \text{if the } j\text{-th variable is in the } q\text{-th segment} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Notice that, differently from standard biclustering techniques based on so called grid-clustering, such a modelling approach allows for a different number of variable-specific segments for each mixture component avoiding the unnecessary assumption of independence between the unit and the variable-specific partitions.

### 3 Extension to longitudinal data

Suppose that a set of  $p$  variables is observed at  $T$  consecutive occasions to a sample of  $n$  units and let  $y_{ijt}$  denote the value of the  $j$ -th variable at occasion  $t$  from unit  $i$ ,  $i = 1, \dots, n, j = 1, \dots, p, t = 1, \dots, T$ . Clearly, it is binary in the case of dichotomously-scored variables and categorical, with more than two categories, in the case of polytomously-scored variables.  $\mathbf{y}_{it} = (y_{i1t}, \dots, y_{ipt})'$  represents the vector of individual variables at occasion  $t$  from unit  $i$  ( $i = 1, \dots, n, t = 1, \dots, T$ ) while

$\mathbf{y}_i = (\mathbf{y}_{i1}, \dots, \mathbf{y}_{iT})'$  is the vector of individual variables recorded over all occasions from unit  $i$ .

In order to extend the model introduced in the previous section to the longitudinal framework, we assume that  $\mathbf{y}_i$  is drawn from a population  $\mathcal{P}$  formed by  $K$  subpopulations  $\mathcal{P}_k, k = 1, \dots, K$ , where  $\pi_k = \Pr(i \in \mathcal{P}_k), 0 < \pi_k \leq 1, k = 1, \dots, K$ , with  $\sum_{k=1}^K \pi_k = 1$ . Moreover, we assume local independence on variables and times; that is, conditional on  $i \in \mathcal{P}_g$ , the  $p$  variables recorded at the different occasions are independent. Thus, the conditional density (1) may be extended in a natural way as follows:

$$\begin{aligned}
 f(\mathbf{y}_i | z_{ik} = 1) &= \prod_{t=1}^T \prod_{j=1}^p f(y_{ijt} | \boldsymbol{\theta}_{jt(k)}; \boldsymbol{\sigma}_k) = \\
 &= \prod_{t=1}^T \prod_{j=1}^p \exp \left\{ \frac{y_{ijt} \boldsymbol{\theta}_{jt(k)} - c(\boldsymbol{\theta}_{jt(k)})}{b(\boldsymbol{\sigma}_k)} + d(y_{ijt}; \boldsymbol{\sigma}_k) \right\}
 \end{aligned}
 \tag{4}$$

where  $\boldsymbol{\theta}_{jt(k)}$  represents the canonical parameter corresponding to the  $j$ -th response for a unit at occasion  $t$  in the  $k$ -th component.

To induce variable partitioning, different parametrizations may be considered. Here, we focus on the following interesting reparametrization:

$$\boldsymbol{\theta}_{jt(k)} = \boldsymbol{\phi}_k + \mathbf{a}'_{kj} \boldsymbol{\beta}(t)
 \tag{5}$$

where the partition of the  $p$  variables is constant over all occasions within the  $k$ -th component, but the model takes into account of repeated measures on the same variables by defining a  $Q$ -dimensional time function vector  $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_Q(t))'$  with  $\beta_q(t)$  describing the segment-specific temporal evolution of the variables,  $t = 1, \dots, T$ . According to this parameterization, units do not change component over time, and variables are clustered with respect their temporal dynamics.

A simple way to model the segment-specific function  $\beta_q(t)$  is to assume it being a polynomial time function of degree  $R$  as follows:

$$\beta_q(t) = \sum_{r=0}^R \lambda_{qr} t^r
 \tag{6}$$

where  $\lambda_{qr}$  represents the polynomial coefficient. Thus, equation (5) can be rewritten as:

$$\boldsymbol{\theta}_{jt(k)} = \boldsymbol{\phi}_k + \mathbf{a}'_{kj} \Lambda \boldsymbol{\omega}(t)
 \tag{7}$$

where  $\boldsymbol{\omega}(t) = (1, t^1, t^2, \dots, t^R)'$  is an  $(R + 1)$ -dimensional design vector and  $\Lambda = (\lambda_1, \dots, \lambda_Q)'$  is a  $Q \times (R + 1)$  matrix of polynomial coefficients. Notice that, each row of the above matrix,  $\lambda_q$ , is an  $(R + 1)$ -dimensional vector containing segment-specific effects  $\lambda_{qr}, r = 0, \dots, R$ . For large enough degree  $R$ , a polynomial function allows to produce an extremely non-linear curve. Clearly, different choices of the degree  $R$  lead to different polynomials.

Biclustering longitudinal trajectories through a model-based approach

In order to add flexibility to the smooth segment-specific function, we may refer to the more general basis function approach, which includes polynomials as special cases. The idea is to represent the segment-specific function as a linear combination of  $L$  basis functions  $\{\phi_{q1}(t), \dots, \phi_{qL}(t)\}$  taking

$$\beta_q(t) = \sum_{l=1}^L \lambda_{ql} \phi_{ql}(t) \quad (8)$$

where  $\alpha_{qr}$  are basis function coefficients and  $\phi_{q1}(t), \dots, \phi_{qL}(t)$  are fixed and known. Note that, for polynomial, the basis functions are  $\phi_{ql}(t) = t^l$ . Several basis functions available in the specialized literature may be adopted (cubic splines, truncated power, B-splines, Fourier, P-splines, among others).

#### 4 Parameter estimation

For parameter estimation, we consider the following observed log-likelihood function for  $n$  independent observations:

$$\ell(\Psi) = \sum_{i=1}^n \log f(\mathbf{y}_i | \Psi) = \sum_{i=1}^n \log \sum_{k=1}^K \pi_k f(\mathbf{y}_i | z_{ik} = 1) \quad (9)$$

where  $\Psi$  represents the vector of model parameters and  $f(\mathbf{y}_i | z_{ik} = 1)$  is defined according to Equations (4) and (5) and by one among Equation (6) or (8) describing a different segment-specific function. To compute the maximum likelihood (ML) estimate of  $\Psi$ , we adopt a EM type algorithm [1] which makes use of the complete data framework. Considering the  $z_{ik}$ 's as missing data and adopting a Multinomial distribution for the vector  $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})'$ , the complete-data log-likelihood function of  $\Psi$  is

$$\ell_c(\Psi) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log(\pi_k) + \sum_{i=1}^n \sum_{k=1}^K \sum_{j=1}^p \sum_{t=1}^T z_{ik} \log f(y_{ijt} | z_{ik} = 1). \quad (10)$$

Denoting with  $\mathbf{W} = \{w_{ik}\}_{(k=1, \dots, K, i=1, \dots, n)}$  and  $\mathbf{A} = \{\mathbf{a}_{kj}\}_{(k=1, \dots, K, j=1, \dots, p)}$  the matrix of posterior weights used to cluster units and the matrix used to cluster variables, respectively; the EM algorithm proceeds initializing  $\mathbf{A}$  and  $\mathbf{W}$  (randomly or in a deterministic way) and iterating the following steps till convergence:

1. **Updating  $\beta(t)$  and  $\phi$**   
Conditional on  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{W}}$ , the likelihood is a weighted version of a standard generalized linear (mixed-effects in case of P-spline) model likelihood. Therefore, we update  $\hat{\beta}(t)$  ( $t = 1, \dots, T$ ) and  $\hat{\phi} = (\hat{\phi}_1, \dots, \hat{\phi}_K)$  via Newton-Raphson algorithm with augmented data  $(\mathbf{Y}, \mathbf{W}, \mathbf{A})$  (no closed form);
2. **Updating  $\pi$**   
Conditional on  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{W}}$ , update  $\hat{\pi} = \{\hat{\pi}_k\}_{(k=1, \dots, K)}$ , via

$$\hat{\pi}_k = \frac{\sum_{i=1}^n \hat{w}_{ik}}{n} \quad k = 1, \dots, K;$$

**3. Updating A**

Conditional on  $\hat{\mathbf{W}}$ ,  $\hat{\beta}(t)$ ,  $\hat{\phi}$  and  $\hat{\pi}$ , update  $\hat{\mathbf{A}}$  as follows

- a. Consider the  $j$ -th variable and the  $k$ -th component,  $j = 1, \dots, p, k = 1, \dots, K$  and compute the log-likelihood contribution  $l_{jkq}$  for all  $q = 1 \dots, Q$ ;
- b. Fixed  $j$  and  $k$ , compute the maximum of the log-likelihood values  $l_{jkq}$  over  $q = 1, \dots, Q$  and denote it by  $\ell_{jk}^{max}$ ;
- c. In the  $k$ -th component, allocate the  $j$ -th variable to according

$$a_{jkq} = 1 \Leftrightarrow \ell_{jkq} = \ell_{jk}^{max},$$

$$j = 1, \dots, p, k = 1, \dots, K.$$

4. Conditional on  $\hat{\Psi}$ , update  $\hat{\mathbf{W}}$  as follows

$$w_{ik} = \frac{\pi_k f(\mathbf{y}_i | z_{ik} = 1)}{\sum_{k=1}^K \pi_k f(\mathbf{y}_i | z_{ik} = 1)}.$$

At convergence, each unit is assigned to the component with the highest posterior probability by using  $\mathbf{W} = \{w_{ik}\}$  and each variable is assigned to the  $q$ -th cluster by using  $\hat{\mathbf{a}}_{kj}$  ( $j = 1, \dots, p, k = 1, \dots, K$ ).

Further details about initialization strategy, convergence and model selection as well as simulation and real data results will be given in the extended version of the paper.

**References**

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.*, **39**, 1–38 (1977)
2. Fernández, D., Arnold, R., Pledger, S., Liu, I., Costilla, R.: Finite mixture biclustering of discrete type multivariate data. *Adv. Data Anal. Classif.*, **13**, 117–143 (2019)
3. Good, I.: *Categorization of classification Mathematics and Computer Science in Biology and Medicine* (London: Her Majesty’s Stationary Office), 115–128 (1965)
4. Govaert, G., Nadif, M.: *Co-Clustering: Models, Algorithms and Applications*, Wiley, New York (2014)
5. Martella, F., Alfó, M.: A finite mixture approach to joint clustering of individuals and multivariate discrete outcomes. *J. Stat. Comput. Simul.*, **87(11)**, 2186–2206 (2017)
6. Mankad, S., Michailidis, G.: Biclustering Three-Dimensional Data Arrays With Plaid Models. *J. Comput. Graph. Statistica*, **23(4)**, 943–965, (2014)
7. Turner, H., Bailey, T., Krzanowski, W., and Hemingway, C.: Biclustering Models for Structured Microarray Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **2**, 316–329 (2005)
8. Vicari, D., Alfó, M.: Model based clustering of customer choice data. *Comput. Statist. Data Anal.*, **71**, 3–13 (2014)
9. Wyse, J., Friel, N.: Block clustering with collapsed latent block models. *Stat. Comput.*, **22(2)**, 415–428 (2012)

# Monitoring tools for robust estimation of Cluster Weighted models

## *Strumenti di monitoring per la stima robusta del modello Cluster Weighted*

Andrea Cappelletto and Francesca Greselin

**Abstract** In a robust approach to model fitting for the cluster weighted model, many choices are to be made by the statistician: specifying the shape of the clusters in the explanatory variables, assuming (or not) equal variance for the errors in the regression lines, and setting hyper-parameter values for the robust estimation to be protected from outliers and contamination. The most delicate hyper-parameter to specify is perhaps the percentage of trimming, or the amount of data to be excluded from the estimate, to ensure reliable inference. In this work we introduce diagnostic tools to help the professional, or the scientist who needs to group the data, to make an educated choice about this hyper-parameter, after a first exploration of the resulting model space.

**Abstract** *Nella stima robusta di un cluster weighted model, lo statistico deve fare molte scelte: specificare la forma dei cluster nelle variabili esplicative, assumere (o meno) varianza uguale per gli errori nelle linee di regressione e impostare i valori degli iper-parametri per la stima robusta, per evitare la distorsione generata da valori anomali e contaminazione. L'iper-parametro più delicato da specificare è la percentuale di trimming, ovvero la quantità di dati da escludere nella stima per garantirne l'affidabilità. In questo lavoro introduciamo specifici strumenti diagnostici per aiutare il professionista, o lo scienziato che ha bisogno di classificare i dati, a compiere una scelta ragionata a riguardo di tale iper-parametro, anche in base ad una prima esplorazione dello spazio delle soluzioni.*

**Key words:** Cluster-weighted modeling, Outliers, Trimmed BIC, Eigenvalue constraint, Monitoring

---

Andrea Cappelletto  
University of Milano Bicocca, Department of Statistics and Quantitative Methods e-mail:  
a.cappelletto@unimib.it

Francesca Greselin  
University of Milano Bicocca, Department of Statistics and Quantitative Methods e-mail:  
francesca.greselin@unimib.it

## 1 Introduction

Clustering is a well known ill-posed problem, where the number of groups, their shape, and their parameters depend, in general, on a multiplicity of subjective choices [4]. Generally, selecting the unknown number of groups  $G$  defines the most challenging task. The most popular method adopted in model-based clustering for tackling the problem is based on penalized likelihoods, but the presence of data contamination and outliers could severely undermine such powerful criteria. In addition, when it comes to cluster weighted modeling, many other choices need to be performed: whether to constrain the cluster shapes in the explanatory variables, to impose or not equal variances in the regression errors, how to set hyper-parameters for discarding spurious solutions and how to protect against outliers.

We introduce here a semiautomatic procedure for selecting a reduced set of solutions, extending to the cluster weighted model the methodology developed in [1] for the Gaussian mixture models. Such an extension is far from being straightforward. A new penalized likelihood criterion will be devised to account for the constraint imposed on the regression term and on the covariates, varying trimming levels and number of cluster. The remainder of the article proceeds as follows. Section 2 provides a brief overview of the Cluster Weighted Model (CWM) and its robust estimation. Section 3 reports the two-stage monitoring strategy, based on (i) a first exploration of the model space with a dedicated information criterion and (ii) usage of new “trimming-based” tools, tailored for CWM. Section 4 concludes the paper by showcasing the validity of our proposal within a controlled experiment.

## 2 The cluster weighted model

Let  $\mathbf{X}$  be a vector of *explanatory* variables with values in  $\mathbb{R}^d$ , and let  $Y$  be a *response* or *outcome* variable, with values in  $\mathbb{R}$ . Suppose that the regression of  $Y$  on  $\mathbf{X}$  varies across the  $G$  levels (group or clusters) of a categorical latent variable. The CWM, introduced in [3], decomposes the joint p.d.f. of  $(\mathbf{X}, Y)$  in each component of the mixture as the product of the marginal and the conditional distributions as follows

$$p(\mathbf{x}, y; \theta) = \sum_{g=1}^G \pi_g p(y|\mathbf{x}; \xi_g) p(\mathbf{x}; \psi_g). \quad (1)$$

In the cluster-weighted approach the marginal distribution of  $\mathbf{X}$  and the conditional distribution of  $Y|\mathbf{X} = \mathbf{x}$  may have different scatter structures in each group. In this work, we focus on the *linear Gaussian CWM*:

$$p(\mathbf{x}, y; \theta) = \sum_{g=1}^G \pi_g \phi(y; \mathbf{b}'_g \mathbf{x} + b_g^0, \sigma_g) \phi_d(\mathbf{x}; \mu_g, \Sigma_g), \quad (2)$$

where  $\phi_d(\cdot; \mu_g, \Sigma_g)$  denotes the density of the  $d$ -variate Gaussian distribution with mean vector  $\mu_g$  and covariance matrix  $\Sigma_g$ , and  $Y$  is related to  $\mathbf{X}$  by a linear model,

that is,  $Y = \mathbf{b}'_g \mathbf{x} + b_g^0 + \varepsilon_g$  with  $\varepsilon_g \sim N(0, \sigma_g^2)$ ,  $\mathbf{b}_g \in \mathbb{R}^d$ ,  $b_g^0 \in \mathbb{R}$ ,  $\forall g = 1, \dots, G$ . Given a sample of  $n$  i.i.d. pairs drawn from  $(Y, \mathbf{X})$ , the ML estimation of the linear Gaussian CWM is based on the maximization of the following log-likelihood function

$$\mathcal{L} = \sum_{i=1}^n \log \left[ \sum_{g=1}^G \pi_g \phi(y_i; \mathbf{b}'_g \mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right]. \quad (3)$$

Unfortunately, ML inference on models based on normal assumptions suffers from lack of robustness. Another important concern is the unboundedness of the likelihood function to be maximized. To overcome these issues, a robust version of the CWM has been presented in the literature by considering impartial trimming and constrained estimation of the scatter variances [2]. The robust approach to CWM (CWRM) is based on the maximization of the *trimmed* log-likelihood function [6]

$$\mathcal{L}_{\text{trimmed}} = \sum_{i=1}^n z(\mathbf{x}_i, y_i) \log \left[ \sum_{g=1}^G \pi_g \phi(y_i; \mathbf{b}'_g \mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right], \quad (4)$$

where  $z(\cdot, \cdot)$  is a 0-1 trimming indicator function that tells us whether observation  $(\mathbf{x}_i, y_i)$  is trimmed off ( $z(\mathbf{x}_i, y_i)=0$ ), or not ( $z(\mathbf{x}_i, y_i)=1$ ). A fixed fraction  $\alpha$  of observations is unassigned by setting  $\sum_{i=1}^n z(\mathbf{x}_i, y_i) = \lfloor n(1 - \alpha) \rfloor$ , and the parameter  $\alpha$  denotes the trimming level.

We introduce two constraints on the maximization in (4). The first one concerns the set of eigenvalues  $\{\lambda_l(\boldsymbol{\Sigma}_g)\}_{l=1, \dots, d}$  of the scatter matrices  $\boldsymbol{\Sigma}_g$  by imposing

$$\lambda_{l_1}(\boldsymbol{\Sigma}_{g_1}) \leq c_X \lambda_{l_2}(\boldsymbol{\Sigma}_{g_2}) \quad \text{for every } 1 \leq l_1 \neq l_2 \leq d \text{ and } 1 \leq g_1 \neq g_2 \leq G. \quad (5)$$

The second constraint refers to the variances  $\sigma_g^2$  of the regression error terms, by requiring

$$\sigma_{g_1}^2 \leq c_\varepsilon \sigma_{g_2}^2 \quad \text{for every } 1 \leq g_1 \neq g_2 \leq G. \quad (6)$$

The constants  $c_X$  and  $c_\varepsilon$ , in (5) and (6) are finite (not necessarily equal) real numbers, such that  $c_X \geq 1$  and  $c_\varepsilon \geq 1$ . They automatically guarantee that we are avoiding the  $|\boldsymbol{\Sigma}_g| \rightarrow 0$  and  $\sigma_g^2 \rightarrow 0$  degenerate cases.

### 3 Monitoring the setting of CWM hyper-parameters

We propose a semi-automatic approach to provide adaptive values for the hyper-parameter  $\alpha$  involved in the robust fitting of CWMs. By building upon previous work developed for robust clustering [7], a two-stage monitoring procedure is devised. First off, for each trimming level  $\alpha \in \{0, \dots, \alpha_{MAX}\}$  ( $\alpha_{MAX} = 0.15$  in the analysis of Section 3) the most appropriate model, varying  $G$ ,  $c_X$  and  $c_\varepsilon$ , is determined. Secondly, exploratory tools are employed to compare solutions for different levels of  $\alpha$ , providing aid in assessing the true contamination level present in a dataset.

In details, in the first phase a constrained estimation criterion is devised for comparing models when  $\alpha$  is kept fixed. As in the well known Bayesian Information Criterion ( $BIC = -2\mathcal{L} + v_G$ ) and along the lines of [1], the dedicated penalty term  $v_G$  depends on the number of free parameters in the model:

$$v_G = \{(G - 1) + Gp + G(p + 1) + 1 + ((Gp - 1) + Gp(p - 1)/2)(1 - 1/c_X) + 1 + (G - 1)(1 - 1/c_\epsilon)\} \log(\lceil n(1 - \alpha) \rceil). \quad (7)$$

The first three terms in (7) respectively refer to the  $(G - 1)$  mixture weights, the  $Gp$  cluster means of the covariates, and the  $G(p + 1)$  beta coefficients for the regression  $\mathbf{b}_g + b_g^0$ ,  $g = 1, \dots, G$ . The second group of terms is related to the modelling of  $\mathbf{X}$ , where we have 1 free eigenvalue,  $Gp - 1$  constrained eigenvalues and  $Gp(p - 1)/2$  rotation matrices for  $\Sigma_g$ . Except the first one, all terms are multiplied by  $(1 - 1/c_X)$  to take into account the enforced constrained estimation. Lastly, in the third line of (7), the part relative to modelling  $Y|\mathbf{X}$  induces one free  $\sigma_g^2$  and  $G - 1$  constrained  $\sigma_g^2$ . Notice that, while in [1] the authors distinguish between rotation and eigenvalue parameters multiplying only the latter by the factor  $(1 - 1/c_X)$ , we opt here for penalizing all the variance parameters, as rotation loses its meaning for  $c_X \rightarrow 1$ .

In the second phase, we extend the monitoring introduced in [7], where a plot of the Adjusted Rand Index (ARI) between consecutive cluster allocations for a grid of  $\alpha$  values is proposed, to determine an optimum trimming level. This tool can be effective in detecting noise in the form of bridges, where only a correct level of trimming uncovers the true underlying structure. In the case of scattered noise, however, the clustering structure could evolve very smoothly from an initial partition, obtained without trimming, and a pretty different final partition, yielding an ARI pattern between consecutive solutions with no apparent abrupt change. Motivated by this argument, we widen the monitoring tools accompanying the ARI plot with regression coefficients and mixture weights paths, to highlight specific CWM features. Further, we are interested in monitoring the CWM validation measure based on the decomposition of the total sum of squares  $TSS = BSS + RWSS + EWSS$  [5].  $BSS$  is the (soft) between-group sum of squares, while  $EWSS$  is the portion of the (soft) within-group sum of squares  $WSS$  explained by the model, thanks to the covariate, and  $RWSS$  is the residual portion of  $WSS$ . In terms of cluster validation, therefore,  $BSS$  can be seen as a separation measure on the  $Y$ -axis, and  $WSS$  can be seen as a cluster compactness measure. To overcome the non-identifiability issue due to invariance of mixture components, a relabeling strategy based on data depth [8] is adopted. In this way, component-dependent metrics, estimated varying trimming levels, are directly comparable: an application is provided in the next section.

## 4 Illustrative experiment

A dataset with 180 genuine samples is generated according to (2) with the following parameters:



$$\begin{aligned} \pi &= (0.5, 0.5)', \quad \mu_1 = (2, 2)', \quad \mu_2 = (5, 5)', \quad \Sigma_1 = \Sigma_2 = I_2 \\ b_1^0 &= 30, \quad b_2^0 = 50, \quad \mathbf{b}_1 = (-1, -1)', \quad \mathbf{b}_2 = (10, 10)', \quad \sigma_1^2 = \sigma_2^2 = 1, \end{aligned} \quad (8)$$

in addition, 20 uniformly distributed outliers are appended to the uncontaminated observations, resulting in a total of  $n = 200$  data units with a true contamination level equal to 0.1. In the first phase, models with  $c_X, c_E \in \{1, 4, 16, 64\}$  and  $G = \{2, 3, 4\}$  are fitted to the considered dataset: Table 1 reports the best model, selected by minimizing the information criterion introduced in the previous section (denoted with TBIC in the table), conditioning on the trimming value  $\alpha$ . Notice that, whenever  $\alpha$  is set below the true contamination rate, some erroneous solutions are preferred:  $G$  is selected to be greater than 2, with spurious groups fitting the portion of untrimmed noise. The second phase of our procedure encompasses the plots reported in Figure

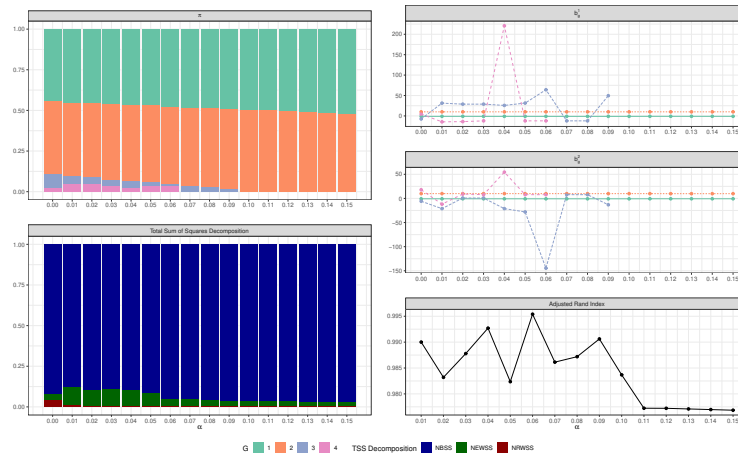
**Table 1** Best models, as a function of  $G$ ,  $c_X$  and  $c_E$ , selected via TBIC minimization conditioning on the trimming value  $\alpha$  (only a subset of the entire  $\alpha$  grid considered in the experiment is reported) for the first phase of the monitoring procedure.

$\alpha$	0.00	0.03	0.06	0.09	0.10	0.11	0.12	0.13	0.14	0.15
$G$	4	4	4	3	2	2	2	2	2	2
$c_X$	4	4	64	4	4	4	4	4	4	4
$c_E$	64	64	64	64	1	1	1	1	1	1
TBIC	2801.82	2363.97	2157.08	1998.03	1940.30	1885.77	1848.96	1812.12	1776.44	1741.11

1: by monitoring the changes in mixing proportions, regression parameters, total sum of squares and ARI between consecutive cluster allocations the analyst may reasonably observe how the solutions stabilize as soon as  $\alpha$  is higher than the true contamination level 0.1. Particularly, given the ARI almost constant high value (bottom right plot), this metric alone would not have been sufficient to properly address the complexity of the problem.

## 5 Conclusions

The present article provides a two-stage monitoring procedure for aiding in the hyper-parameters selection when fitting robust CWM to contaminated datasets. We opted for providing the user with sensible information to make the required tuning decisions: ultimately an optimal tuning of model parameters should also depend on knowledge about the subject matter background and the aim of clustering. The procedure takes over and extends the state-of-the-art methods proposed for robust clustering by including a wider range and component-dependent metrics, essential for thoroughly understanding the true data generating mechanism.



**Fig. 1** Monitoring the mixing proportions (top left plot), regression parameters (top right plots), total sum of squares decomposition (bottom left plot) and ARI between consecutive cluster allocations (bottom right plot, please be aware of the  $Y$  axis range) as a function of the trimming proportion  $\alpha$ .

## References

- [1] A. Cerioli, L. A. García-Escudero, A. Mayo-Iscar, and M. Riani. Finding the number of normal groups in model-based clustering via constrained likelihoods. *Journal of Computational and Graphical Statistics*, 27(2):404–416, apr 2018.
- [2] L. A. García-Escudero, A. Gordaliza, F. Greselin, S. Ingrassia, and A. Mayo-Iscar. Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Statistics and Computing*, 27(2):377–402, mar 2017.
- [3] N. Gershenfeld. Nonlinear Inference and Cluster-Weighted Modeling. *Annals of the New York Academy of Sciences*, 808(1 Nonlinear Sig):18–24, jan 1997.
- [4] C. Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015.
- [5] S. Ingrassia and A. Punzo. Cluster Validation for Mixtures of Regressions via the Total Sum of Squares Decomposition. *Journal of Classification*, 37(2):526–547, jul 2020.
- [6] N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299–308, sep 2007.
- [7] M. Riani, A. C. Atkinson, A. Cerioli, and A. Corbellini. Efficient robust methods via monitoring for clustering and multivariate data analysis. *Pattern Recognition*, 88:246–260, apr 2019.
- [8] K. Singh, J. M. Parelus, and R. Y. Liu. Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics*, 27(3):783–858, jun 1999.

# Co-clustering Models for Spatial Transcriptomics: Analysis of a Human Brain Tissue Sample

## *Modelli di Co-clustering per la Trascrittomica Spaziale: Analisi di un Campione di Tessuto di Cervello Umano*

Andrea Sottosanti and Davide Risso

**Abstract** In the last few years, we have witnessed a substantial improvement in the efficiency of DNA sequencing techniques with the advent of *10X-Visium*, a new technology that is capable of providing the expression of tens of thousands of genes inside thousands of cells from a tissue sample. From a statistical perspective, this technology represents an astonishing step forward in the analysis of single cell data, as it gives access to a huge amount of information inaccessible to us until now. In this paper, we apply some innovative statistical methods that cluster both the rows and the columns of a data matrix to a human brain tissue sample processed with *10X-Visium*. This operation is known as *co-clustering* and aims to detect groups of genes whose expression activity is similar in some specific areas of the brain tissue.

**Abstract** Negli ultimi anni, si è assistito ad un sostanziale aumento dell'efficienza di tecnologie per il sequenziamento del DNA con la nascita di *10X-Visium*, una nuova tecnologia in grado di fornire l'espressione di decine di migliaia di geni misurata in migliaia di cellule provenienti da un campione di tessuto. Da un punto di vista statistico, questa tecnologia rappresenta un notevole passo in avanti nell'analisi dei dati a singola cellula, in quanto permette di accedere a una grande quantità di informazione fino ad ora inaccessibile. In questo articolo, applichiamo alcuni metodi statistici che identificano gruppi di righe e di colonne di una matrice di dati su un campione di tessuto di cervello umano processato con *10X-Visium*. Tale operazione, detta *co-clustering*, ha l'obiettivo di individuare gruppi di geni la cui attività di espressione è simile in alcune aree specifiche del tessuto di cervello raccolto.

**Key words:** LASSO regularization, Model-based Co-clustering, Spatial Transcriptomics

---

Andrea Sottosanti

University of Padova, Department of Statistical Sciences, via Cesare Battisti 241/243, 35121 Padova - Italy, e-mail: andrea.sottosanti@unipd.it

Davide Risso

University of Padova, Department of Statistical Sciences, via Cesare Battisti 241/243, 35121 Padova - Italy, e-mail: davide.risso@unipd.it

## 1 Introduction

*Spatial Transcriptomics* is a modern sequencing technology that allows to measure the activity of all the genes in a tissue sample and map where the activity is occurring. The recent *10X-Visium*, developed by *10X-Genomics*, is a new outstanding protocol that furnishes the location of thousands of cells from a tissue sample, thus providing a complete reconstruction of its morphology.

The data are collected using a grid of spots, each of which gathers a single cell or at most few neighbour cells. Then, for each spot, the expression of thousands of genes is measured.

The huge amount of information carried by the data processed with *10X-Visium* opens its doors to relevant and impactful data analyses. First, we can exploit the statistical dependency of the cells to achieve a better classification of their nature (for example, distinguishing *tumoral* cells from *stromal* and *immune* cells), and second, we can determine if there exist groups of genes that are particularly active only in some specific cell types. Such conclusions can be achieved by performing a double clustering on the data: first, on the cells of the tissue, and then on the genes.

In this manuscript, we analyse a human brain tissue sample processed with *10X-Visium* using some advanced clustering models. Our goal is checking if the existing methods are capable of extracting the huge amount of information contained into the dataset, and eventually try to determine which directions the statistical research should take in order to help comprehending the human genome functions.

## 2 Dorsolateral Prefrontal Cortex Data

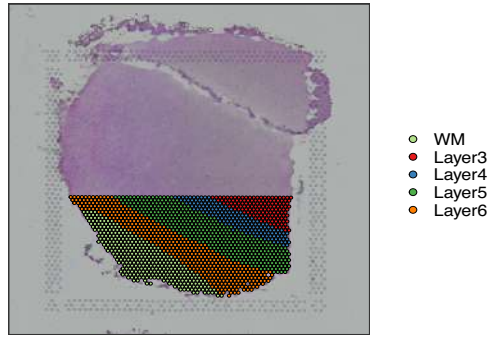
We consider the data contained in the R package `spatialLIBD` by [2]. It collects some tissue samples of the dorsolateral prefrontal cortex taken in 12 subjects. In this analysis, we focus only on the data from the first subject, coded as *151507*. The spots containing the tissue cells have been manually annotated by the researchers according to their functionality in the organism and divided in 7 layers: *Layers 1–6* and *White Matter (WM)*. To reduce computational cost, we focus our attention only on a subregion in the lower part of the tissue containing 1534 cells (see Figure 1). In this area, there are only 5 of the 7 layers.

### Pre-processing analysis

Genomic datasets need first to be pre-processed. This phase includes three steps: *i*) removing too low expressed genes, *ii*) selecting the most informative genes, and *iii*) normalizing data at a single-cell level. Each spot of the grid contains the information on the gene expression in terms of UMI counts (*unique molecular identifier*, see [6]).

We performed Step *i*) by removing those genes whose total counts in the 1534 cells was smaller than 500. Then, to perform Step *ii*), we fitted a binomial GLM

**Fig. 1** Tissue sample from subject 151507. The region we considered for the analysis is in the lower part of the figure, denoted by coloured spots. In total, we consider 1534 cells. The manual annotation of the cells in this region has determined the presence of 5 biological layers in this area.



on each gene assuming a constant rate, and for each one we computed the deviance statistic. [6] showed that genes with a large deviance are likely to be informative. We further reduced the data size by picking the first 600 genes with the largest deviance value. Last, we accomplished Step *iii*) by considering for our analysis the Pearson residuals of the GLM regressions applied in Step 2. [6] showed that, in the presence of UMI counts, the three steps we listed guarantee a good normalization of the data at a single-cell level, and allow to work with properly symmetrized continuous data.

### 3 Co-clustering Methods

The scope of our analysis is to determine a suitable statistical method to perform the two types of clustering operations on 10X-Visium-type data that we discussed in Section 1: first, a clustering of the cells to infer on their nature, and second, a clustering of the genes with respect their expressions across cells. *Co-clustering* [1] is a family of statistical techniques whose scope is to perform a simultaneous clustering of rows and columns, thus partitioning the data matrix into multiple non-overlapping sub-matrices, called *co-clusters* or *blocks*. A clear aspect that characterizes a proper co-clustering technique is that the allocation of the rows depends on the allocation of the columns, and vice versa. In our application, every co-cluster represents a specific group of genes which are, in some ways, similar to each other in a specific cell type.

Let  $\mathbf{X}$  be the  $n \times p$  matrix whose element  $(i, j)$  gives the expression of gene  $i$  into cell  $j$ . As we discussed earlier in Section 2, our dataset has  $n = 600$  and  $p = 1534$ . Additionally, given  $k = 1, \dots, K$  and  $r = 1, \dots, R$ , let  $\mathbf{z} = \{z_{ik}\}_{i,k}$  and  $\mathbf{w} = \{w_{jr}\}_{j,r}$  be the clustering indicators for the rows and for the columns: thus,  $z_{ik} = 1$  means that the gene  $i$  belongs to gene cluster  $k$ , and  $w_{jr} = 1$  means that the cell  $j$  belongs to the cell cluster  $r$ . Finally, a *block*  $\mathbf{X}_{kr}$  is a submatrix of  $\mathbf{X}$  made by the rows and the columns which have  $z_{ik} = 1$  and  $w_{jr} = 1$ .

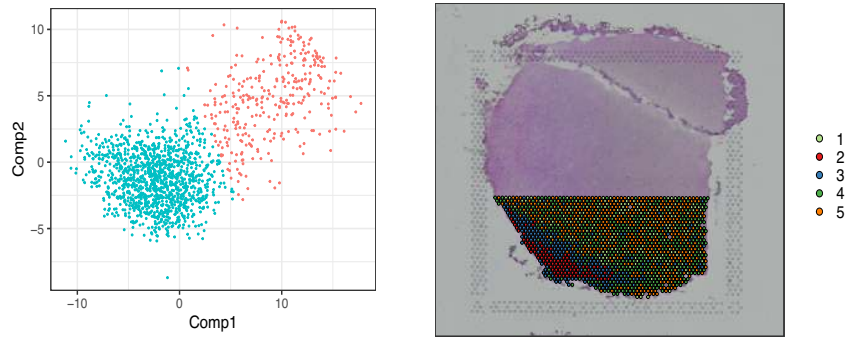
In the following, we present three powerful co-clustering methods proposed in the statistical literature within the last decade.

1. **Sparse Biclustering:** introduced by [5], it extends the well-known K-means algorithm in a co-clustering framework, allocating the rows and the columns to the closest co-cluster based on the distance from the centroids. The model allows also a LASSO regularization of the cluster centroids controlled by a sparsity parameter  $\lambda$ . If  $\lambda = 0$ , the model is equivalent to two independent K-means procedures applied independently to the rows and to the columns of the data matrix. From a statistical perspective, this model is equivalent to assuming that each observation in a given block is *iid* and comes from a Gaussian distribution, with a block-specific mean and a variance shared with all the other blocks: thus,  $x_{ij}|z_{ik} = 1, w_{jr} = 1 \sim \mathcal{N}(\mu_{kr}, \sigma^2)$ .
2. **Latent Block Model (LBM):** it is a vast class of statistical models for co-clustering. The basic idea behind the LBM is to extend the classical mixture model by allowing for two distinct clustering labels, for the rows and for the columns. We assume every observation in the  $kr$ -th block to be Gaussian distributed,  $x_{ij}|z_{ik} = 1, w_{jr} = 1 \sim \mathcal{N}(\mu_{kr}, \sigma_{kr}^2)$ , and we further assume that  $z_{ik} \sim Be(\pi_k)$  and  $w_{jr} \sim Be(\rho_r)$ . The LBM assumes both independence of the clustering labels (so,  $p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$ ), and of the data inside the blocks. Inference on this model can be carried out using the variational Bayes algorithm showed by [1].
3. **sparse Matrix Variate Normal (MVN) biclustering:** proposed by [5], it extends the Sparse Biclustering by considering a matrix  $\Sigma_k$  that governs the covariance of the genes inside the  $k$ -th row cluster, and a matrix  $\Delta_r$  that governs the covariance of the cells inside the  $r$ -th column cluster. These assumptions formalize into a statistical model over the entire block:  $\mathbf{X}_{kr}|\mathbf{z}_k, \mathbf{w}_r \sim \mathcal{M}\mathcal{V}\mathcal{N}(\mu_{kr}\mathbf{1}, \Sigma_k, \Delta_r)$ , where  $\mathcal{M}\mathcal{V}\mathcal{N}$  denotes the matrix-variate normal distribution,  $\mu_{kr}$  is the mean matrix and  $\mathbf{1}$  is a matrix of ones of the same size of  $\mathbf{X}_{kr}$ . A LASSO penalization regulates both the covariance matrices to avoid singularity problems. Thus, the contribute to the log-likelihood given by the  $(k, r)$ -th block is

$$\log \mathcal{M}\mathcal{V}\mathcal{N}(\mathbf{X}_{kr}; \mu_{kr}, \Sigma_k, \Delta_r) - \lambda |\mu_{kr}| - \alpha \sum_{i,i'} |\Sigma_{k,ii'}| - \beta \sum_{j,j'} |\Delta_{r,jj'}|.$$

The authors propose also an estimation algorithm which performs the allocation step by treating each element (row/column) as independent from the others. For this reason, the monotonically increase of the log-likelihood is not guaranteed.

Let us now look at how the three discussed models account for the dependency across rows and columns. The first model forms the clusters based on the Euclidean distance of the observations from the centroids. So, any observation is independent from the others. The second assumes that the rows and columns in a block are independent, but extends the previous model allowing for a block specific variance ( $\sigma_{kr}^2$ ) and a different probability for each block ( $\pi_k \rho_r$ ). Last, the sparse MVN method assumes a distribution over the entire block matrix, and so the observations are no longer independent. Since it is widely known that genes often correlate and can be clustered into groups [4, 3], and there exist also cells of different nature, the latter



**Fig. 2** Left: dimension reduction of the gene space using the GLM-PCA. We distinguish two groups of observations, that we highlight using a K-means classification. Right: cell clustering obtained from the sparse MVN procedure.

method seems to be the most appropriate to model a tissue sample generated with the 10X-Visium protocol.

## 4 Results

In this section, we confront the performance of the three co-clustering models discussed earlier on the brain tissue sample illustrated in Section 2. The first step of every unsupervised analysis is determining the number of clusters; in our problem, we need to find the values for  $K$  and  $R$ . Although there is a vast literature on methods for computing the most appropriate number of clusters, we chose here to follow a completely data-driven approach.

To determine the number of gene clusters, we applied the GLM-PCA procedure discussed by [6] taking  $\mathbf{X}^T$  and performing a dimension reduction of the variables, which are now the genes. The GLM-PCA was developed for generalizing the classic Principal Component Analysis to count data. Figure 2 shows the observations remapped in the space given by the first two principal components of the genes. We evidently distinguish two groups of observations, on the bottom left and on the top right of the figure, that we highlighted with different colours using a K-means classification. So,  $K = 2$  looks a reasonable choice for our dataset.

Regarding the cell clusters, we set  $R = 5$  according to the number of Layers in our tissue sample, as shown in Figure 1. In addition, we applied the algorithm of [5] for determining the number of blocks based on the cross-validation procedure. Although this approach is designed to work only with the Sparse Biclustering model, and so it does not account for the covariance of the genes and of the cells, it turned out to be useful and computationally efficient, especially in a first explanatory analysis of the data. The algorithm returned that the best choice is  $R = 5$ , fitting the Sparse Biclustering with both  $\lambda = 0$  and  $\lambda = 1$ .

**Table 1** Clustering of the genes given by the three methods discussed in Section 3. The label `sparseBiclust` denotes the Sparse Biclustering model.

sparse MVN:	<b>1</b>		<b>2</b>	
sparseBiclust / LBM	<b>1</b>	<b>2</b>	<b>1</b>	<b>2</b>
<b>1</b>	158	2	0	0
<b>2</b>	9	8	12	411

We then fitted the three models discussed. For Models 1 and 3, we set  $\lambda = 1$ ,  $\alpha = 1$  and  $\beta = 0.5$ , as we wanted to shrink some centroids toward zero, and at the same time we wanted to induce a small sparsity on the column covariances, to capture the dependency of the cells. However, the procedure returned a not-diagonal estimate of the  $\Sigma_k$  and a diagonal estimate of the  $\Delta_r$ . From this result, we conclude that the sparse MVN has effectively captured some information that could not be handled by the first two models, but some extensions of this method are still necessary (for example, by expressing  $\Delta_r$  as a function of a sparial kernel matrix, see [4, 3]). The right plot of Figure 1 shows the cell clusters obtained by the sparse MVN model. The groups denoted as 2 and 3, which are part of the White Matter layer (see Figure 1), were detected also by the two other competitor models. However, regarding the structure of the three remaining clusters, the three methods are not fully concordant.

Finally, Table 1 shows the results obtained from the classification of the genes: the three models agree in the classification of almost the 95% of the available observations, with 158 genes pulled in Cluster 1, and 411 genes in Cluster 2. The analysis of the cluster centroids allows us to interpret the results: we discovered that the genes in Cluster 1 are mainly active only in the Cell Cluster 2 (red points in the right plot of Figure 2), while the genes in Cluster 2 are considerably expressed in the Cell Clusters 1, 2 and 5.

## References

1. Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E.: Model-based clustering and classification for data science: with applications in R. Cambridge University Press, Volume 50 (2019)
2. Collado-Torres, L., Maynard, K. R., and Jaffe, A. E.: LIBD Visium spatial transcriptomics human pilot data inspector. R package version 1.2.0 (2020). <https://github.com/LieberInstitute/spatialLIBD>
3. Sun, S., Zhu, J., and Zhou, X.: Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, **17**(2):193–200 (2020)
4. Svensson, V., Teichmann, S. A., and Stegle, O.: SpatialDE: identification of spatially variable genes. *Nature methods*, **15**(5):343–346 (2018)
5. Tan, K. M. and Witten, D. M.: Sparse biclustering of transposable data. *Journal of Computational and Graphical Statistics*, **23**(4):985–1008 (2014)
6. Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A.: Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, **20**(1):1–16 (2019)



# Graph nodes clustering: a comparison between algorithms

## *Clustering di nodi in un grafo: un confronto tra algoritmi*

Ilaria Bombelli

**Abstract** Networks represent an important tool to describe problems and applications in various fields, such as science, technology and economics. Statistics can play a role in a network framework, for example using some clustering techniques to detect clusters of nodes. This work focuses on reviewing the existing algorithms designed specifically for this aim and on suggesting the application of other clustering techniques that require a matrix of distances or dissimilarities between units: a description of how to get such matrix is also provided. A comparison between the aforementioned algorithms is given, by applying them to a benchmark network.

**Abstract** *I network (o grafi) sono uno strumento importante per rappresentare problemi e applicazioni in vari campi, come quello scientifico, tecnologico e economico. La statistica può giocare un ruolo importante in questo contesto, per esempio applicando alcune tecniche di clustering per identificare clusters di nodi. Tale lavoro passa in rassegna gli algoritmi appositamente costruiti per raggiungere questo scopo e suggerisce anche l'applicazione di altri algoritmi che prendono in input una matrice di distanze o dissimilarità tra le unità: viene fornita anche una descrizione su come ottenere questa matrice. Gli algoritmi suddetti vengono confrontati, applicandoli ad un network di riferimento (benchmark).*

**Key words:** Clustering, Network, Nodes Clustering, Fuzzy clustering

---

Ilaria Bombelli  
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 (00185)  
Roma, e-mail: [ilaria.bombelli@uniroma1.it](mailto:ilaria.bombelli@uniroma1.it)

## 1 Introduction

Networks (graphs) can be found in many fields: for example, in a society, a network may represent how people interact one another; in technological field, networks may represent email exchange. A graph or a network is a mathematical tool representing connection or relationship between several objects. Formally, a graph  $G$  is defined as an ordered tuple of 2 sets, i.e.  $G = (V, E)$ , where  $V$  is the set of  $n$  unique nodes, i.e.  $V = \{v_1, \dots, v_n\}$  and  $E$  is the set of  $m$  edges, i.e.  $E = \{e_1, \dots, e_m\}$ .

In a network framework it can be of interest the application of clustering techniques: indeed, we can consider the nodes as the statistical units and the aim is to detect clusters of nodes. Usually in this framework, the starting point of many clustering algorithms, i.e. the distance matrix  $\mathbf{D}$ , is not provided and therefore, given a network object, a measure of distance between nodes has to be considered.

The paper will be organized as follows: in Section 2 the description of the clustering algorithms that have been used is provided, as well as the explanation of how to build a distance matrix  $\mathbf{D}$  from a network object; Section 3 shows an application of the clustering algorithms to a benchmark network. Finally in Section 4 final remarks are given and further possible developments are sketched.

## 2 Methodology

In this section, an overview on the clustering techniques is presented, followed by an explanation of how to build a distance matrix from a given graph.

### 2.1 Clustering algorithms

The main two classes of clustering algorithms are hierarchical and non-hierarchical: the former is a class of algorithms that generate  $n$  different nested classifications, the latter is a class of algorithms that give rise to a single partition with  $k$  (fixed before running the algorithm) groups. Hierarchical clustering algorithms can be either agglomerative or divisive.

In graph framework, agglomerative methods start with the only set of nodes  $V$ ; the whole network  $G = (V, E)$  will be progressively constructed by adding edges between nodes and involving nodes into nested larger and larger *communities* (subsets of the network).

Divisive methods instead start from the whole network and progressively cut edges and divide the network into smaller and smaller communities.

As examples of the two aforementioned different techniques two algorithms designed to be applied to a network in order to detect clusters of nodes are described and then applied in Section 3.

Louvain algorithm [2] belongs to the class of agglomerative algorithms. Accord-

ing to [2], this kind of algorithm finds high quality communities and it is based on modularity optimization: modularity index for clustering evaluation was deeply discussed by [3].

The algorithm consists of two phases, that are repeated alternatively. The starting point is the set of  $n$  nodes: it assigns a different membership community to each node of the network, so that in the initial partition there are as many communities as there are nodes. The first phase of the algorithm consists in considering each node  $i$  and its neighbors  $j$ : it computes the gain of modularity that would be obtained by removing the node  $i$  from its community and placing it into the community of neighbor  $j$ . After having evaluated all these gains for each node  $i$ , the algorithm places node  $i$  in the community for which the gain is maximum. This procedure is repeated for each node  $i$  in the set of nodes  $V$  until no further improvement can be achieved. The second phase of the algorithm consists in building a new graph whose nodes are the communities detected in phase 1. In order to achieve this goal, the algorithm firstly computes the weights of the edges between any two new nodes (i.e. any two communities identified in phase 1), by summing up all the weights of the links between nodes in the corresponding two communities.

After phase 2 is completed, the algorithm applies again phase 1 to the new network and to iterate. Hence, this type of algorithm is an agglomerative hierarchical procedure, as communities of communities are built during the process: the last community will be the one that involves all the nodes and aggregates all the communities detected in the previous step in only one.

Girvan-Newman algorithm [5] is one of the most known and used algorithm for communities detection problem. This algorithm is divisive and therefore it starts with the whole network and progressively cuts edges (most likely between communities) and reveals the community structure of the graph.

In order to find such edges, Girvan and Newman generalized the idea of node betweenness centrality, defining the edge betweenness centrality of an edge as the proportion of shortest paths connecting two vertices in the graph and passing through the edge. More formally, the edge betweenness centrality of edge  $e$  is

$$C_B(e_i) = \sum_{i \neq j \in V} \frac{\sigma_{jk}(e_i)}{\sigma_{jk}} \quad (1)$$

where  $\sigma_{jk}$  is the number of shortest paths connecting node  $v_j$  and node  $v_k$ , and  $\sigma_{jk}(e_i)$  is the number of shortest paths connecting node  $v_j$  and node  $v_k$  that run along edge  $e_i$ .

The algorithm proceeds as follows: first of all the edge betweenness for all edges in the network are computed; then, the edge with the highest betweenness is removed: indeed, such edge is an inter-communities edge, since all the paths linking any two nodes belonging to different communities go through it. Finally the algorithm computes again the betweenness for all the edges affected by the removal and repeats itself from the second step until no edges remain. As it is clear from the algorithm, the Girvan-Newman procedure belongs to the so-called divisive methods: indeed, it starts by taking into consideration the whole network  $G = (V, E)$ ; then, according

to the decreasing order of the edge betweenness, edges are cut progressively and therefore the whole network is splitted into smaller and smaller communities until we get  $n$  communities, as many as there are nodes.

These aforementioned algorithms were built such that they can be applied directly to a network object; it is of interest to notice that actually any clustering algorithm that takes as input a distance matrix  $\mathbf{D}$  can be applied, provided that  $\mathbf{D}$  can be built from the network  $G = (V, E)$ .

Clustering algorithms differ from each other also depending on the approach they take. More in details, the *hard* (or *crisp*) approach assigns any single object either to one cluster or to another one. Instead the *fuzzy* approach, introduced by [1], assigns to each object  $k$  membership degrees, one for each cluster. Each membership degree takes values in  $[0, 1]$ , instead of in  $\{0, 1\}$ , as occurs in hard approach, and it is such that the membership degrees of each unit sum up to 1.

Among fuzzy clustering algorithms, we focus on the Non-Euclidean Fuzzy Relational Clustering (NEFRC) algorithm, introduced by [4]. [4] proposed a fuzzy clustering algorithm, whose objective function is the following: let  $i$  and  $j$  identify units,  $i, j \in \{1, 2, \dots, n\}$  and  $c$  identify clusters, ranging in  $\{1, 2, \dots, k\}$ , where  $k$  is the desired number of clusters,

$$F_{NEFRC} = \sum_{c=1}^k \frac{\sum_{j=1}^n \sum_{i=1}^n u_{ic}^m u_{jc}^m d_{ji}}{2 \sum_{t=1}^n u_{tc}^m} \quad (2)$$

subject to constraints:

$$\sum_{c=1}^k u_{ic} = 1, \quad i = 1, 2, \dots, n \quad (3)$$

$$u_{ic} \geq 0 \quad i = 1, \dots, n \quad c = 1, 2, \dots, k \quad (4)$$

where  $m$  is *fuzzifier* or *fuzzyness parameter* that controls how fuzzy the clusters tend to be;  $u_{ic}$  is the membership degree of unit  $i$  to cluster  $c$ . Noteworthy that relational data in  $\mathbf{D}$  can be from any dissimilarity measure: indeed, most dissimilarity data are non-Euclidean and, as [6] showed, original relational fuzzy clustering methods that only require Euclidean distances often failed.

## 2.2 Distance Matrix

In order to apply the NEFRC algorithm, it is necessary to build distance matrix, having dimension  $n \times n$  and as generic element  $d_{ij}$  the distance between node labeled with  $i$  and node labeled with  $j$ ;  $\mathbf{D}$  must be symmetric (i.e.  $d_{ij} = d_{ji} \forall i, j = 1, \dots, n$ ), must have null diagonal (i.e.  $d_{ii} = 0 \forall i = 1, \dots, n$ ) and must have non-negative entries (i.e.  $d_{ij} \geq 0 \forall i, j = 1 \dots, n$ ). In order to build such matrix, the *geodesic*

distance, i.e. the length of the shortest path linking any two nodes, is used as measure of distance between the nodes: in this way, the higher the length of the shortest path between any two nodes, the more distant the nodes.

### 3 Application

The network object of our study is well known in literature and widely used for analysis; it belongs to the category of social networks. The network is called "Zachary karate club network" and it was downloaded from the network repository: it contains social ties among the members of a university karate club collected by Wayne Zachary in 1977. Each member of the club is represented by a node and each tie is represented by an edge. More in details, it is a unweighted, undirected graph, having  $|V| = n = 34$  and  $|E| = m = 77$ : so among the 34 club members there have been 77 bonds of friendship.

This network is well known in literature (see, for example, [9]), since it is of interest for detecting communities: indeed, an argument between the president and the instructor regarding some pay causes actually occurred and divided the group in two parts. The real clustering structure of the problem is therefore known and available (Figure 1 (d)) and hence it is possible to compare the obtained partition with the ground truth one, by using external validation indices to evaluate the performance of the methods used.

The results of the applications of the Girvan-Newman, Louvain and NEFRC algorithms are provided in Figure 1: it shows that the first two, i.e. the hard clustering algorithms, failed to recognize the clustering structure, since they identify four clusters instead of two; the fuzzy clustering algorithm, instead, detects exactly the true partition, leading to an Adjusted Rand Index ([8]) of 1.

### 4 Conclusion

This contribution aimed to review the most important and known clustering techniques that can be used in order to detect clusters of nodes and to suggest also the use of other clustering algorithms, that may be more successful, as occurred in our application. Other authors focused on applying fuzzy clustering algorithms to a network to detect the underlying community structure: we recall for example [9] that applied Non-Euclidean Relational Fuzzy C-Mean to the distance matrices resulting from the application of hard algorithms and [7] that applied Fuzzy c-Means to the spectral features extracted by spectral clustering from the graph.

Further developments regarding the application of clustering in a network framework is to consider a graph as a statistical unit and look for clusters of networks. This idea open up the field of possible questions regarding which measure of distance between networks can be used and all other instances related to it.

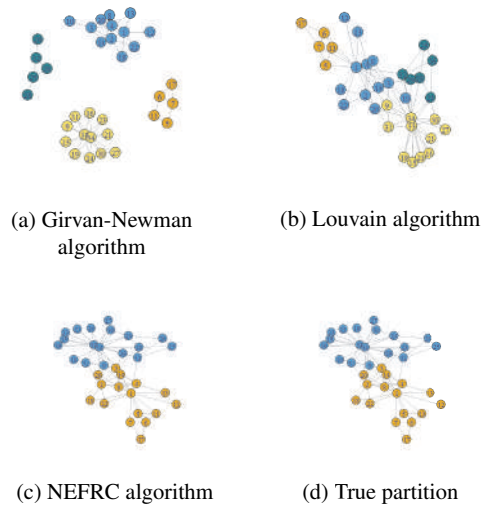


Fig. 1: Zachary Karate Network: Clustering results and true partition

## References

1. Bezdek, J.C.: Objective function clustering. In: Pattern recognition with fuzzy objective function algorithms, pp. 43–93. Springer (1981)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**(10) (2008)
3. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hofer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. *IEEE transactions on knowledge and data engineering* **20**(2), 172–188 (2007)
4. Davé, R.N., Sen, S.: Robust fuzzy clustering of relational data. *IEEE Transactions on Fuzzy Systems* **10**(6), 713–727 (2002)
5. Girvan, M., Newman, M.E.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**(cond-mat/0112110), 8271–8276 (2001)
6. Hathaway, R.J., Bezdek, J.C.: Nerf c-means: Non-euclidean relational fuzzy clustering. *Pattern recognition* **27**(3), 429–437 (1994)
7. Havens, T.C., Bezdek, J.C., Leckie, C., Chan, J., Liu, W., Bailey, J., Ramamohanarao, K., Palaniswami, M.: Clustering and visualization of fuzzy communities in social networks. In: 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–7. IEEE (2013)
8. Morey, L.C., Agresti, A.: The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement* **44**(1), 33–37 (1984)
9. Runkler, T.A., Ravindra, V.: Fuzzy graph clustering based on non-euclidean relational fuzzy c-means. In: 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15). Atlantis Press (2015)

## 4.26 Social demography

# Childcare among migrants: a comparison between Italy and France

## *La cura dei figli tra i migranti: un confronto tra Italia e Francia*

Eleonora Trappolini, Elisa Barbiano di Belgiojoso, Stefania Maria Lorenza Rimoldi, Laura Terzera<sup>1</sup>

**Abstract** Using the ‘Social Condition and Integration of Foreign Citizen’ survey (2011-2012) for Italy and the ‘Trajectoires et Origines’ survey (2008-2009) for France, we examined differences in the need for informal childcare and parental preferences in childcare between migrants in Italy and France according to family demographic and socio-economic characteristics, migrants’ background and contextual factors. Results suggest that migrants’ choices and preferences are strongly affected by the household composition, mother occupational status and migrants’ country of origin. Furthermore, our results highlight the role played by the destination country, which can shape migrants’ choices and options. The use of informal childcare is higher for migrants in Italy than in France, even among the same origin areas.

**Abstract** Utilizzando le indagini ‘Condizione e integrazione sociale dei cittadini stranieri’ (2011-2012) per l’Italia e ‘Trajectoires et Origines’ (2008-2009) per la Francia, si analizzano le differenze tra migranti nel ricorso alla cura informale, e per specifiche soluzioni di cura, considerando le caratteristiche demografiche e socio-economiche delle famiglie, il passato migratorio e fattori contestuali. I risultati suggeriscono che le scelte e le preferenze dei migranti sono influenzate dalla composizione familiare, dall’occupazione della madre e dal paese di origine. Si sottolinea anche il ruolo del paese di destinazione nel definire le scelte e le soluzioni di cura. Rispetto ai migranti in Francia, i migranti in Italia mostrano maggior ricorso alla cura informale, anche confrontando migranti dalla stessa area di origine.

**Key words:** childcare, migrants, France, Italy, welfare

---

<sup>1</sup> Eleonora Trappolini, University of Milan-Bicocca; email: [eleonora.trappolini@unimib.it](mailto:eleonora.trappolini@unimib.it)  
Elisa Barbiano di Belgiojoso, University of Milan-Bicocca; email: [elisa.barbiano@unimib.it](mailto:elisa.barbiano@unimib.it)  
Stefania M. L. Rimoldi, University of Milan-Bicocca; email: [stefania.rimoldi@unimib.it](mailto:stefania.rimoldi@unimib.it)  
Laura Terzera, University of Milan-Bicocca; email: [laura.terzera@unimib.it](mailto:laura.terzera@unimib.it)



## 1 Introduction

Care for children influences the life chances of parents, in particular in terms of female labour market participation, and families' fertility behaviours. This issue becomes even more challenging when parents are migrants and settle in a new context, because they are often deprived of the informal support network of family and friends (Bojarczuck and Mühlau, 2018). Most of the studies on childcare concern the overall population considering migrants as providers and not as potential consumers of such service (William and Gavanoas, 2008). The few studies available on migrants' childcare preferences suggest they are more likely to use informal than formal care (Barglowski and Pustulka, 2018; Bonizzoni, 2014). In addition, an important finding from the literature is the role played by migrants' country of origin in childcare preferences (Furfaro et al., 2020; Seibel and Hedegaard, 2017). However, it should be noted that childcare choice involves multiple elements, such as social protection availability, gender norms, social class, labour market and parental leave policies (Barglowski and Pustulka, 2018) which can limit or facilitate parents' preferences. Another important factor to be considered when analysing migrants' childcare preferences is the context in which the migration process occurred. The availability of childcare services varies prominently between countries (Rindfuss and Brauner-Otto, 2008), and in several countries informal childcare (especially provided by grandparents) plays a relevant role in balancing work and family responsibilities (Aasve et al., 2012; Arpino et al., 2014; Zamberletti et al., 2018). This study contributes to the understanding of the use of multiple sources of informal childcare (grandparents, other family and extra-family members) including child(ren) in pre-school and school age (0-11 years old), by comparing migrant parents' preferences in Italy and France. This comparative analyses represents an interesting case study because we assume that parental preferences may depend both on migrants' origin countries, thus on cultural and behavioural elements, and on the considered context and institutional setting. Women's labour market participation differs between France and Italy (Del Boca et al., 2005). In France, policies are more generally family-friendly respect to Italy. In this country, family policies have a long tradition (Thevenon, 2016) and public expenditures on families are among the highest of OECD countries<sup>2</sup> (3.9% of GDP in 2009). Conversely, in Italy, public support for families is very limited (1.6% of GDP in 2009) and policies are far from being family-friendly. Public childcare in France is extensive, and provides a great social support, while Italy, belonging to the Southern European welfare model, is strongly characterized by family support and, consequently, by a lack of public services (Esping-Andersen, 1990). The work-life balance is certainly more difficult in Italy, due to the limited supply of public childcare, in terms of availability and cost (Tanturri, 2016). In France, the enrolment rates in early childhood education and care services (0-2 and 3-5 years old) are 47.9% and 100%, respectively; while in Italy they are 23.3% and 97.4%. Looking at the use of informal childcare, Italy shows higher

---

<sup>2</sup> OECD average is 2.6% in 2009.

percentages with respect to France<sup>3</sup> (OECD, 2010). In this context, in countries where family support plays a key role in informal childcare, being migrant constitutes a further constraint on the choice of childcare arrangements, having a limited parental network, they are forced to opt for different types of informal childcare. This study has three aims: 1) to explore the determinants of migrants' childcare preferences in the contexts analysed; 2) to investigate the existence of differences by migrants' country of origin; 3) to explore differences in the need for informal childcare and parental preferences in childcare between migrants in Italy and France.

## 2 Data and Methods

We used two surveys conducted in Italy and France and we created a pooled dataset. For Italy we used the survey 'Social Condition and Integration of Foreign Citizen, 2011-2012' (hereafter SCIF), collected by Istat. It contains information on households with at least one foreign-born member. The French survey is 'Trajectoires et Origines, 2008-2009' (hereafter TeO), conducted by Ined-Insee, which contains information on households living in France, and we selected only immigrants<sup>4</sup>. We excluded the second generation due to its small sample size in the Italian survey. Since our study focuses on informal childcare among foreign-born households, and in the French survey information about childcare were asked only to households with at least one child aged 11 or less, in both surveys we restricted our sample to foreign-born households with at least one child aged 11 or less and at least one parent, dropping also mixed couples and single-parent families with only Italian or French children, because the Italian survey does not include childcare information on Italian children. The pooled sample consists of 4,885 households.

We employed some indicators to describe the context at the local level<sup>5</sup> (contextual factors): the availability of kindergarten for any 100 children in age 0-2; the percentage of immigrants; the activity rate and the percentages of unemployed. We used data referred at the time of the surveys published by Istat, Drees and Insee<sup>6</sup>.

We have four dependent variables. The first one is 'need of childcare' with 'no' reference category. We used other three variables that describe childcare choices derived from the question 'Who are the people your child is with when he/she is not

<sup>3</sup> France: 0-2 years old (17.7%), 3-5 years old (19.6%), and 6-12 years old (13.6%); Italy: 0-2 years old (31.5%), 3-5 years old (37.0%), and 6-12 years old (29.2%). *Source:* OECD family database <http://www.oecd.org/social/family/database.htm>.

<sup>4</sup> Despite the time-periods covered by the two surveys are not perfectly overlapping, they refer to two close time-periods, and they both refer to the first period of the financial crisis that in Italy showed its effects on migrants later compared to the overall population (Barbiano di Belgiojoso and Ortensi, 2013). In addition, they have been successfully used to compare the labour market integration in Italy and France (see e.g. Fellini and Guetto, 2019).

<sup>5</sup> The geographical aggregation used refers to NUTS1, for the years considered.

<sup>6</sup> To calculate the activity rate, percentage of unemployment and percentage of foreign-born individuals, we used data from Istat.it for Italy, and the Census data (Insee, 2009) for France. While the data for estimating the number of places in the kindergartens we used the data Istat (2013) for Italy and for France the data available by the Drees.

with his/her parents or at school?'. Due to the available options in the two surveys, we defined three dummy variables with each variable referred to a provider of childcare: grandparents, other relatives, and extra-family members.

We used as an independent variable the variable referred to the country coded 0 'Italy' (reference category) and 1 'France'. As control variables, we included four sets of factors:

- 1) Family demographic and socio-economic characteristics: the highest educational level in the family and the employment status of the parent(s);
- 2) Household composition: household arrangement; children aged 11 or more; and at least one child in preschool age;
- 3) Migrants' background: duration of stay as the years spent in the host country by the forerunner; and the area of origin;
- 4) Contextual factors at the local level as previously described.

We used logistic regression models separately for the four outcomes. We also estimated predicted probabilities to analyse and compare differences between migrant subgroups.

### 3 Results

Our results highlight some crucial points. Four elements shape migrants' choices and preferences: household composition, mother employment, country of origin and country of destination. Household composition strongly affects the choices: if available within the household, grandparents and older siblings are frequently the preferred option, with the age of the child(ren) to take care of driving this decision: grandparents in case of young children in preschool age, other relatives otherwise. Couples can share rearing duties as well as multi-family households, while single-parent families can rely on one parent. Thus, couples and multi-family households are less likely to need childcare, but more likely to use family childcare; while single-parent families are more likely to rely on extra-family solutions. Confirming previous results (Santero and Naldini, 2017), mother occupational status is crucial. An unemployed or inactive mother has as her main duty the rearing of her offspring, while an employed mother entrusts childcare to someone else to reconcile work and family (Bonizzoni, 2014). Empirical evidence unanimously shows that the country of origin, as a proxy of cultural norms and beliefs, has a crucial role in orienting childcare choices and preferences (Seibel and Hedegaard, 2017). Our results, while confirming this effect, highlight the effect of the country of destination in the use of childcare. Italian and French welfare states are particularly different, limiting or facilitating the options available to migrants. Unsurprisingly, the need for informal childcare is higher in Italy compared to France, due to limited availability of services for children in Italy. Moreover, for each area of origin we found a regular difference between the two countries: migrants from the same area have a higher childcare rate use in Italy compared to France.

**Table 1:** Logistic regression. Odds ratios and significance

<i>Variables</i>	<i>Childcare use</i>	<i>Grandparents</i>	<i>Other relatives</i>	<i>Extra-family</i>
<i>Country France (ref. Italy)</i>	0.375*	1.021	0.310*	0.490
<i>Household composition (ref. single-parent family)</i>				
Couple	0.505***	0.621*	0.874	0.594*
Multi-family household	1.107	3.450***	0.814	0.264***
<i>Length of stay (in years)</i>	1.009	1.048***	0.984	0.994
<i>Household occupational status (ref. both parents employed)</i>				
Mother unemployed or inactive	0.512***	0.575***	0.642***	0.763
Father unemployed or inactive	0.748	0.629	1.118	0.596
Both parents unemployed or inactive	0.663*	0.708	0.803	0.912
<i>Highest educational level in the household (ref. primary or none)</i>				
Secondary	0.912	1.273	0.719*	0.973
Tertiary	0.825	1.093	0.636*	0.997
<i>Scarce or insufficient household economic condition (ref. adequate or very good)</i>	0.836*	0.885	0.905	0.920
<i>Children pre-school age (ref. No)</i>	1.001	2.064***	0.677**	1.001
<i>Children aged 11 and over (ref. No)</i>	1.379**	0.604**	2.613***	0.565**

Source: Authors' elaboration on SCIF (2011-2012) and TeO (2008-2009) data.

The models control also for area of origin and context variables.

Legend: \* p<0.05; \*\* p<0.01; \*\*\*p<0.001.

**Table 2:** Predicted probabilities for childcare based on logistic regression models

	<i>Italy</i>	<i>France</i>	<i>Difference and significance</i>
East Europe	0.574	0.336	0.239*
Asia	0.506	0.277	0.229*
Morocco	0.475	0.253	0.222*
Other North Africa	0.497	0.270	0.227*
Sub-Saharan Africa	0.543	0.308	0.235*
Latin American	0.635	0.395	0.240*
Wester migrants	0.445	0.231	0.214*

Legend: \* p<0.05; \*\* p<0.01; \*\*\*p<0.001.

## 4 References

1. Aassve, A., Arpino, B., & Goisis, A. (2012). Grandparenting and mothers' labour force participation: A comparative analysis using the generations and gender survey. *Demogr. Res.*, 27, 53-84
2. Arpino, B., Pronzato, C. D., & Tavares, L. P. (2014). The effect of grandparental support on mothers' labour market participation: An instrumental variable approach. *Eur. J. Popul.*, 30(4): 369-390
3. Barbiano di Belgiojoso, E., & Ortensi, L. E. (2013). Should I stay or should I go? The case of Italy. *RIEDS*, 67(3/4), 31-38
4. Barglowski, K. & Pustulka, P. (2018). Tightening early childcare choices – gender and social class inequalities among Polish mothers in Germany and the UK. *Comp. Migr. Stud.*, 6(36)
5. Bojarczuk, S., & Mühlau, P. (2018). Mobilising social network support for childcare: The case of Polish migrant mothers in Dublin. *Soc. Netw.*, 53:101–110
6. Bonizzoni P. (2014). Immigrant Working Mothers Reconciling Work and Childcare: the Experience of Latin American and Eastern European Women in Milan, *Soc. Politics*, 1-14
7. Del Boca, D. D., Pasqua, S., & Pronzato, C. (2005). Fertility and employment in Italy, France, and the UK. *Labour*, 19, 51-77
8. Drees, L'offre d'accueil collectif des enfants de moins de trois ans en 2010. Document de travail, Série statistiques n° 174 – octobre 2012: [http://www.data.drees.sante.gouv.fr/ReportFolders/reportFolders.aspx?IF\\_ActivePath=P.330.331](http://www.data.drees.sante.gouv.fr/ReportFolders/reportFolders.aspx?IF_ActivePath=P.330.331)
9. Esping-Andersen, G. (1990). *The three worlds of welfare capitalism*. Cambridge: The Policy Press
10. Fellini, I., & Guetto, R. (2019). A “U-shaped” pattern of immigrants' occupational careers? A comparative analysis of Italy, Spain, and France. *Int. Migr. Rev.*, 53(1), 26-58
11. Furfaro, E., Rivellini, G. & Terzera, L. (2020). Social Support Networks for Childcare Among Foreign Women in Italy. *Soc. Indic. Res.*, 151: 181–204
12. Ined-Insee, Trajectories and Origins Survey, Ined-Insee, 2008. A Survey on Population Diversity In France User's Guide & Code Dictionary (2011), [https://teo-english.site.ined.fr/fichier/s\\_rubrique/20309/user.s.guide.and.code.dictionnaire.teo.fr.pdf](https://teo-english.site.ined.fr/fichier/s_rubrique/20309/user.s.guide.and.code.dictionnaire.teo.fr.pdf)
13. Insee, lil-0709: Recensement de la population (2009): tableaux détaillés, 2009
14. Istat, L'offerta comunale di asili nido e altri servizi socio-educativi per la prima infanzia. Anno scolastico 2011/2012. Statistiche Report (2013): <https://www.istat.it/it/archivio/96663>
15. Rindfuss, R. R., & Brauner-Otto, S. R. (2008). Institutions and the transition to adulthood: Implications for fertility tempo in low-fertility settings. *VYPR/Vienna Institute of Demography, Austrian Academy of Sciences*, 2008, 57
16. Santero, A., & Naldini, M. (2017). Migrant parents in Italy: gendered narratives on work/family balance. *J. Fam. Stud.*, 26(1), 126-141
17. Seibel V., & Hedegaard, T. F. (2017). Migrants' and natives' attitudes to formal childcare in the Netherlands, Denmark and Germany. *Child. Youth Serv. Rev.*, 78, 112-121
18. Tanturri, M. L. (2016). Aging Italy: Low fertility and societal rigidities. In M. Choe & R. Rindfuss (Eds.), *Low fertility, institutions, and their policies: Variations across developed countries* (pp. 221–257). Berlin: Springer
19. Thévenon, O. (2016). The influence of family policies on fertility in France: Lessons from the past and prospects for the future. In *Low Fertility, Institutions, and their Policies* (pp. 49-76). Springer, Cham
20. Williams, F., & Gavanoas, A. (2008). The intersection of childcare regimes and migration regimes: a three-country study. In H. Lutz (Ed.), *Migration and domestic work: A European perspective on a global theme* (pp. 13–28). Aldershot: Ashgate
21. Zamberletti, J., Cavrini, G., & Tomassini, C. (2018). Grandparents providing childcare in Italy. *Eur. J. Ageing*, 15(3): 265-275

# Employment Uncertainty and Fertility in Italy: The Role of Union Formation

## *Incertezza lavorativa e fecondità in Italia: il ruolo della formazione dell'unione*

Giammarco Alderotti, Valentina Tocchioni, Alessandra De Rose

**Abstract** The relationship between employment uncertainty and fertility is a prominent topic in demographic research. Evidence about Italy shows that men and women with precarious employment have fewer children and later. However, such evidence is outdated, with most recent studies based on data collected in 2009. Besides, the role of mediator that union formation may play in the relationship between employment uncertainty and fertility is largely neglected. With this work, we overcome the lack of a study with recent data about this topic for Italy, considering also union formation. Preliminary results suggest that employment uncertainty negatively affects fertility, especially among men, but when union formation is considered, the effect of employment uncertainty becomes much smaller.

**Abstract** *La relazione tra incertezza lavorativa e fecondità è un tema importante in demografia. Gli studi sull'Italia mostrano che uomini e donne con lavori precari hanno meno figli e li fanno più tardi. Tuttavia, questi studi non sono aggiornati, considerando che la maggior parte è realizzata con dati raccolti nel 2009. Inoltre, il ruolo giocato dalla formazione dell'unione all'interno della relazione tra incertezza lavorativa e fecondità è spesso trascurato. Con questo lavoro, ci proponiamo di sopperire alla mancanza di studi effettuati su dati recenti per l'Italia, tenendo conto anche dell'effetto di mediazione dell'unione. I primi risultati mostrano che l'incertezza lavorativa ha effetti negativi sulla fecondità, specie tra gli uomini, ma quando si considera anche lo stato di unione, tali effetti si riducono notevolmente.*

**Key words:** fertility, union formation, employment uncertainty, Italy

---

<sup>1</sup>

Giammarco Alderotti, University of Florence; email: [giammarco.alderotti@unifi.it](mailto:giammarco.alderotti@unifi.it)

Valentina Tocchioni, University of Florence; email: [valentina.tocchioni@unifi.it](mailto:valentina.tocchioni@unifi.it)

Alessandra De Rose, Sapienza University of Rome; email: [alessandra.derose@uniroma1.it](mailto:alessandra.derose@uniroma1.it)

## 1. Introduction

The relationship between employment uncertainty and fertility has become an increasingly important issue in demographic research. The notion of employment uncertainty refers to the lack of knowledge about what will happen in the labour market and the availability of a stable job or, indeed, any job to cover household expenses (Scherer 2009; Bloom 2014). Generally speaking, employment uncertainty is usually deemed to negatively impact fertility, with individuals in more precarious positions being more likely to postpone or forego parenthood (Blossfeld et al. 2006).

Theoretical premises suggest that individuals tend to postpone childbearing until more certain times, because childbearing is an irreversible event and people might wait to be sure about their income level before deciding to have a child (Ranjan 1999). Moreover, the deregulation of the labour market that started during the 1980s generated unprecedented levels of employment uncertainty, which became an inherent part of adults' life-courses in Western countries (Mills and Blossfeld 2013).

Empirical studies usually operationalise the forces of employment uncertainty through objective indicators of individuals' labour market situation, such as holding a temporary contract (e.g., Kreyenfeld et al. 2012). More specifically, employment uncertainty may stem from time-limited working contracts, because they are often connected to wage penalties and low employment protection (Schmitt 2012), low levels of labour market integration and little control over working hours (Pirani 2017).

The direction and magnitude of the relationship between employment uncertainty and fertility are still debated in the literature, because empirical evidence is highly context-dependent (see Alderotti et al. 2021 for a review). Whilst some studies did not identify any relationship (de Lange et al. 2014), others found a positive effect of employment uncertainty on fertility, mostly limited to countries with liberal labour markets (e.g., the United Kingdom). On the contrary, studies about Southern Europe point out that employment uncertainty is strongly detrimental for fertility (e.g. Barbieri et al. 2016).

In the Southern European context, Italy represents an intriguing case study because of the quick rise of uncertainty levels in the labour market and for the peculiarity of its family formation dynamics. From 1996 to 2016, the share of temporary employment among dependent workers grew from 7.0% to 14.0%, whereas the EU-28 average slightly moved from 11.5% to 14.0% (OECD 2021). Italy is also well-known for its *latest-late* transition to adulthood (Billari et al. 2002). With respect to their European counterparts, young Italians are among the oldest ones to complete their education, to enter the labour market, to leave the parental family, to start a union and to have a child – women's mean age at first birth is 31.3 years old in 2019 (De Rose et al. 2008; Istat 2020). This might (also) be due to the fact that Southern European youth and their families attach strong importance to employment stability as a prerequisite to start a family (Vignoli et al. 2020a).

A number of studies about the micro-level relationship between employment uncertainty and fertility in Italy are available; however, such evidence suffers from two main limitations. First, the timing: most recent studies are based on the survey

“Family and Social Subjects”, released by the Italian National Statistical Institute (Istat), which includes data collected in 2009 (e.g., Vignoli et al. 2020b; Busetta et al. 2019). Such studies are clearly outdated. The lack of recent evidence is particularly problematic in light of the extraordinary economic events that have taken place in the country after 2009 such as the Great Recession, started already in 2008, but which affected European countries most severely in the following years, or the crisis of the sovereign debt of 2012, which introduced further economic instability in Italy. Also, a substantial labour market reform occurred in 2015 – the so-called “Jobs Act” – which has essentially reduced the employment protection for large firm employees and left largely unchanged that for small-firm ones (De Paola and Scoppa 2020). The second limitation is that virtually no study about the relationship between employment uncertainty and fertility considers explicitly the role played by union status. However, fertility does not occur in isolation, but within relationships, especially in a country like Italy where *out-of-union* births are very scant. In Italy the share of children born outside marriage has increased over the last years (from about 20% in 2008 to about 34% in 2019; Istat 2020), but children born within marriages are still the large majority, and such share is larger than in most other European countries. Uncertain employment conditions might first jeopardise one’s chance to enter a stable union, and then affect childbearing. We advocate that, in order to have a comprehensive picture of how employment uncertainty affects fertility dynamics, union formation processes must be taken into account. A handful of studies faced this issue by analysing and modelling jointly the processes of union formation and transition to parenthood (e.g., Trimarchi and Van Bavel 2017), and proved that analysing the uncertainty-fertility link disregarding union status leads to incorrect results.

The aim of this work is twofold: first, by applying event history analysis we provide a timely evidence of how employment uncertainty relates to transition to parenthood by using the most up-to-date data available for Italy; second, we analyse the interrelationships between employment uncertainty, union formation and transition to parenthood, by looking at how the link between employment uncertainty and fertility changes when union status is included in our analysis.

## 2. Data and Methods

We use data from the 2016 survey of Household Multipurpose Survey Family, Social Subjects and Life Cycle (FSS), released by ISTAT. This survey includes detailed retrospective information on men’s and women’s partnership, employment and fertility histories, on a monthly basis. Our sample is formed by 5,609 men and 5,600 women aged 18-49 at interview (we excluded individuals with missing information on their childbearing or union history).

We employ event history analysis to study the transition to parenthood among childless men and women. We run gender-specific Cox models in order to investigate the relationship between employment uncertainty and fertility among men and women, and the role of mediator played by union formation. Since we are dealing with union formation and first childbirth, it is important to account for potential



pregnancies outside union, because some marriages and consensual unions might be a consequence of conceptions. Accordingly, in order to minimise possible reversed causation, the timing is specified as a transition to first-child conception. Individuals enter at risk of having a first child at age 16 and exit from the study at first child's conception turned into a live birth, at the interview date or at the age of 49, whichever occurred first. The main explanatory variable is the employment condition together with the type of job contract. Information on employment history allows to distinguish between unlimited-time contracts, time-limited contracts, self-employment and non-employment. The 'time-limited contracts' category includes both fixed-term jobs and 'atypical work', which identify the most precarious forms of employment (e.g., jobs on call, seasonal work). Unfortunately, it is not possible to distinguish between unemployed and inactive individuals. The mediation variable is the union status (single, cohabiting or married). Other control variables include macro-area of residence, time-varying educational level and calendar period.

### 3. Results

Table 1 shows model results on the transition to parenthood for men and women, separately, with and without controlling for union status. For the sake of brevity, we report information about the employment variable only (full models available upon request). The results clearly confirm that employment uncertainty is generally detrimental for the transition to parenthood in Italy, especially among men.

When not controlling for union status, men with time-limited contract or employed in atypical jobs have a hazard rate of conceiving a first child that is 27% lower than that of men with unlimited-time contract. Non-employment is even more detrimental for men's transition to fatherhood, with a relative risk reduced by 57%. Among women, time-limited contract only is related to a lower hazard of transition to motherhood, with a relative risk reduced by 13%. Self-employed women and especially non-employed women have a similar hazard rate of conceiving a first child as employed women (hazard ratios are not significant and very close to one).

Once introduced union status in the model, all hazard ratios become bigger (except one) and most of them lose significance, thus showing the mediating role played by union status in the relationship between employment uncertainty and fertility. More precisely, time-limited contract is not significantly associated to a lower risk of transition to parenthood both among men and women, while non-employment still reduces the risk of conceiving a first child among men, but to a smaller extent. On the other hand, being not-employed slightly increases by 6% the risk of conceiving a first child among women. Interestingly, self-employment is significantly associated to the lowest risk of conceiving a first child among women when union status is taken into account, thus suggesting that those women could be the ones who have more difficulties in conciliating their professional careers with childbearing.

**Table 1:** Cox model on the transition to parenthood. Gender-specific models with and without control for union status. Hazard ratios.

<i>Employment status</i>	<i>Without control for union</i>		<i>With control for union</i>	
	<i>Men</i>	<i>Women</i>	<i>Men</i>	<i>Women</i>
Ref.: unlimited-time contract				
time-limited contract	0.73 (<0.01)	0.87 (0.04)	0.90 (0.17)	0.96 (0.55)
self-employed	0.93 (0.15)	0.91 (0.20)	0.98 (0.78)	0.87 (0.04)
not employed	0.43 (<0.01)	0.99 (0.85)	0.63 (<0.01)	1.06 (0.114)

Note: p-value in brackets. Source: authors' elaboration on 2016 FSS data.

The models include control for macro-area of residence, educational level, calendar year, and (only in the second panel) union status.

#### 4. Discussion

We used the most recent data available for Italy to study the relationship between employment uncertainty and fertility, with a special focus on the role played by union status. These first results clearly indicate that considering union formation is fundamental when studying the relationship between employment uncertainty and fertility. Having an uncertain employment status seems to have a negative effect on childbearing among men and women, but once controlled for union status this effect markedly reduces: as a consequence, we may hypothesize that, in Italy, employment uncertainty has a detrimental effect on union status, whilst the direct effect of employment uncertainty on fertility is virtually null, and mainly restricted to men's non-employment.

These results clearly suggest – in line with our expectations – that uncertainty may select individuals into union before directly affecting childbearing. We claim that a deeper analysis of the mediating role of union status needs to be carried out for the Italian context. Accordingly, our next step will be performing a mediation analysis (e.g. Breen et al. 2013; Vignoli et al. 2020c) to assess the extent to which the effect of employment uncertainty on fertility is mediated by union formation processes, and if and to what extent a direct effect between employment uncertainty and fertility persists.

#### References

1. Alderotti, G., Vignoli, D., Baccini, M., & Matysiak, A.: Instability of employment careers and fertility in Europe: A meta-analysis. *Demography* (forthcoming), (2021)

Giammarco Alderotti, Valentina Tocchioni, Alessandra De Rose

2. Barbieri, P., Bozzon, R., Scherer, S., Grotti, R., & Lugo, M.: The rise of a Latin model? Family and fertility consequences of employment instability in Italy and Spain. *European Societies*, 17(4), 423–446 (2016)
3. Billari, F., Castiglioni, M., Castro Martin, T., Michielin, F., & Ongaro, F.: Household and union formation in a Mediterranean fashion: Italy and Spain, (2002)
4. Bloom, N: Fluctuations in Uncertainty. *Journal of Economic Perspectives*, 28(2), 153–176 (2014)
5. Blossfeld, H. P., Klijzing, E., Mills, M., & Kurz, K.: Globalization, uncertainty and youth in society: The losers in a globalizing world. Routledge (2006)
6. Breen, R., Karlson, K. B., & Holm, A.: Total, direct, and indirect effects in logit and probit models. *Sociological Methods & Research*, 42(2), 164-191, (2013)
7. Busetta, A., Mendola, D., & Vignoli, D.: Persistent joblessness and fertility intentions. *Demographic Research*, 40, 185-218, (2019)
8. De Lange, M., Wolbers, M. H., Gesthuizen, M., & Ultee, W. C.: The impact of macro-and micro-economic uncertainty on family formation in The Netherlands. *European Journal of Population*, 30(2), pp.161-185, (2014)
9. De Paola, M., & Scoppa, V.: Fertility Decisions and Employment Protection: The Unintended Consequences of the Italian Jobs Act, IZA Discussion Papers, No. 12991, (2020)
10. De Rose, A., Racioppi, F., & Zanatta, A. L.: Italy: Delayed adaptation of social institutions to changes in family behaviour. *Demographic research*, 19, 665-704, (2008)
11. Istat: Natalità e fecondità della popolazione residente. Anno 2019, (2020)
12. Kreyenfeld, M., Andersson, G., & Pailhé, A.: Economic uncertainty and family dynamics in Europe: Introduction. *Demographic Research*, 27, 835-852, (2012)
13. Kreyenfeld, M.: Uncertainties in female employment careers and the postponement of parenthood in Germany. *European sociological review*, 26(3), 351-366, (2010)
14. Mills, M., & Blossfeld, H. P.: The second demographic transition meets globalization: A comprehensive theory to understand changes in family formation in an era of rising uncertainty. In *Negotiating the life course* (pp. 9-33). Springer, Dordrecht., (2013)
15. OECD. Temporary employment (indicator) [electronic resource]. Paris: OECD, (2021)
16. Pirani, E.: On the relationship between atypical work (s) and mental health: New insights from the Italian case. *Social Indicators Research*, 130(1), 233-252, (2017)
17. Ranjan, P.: Fertility behaviour under income uncertainty. *European Journal of Population/Revue Européenne de Démographie*, 15(1), 25-43, (1999)
18. Scherer, S.: The social consequences of insecure jobs. *Social Indicators Research*, 93(3), pp. 527- 547 (2009)
19. Schmitt, C.: Labour market integration, occupational uncertainties, and fertility choices in Germany and the UK. *Demographic Research*, 26, 253-292, (2012)
20. Trimarchi, A., & Van Bavel, J. (2017). Education and the transition to fatherhood: The role of selection into union. *Demography*, 54(1), 119-144.
21. Vignoli, D., Guetto, R., Bazzani, G., Pirani, E., & Minello, A. (2020a). A reflection on economic uncertainty and fertility in Europe: The Narrative Framework. *Genus*, 76(1), 1-27.
22. Vignoli, D., Tocchioni, V., & Mattei, A.: The impact of job uncertainty on first-birth postponement. *Advances in Life Course Research*, 45, 100308, (2020b)
23. Vignoli, D., Mencarini, L., & Alderotti, G.: Is the effect of job uncertainty on fertility intentions channeled by subjective well-being?. *Advances in Life Course Research*, 46, 100343, (2020c)

## Acknowledgements:

The authors acknowledge the financial support provided by the Italian Ministry of University and Research under the 2017 MiUR-PRIN Grant Prot. N. 2017W5B55Y (“The Great Demographic Recession,” PI: Daniele Vignoli).

## **Determinants of union dissolution in Italy: Do children matter?**

### ***Determinanti della dissoluzione dell'unione: il ruolo dei figli***

Valentina Tocchioni, Daniele Vignoli, Eleonora Meli, Bruno Arpino

**Abstract** Cohabitation as a precursor or as an alternative to marriage has spread in middle- and high-income countries, and childbearing within cohabitation is increasingly common. Nonetheless, cohabitation remains less stable than marriage. Most studies concentrate on marital disruption in the presence of children, whilst the present paper focuses on the relation between childbearing and union dissolution in Italy, being it marriage or cohabitation. Results from the 2016 Families, Social Subjects and Life Cycle Survey suggest that the presence of children within a union has a binding effect for the couple – especially at higher parities, irrespective of whether the dissolution involves a marriage or a cohabitation.

**Abstract** *La convivenza come precorritrice o alternativa al matrimonio è sempre più diffusa nei paesi a medio e alto reddito, e la nascita di figli all'interno della convivenza è sempre più comune. Tuttavia, la convivenza rimane una forma di unione meno stabile del matrimonio. La maggior parte degli studi si concentra sul rischio di scioglimento dei matrimoni in base alla presenza di figli, mentre questo studio si concentra sulla relazione tra fecondità e scioglimento dell'unione in Italia, sia essa matrimonio o convivenza. I risultati basati sull'Indagine Multiscopo Famiglie, Soggetti sociali e ciclo di vita 2016 suggeriscono che la presenza di figli all'interno dell'unione ha un effetto saldante per la coppia, specialmente alle parità più alte, indipendentemente dal fatto che a sciogliersi sia un matrimonio o una convivenza.*

**Key words:** Union dissolution, childbearing, cohabitation, marriage, Italy

---

<sup>1</sup> Valentina Tocchioni, University of Florence; e-mail: [valentina.tocchioni@unifi.it](mailto:valentina.tocchioni@unifi.it)  
Daniele Vignoli, University of Florence; e-mail: [daniele.vignoli@unifi.it](mailto:daniele.vignoli@unifi.it)  
Eleonora Meli, Istat; e-mail: [elmeli@istat.it](mailto:elmeli@istat.it)  
Bruno Arpino, University of Florence; e-mail: [bruno.arpino@unifi.it](mailto:bruno.arpino@unifi.it)

## 1 Introduction

Notable changes in family-related behaviours started to emerge in Western and Northern Europe in the early 1960s [11]. Over the last decades, cohabitation rose whereas *direct* marriage (i.e., without a premarital cohabitation) decreased, and the number of marriages ending in divorce boomed. Next to family formation changes, Western and Northern European societies also faced a marked fertility decline since the late 1960s and early 1970s, with women having fewer children, and at later ages [6, 10].

In Italy, the diffusion of new family behaviours started with some delay compared to the rest of Western and Northern Europe [7]. Choosing the fertility decline as the onset of family formation practices, we can consider the mid-sixties as a starting point; other notable changes started in the early eighties, intensified in the nineties, and accelerated in the first decade of twenty-first century. The main processes observed during this period were the marked drop in the number of marriages, the increase in marital instability, and the rise of non-marital cohabitations.

On this backdrop, the majority of studies points out that cohabiting couples are more likely to terminate a union than married ones. Among the several demographic factors that might affect union stability, some are related to childbearing: whilst the presence of children is usually recognised to contribute to marital stability [1, 23], it is substantially ignored for the stability of cohabiting couples.

This study addresses the (in)stability of co-residential relationships, focusing on the effects of having children on union dissolution both for men and women, comparing marriage and cohabitation in Italy.

## 2 Union dissolution and childbearing within the union

Marital dissolution, and, more recently, union dissolution, have been extensively studied [15]. It is well known that cohabitation is a less stable union than marriage [4], whilst the majority of studies places *indirect* marriages some way in between cohabitations and marriages, in terms of union instability [3]. But such a difference in the “survival” of cohabitation and marriage is arguably weakening over time: Especially in countries where cohabitation is increasingly widespread, this form of union should become more normative, and childbearing within cohabitation popularized and institutionalized, making altogether cohabitation increasingly stable [22]. In this vein, cohabiting partners are more and more similar to married couples [17]. Consistently, childbearing within cohabitation has increased since the 1970s, and transitions to marriage among cohabitators have declined [9, 13].

In the socio-demographic literature, the relationship between childbearing and union dissolution has been analysed from different perspectives. Different theories explain this association through different explanations, but they all anticipate greater stability of unions with children, compared to those without children. The effect of childbearing may be direct or due to selection. Children represent a “union-specific

Determinants of union dissolution in Italy: Do children matter?

capital”, therefore contributing to union stability. According to the selection effect, family-oriented individuals are less likely to terminate their union and more likely to have children; conversely, less family-oriented partners are more likely to put an end to an unhappy union and less likely to have children [12].

Empirically, however, the impact of childbearing on union dissolution needs some clarifications. First, the union type matters. In the existing literature, the binding effect that children have on both marriages and cohabitations is well-recognized. For cohabitators, however, the relationship between fertility and separation is more difficult to be detected than for spouses, because data are often lacking, and selection plays an even stronger role. Non-married couples are only a few (and socially selected) in some countries, and common in others. The same holds for childbearing within cohabitation [18]. Cohabiting couples with children express high hopes that their relationships will last [8, 24], and experience higher emotional distress following separation than those without children [21]. Nevertheless, cohabitation usually emerges as less resilient than marriage, even when there are children involved [2, 20]. Couples who were cohabiting at the time of the birth of their child, and subsequently married, appear more stable than those who remained cohabiting [16, 25].

The influence of other features of childbearing on union dissolution have been primarily investigated for marriages, and usually ignored for cohabitations (and marriages preceded by cohabitation). For example, parity matters: for the United States, Lillard and Waite [14] found a lower risk of divorce associated with the first child, whereas higher order children had the opposite effect. The same result is confirmed for Denmark [19], but not in Italy and Spain, where second or higher order births apparently decrease the risk of union dissolution [5].

On this backdrop, we aim to answer the following research questions: *Does the presence of children affect the stability of cohabiting and married couples in a similar fashion? Does the number of children play a role in influencing their risk of union disruption? If yes, does this effect differ between marriage and cohabitation?*

### 3 Data and Methods

The analysis was carried out on retrospective data stemming from the 2016 survey “Families, Social Subjects and Life Cycle” (FSS), released by the Italian National Statistical Office (ISTAT). The timing of the survey and the information collected made it an ideal source for the study of marital and non-marital unions and dissolutions, as well as fertility careers of both genders: first, it occurred after two important normative changes that occurred in Italy, namely the Divorce law (L.55/2015) and the Civil Union law (L.76/2016). Second, it included detailed men’s and women’s partnership and childbearing histories recorded on a monthly basis. Our analytical sample consisted of 12,819 individuals aged 18-66 (whose 6,830 were women; we excluded individuals with missing information on their first union, on their childbearing history, or those having their first child before union formation).

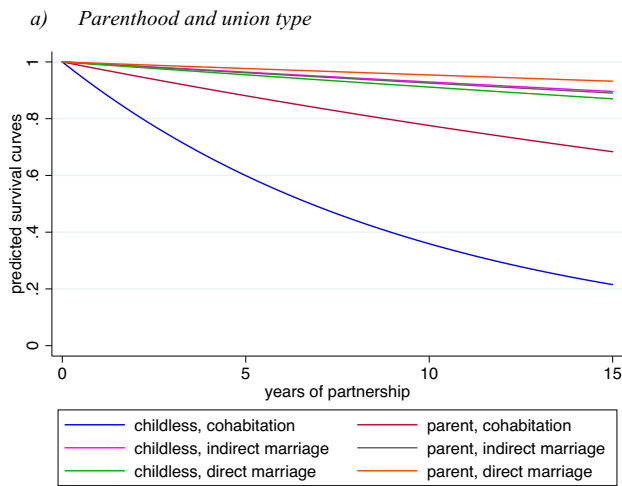
We employed piecewise-constant event history analysis (with knots at 3, 7, 10, 20, 35 years of partnership) to study the transition to first union dissolution. Individuals

entered at risk when their first union started and were observed until their disruption, at the interview date or at widowhood time<sup>1</sup>, whichever occurred first. The main explanatory variables were the presence and number of children (0, 1, 2 or more children) within the couple, and the type of union (direct marriage, indirect marriage and cohabitation). For answering our research questions, first we included an interaction between parenthood and the type of union to provide evidence of how the presence of children relates to transition to union disruption by union type; second, we included an interaction term between the number of children and union type. Control variables included gender, respondent's highest educational level, parental education, parental separation, partnership cohort, and macro-area of residence.

#### 4 Results

Figure 1 displays predicted survival curves of the transition to first union dissolution for our main explanatory variables (full models available upon request). Figure 1a) shows that childless cohabiting partners have the highest risk of disrupting their union, followed by cohabitators with children: after ten years of partnership, only 36% of cohabiting unions without children survives, in contrast with 78% of those with children. Instead, survival curves for marital dissolution of direct and indirect marriages are very similar, with a parenthood gap that seems to be less relevant, especially for indirect marriage that have nearly the same survival curves for childless couples or parents. Having more than one child has an even stronger binding effect for couples' stability, increasing the survival time for all types of union (Figure 1b).

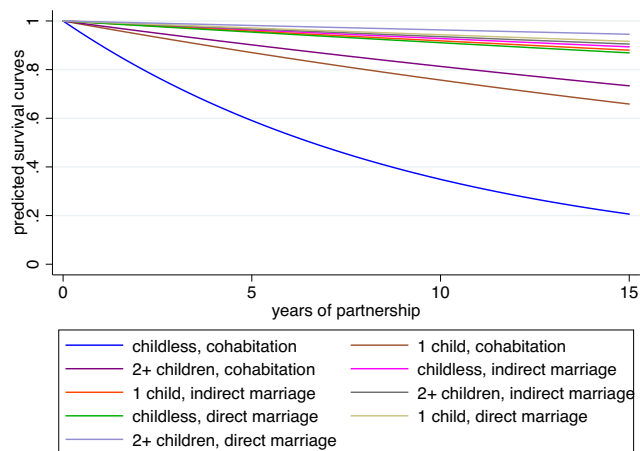
**Figure 1:** Predicted survival curves for the transition to union dissolution by parenthood/number of children and type of union.



<sup>1</sup> The information about the death of the partner was available for marriages only.

Determinants of union dissolution in Italy: Do children matter?

b) *Number of children and union type*



Source: authors' elaboration on FSS 2016 data. The models include control for the number of years of partnership (categorised in six periods), gender, highest educational level, parental education, parental separation, partnership cohort, macro-area of residence. Predicted survival curves are estimated according to the mean values of control covariates.

## 5 Discussion and Further Steps

Using the best data available for Italy, we study the relationship between childbearing and union disruption. Preliminary evidence clearly indicates that having child(ren) within a couple has a binding effect for partnership stability, lowering the risk of disruption both for cohabitation and marriage. Notably, the highest gap has been identified between childless/parent cohabitators. Cohabitors with child(ren) seem to be strikingly similar to spouses in their attitudes towards childbearing and union dissolution. This result is particularly relevant for Italy, a country all-to-often pitched as very traditional.

This study provides initial evidence of the relation between childbearing and union dissolution across marriage and cohabitation in Italy. Our next step will be adopting a causal inference approach to assess the extent to which the binding effect of children in a union may be driven by self-selection of people into the various types of partnership (cohabitation, indirect and direct marriage). Detecting changes over time of the associations highlighted in this study represents an additional important path for future investigations.

## References

1. Andersson, G.: The impact of children on divorce risks of Swedish women. *Eur. J. Popul.* 13(2),109-45 (1997).



2. Andersson, G.: Children's experience of family disruption and family formation: Evidence from 16 FFS countries. *Dem. Res.* 7(7),343-64 (2002).
3. Berrington, A., Diamond, I.: Marital dissolution among the 1958 British birth cohort: The role of cohabitation. *Popul. Stud.* 53(1),19-38, (1999).
4. Booth, A., Johnson, D.: Premarital Cohabitation and Marital Success." *J. Fam. Issues* 9(2),255-72 (1988).
5. Coppola, L., Di Cesare, M.C.: How fertility and union stability interact in shaping new family patterns in Italy and Spain. *Dem. Res.* 18, 117-44 (2008).
6. Frejka, T., Sobotka, T. Hoem, J.M., Toulemon, L.: Summary and general conclusions: Childbearing trends and Policies in Europe. *Dem. Res.* 19(2),5-14 (2008).
7. Gabrielli, G., Vignoli, D.: The breaking-down of marriage in Italy: Trends and trendsetters. *Popul. Review*, 52(1) (2013).
8. Gibson-Davis, C.M., Edin, K., McLanahan, S.: High hopes but even higher expectations: The retreat from marriage among low-income couples. *J. Marriage Fam.* 67,1301-1312 (2005).
9. Kennedy, S., Bumpass, L.: Cohabitation and children's living arrangements: New estimates from the United States." *Dem. Res.* 19, 1663-92 (2008).
10. Kohler, H. P., Billari, F.C., Ortega, J.A.: The emergence of lowest-low fertility in Europe during the 1990s. *Popul. Dev. Rev.* 28(4), 641-80 (2002).
11. Lesthaeghe, R.: The second demographic transition, 1986–2020: sub-replacement fertility and rising cohabitation—a global update. *Genus*, 76(1), 1-38 (2020).
12. Lesthaeghe, R., Moors, G.: Recent Trends in Fertility and Household Formation in the Industrialised World. *Rev. Popul. Social Policy* 9, 121-70 (2000).
13. Lichter, D. T., Qian, Z., Mellott, L.M.: Marriage or dissolution? Union transitions among poor cohabiting women." *Demography* 43(2), 223-40 (2006).
14. Lillard, L. A., Waite, L. J.: A joint model of marital childbearing and marital disruption. *Demography* 30(4), 653-81 (1993).
15. Lyngstad, T. H., Jalovaara, M.: A review of the antecedents of union dissolution." *Dem. Res.* 23, 257-91 (2010).
16. Manning, W.D., Smock, P.J., Majumdar, D. The relative stability of cohabiting and marital unions for children. *Popul. Res. Policy Rev.* 23(2), 135-59 (2004).
17. Pirani, E., Vignoli, D.: Changes in the satisfaction of cohabitators relative to spouses over time. *J. Marriage Fam.*, 78(3), 598-609 (2016).
18. Sobotka, T., Toulemon, L.: Changing family and partnership behaviour: Common trends and persistent diversity across Europe. *Dem. Res.* 19, 85-138 (2008).
19. Svarer, M., Verner, M.: Do Children Stabilize Danish Marriages? *J. Popul. Econ.* 21(2), 395-417 (2006).
20. Tach, L., Edin, K.: The compositional and institutional sources of union dissolution for married and unmarried parents in the United States. *Demography* 50(5), 1789-818 (2013).
21. Tavares, L. P., Aassve, A.: Psychological distress of marital and cohabitation breakups. *Soc. Sci. Res.* 42(6), 1599-611 (2013).
22. Van de Kaa, D.J.: Europe's Second Demographic Transition. *Popul. Bull.* 42, 1-57 (1987).
23. Waite, L. J., Lillard, L.A.: Children and Marital Disruption. *Amer. J. Sociol.* 96(4), 930-53 (1991).
24. Waller, M.R.: High hopes: Unwed parents' expectations about marriage. *Child. Youth Serv. Rev.* 23, 457–84 (2001).
25. Wu, L., Musick, K.: Stability of Marital and Cohabiting Unions Following a First Birth. *Popul. Res. Policy Rev.* 27(6), 713-27 (2008).

**Acknowledgements:** The authors acknowledge the financial support provided by the Italian Ministry of University and Research under the 2017 MiUR-PRIN Grant Prot. N. 2017W5B55Y ("The Great Demographic Recession," PI: Daniele Vignoli).

# Working schedules and fathers' time with children: A Sequence Analysis

## *Gli orari di lavoro e il tempo dei padri con i figli: un'analisi delle sequenze*

Annalisa Donno, Maria Letizia Tanturri

**Abstract** Mothers' labour market participation requires fathers to find new schemes for time allocation among paid work and fathering activities. This paper investigates whether typology of fathers' engagement depends on their time availability or on other characteristics. We carry out a Sequence Analysis on data from the 2008-09 Italian Time Use Survey, allowing to identify some "fathering profiles". Multinomial logit models are used to understand which factors influence the risk to be included in the groups identified. Five 'fathering profiles' are identified, strongly shaped by the father's work schedules. A cultural threshold of "minimum compulsory childcare" emerges: even when fathers are more available to children, they spend their time in scarcely engaged activities.

**Abstract** *L'aumento della partecipazione femminile al mercato del lavoro richiede che anche i padri adottino una diversa allocazione del loro tempo tra lavoro retribuito e attività con i figli. Questo lavoro intende verificare se il modo in cui i padri gestiscono il loro tempo con i figli dipende dalla loro disponibilità di tempo o piuttosto da altri fattori. Con i dati dell'Indagine ISTAT sull'Uso del Tempo 2008-09, tecniche di analisi delle sequenze ci hanno permesso di identificare 5 diversi "profili di uso del tempo con i figli", fortemente caratterizzati dal diverso impegno lavorativo dei padri. Tuttavia i risultati dei modelli multinomiali, mostrano che i padri che hanno più tempo per stare con i figli, tendono a trascorrerlo in attività poco impegnative e non di child-care vero e proprio.*

**Key words:** fathering, time use, working schedules, Sequence Analysis, cluster analysis.

---

<sup>1</sup> Annalisa Donno, University of Padova; email: donno@stat.unipd.it

Maria Letizia Tanturri, University of Padova; email: tanturri@stat.unipd.it

## 1 Introduction

The increase in the female labor force participation occurred since the 60s has given origin to a process of gender roles redefinition. Fathers are no more expected to be financial providers only, but rather to be actively engaged in parenting activities too. Such changes require fathers to find new schemes for the allocation of their time thus breaking with consolidate daily rhythms and social norms.

The hypothesis driving this paper is that fathers tend to schedule their time with children by following a ‘crowd effect’, that is, by performing parenting activities in a quite homogeneous way, influenced by the collective rhythm. We expect that father's roles, indeed, are driven mostly by the workplace rules, but also by cultural elements, like the social expectations. We thus want to answer the following research questions: 1) Which elements influence the way fathers allocate their time in parenting activities? 2) Who are the fathers who do not ‘follow the crowd’, thus showing untraditional schedules? 3) Do fathers’ commitment and engagement depend on their time availability?

Previous studies in this field analyze fathers’ time use in terms of durations mainly: i.e. the mean time devoted to different activities all along the day in a traditional time budget approach. We propose an original approach, by adopting a time reckoning system based on a chronological method simultaneously focusing on the *duration*, on the *timing* and on the *sequencing* of activities performed. We focus our analysis on Italy, a quite traditional country in terms of gender role-set where the male breadwinner model is still well rooted and the type of occupation remains a pivotal trait shaping men’s identity.

## 2 Data and methods

We rely on data from the Italian Time Use Survey carried out by the National Institute of Statistics (ISTAT) in 2008-2009. We select a sub-sample of 2,481 men who self-identified as a biological, adoptive, step or foster parent or guardian of at least one co-resident child aged 0-14 years. By using the daily activity diary of the survey we quantify not only the duration of different individual activities in a sample day, but also the specific sequence of activities. Indeed, the diary data are based on a grid of 10 minute-intervals of time, describing the main activity carried out by the respondent, the concurrent activity, and where and with whom the activity is carried out. Therefore, the time allocation scheme of each individual in the sample is a sequence of 144 ordered events (each lasting 10 minutes), describing in a detailed way how individuals spend their time all along the day. Aside from the diary, a data set containing rich information on the background and socio-economic situation of individuals and their households is available. In this paper, we take into account seven kind of activities: sleeping, eating, primary childcare, housework, leisure, work, and a residual category. The timing of fathering activities, is defined as doing

any of the previous activity in presence of children. Given the interesting structure of our data source, we use Sequence Analysis techniques, in order to identify homogeneous groups of fathers, according to their parenting time use patterns. We expect that most fathers tend to spend fathering activities at about the same time, and to show standardized behaviors.

Traditional Sequence Analysis techniques are based on the Optimal Matching Analysis algorithm (Abbott, (1995), Abbott and Hrycak, (1990)) that allows to compare sequences as a whole, to measure the degree of dissimilarity between two sequences, i.e. two sets of ordered events, and to transform sequences into distances between individuals which can then be clustered in order to uncover homogeneous patterns. Dissimilarity is conceptualized as the cost required to make identical two sequences with the help of three basic operations:

- Insertion, deletion (indel operations) – traditionally each indel operation costs one unit
- Substitution. The choice of substitution costs depends on the interpretation of replacing a state ( $a$ ) by another one ( $b$ ). If transitions don't have a meaning, the substitution cost is set to  $SC(a,b) = 2 - p(a,b) - p(b,a)$ ;

The dissimilarity produced by OMA is the minimum total cost required to match two sequences. Choosing the cost parameters represents the crucial point of Sequence Analysis applied to time use analysis and requires a tailor-made method. Specifically, we need to consider that the timing of the events (influenced by societal rhythms) are of paramount importance in the study of fathers daily schedules. It is not possible to separate activities from their temporal setting, and thus warp the temporal structure of the data. Three elements are important in this work, in order to analyze fathers' behaviors taking into account the social nature of time use: 1) the individual activity sequence; 2) its temporal setting; 3) the level of synchronization with the other fathers in the sample.

This is the reason why we use the Dynamic Hamming Approach (Lesnard, (2004)) to Sequence Analysis. Such an approach is based on the idea that the indel operations tend to separate events from their moment of occurrence since each indel operation has all the earmarks of inserting or deleting time, thereby warping the temporal structure. While, in our case, the cost system should be able to discriminate between two sequences which are quite similar from the point of view of the ordering of states but moved forward or put back in time, because this kind of shift is crucial in time use analysis. The solution is to use only substitution costs, and to derive them from time-varying observed transition between states. At each time point,  $t$ , the cost of substituting the state  $a$  with the state  $b$ , in order to transform one sequence in another one, is computed as follow:

$$s_t(a, b) = \begin{cases} 4 - [p(X_t = a|X_{t-1} = b) + p(X_t = b|X_{t-1} = a) + \\ p(X_{t+1} = a|X_t = b) + p(X_{t+1} = b|X_t = a)] & \text{if } a \neq b \\ 0 & \text{otherwise} \end{cases}$$

As a consequence, the distance at every moment between two individuals depends on what the entire population has done at the last stage and is about to do in the next one, which is a way to have both a dynamic and a relative definition of

which behaviour is common and uncommon (for instance, if two fathers (A and B) perform leisure activities with their children in two different moment of the day – father A at 11 a.m. and father B at 8 p.m. – the substitution cost for father A (to go from not being with children to perform leisure activity with them) will be higher, as 11 a.m. is quite a unusual time to have leisure with children, while father B will have a lower substitution cost as he performs such an activity in a moment of the day in which the proportion of fathers carrying out the same activity is very high).

In such a way each activity is assigned to a different meaning, depending on its temporal setting, and on the time patterning of all the other fathers, as substitution costs vary with the time and with the probability of transition between two states for the particular time considered.

Once the dissimilarity (distance) matrix has been computed, Cluster Analysis techniques (Ward's Method) are used to see if the sequences belong to a small number of distinct types. Such an approach will allow us to identify some "fathering profiles" and to differentiate between 'mainstream' and 'uncommon' childrearing scheduling. Multinomial logit models are then used to understand which factors influence the risk to be included in the groups identified.

### 3 Results

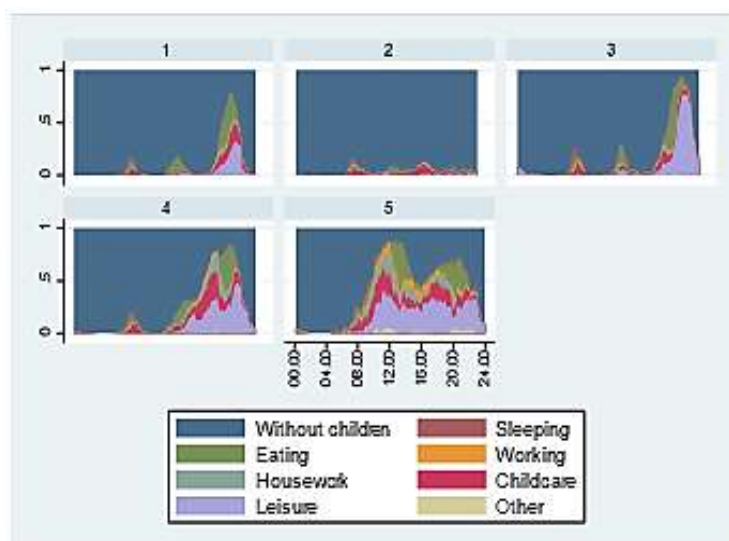
Five are the groups identified, by the sequence analysis. The corresponding five 'fathering chronograms' in Figure 1 report the proportion of fathers who spend time with their children, in each of the activities considered, in each moment of a 24 hours day. In order to understand the characteristics of fathers included in the five profiles', we run a multinomial logistic model, by taking into account two factors that are hypothesized to influence the allocation of fathers' time with children:

- *Fathers' work-related characteristics* (daily working hours and working schedules, evening/night/shift work), under the hypothesis that working long hours or with non-standard working schedules can be detrimental for the time spent with children, reducing the possibility to stay with them, or making them following non-standard fathering allocation schemes.
- *Partner's characteristics* (daily working hours, education, contribution to household income) in order to test both the *time availability* and the *relative resources theory*.

We also control for *fathers age and education*, *children characteristics* (age of the youngest child, number of children), *availability of external aid* (outsourcing of childcare or domestic chores).

Multinomial logistic regression results show that the way fathers allocate their time with children is mainly determined by their *work-related characteristics*. The children features seem to play a role, too. While, interestingly, the partners' work-related characteristics do not influence the 'fathering profiles'. Therefore, it seems that in Italy, roles are still predetermined by gender, and are not defined in relation to the relative earning power/time availability of each partner. A more detailed

Working schedules and fathers' time with children: A Sequence Analysis  
description of each groups, on the basis of both the Sequence Analysis (Figure 1)  
and multinomial regression results are presented in the following part.



**Figure 1:** Fathering typologies, Sequence Analysis results.

*Group 1: Full time workers, 'evening' fathers.* The 'modal' group (40% of fathers) is composed of fathers spending time with their children mainly in the evening hours. They share their time with children during meals (family time), and also perform leisure activities and primary childcare. Fathers working more than 10 hours/day, by following standard working schedules, with a partner working more than 10 hours/day, and with more than 1 child are more likely to belong to such a group. More likely in a dual earner couple, fathers in this group are those who, in relative terms, spend on average more time in childcare activities. They contribute to a greater extent to parental tasks in order to alleviate the 'dual burden' of their working partner.

*Group 2. Not available fathers* (15% of fathers in the sample) spend limited time with their children, but they mainly perform primary childcare activities. Fathers in this group are more likely to have non-standard working schedules (evening/shift workers) and to have more than one child. Given their scarce time availability, when present, they try to share 'quality' time with their children, by performing committing and interacting activities.

*Group 3. Leisure-mate fathers* (representing 17% of fathers in the sample) spend time with children in the late evening hours, mainly performing leisure activities, not necessarily implying a direct interaction and engagement with children. Fathers belonging to this group are more likely to work more than 10 hours/day and to have low education levels. Differently from fathers in group 1, showing a very similar time allocation scheme, a low proportion of childcare activities is registered in this

group. It is thus possible to hypothesize that fathers with lower educational levels are less concerned about the importance of performing childcare activities with their children, or more likely to behave by following traditional gender roles.

*Group 4. Part-time fathers* (22% of fathers in the sample) spend time with their children in the afternoon and evening hours. Fathers working less than 6 hours/day and shift workers are more likely to belong to this group. They are the most engaged fathers (together with fathers in the first group) as, relatively to their time availability, they perform the highest amount of primary childcare activities.

*Group 5. Full time Fathers* (6% of fathers in the sample) are available to their children all along the day, as they are more likely to work less than 6 hours/day or not to work at all. Even if they spend more time with their children, however, they do not show a greater level of engagement and involvement in primary childcare activities.

#### 4 Conclusions

Summing up, two main structures have emerged, shaping the fathers time allocation with children: the workplace-related schedules, mainly defining the fathers' availability during the week day, as well as some social norms, expecting fathers to spend time during the main family socializing moments, as meals or evening. Italian fathers' time use is strongly shaped by their workplace organization: work-related constraints to time with children seems to mainly determine their parenting patterns. However, even when fathers are more available to children, they spend their time in scarcely interactive activities. There seems to exist a threshold of 'minimum compulsory childcare', once it has been reached, fathers spend the rest of time with children in less demanding activities. It is possible to hypothesize the existence of gender display mechanism: those men whose participation to the labour market is scarce perceive themselves as deviant, and they do not increase in a consistent way their participation to childcare activities (as predicted under relative resources perspective) in an effort to reassert their masculinity in the face of their failure as good providers (deviance neutralization).

#### References

1. Abbott, A.: Sequence Analysis. *Annual Review of Sociology*, 21:93-113 (1995)
2. Abbott, A., Hrycak A.: Measuring Resemblance in Social Sequences. *American Journal of Sociology*, 96:144-185 (1990)
3. Lesnard, L.: Schedules as sequences: A new method to analyze the use of time based on collective rhythm with an application to the work arrangements of French dual-earner couples. *Electronic International Journal of Time Use Research*, 1:63-88 (2004)

# Correlates of the non-use of contraception among female university students in Italy

## *Il mancato uso della contraccezione tra le studentesse universitarie in Italia*

Busetta Annalisa, Alessandra De Rose and Daniele Vignoli

**Abstract** The present paper scrutinizes the correlates of the non-use of contraception among young women in Italy. To this end we consulted two releases of *Selfy (Sexual and Emotional LiFe of Youths) survey*, which offer information for a nationally sample of university students. The results reveal that the likelihood of non-using contraception is higher among women living in the South and among those with previous experience of unsafe sexual behaviours. The risk of unprotected sex is also higher within a group of students who are relatively older than the others and live in a stable cohabiting union.

**Abstract** *Questo paper studia il comportamento contraccettivo delle giovani donne in Italia sulla base di due indagini campionarie condotte su un campione nazionale di studenti universitari: Selfy (Sexual and Emotional LiFe of Youths). I risultati rivelano che la probabilità di non utilizzare la contraccezione è più alta tra le donne che vivono nel Sud Italia e tra quelle con precedenti esperienze di comportamenti sessuali non sicuri. Il rischio di rapporti sessuali non protetti è anche più alto tra le studentesse più grandi e che convivono con il partner.*

**Key words:** Contraception; University students; Risky behaviours; Low fertility; Selfy Survey; Ital

---

<sup>1</sup> Annalisa Busetta, Department of Economics, Business and Statistics (University of Palermo, Italy); e-mail: [annalisa.busetta@unipa.it](mailto:annalisa.busetta@unipa.it);

Alessandra De Rose, Department of Methods and Models for Economics, Territory and Finance (Sapienza University of Rome); e-mail: [alessandra.derose@uniroma1.it](mailto:alessandra.derose@uniroma1.it)

Daniele Vignoli, Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti" (DiSIA), University of Florence; email: [daniele.vignoli@unifi.it](mailto:daniele.vignoli@unifi.it)



## 1 Introduction and aim

Information about the contraception behaviours of Italian youths are scarce or outdated [2,3]. Despite the persistent use of “non-technological” contraceptive methods, Italian fertility declined to lowest-low fertility levels. The contraceptive behaviours of Italians remained very different from that of most European countries for a long time, and the “contraceptive revolution” is far to be completed. During the last decades, the sunset of *coitus interruptus* was very slow, the diffusion of the pill postponed, and the IUD (Intra-Uterine Device) never caught on. Within couples, withdrawal had been the most popular method until mid-1990s, when Italy became the country with the lowest fertility in the world with an average of 1.19 children per woman in 1996. We know very little about contraceptive behaviours in more recent years, and, above all, few information are available on the youngest people behaviour. Namely, an issue of concern is the persistency of a relative high frequency of unprotected sex which is often associated with other dangerous sexual behaviours [8,9] or with drug and alcohol abuse; as such, it has to be considered as a public health policy issue [4].

The present paper addresses this knowledge gap by exploring the correlates of the non-use of contraception among young women on a nationally representative survey of Italian university students (*Selfy - Sexual and Emotional LiFe of Youths survey*).

## 2 Data on contraception in Italy

According to the United Nations [11], a relevant share of Italian women in a union (37.3%) were not using modern methods<sup>1</sup> in 1996. This is in sharp contrast to other European countries: for instance, the value is 29.9% in Germany in 1992; or 25.5% in France in 1994; or 21.2% in Finland in 1992; or 18.9 in Spain in 1995. In 2013, the percentage declined in Italy to 34.9%, still much higher than in the majority of European countries: the percentage of people in couple who did not use any modern method is quite high (51.8%), and 13.3% were still relying traditional contraception.

Focusing on women exposed to the risk of a pregnancy – that is women aged 18-49 who had at least one sexual intercourse in the last 12 months and who declare to be not sterile, not pregnant and not in menopause – data from the Italian National Office of Statistics 2013 survey [7] revealed that the contraceptive prevalence raised to 76% (82% excluding women actively seeking for a pregnancy) and that the most used contraceptive method was the pill (27%), followed by condom (25%); still 20% relied on *coitus interruptus*. Among sexually active women but not in a cohabiting union, the percentage of not using any contraceptive methods is lower but not

---

<sup>1</sup> Modern methods include: female sterilization; male sterilization; IUD; implant; injectable; pill; male condom; female condom; vaginal barrier methods; lactational amenorrhea method (LAM); emergency contraception; other modern methods. Traditional methods include: rhythm; withdrawal; other traditional methods [11].

The non-use of contraception among Italian university students

negligible (14.4%), and the practice of withdrawal was as equally diffused as among cohabiting women. The no-use of any methods increases with age, reaching 31% among women aged 39-45. The younger segments of the population – aged 18-29 – show a relatively high share of (effective) contraceptive prevalence, with more than 70% using condom, but still 11% do not use any methods.

The nationally representative<sup>1</sup> survey of Italian university students (*Selfy*), reveals that the usage of modern contraceptive methods increased up to 77.3% in 2017 and the “traditional” methods (e.g., withdrawal and other natural methods, such as the billings ovulation method) reduced from 23.5% in 2000 to 15.1% in 2017, and the percentage of female students who did not use any method slightly increased from 4.5% in 2000 to 7.6% in 2017. This *Selfy* survey was carried out in the first half of 2017 on undergraduate students attending undergraduate courses in Economics and Statistics in 28 Italian public universities, and is almost identical to a survey carried out 17 years before [1]. The two sample involve 12,604 students (4,998 students in 2000 and 8,094 in 2017) who self-completed an anonymous<sup>2</sup> questionnaire during a one-hour lesson of a compulsory course. This process resulted in a practical nonexistence of refusals to fill out the questionnaire in class in both the surveys that, at the end, is representative<sup>3</sup> of the university students of the Italian undergraduate course in economics and statistics.

### 3 Correlates of non-use of contraception

Based on *Selfy* data, we focus on female students who had engaged in sexual intercourse at least once over the three months preceding the interview, and who declared themselves as being in a stable and intimate relationship. Our analytical sample comprised a total of 2,915 female students (N = 1,224 in 2000 and N = 1,691 in 2017). We explored several factors potentially associated to the non-use of contraception, as well as the reasons behind this choice. In particular, we estimated a logit model to predict the likelihood of non-use of contraception during the last sexual intercourse among female students. We include a set of covariates referring to key socio-demographic factors, i.e. age (from 17 to 26 ys); area of residence; year of the survey (2000 or 2017) and living arrangement at the time of interview (in the parental home, alone, with friends or in a residence, and with a partner). We also took into account students’ characteristics and lifestyles (i.e. diploma graduation mark; drug habits; importance attached to religion) and other covariates concerned

---

<sup>1</sup> For both 2000 and 2017, the data were post-stratified at the macro-regional level to obtain representative results of these university students at the national level [2,5].

<sup>2</sup> Students were reassured about anonymity and the use of the data (note that after completion, the questionnaires were sealed in an envelope by the students and all the envelopes were mailed to the directors of the survey for data entry).

<sup>3</sup> It is noteworthy to say that interviewing students in their first year of undergraduate studies minimizes the selectivity of future eventual dropouts, but affect the survey by limited external validity as the sample is not representative of the population of young Italians as a whole. Minello *et al.* [1] deeply discuss pro and cons of their sampling choices.

the sexual life of female students (i.e. the frequency of sexual activity in the last three months; having experienced first sexual intercourse with an occasional/casual partner; and having used contraception during the first sexual intercourse). Finally, we included a dummy variable indicating whether the student had experienced their parents' breakup.

The logit model (table 1) shows the profile of female students that have had unprotected sex: older women (24–26 years) and those cohabiting with their partner were more likely to have not used contraception at the last sexual intercourse compared, respectively, to younger women and those who lived in the parental home.

The results of our models show that students' non-use of contraception at first intercourse is strongly associated with a higher risk of not using any contraception at the last sexual intercourse. Conversely, having had sex with an occasional partner at the first experience is associated with a lower risk of not using contraception. The frequency of sexual intercourse is also significant: the higher the number of sexual intercourses during the previous three months, the higher the probability of the non-use of contraceptive methods. The risk of having unprotected sex is elevated among those living in the south of Italy and in the islands (i.e., Sicily and Sardinia). In 2017, the likelihood of not using contraception was significantly pronounced among female students, and the situation is slightly but significantly worse than in 2000. Interestingly, neither religiosity nor parental breakups display statistically precise estimates.

The non-use of contraception among Italian university students

**Table 1** The non-use of contraception at the last sexual intercourse: results of the logistic regression model (odds ratio).

<b>Y=1 Non-use of contraception at the last sexual intercourse</b>	<b>OR</b>
<b>Y=0 use of contraception</b>	
<b>Age (ref. 17-19 years)</b>	
20-21 years	0.855
22-23 years	0.894
24-26 years	2.011**
<b>Geographical Area (ref. North-West)</b>	
North-East	0.912
Centre	0.967
South	1.982**
Islands	2.623***
<b>Year of the survey 2017 (ref. 2000)</b>	1.036***
<b>Living arrangement (ref. living in parental home)</b>	
Alone, with friends or in a residence	1.089
With partner	1.616**
<b>Graduation mark (ref. 60-75)</b>	
76-89	0.999
90-100	0.856
<b>Drugs (ref. Never)</b>	
Sometimes	1.113
Often	1.376
<b>Religion is important (ref. Non important)</b>	0.902
<b>Frequency of sexual intercourses</b>	3.102***
<b>First sexual intercourse with occasional partner</b>	0.567*
<b>Non-use of contraception at 1<sup>st</sup> sexual intercourse</b>	12.11***
<b>Parents live together</b>	1.007
Wald chi2(19)	364.54
Pseudo R2	0.1825
Log pseudolikelihood	-567.86205
Observations	2,693

Note: Since not all the female students answered every questions considered in the model, the number of observations in the logit model is lower than in the descriptive analysis.

Significance level: \*\*\* p < 0.01, \*\* p < 0.05, \* p < 0.1

## 4 Discussion

The non-use of contraception is not an emergency in Italy, at least among University students. Less than 10% of sexually active people take no measures to avoid pregnancy or an STI, and the same holds for young adults. Having unprotected sex is however associated with other dangerous sexual behaviours [8,9].

Our results highlight the salience of several demographic and social correlates of the non-use of contraception among female Italian university students. We find traces of cumulative disadvantage over the life course with respect to the risk of contracting a sexually transmitted infection: our findings show that the non-usage of contraception at first intercourse is strongly associated with a higher risk of not using any contraception at the last sexual intercourse. This accords with prior

research. Indeed, multiple sexually-related health risk behaviours have been identified as a problem during early adulthood [4], especially among university students, an exceptionally high-risk category for sexual behaviour and reproductive health [6]. Some studies suggested that sexual risky behaviours are reciprocally associated among them. In particular, having a non-protected sexual debut, and with an occasional partner, and a higher number of partners over a person's lifetime are predictive of a risky behaviour later in life, that is an exposition to sexually transmitted infection, to the non-use of contraceptives and also to unwanted pregnancy [10].

## References

1. Minello, A., Caltabiano, M., Dalla-Zuanna, G., & Vignoli, D. (2020). Catching up! The sexual behaviour and opinions of Italian students (2000–2017). *Genus*, 76(1), 16. <https://doi.org/10.1186/s41118-020-00085-4>
2. De Rose, A., & Dalla Zuanna, G. (Eds.). (2013). *Rapporto sulla popolazione: sessualità e riproduzione nell'Italia contemporanea* [Report on population. Sexuality and reproduction health in contemporary Italy]. Il mulino.
3. Dalla Zuanna, G., De Rose, A., & Racioppi, F. (2005). Low fertility and limited diffusion of modern contraception in Italy during the second half of the twentieth century. *Journal of Population Research*, 22(1), 21–48. <https://doi.org/10.1007/BF03031802>
4. Bajos, N., Bozon, M., Beltzer, N., Laborde, C., Andro, A., Ferrand, M., Goulet, V., Laporte, A., Le Van, C., Leridon, H., Levinson, S., Razafindratsima, N., Toulemon, L., Warszawski, J., & Wellings, K. (2010). Changes in sexual behaviours: from secular trends to public health policies: *AIDS*, 24(8), 1185–1191. <https://doi.org/10.1097/QAD.0b013e328336ad52>
5. Billari, F. C., Caltabiano, M., & Dalla-Zuanna, G. (Eds.). (2007). *Sexual and affective behaviour of students. An international research*. Padua: Cleup.
6. Fromme, K., Corbin, W. R., & Kruse, M. I. (2008). Behavioral risks during the transition from high school to college. *Developmental psychology*, 44(5), 1497–1504. <https://doi.org/10.1037/a0012614>
7. Istat. (2017). *La salute riproduttiva della donna* [Women's reproductive Health].
8. Pirani, E., & Matera, C. (2020). Who is at risk? Gendered psychological correlates in Italian students' sexual risk profiles. *Genus*, 76(1), 11. <https://doi.org/10.1186/s41118-020-00080-9>
9. Potard, C., Lancelot, C., & Courtois, R. (2019). Examining Relationships Between Sexual Risk–Safety Behaviors and Physical Self-Concept by Gender: A Cluster Analytical Approach. *Emerging Adulthood*, 7(1), 31–44. <https://doi.org/10.1177/2167696817750802>
10. Tafuri, S., Martinelli, D., Germinario, C., & Prato, R. (2011). A study on the sexual and contraception behaviours of the pre-university students in Puglia (South-Italy). *Journal of preventive medicine and hygiene*, 52(4): 219–223. <https://core.ac.uk/download/pdf/234784989.pdf>
11. United Nations. (2018). *World Contraceptive Use 2018*.

## 4.27 Social indicators applications and methods

# A logistic regression model for predicting child language performance

## *Un modello di regressione logistica per la previsione dell'apprendimento del linguaggio nel bambino*

Andrea Briglia, Massimo Mucciardi and Giovanni Pirrotta

**Abstract** In this paper we propose a logistic regression model to evaluate how different components of language contribute to its acquisition over time. The empirical basis consists of a corpus which can be considered as a series of statistically representative samples taken at regular time intervals. The aim is to show how quantitative methods can contribute to solving linguistic puzzles.

**Abstract** *In questo articolo si propone un modello di regressione logistica per valutare come differenti componenti del linguaggio contribuiscono alla sua acquisizione nel corso del tempo. La base empirica consiste in un corpus, considerabile come una serie di campioni statisticamente rappresentativi presi ad intervalli di tempo regolari, l'obiettivo è mostrare come fenomeni apparentemente qualitativi possano essere spiegati con metodi quantitativi.*

**Key words:** Natural Language Processing; Logistic Regression; Phonetic Variation, Frequency Effects on Learning

---

<sup>1</sup> Andrea Briglia, Univ. of Montpellier "Paul Valéry"; email: [andrea.briglia@univ-montp3.fr](mailto:andrea.briglia@univ-montp3.fr)  
Massimo Mucciardi, Dep. of Cognitive Science, Univ. of Messina; email: [mucciard@unime.it](mailto:mucciard@unime.it);  
Giovanni Pirrotta, Univ. of Messina; email: [giovanni.pirrotta@unime.it](mailto:giovanni.pirrotta@unime.it)

## 1 Introduction

This paper is to be considered as a continuation of a previous research project [1] [4] in which the phonetic development of children was explored. In the current paper we have extended the level of analysis from a merely phonetic one to a more global view on how phonemes turn into words. The elementary units are Part Of Speech tags (from now POS tags). If phonemes could be metaphorically considered as the atoms of language, in a similar fashion words could be viewed as playing the role of the molecules: though the latter are far bigger than the former, they combine in different ways to form more complex meaningful entities in an analogous manner. Children always need to infer rules and regularities of their native language from a limited amount of input. This is the process task they need to follow:

« [...] discover the underlying structure of an immense system that contains tens of thousands of pieces, all generated by combining a small set of elements in various ways. These pieces, in turn, can be combined in an infinite number of ways, although only a subset of those combinations is actually correct. However, the subset that is correct is itself infinite. Somehow you must rapidly figure out the structure of this system so that you can use it appropriately early in your childhood» [6].

Learning a language means learning how to combine in a creative way a set of units that must respect a conventional order (*i.e.* phonotactic constraints and grammar): to reach this cognitive ability, children do not acquire their native language by simply repeating the input received because doing it this way would require much more time than it actually does. Children actively optimize the input received by trying to check whether it fits with adult language: English speaking children often pronounce an irregular verb conjugation by adding the regular “-ed” suffix on it. As “-ed” is the most frequent one, when you do not know how to conjugate a verb that you have never heard, the best option is to not take any risk and treat it as if it was part of the most common category. Another example, Italian speaking children do not need to hear all verbs listed in the “-are”, “-ere”, “-ire” forms to be sure how to conjugate the corresponding suffixes: once they know which rule applies to the different singular and plural forms of each person, they reach the ability to derive this rule even to verbs that they have never heard (that means the vast majority).

Modeling first language acquisition is a challenging scientific puzzle impossible to tackle in a short paper: what is at stake here is to propose a logistic regression model that has shown good performance in predicting grammatical development.

## 2 Data optimization strategy and statistical model

CoLaJE [3] is an open access French database part of the broader CHILDES project: seven children have been recorded in a natural setting one hour every month, from their first year of life approximately until five years of age. Data is



A logistic regression model for predicting child language performance available in three different formats: IPA, orthographic norm and CHAT (acronym for Code for the Human Analysis of Transcription), each of them is aligned to the correspondent video recording, allowing researchers to see the original source and to eventually reinterpret every utterance on their own. The main coding structure of the database consists in the fundamental division between “*pho*” (what the infant says) and “*mod*” (what the infant should have said according to the adult standard phonetic/phonological norm): we define every occurrence in which “*pho*” differs from “*mod*” as variation. The sampling scheme can influence the range of deductions and generalizations that we could draw from data: for this reason we check if this corpus sampling scheme meets internationally recognized reliability criteria [8] and it does: this means that it is considered as statistically representative in respect to the frequency of the linguistic structures targeted. We transformed all the sentences for the child named *Adrien* from 2 to 5 years old (8,000 sentences, 20,000 words approximately) to machine-readable strings of characters to make them computable by the Python STANZA library [5] software. STANZA library features a language-agnostic fully neural pipeline for text analysis, including tokenization, multiword token expansion, lemmatization, part-of-speech, morphological feature tagging, dependency parsing, and named entity recognition. We tagged all the words of the sentences by assigning a part-of-speech (POS) tag to each. In the second step, we calculated the “Word Phonetic Variation” (WPV) for each word by setting a specific algorithm to compute this difference. At this step, we assume that a correct word is a word that has been correctly pronounced though we are aware that -grammatically speaking- a word needs also to be pronounced in the correct place to be considered fully correct (*i.e* in the correct order). This is the case for the majority of sentences (especially shorter ones). This assumption is to be considered as acceptable because programming a set of grammar-sensitive and context-dependent algorithms is a hard challenge, especially for evolving linguistic structures such as those of children, besides the fact that every language has its own specific grammar. In other words, this analysis has been made *a priori* from the cardinality of the original sentences to which the words belong. We then organize the data in a spreadsheet structure on which we have built the statistical model.

From a statistical point of view, we developed a logistic regression model [2] to examine which factors can predict a child’s performance as part of our methodology. The binary variable was set as follows: WPV=1 if there is a phonetic variation in the spoken word and WPV=0 if there is no phonetic variation.

We choose 3 predictors to explain WPV: AGE, COMPLEX and CLASS.

-AGE is the main driver of development: the more a child is exposed to his environment, the more he will learn from it.

-COMPLEX relates to the difficulty a child has to get over long and semantically rich sentences where more cues need to be spotted.

-CLASS provides a way to evaluate whether a child can or cannot use a given grammatical element, giving an indirect measure of his grammatical development.

In particular:

- 1) AGE - It represents Adrien’s age from 1 to 5 year<sup>1</sup>;

---

<sup>1</sup> We transform the variable from months to years for a better representation of the data.

- 2) COMPLEX - It represents the type/token ratio, meaning the percentage of different or distinct words per sentence (*i.e* a proxy of lexical richness);
- 3) CLASS - Specifies the class to which the word belongs to: Open, Closed and Other. This framework of three classes has been taken from the Universal Dependencies project<sup>1</sup>.

The following “logit” equation (1) will give the probability of WPV based on the three regressors:

$$P(Y | \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 AGE + \beta_2 COMPLEX + \beta_3 CLASS)}} \quad (1)$$

Based on the equation analysis results, we can see (Table 1) which variable among AGE<sup>2</sup>, COMPLEX and CLASS variables is statistically significant. The likelihood ratio (LR) test is significant indicating that the logistic model provides a better fit to the data than the intercept-only model. Furthermore, with a cut-off = 0.5, overall correctly classified cases are equal to 72.2%. Odds ratios (OR) in Table 1 suggest that AGE is the main regressor: as it increases, the likelihood of reporting a higher WPV decreases consistently. COMPLEX works differently: an increase in lexical richness causes an increase in WPV too, but this relation becomes weaker over time, as it is confirmed by the graph in Figure 1.

CLASS is composed by three categories. “Open” contains lexical words such as verbs, common and proper nouns, adjectives. These classes contain large numbers of elements and are subject to change (a new entry can be added, another can be deleted). “Closed” contains functional words such as auxiliaries, pronouns and determiners. These classes contain few but highly occurring elements that are not subject to change (no new entries at all). “Other” contains everything that cannot be classified in the previous categories (punctuation, acronyms, etc). Figure 1 plots the WPV probability profiles with respect to the variable CLASS based on nine pre-defined scenarios. We consider a range for the AGE variable from 1 to 5 years while for the COMPLEX variable a (hypothetical) range from 0.1 to 0.9. To give an example “A1-C0.1” means one year of age, and a percentage of distinct words of 10% per sentence. “A” stands for age and “C” stands for complexity. It can be

---

<sup>1</sup> Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages [9].

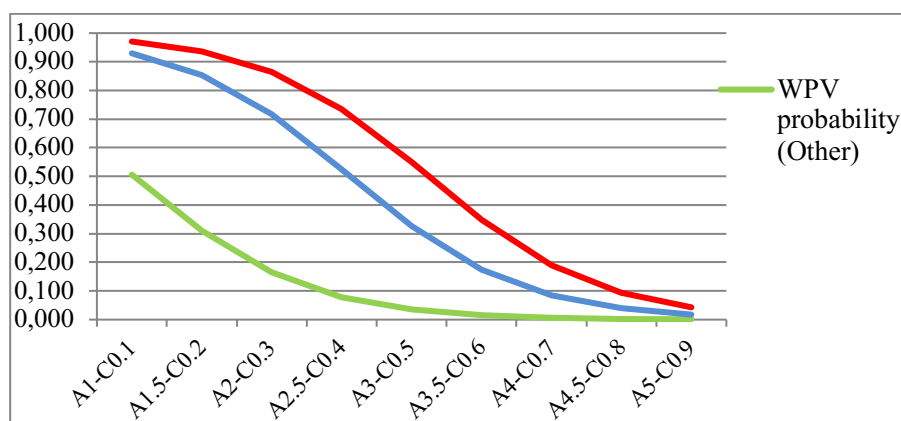
<sup>2</sup> Before modelling AGE as linear, we tried to create three successive yearly time slots to see how the two other regressors behave if taken apart, but the resulting correctly classified number of cases was lower than the model proposed. We then choose to model it as linear because first language acquisition is a highly non-linear phenomenon and the only certainty linguists have is that – roughly speaking – it develops in a cumulative way over time. We tried to model the interaction effects between COMPLEX and CLASS too, but it turned out to be less precise than the model proposed: in fact, COMPLEX showed a counterintuitive result in which its increase in value causes a decrease in WPV (models are available on request).

A logistic regression model for predicting child language performance observed how Open classes words are easier to learn compared to Closed class words: the difference between the two profiles shows an (almost) constant value of 0.2 up to 4 years old. When the child has almost completed his growth (after 4 years) the two profiles tend to be similar. This is because children are more at ease by naming things and persons with their names instead by using more abstract pronouns which impersonally refers to them, and because verbs are easier to put in a sentence rather than auxiliaries, whose place must respect precise grammar rules that requires time to be learned.

**Table 1:** Logistic regression estimates<sup>1</sup>

Variables in the equation	<i>B</i>	<i>SE</i>	<i>WALD</i>	<i>df</i>	<i>OR-Exp(B)</i>
AGE	-1.854*	0.033	3189.8	1	0.157
COMPLEX	1.007*	0.087	132.6	1	2.739
CLASS#			602.3	2	
Class (Open)	2.555*	0.374	46.6	1	12.875
Class (Closed)	3.478*	0.375	85.9	1	32.410
Constant	1.777*	0.389	20.8	1	5.914

# baseline CLASS = "Other" \*p<0.01- Overall percentage correct = 72.2%  
 Nagelkerke R Square = 0.29 - Initial -2 Log Likelihood =23846.685 - Final -2 Log Likelihood = 19560 - (LR test p<0.01) - Sample size=19093 words



**Figure 1:** Predicted probability according to 9 scenarios by CLASS category – In abscissa A = Age (1 – 5 years); C = Complex Index (range 0.1 – 0.9)

<sup>1</sup> All calculations are performed with STATA ver. 15

### 3 Conclusion and future directions

The logistic regression model could be a fair way to represent child language acquisition through quantitative and graphical tools. The predicted outcome is fairly good but needs to be improved by taking into account how the place a word occupies in the sentence structure influences the WPV. Attempts to create a model closer to child development, in which age is modelled in a non-linear way and in which the complexity of a sentence influences (and is influenced by) the grammatical elements contained in it turned out to be too difficult and unpredictable. These difficulties could be interpreted in the following way: as we do not know exactly how AGE influences learning (WPV) and how COMPLEX and CLASS interact with each other, it seems to be better to model these regressors in the simplest possible way (AGE as linear, COMPLEX and CLASS as independent from each other). Having said so, this can be true only at an initial stage of research: this statistical model should be applied to other similarly sampled children. By doing so, it would become possible to test the generalizability of the claims made in this paper and improve current knowledge on first language acquisition by comparing children between them and children learning similarly grammatical structured language between them [7]. A new research project on these themes is currently in progress, new tests on accounting for non-linear effects of age and the interaction between regressors will be made.

### References

1. Briglia, A., Mucciardi, M., Sauvage, J.: Identify the speech code through statistics: a data-driven approach, Book of Short Papers SIS (2020)
2. Hosmer, D., Lemeshow, S.: Applied logistic regression. New York: Wiley (1989)
3. Morgenstern, A., Parisse, C.: The Paris Corpus. French language studies 22. 7-12. Cambridge
4. Mucciardi, M., Pirrotta, G., Briglia, A.: EM Clustering method and first language acquisition. Workshop in Models and Learning for Clustering and Classification, Poster Session, Catania (2020)
5. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.J.: Stanza: a Python Natural Language Processing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020)
6. Saffran, J.: Statistical language learning: mechanisms and constraints. Current directions in Psychological Science. Vol.12 No 4. P 110-114. (2003)
7. Sekali, M.: First language acquisition of French grammar (from 10 months to 4 years old). French Language Studies 22, 1-6 . (2012)
8. Tomasello, M., Stahl, D.: Sampling children's spontaneous speech: How much is enough?. Journal of Child Language, 31:101-121. (2004)
9. UD (Universal Dependencies): Retrieved from <https://universaldependencies.org> (2021)

# Subject-specific measures of interrater agreement for ordinal scales

## *Misure di accordo su caso singolo per valutazioni espresse su scala ordinale*

Giuseppe Bove

**Abstract** Interrater agreement for ratings on ordinal scales is usually evaluated by overall measures across subjects like weighted Kappas for multiple raters or intraclass correlations. In this paper, a new index is presented that allows to evaluate the agreement between raters for each single case (subject or object), and to obtain also a global measure of the interrater agreement for the whole group of cases evaluated. The index is not affected by the possible concentration of ratings on a very small number of levels of the ordinal scale.

**Abstract** *L'accordo assoluto tra valutazioni espresse su scala ordinale viene solitamente valutato mediante misure globali di accordo come Kappa pesato per più valutatori o i coefficienti di correlazione intraclassa. In questo lavoro viene presentato un nuovo indice che consente di valutare l'accordo tra valutatori per ogni singolo caso (soggetto o oggetto), e di ottenere anche una misura globale dell'accordo tra i valutatori per l'intero gruppo di casi valutati. L'indice non risente della possibile concentrazione delle valutazioni su un numero molto ridotto di livelli della scala ordinale.*

**Key words:** ordinal rating scales, interrater agreement, educational assessment.

### Introduction

Classifying subjects or objects into predefined classes or categories is a rather common activity in many domains in the areas of education, psychology, sociology,

---

<sup>1</sup> Giuseppe Bove, Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre; email: giuseppe.bove@uniroma3.it

and medical research (e.g., Shoukri (2011), von Eye and Mun (2005)). For instance, the agreement of a group of raters who assess on a new rating scale (e.g., Likert scale) the language proficiency of a corpus of argumentative (written or oral) texts is analysed to test reliability of the scale. In medical studies, agreement between diagnoses provided by more than one doctor is considered for identifying the best treatment for the patient. In all these applications, the extent to which the categorizations of raters coincide, the rating procedure (or scale) can be used with confidence without worrying about which raters produced those categorizations. So, the main interest here is in analysing *interrater absolute agreement*, that is the extent that raters assign the same (or very similar) values on the rating scale.

Many of the measures of absolute agreement used in the case of ordinal scales are extensions of the Cohen's weighted Kappa index (e.g., Warrens, (2012)) or particular intraclass correlations coefficients (e.g., Mitani *et al.*, (2017)). They allow to analyse agreement between multiple raters for a whole group of subjects but not for a single subject. The possibility of having evaluations of the agreement on the single case is particularly useful, for example, in situations where the rating scale is being tested, and it is necessary to identify any changes to improve it, or to request the raters for a specific comparison on the single case in which agreement is poor.

There is a different kind of agreement that can be of interest in some applications, that is when we want to know which subject is the best, which one is the second, and so on. This is the agreement (or *consistency*) between the rankings of subjects provided by the raters, usually evaluated by measures of association or concordance. This second type of agreement is not addressed in this paper.

In the following, attention will be restricted to the case of evaluations expressed on an ordinal scale, in which each rater evaluates all the subjects (or objects) to be evaluated.

## Subject-specific measures of interrater agreement

A subject-specific measure of agreement was proposed in O'Connell and Dobson (1984) for several raters using nominal or ordinal categories. A small simulated example will show the features of the proposal. Let us suppose the level of agreement between seven judgments on a student's written text (provided in table 1) has to be evaluated. The judgements are the categories assigned by seven raters on a four-level ordinal scale, each category of the scale is defined by describing the features that the rater has to check in the text, the higher the level the better the evaluation of the written text.

According to O'Connell and Dobson, first we need to assign a score to each of the four categories of the scale to reflect the spacing between the categories (a common choice is to assign the first four naturals 1,2,3,4). Then, a disagreement function between the scores has to be defined in order to compare the ratings in each pair of raters (common choices are their Euclidean distance or its square). So, for the response profile in table 1, assuming the squared Euclidean distance between the scores (naturals numbers) as disagreement function, the disagreement between rater 1 and

Subject-specific measures of interrater agreement

rater 2 is 4, the disagreement between rater 1 and rater 3 is 0, and so on. The overall disagreement on the whole judgement profile of the student  $D_i$  is the sum of the disagreements obtained in this way for all the combinations of pairs of raters.

**Table 1:** Judgements of seven raters on a student written text (4-level scale)

<i>Student</i>	<i>Rater 1</i>	<i>Rater 2</i>	<i>Rater 3</i>	<i>Rater 4</i>	<i>Rater 5</i>	<i>Rater 6</i>	<i>Rater 7</i>
1	Level 4	Level 2	Level 4	Level 2	Level 3	Level 4	Level 3

O’Connell and Dobson propose a chance-corrected measure of agreement on a single subject  $i$ , given by

$$S_i = 1 - \frac{D_i}{\Delta}$$

where  $D_i$  is the overall disagreement on the whole response profile  $i$  and  $\Delta$  is the disagreement expected by chance (O’Connell and Dobson (1984), equation (6)). The measure takes the value 1 when there is perfect agreement; it is positive when the agreement is better than chance, and negative otherwise. Besides, an overall measure of agreement across subjects  $S_{av}$  can be obtained as the arithmetic average of the  $S_i$  individual values. The size of the agreement can be evaluated similarly to other Kappa-type measures, values lower than 0.6 are found in correspondence with low or moderate levels of agreement, values between 0.6 and 0.8 indicate a good level of agreement, values above 0.8 an excellent level of agreement.

The approach described has some drawbacks. Scores on the categories and a disagreement function (or weights) have to be defined for computing the disagreement between raters, this choice may vary in relation to the different contexts of application and the ability of the user. This can make it difficult to compare the results obtained in different applications. The index can not be computed for only one observation like that in table 1, because in that case the disagreement expected by chance  $\Delta$  is not defined. Besides, the disagreement expected by chance depends on the observed proportions of subjects allocated to the categories of the scale by each rater, and this imply that the measure of agreement depends on the marginal distributions of the categories of the scale observed for each rater (a limitation common to other Kappa-type measures).

Bove *et al.* (2018) follow a different approach to define a subject-specific measure of agreement for several raters using ordinal categories. Rather than working on the agreement between pairs of raters, it is proposed to work on a measure of the tendency to assume different levels in the frequency distribution of the seven raters with respect to the levels of the ordinal scale (table 2).

**Table 2:** Judgements of the seven raters respect to the levels of the scale

<i>Scale levels</i>	<i>Raters</i>
Level 2	2
Level 3	2
Level 4	3
<b>Total</b>	<b>7</b>

A measure of interrater agreement (in analogy with the measure of dispersion for ordinal categorical variables provided in Leti (1983)) is given below with reference to a number  $N$  of raters and to a  $K$ -levels ordinal scale:

$$\delta_i = 1 - \frac{2 \sum_{j=1}^{K-1} F_{ij}(1 - F_{ij})}{D_{max}}$$

where  $F_{ij}$  is the cumulative proportion associated with category level  $j$  of the scale in the response profile  $i$ , for  $j=1,2,\dots,K$ ,  $D_{max}$  is the maximum value of  $2 \sum_{j=1}^{K-1} F_{ij}(1 - F_{ij})$ , obtained when all raters are concentrated in the two extreme categories of the variable (maximum dispersion), and it is  $D_{max} = \binom{K-1}{2}$  as  $N$  is even, and  $D_{max} = \binom{K-1}{2} (1 - \frac{1}{N^2})$  as  $N$  is odd.

The index  $\delta_i$  is non-negative, it assumes a value of one in the case of maximum absolute agreement, in which all the evaluations coincide, and assume a value of zero in the case of maximum disagreement, in which the evaluations are equally distributed only on the two extreme levels of the scale. The values assumed by the  $\delta_i$  index can be interpreted similarly to other Kappa-type measure (i.e., lower than 0.6 low or moderate, between 0.6 and 0.8 good, above 0.8 excellent). For table 2, the value assumed by the index is  $\delta_i = 0.39$  which shows a low level of agreement between the seven raters.

In applications to a group of subjects (or objects),  $\delta_i$  allows to identify all the particular observations for which the agreement between the raters is low. Furthermore, a global measure of agreement on the whole group (indicated with  $\bar{\delta}$ ) can be easily obtained as the arithmetic average of the individual values  $\delta_i$ . An interesting property of  $\bar{\delta}$  is that it does not depend on the marginal distribution of raters (e.g., raters' tendency to classify the written texts in a very restricted number of levels). Besides, under certain conditions, it is possible to construct confidence intervals for the estimate of the population value of  $\bar{\delta}$ , without resorting to "computer intensive" resampling techniques (details on statistical aspects can be found in Bove *et al.* (2020), where the index is indicated by  $1-d$ ).

## Application

Data considered regard a research concerning the assessment of language proficiency, conducted at the Roma Tre University in 2017 (for details, see Nuzzo and Bove (2020)). The main aim of the study was to investigate the applicability of a six-point Likert scale developed by Kuiken and Vedder (2017) to texts produced by native and non-native writers, and to three different task types (narrative, instruction, and decision-making tasks). The scale comprises four subscales, corresponding to the four dimensions of functional adequacy identified by the authors of the scale: content, task requirements, comprehensibility, coherence and cohesion. Twenty native speakers of



Subject-specific measures of interrater agreement

Italian (L1) and twenty non-native speakers of Italian (L2) participated in the study as writers. All the texts produced by L1 and L2 writers (120 texts in total for the three tasks) were assessed by 7 native speakers of Italian on the Kuiken and Vedder's six-point Likert scale.

The interrater agreement between the seven raters for the comprehensibility subscale in the decision-making task will be analysed for the twenty L1 students, data are showed in columns labelled rater 1-rater 7 in table 3 (levels are reported by the first six natural). The values of  $S_i$  and  $\delta_i$  (last two columns of table 3) and some descriptive statistics (table 4) are also provided. The values of  $S_i$  are obtained by assuming the quadratic Euclidean disagreement function (quadratic weights).

**Table 3:** Ratings for the twenty L1 students on the comprehensibility dimension in the decision making task.  $S_i$  and  $\delta_i$  values in the last two columns.

L1 student	rater 1	rater 2	rater 3	rater 4	rater 5	rater 6	rater 7	$S_i$	$\delta_i$
1	5	5	5	6	4	5	4	0.48	0.73
2	3	3	3	3	3	4	3	0.84	0.9
3	5	5	4	4	4	3	5	0.38	0.7
4	3	4	4	5	4	6	4	0.02	0.63
5	5	3	4	5	4	4	5	0.38	0.7
6	4	6	4	5	4	5	5	0.38	0.7
7	3	4	4	4	4	4	4	0.84	0.9
8	5	6	4	6	4	6	6	0.02	0.63
9	5	6	6	6	4	6	5	0.33	0.7
10	5	6	6	5	4	4	5	0.27	0.67
11	5	4	6	4	4	5	4	0.33	0.7
12	3	2	2	3	3	3	3	0.74	0.83
13	5	5	4	5	4	5	5	0.74	0.83
14	5	4	5	6	4	5	5	0.48	0.73
15	4	5	4	4	4	3	4	0.64	0.8
16	5	6	4	6	4	5	6	0.12	0.63
17	5	6	5	6	4	5	5	0.48	0.73
18	4	4	5	4	4	4	4	0.84	0.9
19	5	5	4	4	4	4	5	0.69	0.8
20	5	6	5	6	5	6	6	0.69	0.8

**Table 4:** Some descriptive statistics for  $S_i$  and  $\delta_i$  values

	Mean	Std. Dev.	CV
$S_i$	<b>0.48</b>	<b>0.26</b>	<b>52.8</b>
$\delta_i$	<b>0.75</b>	<b>0.08</b>	<b>11.6</b>

The measures are strongly correlated ( $r=0.97$ ), however the  $S_i$  values show higher dispersion respect to the  $\delta_i$  values and seem difficult to interpret in some cases. For instance, to the very low values of  $S_i$  for student 4 (0.02) and student 8 (0.02) do not seem to correspond a very high disagreement (for the same students  $\delta_i=0.63$  indicates a moderately good agreement). These and other low values of  $S_i$  for other students generally do not seem to correspond to cases that need a revision by the raters.

The global agreement measured by  $S_{av}=0.48$  and  $\bar{\delta}=0.82$  reflects the differences already considered between  $S_i$  and  $\delta_i$  values. The low value of  $S_{av}$  may be a consequence of the strong concentration of judgments in levels 4 and 5 in the observed rater marginal distributions for the levels of the scale, especially for raters 3, 5 and 7. Besides, low levels of the scale are almost absent in table 3. These can also be the reasons for the very low value of the intraclass correlation coefficient  $ICC(A,1)=0.056$  (two-way random effects model, McGraw and Wong (1996)).

## Conclusions

According to the values of  $\delta_i$  and  $\bar{\delta}$  the scale developed by Kuiken and Vedder (2017) seems to be characterized by good levels of absolute agreement for L1 students, contrary to what suggested by  $S_i$  and  $S_{av}$  values. However, the lack of low evaluations deserves particular attention from language researchers. Future developments may concern sensitivity analyses to further assess properties and behaviours of the index with respect to different distributions. Finally, we notice that measures of interrater agreement for interval data proposed in organizational studies follow an approach similar to the present proposal (e.g., LeBreton and Senter (2008)).

## References

1. Bove, G., Nuzzo, E., Serafini, A.: Measurement of interrater agreement for the assessment of language proficiency. In: S. Capecchi, F. Di Iorio, R. Simone (Eds.), *ASMOD 2018: Proceedings of the Advanced Statistical Modelling for Ordinal Data conference* (pp. 61-68). Università di Napoli Federico II, FedOAPress, Napoli (2018).
2. Bove, G., Conti, P.L., Marella, D.: A measure of interrater absolute agreement for ordinal categorical data. *Statistical Methods & Applications* (2020) doi.org/10.1007/s10260-020-00551-5.
3. Kuiken, F., Vedder, I.: Functional adequacy in L2 writing. Towards a new rating scale. *Language Testing* (2017), 34(3), 321-336.
4. LeBreton J.M., Senter, J.L.: Answers to 20 questions about interrater reliability and interrater agreement, *Organizational Research Methods* (2008), 11 (4), pp. 815-852
5. Leti, G.: *Statistica descrittiva*. Il Mulino, Bologna (1983)
6. McGraw, K. O., Wong, S. P.: Forming inferences about some intraclass correlation coefficients. *Psychological Methods* (1996), 1(1), 30-46
7. Mitani, A.A., Freer, P.E., Nelson, K.P.: Summary measures of agreement and association between many raters' ordinal classifications. *Annals of Epidemiology* (2017), pp. 677-685
8. Nuzzo, E., Bove, G.: Assessing Functional Adequacy across Tasks: A Comparison of Learners' and Native Speakers' Written Texts. *EuroAmerican Journal of Applied Linguistics and Languages* (2020), 7 (2), 9-27.
9. O'Connell, D.L., Dobson, A.J.: General Observer-Agreement Measures on Individual Subjects and Groups of Subjects. *Biometrics* (1984), 40 (4), 973-983.
10. Shoukri, M. M.: *Measures of interobserver agreement and reliability*. Taylor and Francis Group, Boca Raton, Florida (2011)
11. von Eye, A., Mun, E. Y.: *Analyzing rater agreement. Manifest variable methods*. Lawrence Erlbaum Associates, Mahwah, New Jersey (2005)
12. Warrens, M., J.: Equivalences of weighted kappas for multiple raters. *Statistical Methodology* (2012), 9, 407-422

# **A Tucker3 method application on adjusted-PMRs for the study of work-related mortality**

## ***Applicazione della metodologia Tucker3 per lo studio della mortalità da esposizione occupazionale***

Vittoria Carolina Malpassuti, Vittoria La Serra, Stefania Massari

**Abstract** Principal Component Analysis is a widely used technique on two-way matrices for both dimensionality reduction and interpretation of latent relations among observed variables. The Tucker3 method is a generalization of PCA for three-way matrices, which not only runs classical PCA on each mode (way) of the data but also gives an estimate of the interrelation among the three modes. In the current analysis, the Tucker3 method is applied on data concerning mortality in the male population of Italy, in 2005-2015, specified for different causes of death, in people who have been working in different sectors and have had different levels of education; the main goal of this analysis is to understand the underlying relations among the three variables, in the considered population.

**Abstract** *L'analisi in Componenti Principali è una metodologia statistica molto utilizzata su matrici a due vie per la riduzione dimensionale e per ottenere informazioni circa le relazioni latenti tra le variabili osservate. Il metodo Tucker3 è una generalizzazione della PCA, utilizzabile su matrici a tre vie; questo permette di eseguire la PCA sulle tre dimensioni dei dati e allo stesso tempo fornisce una stima delle interrelazioni latenti che sussistono tra queste. Nella presente analisi, il metodo viene applicato su dati riguardanti la mortalità nella popolazione maschile in Italia, dal 2005 al 2015, registrata per diverse cause di morte, per individui che hanno operato in diversi settori lavorativi e che avevano diversi livelli d'istruzione. L'obiettivo principale dell'analisi è capire come queste tre variabili, nella popolazione in analisi, siano legate.*

**Key words:** Tucker3, PCA, PMR, mortality, working sectors, education

**Acknowledgements** We thank Italian Workers' Compensation Authority (INAIL) for providing us with mortality data in Italy in 2005-2015 and occupational INPS data.

---

Vittoria Carolina Malpassuti  
Sapienza Università di Roma, e-mail: vittoriacarolina.malpassuti@uniroma1.it

Vittoria La Serra  
Sapienza Università di Roma, e-mail: vittoria.laserra@uniroma1.it

Stefania Massari  
INAIL, e-mail: s.massari@inail.it

## 1 The data

Mortality due to occupational exposure still concerns a huge amount of workers. International Labour Organisation (ILO) in 2019 assessed around 2.4 million workers dying from work-related diseases every year in the world [1]. The identification of variables which help understand the etiology of causes of death among workers is essential to develop occupational health interventions.

The analyzed data in this study concern mortality in Italy, for the male population, in the range of years from 2005 to 2015, specified for three variables of interest: causes of death, economic activity and educational level, which is used as a proxy of the profession, which is lacking in the original dataset. The data include deaths on 1,787,570 men aged more than 20 years old, because younger people represent a small number in the considered working sectors. This study is part of the INAIL project (IAI-00032) included in the National Statistics Programme (PSN) 2020-2022 and occupational data were acquired thanks to a specific agreement between INPS and INAIL.

For privacy reasons, each combination of the three variables - cause of death, working sector and education - shall not include less than three units, therefore the considered classes verify this condition.

The causes of death have been reclassified from ICD-10 classification into thirty categories<sup>1</sup>, including different malignant tumors, such as lung cancer, stomach cancer, mesothelioma, etc. and other causes like respiratory system diseases and psychic disorders.

Economical activities have been grouped, too, from ATECO 81 classification into forty-nine sectors<sup>2</sup>, such as mineral extraction, chemical sector, agriculture, business sector, etc.

The people that have been included in the analysis have also been divided into three groups, with respect to their educational level: “low level” refers to people that have had no education or primary school diploma, for those being born before 1952, and people that have had no middle school diploma, for those being born after 1952; “medium level” refers to people that have a middle school diploma, if born before 1952, or a high school diploma, if born later, at most; “high level” refers to people having a degree or a higher qualification.

In the considered sample, 77.09% concerns men with a low level of education, 15.76% concerns men with a medium level and the remaining percentage refers to men with a high level.

Deaths are distributed among the different mortality causes with relative frequencies ranging between 0.9% and 4.4%; 77% of the causes has a frequency that is larger than 3%.

The considered working sectors present frequencies of observations ranging between 0.56% and 2.63%; the sectors with frequencies that are smaller than 1% are just four, out of forty-nine.

The age of death variable, observed on the sample, has been used as a covariate for the following computations; its mean value in the sample is 75 and it ranges between 20 and 108.

## 2 The methodology

### 2.1 PMR computation

An adjusted Proportional Mortality Rate (PMR) is applied as indicator. The “classical” PMR is commonly used in occupational epidemiological studies and it examines the pattern of mortality with respect to specific causes [2].

For each combination of cause  $i$  and sector  $j$ , if  $D_{ij}$  is the number of deaths associated to the combination,  $D_j$  is the number of deaths in the  $j$  sector and  $\bar{j}$  is “all sectors but  $j$ ”, the PMR is defined as:

<sup>1</sup> International Statistical Classification of Diseases and Related Health Problems, WHO.

<sup>2</sup> Economical Activities Classification, ISTAT.

A Tucker3 method application on adjusted-PMRs for the study of work-related mortality

$$PMR_{ij} = \frac{D_{ij}}{D_j} / \frac{D_{x\bar{j}}}{D_{\bar{j}}} \quad (1)$$

If the age covariate is available, PMRs can be estimated through a GLM procedure where this is used as an adjustment variable.

For a specific cause of death  $i$ , if  $Z_i$  is the dicotomical variable indicating death for this cause, we want to model  $Z_i$  as a function of the working sector, the age and a fixed offset for each sector  $j$ , obtained as the logarithm of the ratio between the number of deaths for cause  $i$  in  $\bar{j}$  and the total number of deaths in  $\bar{j}$ ; the used link function is a logarithmic one. Here is the model formula:

$$E[Z_i] = f(\text{sector} + \text{age} + \text{offset}) + \varepsilon \quad (2)$$

The estimated regression coefficients are the estimated adjusted-PMRs.

In the current analysis, this model was fitted for the three levels of education, separately, resulting in  $30 \times 49 \times 3$  coefficients, for each cause-sector-education combination. Dimensionality reduction is therefore needed.

## 2.2 Tucker3 method

When working with table matrices with large numbers of variables, dimensionality reduction can be of interest. A widely used methodology is Principal Component Analysis (PCA), which helps reducing dimensionality and also provides interpretability on the relations among units and among variables.

In epidemiological or medical studies, data for some set of units and variables can be replicated in different occasions; in such cases, data are represented as three-way tensors instead of 2-way matrices.

Reducing dimensionality and gaining interpretability can be done via simple PCA, but there are also tensor-based methods that can be useful for this purpose and can actually produce more information on the data than simple methods for two-way matrices.

A useful method for this purpose is the Tucker3 method [3]. If three modes (ways) are defined on the data, being  $A$  assuming values  $i \in [1 \dots I]$ ,  $B$  assuming values  $j \in [1 \dots J]$  and  $C$  assuming values  $k \in [1 \dots K]$ , the following steps are run in the Tucker3 methodology:

1. simple PCA is run on mode  $A$ : its  $[1 \dots I]$  dimensions are reduced by estimating a smaller and fixed number of principal components,  $P < I$ ;
2. simple PCA is run on mode  $B$ : its  $[1 \dots J]$  dimensions are reduced by estimating a smaller and fixed number of principal components,  $Q < J$ ;
3. simple PCA is run on mode  $C$ : its  $[1 \dots K]$  dimensions are reduced by estimating a smaller and fixed number of principal components,  $R < K$ ;
4. a "core" three-way tensor  $G$  of dimensions  $P \times Q \times R$  is built as an expression of the triple interaction among the three modes. Each element  $g_{pqr}$  of the core tensor explains the relation among the  $p^{th}$ ,  $q^{th}$ ,  $r^{th}$  components of modes  $A, B, C$ , respectively.

## 3 Application

### 3.1 Model fitting

Our data are collected in three matrices, one for each level of education; for each matrix, rows refer to the death causes and columns refer to the working sectors. In the matrix referring to a level of education  $k$ , for a fixed row  $i$  and a fixed column  $j$ , the element  $x_{ijk}$  is the adjusted-PMR of the  $i^{th}$

cause of death in the  $j^{\text{th}}$  sector. The array data is normalized with respect to the death causes, so that large values of the PMR for some causes do not influence the results.

The three modes in our data are: (A) the causes of death, (B) the working sectors and (C) the levels of education. We want to fit the Tucker3 model [4, 5] on the adjusted-PMR data tensor in order to reduce the dimensionality of the modes, especially for the causes of death and the working sectors and we also want to gain information on the interrelations that exist among these three variables.

For dimensionality reduction, the maximum allowed numbers of components for each mode are, respectively, 10 for the A-mode, 10 for the B-mode and 2 for the C-mode.

Each combination of different numbers of components shows a different percentage of fitting on the data. We want to choose the combination that verifies parsimony, so that not too many components are considered, but also shows a good fitting on the data. The chosen combination is: 6 components for the A-mode, 7 for B-mode and 2 for C-mode; the fitting percentage is 80.03%.

After choosing the combination, we run the ALS algorithm using a convergence criterion equal to  $10^{-6}$  and 5 random starts are considered, in order to limit the risk of attaining local optima.

In order to better interpret the solutions, we decide to rotate it. A good compromise for the simple structure of A, B, C and G was found, after many empirical tries, when  $w_A = 0$ ,  $w_B = 5$  and  $w_C = 5$  (where  $w_A$ ,  $w_B$  and  $w_C$  are the weights of rotation for A, B and C, respectively). This combination of weights gives us interpretable results for the computed principal components of the three modes and the core matrix, as shown in the following section.

### 3.2 Results

In each component-scores matrix, which can be observed in Tables 1, 2 and 3, we have used the scores that were larger than 0.15 in absolute value for providing the principal components with the following interpretations.

- The six components obtained for the causes of death (Table 1) can be interpreted as: A1) asbestos related diseases; A2) any cause of death; A3) accidental falls and other traumas; A4) silicosis; A5) nasal cavity and sinuses cancer; A6) neoplasm of nasopharynx, connective and soft tissue cancer and breast cancer and malignant neoplasm of eye vs silicosis and chronic obstructive pulmonary diseases.
- The seven components obtained for the working sectors (Table 2) can be interpreted as: B1) manufacture of glass; B2) manufacture of ceramic products, photographic activities, hairdressers, water treatments and machinery repair vs pottery and mining; B3) any working sector; B4) manufacture of leather and manufacture of wood; B5) building of sheeps and boats; B6) pottery, mining, textile cleaning vs manufacture of electrical equipment and office works; B7) forestry.
- The two components for the educational level (Table 3) are: C1) high level and C2) any level.

Interesting conclusions on the existing interrelations among the three modes can be drawn by looking at the core matrix in Table 4.

The scores for the combinations of components  $A4 \times B1 \times C1$  and  $A4 \times B1 \times C2$  are similar, showing that silicosis (A4) is correlated to the glass manufacture sector (B1) with the same strength in both high level (C1) and any level of education (C2).

A comparison can be made between the combinations  $A5 \times B4 \times C1$  and  $A5 \times B4 \times C2$ : nasal cavity and sinuses cancer (A5) show positive correlation with the leather manufacture and the wood manufacture sector (B4) for any level of education (C2) but have a negative correlation, with a similar strength, for high level of education (C1), meaning that positive correlation exists for low-medium levels of education.

The  $A2 \times B3 \times C2$  element of the core matrix, referring to any cause of death (A2), any working sector (B2) and any level of education (C2) has a very high value, which is probably due to unobserved factors; in fact, most of the causes of death are not directly related to the working sector, which is valid for all levels of education.

**Table 1** Components for causes of death

Causes of death	A1	A2	A3	A4	A5	A6
Other malignant tumors	-0.035	0.210	-0.061	-0.048	-0.012	-0.092
Neoplasm of nasopharynx	-0.062	0.134	-0.046	0.071	-0.100	0.513
Stomach cancer	-0.020	0.207	-0.061	-0.046	0.046	-0.075
Colorectal cancer	-0.037	0.204	0.099	-0.036	-0.023	-0.068
Liver cancer	-0.005	0.210	-0.045	0.036	-0.029	-0.049
Nasal cavity and sinuses cancer	-0.034	0.095	-0.020	0.002	0.964	0.176
Lung cancer	-0.015	0.211	-0.038	-0.027	-0.007	-0.089
Mesothelioma	0.516	0.148	-0.116	-0.009	-0.095	0.032
Connective and soft tissues cancer	0.062	0.170	-0.114	-0.003	-0.034	0.231
Breast cancer	0.046	0.133	-0.123	0.070	-0.162	0.422
Kidney cancer	-0.008	0.207	-0.097	-0.041	-0.037	-0.017
Malignant neoplasm of eye	-0.145	0.115	0.033	0.071	-0.062	0.526
Brain cancer	-0.027	0.203	-0.118	0.019	-0.026	-0.045
Neoplasm of lymphoid, haematopoietic and related tissue	-0.021	0.208	-0.106	-0.042	-0.018	-0.055
Mental and behavioural disorders	-0.024	0.203	-0.136	-0.041	0.011	-0.092
Diseases of the nervous system	-0.033	0.211	-0.103	-0.023	-0.013	-0.064
Disease of the circulatory system	-0.050	0.210	-0.004	-0.062	-0.004	-0.105
Diseases of the respiratory system	-0.047	0.209	0.071	0.005	0.001	-0.082
Chronic obstructive pulmonary diseases	-0.043	0.200	0.062	-0.002	0.016	-0.132
Asbestosis	0.817	0.056	0.059	-0.031	0.089	0.000
Silicosis	0.026	0.059	0.001	0.973	0.029	-0.160
Accidental falls	0.020	0.162	0.772	0.018	-0.043	0.082
Other traumas	-0.030	0.196	0.406	-0.025	-0.017	-0.044
Other causes	-0.033	0.212	0.027	-0.049	-0.005	-0.105

<sup>a</sup> Rows with component-scores that were all smaller than 0.2 in absolute value have been excluded from the table.

**Table 2** Components for working sectors

Working sectors	B1	B2	B3	B4	B5	B6	B7
Pottery	0.048	-0.267	0.152	0.024	-0.057	0.241	0.019
Manufacture of ceramic products	0.022	0.406	0.071	0.016	0.017	-0.020	0.041
Wholesale and retail trade	-0.030	-0.061	0.170	-0.011	-0.069	-0.250	-0.007
Manufacture of electrical equipment	-0.017	0.005	0.180	-0.198	-0.042	-0.422	0.021
Building of ships and boats	-0.014	0.000	0.115	0.002	0.924	0.003	0.007
Financial and insurance activity	-0.008	-0.022	0.162	-0.015	-0.058	-0.226	-0.009
Manufacture of leather and related products	0.009	0.024	0.126	0.733	0.008	0.027	-0.036
Mining and quarrying	0.108	-0.302	0.148	-0.052	-0.011	0.365	0.051
Forestry	-0.004	0.004	0.061	-0.021	0.006	0.002	0.933
Photographic activities	0.042	0.283	0.100	-0.022	-0.020	0.055	-0.029
Washing and dry-cleaning of textile and fur products	-0.135	-0.074	0.150	-0.152	-0.124	0.215	-0.027
Manufacture of wood	-0.040	-0.094	0.153	0.426	-0.030	-0.016	-0.027
Public administration and other services activities	-0.015	-0.009	0.158	-0.014	-0.061	-0.205	0.000
Hairdressers	0.006	0.252	0.094	0.006	-0.032	0.033	0.129
Water collection, treatment and supply	-0.039	0.208	0.116	-0.013	-0.029	0.125	-0.051
Repair and installation of machinery and equipment	0.033	0.414	0.076	0.030	-0.015	-0.024	0.025
Manufacture of basic metals	0.025	-0.130	0.170	-0.035	0.019	-0.243	-0.005
Manufacture of glass and glass products	0.937	0.002	0.101	0.001	-0.014	0.008	-0.004

<sup>b</sup> Rows with component-scores that were all smaller than 0.2 in absolute value have been excluded from the table.

**Table 3** Components for educational level

Educational Level	C1	C2
Level 1	-0.247	0.614
Level 2	-0.457	0.608
Level 3	0.855	0.502

**Table 4** Core matrix

	C1							C2						
	B1	B2	B3	B4	B5	B6	B7	B1	B2	B3	B4	B5	B6	B7
A1	0	0	-1	0	-3	0	1	0	-1	0	-2	12	1	-1
A2	0	-2	2	-1	0	-1	-2	3	0	55	2	3	0	2
A3	0	0	0	0	-1	-1	-3	0	-2	0	0	0	0	5
A4	6	2	-1	0	0	-2	0	7	-4	0	0	0	1	0
A5	0	0	0	-5	-1	-2	0	0	-1	0	6	0	-1	0
A6	-1	1	-4	-2	0	-1	1	-2	-4	0	2	0	-10	-1

The negative value of A6xB6xC2 shows an interesting phenomenon. The A6 component is highly and positively explained by neoplasm of nasopharynx, connective and soft tissue cancer, breast cancer and malignant neoplasm of eye, and it is highly and negatively explained by silicosis and chronic obstructive pulmonary diseases. The observed coefficient ( $-10$ ) shows that the adjusted-PMRs are higher in the second group of causes for the pottery, mining and textile cleaning sectors and higher in the first group of causes for the manufacture of electrical equipment and office jobs sectors.

This final result is interesting, since it has been explained in scientific literature; in fact: i) breast cancer in men is related to obesity which is linked to sedentary habits, such as working in offices [6]; ii) neoplasm of the eye may be linked to office jobs as well because this kind of workers spend much time in front of computer screens [7]; iii) neoplasm of nasopharynx can be related to inhaled chemical substances [8], which are present in the fields of electrical equipment manufacture.

## 4 Conclusions

These results have given us new information on the relations that lay among different working sectors and causes of death, for different educational levels. Some already known relations, in the scientific literature, have been confirmed; some new interesting ones have been discovered. The application can surely be improved by trying with different parameters in the Tucker3 method, but the obtained results are promising.

## References

1. Hämäläinen P, Takala J., Tan BK. Global Estimates of occupational accidents and fatal work-related diseases. XXI World Congress on Safety and Health at Work, Singapore. Workplace Safety and Health Institute. doi: 10.1002/ajim.20411
2. Wong O., Morgan R.W., Kheifets L., Larson S.R. Comparison of SMR, PMR, and PCMR in a cohort of union members potentially exposed to diesel exhaust emissions. *British journal of industrial medicine*. 1985; 42: 449-460.
3. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika*. 1966;31:279-311.
4. Maria Antonietta Del Ferraro, Henk A.L. Kiers, Paolo Giordani. Three-Way Component Analysis: Component analysis for three-way data arrays by means of Candecomp/Parafac, Tucker3, Tucker2 and Tucker1 models. 2015-09-07.
5. Giordani P., Kiers HAL. A review of tensor-based methods and their application to hospital care data. *Statistics in Medicine*. 2017;1–20. <https://doi.org/10.1002/sim.7514>
6. Lee K, Kruper L, Dieli-Conwright CM, Mortimer JE. The Impact of Obesity on Breast Cancer Diagnosis and Treatment. *Curr Oncol Rep*. 2019;21(5):41. Published 2019 Mar 27. doi:10.1007/s11912-019-0787-1
7. Di Cesare S, Maloney S, Fernandes BF, Martins C, Marshall JC, Antecka E, Odashiro AN, Dawson WW, Burnier MN Jr. The effect of blue light exposure in an ocular melanoma animal model. *J Exp Clin Cancer Res*. 2009 Apr 7;28(1):48. doi: 10.1186/1756-9966-28-48.
8. Armstrong RW, Imrey PB, Lye MS, Armstrong MJ, Yu MC, Sani S. Nasopharyngeal carcinoma in Malaysian Chinese: occupational exposures to particles, formaldehyde and heat. *Int J Epidemiol*. 2000 Dec;29(6):991-8. doi: 10.1093/ije/29.6.991. PMID: 11101539.



# Two case-mix adjusted indices for nursing home performance evaluation

## *Due indici per la valutazione della performance di case di riposo aggiustati per il case-mix*

Giorgio E. Montanari and Marco Doretti

**Abstract** Two indices for the performance evaluation of nursing homes are introduced. These indices properly take into account the case-mix, that is, the different complexity each nursing home copes with at baseline. Two estimators are consequently derived, whose finite-sample behaviors are studied in simulation.

**Abstract** *In questo lavoro si propongono due indici di valutazione della performance di case di riposo per anziani. Questi indici tengono conto del case-mix, ovvero della diversa complessità assistenziale in ingresso fronteggiata dalle case. Ne derivano due diversi stimatori, i cui comportamenti in campioni di ampiezza finita sono analizzati in uno studio di simulazione.*

**Key words:** binomial distribution, case-mix adjustment, mixed effect model, performance index, rate ratio

## 1 Introduction

In the last years, Nursing Home (NH) services have been receiving a growing attention due to population aging and, more recently, to the Covid-19 pandemic. In particular, the statistical evaluation of public health services is nowadays a well-established paradigm for both monitoring and improvement purposes across Western countries. Within this evaluation framework, a number of methods for defining quality standards have been developed. When the comparison of NH performances is of interest, it is well-known that adjustments are necessary to properly account

---

Giorgio E. Montanari  
University of Perugia, Department of Political Science, via A. Pascoli 20, 06123 Perugia (Italy)  
e-mail: giorgio.montanari@unipg.it

Marco Doretti  
University of Perugia, Department of Political Science, via A. Pascoli 20, 06123 Perugia (Italy)  
e-mail: marco.doretti@unipg.it

for the case-mix, that is, residents' clinical complexity each facility has to face at baseline. In light of this, a classification of residents based on Resources Utilization Groups (RUGs) has been introduced and subsequently refined [2], being now commonly adopted for public funding of care services. Such a classification system takes an economic perspective, in the sense that NH residents in the same RUG demand the same care burden.

In this paper, two case-mix adjusted performance indices are introduced that could be useful for NH comparison and ranking aims. These indices are based on residents' one-year ahead probability of death. This means that a framework is considered where a set of NH residents is observed at baseline and then followed up, recording whether or not they have died after one year. Specifically, we here assume a longitudinal rather than a cross-sectional perspective. However, since at baseline residents may be in quite different health conditions across NHs, performance measures based on the raw death probabilities in each NH are unfair. To overcome this problem, we propose to benchmark, for each RUG, the NH specific death probabilities with the marginal one. Case-mix adjustment is then obtained by averaging such RUG specific quantities with weights equal to the relative frequencies of residents across RUGs. In this way, synthetic indices are obtained.

In finite samples, death probabilities can be estimated by mortality rates observed at the one-year follow-up. To address the uncertainty deriving from additional factors, these mortality rates are taken as realizations from binomial random variables divided by the number of trials. In this setting, the two indices are associated to two estimators, whose approximate main moments are derived. The performance of these estimators is also assessed in simulation.

## 2 Two case-mix adjusted performance indices

### 2.1 Preliminaries

Let  $N_{hj}$  denote the number of residents in RUG  $j$  the  $h$ -th NH takes charge of at baseline, with  $j \in \mathcal{J} = \{1, \dots, J\}$  and  $h \in \mathcal{H} = \{1, \dots, H\}$ . Among these residents,  $K_{hj}$  denote those who are dead after one year, so that one estimator of the one-year ahead death probability,  $p_{hj}$ , is given by  $M_{hj} = K_{hj}/N_{hj}$ . Assume that  $K_{hj}$  can be regarded as a binomial random variable based on  $N_{hj}$  independent trials with a probability of success equal to  $p_{hj}$ . As mentioned above, in finite samples the estimate of such a probability is the observed mortality rate  $m_{hj} = k_{hj}/N_{hj}$ , with  $k_{hj}$  being the realized value of  $K_{hj}$ . Since NHs are treated like independently sampled units, counts/rates of different NHs can be assumed to be uncorrelated random variables. In what follows, we will also assume that counts/rates of different RUGs within the same NH are uncorrelated. In principle, such an assumption is more questionable and should be somehow checked in the data at hand. This is because measures of the same facility might be influenced by, say, accidental factors besides its overall quality level, that

we assume represented by  $p_{hj}$ . For each RUG, the variability in care quality levels reflects itself in the differentials between NH specific and marginal death probabilities.

Marginal death counts over NHs and RUGs are expressed by  $K_{.j} = \sum_{h=1}^H K_{hj}$  and  $K_h = \sum_{j=1}^J K_{hj}$ , respectively. The same notation applies to the number of residents, i.e.,  $N_{.j} = \sum_{h=1}^H N_{hj}$  and  $N_h = \sum_{j=1}^J N_{hj}$ , which leads to define the normalized weights  $W_{hj} = N_{hj}/N_h$  and  $Q_{hj} = N_{hj}/N_{.j}$ . Notice that  $K_{.j}$  corresponds to the sum of  $H$  independent binomial random variables with different success probabilities. A rather good approximation of its distribution is given by a  $\text{Bin}(N_{.j}, p_{.j})$  law, where  $p_{.j} = \sum_{h=1}^H Q_{hj} p_{hj}$  is the weighted average of the RUG specific success probabilities [1]. Clearly, the random variable indicating the relative marginal death rate is  $M_{.j} = K_{.j}/N_{.j}$ , with  $m_{.j}$  being its realized value in the sample.

### 2.2 The $\text{API}_h^{(1)}$ index

In this setting, for each NH the first Adjusted Performance Index (API) is defined as

$$\text{API}_h^{(1)} = \sum_{j \in \mathcal{J}_h^*} W_{hj} \frac{p_{hj}}{p_{.j}}, \tag{1}$$

where  $\mathcal{J}_h^* = \{j \in \mathcal{J} : N_{hj} > 0\}$  is the subset of RUGs for which the  $h$ -th NH hosts at least one resident. From (1) it is clear that case-mix is accounted for, since for each NH the RUG specific performance measures  $p_{hj}/p_{.j}$  are aggregated using the weights  $W_{hj}$ . This index fluctuates around 1, corresponding to a performance equal to the average overall performance of the NHs in the population: the lower its value, the better the performance with respect to the one-year ahead mortality rate.

The index in (1) can be estimated by

$$\widehat{\text{API}}_h^{(1)} = \sum_{j \in \mathcal{J}_h^*} W_{hj} \frac{M_{hj}}{M_{.j}}. \tag{2}$$

Performing the first-order Taylor linearization of the rate ratio  $M_{hj}/M_{.j}$  in the pair  $(p_{hj}, p_{.j})$  leads to conclude, after some algebra, that (2) is a consistent and asymptotically (i.e., for  $N_{hj} \rightarrow \infty$ ) unbiased estimator of (1), with an asymptotic variance, under the within-NH independence assumption, equal to

$$V(\widehat{\text{API}}_h^{(1)}) \approx \sum_{j \in \mathcal{J}_h^*} W_{hj}^2 \left\{ \frac{p_{hj}(1-p_{hj})}{N_{hj} p_{.j}^2} \left( 1 - 2Q_{hj} \frac{p_{hj}}{p_{.j}} \right) + \frac{p_{hj}^2(1-p_{.j})}{N_{.j} p_{.j}^3} \right\}. \tag{3}$$

The approximation in the expression above follows from  $K_{.j} \approx \text{Bin}(N_{.j}, p_{.j})$  and from  $\text{Cov}(M_{hj}, M_{.j}) = Q_{hj} V(M_{hj})$ . This variance can be consistently estimated by replacing  $p_{hj}$  and  $p_{.j}$  by  $m_{hj}$  and  $m_{.j}$ , respectively. However, it can be shown that an

alternative estimator using  $N_{hj} - 1$  and  $N_{.j} - 1$  in place of  $N_{hj}$  and  $N_{.j}$  is, while still consistent, less prone to bias in finite samples.

### 2.3 The $\widehat{\text{API}}_h^{(2)}$ index

The  $\widehat{\text{API}}_h^{(1)}$  estimator cannot be computed when  $m_{.j} = 0$  for some  $j \in \mathcal{J}_h^*$ . A viable alternative is represented by the index

$$\widehat{\text{API}}_h^{(2)} = \sum_{j \in \mathcal{J}_h^*} W_{hj} \log \frac{1 + p_{hj}}{1 + p_{.j}}, \tag{4}$$

which operates on the logarithmic scale. Such an index fluctuates around 0, which, again, corresponds to an average overall performance. Like in (1), lower values denote better performances with respect to the one-year ahead mortality rate. The corresponding estimator is

$$\widehat{\text{API}}_h^{(2)} = \sum_{j \in \mathcal{J}_h^*} W_{hj} \log \frac{1 + M_{hj}}{1 + M_{.j}}, \tag{5}$$

whose approximate moments can also be derived by expanding the inner term  $\log\{(1 + M_{hj})/(1 + M_{.j})\}$  in the point  $(p_{hj}, p_{.j})$  and exploiting the same assumptions as in Section 2.2. It follows that (5) is a consistent and asymptotically unbiased estimator of the index in (4), with an asymptotic variance given by

$$V\left(\widehat{\text{API}}_h^{(2)}\right) \approx \sum_{j \in \mathcal{J}_h^*} W_{hj}^2 \left\{ \frac{p_{hj}(1 - p_{hj})}{N_{hj}(1 + p_{hj})^2} \left( 1 - 2Q_{hj} \frac{1 + p_{hj}}{1 + p_{.j}} \right) + \frac{p_{.j}(1 - p_{.j})}{N_{.j}(1 + p_{.j})^2} \right\}. \tag{6}$$

For the estimation of this variance, the same considerations of Section 2.2 hold.

## 3 Simulation study

As mentioned in Section 1, a simulation study is performed to evaluate the finite-sample behavior of the estimators of the two performance indices as well as of the estimators of their variability measures. In particular, we present a study designed from a real dataset concerning  $H = 47$  NHs based in Umbria, a region of central Italy. These NHs host 1748 residents divided in  $J = 30$  RUGs. The distribution of these residents across both NHs and RUGs is rather variable. Specifically, each NH hosts from 15 to 84 residents (with an average of 37.2), whereas residents assigned to the same RUG range from 7 to 193 (58.3 on average). Also, mortality is quite different across RUGs, with the distribution of the  $m_{.j}$  rates ranging from 0.040 to

	RMSE							
	$\widehat{\text{API}}_h^{(1)}$				$\widehat{\text{API}}_h^{(2)}$			
	2	4	8	16	2	4	8	16
<b>SSF</b>								
<b>Min</b>	0.122	0.086	0.062	0.043	0.022	0.016	0.011	0.008
<b>Mean</b>	0.246	0.179	0.127	0.088	0.036	0.026	0.018	0.013
<b>Max</b>	0.416	0.332	0.234	0.159	0.055	0.039	0.027	0.019

	Empirical Coverage (95% CIs)							
	$\widehat{\text{API}}_h^{(1)}$				$\widehat{\text{API}}_h^{(2)}$			
	2	4	8	16	2	4	8	16
<b>SSF</b>								
<b>Min</b>	0.815	0.853	0.873	0.881	0.861	0.902	0.925	0.934
<b>Mean</b>	0.909	0.925	0.937	0.942	0.914	0.933	0.943	0.947
<b>Max</b>	0.954	0.959	0.959	0.960	0.947	0.956	0.954	0.957

**Table 1** Results of the simulation study.

0.467 (with a standard deviation of 0.094). Thus, it is evident that case-mix needs to be accounted for. To obtain the true values of the indices as well as the first-order approximate variances of the corresponding estimators, the  $m_{hj}$  rates observed in the dataset are taken as the true death probabilities  $p_{hj}$ . To avoid extreme values, these probabilities are then capped to vary between  $m_{.j}/2$  and  $2m_{.j}$ . The simulation study is run for different sample sizes as follows. For each  $(h, j)$  pair,  $N_{hj}$  is obtained by multiplying the value in the original data by a factor of 2, 4, 8 and 16 (Sample Size Factor (SSF)). In this way, the overall case-mix composition is unchanged. Then, for each  $N_{hj}$ , the value of  $k_{hj}$  is drawn from a  $\text{Bin}(N_{hj}, p_{hj})$ . This scheme is replicated so that  $N = 5000$  simulated datasets are generated for each value of SSF.

As stated in Section 2.1, the absence of correlation among measures of the same NH is assumed to derive the variance expressions in (3) and (6). Since the structure of the simulated data reflects that of the Umbrian dataset, it is worth to check whether such an assumption is met in the real data. To this end, we have fitted a logistic mixed model [3, 4] where the observed mortality rates are regressed against RUG and NH membership. The former is included as a fixed effect, whereas the latter as a random effect with a  $N(0, \sigma_\alpha^2)$  distribution. In this framework, the intraclass correlation coefficient can be used to evaluate the correlation between rates in the same NH, adjusting for the presence of RUG effects. Specifically, this coefficient is given by  $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \pi^2/3)$ . In the Umbrian dataset, its estimate is  $\hat{\rho} = 0.024$ , denoting an almost null level of correlation. Despite the absence of intraclass correlation, it is worth to underline that the estimate  $\hat{\sigma}_\alpha^2 = 0.082$  shows that a quite robust NH effect is present in the data.

Table 1 contains an overview of the simulation results. Specifically, for each index a summary of the distribution across the NHs of the Root Mean Squared Error (RMSE) and the empirical coverage of the 95% Confidence Intervals (CIs) based on the normal distribution is reported for every value of the SSF. As expected, the values of the RMSEs approach zero as the SSF grows. However, some degree of

under-coverage is spotted, especially for the  $API_h^{(1)}$  index. Since the biases of the two index estimators always appear to be negligible, this is essentially due to the under-estimation of the variance or to substantial deviation from normality, which are more likely to occur for small NHs. We recall that such an under-estimation is the result of two components: the bias in the estimation of the right-hand sides of (3) and (6) and the distortion induced by the Taylor approximations in small samples. The above result can be applied in different ways: for example, significant departures from the mean NH performance ( $API_h^{(1)} = 1$  or  $API_h^{(2)} = 0$ ) can be easily detected.

#### 4 Further extensions

The framework proposed in this paper can be extended in a number of directions. First, the derivation of the covariance between estimators of different NHs would be useful to make proper inference on the difference between the corresponding indices, both in a pairwise and in a multiple comparison setting. Also, suitable normalizations could be introduced to allow the indices to vary within the same interval, thereby enhancing the comparability of their estimators' performances. Finally, the variances in (3) and (6) could be reformulated to account for the presence of intraclass (within-NH) correlation. This extension would be appropriate when such a correlation is non-negligible and would result in computing the covariance between any pair of measurements referring to the same NH. An approximate expression for this covariance can be obtained via the delta method.

**Acknowledgements** We are thankful to Cassa di Risparmio di Perugia for financial support.

#### References

- [1] Ken Butler and Michael A. Stephens. The distribution of a sum of independent binomial random variables. *Methodology and Computing in Applied Probability*, 19(2):557–571, 2017.
- [2] Brant E. Fries, Don P. Schneider, William J. Foley, Marie Gavazzi, Robert Burke, and Elizabeth Cornelius. Refining a case-mix measure for nursing homes: Resource Utilization Groups (RUG-III). *Medical Care*, 32(7):668–685, 1994.
- [3] C. E. McCulloch and S. R. Searle. *Generalized, linear and mixed models*. Wiley, 2002.
- [4] Shinichi Nakagawa and Holger Schielzeth. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, 85(4):935–956, 2010.

# The ultrametric covariance model for modelling teachers' job satisfaction

## *Il modello di covarianza ultrametrica per lo studio della soddisfazione lavorativa dei professori*


Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria

**Abstract** Multidimensional phenomena are often characterised by nested latent concepts ordered in a hierarchical structure, from the most specific to the most general ones. In this paper, we model a nonnegative data covariance matrix by extending the Ultrametric Correlation Model to covariance matrices. The proposal is a parsimonious model which identifies a partition of variables in a reduced number of groups, and the relationships among them via the ultrametric property. The proposed model is applied to investigate the relationships among the dimensions of the Teachers' Job Satisfaction in Italian secondary schools.

**Abstract** *I fenomeni multidimensionali sono spesso caratterizzati da concetti latenti ordinati in una struttura gerarchica, dai più specifici al più generale. In questo articolo ci proponiamo di modellare una matrice di covarianza nonnegativa, estendendo il modello chiamato Ultrametric Correlation Model alle matrici di covarianza. La proposta metodologica si esplica in un modello parsimonioso che identifica una partizione di variabili in un numero ridotto di gruppi e le loro relazioni mediante la proprietà di ultrametricità. Il modello proposto è applicato allo studio delle relazioni tra le dimensioni della soddisfazione lavorativa dei professori nelle scuole italiane superiori di secondo grado.*

**Key words:** Ultrametric matrices, hierarchical structures, teachers' job satisfaction, confirmatory analysis, dimensionality reduction

---

Carlo Cavicchia   
Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands  
e-mail: cavicchia@ese.eur.nl

Maurizio Vichi   
Department of Statistical Sciences, University of Rome La Sapienza, Rome, Italy  
e-mail: maurizio.vichi@uniroma1.it

Giorgia Zaccaria   
Department of Statistical Sciences, University of Rome La Sapienza, Rome, Italy  
e-mail: giorgia.zaccaria@uniroma1.it

## 1 Introduction

Multidimensional phenomena are often characterised by a hierarchy of nested latent concepts (dimensions) with different levels of abstraction, from the most specific to the most general ones. The study of these phenomena needs specific models since the traditional ones, usually used to reconstruct the relationships among variables (e.g., Factor Analysis, FA, [1]), fail in the definition of a hierarchical structure over them. Cavicchia et al. [4] introduced a parsimonious simultaneous model, named Ultrametric Correlation Model (UCM), to reconstruct a nonnegative data correlation matrix of order  $p$  via an ultrametric correlation one. The ultrametric property allows both to identify a partition of variables in  $Q \leq p$  groups and the relationships among them by defining two difference features: the within-concept consistency and the correlation between groups.

In this paper, we introduce a new model, called *Ultrametric Covariance Model* (UCovM), to reconstruct a nonnegative covariance matrix by extending the one proposed by Cavicchia et al. [4] for nonnegative correlation matrices. Similarly to UCM, UCovM defines a hierarchy of latent concepts by pinpointing a variable partition in  $Q$  groups characterised by three features: the variance of a group, the covariance within the group and the covariance between groups. Since a decreasing order is imposed on these features, two variables belonging to the same group are more concordant than two belonging to different groups. Although the nonnegativity assumption might seem restrictive, it turns out to be realistic in many real-data applications. We apply UCovM to Teachers Job Satisfaction data set [7] in order to investigate the hierarchical relationships between the six dimensions defining the job satisfaction for teachers.

## 2 Background

Let us recall the definition of an ultrametric matrix [6, pp. 58-59], which differs from an ultrametric distance matrix even if there exists a relationship between the two.

**Definition 1.** A nonnegative matrix  $\mathbf{U}$  of order  $p$  is said to be ultrametric if

- (i)  $u_{jl} = u_{lj}$  for all  $j, l = 1, \dots, p$  (symmetry);
- (ii)  $u_{jj} \geq \max\{u_{lj} : l = 1, \dots, p\}$  for  $j = 1, \dots, p$  (column pointwise diagonal dominance);
- (iii)  $u_{jl} \geq \min\{u_{ji}, u_{il}\}$ , for  $i, j, l = 1, \dots, p$  (ultrametric inequality).

Every ultrametric matrix turns out to be positive semi-definite, as demonstrated by Dellacherie et al. [6, pp. 60-61]. Considering a nonnegative data covariance matrix  $\mathbf{S}$  of order  $p$ , with elements  $s_{jl} \in \mathbb{R}^+$  (the set of nonnegative real numbers),  $j, l = 1, \dots, p$ , it is (i) symmetric and (ii) positive semi-definite by definition. If conditions (ii) and (iii) hold,  $\mathbf{S}$  is an ultrametric covariance matrix.



### 3 Methodology

Let  $\mathbf{S}$  be a nonnegative data covariance matrix of order  $p$ . The problem we want to deal with can be formalised as

$$\mathbf{S} = \mathbf{S}_u + \mathbf{E}, \quad (1)$$

where  $\mathbf{S}_u$  is an ultrametric covariance matrix of order  $p$  and  $\mathbf{E}$  is an error matrix of the same order.

The Ultrametric Covariance Model (UCovM) defines an ultrametric covariance matrix for modelling hierarchical latent concepts, which is formally specified as follows

$$\mathbf{S}_u = \mathbf{V}(\mathbf{S}_W + \mathbf{S}_B)\mathbf{V}' - \mathbf{V}\mathbf{S}_W\mathbf{V}' \odot \mathbf{I}_p + \mathbf{V}\mathbf{S}_V\mathbf{V}' \odot \mathbf{I}_p, \quad (2)$$

subject to constraints

$$\mathbf{V} = [v_{jq} \in \{0, 1\} : j = 1, \dots, p, q = 1, \dots, Q]; \quad (3)$$

$$\mathbf{V}\mathbf{1}_Q = \mathbf{1}_p \quad \text{i.e.} \quad \sum_{q=1}^Q v_{jq} = 1 \quad j = 1, \dots, p; \quad (4)$$

$$\mathbf{S}_B = \mathbf{S}'_B, \text{diag}(\mathbf{S}_B) = \mathbf{0}, {}_B s_{qh} \geq \min\{{}_B s_{qt}, {}_B s_{ht}\} \quad q, h, t = 1, \dots, Q, t \neq h \neq q; \quad (5)$$

$$\min\{{}_W s_{qq} : q = 1, \dots, Q\} \geq \max\{{}_B s_{qh} : q, h = 1, \dots, Q, h \neq q\}; \quad (6)$$

$${}_V s_{qq} \geq {}_W s_{qq}, q = 1, \dots, Q, \quad (7)$$

where  $\mathbf{I}_p$  is an identity matrix of order  $p$ ,  $\odot$  is the Hadamard (element-wise) product and  $\text{diag}(\mathbf{S}_B)$  identifies the main diagonal of  $\mathbf{S}_B$ .

$\mathbf{S}_V$  and  $\mathbf{S}_W$  are diagonal matrices, whose diagonal elements represent the variances of and the covariances within the  $Q$  variable groups, respectively, whereas the covariances between them are expressed by the off-diagonal elements of  $\mathbf{S}_B$ . Since constraint (5), (6) and (7) hold, an ordering between the elements of  $\mathbf{S}_V$ ,  $\mathbf{S}_W$  and  $\mathbf{S}_B$  exists. This leads to a hierarchy of latent concepts, each one associated with a variable group, whose hierarchical levels are defined by the covariances within and between groups, i.e., the diagonal and off-diagonal elements of  $\mathbf{S}_W$  and  $\mathbf{S}_B$ , respectively. Specifically, the higher the covariance among two variable groups (or variables themselves), the stronger the concordance among them and the earlier they are merged together.

UCovM allows pinpointing groups of variables, each one associated with a dimension, by reducing the dimensionality of the phenomenon under study, and identifying new latent concepts and the hierarchical relationships among them. Thus, UCovM is an exploratory, parsimonious and simultaneous model. If  $\mathbf{V}$  is set a priori, i.e., the variable partition is fixed, then the model can be applied in a confirmatory approach.

The proposal is estimated in a least-squares framework and implemented with a coordinate descent algorithm.

Table 1: List of variables for each dimension of the Teachers' Job Satisfaction data set<sup>a</sup> and the corresponding Cronbach's  $\alpha$ .

Dimension	Dimension ID	Variables	$\alpha$
Communication	Comm	1, 2, 3, 4, 5	0.8136
External School Image	Imag	6, 7	0.8582
Involvement	Invo	8, 9, 10, 11, 12, 13	0.8999
Leadership	Lead	14, 15, 16, 17	0.9021
School Climate	Clim	18, 19, 20, 21, 22, 23	0.8817
Infrastructure	Infr	24, 25	0.7052

<sup>a</sup> See [7, Table 1] for a complete description of the variables.

#### 4 Teachers' Job Satisfaction: differences between the overall ultrametric covariance structure and those by gender

Job Satisfaction is a multidimensional phenomenon characterised by different dimensions affecting feelings and emotions of employees towards their job. We apply the UCovM to study Teachers' Job Satisfaction (TJS) and investigate the hierarchical relationships among the factors that contribute to define TJS. The analysis is based upon the survey conducted by Sarnacchiaro et al. [7] in four Italian state secondary schools. Table 1 shows the dimensions of the TJS and the partition of variables in six groups, each one associated with the corresponding dimension. Cronbach's  $\alpha$  [5] for each group is also computed and all dimensions result reliable. Moreover, Cronbach's  $\alpha$  for the whole data set turns out to be 0.9528, revealing the strong reliability of the general latent concept, i.e. the TJS. Additional variables pertaining socio-demographic features are also measured; among them, we consider the variable *Gender* in order to compare the hierarchical structure defining TJS on the aforementioned data set with those estimated differently for female and male.

Firstly, we performed UCovM in a confirmatory approach on the covariance matrix - containing nonnegative values - of the whole data set. The partition in six groups of variables corresponding to the dimensions of the TJS is clearly visible in the covariance matrix (Figure 1a). The groups which are mostly concordant within them are those associated with *Leadership* and *External School Image*. As shown in Figure 1b, the first aggregation lumps together *Involvement* and *Leadership*, which, indeed, have a high impact on TJS [7]. The following aggregations show a constant trend by adding one at a time the remaining dimensions - connected with the school-based factors - to the first group, up to the *Infrastructure*, which is the less concordant dimension with the others (the covariance between the broader group with five dimensions and *Infrastructure* is equal to 0.2045).

Comparing these results with those obtained by implementing UCovM by gender - both the covariance matrices are nonnegative - we can notice some differences between TJS for female and male (Figure 2). The hierarchy over the six dimensions of TJS for female (Figure 2a) is similar to that obtained on the whole data set. Indeed, even if the covariances within and between groups are less strong than

The ultrametric covariance model for modelling teachers' job satisfaction

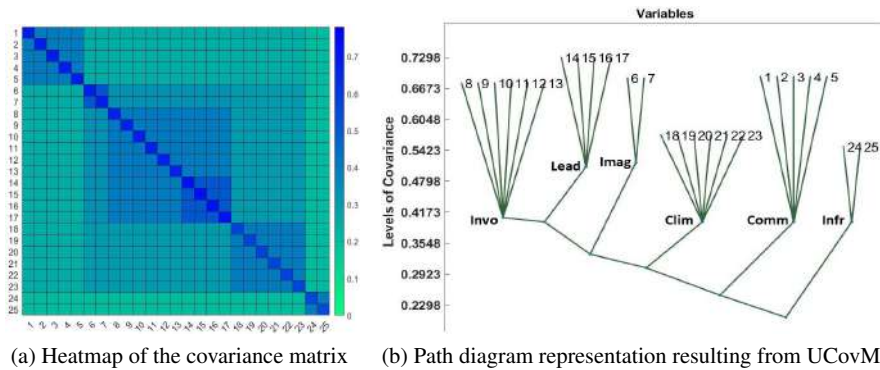


Fig. 1: Graphical representations of relationships among the dimensions of TJS for the whole data set.

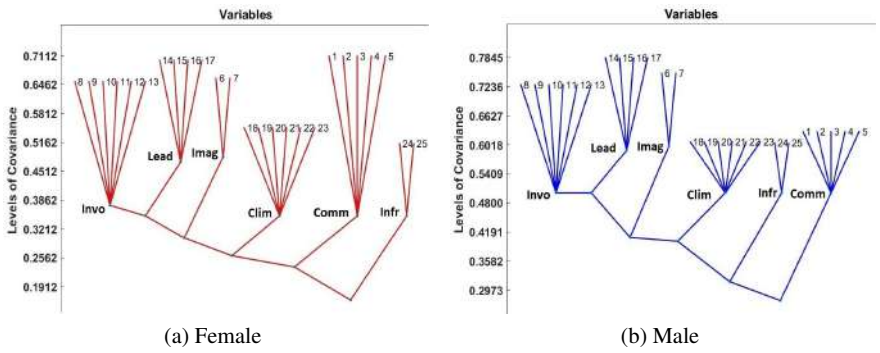


Fig. 2: Path diagram representation of the TJS resulting from UCovM by gender.

those on the whole data set, the aggregations are the same. This happens also because the percentage of women in the data set is greater than that of men. On the other hand, the six dimensions of TJS for male show a slightly different hierarchical structure (Figure 2b). The first aggregation lumps together *Involvement* and *Leadership* as well; therefore, the variables pertaining *Involvement* are merged with those associated with *Leadership* such that the covariance within the former group is equal to that between the two variable groups. Looking at Figure 1a, the difference between the covariance magnitude of these two variable groups and that of the variable group associated with *Involvement* seems to be slight. The other aggregations show a constant trend, with covariances between *Involvement*, *Leadership*, *External School Image* and *School Climate* greater than 0.38. The last two aggregations are reversed with respect to those for female.

## 5 Conclusions

The model proposed herein, called Ultrametric Covariance Model, is an extension of the Ultrametric Correlation Model, introduced by Cavicchia et al. [4], to covariance matrices. It aims at reconstructing the hierarchical relationships existing among variables by modelling a nonnegative covariance matrix via an ultrametric covariance one.

UCovM was applied on a real data set in order to study the hierarchical relationships among the six dimensions of the Teachers' Job Satisfaction. The analysis is conducted on the overall data set [7] and differently by gender. The hierarchy of the TJS dimensions is slightly different between male and female. Comparing the results obtained by UCovM with those attained by UCM, we can highlight that for the whole data and the males' ones the second and the third aggregations are swapped, whereas the hierarchy remains the same for the females' data. Conversely to UCM, UCovM allows to inspect the variability of each group of the variable partition. Some comparisons with other methodologies, as Higher-Order models [2] and hierarchical clustering methods, were carried out: in both cases the models' fit pointed out that a simultaneous methodology was needed. Cavicchia et al. [3] in turn demonstrated that hierarchical clustering techniques had some limitations in detecting hierarchical relationships among variables if compared to simultaneous methodologies as UCM.

Our goals for future studies are to implement a bootstrap test to assess if the difference between the parameters of the UCovM estimated by gender is statistically significant; to study the TJS according to other socio-demographic features and to build an R and/or Matlab package to implement the proposal.

**Acknowledgements** The authors would like to thank Prof. Pasquale Sarnacchiaro, and the authors of [7], to have shared with us their data.

## References

1. Anderson, T., Rubin, H.: Statistical inferences in factor analysis. *Proceedings of the Third Symposium on Mathematical Statistics and Probability* **5**, 111–150 (1956)
2. Cattell, R.: *The scientific use of factor analysis in behavioral and life sciences*. Plenum (1978)
3. Cavicchia, C., Vichi, M., Zaccaria, G.: Exploring hierarchical concepts: theoretical and application comparison. In: T. Imaizumi, A. Nakayama, S. Yokoyama (eds.) *Advanced Studies in Behaviormetrics and Data Science*. *Behaviormetrics: Quantitative Approaches to Human Behavior*, vol. 5, pp. 315–328. Springer, Singapore (2020)
4. Cavicchia, C., Vichi, M., Zaccaria, G.: The ultrametric correlation matrix for modelling hierarchical latent concepts. *Adv Data Anal Classif* **14**(4), 837–853 (2020)
5. Cronbach, L.: Coefficient alpha and the internal structure of tests. *Psychometrika* **16**(3), 297–334 (1951)
6. Dellacherie, C., Martinez, S., Martin, J.S.: *Inverse M-matrices and ultrametric matrices*. *Lecture Notes in Mathematics*. Springer International Publishing (2014)
7. Sarnacchiaro, P., Scippacercola, S., Malafrente, P.: A statistical model for self-evaluation of teacher's satisfaction: a study in an Italian secondary school. *Electron. J. Appl. Stat. Anal.* **13**(3), 637–656 (2019)

## 4.28 Some recent developments in compositional data analysis

# A Robust Approach to Microbiome-Based Classification Problems

## *Un Approccio Robusto ai Problemi di Classificazione nell'Analisi del Microbioma*

Gianna Serafina Monti and Peter Filzmoser

**Abstract** In this paper we consider a two-class classification problem based on microbiome data, which play a central role in the diagnosis of diseases. The log-contrast model, involving log-transformed ratios of two parts of a composition, was successfully employed for modeling a continuous response as a function of a compositional vector. Then the log-contrast model was generalized by allowing the linear model to be related to the response variable, belonging to the exponential family, via a link function. Afterwards, to face with the analysis of high-dimensional data, regularization was used. Monti and Filzmoser [7] had recently proposed a robust version of the generalized log-contrast model, the so-called RobLZS estimator, to cope with the dual purpose of variable selection and classification task. In this contribution, a microbiome data application is considered to compare the performance of the RobLZS estimator with two non-linear methods, namely Decision Trees and Random Forests, which are widely used in machine learning.

**Abstract** *In questo articolo consideriamo un problema di classificazione a partire dalla composizione del microbioma, che gioca un ruolo centrale nella diagnosi delle malattie. Il modello basato sui log-contrast, ovvero rapporti di trasformate logaritmiche di due parti di una composizione, è stato impiegato con successo per modellare una risposta continua in funzione di un vettore di covariate composizionali. Tale modello è stato poi generalizzato consentendo di modellizzare una variabile di risposta, appartenente alla famiglia esponenziale, tramite una funzione link. Successivamente, per far fronte all'analisi di dati ad alta dimensione si è ricorso alle tecniche di regolarizzazione. Monti and Filzmoser [7] hanno recentemente proposto una versione robusta del modello basato sui log-contrast, il cosiddetto stimatore RobLZS, con il duplice scopo di selezione delle variabili rilevanti e di classifi-*

---

Gianna Serafina Monti

Department of Economics, Management and Statistics, University of Milano Bicocca, e-mail: gianna.monti@unimib.it

Peter Filzmoser

Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology  
e-mail: P.Filzmoser@tuwien.ac.at

*cazione delle unità statistiche. In questo contributo si considera un'applicazione ai dati del microbioma per confrontare le prestazioni dello stimatore RobLZS con due metodi non lineari, gli alberi decisionali e le Random Forests, largamente utilizzati nell'ambito del machine learning.*

**Key words:** Log-contrast model; Model selection; Regularization; Sparsity; Random Forest.

## 1 Introduction

The microbiome analysis is of great importance in understanding the human state of health, and it is becoming pivotal in the comprehension of several diseases. Microbiome studies are based on high-throughput DNA sequencing technologies. The sequencing reads data are often summarized into sparse compositional counts of operational taxonomic units (OTUs).

The total number of reads per sample depends on the capacity of the instrument, in other words is constrained by the maximum number of sequence reads, resulting in constrained compositional data. The abundance values lose their informative value, and the relevant information is incorporated in the ratio between the several components of microbioma.

The analysis of this data is particularly challenging as a microbiome dataset is typically high dimensional, zero inflated due to the presence of a lot of rare taxa and just compositional. This latter characteristic makes traditional statistical methods inadequate, leading to inconsistent results and spurious associations, and proper methods that take into account the compositional character are required.

In the generalized linear log-contrast model [1] a set of compositional covariates is linked to a non-gaussian outcome to successfully cope with classification problems. The log-contrast model has been extended tailored to the high-dimensional setting, in which the number of features is typically greater than the sample size, via regularization [10, 6].

This paper refers to a recently proposed method [7] with the aim to robustify the estimation procedure of a penalized generalized linear log-contrast model for a binary outcome by means of a trimmed estimator, the so-called RobLZS estimator.

The paper is organized as follows: in Section 2 the log-contrast model and the robust generalized log-contrast model are outlined, and Section 3 illustrates an application to a real microbiome dataset to compare the performance of the RobLZS estimator with two non-linear methods, namely Decision Trees and Random Forests.

## 2 Robust Generalized Log-Contrast Model

Let  $y_i$  ( $i = 1, \dots, n$ ) be binary responses that follow independent Bernoulli distributions, i.e.,  $y_i \sim Be(1, \pi_i)$ , where the success probability  $\pi_i$  depends on the value of a  $p$ -dimensional covariate vector  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$  which lies in the unit simplex  $S^p = \{\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T : x_{ij} > 0, \text{ for } j = 1, \dots, p, \text{ and } \sum_{j=1}^p x_{ij} = 1\}$ .

Let's consider the elastic-net-type penalized logistic regression estimator (hereafter ZS estimator) [10, 6]

$$\hat{\boldsymbol{\beta}}_{ZS} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^n d(\mathbf{z}_i^T \boldsymbol{\beta}, y_i) + n\lambda P_\alpha(\boldsymbol{\beta}) \right\} \quad \text{subject to } \sum_{j=1}^p \beta_j = 0, \quad (1)$$

where  $\mathbf{z}_i = \log(\mathbf{x}_i)$ ,  $d(\mathbf{z}_i^T \boldsymbol{\beta}, y_i) = -\log(\ell(y_i, \mathbf{z}_i^T \boldsymbol{\beta})) = -y_i \mathbf{z}_i^T \boldsymbol{\beta} + \log(1 + \exp(\mathbf{z}_i^T \boldsymbol{\beta}))$  is the negative log-likelihood, or deviance, for the  $i$ th component,  $P_\alpha(\boldsymbol{\beta})$  is the elastic-net regularization penalty [11], defined as

$$P_\alpha(\boldsymbol{\beta}) = \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1,$$

where  $\alpha \in [0, 1]$  and  $\lambda \in [0, \infty)$  are the tuning parameters:  $\alpha$  balances the  $\ell_2$  and  $\ell_1$  penalizations, whereas  $\lambda$  controls the sparsity of the solution.

As likelihood-based procedures are highly sensitive to the presence of outliers, leading to biased estimations and unreliable conclusions, Monti and Filzmoser [7] recently proposed the Robust Logistic Zero-Sum estimator (RobLZS), i.e. a penalized maximum trimmed likelihood estimator for robust high-dimensional generalized linear models [2, 8, 5]. The RobLZS estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{RobLZS}} = \arg \min_{H \subseteq \{1, \dots, n\}: |H|=h} \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i \in H} d(\mathbf{z}_i^T \boldsymbol{\beta}, y_{i_i}) + n\lambda P_\alpha(\boldsymbol{\beta}) \right\}, \quad (2)$$

subject to  $\sum_{j=1}^p \beta_j = 0,$

where  $d(\mathbf{z}_{i_1}^T \boldsymbol{\beta}, y_{i_1}) \leq d(\mathbf{z}_{i_l}^T \boldsymbol{\beta}, y_{i_l}) \leq d(\mathbf{z}_{i_h}^T \boldsymbol{\beta}, y_{i_h})$ ,  $i_l \in \{1, \dots, n\}$ ;  $d(\mathbf{z}_{i_k}^T \boldsymbol{\beta}, y_{i_k})$  are the ordered deviances  $h = \lfloor \alpha n \rfloor$ ,  $\alpha \in [0.5, 1]$  with  $1 - \alpha$  the trimmed portion, and  $\lfloor \cdot \rfloor$  means rounding down to the nearest integer. RobLZS is based on adaptively trimming observations for which the contribution to the likelihood are the least to get a more robust estimation. A coordinate descend algorithm in conjunction with an analog of the fast-LTS algorithm [9, 2] was implemented to obtain the parameter estimates of (2). RobLZS regression is accessible as an R package in GitHub <https://github.com/giannamonti/RobZS>.



### 3 A Microbiome Application

We study the performance of the RobLZS estimator through an application to real microbiome data, with respect to a plausible competitor, the ZS estimator, i.e. its non-robust counterpart, and with the two non-linear methods Decision Trees and Random Forests [3]. The latter are widely used in machine learning modeling thanks to their flexibility, which leads to high prediction accuracy.

The analysed data refers to a published study on the role of microbioma components in colorectal carcinoma pathogenesis [4]. The goal is to classify individuals as having noncancerous colon or colorectal carcinoma based on 16S rDNA sequences, as well as to identify which OTUs are relevant in cancer diagnosis. The dataset was preprocessed by filtering out OTUs which had more than 90% of zeros. The remaining zero counts were then replaced by a pseudo-count value 0.5 to allow for a logarithmic transformation. The dimension of the microbiome dataset originally was  $n \times p = 172 \times 409$ , and after preprocessing the final number of OTUs is  $p = 130$ . The two class of subjects, healthy and sick, are equally represented.

To compare the predictive performance of the four methods, we used a validation-set approach partitioning of the data in 80% training and 20% testing. We trained the models using the training dataset and applied the final estimated model to the held-out data to evaluate the testing predictive performance of each model. The data split, training, and testing steps were repeated 100 times to obtain a more objective picture of models performances. To assess the classifiers we employed as evaluation measures: precision, recall, accuracy, specificity and the AUC, namely the area under the Receiver Operating Characteristic (ROC).

**Table 1** Confusion matrix for binary classification. TP corresponds to true positive, FP false positive, FN false negative, and TN true negative counts.

Actual \ Predicted	Negative	Positive
Negative	TN	FP
Positive	FN	TP

The different accuracy measures, that can be interpreted from the confusion matrix (see Table 1), are calculated using the given formulas below.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP+FP}, & \text{Recall(aka Sensitivity)} &= \frac{TP}{TP+FN}, \\ \text{Accuracy} &= \frac{TN+TP}{TN+FP+FN+TP}, & \text{Specificity} &= \frac{TN}{TN+FP}, \\ \text{AUC} &= \frac{\text{Sensitivity}+\text{Specificity}}{2}. \end{aligned}$$

Table 2 presents the results of the CV procedure, over 100 repetitions. Decision Trees result in the worst performance, while ZS successes only in Precision, but

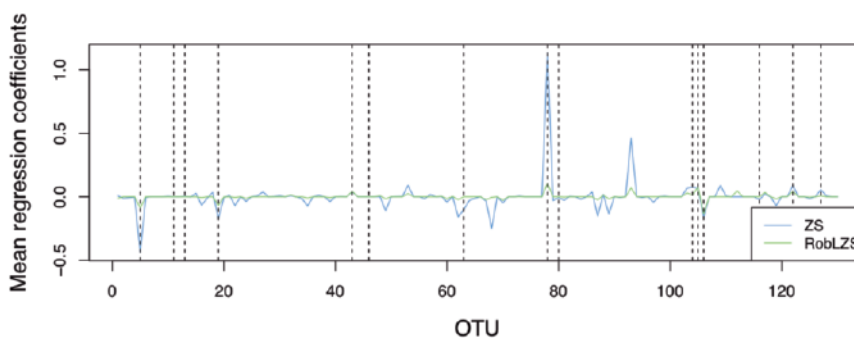
interestingly, these results demonstrate that RobLZS estimator performs nearly as well as the most complex non-linear model (Random Forest).

When sensitivity and specificity are diagnostically equally important, a good summary is represented by AUC, aka balanced accuracy. The Random Forests gave a mean AUC (se) of 0.665 (0.01), slightly better than those from the RobLZS estimator 0.64 (0.01).

**Table 2** Comparison of prediction performance among different methods from CV.

	<b>ZS</b>		<b>RobLZS</b>		<b>Rf</b>		<b>Tree</b>	
	mean	se	mean	se	mean	se	mean	se
Precision	<b>0.734</b>	0.015	0.679	0.014	0.694	0.012	0.591	0.014
Recall	0.275	0.007	0.590	0.015	<b>0.647</b>	0.014	0.566	0.016
Accuracy	0.558	0.006	0.642	0.010	<b>0.659</b>	0.009	0.564	0.012
Specificity	0.879	0.006	<b>0.691</b>	0.010	0.682	0.009	0.568	0.012
AUC	0.577	0.004	0.640	0.010	<b>0.665</b>	0.010	0.567	0.011

From the interpretability point of view, in the identification of potential biomarkers for colorectal carcinoma, we computed and visualised in Figure 1 the mean regression coefficients over all CV replications for ZS and RobLZS. For Random Forests we quantified the importance of each OTU, based on how much the accuracy in making predictions decreases when such feature is excluded. We computed the averaged importance of each OTU over 100 repetitions of CV, and ranked the OTUs from the most to the least important. The vertical dashed lines in Figure 1 correspond to the first 15 important OTUs. In summary, 10 out of the 15 most important OTUs are shared by the three considered methods.



**Fig. 1** Mean regression coefficients over all CV replications for ZS and RobLZS. The vertical dashed lines indicate the most 15 important OTUs identified by Random Forests.

## References

- [1] Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*, **71**(2):323–330.
- [2] Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann Appl Stat*, **7**(1):226–248.
- [3] Breiman, L. (2001). Random forests. *Mach Learn*, **45**:5–32.
- [4] Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Taberero, J., Baselga, J., Liu, C and Shivdasani, R. A., Ogino, S., Birren, B. W., Huttenhower, C., Garrett, W. S., and Meyerson, M. (2012). Genomic analysis identifies association of fusobacterium with colorectal carcinoma. *Genome Res*, **22**(2):292–8.
- [5] Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemom Intell Lab Syst*, **172**:211 – 222.
- [6] Lu, J., Shi, P., and Li, H. (2019). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics*, **75**(1):235–244.
- [7] Monti, G. S. and Filzmoser, P. (2021). Robust Logistic Zero-Sum regression for microbiome compositional data. *Submitted*.
- [8] Neykov, N. M., Filzmoser, P., and Neytchev, P. N. (2014). Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. *Stat Pap*, **55**(1):187–207.
- [9] Rousseeuw, P. J. and Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Min Knowl Discov*, **12**(1):29–45.
- [10] Zacharias, H. U., Rehberg, T., Mehrl, S., Richtmann, D., Wettig, T., Oefner, P. J., Spang, R., Gronwald, W., and Altenbuchinger, M. (2017). Scale-invariant biomarker discovery in urine and plasma metabolite fingerprints. *J Proteome Res*, **16**(10):3596 – 3605.
- [11] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*, **67**(2):301 – 320.

# What is a convex set in compositional data analysis?

*Che cos'è un insieme convesso nell'analisi dei dati composizionali?*

Saperas, J., Martín-Fernández, J. A.

**Abstract** Compositions are vectors which components represent parts of a whole. Historically, they have been defined as vectors with strictly positive components with constant sum. The sample space of this data is the simplex which has a particular geometric structure known as the Aitchison geometry. The basic operations of Aitchison geometry are the perturbation and the powering. Consequently, concepts and statistical techniques in the analysis of compositional data must be consistent with Aitchison's geometry. In this work, we rigorously define geometric objects related to the concept of convexity to ensure that they are compatible with the geometry of the simplex. Examples of most common sets used in statistical and operational research techniques will be presented.

**Abstract** *Le composizioni sono vettori le cui componenti rappresentano parti di un tutto. Storicamente, sono stati definiti come vettori con componenti strettamente positive e con somma costante. Lo spazio campionario di questi dati è il semplice, dotato di una particolare struttura geometrica nota come geometria di Aitchison. Le operazioni di base della geometria di Aitchison sono l'operazione di perturbazione e l'operazione di potenza. Di conseguenza, è auspicabile che un'analisi statistica dei dati composizionali sia coerente con la geometria di Aitchison. In questo lavoro definiamo rigorosamente oggetti geometrici legati al concetto di convessità, verificando la loro compatibilità con la geometria del semplice. Verranno presentati alcuni tra i più comuni esempi di dati utilizzati nelle comuni tecniche di analisi statistica e di ricerca operativa.*

**Key words:** Convex set, simplex, log-ratio, Aitchison geometry

---

Jordi Saperas Riera  
Dpt. d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona, Girona (Spain)  
e-mail: jordi.saperas@udg.edu

Josep Antoni Martín Fernández  
Dpt. d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona, Girona (Spain)  
e-mail: josepantoni.martin@udg.edu

### 1 Introduction

Many statistical techniques such as, among others, mixtures on design of experiments, optimal partitions on clustering, convex hulls on outlier analysis, and any method including optimization are based on the concept of convexity ([6]). Importantly, geometric objects related to the concept of convexity should be compatible with the geometry of the data sample space. In our case, the simplex is the sample space of compositional data (CoDa), where it is defined the Aitchison geometry.

A D-part composition is a vector which D components represent parts of a whole ([1]). The compositional space is the quotient space defined by vectors with components strictly positive and the scalar invariance as an equivalence relation ([2]).

$$\mathbb{K}^D = \mathbb{R}^{+,D} / \mathcal{V} = \{ \underline{\mathbf{x}} \mid \mathbf{x} \in \mathbb{R}^{+,D}, \underline{\mathbf{x}} = \underline{\mathbf{y}} \leftrightarrow \exists \mu \in \mathbb{R}^+, \mathbf{y} = \mu \mathbf{x} \}.$$

The two operations defined over  $\mathbb{K}^D$  are the perturbation ( $\otimes$ ) and the powering ( $\odot$ ).

$$\begin{aligned} \underline{\mathbf{x}} \otimes \underline{\mathbf{y}} &= (x_1 y_1, \dots, x_D y_D), \underline{\mathbf{x}}, \underline{\mathbf{y}} \in \mathbb{K}^D \\ \lambda \odot \underline{\mathbf{x}} &= (x_1^\lambda, \dots, x_D^\lambda), \lambda \in \mathbb{R}, \underline{\mathbf{x}} \in \mathbb{K}^D. \end{aligned}$$

The most common representatives of CoDa are the vectors of proportions forming the unit simplex.

$$\mathcal{S}^D = \left\{ (x_1, x_2, \dots, x_D)' \in \mathbb{R}^D \mid x_i > 0, \sum_{i=1}^D x_i = 1 \right\}.$$

In this context, normalization to the unit simplex is known as closure and is denoted by  $\mathcal{C}$ .

$$\begin{aligned} \mathcal{C} : \mathbb{K}^D &\longrightarrow \mathcal{S}^D \\ \underline{\mathbf{x}} &\longrightarrow (x_1, \dots, x_D) / \sum_{i=1}^D x_i \end{aligned}$$

Unit simplex with the before operations, perturbation and powering, is a vector space of dimension D-1 ([9]).

Given  $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ , the Aitchison inner product  $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{A}}$  is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{A}} = \sum_{i=1}^D \ln \left( \frac{x_i}{g(\mathbf{x})} \right) \ln \left( \frac{y_i}{g(\mathbf{y})} \right) = \langle clr(\mathbf{x}), clr(\mathbf{y}) \rangle_E \tag{1}$$

where  $g(\mathbf{x}) = (x_1 x_2 \dots x_D)^{1/D}$  is the geometric mean,  $\langle \cdot, \cdot \rangle_E$  is the typical euclidean inner product, and  $clr(\mathbf{x}) = \left( \ln \left( \frac{x_1}{g(\mathbf{x})} \right), \dots, \ln \left( \frac{x_D}{g(\mathbf{x})} \right) \right)$ . From this inner product (1) is derived the definition of Aitchison norm and distance to complete the Aitchison geometry concepts:

$$\|\mathbf{x}\|_{\mathcal{A}}^2 = \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{A}} = \sum_{i=1}^D \left( \ln \left( \frac{x_i}{g(\mathbf{x})} \right) \right)^2$$

What is a convex set in compositional data analysis?

$$d_{\mathcal{A}}^2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{\mathcal{A}}^2 = \sum_{i=1}^D \left( \ln \left( \frac{x_i}{g(\mathbf{x})} \right) - \ln \left( \frac{y_i}{g(\mathbf{y})} \right) \right)^2$$

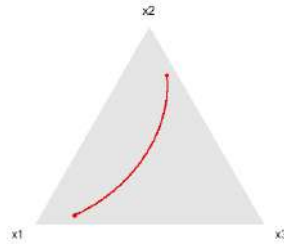
These elements are used for defining orthonormal log-ratio (olr) basis in the simplex ([8]). Once such a basis is created, typical statistical methods are applied to the compositions expressed in olr-coordinates.

## 2 $\mathcal{A}$ -convex set

To ensure that convexity is compatible with the geometry of the simplex, the following definitions will be given using perturbation and powering operations, and therefore, those definitions will differ from the standard euclidean definitions of convexity ([3]).

**Definition 1.** Given  $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}^D$ , the segment that joins them is

$$\overline{\mathbf{x}_1 \mathbf{x}_2} = \{ \mathbf{y} \in \mathcal{S}^D \mid \mathbf{y} = \lambda \odot \mathbf{x}_2 \otimes (1 - \lambda) \otimes \mathbf{x}_1, \lambda \in [0, 1] \}.$$

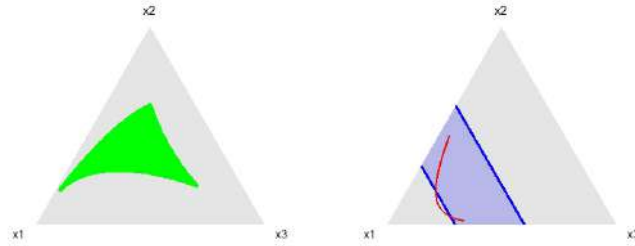


**Fig. 1** Ternary diagram: a segment in  $\mathcal{S}^3$ .

**Definition 2.** A set  $\mathcal{B} \subseteq \mathcal{S}^D$  is  $\mathcal{A}$ -convex if  $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B} \longrightarrow \overline{\mathbf{x}_1 \mathbf{x}_2} \in \mathcal{B}$ .

## 3 Some convex sets formed using linear borders

When we look for set definitions in the simplex, the most common and basic sets are defined constraining one component,  $a < x_i < b$  ([4][5]), or constraining a ratio of two components,  $a < x_i/x_j < b$  ([7][10]).



**Fig. 2** Ternary diagram with typical sets: (Left)  $\mathcal{A}$ -convex triangle (green); (Right) Non  $\mathcal{A}$ -convex strip (blue)

**Definition 3.** An  $\mathcal{A}$ -hyperplane in the simplex is the set

$$\begin{aligned} \Pi_{\alpha_1, \dots, \alpha_D}(k) &= \{ \mathbf{x} \in \mathcal{S}^D \mid \prod_{i=1}^D x_i^{\alpha_i} = k, \sum_{i=1}^D \alpha_i = 0, k \in \mathbb{R}^+ \} \\ &= \{ \mathbf{x} \in \mathcal{S}^D \mid \sum_{i=1}^D \alpha_i \ln(x_i) = \ln(k), \sum_{i=1}^D \alpha_i = 0, k \in \mathbb{R}^+ \}. \end{aligned}$$

A particular case of  $\mathcal{A}$ -hyperplane is  $\Pi_{i,j}(k) = \{ \mathbf{x} \in \mathcal{S}^D \mid x_i/x_j = k \}$ . This  $\mathcal{A}$ -hyperplane  $\Pi_{i,j}(k)$  splits the simplex into two half-spaces:

$$\begin{aligned} \Pi_{i,j}^+(k) &= \{ \mathbf{x} \in \mathcal{S}^D \mid x_i/x_j > k \} \\ \Pi_{i,j}^-(k) &= \{ \mathbf{x} \in \mathcal{S}^D \mid x_i/x_j < k \} \end{aligned}$$

**Proposition 1.** The sets  $\Pi_{i,j}^+(k)$  and  $\Pi_{i,j}^-(k)$  are  $\mathcal{A}$ -convex sets.

For example, Figure 3 (left) shows a red parallelogram defined as intersection of four  $\Pi_{i,j}^{+/-}(k_n)$ . This parallelogram is an  $\mathcal{A}$ -convex set. Figure 3 (right) shows the parallelogram in the olr-space, that is, expression the data in olr-coordinates. It's easy to generalize the result given in Proposition 1 to a generic half-space  $\Pi_{\alpha_1, \dots, \alpha_D}^{+/-}(k)$ .

**Proposition 2.** The sets  $\Pi_{\alpha_1, \dots, \alpha_D}^{+/-}(k)$  are  $\mathcal{A}$ -convex sets.

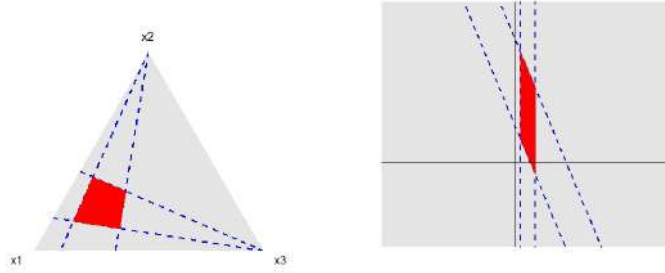
**Definition 4.** We define the hypersurface with constant component  $x_i$  as

$$\Sigma_i(k) = \{ \mathbf{x} \in \mathcal{S}^D \mid x_i = k, 0 < k < 1 \}$$

In this case, the hypersurface splits the simplex into two complementary sets.

$$\begin{aligned} \Sigma_i^-(k) &= \{ \mathbf{x} \in \mathcal{S}^D \mid x_i < k, 0 < k < 1 \} \\ \Sigma_i^+(k) &= \{ \mathbf{x} \in \mathcal{S}^D \mid x_i > k, 0 < k < 1 \} \end{aligned}$$

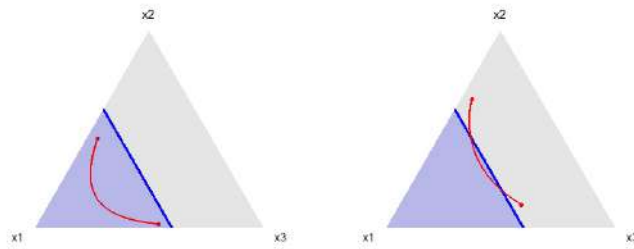
What is a convex set in compositional data analysis?



**Fig. 3** A typical parallelogram: (Left) in the ternary diagram; (Right) in the olr-space.

**Proposition 3.** *The set  $\Sigma_i^+(k)$  is  $\mathcal{A}$ -convex set.*

Figure 4 shows two examples for the above sets. Figure 4 (left) shows the  $\mathcal{A}$ -convex set  $\Sigma_1^+(0.4) = \{\mathbf{x} \in \mathcal{S}^3 | x_1 > 0.4\}$ . On the other hand, Fig. 4 (right) shows a typical non  $\mathcal{A}$ -convex set  $\Sigma_1^-(0.4) = \{\mathbf{x} \in \mathcal{S}^3 | x_1 < 0.4\}$ , the gray region. Note that the red segment is not fully included in the set.



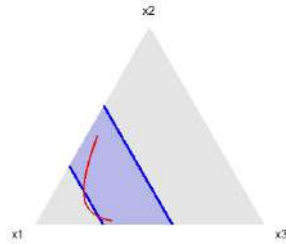
**Fig. 4** Ternary diagram: (Left)  $\Sigma_1^+(0.4) = \{\mathbf{x} \in \mathcal{S}^3 | x_1 > 0.4\}$ ; (Right)  $\Sigma_1^-(0.4) = \{\mathbf{x} \in \mathcal{S}^3 | x_1 < 0.4\}$ .

In addition, let  $\mathcal{B} = \{\mathbf{x} \in \mathcal{S}^3 | 0.4 < x_1 < 0.7\}$  be a typical strip in the simplex. This set is  $E$ -convex but it is not an  $\mathcal{A}$ -convex set. Note that  $\mathcal{B} = \Sigma_1^+(0.4) \cap \Sigma_1^-(0.7)$ . Because  $\Sigma_1^-(0.7)$  is not  $\mathcal{A}$ -convex,  $\mathcal{B}$  is not an  $\mathcal{A}$ -convex set. So, the lack of  $\mathcal{A}$ -convexity is due to the border  $\{\mathbf{x} \in \mathcal{S}^3 | x_1 = 0.7\}$ .

## References

1. Aitchison, J.: The Statistical analysis of compositional data. The Blackburn Press, (1986).





**Fig. 5** Ternary diagram: The blue strip is not an  $\mathcal{S}$ -convex set.

2. Barceló, C., Martín, J. A.: The mathematics of compositional analysis. *Austrian Journal of Statistics*. (2016) doi: 10.17713/ajs.v45i4.142
3. Boyd, S. P., Vandenberghe, L.: *Convex optimization*. Cambridge University Press, (2004).
4. Chen, R., Zhang, Z., Feng, C., Hu, K., Li, M., Li, Y., Shimizu, K., Chen, N., Sugiura, N.: Application of simplex-centroid mixture design in developing and optimizing ceramic adsorbent for As(V) removal from water solution. *Microporous and Mesoporous Materials*. (2009) doi: 10.1016/j.micromeso.2009.12.010
5. Coetzer, R., Haines, L. M.: The construction of D- and I-optimal designs for mixture experiments with linear constraints on the components. *Chemometrics and Intelligent Laboratory Systems*.(2017) doi: 10.1016/j.chemolab.2017.10.007
6. Juditsky, A.;Nemirovski, A.: *Statistical Inference via Convex Optimization*., Princeton University Press, (2020).
7. Lo Huang, M.-N., Huang, M.-K.:  $\phi$  p-optimal designs for a linear log contrast model for experiments with mixtures. *Metrika*. (2009) doi: 10.1007/s00184-008-0190-7
8. Martín-Fernández, J. A.: Comments on: Compositional data: the sample space and its structure. *Test*. (2019) doi: 10.1007/s11749-019-00672-4
9. Pawlowsky-Glahn,V., Egozcue, J.J., et al.: *Modeling and Analysis of Compositional Data*. Wiley, (2015).
10. Wang, X., Wang, H., Wang, Z., Yuan, J.: Convex clustering method for compositional data modeling. *Soft Computing*. (2020) doi 10.1007/s00500-020-05355-z

# Compositional Analysis on the Functional Distribution of Extended Income

## *Analisi composizionale della distribuzione funzionale del reddito esteso*

Elena Dalla Chiara and Federico Perali<sup>1</sup>

**Abstract** This study analyses the functional distribution of extended income implementing a compositional analysis procedure. We specify a linear regression model with the three parts of the extended income (labor, wealth and time or unpaid labor) as compositional response variables to investigate the influence of both family background and the socio-demographic characteristics on the three parts.

**Abstract** *Questo studio analizza la distribuzione funzionale del reddito esteso applicando un'analisi composizionale. Abbiamo specificato un modello di regressione lineare con le tre componenti del reddito esteso (lavoro, ricchezza e tempo o lavoro non pagato) come variabili esplicative per studiare l'influenza delle caratteristiche socio-demografiche nel determinare le tre parti.*

**Key words:** compositional analysis, current extended income, functional distribution of income, isometric log-ratio transformation

## 1 Introduction

The analysis of the functional distribution of income, that refers to the amounts of income paid to various individuals or households, is recently receiving renewed attention because increases in the share of capital in the income portfolio of an individual may be associated with an increase in inequality. An individual may compose her/his own income with the remuneration of her/his labour endowment, with rents from property such as land or apartments and from her/his holdings of company shares or government bonds.

---

<sup>1</sup> Elena Dalla Chiara, Interdepartmental Center of Economic Documentation (CIDE), University of Verona, email: elena.dallachiara@univr.it

Federico Perali, University of Verona (Italy), Department of Economics and CHILD, email: federico.perali@univr.it

To appreciate the relevance of the functional analysis of the distribution of income, Atkinson (2009) cites Mark Blaug who states that ‘the great mystery of the modern theory of distribution is why anyone regards the share of wages and profits as an interesting problem’. It is normally recognized that changes in income inequality across a wide range of countries have been driven mainly by changes in the inequality of wages, while the distribution of income between labor and capital is not considered to be a major factor. Atkinson (2009) contends that since the 1960s, factors shares have been downplayed and that researchers concerned with the personal distribution of income make no direct link with factor shares.

The most common functional decomposition is between the labor share and the capital share. Normally, labor shares also include benefits, pensions and self-employment income, while capital shares cover interests, rents, and other business payments. Our novel contribution is to analyze the role of factor shares, such as labor and wealth, explaining the link between the personal and functional distribution of income by including a new element that we obtain by distinguishing the use of the labor factor in paid and unpaid work activities such as housework and care of young and old people. The sum of these three components, disposable labor income, current wealth and the value of unpaid work invested in domestic production, forms the measure of current extended income that we study. In this framework, observations can be treated as compositional data since the three parts represent relative contributions of the current extended income of the household.

## 2 Methodology

Compositional data are vectors whose components represent the proportion of percentage of a whole, characterized by the constraint of the constant sum, equal to 1 for the proportions and 100 for the percentages. The sum constraint implies that the variance matrices are singular. Further, component distributions cannot be normally distributed due to the bounded range of values. Therefore, the sample space for compositional vectors differs from the real Euclidean space associated with the unconstrained data. The sum constraint problem can be overcome by implementing the log-ratio transformation of compositional data proposed by Aitchison (1986) because compositions often provide information about the relative values of the components, so composition can be specified in terms of component ratios. Consequently, since log-ratios are mathematically easier to deal with than ratios and a log-ratio transformation provides a one-to-one correspondence on real space, Aitchison developed a methodology based on a variety of log-ratio transformations that allow the use of standard unconstrained multivariate statistics. The choice of divisor is one of the most important aspect to evaluate the log-ratio transformation to adopt: the additive log-ratio transformation, the centred log-ratio transformation or the isometric log-ratio transformation (henceforth *ilr*). Egozcue et al. (2003) suggested the *ilr* transformation that uses orthonormal bases to convert compositions into real coordinates preserving all metric properties. Its disadvantage is the choice of the basis because there is not a canonical basis (Pawlowsky-Glahn et

Compositional Analysis of the Functional Distribution of Extended Income al., 2011). Further, the interpretation of estimated parameters in orthonormal coordinates is rather complex because they have to be interpreted in terms of *scaled* log-ratio under natural logarithm. This issue can be sidestepped by considering orthogonal coordinates that removes scaling constants in orthonormal coordinates.

Many compositional data present missing values or zero values, that can be “rounded zeros” or “structural zeros”, in some compositional parts, so it is not possible to apply a log transformation. In our data we have to deal with structural zeros, that are true zeros, only in the wealth component. There are different methodologies to replace structural zeros before computing any type of ratio or log-ratio transformations. We use the transformation  $y^* = [y(N-1) + 1/C]/N$  proposed by Smithson and Verkuilen (2006) where  $y$  is the share,  $N$  is the number of observations and  $C$  is the number of components. After the zero values transformation in wealth part, we consider the *ilr* transformation in orthogonal coordinates in order to interpret easier the parameter estimations. Estimated parameters have to be interpreted as  $2^{\text{estimates}}$  times (Fišerová et al., 2016; Müller et al., 2018). All compositional analyses are performed using the R-package “compositions” (van den Boogaart et al., 2021) improved by the authors to adjust for inconsistencies in the order of execution of the *ilr* transformation.

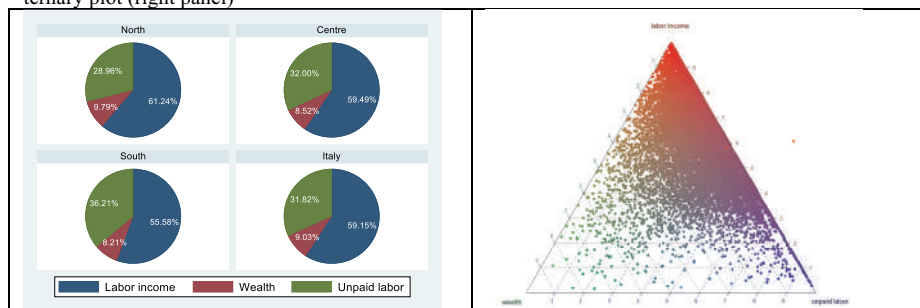
### 3 Results

Our components are disposable labor income, current wealth and the value of unpaid labor invested in domestic production that make up the household current extended income. Disposable labor income is the total household income after tax deduction, current wealth includes both assets and return on assets multiplied by the prevailing interest rate in the economy, while the value of household domestic production is the values of goods or services produced by household members using their unpaid activities. We use the integrated database created by Dalla Chiara et al. (2019) to measure the standard of living in Italy with information collected on 2010 because it is the only year, as per our knowledge, that has survey information about relational well-being, income, consumption and time use of the corresponding year (Dalla Chiara and Perali, 2021).

We first analyze the raw share of the three parts to have an overview of the magnitude of each one. Figure 1 (left panel) shows the portfolio of the average Italian family which is composed of 59% of disposable labor income, 32% of the value of unpaid labor from domestic production and 9% of the current value of wealth. In all macro area, labor income plays the main role in defining the current extended income composition. In the South, labor income shows the lower percentage compared to the other geographical areas. The biggest gap in the unpaid labor is detected when comparing the average functional distribution of the North and South with 7 percentage points difference. Each observation in compositional term can be displayed in the ternary plot. As shown in Figure 1 (right panel), most of the compositional data points lie from the high amount of income to a 40/50

composition of the value of unpaid labor. This representation supports the results highlighted in Figure 1 – left panel.

**Figure 1:** The three components of household current extended income: average share (left panel) and ternary plot (right panel)



Based on this evidence, we intend to understand how these compositions are related to household socio-demographic characteristics describing different organizational strategies of the family.

We specify a linear regression model with the three parts of the current extended income as compositional response variables in *ilr* orthogonal coordinates and non-compositional explanatory variables for exploring influences among several socio-demographic information. We use the following explanatory variables: region of residence (three dummies coded as South, North and Centre), age of the household head, education of the household head (three dummies coded as Primary, Middle and High), number of earners in the family, number of family members, the ability to make ends meet (three dummies coded as With difficulty, With some difficulty and Easily), number of sons, dummy variables refer to the household head for Italian nationality, home-owner and older than 64 years old, self-employed and a single-parent. Covariates that have the greatest positive impact on the response coordinates are single-parent and the number of earners for labor income and nationality of the household head, home-owner greater than 64 years old and ability to make ends meet for wealth component (Table 1 – left panel). Being a single parent increases the relative dominance of labor income approximately by 58% ( $2^{0.658}$ ), while for a unit additive change in the number of earners, the ratio of labor income to the mean value of the other compositional responses increases about 26%. The relative dominance of wealth to the “mean value” of the other compositional response increases by 26.5 times for Italian household head and by 15.6 times for home-owner greater than 64 years old. The relative dominance of wealth for the ability to make ends meet “With some difficulty” and “Easily” is 2.6 and 10 times greater than the ability to make ends meet “With difficulty” respectively. In compositional data, the first level of the categorical variable is assumed to have an effect of zero due to the implicit use of the *contrasts* (van de Boogaart et al., 2011). The age of the household head has the same slightly negative influence both in labor income and wealth, this mean that age has a higher impact on unpaid labor. The same considerations apply to the number of household members, but in this case the impact on the unpaid labor is greater (Table 1 – right panel). Right panel of Table 1

Compositional Analysis of the Functional Distribution of Extended Income

reports the influence of each socio-demographic information in the composition of the household current extended income. Paid labor is the most important part in the composition of the household current extended income. Indeed, the expected composition of the current extended income, when all explanatory variables are equal to 0, is made up of 70% from paid labor, about 30% from unpaid labor while the impact of wealth is null. The nationality of the household head and the home-owner household greater than 64 years old have a great impact on wealth composition (92.3% and 88.3% respectively). Wealth is also affected by the ability to make ends meet, it increases with the progressive easily ability to make ends meet, by the household head working condition (if the household head is self-employed wealth accounts for 41.6% of household current extended income). Education level affects wealth part and its impact increases as the level of education increases with a consequent sharp reduction on unpaid labor compared to labor income. Labor income is the most important part for single-parent families (43.9%), while unpaid labor grows with the increase of the number of family members (43.2%). The age of the household head effects with the same intensity the three parts, so its influence is equally important for all components.

**Table 1:** Compositional regression (left panel) and contribution of each variable (right panel)

<i>Variable</i>	<i>Labor Income</i>		<i>Wealth</i>		<i>Labor Income</i>	<i>Wealth</i>	<i>Unpaid Labor</i>
Intercept	7.299	***	-12.129	***	0.702	0.000	0.298
<i>Geographic area</i>							
North	-0.016		0.339	***	0.329	0.375	0.296
Centre	0.003		0.215	*	0.333	0.358	0.309
Age of the hh	-0.139	***	-0.131	***	0.312	0.328	0.360
<i>Education of the hh</i>							
Middle level	-0.050		0.510	***	0.322	0.398	0.280
High level	-0.052		1.305	***	0.304	0.496	0.200
Number of earners	0.330	***	0.761	***	0.378	0.391	0.231
Number of family members	-0.342	***	-0.577	***	0.279	0.289	0.432
<i>Ability to make ends meet</i>							
With some difficulty	-0.723	***	1.970	***	0.196	0.640	0.164
Easily	-1.088	***	3.361	***	0.118	0.804	0.078
Number of sons	0.087		0.464	***	0.344	0.380	0.276
Nationality of the hh	-2.076	***	4.730	***	0.042	0.923	0.035
Home-owner and hh >64	-1.882	***	3.967	***	0.061	0.883	0.056
Self-employed	-0.018		0.683	***	0.325	0.416	0.259
Single-parent	0.658	***	-0.344	*	0.439	0.247	0.314

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1; Hh: household head

## 4 Conclusions

This work describes a compositional analysis approach on the composition of the household current extended income that is made up of labor income, current income

from wealth and the value of unpaid labor from the investment on domestic production. Our aim is to investigate the impact of single socio-demographic variables in the composition of the three parts of household current extended income in order to assess not only the pure monetary aspect but also non-monetary dimensions such as paid and unpaid work in the household. Using compositional analysis methodology, we intend to gain a policy-relevant picture of both family background features and socio-demographic characteristics playing a significant role in explaining both monetary and non-monetary aspects.

## References

1. Aitchison, J.: *The Statistical Analysis of Compositional Data*. Chapman and Hall, London (1986)
2. Atkinson, A.B: Factor shares: the principal problem of political economy? *Oxford Review Economic Policy* (2009) doi: 10.1093/oxrep/grp007
3. Dalla Chiara, E., Menon, M., Perali, F.: An Integrated Database to Measure Living Standards. *Journal of Official Statistics* (2019) doi: 10.2478/jos-2019-0023
4. Dalla Chiara, E., Perali, F.: Relational Well-Being and Poverty in Italy. Working Paper, Department of Economics, University of Verona (2021)
5. Egozcue, J.J., Pawłowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C.: Isometric Logratio Transformation for Compositional Data Analysis. *Mathematical Geology* (2003) doi: 10.1023/A:1023818214614
6. Fišerová, E., Donevska, S., Hron, K., Bábek, O., Vankátová, K.: Practical Aspects of Log-ratio Coordinate Representations in Regression with Compositional Response. *Measurement Science Review* (2016) doi: 10.1515/msr-2016-0029
7. Müller, I., Hron, K., Fišerová, E., Šmahaj, J., Cakirpaloglu, P., Vančáková, J.: Interpretation of Compositional Regression with Application to Time Budget Analysis. *Austrian Journal of Statistics* (2018) doi: 10.17713/ajs.v47i2.652
8. Pawłowsky-Glahn, V., Buccianti, A.: *Compositional Data Analysis. Theory and Applications*. Wiley (2011)
9. Smithson, M., Verkuilen, J.: A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* (2006) doi: 10.1037/1082-989X.11.1.54
10. van den Boogaart, K.G., Tolosana-Delgado, R., Bren, M.: *Compositions: Compositional data analysis*. R package (2021)

# Evaluating seasonal-induced changes in river chemistry using Principal Balances

## *Valutazione dei cambiamenti stagionali nella chimica dei fiumi mediante i Bilanci Principali*

Caterina Gozzi and Antonella Buccianti

**Abstract** Seasonal cycles significantly impact inland water resources, altering their availability, quality and chemical composition. The Principal Balances (PBs) approach was used to assess seasonal-induced compositional changes in the surface water chemistry of the Tiber River Basin (central Italy). The same sequence of PBs obtained for the whole dataset was applied separately to data subsets sampled in different hydrological regimes. The comparison of PBs' density distributions and basins of attraction revealed a higher variability and vulnerability in dry periods and a key role of silicate weathering processes. The method proved to be effective and offers new insights to evaluate seasonal variations and system stability.

**Abstract** *I cicli stagionali hanno un impatto significativo sulle risorse idriche, alterandone la disponibilità, la qualità e la composizione chimica. Il metodo dei Bilanci Principali (PBs) è stato utilizzato per valutare i cambiamenti composizionali indotti dalla stagionalità nella chimica delle acque superficiali del Bacino del Fiume Tevere (Italia centrale). La stessa sequenza di PBs ottenuta per l'intero set di dati è stata applicata separatamente a gruppi di dati campionati durante regimi idrologici diversi. Il confronto delle distribuzioni di densità e dei bacini di attrazione dei PBs ha evidenziato una maggiore variabilità e vulnerabilità nei periodi di magra e un ruolo chiave dei processi di alterazione dei silicati. Il metodo si è dimostrato efficace e offre nuove prospettive per esplorare le variazioni stagionali e la stabilità del sistema.*

**Key words:** River chemistry, Compositional Data, Principal Balances, Seasonal Changes

---

Caterina Gozzi  
University of Florence, Department of Earth Sciences, Via G. La Pira 4, 50121 Firenze, Italy; e-mail: caterina.gozzi@unifi.it

Antonella Buccianti  
University of Florence, Department of Earth Sciences, Via G. La Pira 4, 50121 Firenze, Italy; e-mail: antonella.buccianti@unifi.it



## 1 Introduction

Seasonal variability of hydro-geochemistry is determined by several factors, e.g. changes in precipitation, climate, river discharge, intensity of anthropogenic activities and surface run-off. Seasonal variations are expected to have strong consequences on the concentration of solutes, nutrients and pollutants transported by rivers [12]. In this framework, the use of new compositional methods represents a powerful means to uncover chemical changes from a holistic perspective [7]. The aim of this study is to understand how the seasonality affects the hierarchical variability structure monitored in the surface water composition of the Tiber River catchment (central Italy) [6, 5]. The basin is characterized by a heterogeneous geological, hydro-geological and morphological setting [2], representing an interesting area to assess the effects of seasonal fluctuations.

## 2 Materials and methods

### 2.0.1 Geochemical dataset

The dataset consists of:

1. A total of 160 water samples, belonging to the Tiber river, its major and minor tributaries, collected in 2017 during different hydrological conditions (winter-spring and summer) as a first comprehensive survey of the catchment;
2. 62 samples collected during additional monitoring campaigns in 2018 (winter-spring and summer) from selected locations (i.e. Tiber river and main tributaries).

The analysis was performed by considering 10 major elements which define the main composition of the waters. Additional information regarding sampling and analytical methods can be found in [5].

### 2.0.2 Principal Balances Approach

Geochemical data are a typical example of compositional data. They are positive and closed data in which the relevant knowledge is enclosed in the ratios between the components [1]. In order to transform compositional data into real coordinates to be analyzed with classical statistical and geostatistical methods, a solution is to adopt the isometric log-ratio (ilr) transformation [4]:

$$\text{ilr}_i(\mathbf{x}) = \sqrt{\frac{r_{i+} \cdot r_{i-}}{r_{i+} + r_{i-}}} \ln \frac{g(\mathbf{c}_{i+})}{g(\mathbf{c}_{i-})}, \quad i = 1, 2, \dots, D-1, \quad (1)$$

where  $\mathbf{c}_{i+}$  and  $\mathbf{c}_{i-}$  are the groups of components separated in the  $i$ -th step of the Sequential Binary Partition (SBP);  $r_{i+}$  and  $r_{i-}$  are the numbers of parts included in

$c_{i+}$  and  $c_{i-}$ , respectively; and  $g(\cdot)$  is the geometric mean of its argument.

A basic point is how to choose the best SBP to get ilr-coordinates easily interpretable geochemically. In the research community, several methods have been proposed for this purpose, such as to select a partition that follows, where applicable, weathering reactions involving the chemical compounds [3]. An attractive alternative to the above-mentioned criterion is that proposed by [8]. This method enables the creation of a sequence of isometric log-ratio coordinates, named Principal Balances, which sequentially maximize the explained variance in a data set. However, this approach is not robust and outlying observations could strongly influence data variability, thereby affecting the resultant balances. With such awareness, PBs were calculated in R, considering the entire database of 222 water samples. The obtained SBP of the composition in different subsets of chemical variables was then applied to winter-spring and summer (2017-2018) data sets, separately. Following this procedure, it was expected to highlight the impacts of seasonal fluctuations on the resulting PBs. The density distributions of the PBs for the two seasons were then compared in a single plot, investigating their differences in shape and variability. The latter provide valuable information about the governing dynamics of the natural system originating the distribution [10]. Particularly, [11] highlighted that a density distribution or the related histogram can be reversed revealing the presence of basins of attraction in the data structure. From this perspective, each mode depicts the hole of a basin of attraction capturing data, and the lower frequency areas represent the barriers separating different dynamical states.

## 2.1 Results and Discussion

### 2.1.1 Overview on river water variability

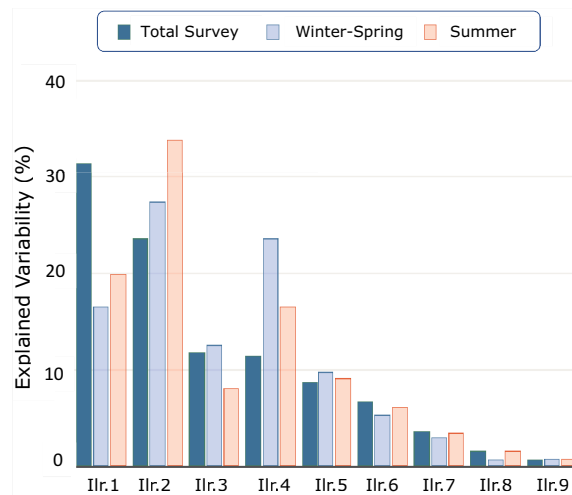
PBs calculated from the entire dataset, and obtained using the decreasing variance criterion, were the following:

- **Ilr.1)**  $\text{NO}_3^- \parallel \text{K}^+, \text{NH}_4^+, \text{Cl}^-, \text{Na}^+, \text{Mg}^{2+}, \text{HCO}_3^-, \text{Ca}^{2+}, \text{F}^-, \text{SO}_4^{2-}$
- **Ilr.2)**  $\text{K}^+ \parallel \text{NH}_4^+, \text{Cl}^-, \text{Na}^+, \text{Mg}^{2+}, \text{HCO}_3^-, \text{Ca}^{2+}, \text{F}^-, \text{SO}_4^{2-}$
- **Ilr.3)**  $\text{NH}_4^+ \parallel \text{Cl}^-, \text{Na}^+, \text{Mg}^{2+}, \text{HCO}_3^-, \text{Ca}^{2+}, \text{F}^-, \text{SO}_4^{2-}$
- **Ilr.4)**  $\text{Cl}^-, \text{Na}^+ \parallel \text{Mg}^{2+}, \text{HCO}_3^-, \text{Ca}^{2+}, \text{F}^-, \text{SO}_4^{2-}$
- **Ilr.5)**  $\text{Mg}^{2+}, \text{HCO}_3^-, \text{Ca}^{2+} \parallel \text{F}^-, \text{SO}_4^{2-}$
- **Ilr.6)**  $\text{F}^- \parallel \text{SO}_4^{2-}$
- **Ilr.7)**  $\text{Mg}^{2+} \parallel \text{HCO}_3^-, \text{Ca}^{2+}$
- **Ilr.8)**  $\text{Cl}^- \parallel \text{Na}^+$
- **Ilr.9)**  $\text{HCO}_3^- \parallel \text{Ca}^{2+}$

The symbol || separates chemical species at numerator and denominator in the log-ratios. The first balance (Ilr.1) accounts for 31% of total variance followed by Ilr.2 and Ilr.3, explaining 24% and 12%, respectively. On the contrary, last balances (Ilr.8 and Ilr.9) show the lowest variability (2% and 1%, respectively). As highlighted by [5], the results monitor a hierarchy in the variability of the water composition characterized by the presence, within the first balances, of major species strongly linked to human pressures which oppose to those related to water-rock interaction processes in the last ones. Similar results were also obtained by [6], using the cascade application of robust compositional biplots.

### 2.1.2 Influence of Seasonal fluctuations

The bar plot represented in Figure 1 highlights the differences in the explained variability between the total survey and the data subsets separated according to the sampling period. The Ilr.1 balance ( $\text{NO}_3^-$  vs. the remaining composition) explains a

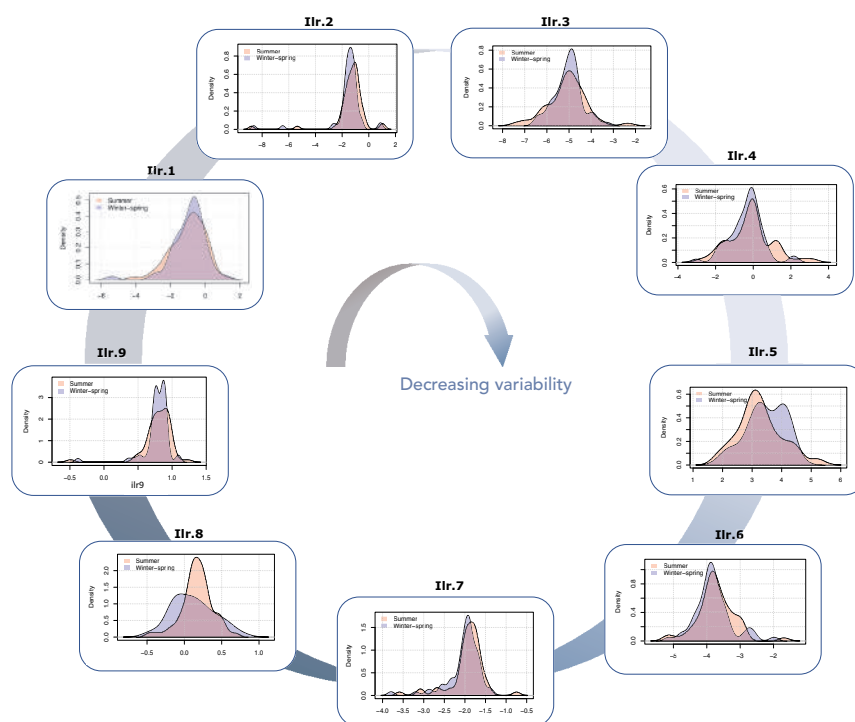


**Fig. 1** Bar plot of the different variability explained by the Principal Balances calculated for the three data sets ( i.e. total survey, winter-spring and summer data subsets).

higher percentage of variability when data from the total survey are considered: 31% compared to 18% and 20% during low and high discharge periods, respectively. The total database differs from the other two because of the large number of samples belonging to small streams and creeks. This leads to assume a greater dominance of nitrates in the composition of these watercourses. Differently, Ilr.2 and Ilr.4 variability is better captured when the datasets are considered separately, thus indicating their stronger seasonal dependence. This might be explained by the influence on river chemistry of silicate weathering processes, which, in turn, strongly depend on runoff fluctuations [9].

The comparison of the density curves (Fig. 2) shows that PBs are characterized by a higher variability in summer with smoother density distributions with respect to

those in the winter-spring season. Consequently, the respective basins of attraction are generally deeper during winter-spring time, suggesting an enhanced resistance to compositional variations [5]. In fact, the deeper is the hole, the greater energy is required to take the system out of the basin, preventing potential transition to alternative states [11]. Conversely, during summer, lower saddles more likely foster the development of new states. This denotes a weaker system predictability and a higher vulnerability of the river chemistry to perturbations.



**Fig. 2** Density curves for Principal Balances (Iir.1-Iir.9) comparing the frequency distributions of the data for the two different sampling periods.

The shape of the density distributions undergoes some modifications as a result of seasonal changes, showing differences in terms of deepening and/or displacement along the x-axis of the corresponding basins of attraction. However, PBs are not subject to dramatic changes, except for the  $\text{Cl}^- \parallel \text{Na}^+$  balance (Iir.8) which instead has a significantly smoother distribution in the high flow periods. This change might be explained by an enhanced water-rock interaction during winter-spring, leading to a higher amount of  $\text{Na}^+$  derived from silicate weathering reactions.

### 3 Conclusions

Principal Balances has led to create ilr-coordinates driven by data variability, which reflect fairly well geochemical processes taking place within the studied area. The results indicate a higher variability and vulnerability of the surface waters of the Tiber River Basin to geochemical threats in dry periods. A relevant seasonal effect on water chemistry was also detected, likely related to silicate weathering processes. In conclusion, the PBs method, coupled with the interpretation of the basin of attractions of the data, offers new insights to evaluate seasonal changes and system predictability. Nevertheless, improvements to the PBs approach in terms of robustness are needed to enhance its reliability for future applications.

**Acknowledgements** Funds of the University of Florence 2020-2021 were used to sustain the research (A.B). The Tuscany Region (Italy) is thanked for funding the three-year Ph.D. scholarship (C.G.).

### References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society Series B* **44**(2), 139–177 (1982)
2. Boni, C., Bono, P., Capelli, G.: Hydrogeological scheme of central Italy. *Memories of the Italian Geological Society* **35**, 991–1012 (1986)
3. Buccianti, A., Zuo, R.: Weathering reactions and isometric log-ratio coordinates: Do they speak to each other? *J. Appl. Geochem* **75**, 189–199 (2016)
4. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric log-ratio transformations for compositional data analysis. *Math. Geol.* **35**(3), 270–300 (2003)
5. Gozzi, C.: Weathering and transport processes investigated through the statistical properties of the geochemical landscapes: the case study of the Tiber river basin (Central Italy). *PLINIUS* **46**, 49–55 (2020)
6. Gozzi, C., Filzmoser, P., Buccianti, A., Vaselli, O., Nisi, B.: Statistical methods for the geochemical characterisation of surface waters: The case study of the Tiber River basin (Central Italy). *Comput Geosci* **131**, 80–88 (2019)
7. Gozzi, C., Sauro Graziano, R., Buccianti, A.: Part–Whole Relations: New Insights about the Dynamics of Complex Geochemical Riverine Systems. *Minerals* **10**(501) (2020)
8. Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R.: Advances in principal balances for compositional data. *Math. Geosci.* **50**, 273–298 (2017)
9. Mortatti, J., Probst, J.L.: Silicate rock weathering and atmospheric/soil CO<sub>2</sub> uptake in the Amazon basin estimated from river water geochemistry: seasonal and spatial variations. *Chem. Geol.* **197**(1), 177–196 (2003)
10. van Rooij, M., Nash, B., Rajaraman, S., Holden, J.: A fractal approach to dynamic inference and distribution analysis. *Front. Physiol.* **4**(1), 1–11 (2013)
11. Scheffer, M., Carpenter, S., Lenton, T., Bascompte, J., Brock, W., Dakos, V., van de Koppel, J., van de Leemput, I., Levin, S., van Nes, E., Pascual, M., Vandermeer, J.: Anticipating critical transitions. *Science* **338**, 344–348 (2012)
12. Vega, M., Pardo, R., Barrado, E., Debán, L.: Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Res.* **32**(12), 3581–3592 (1998)

# Compositional Data Techniques for the Analysis of the Ragweed Allergy

## *Un Approccio Compositzionale all'Analisi dell'Allergia ai Pollini di Ambrosia*

Gianna S. Monti, Maira Bonini, Valentina Ceriotti, Matteo Pelagatti and Claudio M. Ortolani

**Abstract** Ragweed is a primary allergenic risk factor in many countries of the world, in particular the North-West Milan area is one of the most infested zone in Europe. We present here a way of applying regression for compositional data analysis, based on Aitchison geometry, to investigate the influence of airborne ragweed pollen on symptoms composition (conjunctivitis/rhinitis/asthma). We performed compositional regression using both least squares and robust MM estimators to cope with outliers contamination.

**Abstract** *L'ambrosia è un fattore di rischio allergenico primario in molti paesi del mondo, in particolare la zona nord-ovest di Milano è una delle zone più infestate d'Europa. Presentiamo qui un modo di applicare la regressione per l'analisi dei dati composizionali, basata sulla geometria di Aitchison, al fine di studiare l'influenza dei pollini di ambrosia presenti nell'aria sulla composizione dei sintomi (congiuntivite/rinite/asma). Sono considerate sia la regressione classica basata sul principio dei minimi quadrati che la regressione robusta di tipo MM per far fronte alla contaminazione da valori anomali.*

**Key words:** Logratio, Aitchison geometry, Compositional Regression.

---

Gianna Serafina Monti and Matteo Pelagatti  
Department of Economics, Management and Statistics, University of Milano Bicocca, e-mail:  
gianna.monti@unimib.it

Maira Bonini and Valentina Ceriotti  
Agency for Health Protection of Metropolitan Area of Milan, Milan, Italy

Claudio Maria Ortolani  
Istituto Allergologico Lombardo, Casa di Cura Ambrosiana, Cesano Boscone, Milan, Italy

## 1 Introduction

The expansion on a global scale of ragweed pollen and its allergenic potentials is an increasing problem for European countries and in particular for northern Italy [4]. To investigate the influence of airborne ragweed pollen on symptoms an epidemiological observational study was conducted. Ragweed pollen grains, expressed as particles per cubic meter of air (p/m<sup>3</sup>), were sampled by 3Hirst volumetric traps located in the North-West Milan area, respectively in Legnano, Magenta and Rho. Average of the daily ragweed pollen concentrations of the 3 stations were used in the analysis. Ragweed allergic 71 patients (25 allergy immunotherapy tablet (AIT) treated, and 46 not AIT treated) were enrolled by 5 allergy clinics close to the pollen traps. During the ragweed pollen season 2014 (from August, 1st to mid September 2014), they compiled a daily diary of the symptoms and drugs taken.

The aim of this contribution is to assess the influence of airborne ragweed pollen on symptoms composition (conjunctivitis/rhinitis/asthma) in the two cohorts of patients, recognizing its compositional nature. Compositional Data Analysis (CoDa) techniques allow to reveal underlying patterns of symptoms composition compatible with their natural geometry. This is a novel application of CoDa method in this field of research from a relative perspective.

Following the well known compositional geometry, a compositional vector  $\mathbf{x} = (x_1, \dots, x_D)$  is considered a  $D$ -part composition if the relevant information is contained in the ratio of its parts, rather than in their absolute values. The simplex of dimension  $D$ ,  $\mathcal{S}^D = \{\mathbf{x}_i = (x_1, \dots, x_D)^T : x_i > 0, \text{ for } i = 1, \dots, D, \text{ and } \sum_{i=1}^D x_i = 1\}$ , is their appropriate sample space. The application of conventional statistical techniques to compositional data may induce unreliable conclusions, thus an appropriate approach which is capable to take into account their own mathematical structure is recommended. The log-ratio approach provides a set of transformations, or coordinates, that allow to extract the relative information from a composition, and thus to apply conventional statistical techniques to the transformed compositional data samples [1, 2]. The most renowned are the additive logratio (*alr*), the centered logratio (*clr*), and the isometric logratio (*ilr*) transformation [3]. We will refer to the latter in this contribution, albeit linear regression is known to provide exactly the same result whichever logratio transformation is used, thanks to its affine equivariance [9].

In this contribution we consider a compositional regression in which the conjunctivitis/rhinitis/asthma symptoms composition is explained by airborne ragweed pollen load and AIT treatment. As the response variables is compositional, we express the data isometrically in the real Euclidean space, making possible to use the classical statistical tools. Both classical least squares regression and robust MM regression are considered for the parameter estimation, to cope with potential outliers.

The contribution is organized as follows: in Sect. 2 the regression model with compositional response is briefly introduced, and Sect. 3 illustrates an application to the analysis of features of ragweed allergy.

## 2 Linear Model with Compositional Response

Pawlowsky-Glahn [8] introduced the principle of working on coordinates which states that compositions should be expressed as coordinates with respect to an orthonormal basis of the Euclidean structure of  $\{\mathcal{S}^D, \oplus, \odot, \langle \cdot \rangle_a\}$ , where  $\oplus, \odot, \langle \cdot \rangle_a$  represent respectively perturbation and power operations, and the inner product with respect Aitchison geometry. The mentioned *ilr* transformation provides a way to compute orthonormal coordinates

$$\text{ilr}(\mathbf{x}) := \mathbf{x}^* = \mathbf{V}^T \ln \mathbf{x},$$

where  $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_{D-1})$  is a matrix with orthonormal columns. Conversely, given a vector of coordinates  $\mathbf{x}^* = (x_1^*, \dots, x_{D-1}^*)$  the inverse isometric transformation provides a convenient way to apply it to the basis:

$$\text{ilr}^{-1}(\mathbf{x}^*) = C(\exp\{\mathbf{V}\mathbf{x}^*\}),$$

where  $C$  is the closure operator.

We revise briefly the linear model with compositional response in the sense of the Aitchison geometry [9]. A model with compositional response assumes that a random composition  $\mathbf{Y}$  is a linear function of some explanatory random variables, namely  $X_1, X_2, \dots, X_p$ , which gives the expected conditional value of some normally distributed composition,

$$\widehat{\mathbf{Y}} = \bigoplus_{i=0}^p X_i \odot \mathbf{b}_i, \quad \mathbf{Y} \sim N_{\mathcal{S}^D}(\widehat{\mathbf{Y}}, \boldsymbol{\Sigma}_\varepsilon), \quad (1)$$

where  $N_{\mathcal{S}^D}(\widehat{\mathbf{Y}}, \boldsymbol{\Sigma}_\varepsilon)$  stands for the normal distribution on the simplex of  $\mathbf{Y}$  [6].

Following the principle of working in coordinates it is straightforward to transform the model (1) into a multiple regression problem, yielding

$$\widehat{\mathbf{Y}}^* = \bigoplus_{i=0}^p X_i \odot \mathbf{b}_i^*, \quad \mathbf{Y}^* \sim N^{D-1}(\widehat{\mathbf{Y}}^*, \boldsymbol{\Sigma}_\varepsilon). \quad (2)$$

The model parameters are thus the slopes  $\mathbf{b}_0^*, \mathbf{b}_1^*, \dots, \mathbf{b}_p^*$ , and the residual covariance matrix  $\boldsymbol{\Sigma}_\varepsilon$ . Note that it is usual to take  $X_0 \equiv 1$ , and then  $\mathbf{b}_0^*$  is more readily as it represents the model intercept in the logratio coordinate system chosen. Once estimates of the vector coefficients are available, they can be back-transformed to compositional coefficients, e.g.  $\widehat{\mathbf{b}} = \text{ilr}^{-1}(\widehat{\mathbf{b}}^*)$ .



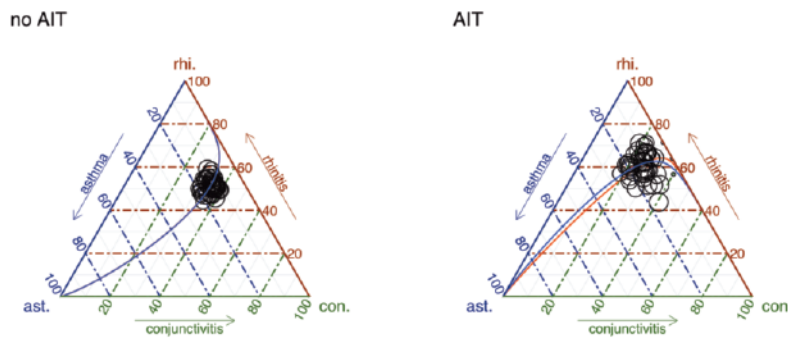
### 3 Compositional Features of Ragweed Allergy

We focus on the effects between the conjunctivitis/rhinitis/asthma composition and the covariates: airborne ragweed pollen, expressed in log-scale, AIT treatment, treated as dummy variable with one for noAIT and zero for AIT, and their interaction.

The available data were transformed in the following *ilr* coordinates:

$$\mathbf{V}^T = \begin{pmatrix} -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{pmatrix}, \quad \begin{aligned} y_1^* &= \frac{1}{\sqrt{6}} \ln \frac{y_{ash}^2}{y_{rhi} y_{con}} \\ y_2^* &= \frac{1}{\sqrt{2}} \ln \frac{y_{rhi}}{y_{con}} \end{aligned} \quad (3)$$

A compositional regression model of the form of equation (2) was fitted by the classical LS and MM method [5].



**Fig. 1** Ternary diagram of the conjunctivitis/rhinitis/asthma symptoms composition in the two groups of patients along with the predictions of both the LS (red) and MM (blue) models. Symbol size is inversely proportional to the weights computed in robust regression.

Fig. 1 shows the model predictions with respect to the Aitchison geometry for the original composition for the classical (red) and the robust (blue) model. Symbol size is inversely proportional to the weights computed in robust regression. LS and the MM approaches give similar coefficients.

The classical LS and robust MM-estimates for the coefficients for the *ilr* coordinates defined in equation (3), and corresponding p-values, are provided in Table 1. The back-transformed coefficients are reported in the last three columns. The LS estimates show that the ratio rhinitis-to-conjunctivitis is not affected by load pollen, but it depends strongly to the AIT treatment, while their relation to ashma does depend on both covariates and their interaction. Similar conclusions could be derived for the MM estimates.

We would interpret the intercept as the expected composition for  $\log(\text{conc}) = 0$  in AIT treated patients. The *ilr* back-transformed intercept provides the conditional

**Table 1** Regression models of conjunctivitis/rhinitis/asthma symptoms against (log) airborne ragweed pollen concentration and AIT treatment, using LS and MM regression. The columns refer to the estimated parameters for the *itr* coordinates, the corresponding p-values, and the back-transformed regression coefficients.

	$y_1^*$	p-value	$y_1^*$	$y_2^*$	p-value.	$y_2^*$	conjunctivitis	rhinitis	asthma
LS: intercept	-1.295	< 0.001	0.573	< 0.001	0.281	0.632	0.086		
LS: log(conc)	0.148	< 0.001	0.017	0.108	0.309	0.316	0.375		
LS: noAIT	0.497	< 0.001	-0.313	< 0.001	0.321	0.206	0.473		
LS: interaction	-0.198	< 0.001	-0.007	0.654	0.361	0.357	0.282		
MM: intercept	-1.013	< 0.001	0.607	< 0.001	0.263	0.620	0.117		
MM: log(conc)	0.063	0.005	0.012	0.213	0.322	0.328	0.351		
MM: noAIT	0.213	0.003	-0.347	< 0.001	0.381	0.233	0.386		
LS: interaction	-0.112	< 0.001	-0.002	0.820	0.349	0.348	0.304		

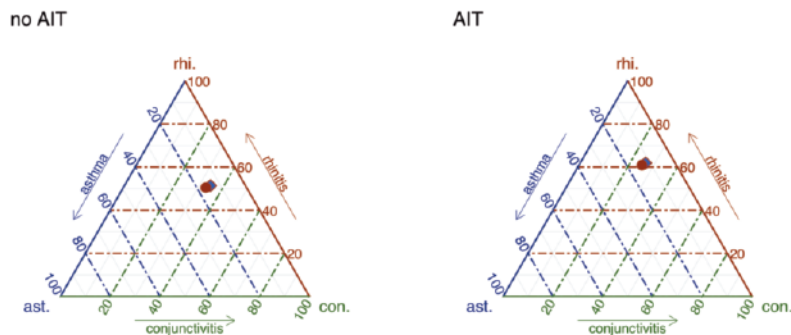
average percentages of symptoms. Rhinitis is the overwhelming majority, while the estimated proportions of asthma is quite low. The slope may be interpreted as the perturbation applied to the composition if  $\log(\text{conc})$  increases one unit [10].

We can use the Type II MANOVA Tests to check for the joint significance of the influence of the two covariates on the two *itr* coordinates (see Table 2). The influences of the two covariates have a p-values respectively equal to 0.013 and < 0.001, there is thus a strong evidence that the pollen load, as well as the AIT treatment, influence the composition of the symptoms.

**Table 2** Multivariate Pillai's trace MANOVA test.

	Df	test stat	approx F	num Df	den Df	Pr(>F)
log(conc)	1	0.096	4.606	2	87	0.013
noAIT	1	0.585	61.218	2	87	< 0.001
interaction	1	0.238	13.606	2	87	< 0.001

This is a clear message for targeted public health interventions: an increase of pollen exposure contributes more to the increase of relative contributions of asthma to the total amount of symptoms. To better understand the effect of increase of one gram of pollen in log-scale, in Fig. 2 is shown the original center, obtained as the closure of the univariate geometric means of each component, in each of the two cohorts, represented by blue diamond. Thus the perturbed center, according to the LS fitted regression model, is shown by red point. In each cohort the center of the data is moved to the baricenter of the ternary diagram, towards a relative increase of asthma symptoms. This effect could intensify the risk of hospitalization for asthma, especially in noAIT patients, as confirmed also by recent studies [7].



**Fig. 2** Ternary diagram of the conjunctivitis/rhinitis/asthma symptoms composition in the two cohorts of patients. Original center is represented by blue diamond, and perturbed center by red point according to the LS fitted regression model.

## References

- [1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.
- [2] Billheimer, D., Guttorp, P., and Fagan, W. (2020). Statistical interpretation of species composition. *J Am Stat Assoc*, **96**(456):1205–1214.
- [3] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Math Geol*, **35**(3):279–300.
- [4] Mandrioli, P., Di Cecco, M., and Andina, G. (1998). Ragweed pollen: The aeroallergen is spreading in Italy. *Aerobiologia*, **14**(13):13–20.
- [5] Maronna, R., Martin, R., and Yohai, V. (2006). *Robust statistics: theory and methods*. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., New York, NY.
- [6] Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2008). A critical approach to probability laws in geochemistry. *Math Geosci*, **40**(5):489–502.
- [7] Osborne, N. J., Alcock, I., Wheeler, B. W., Hajat, S., Sarran, C., Clewlow, Y., McInnes, R. N., Hemming, D., White, M., Vardoulakis, S., and Fleming, L. E. (2017). Pollen exposure and hospitalization due to asthma exacerbations: daily time series in a European city. *Int J Biometeorol*, **61**(10):1837–1848.
- [8] Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In Thió-Henestrosa, S. and Martín-Fernández, J. A., editors, *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*. University of Girona, Girona (Spain).
- [9] van den Boogaart, K., Filzmoser, P., Hron, K., Templ, M., and Tolosana-Delgado, R. (2020). Classical and robust regression analysis with compositional data. *Math Geosci*.
- [10] von Eynatten, H., Pawlowsky-Glahn, V., and Egozcue, J. J. (2002). Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams. *Math Geol*, **34**(3):249–257.

## 4.29 Spatial data analysis

# Spatial multilevel mixed effects modeling for earthquake insurance losses in New Zealand

## *Modello spaziale multilivello a effetti misti per le perdite dell'assicurazione contro i terremoti in Nuova Zelanda*

F. Marta L. Di Lascio and Selene Perazzini

**Abstract** In disaster insurance, the assessment of spatial correlation of regions highly exposed to earthquakes is essential to diversify the risk in a portfolio. Earthquake hazard is extremely relevant in New Zealand due to its proximity to a seismic fault and the diversification of risk-based earthquake insurance premiums is crucial due to the high heterogeneity of population density. In this paper we explore the usefulness of multilevel mixed effects modeling in taking into account spatial correlation of earthquake losses in New Zealand. Total losses due to claims reported between 2000 and 2018 are modeled by assuming that wards in the same regions are correlated, and including geographical and demographic characteristics of New Zealand as well as peculiarities of insured buildings.

**Abstract** *Nell'assicurazione contro i disastri la valutazione della correlazione spaziale delle regioni altamente esposte ai terremoti è essenziale per diversificare il rischio in un portafoglio. Il rischio sismico è estremamente rilevante in Nuova Zelanda a causa della sua vicinanza ad una faglia sismica e la diversificazione dei premi in base al rischio di terremoto è fondamentale data l'elevata eterogeneità della densità di popolazione. In questo lavoro utilizziamo un modello multilivello a effetti misti per cogliere la correlazione spaziale delle perdite da terremoto riportate tra il 2000 e il 2018 in Nuova Zelanda assumendo che le circoscrizioni nelle stesse regioni sono correlate e includendo sia caratteristiche geografiche e demografiche della Nuova Zelanda che alcune peculiarità degli edifici assicurati.*

**Key words:** Earthquake losses, insurance, mixed effects, multilevel model, spatial correlation.

---

F. Marta L. Di Lascio

Faculty of Economics and Management, Free University of Bozen-Bolzano, e-mail: marta.dilascio@unibz.it

Selene Perazzini

Faculty of Economics and Management, Free University of Bozen-Bolzano, e-mail: selene.perazzini@unibz.it

## 1 Introduction

Accounting for the effect of spatial correlation between insured assets has become a prominent issue in disaster insurance, as insurers appear extremely fragile to natural hazards [3]. In particular, disaster insurers are affected by spatial correlation in their risk-portfolios, which collect the risks of several nearby-located immovable assets, creating the potential for extreme losses. As insurance is a risk-sharing mechanism that distributes risk among policyholders proportionally to their exposure, an optimal characterization of spatial correlation would allow insurers to construct risk-based premiums that reflect the specific degree of correlation of the policyholder. In turn, this might help insurers to strengthen their business.

In this paper we propose a two-level mixed effects model describing earthquake-insured building losses in New Zealand. To our knowledge, this is the first time that a multilevel model is applied to insurance losses with the aim to account for the effect of spatial correlation. This approach allows us to capture the correlation between neighbouring areas and therefore to identify those that might put a strain on the business. New Zealand is an interesting case study due to both its proximity to a seismic fault and the heterogeneity of population density, which is very high in three large cities - Auckland, Wellington, and Christchurch - while wide areas are uninhabited. A great effort has been made in identifying the size of the area to observe that should be sufficiently large to detect micro-correlations [1], but small enough to capture the diverse local risk. To this end we considered insured losses at the ward level, i.e. electoral districts, and a second level of aggregation defined by the 16 regions of New Zealand, which occupy much larger territories. In addition to spatial correlation, a series of geographical and demographic characteristics and peculiarities of dwellings have been included in the model. We found that the number of usual resident, the real estate value, and the risk index CRESTA zone are good predictors of wards' earthquake losses. Most of all, correlation between wards has been detected and a clear spatial relationship emerged.

The rest of the paper presents the data set in Section 2, the two-level mixed effects model in Section 3, the empirical findings in Section 4, and conclusion in Section 5.

## 2 Data

Data on losses have been provided by the New Zealand Earthquake Commission (EQC) and refer to the earthquake insurance coverage EQCover. The database collects information about both buildings insured and claims between 2000 and 2018. Given the extraordinary high insurance penetration rate in New Zealand, the dwellings insured approximately correspond to the overall housing estate of the country. Insured properties are localized by longitude and latitude, and have been assigned to the respective wards and regions through reverse geocoding. We refer to the New Zealand 2019 local boundaries map released by Land Information New

Zealand. In case of missing coordinates, records have been referenced by means of postcodes.

As far as claims are concerned, we limited our analysis to open or accepted claims only. A great part of the claims trace back to the 2010-11 Canterbury earthquake sequence. Since extreme events may consistently affect results, the Canterbury region has been excluded from the analysis. Both the properties and the cost of claims have been aggregated at the ward level. In order to overcome issues generated by the time gap between the moment at which the damage occurred and the opening of the claims as well as the effect of earthquakes' sequences, we considered the sum of insurer's losses due to claims reported between 2000 and 2018 in the ward.

The median CRESTA zone of the ward ( $X_1$ ) and the mean value of dwellings in the ward ( $X_2$ ) have also been included in the analysis. Moreover, the number of square kilometers per dwelling ( $X_3$ ) has been computed per ward by combining the information in the New Zealand boundaries map and the number of dwellings insured. Additional information have been taken from Stats NZ, the national statistical institute. In particular, the rate of occupancy of dwellings ( $X_4$ ) and the number of individual usually resident ( $X_5$ ) refer to territorial authorities, while the rate of housing with reported problems (heating, mold, etc..) ( $Z_{1j}$ ) and the average weekly income ( $Z_{2j}$ ) refer to regions. A random effect ( $W$ ) representing the rate of earthquakes in the territorial authority has also been included in the model.  $W$  has been computed considering all the earthquakes of magnitude at least 3.5 happened in New Zealand from January 1900 to May 2020 and reported in the GeoNet earthquake catalogue.

### 3 The Model

The ward's losses per building ( $Y$ ) are represented by a two-level variance component model with  $n$  level 1 units, i.e. the wards, and  $m$  level 2 units, i.e. the regions:

$$\log(y_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 \log(X_{2ij}) + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 \log(X_{5ij}) + \beta_6 Z_{1j} + \beta_7 \log(Z_{2j}) + u_{1j} + u_{2j} W_{ij} + e_{ij} \quad (1)$$

where  $i = 1, \dots, n$  and  $j = 1, \dots, m$ . The model includes independent variables aimed at capturing the variability at both the ward- and region-level, denoted respectively by  $X$  and  $Z$  and described in Sect. 2. Since the distribution of losses per wards is highly skewed, it has been log-transformed. This transformation has been also applied to  $X_2, X_5$  and  $Z_2$  whose range of values is considerably high. Coefficients  $\beta_k$  with  $k = 0, \dots, 7$  are the fixed effects of the model, while  $u_{1j}, u_{2j}$  and  $e_{ij}$  are the random effects. We assume

$$u_j \sim N(0, \Psi_j) \quad \text{with} \quad \Psi_j = \Psi = \begin{pmatrix} \sigma_{u_1}^2 & 0 \\ 0 & \sigma_{u_2}^2 \end{pmatrix} \tag{2}$$

$$e_{ij} \sim N(0, \sigma_e^2), \quad \text{cov}(u_{1j}, e_{ij}) = 0, \quad \text{cov}(u_{2j}, e_{ij}) = 0$$

for all  $i$  and  $j$ . Therefore, the variance of  $y_{ij}$  is

$$\begin{aligned} \text{var}(y_{ij} | \beta_0, \dots, \beta_7, X_{1ij}, \dots, X_{5ij}, Z_{1j}, Z_{2j}, W_{ij}) &= \text{var}(u_{1j} + u_{2j} + e_{ij}) = \\ &= \sigma_{u_1}^2 + \sigma_{u_2}^2 + \sigma_e^2 \end{aligned} \tag{3}$$

The assumptions in Eq. (2) imply homoscedastic variance and an equicorrelated block covariance matrix that exhibits serial correlation between the wards in the same region

$$\text{cov}(u_{1j} + u_{2j} + e_{ij}, u_{1j'} + u_{2j'} + e_{i'j'}) = \begin{cases} \sigma_{u_1}^2 + \sigma_{u_2}^2 + \sigma_e^2, & \text{if } i = i', j = j' \\ \sigma_{u_1}^2 + \sigma_{u_2}^2 + \sigma_{ii'}^{(j)}, & \text{if } i \neq i', j = j' \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Thus, the variance-covariance matrix of  $y_{ij}$  is a block diagonal matrix where each block in the main diagonal is the variance-covariance matrix of a region  $j$  with  $n_j$  wards. Finally, we model the intra-class correlation  $\rho_{ij,i'j}$ , i.e. the correlation between wards within a region, through a function  $h(\cdot)$  of the distance  $d_{ii'}$  between the centroids of two wards  $i$  and  $i'$ , i.e. the averages of the geographical coordinates (longitude and latitude) of all the points located in them, and a parameter  $r$  given by the distance where the variogram first flattens out and reaches the sill [2]

$$\rho_{ij,i'j} = \frac{(\sigma_{u_1}^2 + \sigma_{u_2}^2 + \sigma_{ii'}^{(j)})}{(\sigma_{u_1}^2 + \sigma_{u_2}^2 + \sigma_e^2)} = h(d_{ii'}, r). \tag{5}$$

Since the spatial correlation between two wards is stronger the closer they are and becomes equal to 0 after a certain distance, we assume that wards in the same regions are correlated, while regions are uncorrelated. The between-class correlation is therefore  $\rho_{ij,i'j} = 0$ . As cities are far from regional borders, the correlation between boundary wards in different regions is negligible.

Eq.s (1) and (4) require the estimation of eight fixed coefficients ( $\beta_0, \dots, \beta_7$ ) and three random coefficients ( $\sigma_{u_1}^2, \sigma_{u_2}^2, \sigma_e^2$ ). The model has been fitted by using the restricted maximum likelihood method [4].

#### 4 Empirical analysis

First of all, the model in Eq. (1) with  $\rho_{ij,i'j} = 0$  is estimated in order to investigate the spatial correlation hypothesis. Several functional hypotheses have been tested



to identify the intra-class correlation  $h(d_{iit'}, r)$ . According to the Akaike and the Bayesian information criteria, the best fit is obtained when assuming the following Gaussian correlation structure

$$h(d_{iit'}, r) = (1 - nugg)e^{-\left(\frac{d_{iit'}}{r}\right)^2} \tag{6}$$

where a nugget effect *nugg* [2, 5] is introduced in order to account for abrupt changes at very small distances. Secondly, the model in Eq. (1) has then been estimated including spatial within-region correlation as defined in Eq.s (2-4), and results are reported in Table 1. A few fixed effects appear significant in explaining  $\log(Y)$ : the intercept, the CRESTA zone, the logarithm of the dwellings' value, and the logarithm of the usual residents. Not surprisingly, the average value of dwellings in the ward appears the main determinant of the value of losses. On contrast, the number of usual resident, which might be interpreted as a proxy of the number of buildings in the ward, and the CRESTA zone have much lower effects on  $\log(Y)$ . Overall, the analysis did not reveal any relevant deviation from the assumptions on random effects and residuals in Eq. (2). The standardized residuals (Fig. 1, left) are small, suggesting that the estimated model was successful in explaining the insurer's losses. In addition, a weak evidence of heteroscedasticity emerged, suggesting that there could be possible drivers of different variances in the residuals, such as the inhabited density or other geospatial characteristics of the phenomena. This evidence is further supported by the close agreement between the observed losses and the within-group fitted values (Fig. 1, right). In particular, the figure shows that only a few extreme observations deviate from the fitted values, most of which refer to the Waikato region, and the extent of these misalignment is rather limited.

**Table 1** Estimation results: model in Eq. (1) with within-group correlation  $\rho_{ij, i'j}$  as in Eq. (6).

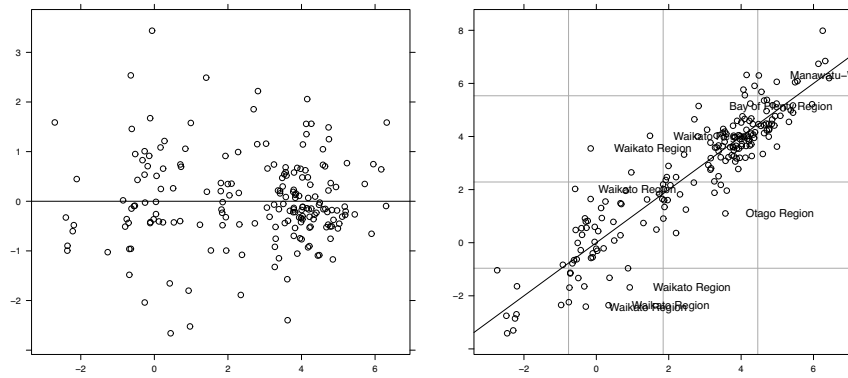
	Intercept	$X_1$	$\log(X_2)$	$X_3$	$X_4$	$\log(X_5)$	$Z_1$	$\log(Z_2)$
$\log(Y)$	-77.696***	0.357***	30.139***	-0.145	-0.896	0.229***	2.100	-0.124
	(14.158)	(0.083)	(5.596)	(0.120)	(1.245)	(0.074)	(11.776)	(4.733)

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Random Effects		Within Correlation	
$\sigma_{u_1}$	1.842	$r$	0.207
$\sigma_{u_2}$	12.389	<i>nugg</i>	0.269
$\sigma_e$	1.058		

## 5 Conclusion

A multilevel mixed effects model for spatially correlated earthquake losses in New Zealand has been presented. We found evidence of the presence of correlations be-



**Fig. 1** Results of two-level mixed effects model in Eq. (1) with within-group correlation  $\rho_{i,j,i'j'}$  as in Eq. (6). Left: Fitted residuals (x-axis) versus standardized residuals (y-axis). Right: Observed losses (x-axis) versus within-group fitted values (y-axis).

tween the losses of the wards, that has been explained by a spatial relationship. Moreover, we found that the number of usual resident, the real estate value, and the risk index CRESTA zone are good predictors of wards' earthquake losses.

**Acknowledgements** Authors acknowledge the New Zealand Earthquake Commission for the provision of the data.

## References

1. Cooke, R.M., Kousky C., and Joe, H.: Micro Correlations and Tail Dependence. In: Kurowicka and Joe (eds.) Dependence Modeling: Handbook on Vine Copulae. World Scientific, Singapore (2010)
2. Cressie, N.A.C.: Statistics for Spatial Data. Wiley, New York (1993)
3. Kunreuther, H. C. and Michel-Kerjan, E.: Climate Change, Insurability of Large-scale Disasters and the Emerging Liability Challenge. NBER Working Papers 12821, National Bureau of Economic Research, Inc (2007).
4. Laird, N.M. and Ware, J.H.: Random-effects Models for Longitudinal Data, *Biometrics* **38**, 963-974 (1982)
5. Pinheiro, J.C. and Bates, D.M.: Mixed-Effects Models in S and S-PLUS. Springer (2000)

# Weighted distances for spatially dependent functional data

## *Misure di distanza per dati funzionali spazialmente dipendenti.*

Andrea Diana, Elvira Romano, Claire Miller and Ruth O'Donnell

**Abstract** In this work we propose optimally weighted  $L^2$  distances for spatially dependent functional data. Two different spatial structures have been considered: a classical georeferenced spatial structure and a connected network one. In these two situations, assuming a penalized basis representation for the functional data, we consider weight functions depending on the spatial location. Real meteorological data have been analyzed in order to show performances of the proposed distances.

**Abstract** *In questo lavoro si propone una distanza per dati funzionali spazialmente dipendenti. A partire da due differenti strutture spaziali, griglia o reticolo, assumendo i dati funzionali rappresentati tramite funzioni di base ottenute da criteri di penalizzazione ottimizzata, definiamo una distanza dipendente dalla componente spaziale. Le caratteristiche della distanza proposta vengono illustrate attraverso l'applicazione della stessa su dati meteorologici.*

**Key words:** Functional data, Spatial dependence, Distance

---

Andrea Diana

Department of Mathematics and Physics, Università della Campania "Luigi Vanvitelli", Caserta, Italy, e-mail: andrea.diana@unicampania.it

Elvira Romano

Department of Mathematics and Physics, Università della Campania "Luigi Vanvitelli", Caserta, Italy, e-mail: elvira.romano@unicampania.it

Claire Miller

School of Mathematics and Statistics, University of Glasgow, Glasgow, UK, e-mail: Claire.Miller@glasgow.ac.uk

Ruth O'Donnell

School of Mathematics and Statistics, University of Glasgow, Glasgow, UK, e-mail: Ruth.Haggarty@glasgow.ac.uk

## 1 Introduction

In this work, we introduce an optimally weighted distance for spatially dependent functional data. Spatially dependent functional data come in many forms. Roughly speaking, such curves, spatially located, refer either to curves observed on points, lines or areal spatial units. The definition of the covariance structure among the functions depends on the spatial structure we observe. For instance, the trace-variogram function [6] enables estimation of the interactions among functions observed on a regular grid in terms of variability; whereas the spatial covariance of Haggarty et al. [7] quantifies the interactions between functions on a connected network. The distance we introduce is a generalization of the distance proposed by [4] for spatially dependent functional data, by considering a regular grid and a connected network. In particular we will focus on distances measuring explicitly differences in terms of spatial dependence such as the distances proposed by [1], [6], [7]. We will do so, first introducing the nature of functional data of interest here, geostatistical functional data and spatial variability measures in general in section 2, then describing a variety of distances and their applications in section 3. We introduce our proposed metric in section 4 and apply to real data monitoring the evapotranspiration problem in the Italian peninsula in section 5.

## 2 Geostatistical functional data and spatial variability measures

Let  $(\chi_{s_1}(t), \dots, \chi_{s_i}(t), \dots, \chi_{s_n}(t))$  be a set  $n$  of geostatistical functional data. The  $n$  points  $(s_1, \dots, s_i, \dots, s_n)$  in  $D \subseteq R^d$  identify the  $n$  locations where the random functions  $\chi_s(t)$  are located. Each function is defined on  $T = [a, b] \subseteq R$  and is assumed to belong to a Hilbert space with the inner product  $\langle \chi_{s_i}, \chi_{s_j} \rangle = \int_T \chi_{s_i}(t) \chi_{s_j}(t) dt$  [9]. For a fixed site  $s_i$ , it is assumed that the observed functions can be expressed according to the model:  $\chi_{s_i}(t) = \mu_{s_i}(t) + \varepsilon_{s_i}(t)$ ,  $i = 1, \dots, n$  where  $\varepsilon_{s_i}(t)$  are zero-mean residuals and  $\mu_{s_i}(t)$  is the mean function.

For each  $t, t \in T$ , the random process is assumed to be second order stationary and isotropic: that is, the mean and variance functions are constant and the covariance depends only on the distance between sampling sites.

It is assumed that the mean function is constant over  $D$  and that the semivariogram function  $\gamma(h, t) = \gamma_{s_i s_j}(t) = \frac{1}{2} V(\chi_{s_i}(t) - \chi_{s_j}(t))$ , according to [2], can be expressed by:

$$\gamma(h, t) = \gamma_{s_i s_j}(t) = \frac{1}{2} V(\chi_{s_i}(t) - \chi_{s_j}(t)) = \frac{1}{2} E [\chi_{s_i}(t) - \chi_{s_j}(t)]^2. \quad (1)$$

Consistently with [5], the estimation of the trace-variogram involves the computation of integrals that can be simplified by considering that the functions are expanded in terms of basis functions. It is the continuous version of the variogram for spatio-temporal data, which provides a helpful framework to do spatial prediction and, with particular relevance for this paper, to provide a mechanism to incorpo-

rate spatial weights in the computation of distance metrics for spatially dependent curves. The trace-variogram does not provide a measure of the covariance between functions. The latter is given by the spatial covariance function defined by [7], which provides a measure to describe the relative variability between functions. The functional covariance is mainly the product of the difference of two areas computed in relation to a reference curve. An area  $\bar{A}$  below the mean curve and an area  $A_i$  below a generic estimated curve  $\chi_i$  with respect to a reference curve  $\chi_l$ . Given a curve of reference  $\chi_l$  corresponding to the horizontal line which is below the minimum value of the set of the curves. The area between the reference line and the mean curve is defined as:

$$\text{Area}(\bar{\chi}_i(r), \chi_l) = \int \{\bar{\chi}_i(r) - \chi_l\}^2 dr = \bar{A}. \quad (2)$$

In the same way, the area between curve  $\hat{\chi}_i(r)$  and reference line  $\chi_l$  is

$$\text{Area}(\hat{\chi}_i(r), \chi_l) = \int \{\hat{\chi}_i(r) - \chi_l\}^2 dr = A_i. \quad (3)$$

Thus a quantification of the difference between a generic function  $\hat{\chi}_i(r)$  and a median curve can be expressed by the difference in terms of magnitude between  $\hat{\chi}_i(r)$  and the mean curve  $\bar{\chi}_i(r)$ , which is the difference of their area as  $A_i - \bar{A}$ . As stated in [7], the area between the mean curve and a reference line can be used both to reflect the direction of the difference between a given location and the overall mean. In addition it can be used to standardize the areas so that the measures of covariance are on a most suitable scale. The estimated functional covariance between the two estimated functions can thus be defined as:

$$\widehat{\text{Cov}}(\hat{\chi}_i(r), \hat{\chi}_j(r)) := \frac{(A_i - \bar{A})(A_j - \bar{A})}{\bar{A}^2}. \quad (4)$$

It is a single covariance value between two functions over a space of interest that can then be used to create an adjusted covariogram cloud.

### 3 Distances for spatially dependent functional data

The most simple distance used between two spatially dependent functional data objects has been proposed in the pioneering work [2]. It is a weighted dissimilarity metric among the geo-referenced curves expressed by

$$d_g(\chi_{s_i}(t), \chi_{s_j}(t)) = d(\chi_{s_i}(t), \chi_{s_j}(t)) \gamma_{s_i s_j}(h) \quad (5)$$

where  $d(\chi_{s_i}(t), \chi_{s_j}(t)) = \sqrt{\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt}$  is the distance between the curves without considering the spatial component, and  $\gamma_{s_i s_j}(h)$  corresponds to the trace-variogram function calculated for the distance between sites  $s_i$  and  $s_j$ . Once the

trace-variogram has been estimated, a parametric model is fitted following classical geostatistical estimation procedures [6].

This distance does not consider the spatial covariance among the functional data, whereas the proposal of [7] is a correlation based distance which groups functions together regardless of the amplitude of their functional variation. It is defined as:

$$d_{ij}^c = d_{i,j} Cov(\chi_{s_i}(t), \chi_{s_j}(t)). \tag{6}$$

where the covariance function is defined as in Eq.(4). It provides differences in terms of relative magnitude, and summarizes in a single value the correlation between two functions over the spatial domain of interest.

#### 4 Optimally weighted distances for spatially dependent functional data

As we have shown, measures of distance between spatially dependent functional data can be distinguished according to the nature of space on which these are defined. Our main aim is to introduce a distance for spatially dependent functional data considering (i) the simple (georeferenced) and then (ii) the more complex spatial domains (e.g. a connected directed network). Using the idea of [4], we define an optimally weighted distance for functional data spatially dependent. Assuming a basis function representation for functional data we propose to consider weight functions including both the spatial and functional component. It is a generalization of [4] to the spatial functional framework for two different spatial domains: the georeferenced and the directed network. As in [4] we define a weighted  $L^2$  distance as follows:

$$d_{\omega_s}(\chi_{s_i}(t), \chi_{s_j}(t)) = \sqrt{\int_T \omega_s(t) (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt} \tag{7}$$

where the weight  $\omega_s$  satisfies  $\omega_s \geq 0$  and  $\int \omega_s dt = 1$ .

The problem is choosing a weight function  $\omega_s(t)$ , such that the seminorm is defined by  $\|(\cdot)\|_{\omega_s} = \sqrt{\int \omega_s \theta(t)^2 dt}$ . We define a spatio-functional smooth function  $\omega_s(t) = [\mathbf{b}_{\omega_s}^T(t)\mathbf{q}]^2$  where  $\mathbf{b}_{\omega_s}(t)$  is a vector of associated basis functions and  $\mathbf{q}$  is the vector of coefficients.

The spatio-functional smooth function is obtained by the following minimization problem:

$$\omega_s(t) = \underset{\|\omega_s\|=1}{\operatorname{argmin}} \frac{\sum_{1 \leq i < j \leq n} V(\|\theta_{i,j}\|_{\omega_s}^2)}{\sum_{1 \leq i < j \leq n} [E(\|\theta_{i,j}\|_{\omega_s}^2)]^2}; \tag{8}$$

with  $\theta_{i,j}(t) = a_{i,j}x_i(t) - a_{j,i}x_j(t)$ , where  $a_{i,j}$  and  $a_{j,i}$  are obtained starting from the structure of the spatial domain of interest.

The coefficient  $a_{j,i}$  is the element reflecting the spatial dependence among functional data and changes according to the spatial grid on which the functional data

are observed. When  $a_{i,j} = a_{j,i} = 1$ , we have a weighted distance  $d_{\omega}$  defined by [4] for functional data without spatial dependence.

In the case of spatially dependent functional data observed on a regular grid, we introduce a weight function depending on the spatial variability expressed by a trace-variogram function. Formally we define:  $a_{i,j} = a_{j,i} = \hat{\gamma}(h_{i,j})$  where  $\hat{\gamma}(h)$  is the estimated trace-variogram. The introduced distance could be viewed in broad terms as a generalization of the dissimilarity measure defined in (5) with the advantage that the distance is optimally calibrated from the functional and spatial point of view.

In the case of spatially dependent functional data observed on a directed network we introduce a weight as a covariance function depending on a structured oriented graph. In particular denote by  $\widehat{Cov}$  the matrix of estimated spatial covariance between the knots of a net (as in Eq.4),  $\Gamma = \text{diag}(\widehat{Cov}_j)$  the diagonal covariance matrix and  $D$  the matrix of the  $L_2$  functional distance  $d(\chi_{s_i}(t), \chi_{s_j}(t)) = \sqrt{\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt}$ , we define:  $a_{i,j} = L_i \cdot \Gamma \cdot D_j^T$ ; where  $L_i$  is a row vector of matrix of contiguity  $L$ . The above measure can be seen as a generalization of the distance introduced by [7] to a directed network by considering the complex spatial interrelationship between curves. According to this distinction we can rewrite our distance (Eq. 7) as

$$d_{\omega_s}(\chi_{s_i}(t), \chi_{s_j}(t)) = \begin{cases} d_{\omega_\gamma}(\chi_{s_i}(t), \chi_{s_j}(t)) & s = \gamma \\ d_{\omega_C}(\chi_{s_i}(t), \chi_{s_j}(t)) & s = C \end{cases} \quad (9)$$

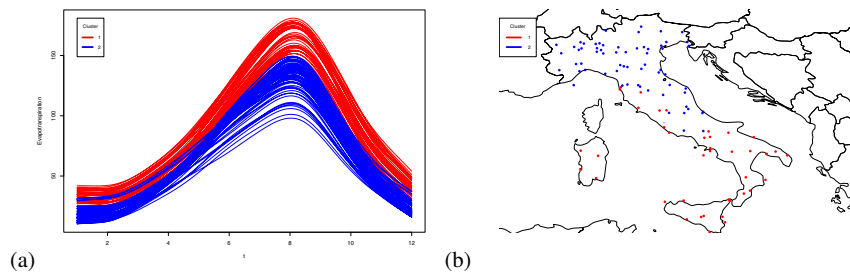
Where  $s = \gamma$  and  $s = C$  correspond to weight functions for spatially dependent functional data observed respectively on a regular spatial grid and on a directed network.

## 5 Real data Analysis: a meteorological study on evapotranspiration in Italy

In this section we show an application of the distance introduced in section 4 to a non-regular grid. We focus on a hierarchical classification of the meteorological time series of evapotranspiration for 12 months, from December 2016 to November 2017, in 103 provinces of Italy. One aim of our analysis was to obtain groups of stations which are similar in terms of evapotranspiration of the determinand of interest.

In Figure 1 it is possible to observe the evapotranspiration curves for 103 Italian provinces and the results of a hierarchical classification using the distance  $d_{\omega_\gamma}$ .

We identify the number of clusters by considering three indices among many proposed in the literature, that is the Average silhouette width [10], Calinski and Harabasz index [3] and Dunn2 index [8]. The two groups of curves respectively represented in red and blue reflect the geographic conformation of the provinces (Fig-



**Fig. 1**

ures 1(a), 1(b)); the blue curves are associated with the northern Italian provinces while the red curves are associated with the Southern ones. Looking at the two families, we can say that the first cluster, from a functional point of view, is characterized by high values of water evapotranspiration and, from a geographical point of view, covers all the provinces of central-southern Italy; the second cluster, from the functional point of view, is characterized by lower values of water evapotranspiration and, geographically, covers all the provinces of central-northern Italy. It could be expected that the clusters have this configuration, and are coherent by considering the spatial correlation. The reasons for which this could be expected is that locations that are close to the north are more similar in term of temperature.

## References

1. Balzanella A., Romano E., Verde R.: Modified half-region depth for spatially dependent functional data, *Stochastic Environmental Research and Risk Assessment*, 31: 87-103, (2017)
2. Caballero, W., Giraldo, R., Mateu, J.: A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment*. Volume 27, Issue 7, pp. 1553-1563, (2013)
3. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis, *Communications in Statistics*, 3, no. 1:1-27, (1974)
4. Chen H., Reiss P.T., Tarpey T.: Optimally Weighted L2 Distance for Functional Data. *Biometrics*, 70(3): 516-525, (2014).
5. Delicado, P., Giraldo, R., Comas, C. and Mateu, J.: Statistics for spatial functional data: some recent contributions. *Environmetric*, 21: pp.224-239, (2010)
6. Giraldo, R., Delicado, P., Comas, C., Mateu, J.: Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, (2011)
7. Haggarty, R., Miller, C., Scott, E.M.: Spatially Weighted Functional Clustering of River Network Data. *Journal of the Royal Statistical Society, Series C.*,(2015)
8. Halkidi, M., Batistakis, Y., Vazirgiannis M.: On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, 17:2/3, 107-145 (2001)
9. Ramsay, J.E., Silverman, B.W.: *Functional Data Analysis*, (Second ed.) Springer (2005)
10. Rousseeuw, P. J., Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, *Computational and Applied Mathematics*. 20: 53-65 (1987)



# Spatial modeling of childcare services in Lombardia

## *Modellazione spaziale dei servizi per l'infanzia in Lombardia*

Emanuele Aliverti, Stefano Campostrini, Federico Caldura and Lucia Zanotto

**Abstract** We are interested in mapping the level of childcare services in Lombardia. As a first step, we focus on modeling the flows across municipalities, measuring the number of children that moves from a municipality to another one for using childcare services. Then, we model the coverage rate as the ratio between the number of childcare services and the number of children in a given municipality using a zero-inflated spatial Poisson regression model, providing a model-based map of the level of services in the region. Results allow to approximate the level of services in Lombardia, providing preliminary insights on how resources should be addressed to improve such an aspect.

**Abstract** *In questo articolo ci concentriamo sulla modellazione spaziale del livello di servizi per l'infanzia in Lombardia. In una fase preliminare, vengono modellati i flussi tra i diversi comuni, in modo da misurare il numero di bambini che si spostano da un comune ad un altro per utilizzare i servizi per l'infanzia. Successivamente, il tasso di copertura (inteso come il rapporto tra il numero di posti ed il numero di bambini) viene modellato tramite un modello di regressione spaziale con risposta Poisson sovra-dispersa, per fornire una stima del livello di copertura all'interno della regione. I risultati di questo approccio permettono di fare chiarezza sulla diffusione di questo fenomeno.*

**Key words:** Bayesian modeling; Childcare services; Spatial model; Poisson regression.

---

Emanuele Aliverti  
Università Ca' Foscari di Venezia, e-mail: emanuele.aliverti@unive.it

Federico Caldura  
Università Ca' Foscari di Venezia, e-mail: federico.caldura@unive.it

Stefano Campostrini  
Università Ca' Foscari di Venezia, e-mail: stefano.campostrini@unive.it

Lucia Zanotto  
Università Ca' Foscari di Venezia, e-mail: lucia.zanotto@unive.it

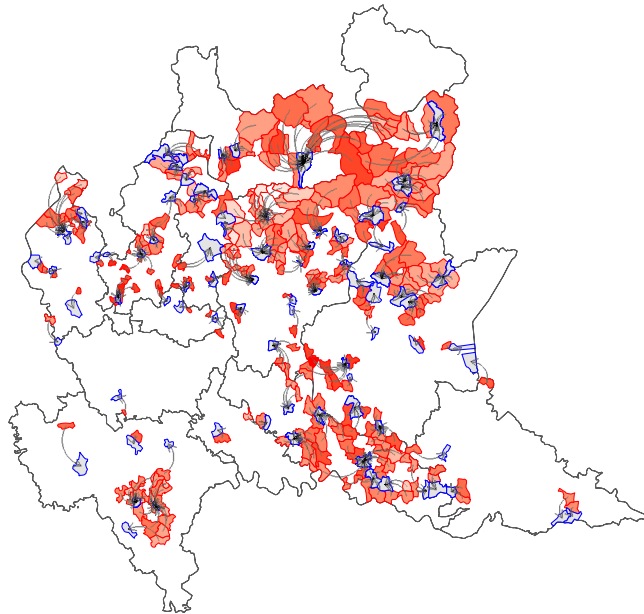
## 1 Introduction

Childcare services were firstly introduced in Italy during 1971 by Law 1044/1971 “servizi sociali di interesse pubblico”. Their main aim was assisting parents — in particular women — during childcare, in order to facilitate their participation in the labor market and promote gender equality. Over the years, their role in infant education has been highlighted, since they contribute to cognitive, emotional and social development; in addition, childcare services can also reduce socio-economic inequalities, generating equal education opportunities for men and women. Even if their benefits are generally recognized and supported by national and local fundings, there are huge differences among areas: in northern regions, childcare network services are more developed, while southern ones still have some difficulties to implement them.

A useful index that can help to understand this phenomena is the level of coverage, which can be obtained as the ratio between the number of available childcare services and the number of children between 0-3 years old. In 2010, the European Union in the Barcelona European Council has fixed this parameter at 33% for all the European countries. In Italy, this goal in some municipalities is far exceeded, while in many others it is much lower and the differences are relevant also within regions [3]. One aspect regarding the estimation of the amount of coverage at a local level is that the raw division between available places and children in a single municipality generally underestimates the quantity of interest; for examples, in areas without childcare services, parents can decide to move to close kindergartens and municipalities (particularly those small in population) could decide to support the services of nearby areas, instead of opening a new one. In this work, we try to offer a better estimate of this quantity focusing on Lombardy region in 2018 and taking into account the possible flows between municipalities, relying on the survey “Asili nido e servizi integrativi per la prima infanzia” carried out by Istat. The dataset collects information about all kindergartens (public and private), including the spending of each municipality for childcare services.

## 2 Data pre-processing

We focus on data from 2018, measured at the municipal level and covering information on the number of children, the number of active childcare services and the overall municipal expenses for childcare services. Potentially, these data allow to measure the coverage rate by taking the ratio between the number of childcare places and the number of children in a given municipality. However, it is well known that several families bring their children to a different municipality, due to the lack of available places in the municipality of residence. We illustrate a simple procedure to estimate, at least partially, this phenomena. We define a municipality as an *out-taker* if it satisfies the following conditions:



**Fig. 1** Estimated childcare services flows.

- absence of childcare services within its boundary;
- it reports a positive number of children, and
- it spends more than 50 euro for each child.

These criteria provide a preliminary rule of thumb for detecting municipality from where there might be outgoing flows. We are aware that we are underestimating an important part of the phenomena, since there might be a flow without the financial contribution from the municipality. However, we found that these criteria are sufficiently simple to exclude a large number of non-interesting municipality, and eventually allow to introduce more refined criteria within a second step, described below.

As a further step, we need to determine where these flows might be directed to. We define as a potential “in-taker” a municipality that registers a ratio between the number of childcare places and the number of children above the 0.8 provincial quantile. Note that we define these municipalities as “potential” since, in order to be included in the analysis, they need to be matched with at least one *out-taker*. Therefore, it is reasonable to take a fairly conservative proportion of municipalities at this step, since many potential are likely to be discarded.

As a final step of this process, we match each *out-taker* with the closest potential *in-taker* in terms of temporal distance. These quantity has been measured in min-

utes from reaching the two municipalities, using the distance matrices provided by ISTAT.

Figure 1 shows the estimated movements for using childcare services. These are particularly evident in the north-east of the region, where there are some poles which cover the demand of neighboring territories. Also in the south-east of the map the movements are quite significant even if of minor intensity.

As a result, we obtain for each municipality  $i = 1, \dots, n$  the number of active childcare places  $y_i$  and the number  $e_i$  of children referring to such municipality, adjusted through the procedure just outlined. This approach has some drawback, since it ignores the fact that often parents bring children to childcare while commuting to work, non necessary to the closest municipality. This issue could be mitigated including information on commuting into the analysis. Unfortunately, most recent data on this aspect refer to 2011, and in the last 10 years several municipalities have been merged, and the socio-economic landscape has definitely changed. As an alternative, the drawback of our procedure can be restricted considering the heterogeneity of labor market areas. In the next section, we follow this approach and use this information within a spatial model for the level of coverage.

### 3 Spatial modeling

We model the number of childcare places  $y_i$  through a Bayesian Zero-Inflated Poisson (ZIP) spatial model, introducing number of children  $e_i$  as an offset and considering the effect of the labor market area of each municipality. Specifically, we let

$$y_i \sim \text{ZIP}(\lambda_i, \pi_0)$$

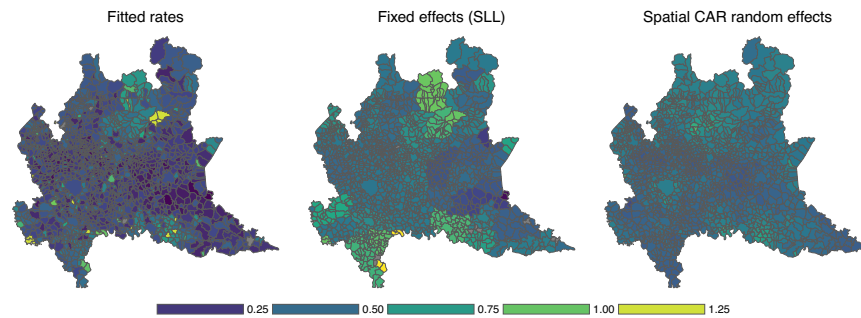
$$\log(\lambda_i) = \alpha + u_i + v_i + x_i^T \beta + \log(e_i), \quad (1)$$

where  $\lambda_i$  denotes the Poisson mean parameter,  $\alpha$  and intercept term and  $u_i$  and  $v_i$  denote municipality-specific spatial and exchangeable random effects, respectively, while  $\log(e_i)$  introduces the number of children as an offset; see [1] for a practical application of this model in epidemiology. In addition, we account for the heterogeneity of labor market areas (SLL) including a set of fixed effects  $\beta = (\beta_1, \dots, \beta_p)$ , where  $x_i$  denotes a  $p$ -dimensional indicator vector of the labor market area for the  $i$ -th municipality.

The probability mass function for the response is given by

$$p(y_i | \lambda_i, \pi_0) = \pi_0 I(y_i = 0) + (1 - \pi_0) \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i},$$

where the additional parameter  $\pi_0$  controls the amount of inflation. This specification allows to account for the large number of zero observations in our data, characterizing municipalities without childcare services.



**Fig. 2** Left plot: fitted rates of number of places over number of children. Middle and right plot: estimated fixed effects and smooth spatial components, transformed in the rate scales.

In order to account for the spatial dependence across observations, we follow a standard approach in modeling areal data and specify an intrinsic conditional autoregressive structure (iCAR) with precision  $\tau_u$  on the random effects  $\mathbf{u} = (u_i, \dots, u_n)$ ; see, for example, [1, sec 6.1] for details. According to such a specification, each location  $i$  is modeled as conditionally independent from the others, given its neighbors (corresponding, in our settings, to the municipalities with whom  $i$  shares a border); the random effects  $v_i$  are instead assumed from a common Gaussian with precision  $\tau_v$ . This random-effects specification allows to take into account the spatial structure of the data, borrowing information across municipalities. We further specify non-informative Gaussian priors on  $\alpha$  and non-informative log-Normal distributions on  $\tau_u$  and  $\tau_v$ .

We conduct approximate posterior inference through the R package `INLA`, which performs an Integrated Nested Laplace Approximation of the posterior distribution of the model's parameters [2, 4]. We obtain estimates — via posterior mean — equal to  $\hat{\pi}_0 = 0.327$ , suggesting a modest amount of zero inflation, and for  $\hat{\tau}_v = 7.58$  and  $\hat{\tau}_u = 2.911$ . Estimates for the random and fixed effects are reported in Figure 2, which also illustrates the fitted values for the expected rates  $\hat{\lambda}_i/e_i$ , as well as the fixed effects  $\exp(\beta)$  and the spatial random effects in the rates scale  $\exp(\hat{u}_i)$ . Results indicate an interestingly heterogeneous coverage in Lombardy: its level is generally greater than 33% for most municipalities, except for the area located in north of Milan, the territories in south-east Mantova and especially around Brescia (20% or less). It seems particularly good in the Sondrio area and between Cremona and Mantova, where the childcare services are shared.

The spatial approach takes into account the proximity, and it has the advantage of providing an easier interpretation of the phenomena, but, of course, the special

features of municipalities can be lost, such as the regional excellences and the major shortcomings. The estimation is very satisfactory for close homogeneous area, like Milan and surroundings, but probably less accurate where there are municipalities covering several neighboring territories, such as Sondrio areas, where the coverage levels are probably overestimated.

## 4 Discussion

In this article, we have focused on modeling the coverage of childcare services in Lombardia, using a simple spatial model. These analysis have some limitations, but, at the light of these first positive results, can be further improved. One important feature of the proposed work is taking into account parents movements for childcare services, which is essential to provide a more realistic view of the actual offer of services in a territory. Moreover, from a methodological perspective, it might be useful to consider a more elaborate spatial specification taking into account isolated peaks of coverage, characterizing hubs municipality. These aspects are currently under investigation.

## 5 Acknowledgments

This work has been possible and particularly supported by and in agreement between Dipartimento per le Politiche della Famiglia of the Presidenza del Consiglio dei Ministri, Istat and the Department of Economic at the University Ca' Foscari of Venice.

## References

- 1 Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.
- 2 Finn Lindgren and Håvard Rue. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63(19):1–25, 2015.
- 3 Dipartimento per le politiche della famiglia, Istat, Università Ca' Foscari, and MIPA. *Nidi e servizi educativi per la prima infanzia, stato dell'arte, criticità e sviluppi del sistema educativo integrato 0-6*. Presidenza del Consiglio dei Ministri, 2020.
- 4 Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392, 2009.

# On the use of a composite attractiveness index for the development of sustainable tourist routes

## *Sull'uso di un indicatore di attrattivit composto per lo sviluppo di circuiti turistici sostenibili*

Cappello C. and De Iaco S. and Maggio S. and Palma M.

**Abstract** Italy boasts the most heritage sites in the world. The cultural heritage is widespread throughout the Italian peninsula, indeed one out of three municipalities hosts at least one museum or an analogous structure. In this context, it is relevant to identify new city networks in order to promote their museums and their points of attraction. In the paper, geo-referenced socio-demographic data together with a new statistical indicator, obtained through the linear combination of visitors arrivals, number of tourist accommodations, indexes of cultural and environmental attractions are used to define various touristic sustainable routes in Lecce district. The final results will help policy makers in planning some possible actions aimed at both the cultural development of the territory and definition of strategies for tourism deseasonalisation.

**Abstract** *L'Italia è il Paese che vanta il più vasto patrimonio culturale al mondo. Esso è diffuso capillarmente sul territorio, infatti in un comune su tre vi è almeno un museo o una struttura simile. Risulta, pertanto, indispensabile individuare delle reti di comuni per promuovere i musei e gli istituti culturali che sorgono in tali comuni. In questo contributo si propone un utilizzo integrato di dati georeferenziati ed un indicatore statistico, ottenuto mediante la combinazione lineare di indici relativi agli arrivi turistici, ai posti letto e ai punti di attrazione ambientale e culturale, per identificare percorsi museali e culturali in Provincia di Lecce. Il risultato finale aiuterà i policy makers per lo sviluppo di una rete culturale locale, nonché per individuare strategie alternative volte alla destagionalizzazione del turismo locale.*

**Key words:** Cultural network, GIS, Principal component analysis

---

Universit del Salento  
Dip.to di Scienze dell'Economia, e-mail: cappello.claudia@unisalento.it

## 1 Introduction

The Italian cultural heritage includes almost 5 thousand state and non-state museums, archaeological areas and monuments or monumental complex, open to the public (4900 in 2018 [2]). The cultural properties are rife in the municipalities, since one out of three hosts at least one museum or an analogous structure. However, further characteristics that distinguish the national cultural heritage are the dimensional polarization and the concentration of visitor flows [6].

For this reason, it is interesting to identify cultural itineraries to enhance those museums and attraction points that are not able to exploit their full potential.

The basic hypothesis is to identify paths that from the most attractive municipalities, in terms of number of tourists, number of accommodations and cultural and environmental points of attraction, stop in less known municipalities, providing at the same time the opportunity to visit the various cultural and environmental point of attractions in each selected town. These above-mentioned data have been used to define a GIS (geographic information system) project. Therefore, in the paper the design and the development of a GIS database as well as customized maps, based on the properly defined statistical indicator on the level of attraction of the cities, are proposed for promoting the whole territory. Indeed, the created thematic maps can be used by policy makers or travel agencies to encourage alternative tourist itineraries [7], which will help to attract travelers to visit locations outside the most known and crowded areas in favor of undiscovered destinations. Moreover, the development of networks among municipalities with different levels of attractiveness will mitigate the pressure on local population and it will promote the museums and other cultural attraction points located in each municipality which belongs to the network.

## 2 Data and methods

The development of new networks which connect the most known and attractive municipalities with the less known cities in the Lecce district represents an attempt to propose tourist itineraries, which are alternative to the well-known places of interest in the Salento peninsula. In particular, the main goals are to reduce the pressure on the locations which are overrun by tourists and redirect the public in the cities which are less known. At the same time, it is relevant to identify, for each proposed city to be visited, a museum and cultural route for discovering the points of attraction of the town. To this aim, data from different sources, such as public institutes (i.e. the Italian National Institute of Statistics - ISTAT, Italian Ministry of the environment) and private institutions have been collected. The spatial information regarding the administrative features (i.e. the municipal boundaries and the delimitation of their urban centers), the municipal road network (i.e. the main and local roads, as well as the major thoroughfares) and the orthophoto of the investigated area have been included in a GIS project. The aforementioned spatial data have been combined with



Title Suppressed Due to Excessive Length

others variables, available at the municipal level in 2018, which have been classified in the following categories:

- demographic data (i.e. resident population);
- geographic data (i.e. length of coastline, surface covered by parks, protected areas and marine protected areas, land areas);
- tourism data (i.e. tourist arrivals and overnight stays, number of tourism accommodations);
- cultural sites (i.e. number of museums, theaters, historic houses, churches, farmhouses);
- other data (i.e. number of museum visitors, number of hotel and non-hotel accommodations).

It is important pointing out that all the above data, which come from public sources, provide geo-referenced information or, at least, geo-coded information at the municipal level. Moreover, it should be noted that only the museums database provided by the ISTAT [2] has been integrated with a more detailed list of the museums which were open to the public in 2018 in the Lecce district, since the initial database was incomplete. Overall, a list of 39 museums has been obtained.

The aforementioned information, recorded at municipal level, has been converted into a format supported by QGIS (an open source software) with the corresponding shapefile, in which the related geographic information is stored.

In order to properly use all the collected data in a GIS project, it was necessary to define a relational model between the databases. This model is useful to manage the relationship between the different, independent and unrelated databases.

## ***2.1 A composite indicator of tourism attraction***

In order to identify new networks which connect the most known and attractive municipalities with the less known cities, a composite indicator of tourism attraction has been developed on the basis of the following variables measured at 97 municipalities of Lecce district

1. index of domestic tourist arrivals (DomTA),
2. index of international tourist arrivals (IntTA),
3. index of cultural points of attraction (CPA),
4. index of environmental points of attraction (EPA),
5. index of hotel and non-hotel accommodations (Beds),

where the CPA variable has been obtained as the sum of the number of museums, theaters, historical houses, churches and buildings, whereas the EPA variable corresponds to the average of the percentage of kilometers of coastline and the percentage of hectares covered by parks, protected areas and marine protected areas.

In particular, by applying principal component analysis (PCA) few uncorrelated linear combinations of original variables, that explain most of the total variance in the

data, can be obtained.

Let  $\mathbf{Z}$  be the  $(n \times p)$  data matrix, where  $n = 97$  represents the number of the location points and  $p = 5$  the analyzed variables above mentioned, denoted respectively  $Z_1, Z_2, \dots, Z_5$ . The PCA [3, 5] linearly transforms the variables into uncorrelated principal components as follows:

$$\mathbf{X} = \mathbf{Z} \mathbf{Q},$$

where  $\mathbf{X}$  is  $(n \times p)$  matrix of principal components  $X_i, i = 1, 2, \dots, 5$ , while  $\mathbf{Q}$  is a  $(p \times p)$  matrix, whose elements are the eigenvectors of the correlation matrix corresponding to the analyzed variables.

As is well known, the eigenvalues of the correlation matrix of the variables under study, indicate the proportion of total variance explained by each principal component  $X_i, i = 1, 2, \dots, 5$ . Hence, by considering the eigenvalues from the largest to the lowest, it is possible to identify those uncorrelated components which explain most of the total variance characterizing the observed data.

In this case, the first principal component  $X_1$  explains about 83% of the total variance in the data and it has been interpreted as a composite indicator of tourism attraction (cITA) since it is a weighted linear combination of the initial variables  $Z_i, i = 1, 2, \dots, 5$ , with positive weights, as follows

$$X_1 = 0.488 Z_1 + 0.464 Z_2 + 0.391 Z_3 + 0.437 Z_4 + 0.453 Z_5$$

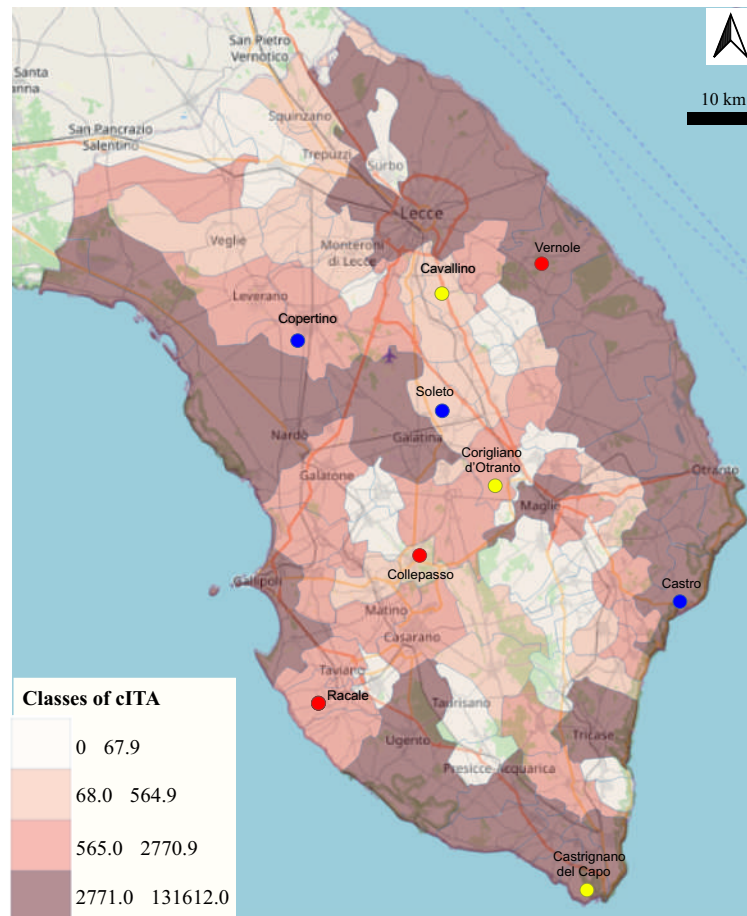
Starting from the cITA, a city's ranking, from poorly attractive to very attractive, has been derived. Note that the proposed composite indicator helps to identify the different levels of tourist attractiveness of the municipalities and it is useful to propose tourist itineraries, which are interesting and sustainable alternatives to the well-known and popular places in the Salento peninsula.

### 3 Results

In Fig. 1 the thematic map of cITA is proposed, where the municipalities have been grouped in four categories according to the quartiles of the cITA. It is evident that almost all the coastal municipalities as well as Galatina and Maglie are the most attractive places.

By way of illustration, starting from the city's ranking three routes over the Salento area have been identified and each route proposes three stops in three different municipalities, which are characterized by different levels of attraction. The first route has been defined along the North-South direction (Cavallino, Corigliano d'Otranto and Castrignano del Capo municipalities, denoted by yellow points in Fig. 1), the second one along the NorthEast-SouthWest direction (Vernole, Collepasso and Racale municipalities, denoted by red points in Fig. 1) and the last one along the NorthWest-SouthEast direction (Copertino, Soleto and Castro municipalities, de-

Title Suppressed Due to Excessive Length



**Fig. 1** Color maps of the cITA.

noted by blue points in Fig. 1). Finally, for each municipality involved in the network, a map of the museums and other cultural points of attraction could also be provided.

## 4 Conclusions

In this paper, some touristic routes were defined by considering the level of appeal of the municipalities, on the basis of relevant variables on tourism and points of attractions. Moreover, the related GIS project represents a useful tool to support the identification of alternative paths among different municipalities, characterized by

the presence of museums, or other cultural sites in nearby or environmental attractiveness. This can be considered a first step towards the construction of a cultural network at a local level which will contribute to improve the local identity of the territory and to support alternative tourist itineraries (in a sustainable vision), which are able to attract more visitors to locations outside the most-visited areas and to increase the visibility of less-known cultural and environmental points of attractions and museums.

**Acknowledgements** This research has been partially supported by the Consorzio Universitario Interprovinciale Salentino - CUIS (grant given to the authors on 2018).

## References

1. Campbell, H., Masser, I.: GIS in local government: some findings from Great Britain. *Int. J. Geogr. Inf. Sci.* (1992) doi: 10.1080/02693799208901933
2. Istat: Indagine sui musei e istituzioni similari. (2017) <https://www.istat.it/it/archivio/167566>
3. Lebart, L., Morineau, A., Warwick, K. M.: *Multivariate descriptive statistical analysis*, John Wiley & Sons, New York (1984)
4. Lo, C.P., Yeung Albert K.W.: *Concepts and Techniques of Geographic Information Systems*, 2nd ed., Pearson Education Canada, Inc., Toronto (2007)
5. Mardia, K. V., Kent, J. T., Bibby, J. M.: *Multivariate Analysis*, London Academic Press (1997)
6. Minucciani, V.: The territory and the small museums: The Case of Piemonte. *Tafer J.* **92**, 1–10 (2017)
7. Sieber, R.: Public participation geographic information systems: A literature review and framework. *Ann. Am. Assoc. Geogr.* (2006) doi: 10.1111/j.1467-8306.2006.00702.x

## 4.30 Statistical applications in education

# Does self-efficacy influence academic results? A separable-effect mediation analysis

*Il senso di autoefficacia influenza i risultati scolastici?  
Un'analisi di mediazione a effetti separabili*

Chiara Di Maria

**Abstract** In causal mediation analysis, natural effects are identified only under strict assumptions involving cross-world counterfactuals. An alternative approach recently developed, called separable, allows for identification of mediational effects in a wide range of models, since it relies on weaker assumptions than those required by natural effects. In this paper, the separable-effect approach is revised and an application to data is presented.

**Abstract** *Nell'ambito della mediazione causale, gli effetti naturali sono identificabili solo sotto rigide assunzioni che coinvolgono controfattuali cross-world. Un approccio alternativo sviluppato di recente, detto separabile, consente di stimare gli effetti di mediazione in un'ampia gamma di modelli, poiché si basa su ipotesi più deboli di quelle richieste dagli effetti naturali. In questo articolo si discute l'approccio a effetti separabili e viene presentata un'applicazione ai dati.*

**Key words:** causal mediation analysis, separable effects, iSP study

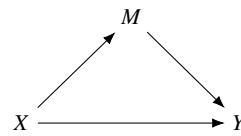
## 1 Introduction

In the last decades, mediation analysis has rapidly grown in popularity and usage among researchers of several areas, due to the wide range of fields in which it can be applied. The aim of mediation analysis is to understand the mechanisms underlying phenomena, by decomposing the total effect of an exposure on a response into a direct effect, and an indirect effect conveyed by a third variable called mediator.

Researchers have proposed a number of definitions for direct and indirect effects, and different methods have been developed to estimate them. In the causal approach, formalised through counterfactuals or potential outcomes, one of the earliest defini-

---

Chiara Di Maria  
University of Palermo, Viale delle Scienze, Building 13, Palermo 90128, e-mail: chiara.dimaria@unipa.it



**Fig. 1** Basic mediation model

tions of direct and indirect effects is that provided by [7], who called them pure and total effects. They were later renamed natural by [4].

The identifiability of natural effects, i.e. the ability to express them as functions of observed variables, has long been discussed in the literature. [7] note that these effects can be identified only under strict, and often implausible, assumptions. [4] proposes a set of assumptions sufficient for identifying natural mediational effects in non-parametric structural equation models (NPSEM). The last of this assumptions, called *cross-world independence assumption*, is rather controversial, as it is impossible to test. Since it is also easy to be violated, in many real contexts natural effects cannot be identified.

For this reason, other definitions of mediational effects have been proposed, relying on weaker, single-world assumptions. [8] introduce a new kind of effects called *separable effects*, which can be identified also in models relying on sets of assumptions weaker than that of NPSEMs, and in more cases. In this paper, we show how separable effects differ from the traditionally used natural effects, why this new conceptualisation is useful, and apply the method to data concerning a randomised experiment related to academic achievements of a group of Japanese students.

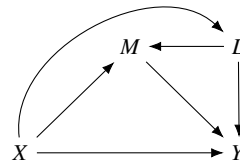
The remainder of the article is organised as follows: in Section 2 we formalise the notion of natural effects and Pearl's assumptions sufficient for their identification, showing why they are somewhat problematic; in Section 3 we illustrate the separable effects approach; Section 4 is devoted to the data analysis and in Section 5 we draw some conclusions.

## 2 Natural effects and the cross-world independence assumption

Causal effects can be expressed through counterfactuals. Consider a simple mediation model as that in Figure 1, where  $X$  denotes the exposure,  $M$  the mediator and  $Y$  the response. Let  $M(x)$  and  $Y(x)$  denote the value that the mediator and the response would assume if  $X$  were set to  $x$ , respectively. Similarly, let  $Y(x, m)$  be the potential outcome value of  $Y$  if  $X$  were set to  $x$  and  $M$  to  $m$ .

The natural direct effect (NDE) and the natural indirect effect (NIE) on the difference scale are defined as  $\mathbb{E}[Y(x, M(x^*)) - Y(x^*, M(x^*))]$  and  $\mathbb{E}[Y(x, M(x)) - Y(x, M(x^*))]$ , respectively, where  $x$  and  $x^*$  are two different values of the exposure and  $Y(x, M(x^*))$  is the value that the outcome would assume if  $X$  were set to  $x$  and the mediator to the value it would take under an intervention setting  $X$  to  $x^*$ .  $Y(x, M(x^*))$  is a nested cross-world counterfactual, since it encompasses an intervention setting  $X$  to two different values, which can never be performed.

**Fig. 2** Basic mediation model including an exposure-induced mediator-outcome confounder



The identifiability of  $Y(x, M(x^*))$  is discussed in [6, 7, 8]. In 2001, [4] proposed four assumptions sufficient to ensure the identifiability of cross-world quantities in NPSEMs. The last of them states the independence, possibly conditional on a set of covariates  $L$ , between  $Y(x, m)$  and  $M(x^*)$ , formally  $Y(x, m) \perp\!\!\!\perp M(x^*) \mid L$ . This is a cross-world independence assumption, since it postulates the independence of two counterfactual variables never observable together.

This assumption is controversial for many reasons. First, since it involves cross-world quantities, it may happen to be identified only in NPSEM, not in models based on weaker sets of assumptions, like the FFRCISTG or the MCM [5, 8]. Second, for this assumption to be satisfied, no exposure-induced mediator outcome confounders have to be present. Figure 2 shows a graph including such a confounder. The reasons why  $L$  leads to a lack of identifiability have been described elsewhere [2, 10, 11]. Unfortunately, it is not uncommon at all to have exposure-induced confounders, thus, natural effects are often non-identifiable. Third, more recently, [1] have shown that the cross-world independence assumption can be violated even if there are no post-treatment confounders. In addition, sometimes natural effects are not the parameters a researcher is interested in, since they encompass an intervention on the mediator, which may be difficult to devise.

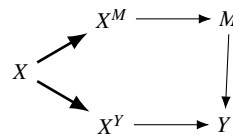
[1, 6] propose alternative assumptions to the cross-world independence, like the no-interaction assumption or parametric assumptions, which ensure identifiability of natural effects. [8] introduce a new kind of mediational effects which overcome the issues related to the definition of natural effects. They are the focus of the next section.

### 3 The separable-effect approach

The separable-effect approach entails associating the graph representing the alleged causal structure to an expanded graph constructed as follows. The exposure  $X$  is split into different components, which become its only children. Each of the children of  $X$  in the original graph becomes child of only one of the different components of  $X$ . Figure 3 depicts the expanded graph corresponding to the mediation model in Figure 1. Bold arrows indicate a deterministic relationship, i.e.  $X \equiv X^M \equiv X^Y$ .

In this representation, direct and indirect effects can be expressed as  $\mathbb{E}[Y(X^M = x^*, X^Y = x) - Y(X^M = x^*, X^Y = x^*)]$  and  $\mathbb{E}[Y(X^M = x, X^Y = x) - Y(X^M = x^*, X^Y = x)]$ , respectively. They will be called separable effects. Notice that these definitions





**Fig. 3** Separable-effect mediation model

do not involve any cross-world quantity. In this framework, the parameter of interest is  $Y(X^M = x^*, X^Y = x)$ , which [9] show to be identified under two sets of non cross-world assumptions. They allow us to identify separable effects not only in NPSEM, but also in weaker models. Moreover, the identifiability is retained even in the presence of an exposure-induced mediator-outcome confounder. [8, 9] show also that separable effects are equivalent to natural ones, and  $\mathbb{E}[Y(X^M = x^*, X^Y = x)]$  can be identified through the g-formula [5] and is given by  $\sum_m \mathbb{E}[Y|x, m]P(m|x^*)$ .

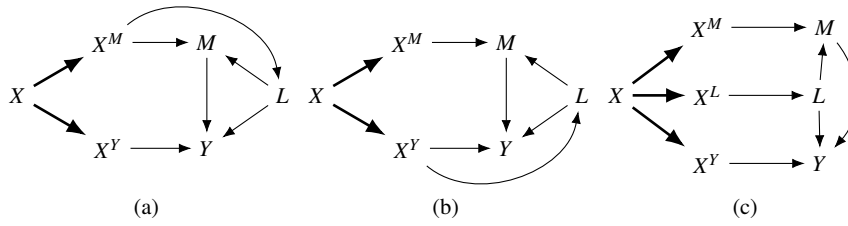
Although separable effects are a useful representation, they are not just a machinery. They have substantive meaning, since they help in gaining insights into the mechanism linking  $X$  to the response. Consider, as an example, a randomised trial to understand the effects of economic support to poor college students on their academic achievements. The effect can be mediated by stress. In Figure 3,  $X^Y$  can be interpreted as the substantive component of the intervention: having more money allows students to buy books or other supplies.  $X^M$  is the psychological component of the intervention, since not having economic issues reduces stress levels and this may increase concentration and willingness to study, leading to better academic results.

## 4 Data analysis

We analyse data from the iSP study [3], which investigates the relationship between perceived self-efficacy and academic success among Japanese students. A group of students were asked to solve 30 anagrams: 27% of pupils were assigned easier anagrams (treatment group) while the rest had to solve standard anagrams. Outperformance of students in the treatment group was expected to foster their perceived self-efficacy and made their academic results improve.

Self-efficacy is a factor with five levels, assessed by students before taking the test, immediately after, two weeks, one month, and two, three, six and twelve months later. Academic achievements are standardised scores measured before the test and two, five, ten, fourteen and seventeen months later. The data set contains also students' gender, the class they attended, and the score they obtained in the test.

We consider a mediational perspective and first assume a causal structure as in Figure 1, where  $X$  is the treatment,  $M$  is self-efficacy, and  $Y$  represents academic results. Then, we include an exposure-induced mediator-outcome confounder  $L$  as in Figure 2, where  $L$  is the score obtained in the test, ranging from 0 to 30. We fit a great variety of models, trying different combinations of variables, and select those showing the best fit. Specifically, we use a continuous version of the test score, between 0 and 1, obtained dividing  $L$  by its maximum. We choose the difference be-



**Fig. 4** Possible expanded graphs associated to the one shown in Fig. 2

tween self-efficacy after a month and self-efficacy level pretest as mediator, and the difference between achievement scores after two months and pretest achievements as outcome. We model the mediator and the outcome via linear regressions and  $L$  through a Beta regression.

There are three expanded graphs associated to that in Figure 2 under which separable effects are identified. It can be proved that the intervention setting  $X^L = x, X^M = x, X^Y = x^*$  in Figure 4(c) is equivalent to the intervention setting  $X^M = x, X^Y = x^*$  in 4(a) and the identifying formula for  $P(Y(X^M = x, X^Y = x^*))$  is

$$\sum_{m,l} P(Y | m, l, x^*) P(m | l, x) P(l | x). \tag{1}$$

Similarly, the intervention setting  $X^L = x, X^M = x^*, X^Y = x$  is equivalent to that setting  $X^M = x^*, X^Y = x$  in 4(b) and the identifying formula is

$$\sum_{m,l} P(Y | m, l, x) P(m | l, x^*) P(l | x). \tag{2}$$

In this context,  $X^Y$  can be interpreted as the substantive component of the test, while  $X^M$  as the psychological component. In this light, we believe that, since  $L$  is the test score, representation in Figure 4(b) may be more plausible.

## 5 Results

We estimate the separable effects for the entire sample ( $n = 267$ ), first not considering  $L$  and then including it. In the former analysis, the total and the direct effect result significant and positive, thus, the treatment has a positive effect on academic achievements, but self-efficacy seems not to have any mediating role. The situation changes when  $L$  is included. We adjust for gender and class, and find that, under the model in Figure 4(b), all effects are significant: the total and the direct ones are positive, the indirect effect negative. This means that the treatment affects academic success both directly, and indirectly through self-efficacy. The negative sign of the

indirect effect may indicate that the treatment negatively influences academic results via a misconception of one's own self-efficacy, probably leading students to be more confident about their capacities, overestimating them.

## 6 Conclusions

In this paper we discussed the separable-effect approach to estimate direct and indirect effects in a causal mediation setting. This can be useful when cross-world independence is not believed to hold. We applied the method to data from a randomised trial, which turns out to have a significant positive direct effect on students' performance and a negative indirect effect through self-efficacy.

## References

1. Andrews, R.M., Didelez, V.: Insights into the "cross-world" independence assumption of causal mediation analysis (2020) Available via ArXiv.org. <https://arxiv.org/abs/2003.10341>. Cited 3 Feb 2021
2. Avin, C., Shpitser, I., Pearl, J.: Identifiability of Path-Specific Effects. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 357-363. (2005) Research: Conceptual, Strategic, and Statistical Considerations. *J. Person. Soc. Psychol.* **51**(6), 1173–1182 (1986)
3. Mori, K., Uchida, A.: Can contrived success affect self-efficacy among junior high school students?. *Res. Educ.* **82**(1), 60–68 (2009)
4. Pearl, J.: Direct and Indirect Effects. In: Breese, J., Koller, D. (eds.) Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pp. 411-420. Morgan Kaufmann, San Francisco (2001)
5. Robins, J.M.: A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect. *Math. Model.* **3**, 1393–1512 (1986)
6. Robins, J.M.: Semantics of Causal DAG Models and The Identification of Direct and Indirect Effects. In: Green, P., Hjort, N., Richardson, S. (eds.) Highly Structured Stochastic Systems, pp. 1–12. Oxford University Press, Oxford (2002)
7. Robins, J.M., Greenland, S.: Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiol.* **3**(2), 143–155 (1992)
8. Robins, J.M., Richardson, T.S.: Alternative Graphical Causal Models and the Identification of Direct Effects. In: Shrout, P., Keyes, K., Ornstein, K. (eds.) Causality and psychopathology: finding the determinants of disorders and their cures, pp. 103–158. Oxford University Press, Oxford (2011)
9. Robins, J.M., Richardson, T.S., Shpitser, I.: An interventionist approach to mediation analysis (2020) Available via ArXiv.org. <https://arxiv.org/abs/2008.06019>. Cited 3 Feb 2021
10. Tchetgen Tchetgen, E.J., VanderWeele, T.J.: On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiol.* **25**(2), 282–291 (2014)
11. VanderWeele, T.J., Vansteelandt, S., Robins, J.M.: Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder. *Epidemiol.* **25**(2), 300–306 (2014)

# Statistics Knowledge assessment: an archetypal analysis approach

## *Valutare l'apprendimento della statistica: un approccio basato sull'analisi archetipale*

Adabbo B., Fabbriatore R., Iodice D'Enza A. and Palumbo F.

**Abstract** As a complement to the traditional learning methodologies, tutoring systems allow tailoring learning activities for students according to their characteristics and abilities, improving the learning outcome. In this framework, the Adaptive Learning in Statistics (ALEAS) ERASMUS+ Project aims to implement an adaptive system for advising in learning Statistics, supporting students' learning process. This article focuses on the ALEAS assessment of students' statistical knowledge, as an essential step to build an appropriate recommender system to provide students with suggestions based on their abilities. Results from a simulation study were used to describe the proposed approach.

**Abstract** *I sistemi automatici di assistenza all'apprendimento, complementari alla didattica tradizionale, consentono di tarare gli strumenti utilizzati sulla base delle caratteristiche e delle abilità degli studenti. In questo contesto, il progetto ERASMUS+, denominato Adaptive LEARNING in Statistics (ALEAS), ha l'obiettivo di implementare un sistema adattivo che supporti gli studenti nell'apprendimento della statistica. Il presente contributo si concentra sulla fase di valutazione delle competenze degli studenti, quale step necessario per la messa a punto di un sistema automatico di consigli da inviare agli studenti per migliorare il proprio livello di abilità. Alcuni risultati su dati simulati sono riportati per descrivere l'approccio proposto.*

**Key words:** Archetypal Analysis, Learning Statistics, Knowledge Assessment

## 1 Adaptive Learning in Statistics

Learning Statistics often represents a problematic task, especially for students enrolled in social and human degree programs. Coping with Statistics makes some students feel unconfident, discouraged, and anxious [10]. The use of a virtual environment for teaching Statistics [9] may help the students learning process by in-

creasing their aptitude and motivation via the support of graphical tools [8]. Tutoring systems allow tailoring learning activities for students according to their characteristics and abilities. In this framework, the Adaptive Learning in Statistics (ALEAS) ERASMUS+ Project<sup>1</sup> aims to implement an adaptive system for advising in learning Statistics, supporting students in the learning process. ALEAS system is designed to support students enrolled in higher education courses in learning Statistics. It is a self-containing digital environment that can make an accurate assessment of students' abilities and support her/him in filling the gaps by allowing access to some learning materials that the student has at hand. According to the Dublin descriptors, the ALEAS system provides a multidimensional evaluation representing a general framework to qualify the expected learning outcome in Higher Education [7]. In particular, it considers the following three out of five Dublin descriptors: knowledge and understanding, applying knowledge and understanding, making judgments. ALEAS includes a knowledge structure for an introductory Statistics course in social and human degree programs, organized in a directed acyclic graph structure by exploiting the Knowledge Space Theory [6]. The knowledge structure consists of ten main Topics (central tendency, variability, etc.) containing several more specific subjects, namely Units, such as mode, median, and arithmetic mean for the central tendency Topic. One or more Topics constitute an Area, the most general subjects' classification (e.g., descriptive measures Area includes central tendency, variability, and graphical displays and tables Topics).

The ALEAS consortium aims to implement a recommender system for students that can be used on a mobile device (a smartphone or a tablet) that provides suggestions to students according to their abilities. To this aim, the system is intended to identify small homogeneous reference sets of students with very similar combinations of abilities and build a corresponding recommendation for each reference small set. Then each of the remaining students will receive a recommendation based on the most similar reference set. In particular, ALEAS assesses students' ability upon completing each Topic and at the end of each learning Area. A multidimensional latent class Item Response Theory (IRT) model [1] is used to estimate the student's ability level score for each Topic, concerning the three considered Dublin descriptors. The Area-level score is obtained by aggregating the topic-level scores. Finally, the reference sets of students are obtained by applying an archetypal analysis (AA, [3]) to the Area-level scores.

The contribution is structured as follows: Section 2 describes the multidimensional latent class IRT model-based scoring system; Section 3 briefly recalls the AA, Section 4 describes the application on a set of synthetic students. The last Section is for conclusion.

---

<sup>1</sup> <https://aleas-project.eu/>

## 2 Students ability assessment: topic-level scores

Given the matrix of students' response patterns, the multidimensional latent class IRT model allows detecting homogeneous groups of students according to their performance in each of the Dublin descriptors [4]. The multidimensional latent class IRT models represent an extension of the traditional IRT models, in that both the constraints of unidimensionality and the continuous nature of the latent trait are released [1]. Thus, the latent trait is defined through a discrete distribution with  $\xi_1, \dots, \xi_k$  support points defining  $k$  latent classes. The prior probability of belonging to the latent classes is expressed by the class weights  $\pi_1, \dots, \pi_k$ , with  $\sum_{c=1}^k \pi_c = 1$  and  $\pi_c \geq 0$ .

A multidimensional latent class model with  $d$  dimensions ( $d = 1, 2, 3$ ), corresponding to the three Dublin descriptors, is applied to obtain the topic-level classification. In the model, each item is related only to one latent trait (between-item multidimensionality) and the probability that the student  $i$ , with the ability vector  $\theta_i = [\theta_{i1}, \theta_{i2}, \theta_{i3}]'$ , correctly answers the dichotomously-scored item  $h$  (with  $h = 1, \dots, H$ ) is:

$$g[P(X_{ih} = 1|\theta_i)] = \log \frac{P(X_{ih} = 1|\theta_i)}{P(X_{ih} = 0|\theta_i)} = a_h \left( \sum_{d=1}^3 \delta_{hd} \theta_{id} - b_h \right),$$

where  $g(\cdot)$  is the logit link function,  $X_{ih}$  is the response of the subject  $i$  to the item  $h$  with realization  $x_{ih} \in [0; 1]$ ,  $\delta_{hd}$  is a dummy variable equal to 1 if the item  $h$  measures the latent trait  $d$ . Moreover, according to the two-parameter logistic (2PL) parametrization, only the item discrimination  $a_h$  and item difficulty  $b_h$  parameters were considered in the model. The Expectation-Maximization (EM) algorithm [5] is used to estimate parameters, then the posterior probability for each student to belong to the latent classes is computed. In fact, the topic-level score corresponds to the expected a posteriori (EAP) estimate of the Dublin descriptors-defined ability. It is worth noting that the number of latent classes can be either taken according to some external knowledge or via a model selection process based on a data-driven approach.

## 3 Defining archetypal students

An archetype is, by definition, an original model such that objects of the same kind are copied from it or based on it. The term 'archetype' is common in the art, behavioral sciences, modern psychological theory and literary analysis. In Statistics, the concept of archetypes was first introduced by Cutler and Breiman [3] in the context of Archetypal Analysis (AA), an unsupervised learning technique designed to synthesize a set of multivariate observations through a few special vectors, the *archetypes*.

To formally describe AA, let  $\mathbf{X}$  be a  $n \times p$  data matrix where each row  $\mathbf{x}_i$  is an ob-

servation on a set of  $p$  covariates. Let  $k \leq n$  and for all  $j = 1, \dots, k$  define the vector  $\mathbf{z}_j := \sum_{i=1}^n b_{ij} \mathbf{x}_i$  to be a convex combination of the data points. The goal is to find an approximation for every  $\mathbf{x}_i$  via a convex combination of the form  $\sum_{j=1}^k a_{ji} \mathbf{z}_j$ . More precisely, if we denote by  $\mathbf{X}^T$  the transpose of  $\mathbf{X}$  and let  $\mathbf{Z}$  to be the  $p \times k$  matrix whose columns are the  $\mathbf{z}_j$  then the problem to solve is the following:

$$\min_{\mathbf{A}, \mathbf{B}} \|\mathbf{X}^T - \mathbf{Z}\mathbf{A}\|^2 = \min_{\mathbf{A}, \mathbf{B}} \|\mathbf{X}^T - \mathbf{X}^T \mathbf{B}\mathbf{A}\|^2$$

under the constraints that  $\mathbf{A} \in \mathbb{R}^{k \times n}$  and  $\mathbf{B} \in \mathbb{R}^{n \times k}$  are column stochastic matrices, i.e., the following conditions hold:

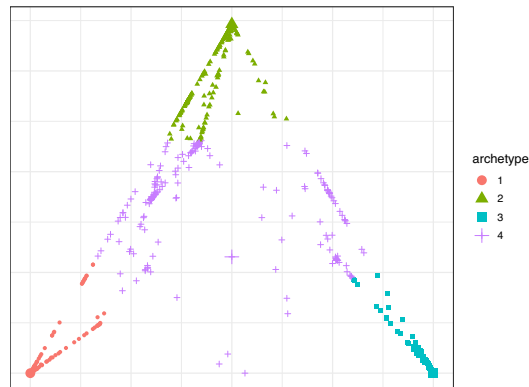
$$\begin{cases} a_{ji}, b_{ij} \geq 0 \quad \forall i, j \\ \sum_{j=1}^k a_{ji} = 1, \sum_{i=1}^n b_{ij} = 1. \end{cases}$$

Although there is no closed-form solution to this problem, the subproblems obtained by alternating optimization of  $\mathbf{A}$  for fixed  $\mathbf{B}$  and vice versa form convex optimization problems in a suitable norm (e.g., Frobenius' norm) so the alternating procedure proposed by Cutler and Breiman always converges, but there is no guarantee that the solution found by the algorithm is global.

The vectors  $\mathbf{z}_j$  obtained after convergence are the archetypes and in [11] it is proven that they lie on the convex hull generated by  $\mathbf{X}$ . Moreover, if the convex hull of the data has  $q \leq n$  vertices and  $k = q$ , then the archetypes are exactly the vertices of the convex hull, while for  $k < q$  they can be thought of as extreme observations defining the vertices of the principal convex hull (PCH), the dominant approximation of the convex hull of  $\mathbf{X}$ .

#### 4 Grouping students using archetypes: an example

A set of 800 students' answer patterns is randomly generated, referring to two hypothetical topics assumed from the same learning Area. According to the Dublin descriptors, the simulation design is based on different levels of ability, consisting of the following four learning outcome combinations. The first three levels are respectively characterized by poor, good, and average performance in all the three dimensions; the last one is characterized by good performance in knowledge (K) and poor performance in both application (A) and judgment (J). For each combination,  $n = 200$  response patterns are generated using the R package MAT [2]. The latent trait is assumed to be normally distributed with  $\sigma = 1$  and  $\mu = 1$  for good performers,  $\mu = 0$  for average performers, and  $\mu = -1$  for poor performers. The latent class IRT model has been applied to the topic-wise generated students data to obtain scores for K, A, and J: the corresponding scores are aggregated together to obtain an Area score. The number of latent classes is  $k = 4$  as many as the considered learning outcome combinations.



**Fig. 1** Ternary map of students: three archetypes on the vertices and a fourth barycentric one

An archetypal analysis approach is applied to define a set of reference students with a characteristic learning outcome. In particular, we considered 3 ( $k - 1$ ) archetypes and one more barycentric point of the archetypal space, where  $k$  is the number of the latent classes in the topic-level classification. Looking at the archetype profiles, we have that: the first archetype represents a student with good performance in all the three Dublin descriptors we considered, the second one corresponds to a student with poor performance in all these dimensions, whereas the third archetype represents a student with good performance in K but poor performance in both A and J. On the other hand, the barycenter refers to an average performer.

Thus, a different recommendation is built for each of the four reference students, for instance, the second archetypal student will have: *“That’s too bad! Your performance in the topics belonging to this Area was poor! I strongly advise you to give more attention to the formal definition of the theoretical concepts and do more exercises to improve your calculation skills and your ability to evaluate information to reach an appropriate judgment in statistical matters”*.

The considered students are assigned to the ‘closest’ archetype. In particular, Figure 1 shows the so-called ternary map depicting both students and archetypes. Each student is described by  $a_{ji}$ , with  $j = 1, \dots, (k - 1)$  and  $i = 1 \dots, n$ . Since,  $\forall i$ , it results that  $\sum_{j=1}^{(k-1)} a_{ji} = 1$ , a two dimensional map can be used: the original coordinates of the  $i^{th}$  student are  $\{a_{1i}, a_{2i}, a_{3i}\}$ ; in the ternary plot the  $i^{th}$  student has coordinates  $x_{1i} = a_{2i} + a_{3i}/2$  and  $x_{2i} = a_{3i}\sqrt{3}/2$ . Note that the  $k - 1$  archetypes are the vertices of the triangle containing all the points, whereas the fourth archetype is simply the barycenter of that triangle. Students with the same label are assigned to the same archetype, therefore they will receive the same recommendation.



While the identification of students with similar abilities is a task shared with cluster analysis, the AA is more suitable for the task at hand: in fact, archetypes have peculiar characteristics, and, by definition, they differ from each other as much as possible: this makes it easier to generate *ad hoc* recommendations for each archetype. The barycentric archetype is inserted to capture all the students with average characteristics. Barycentric students do not lack of any of the considered abilities: they should consolidate their status in the current Area by still keeping practicing.

## 5 Conclusion

Remote learning plays an increasingly crucial role in education. ALEAS system aims to offer a virtual training space where students belonging with similar characteristics can improve their learning abilities in an adaptive way. The system has a client-server architecture, where the client application runs on mobile devices based on the Android OS. The adaptiveness of the ALEAS system is provided by the estimates of the items' parameters being constantly adjusted according to the overall class abilities. In ALEAS, the *typical* Statistics syllabus is divided into knowledge Areas. Once a student completes an Area, the system will send him a recommendation (feed-back) and a quick reference guide on the topics that proved to be more difficult for the student in question.

## References

1. Bartolucci, F.: A class of multidimensional irt models for testing unidimensionality and clustering items. *Psychometrika* **72**(2), 141 (2007)
2. Choi S., W., King D., R.: MAT: Multidimensional Adaptive Testing. R package version 2.2 (2014). URL <https://CRAN.R-project.org/package=MAT>
3. Cutler, A., Breiman, L.: Archetypal analysis. *Technometrics* **36**(4), 338–347 (1994)
4. Davino, C., Fabbriatore, R., Pacella, D., Vistocco, D., Palumbo, F.: Aleas: a tutoring system for teaching and assessing statistical knowledge. *CEUR Workshop Proc.* p. 2730 (2020)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22 (1977)
6. Doignon J., P., Falmagne J., C.: Spaces for the assessment of knowledge. *Int J Man Mach Stud.* **23**, 175–196 (1985)
7. Gudeva L., K., Dimova, V., Daskalovska, N., Trajkova, F.: Designing descriptors of learning outcomes for higher education qualification. *Qual. Rep.* **46**, 1306–1311 (2012)
8. Klein, G., Dabney A., R.: *The Cartoon Introduction to Statistics*. Hill and Wang (2013)
9. Lamezón S., L., López R., R., Aguilar L., M.A., Lorenz L., M.A.: Social significance of a virtual environment for the teaching and learning of descriptive statistics in medicine degree course. *Humanid. méd.* **18**, 50–63 (2018)
10. Malik, S.: Undergraduates' statistics anxiety: A phenomenological study. *Qual. Rep.* **20**, 120–133 (2015)
11. Mørup, M., Hansen, L.K.: Archetypal analysis for machine learning and data mining. *Neuro-computing* **80**, 54–63 (2012)

# Exploring drivers for Italian university students' mobility: first evidence from AlmaLaurea data

## *La mobilità degli studenti universitari italiani: prime analisi sui dati AlmaLaurea*

Giovanni Boscaino and Vincenzo Giuseppe Genova

**Abstract** This article is part of a national Project aimed at studying Italian university students' mobility. The novelty proposed here lies in the use, for the first time, of individual data from the census surveys that the *AlmaLaurea* consortium conducts on all Italian graduates. The advantage of this data consists of information that was previously unavailable. For example, information about the socio-economic conditions of the students' parents, satisfaction with the Bachelor's degree course and university services, and employment status recorded 1, 3, and 5 years after graduation. Here we report the first analyses conducted on the data relating to the University of Palermo Bachelors who enrolled in a Master's degree course in Italy, studying some of their characteristics and connecting them to their mobility.

**Abstract** *Il presente articolo si inserisce in un PRIN dedicato alla mobilità studentesca all'università. Per la prima volta sono disponibili i dati individuali delle indagini AlmaLaurea sui laureati italiani. Il vantaggio di questi dati risiede nelle informazioni prima non disponibili, ad esempio le condizioni socio-economiche dei genitori degli studenti, la soddisfazione verso il corso di laurea triennale e verso l'università, ma anche informazioni sullo stato occupazionale dopo 1, 3, e 5 anni dal conseguimento del titolo. Qui si riportano le prime analisi dei dati relativi ai laureati triennali all'Università di Palermo e che si iscrivono a un corso di laurea magistrale, studiandone alcune caratteristiche e mettendole in relazione alla mobilità.*

**Key words:** university students' mobility, AlmaLaurea, mobility determinants, North-South divide

---

Giovanni Boscaino

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Viale delle Scienze, Edificio 13, Palermo (Italia), e-mail: giovanni.boscaino@unipa.it

Vincenzo Giuseppe Genova

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Viale delle Scienze, Edificio 13, Palermo (Italia) e-mail: vincenzogiuseppe.genova@unipa.it

## 1 Introduction

In the last decades, student mobility has increased significantly. Flows have mainly involved international mobility, driven either by the desire to access an education deemed to be better or by the need to leave an unfavourable economic or social situation. Recent student mobility, which is mainly typical of the university environment, seems to be an anticipation of the migration that has always characterised people's flow to reach more promising labour markets [3, 12, 14]. Often, Students who move to a foreign country to pursue their university studies usually do not return to their country of origin [17]. As a result, research attention to this topic has increased, and the number of scientific publications related to international student mobility has grown exponentially. Indeed, as migration flows have increased, several economic, social and political problems have arisen [11]. For example, universities that attract students act as a boost for the local economy, creating a very favourable spin-off for the surrounding economic environment. On the other hand, countries of origin suffer a loss in their human capital investment when students no longer return home. Moreover, students are often financially supported by their relatives, who transfer capitals from their home country to the destination.

Sometimes, however, the issues related to student mobility not only concern international but also internal movements within a country. There are large flows of students who even move within a country searching for better education and better job opportunities after graduation. What happens between many nations is replicated within a country, like in Italy. For decades, it has been suffering from increasing student mobility, which manifests itself only in one direction: from the South to the Centre-North. Italy is economically and socially divided in two: the poor South and the rich and wealthy North. The policies pursued by governments in recent decades have not succeeded in reducing this divide. So, the former migration of adults who went North to find work now is "evolved". Nowadays, young people move to obtain better education directly in the country's most productive and wealthy areas. Once again, universities are attracting human capital and labour force to the detriment of the southern regions, slowly becoming increasingly depopulated and impoverished because, very often, students who leave never return home.

Usually, student migration studies focused mainly on High School's transition to Higher Education and post-student paths [13]. Besides, the attention is more on international migration than intra-national one [9, 16]. Therefore, the Italian case is an interesting one due to the peculiarities just mentioned. Since the impoverishment of the South due to migration is a political problem and an economic and social one, the attention towards it is high. The Ministry of University has financed a three-year project on this topic<sup>1</sup>. Many studies are starting to be published, intending to quantify the phenomenon and study its determinants [1, 4, 5, 7, 10, 2]. Most of the analyses refer to the MOBYSU.IT database provided by the Ministry and containing

---

<sup>1</sup> Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide.", n. 2017HBTk5P - CUP B78D19000180001

information on all cohorts of university students enrolled in Italy since 2008 [8]. This database is of considerable importance because it makes it possible to link the student's basic socio-demographic information to his/her university history, both in terms of performance and mobility.

This article is part of the Project and also relates to students' mobility. The novelty proposed does not lie in a different methodology, but the new data available. For the first time, the census survey's individual data on graduates conducted by the *AlmaLaurea* consortium has become available. These data are useful because they enrich the information in MOBYSU.IT. with other socio-demographic information and students' opinions. In particular, our attention is focused on those who graduate from a Bachelor's degree course (*BA*) in the South and who then continue their studies in a Master's degree course (*MA*) in the Centre-North. The first data available allows us to focus on the Bachelors at the University of Palermo in Sicily.

## 2 Data and model outcomes

The first data available concern the information collected by the *AlmaLaurea* consortium regarding the cohort of *BA* that gained their degree in the calendar year 2018, at the University of Palermo in Sicily. This cohort is surveyed at least in two moments. A first survey is conducted at graduation, gathering information about student's university experience and post-graduation perspectives (to keep studying, look for work, etc.). Subsequently, one year later the cohort is surveyed to get the graduate's employment status or whether the student is continuing his/her studies. In this paper we show some results about the identification of some possible factors of mobility, introducing some covariate as the student's family's socioeconomic conditions, the student's perception of their university experience, and other contextual variables that can be useful to detect the motivations for mobility out the South. For the sake of brevity, we report here only some of the available information and the results of a first modelling approach to the probability of moving. It should be enough to give an idea of the potential of these data.

Table 1 summarises some basic information. In particular, our analysis was carried out on the 3758 *BA* students who enrolled in an Italian *MA* degree. Most students are Male (59%), and half of the students belong at least to an Upper-middle Social class. Two-thirds of students graduate one year later than planned (i.e. three years), and one in three students lives at such a distance that it is preferable to rent a house near the University of Palermo. This variable has been considered a proxy of the usual definition of "off-site" or "on-site", with a piece of better information about the additional "costs" incurred by the student to attend university courses. Most of the students expressed satisfaction for their *BA* Degree Course. *BA* enrolled mostly in a *MA* course in a Scientific or Social area of study.

In addition, we have built the "Mover Status" variable. It allows distinguishing between those students who enrol at *MA* of Palermo (Stayers) and those who enrol at *MA* of a different university out of Sicily (Movers). Therefore, conditioning by

**Table 1** Synopsis of some 2018 Bachelors' information

		Total	Mover Status	
			Stayers	Movers
Gender	Female	1527	870	332
	Male	2231	567	242
Social class	Upper	554	221	101
	Upper-middle	1188	433	226
	Lower-middle	661	272	99
	Lower	882	350	103
Delay at graduation (years)	0 – 1	2506	1040	410
	2	490	178	60
	3	281	109	41
	4	143	43	16
	5 and more	338	67	47
Rented house	No	2200	831	316
	Yes	1147	464	219
Satisfaction	Not Satisfied	407	100	84
	Satisfied	2936	1193	451
Field of study	Health	319	27	6
	Scientific	1466	660	249
	Social	1196	378	223
	Humanities	777	372	96

Note: Data are reported net of missing information

Movers 58% are Female, 43% belong to Upper-middle Social class, 71% graduate at most one year late, 59% do not take a house for rent, 16% are not satisfied with their BA Degree Course, and almost all the students enrol in a Scientific or Social Degree Course. In terms of “mobility risk”, data reveal that Rented house students have a 16% significant higher probability of moving than others. If we dichotomise Social class in just Upper and Lower, Upper class students seem to have a 36% higher significant probability of moving than the Lower ones. Mobility seems to be 66% higher in unsatisfied students compared to the satisfied ones. Finally, with respect to the Field of study (i.e. Degree Course categorised according to the Ministry of University classification of subjects) we notice more movers in Social field (37%) than Scientific (27%) and Humanities (20%). We have not considered the Health area as this is affected by mobility due to national regulation.

*AlmaLaurea* data could be also used to analyse the probability of moving (i.e. to be Movers), with respect to some covariates. As first analysis, a logit model has been applied. In particular, we are interested in these student's characteristics (baseline category in parentheses): Gender (Male), Rented house (NO), Student satisfaction about his/her BA course (YES), Delay at graduation (0-1), Degree Mark<sup>2</sup>, the Social class (Lower), Satisfaction (Satisfied), and the subject Field of the Mas-

<sup>2</sup> In Italy graduation mark is an integer in [66, 110 cum laude].

ter Degree Course (Humanities). After testing different models, we obtained the more useful model for explaining the probability of moving (in terms of parsimony and AIC) excluding only Gender. In particular, the Social class was re-levelled into three categories (aggregating the two central ones into the "Middle" class) according to Schizzerotto [15], and the Degree mark was dichotomised into the categories " $66 \leq 109$ " and " $> 109$ ". The new Degree mark identifies the very top performer Bachelors from the others. The results in table 2 suggest that being unsatisfied about university experience (OR 2.26), belonging to a non-Lower Social class (OR 1.59 for Middle class and for Upper class), paying a House rent to attend the course (OR 1.37), being a "very top student" (OR 1.53), and enrolling in a non-Humanities field of MA are each significant factors linked to the enrolment out of Sicily. The results allow us to outline the profiles most likely to move and stay. Those students who are very good, belonging to a high social class, do not live in Palermo proximity, are not satisfied with their BA Degree Course, and want to enrol in a Social Degree Course have the highest estimated probability of moving (0.69). On the contrary, the students who likely enrol in a MA Degree Course at Palermo University (estimated probability of moving 0.11) are not the "top student", belonging to a low social class, live in Palermo proximity, are satisfied with their BA Degree Course, and want to enrol in a Humanities Degree Course.

**Table 2** Estimates for logit model on moving probability

Coefficients	Parameters	<i>p-values</i>
Intercept	-2.0385	< 0.001
Social class: Middle	0.4624	< 0.001
Social class: Upper	0.4650	< 0.001
Satisfaction: Unsatisfied	0.8161	< 0.001
House rent: Yes	0.3118	< 0.001
Degree mark > 109	0.4252	< 0.001
Field (Scientific)	0.4703	< 0.002
Field (Social)	0.8420	< 0.001

### 3 Brief concluding remarks

This paper aims to highlight the potential of the *AlmaLaurea* dataset. The individual data coming from the surveys allow us to deepen multiple aspects, enriching the information coming from the Ministry dataset. The two datasets are currently being merged to a unique database having more informative records. *AlmaLaurea* data can be affected by some non-responses that can be covered by *MOBYSU.IT* data. Future studies may concern aspects such as the link between parents' professions and the degree course chosen by the student, or the employment success of Stayers and Movers. Moreover, an in-depth study has to be developed to take into account the fact that we are dealing with cohorts of graduates and not of enrolled students. For instance, the university experience's satisfaction has to be related to the particular

educational offer. Surely, analyses of these enriched data can be useful support for university governance.

**Acknowledgements** The authors would like to thank the Placement Delegate of the University of Palermo, Professor Ornella Giambalvo, for providing the *AlmaLaurea* consortium data.

## References

1. Attanasio M., Enea M., Albano A.: Dalla triennale alla magistrale: continua la “fuga dei cervelli” dal Mezzogiorno d’Italia, Neodemos, ISSN: 2421-3209 (2019)
2. Attanasio, M., Ragozini, G., and Porcu, M.: Verso Nord. Le nuove e vecchie rotte delle migrazioni universitarie, Franco Angeli, ISBN 978-88-35-10562-6 (2020)
3. Beine, M., Noël, R., and Ragot, L.: Determinants of the international mobility of students. *ECON EDUC REV*, 41:40–54 (2014)
4. Boscaino, G., Genova, V.G. (2020) “Pull factors for university students’ mobility: a gravity model approach”. In Book of short papers - SIS 2020 (pp. 73-78). Pearson. ISBN: 9788891910776
5. Boscaino, G., Sottile, G., and Adelfio, G.: Migration and students’ performance: detecting geographical differences following a curves clustering approach. *J APPL STAT*, 1-15 (2020)
6. Breschi, M., Fornasin, A., Gonano, G., and Ruiu, G.: Il capitale umano in Italia: un’analisi empirica a livello micro territoriale. In Quaderni di Ricerca, No 2, Dipartimento di Scienze Economiche e Aziendali, Università degli Studi di Sassari, Forum Editrice, ISBN: 978-88-3283-220-4 (2020)
7. Columbu, S., Porcu, M., Primerano, I., Sulis, I., and Vitale, M. P.: Geography of Italian student mobility: A network analysis approach. *SOCIO-ECON PLAN SCI*, 73, 10091 (2021).
8. Database MOBYSU.IT [Mobilità degli Studi Universitari in Italia], research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenze, Cattolica and Napoli Federico II, Scientific Coordinator Massimo Attanasio (UNIPA), Data Source ANS-MUR/CINECA
9. Ellis, M.: Reinventing us internal migration studies in the age of international migration. *POPUL SPACE PLACE*, 18(2):196–208, (2012)
10. Genova, V. G., Tumminello, M., Enea, M., Aiello, F., and Attanasio, M.: Student mobility in higher education: Sicilian outflow network and chain migrations. *Electronic J APPL STAT ANALYSIS*, 12(4), 774-800 (2019)
11. Gümüş, S., Gök, E., Esen, M.: A Review of Research on International Student Mobility: Science Mapping the Existing Knowledge Base. *Journal of Studies in International Education*, 24(5), 495-517 (2020)
12. King R. and Raghuram, P.: International student migration: Mapping the field and new research agendas. *POPUL SPACE PLACE*, 19(2):127–137 (2013)
13. Mahroum, S.: Highly skilled globetrotters: mapping the international migration of human capital. *R&D MANAGEMENT*, 30(1):23–32 (2000)
14. Raghuram, P.: Theorising the spaces of student migration. *POPUL SPACE PLACE*, 19(2):138–154 (2013)
15. Schizzerotto A.(Ed.): Vite ineguali. Disuguaglianze e corsi di vita nell’Italia contemporanea, Bologna, Il Mulino, pp. 400 (2002)
16. Verbik, L., and Lasanowski, V.: International student mobility: Patterns and trends. *World Education News and Reviews*, 20(10):1–16, (2007)
17. Wu, C., and Wilkes., R.: International students’ post-graduation migration plans and the search for home. *GEOFORUM*, 80:123–132 (2017)

# Can Grading Policies influence the competition among Universities of different sizes?

## *Gli effetti delle politiche di valutazione sulla competizione fra atenei di dimensioni diverse*

Gabriele Lombardi and Antonio Pio Distaso

**Abstract** One of the main commonplaces about students' population at the Higher Education level is that freshmen are attracted by universities which adopt soft grading policies, so to achieve their graduation in the easiest way as possible. At the same time, very little evidences are generally provided about the effect that such a strategy can have on the universities, if adopted. Thanks to the Italian University Register (ANS), we analyze the cohorts of Italian freshmen between 2010-2012. As it will be shown, if universities would compete each other through grading policies, only those which already have a competitive advantage can benefit from this strategy, while the others might only slide down into a vicious circle.

**Abstract** *Si ritiene usualmente che gli studenti universitari preferiscano scegliere università che sono solite assegnare alte valutazioni, così da laurearsi più facilmente. Ciononostante in letteratura si è indagato poco l'effetto che può avere sugli atenei il fatto di alleggerire le proprie politiche di valutazione. Grazie ai dati disponibili all'interno dell'Anagrafe Nazionale Studenti (ANS) si osservano gli immatricolati al primo anno delle coorti 2010-2012. Dall'analisi emerge che se le università competessero attraverso politiche di valutazione più leggere, gli unici a beneficiarne sarebbero gli atenei che presentavano già vantaggi competitivi preesistenti, mentre gli altri peggiorerebbero ulteriormente la propria attrattività.*

**Key words:** Grading Policies, Higher Education, Student Population, Mixed Logit

---

Gabriele Lombardi  
Area Organizzazione e Sistemi Informativi (AOSI), Rettorato, via Banchi di Sotto 55, 53100, Siena (Italy), e-mail: gabriele.lombardi3@unisi.it

Antonio Pio Distaso  
Department of Economics and Statistics (DEPS), P.zza S. Francesco 7-8, 53100, Siena (Italy) e-mail: antoniopio.distaso@student.unisi.it



## 1 Introduction

The present article pursues two main goals. The first aim is to control the effect that softening grading policies by universities have both on the student decision process and on academic institutions' reputation. On the one side, the artificial increasing of grades can be seen as a strategy played by directorates in order to attract students, but also signaling to the job market that very good scholars were trained [1]. On the other side, studies about soft grading policies on students and universities suggest how they can push out the first and cause a reputation loss for the latter [6]. The second aim is to provide useful hints about the universities' attractiveness in the peculiar Italian framework. Indeed, several works show how Italian departments suffer a perverse incentive structure [3] which rewards both the number of students enrolled and the speed needed for reaching graduation [8], so exacerbating the strong disparities between North and South in the country. Consequently, if students positively evaluate soft grading policies both the goals can be easily reached, at the expense of their competence. If this is not true, the two objectives conflict each other. Moreover, several investigations about Italy highlight how students prefer to enrol in northern universities for several reasons: *i*) to anticipate the job market reaching in advance those regions with the lowest unemployment rates [4]; *ii*) links and connections among universities and local areas in the South are below the national average [5]; *iii*) students positively reward heterogeneity in the educational offer [2]. From this last point of view, it is sufficient to think that nowadays only 3 'Giant' Universities (i.e. more than 40,000 enrolled) over 10 are located in the South. On the other side, 7 over 10 'Small' universities (i.e. less than 5,000 enrolled) are located between South and Islands. This could be the umpteenth factor exacerbating the competition among southern and northern universities, with the first dramatically caught in a vicious circle, as the second in a virtuous one, constantly increasing the gap. Consequently, in the next section the McFadden's Choice Model will be briefly explained. Data are jointly obtained from University Student Register (ANS)<sup>1</sup>, Ministry of University and Research (MUR) and National Institute of Statistics (ISTAT). As it will be clarified, a competitive advantage emerges for 'Giant' and 'Medium' universities against 'Big' and 'Small', respectively, in the possibility of using both grading policies and fees as a leverage for attracting students. On the other side, territorial characteristics consistently attract students toward the northern regions.

## 2 Data and Model

In order to explore the determinants of students' university choice, a McFadden's Choice Model [7] will be performed, setting for each student a set choice including

<sup>1</sup> Data - drawn from the Italian "Anagrafe Nazionale della Formazione Superiore"- has been processed according to the research project "From high school to the job market: analysis of the university careers and the university North-South mobility" carried out by the University of Palermo (head of the research program), the Italian "Ministero Università e Ricerca", and INVALSI.

Can Grading Policies influence the competition among Universities of different sizes?

all the Italian public universities which host a degree course in a specific degree class. The intrinsic assumption is that each student decides *what* she wants to study, before than *where* she wants to study it. The model can be summarized as:

$$\max L(\beta, \lambda_j) = \prod_{i=1}^I \prod_{j=1}^J (p_{ij})^c \implies p_{ij} = \frac{e^{x'_{ij}\beta + w'_i\lambda_j}}{\sum_{l=1}^m e^{x'_{il}\beta + w'_i\lambda_l}}, \quad j = 1, \dots, J. \quad (1)$$

Thus,  $p$  is the probability that each student  $i$  chooses a university  $j$ .  $\beta$  and  $\lambda$  are respectively the sets of coefficients associated to the alternative-specific  $x$  and case-specific  $w$  covariates. The dichotomic variable  $c$  identifies the chosen institution.

Four models will be estimated for 'Giant' (i.e. more than 40,000 enrolled), 'Big' (i.e. 20,000 to 40,000 enrolled), 'Medium' (i.e. 5,000 to 20,000 enrolled), and 'Small' Universities (i.e. less than 5,000 enrolled).

Table 1 shows how the three main alternative-specific indicators were calculated. The *Grade Ratio* (GR) measures the average combination between average grade  $v$  and number of credits  $CFU$  earned by each enrolled student  $i$  provided by any degree course  $d$  in a specific academic year  $y$ . Then, it is divided by the same average calculated on all the degree courses of the country in the same degree class  $c$ . From this point of view, GR represents a proxy of how much soft grading policies of a certain course are with regard to all its competitors. Similarly, the *Dropout Ratio* (DR) is calculated on the dichotomous variable  $r$ , which counts the number of students abandon a specific degree course during the first year, being interpreted as a proxy of how hard grading policies are. At last, Average Fees  $\bar{F}$  are calculated as the average fee  $f$  paid from each student enrolled in a certain university  $u$ , as computed by the MUR. Finally, controls are included for the difference between the youth unemployment rate by gender and province of course and residence, and for distance between course and residence (i.e. *shoe leather cost*).

**Table 1** Formulas for the calculation of the three main indexes in the analysis.

<b>Grade Ratio</b>	<b>Dropout Ratio</b>	<b>Average Fees</b>
$GR_{d y} = \frac{\frac{1}{I_{d y}} \sum_{i=1}^{I_{d y}} \bar{v}_{i d,y} CFU_{i d,y}}{\frac{1}{I_{c y}} \sum_{i=1}^{I_{c y}} \bar{v}_{i c,y} CFU_{i c,y}}$	$DR_{d y} = \frac{\frac{1}{n_{d y}} \sum_{i=1}^{n_{d y}} r_{i d,y}}{\frac{1}{I_{c y}} \sum_{i=1}^{I_{c y}} r_{i c,y}}$	$\bar{F}_{u y} = \frac{1}{I_{u y}} \sum_{i=1}^{I_{u y}} f_{i u,y}$

Not shown in the estimations presented in Section 3 for the sake of brevity, also case-specific covariates are included in the analysis: type of High school attended by each student and final grade awarded, and academic year of first academic enrollment. Descriptive Statistics are shown in Table 2, for the entire sample but also differentiated for the size of chosen universities.

Finally, as a robustness check, Table 4 will present the same analysis with the addition of the universities' ranking scores provided by the CENSIS, the cities' cost of life obtained averaging the mean price of coffee, bus tickets, bread and a "pizza

**Table 2** Descriptive Statistics (mean) of alternative and case specific variables.

	Descriptive Statistics				
	Entire Sample	Giant	Big	Medium	Small
<i>Alternative-specific</i>					
Grade Ratio	0.973	0.974	0.947	1.023	0.948
Dropout Ratio	1.009	0.996	1.012	1.022	1.074
Average Fees	920.332	971.402	825.978	976.922	834.435
$\Delta$ Unemp. Rate	-1.221	-1.237	-1.141	-1.618	-0.146
Distance (mt.)	90,014.89	91,853.45	86,873.12	97,122.98	66,946.76
<i>Case-specific</i>					
Female	0.568	0.566	0.577	0.553	0.589
Classyc Lyceum	0.158	0.178	0.153	0.121	0.121
Other HS	0.44	0.407	0.433	0.510	0.528
HS Final Mark	78.14	78.571	78.032	77.620	76.751
A.y. 2010	0.338	0.338	0.346	0.33	0.319
A.y. 2011	0.337	0.338	0.331	0.342	0.342
A.y. 2012	0.325	0.324	0.323	0.328	0.339

and beer” based dinner, and the Student-Teacher ratio for each university (Source: CENSIS for the both of them). This variables are not included in the main analysis since they are accessible only for the a.y. 2010 and 2011, and information is not available for all the considered universities.

### 3 Discussion and conclusion

Before to observe results in Table 3, it is important to clarify how the computed Odds Ratio cannot be compared across the four models, since they refer to different subsamples. Clearly, the first evidence to be pointed out is that soft grading policies are in general a repulsive factor for students’ choice, with the only exception of Medium size universities. At the same time, the Dropout Ratio is not significant for Small universities and always significant and greater than 1 for all the others. So, if soft grading policies would be adopted as a strategical tool, following the hypothesis that students positively evaluate them, it would be a mistake. It is also important to point out that, even if we used a one year lag in computing the two ratios, there is no guarantee that students actually can be aware of the ‘easiness’ of the course they are applying for. Maybe information can be spread from older students by a mouth-to-mouth, but also in this case the conclusion which fits better our results is that students prefer ‘harsher’ courses. This can be explained simply by the will of holding a degree with the highest reputation as possible. On the other side, territorial characteristics provide results much more coherent. Students want to move toward locations with low unemployment rates and distance consistently emerges as a ‘shoe-leather cost’. But, Small universities are a special case. Indeed, looking at Table 2, students who choose them move from their surroundings. Generally, this

Can Grading Policies influence the competition among Universities of different sizes?

choice is due to economic needs, or to the fact that a small university with a high reputation is settled in the nearby, so making irrelevant the issue of moving toward a province with a lower unemployment rate.

**Table 3** Odds Ratios and Standard deviation (between brackets) of alternative specific covariates for Giant, Big, Medium and Small subsamples.

	McFadden's Choice Model							
	Giant		Big		Medium		Small	
	OR	$\sigma$	OR	$\sigma$	OR	$\sigma$	OR	$\sigma$
Grade Ratio	0.636***	(0.019)	0.461***	(0.015)	1.199	(0.050)	0.648***	(0.070)
Dropout Ratio	1.295***	(0.017)	1.063***	(0.013)	1.199***	(0.022)	0.965	(0.044)
Average Fees	1.001***	(0.000)	0.999***	(0.000)	1.001*	(0.000)	0.962***	(0.001)
$\Delta$ Unemp. Rate	0.974***	(0.001)	0.977***	(0.001)	0.974***	(0.001)	0.999	(0.003)
Distance	0.999***	(0.000)	0.999***	(0.000)	0.999***	(0.000)	0.999***	(0.000)
<i>Case-Specific Controls</i>	<i>Yes</i>		<i>Yes</i>		<i>Yes</i>		<i>Yes</i>	
N. Observations	2,048,494		1,722,385		812,629		85,494	

Looking at the results on average fees, an interesting result emerge. First, let us notice that in Table 2 Giant and Medium universities exhibit means much higher with regard to Big and Medium ones. Then, in Table 3 average fees have a positive significant effect for Giant and Medium, and negatively significant for the other two. The possible explanation is that in the Italian framework a double competition appears. Namely, Giant universities compete versus Big ones, while Medium compete versus Small. Apparently, as in the common expression 'the big fish eats the small', bigger universities can impose higher fees on students with regard to their smaller respective competitors, without losing their competitiveness. From this point of view, a larger (w.r.t. the respective competitor) university has a strong competitive advantage, considering that the great majority of them are settled in the Centre/North of Italy. Universities with a competitive disadvantage need to lower their fees in order to attract more students, and they will be further penalized by trying to improve their appeal softening their own grading policies.

Table 4, even suffering for an important loss of information, provides useful hints through inserting more controls about universities' reputation and cities' characteristics. In particular, it seems that Big universities needs lower fees and higher grades in order to compete with the Giant ones. On the other side, grading policies lose almost completely their effect for Medium and Small universities for which is much more important to be settled in the cheapest cities. On the other side, the cost of life does not matter for Giant and Big, probably because they are already settled in the most expensive cities, which are also the most attractive ones. If significant, the Student-Teacher Ratio has a positive effect, probably because it is driven by its numerator: students prefer to apply in universities where they can find a larger number of their peers. In general, the hypothesis of a double competition by size turns up to be reinforced for 'Giant vs Big' and weakened for 'Medium vs Small'.

**Table 4** Robustness check: Odds Ratios and Standard deviation (between brackets) of alternative specific covariates for Giant, Big, Medium and Small subsamples.

	McFadden's Choice Model							
	Giant		Big		Medium		Small	
	OR	$\sigma$	OR	$\sigma$	OR	$\sigma$	OR	$\sigma$
Grade Ratio	0.655***	(0.058)	3.160***	(0.502)	0.868	(0.101)	0.713	(0.200)
Dropout Ratio	1.268***	(0.048)	1.769***	(0.124)	1.059	(0.064)	0.539***	(0.073)
Average Fees	1.005***	(0.000)	0.937***	(0.001)	1.013***	(0.005)	1.011***	(0.002)
$\Delta$ Unemp. Rate	0.972***	(0.004)	0.950***	(0.006)	0.968***	(0.006)	1.007	(0.008)
Distance	0.999***	(0.000)	0.999***	(0.000)	0.999***	(0.000)	0.999***	(0.000)
CENSIS Score	0.993*	(0.004)	0.990***	(0.004)	1.031***	(0.007)	1.007	(0.017)
Student-Teacher Ratio	1.001	(0.001)	1.020***	(0.002)	1.039***	(0.003)	1.021	(0.023)
Cost of Life	1.031	(0.027)	0.944	(0.038)	0.748***	(0.008)	0.450***	(0.055)
<i>Case-Specific Controls</i>	<i>Yes</i>		<i>Yes</i>		<i>Yes</i>		<i>Yes</i>	
N. Observations	233,682		128,854		86,283		18,630	

As in the main literature, the solution to the problems of competitiveness suffered by several Italian universities has to be searched in a policy intervention for mitigating the gap among the two areas of the country. At the same time, interconnections between the Higher Education system and the local areas have to be reinforced, so reducing migrations aimed to reach the healthiest job markets in advantage.

## References

1. Chan, W., Hao, L., Suen, W.: A signaling theory of grade inflation. *International Economic Review*, **45(9)**, 1065–1090 (2007)
2. Columbu, S., Porcu, M., Sulis, I.: University choice and the attractiveness of the study area: Insights on the differences amongst degree programmes in Italy based on generalised mixed-effect models, *Socio-Economic Planning Sciences*, 100926 (2020)
3. De Paola, M.: Are easy grading practices induced by low demand? Evidence from Italy. University Library of Munich, Germany. (2008)
4. D'Agostino, A., Ghellini, G., Longobardi, S.: Out-migration of university enrolment: the mobility behaviour of Italian students, *International Journal of Manpower*, **40(1)**:56–72 (2019)
5. De Angelis, I., Mariani, V., Modena, F., Montanaro, P.: Immatricolazioni, percorsi accademici e mobilità degli studenti italiani (academic enrolment, careers and student mobility in Italy), Bank of Italy Occasional Paper, **354** (2016)
6. Lombardi, G., Ghellini, G.: The effect of grading policies on Italian Universities' attractiveness: A Conditional Multinomial Logit approach. *Electronic Journal of Applied Statistical Analysis*, **12(4)**, 801–825 (2019)
7. McFadden, D. L.: Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka, 105–142. New York: Academic Press (1974)
8. Viesti, G.: La laurea negata: le politiche contro l'istruzione universitaria. Gius. Laterza Figli Spa. (2018)

## **The class A journals and the Italian academic research outcomes in Statistical Sciences**

### ***La lista delle riviste di classe A e i risultati della ricerca accademica in Scienze Statistiche***

Maria Maddalena Barbieri, Francesca Bassi, Antonio Irpino, Rosanna Verde

**Abstract** Since before its first introduction in 2012, the list of journals of “class A” for Statistical Sciences has generated a large debate in the scientific community on its formulation. The list is one of the outcomes of the ranking of journals proposed by Anvur (the Italian National Agency for the Evaluation of Universities and Research Institutes) with the main purpose of calculating the minimum values of the indicators of scientific qualification used in the National Scientific Habilitation (ASN) procedure.

The aim of the paper is to analyze the data currently available on the research outputs of the Italian academic statistical community to check if any of the potentially observed changes may be related to the composition of the rating of journals.

**Abstract** *Fin da prima della sua introduzione nel 2012, la lista delle riviste di “classe A” per le Scienze Statistiche ha generato un ampio dibattito nella comunità scientifica riguardante la sua formulazione. La lista è uno dei prodotti della classificazione delle riviste proposte dall’Anvur (l’Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca) con il fine principale di calcolare i valori minimi degli indicatori di qualificazione scientifica usati nella procedura relativa all’Abilitazione Scientifica Nazionale (ASN).*

*Il lavoro ha l’obiettivo di analizzare i dati attualmente disponibili sulla produzione degli accademici italiani per verificare se i potenziali cambiamenti osservati possano essere messi in relazione con la composizione della lista che classifica le riviste.*

**Key words:** Anvur, Asn, journals rating.

---

<sup>1</sup> Maria Maddalena Barbieri, Università Roma Tre; email: marilena.barbieri@uniorma3.it  
Francesca Bassi, Università di Padova; email: francesca.bassi@unipd.it  
Antonio Irpino, Università della Campania Luigi Vanvitelli; email: antonio.irpino@unicampania.it  
Rosanna Verde, Università della Campania Luigi Vanvitelli; email: rosanna.verde@unicampania.it

## 1 Preliminaries

Starting from 2012, ANVUR (the Italian National Agency for the Evaluation of Universities and Research Institutes) carries out the rating of journals for the scientific fields identified as “not bibliometric”, such as *Economics and Statistics* (Area 13). The rating was first intended for calculating the minimum values of the indicators of scientific qualification used in the National Scientific Habilitation (ASN) procedure both for candidates and for full professors applying for membership in the National Committees which examine applicants. The ASN is a necessary requirement to apply for permanent positions of Full and Associate Professor in Italian Universities. However, the rating has soon been employed for different purposes, e.g. it has become one of the means used by Universities’ Departments to evaluate the outcomes of the research of their members and to define the eligibility criteria to serve in a selection board for competitions for early stage university researchers. More recently, the use of the rating has also been introduced in the accreditation procedure of PhD programs.

The rating procedure has actually two outputs: the list of journals considered with scientific content and the list of top journals, called “class A”. We remind that the first rating was carried out starting from the list of journals where researchers, associate and full professors working in Italian Universities, published in the previous years. The lists were subsequently completed also after a consultation with the Scientific Societies. In 2016, a new procedure to classify the top journals in the area *Economics and Statistics* was introduced and the class A list for the sub-area of Statistical Sciences was consequently emended. This procedure was only based on the choice of a certain number of ASJC (All Science Journal Classification) in the Scopus database referred to the years 1999-2005, followed by the selection of a top percentage of journals for each ASJC. The values of the percentage of picks were specific for different ASJC. Both the lists of “scientific” and “class A” journals were occasionally modified, using criteria which also change from time to time. However, the main core of the current lists is still the one issued in 2016.

It is worth mentioning that, as part of the evaluation of universities research quality (VQR) exercise, carried out by Anvur every five years starting from 2004, in each of the first two exercises the Group of evaluation experts (GEV) for the area of Economic and Statistics compiled their own journal list, following different strategies. In both occasions the resulting lists were different from the concurrent ASN journal rating.

Since before its first introduction, the rating of journals for Statistical Sciences has generated a large debate in the scientific community related both to the structure of the resulting lists and to their sensible uses. Proofs of the intense and still actual activities on this topic are all the actions undertaken by the Italian Statistical Society, whose tracks may be found on the web page of the Society, and the number of papers written on this subject. Main contributions to the discussion were given by Frosini (2008), Carpita (2014), Cocchi (2014), Petrucci (2014) along with the authors of the other articles appeared in two special issues of *Statistica & Società*, published in 2008 and in 2014.

The purpose of this paper is to contribute to the discussion on this topic analyzing the data regarding the publications on journal by the members of the academic staff

The class A journals and the Italian academic research outcomes in Statistical Sciences working in Statistical Sciences in the Italian Universities to check how the habit of publication changed during the last ten years, taking also into account the “class A” list of journals and the criteria used to compose and maintain it.

## 2 The output of research of the Italian academic statistical community in the last decade

Our analysis is based on publicly available data. The first set of information used is the list of members of the academic staff working in Statistical Sciences in the Italian Universities in the time interval 2016-2020. The list was obtained as the union of the lists of the faculty members on December 31 of each year from 2015 to 2019, downloaded from the Italian Ministry of University and Research repository<sup>1</sup>. The consistency of the population takes into account also the members who retired and the new entries on the payrolls of the Italian Universities during the observation period. In order to have information on the research outputs published on journals we resorted to the Scopus dataset, since it is considered the base for the rating of journals for Statistical Sciences carried out by Anvur. For each serial title in the Scopus database, we considered the number of papers published in the years 2010-2020 having as author at least one of the members of the previously referred list. To identify each of them we used the corresponding Scopus author ID, an identifier assigned automatically to every author of at least an article in the index of Scopus. Only in a small number of cases the search was unsuccessful, due to the absence of a Scopus ID. Table 1 contains a first brief summary of the overall population.

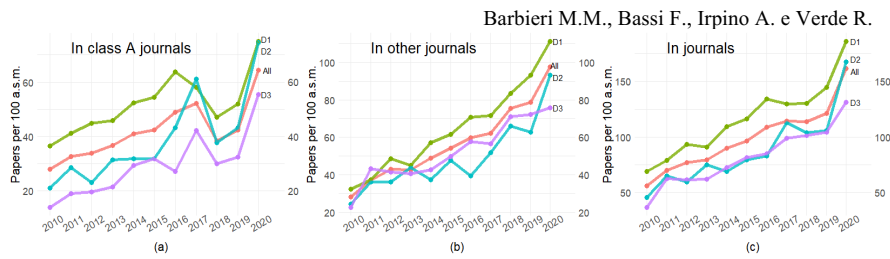
**Table 1:** Distribution of the Italian universities academic staff members in the years 2016-2020 by competition sector (*Settore Concorsuale* – SC) and the availability of a Scopus ID.

<i>SC</i>	<i>With Scopus ID</i>	<i>Without Scopus ID</i>	<i>Total</i>
13/D1 - Statistica	519 (98.3%)	9 (1.7%)	528 (100.0%)
13/D2 – Statistica economica	173 (93.5%)	12 (6.5%)	185 (100.0%)
13/D3 – Demografia e Statistica sociale	162 (93.6%)	11 (6.4%)	173 (100.0%)
<b>Total</b>	<b>854 (96.4%)</b>	<b>32 (3.6%)</b>	<b>886 (100.0%)</b>

Our dataset shows an increase in the amount of research outputs present in the Scopus database over the observation period. Figure 1 depicts the evolution of the number of papers published in journals, expressed per 100 academic staff members, for each competition sector in the field of Statistical Science and overall. Part of the increasing positive trend in time and the resulting different levels of the yearly counts before and after the year 2016 may be explained to some extent noticing that our dataset does not include the publications of those who retired before December 31 2015 and that were not coauthored with a member of our population (i.e. the academic staff in the years 2016-2020). In addition, some of the new entries in the population could have been not yet active in the interval 2010-2015 or at least in the first part of it.

<sup>1</sup> <https://cercauniversita.cineca.it/php5/docenti/cerca.php>





**Figure 1:** Number of papers published expressed per 100 academic staff members: panel (a), (b) and (c) refer to papers published in class A journals, in journals which are not in class A and in the whole Scopus database, respectively. Please note that the vertical scales are different.

From Figure 1, we note that the lines corresponding to each competition sector share a long range common trend, although they also exhibit occasional different behaviors. While the amount of papers appeared on serial titles not classified in class A (represented in panel (b)) exhibits a growth in the whole interval, the number of papers published in class A journals (represented in panel (a)) shows a decrease in the years 2017 and 2018, followed by a quick recovery. To some degree this behavior may be due to the exclusion, announced in 2016, of a large number of journals from the class A list, starting from the issues published in 2018. Some of those journals were quite popular until then. In view of these exclusions, the submissions by Italian authors may have moved to other journals.

In order to check if the increasing trend in the number of publications is common to all areas of research or if some research fields contributed more than others in the changes, we refer to the classification of journals by ASJC. The ASJC scheme is a journal classification produced by Scopus, based on the aims and scope of the journals and on the content they publish. The classification is not bijective in the sense that each journal may be associated to more than one ASJC. We resorted to the Scopus ASJCs since this classification is the base for the procedure fine-tuned by the working group appointed by Anvur to compose the “class A” list of top journals in the area of *Economics and Statistics* (Anvur, 2016). The starting point of this procedure was the choice of a certain number of ASJCs in the Scopus database referred to the years 1999-2005, followed by the selection of a top percentage of journals specific for each ASJC. As a consequence, all journals in the class A list are inside at least one of the selected ASJC, apart on a small number of exceptions represented by journals added to the list during the occasional subsequent revisions. The set of ASJCs used to compile the class A journal list was our starting point.

For each competition sector we selected the top ten elements after sorting the ASJCs into a decreasing order of the total number of papers published in journals classified in each ASJC in the time period 2016-2020 (denoted with  $T_{16-20}$ ). For each ASJC we considered also the number of papers published in class A journals (denoted with  $A_{16-20}$ ), the number of papers published on journals which are not in class A (denoted with  $NA_{16-20}$ , where  $NA_{16-20} = T_{16-20} - A_{16-20}$ ), the number of class A journals involved (denoted with  $N_{16-20}$ ) and, in addition, we computed the Gini index to measure the inequality in the distribution of the amount of papers over class A journals (denoted with  $G_{16-20}$ ). Then we computed the difference in the number of papers published in the time intervals 2016-2020 and 2010-2015 with respect to all journals, to class A journals and to journal which are not in class A. We denoted the resulting

The class A journals and the Italian academic research outcomes in Statistical Sciences differences with  $\Delta T$ ,  $\Delta A$  and  $\Delta NA$ , respectively. Results are reported in Table 2 for the competition sector 13/D1, in Table 3 for the competition sector 13/D2 and in Table 4 for the competition sector 13/D3. The tables contain also the percentages over T, next to the number of products classified as A or NA, and the percentage chances, next to the changes over the two time periods, all in round brackets.

**Table 2:** 13/D1 - Top ten ASJCs with respect to the total number of papers published in 2016-2020

<i>ASJC</i>	<i>N</i> <sub>16-20</sub>	<i>G</i> <sub>16-20</sub>	<i>A</i> <sub>16-20</sub>	<i>NA</i> <sub>16-20</sub>	<i>T</i> <sub>16-20</sub>	$\Delta A$	$\Delta NA$	$\Delta T$
<i>Statistics and Probability</i>	66	0.55	924(69%)	406(31%)	1330(100%)	12(1%)	2(0%)	14(1%)
<i>Statistics, Probability and Uncertainty</i>	44	0.52	629(76%)	202(24%)	831(100%)	45(8%)	39(24%)	84(11%)
<i>Applied Mathematics</i>	26	0.60	281(79%)	75(21%)	356(100%)	-19(-6%)	16(27%)	-3(-1%)
<i>Social Sciences (miscellaneous)</i>	14	0.70	191(75%)	65(25%)	256(100%)	84(79%)	51(364%)	135(112%)
<i>Economics and Econometrics</i>	33	0.58	114(54%)	97(46%)	211(100%)	42(58%)	61(169%)	103(95%)
<i>Modelling and Simulation</i>	15	0.56	120(58%)	87(42%)	207(100%)	11(10%)	-9(-9%)	2(1%)
<i>Computer Science Applications</i>	13	0.59	98(58%)	72(42%)	170(100%)	9(10%)	56(350%)	65(62%)
<i>Management Science and Operations Research</i>	17	0.59	101(76%)	32(24%)	133(100%)	60(146%)	9(39%)	69(108%)
<i>Multidisciplinary</i>	5	0.56	66(61%)	42(39%)	108(100%)	10(18%)	38(950%)	48(80%)
<i>Geography, Planning and Development</i>	12	0.60	44(42%)	61(58%)	105(100%)	33(300%)	38(165%)	71(209%)

**Table 3:** 13/D2 - Top ten ASJC with respect to the total number of papers published in 2016-2020

<i>ASJC</i>	<i>N</i> <sub>16-20</sub>	<i>G</i> <sub>16-20</sub>	<i>A</i> <sub>16-20</sub>	<i>NA</i> <sub>16-20</sub>	<i>T</i> <sub>16-20</sub>	$\Delta A$	$\Delta NA$	$\Delta T$
<i>Economics and Econometrics</i>	40	0.43	127(51%)	121(49%)	248(100%)	9(8%)	17(16%)	26(12%)
<i>Statistics and Probability</i>	31	0.47	99(55%)	81(45%)	180(100%)	14(16%)	30(59%)	44(32%)
<i>Social Sciences (miscellaneous)</i>	13	0.70	103(65%)	56(35%)	159(100%)	49(91%)	20(56%)	69(77%)
<i>Geography, Planning and Development</i>	14	0.44	54(38%)	90(62%)	144(100%)	38(238%)	52(137%)	90(167%)
<i>Statistics, Probability and Uncertainty</i>	25	0.47	86(74%)	30(26%)	116(100%)	35(69%)	5(20%)	40(53%)
<i>Management, Monitoring, Policy and Law</i>	8	0.51	28(35%)	52(65%)	80(100%)	15(115%)	36(225%)	51(176%)
<i>Strategy and Management</i>	13	0.54	44(66%)	23(34%)	67(100%)	33(300%)	0(0%)	33(97%)
<i>Economics, Econometrics and Finance (miscellaneous)</i>	11	0.36	29(45%)	35(55%)	64(100%)	10(53%)	-3(-8%)	7(12%)
<i>Finance</i>	10	0.35	27(50%)	27(50%)	54(100%)	-2(-7%)	-3(-10%)	-5(-8%)
<i>Business and International Management</i>	6	0.52	21(47%)	24(53%)	45(100%)	3(17%)	8(50%)	11(32%)

**Table 4:** 13/D3 - Top ten ASJC with respect to the total number of papers published in 2016-2020

<i>ASJC</i>	<i>N</i> <sub>16-20</sub>	<i>G</i> <sub>16-20</sub>	<i>A</i> <sub>16-20</sub>	<i>NA</i> <sub>16-20</sub>	<i>T</i> <sub>16-20</sub>	$\Delta A$	$\Delta NA$	$\Delta T$
<i>Demography</i>	12	0.51	85(54%)	73(46%)	158(100%)	8(10%)	16(28%)	24(18%)
<i>Social Sciences (miscellaneous)</i>	10	0.76	108(69%)	48(31%)	156(100%)	64(145%)	23(92%)	87(126%)
<i>Statistics and Probability</i>	14	0.47	45(52%)	42(48%)	87(100%)	2(5%)	16(62%)	18(26%)
<i>Geography, Planning and Development</i>	10	0.50	38(53%)	34(47%)	72(100%)	32(533%)	18(113%)	50(227%)
<i>Public Health, Environmental and Occupational Health</i>	4	0.19	9(15%)	50(85%)	59(100%)	-16(-64%)	18(56%)	2(4%)
<i>Statistics, Probability and Uncertainty</i>	12	0.43	36(65%)	19(35%)	55(100%)	13(57%)	10(111%)	23(72%)
<i>Economics and Econometrics</i>	12	0.47	27(51%)	26(49%)	53(100%)	10(59%)	7(37%)	17(47%)
<i>Strategy and Management</i>	5	0.51	22(69%)	10(31%)	32(100%)	20(1000%)	5(100%)	25(357%)
<i>Multidisciplinary</i>	4	0.30	19(61%)	12(39%)	31(100%)	6(46%)	11(1100%)	17(121%)
<i>Health (social science)</i>	4	0.10	10(32%)	21(68%)	31(100%)	-1(-9%)	12(133%)	11(55%)

With regard to the set of the most popular ASJCs, although the competition sectors share a common core, they also show differences in the composition of the aggregates and in the proportion of papers in class A journals in each ASJC, with 13/D1 usually showing larger values.

While the order of magnitude of frequencies referred to 13/D2 and 13/D3 is in most cases too small to allow comparisons, in 13/D1 data seem to show a shift of interest towards research fields different from those usually considered traditional: the

ASJCs with the highest relative changes in the overall number of papers published are *Social Sciences (miscellaneous)*, *Economics and Econometrics*, *Management Science and Operations Research* and *Geography, Planning and Development*. The latter has the largest relative change also in 13/D2 and 13/D3. We also note that *Social Sciences (miscellaneous)* is the common leader in the increase of papers published in class A journals.

The Gini concentration index allows us to highlight the presence of class A journals hosting a number of papers sensibly higher than the others. Starting with 13/D1, we note that in some ASJC the outlier is an usual publication setting for the Italian academic statisticians, while other situations refer to journal whose popularity greatly increased only in recent years. Example of the former case are: *Journal of Computation and Simulation* (where the papers published in 2010-2015 and 2016-2020 were 62 and 46, respectively) and *Statistical Methods and Applications* (62 and 46), both in the ASJC *Statistics, Probability and Uncertainty*, the latter also in *Statistics and Probability*; *Journal of Statistical computation and simulation* (24 and 31) in the ASJC *Modelling and Simulation*; *Journal of Applied Stochastic Models in Business and Industry* (19 and 23) in the ASJC *Management Science and Operations Research*; *Advances in Data Analysis and Classification* (37 and 42) in the ASJCs *Applied Mathematics* and *Computer Science Applications*. While we may classify in the second category *Socio-Economic Planning Sciences* (in the ASJCs *Statistics, Probability and Uncertainty, Economics and Econometrics, Management Science and Operations Research* and *Geography, Planning and Development*), whose transition to 22 papers in 2016-2020, from none in the previous period, is mainly due to the publication of special issues, and *Social indicator research* (in the ASJC *Social Sciences (miscellaneous)*) that hosted 96 papers in 2016-2020 and only 7 in 2010-2015. In 13/D2 and 13/D3 the papers are more scattered between journals and few serial titles hosted more than twenty papers in both time intervals. In 13/D3 this is the case of *Journal of Demographic Research* (in the ASJC *Demography*), passed from 29 papers in 2010-2015 to 32 in 2016-2020. While *Social indicator research* has this feature in both 13/D2 and 13/D3 since this journal passed from 13 papers in 2010-2015 to 60 in 2016-2020 in the former and from 12 to 77 in the latter.

## References

1. Anvur: Revisione generale delle riviste di classe A per l'area 13. Relazione di accompagnamento, a cura del Gruppo di Lavoro su Riviste e Pubblicazioni Scientifiche per l'Area 13 (Silvia Fedeli, Marco LiCalzi, Michelangelo Vasta, Enrico Zaninotto) (2016).  
<https://www.anvur.it/wp-content/uploads/2017/09/GdLArea13RelazioneAccompa~.pdf>
2. Carpita, M.: Valutazione della Ricerca e Classifiche delle Riviste Scientifiche nelle Scienze Statistiche: l'esperienza SIS, *Statistica & Società*, anno III, n. 3. 19-26 (2014).
3. Cocchi, D.: Gli statistici nell'area 13 italiana e nei settori ERC, *Statistica & Società*, anno III, n. 3. 38-40 (2014).
4. Frosini, B.V.: Valutazione della ricerca e valutazione delle riviste scientifiche in ambito statistico. *Statistica & Società*, anno VI n. speciale, 39-47(2008).
5. Petrucci, A.: Il riconoscimento della scientificità delle riviste: l'esperienza del CUN attraverso la Consultazione Pubblica, *Statistica & Società*, anno III, n. 3. 5-9 (2014).

## 4.31 Statistical methods for finance

# Hypotheses testing in mixed–frequency volatility models: a bootstrap approach

## *Test d'ipotesi nei modelli di volatilità a frequenza mista: un approccio bootstrap*

Vincenzo Candila, Lea Petrella

**Abstract** It is widely recognized that standard likelihood–based inference suffers from the presence of nuisance parameters. This problem is particularly relevant in the context of Mixing–Data Sampling (MIDAS) models, when volatility forecasting is the research topic and where often covariates' data are sampled at a different (usually lower) frequency than the asset returns. In this framework, testing the significance of the MIDAS terms brings together the presence of nuisance parameters that under the null hypothesis are not identifiable. This circumstance interferes with the asymptotic distribution of the common statistical tests employed in this framework. In particular, the asymptotic distribution is no more a  $\chi^2$  distribution. The present paper proposes a bootstrap likelihood ratio (BLR) test to overcome this problem, simulating the likelihood ratio test distribution. Using a Monte Carlo experiment, the proposed BLR test presents quite good performances in terms of the test's size and power.

**Abstract** *E' ampiamente riconosciuto che gli approcci inferenziali basati sulla massima verosimiglianza soffrono della presenza di nuisance parameters. Questo problema è particolarmente rilevante nel contesto di modelli Mixing–Data Sampling (MIDAS), usati nell'ambito delle previsioni di volatilità. In questo framework, testare la significatività dei termini MIDAS comporta la gestione dei nuisance parameters che, sotto l'ipotesi nulla, sono non identificabili. Questa circostanza interferisce con la distribuzione asintotica dei test statistici comunemente usati in questo ambito. In particolare, la distribuzione asintotica non risulta più essere una  $\chi^2$ . Il presente lavoro propone un bootstrap likelihood ratio (BLR) test per superare questo problema, simulando la distribuzione del likelihood ratio test. Attraverso una simulazione Monte Carlo, il test BLR proposto presenta ottime performance, in termini di size e potenza.*

**Key words:** Likelihood ratio test, MIDAS, nuisance parameter, bootstrap.

---

Vincenzo Candila, MEMOTEF Department, Sapienza University of Rome, Italy, e-mail: vincenzo.candila@uniroma1.it and Lea Petrella, MEMOTEF Department, Sapienza University of Rome, Italy, e-mail: lea.petrella@uniroma1.it

## 1 Introduction

The financial econometrics literature has paid particular attention to the estimation of asset returns volatility during the last four decades. In this framework, empirical evidences suggest that the volatility has a slow-moving feature around which the conditional second moments of returns oscillate. Starting from this characteristic, a new type of volatility models, based on the decomposition of volatility into two components, namely a short and a long-run component, has been proposed (for more details, see the review of Amado et al., 2019). At the same time, it is quite common in financial data analysis that observations came at a different frequency (usually lower) than the returns' ones. The Mixing-Data Sampling (MIDAS) methods proposed by Ghysels et al. (2007) are designed to solve this problem. When the MIDAS techniques are applied within the GARCH framework, the long-run component of the models can depend on variables observed at different frequencies than daily (see, for example, Engle et al. (2013) and Conrad and Kleen (2020)). Recently, the MIDAS methods have also been applied in the quantile regression framework to forecast the Value-at-Risk (Candila et al., 2020).

Unfortunately, as stated in Ghysels et al. (2007), testing the null hypothesis of no influence of the MIDAS component can be problematic since the weights associated with each realization of the low-frequency variable, seen as nuisance parameters, are not identifiable. This circumstance has a fundamental impact on the asymptotic distribution of the commonly used tests, like the Wald or the Likelihood Ratio (LR) tests (see Hansen (1996) and Andrews (2001) for a complete survey on this topic).

In the context MIDAS variables within a volatility model, our paper aims at investigating the profitability of using a bootstrap LR (BLR) test where the distribution of the test is obtained using a suitable bootstrap procedure. Resorting to the bootstrap to derive the LR test distribution is not new at all: see, for instance, the contributions of Di Sanzo (2009) and Buseti and Di Sanzo (2012). But this is the first time the BLR test is used within the volatility models employing MIDAS components.

In terms of results, the size and power of the proposed BLR are calculated through an extensive Monte Carlo experiment in a GARCH model framework. Comparing the results with the standard LR test, the BLR appears to have an empirical size closer to the nominal one and quite good empirical power.

The rest of the paper is as follows: Section 2 illustrates the models and the proposed BLR test, while Section 3 presents the Monte Carlo experiment.

## 2 Bootstrap Likelihood Ratio test

Let  $r_{i,t}$  be the log-return of an asset representing the first log-difference of the closing prices for the day  $i$  in the period (week or month)  $t$ . Then, let us consider the formalization of the GARCH-MIDAS model proposed by Engle et al. (2013):

Hypotheses testing in mixed-frequency volatility models: a bootstrap approach

$$r_{i,t} = \sigma_{i,t} \varepsilon_{i,t} = \sqrt{\tau_t \times g_{i,t}} \varepsilon_{i,t}, \quad \text{with } i = 1, \dots, N_t \quad \text{and } t = 1, \dots, T, \quad (1)$$

where,  $\sigma_{i,t}$  representing the conditional standard deviation at day  $i$  and period  $t$ , consists of two (multiplicative) components:  $\tau_t$  and  $g_{i,t}$ . In particular,  $\tau_t$  is defined as the long-run component of the volatility at period  $t$  and  $g_{i,t}$  the short-run term at day  $i$  for period  $t$ . Moreover,  $N = \sum_{t=1}^T N_t$  is the total number of days considered with  $N_t$  being the number of days in the period  $t$ . In Eq. (1),  $\varepsilon_{i,t}$  is the *iid* innovation term, with  $E(\varepsilon_{i,t}) = 0$  and  $E(\varepsilon_{i,t}^2) = 1$ , and with a finite fourth moment.

Following the common dynamics specifications of the short- and the long-run components proposed in the GARCH-MIDAS literature, we consider for  $g_{i,t}$  the unit-mean reverting GJR-GARCH(1,1) process given by:

$$g_{i,t} = (1 - \alpha - \gamma/2 - \beta) + \left( \alpha + \gamma \cdot \mathbb{1}_{(r_{i-1,t} < 0)} \right) \frac{(r_{i-1,t})^2}{\tau_t} + \beta g_{i-1,t}, \quad (2)$$

where  $\mathbb{1}_{(\cdot)}$  is an indicator function and  $\alpha > 0$ ;  $\beta \geq 0$ ;  $\gamma \geq 0$ ;  $\alpha + \beta + \gamma/2 < 1$ .

The component  $\tau_t$  is:

$$\tau_t = \exp \left( m + \theta \sum_{k=1}^K \delta_k(\omega) MV_{t-k} \right), \quad (3)$$

where  $m \in R$ ,  $\theta \in R$  represents the response to the one-sided filter of the past  $K$  realizations of the MIDAS terms i.e. the low-frequency variable  $MV_t$  through the weighting function  $\delta_k(\omega)$ . The most common used  $\delta_k(\omega)$  in this context is the Beta function:

$$\delta_k(\omega) = \frac{(k/K)^{\omega_1-1} (1-k/K)^{\omega_2-1}}{\sum_{j=1}^K (j/K)^{\omega_1-1} (1-j/K)^{\omega_2-1}}. \quad (4)$$

Under this configuration, the parameter space is then  $\Theta = \{\alpha, \gamma, \beta, m, \theta, \omega_1, \omega_2\}$ . Given  $K$  and a distributional assumption for  $\varepsilon_{i,t}$  in (1) it is possible to calculate the maximum likelihood (ML) estimator for  $\Theta$ .

In order to test the significance of the MIDAS component in (3), the following null hypothesis is considered:

$$H_0 : \theta = 0. \quad (5)$$

Typically, one can evaluate such a null using the a Wald or a LR test. We focus on this latter case. Let  $\hat{\Theta}_0$  be the ML estimate of  $\Theta$  under the null  $\theta = 0$ , that is the “restricted” model. The correspondent log-likelihood at  $\hat{\Theta}_0$  is denoted by  $\ell(\hat{\Theta}_0)$ . Let  $\hat{\Theta}$  be the ML estimate of  $\Theta$  under the alternative  $\theta \neq 0$  i.e. in the “unrestricted” model. The corresponding log-likelihood at  $\hat{\Theta}$  is denoted by  $\ell(\hat{\Theta})$ . The LR test is:

$$LR = 2 \left[ \ell(\hat{\Theta}) - \ell(\hat{\Theta}_0) \right]. \quad (6)$$

Assuming a significance level  $\alpha$ , test statistic in (6) should reject  $H_0$  when

$$LR > CV_\alpha, \tag{7}$$

where  $CV_\alpha$  is the  $(1 - \alpha)th$  quantile of the LR distribution under the null. Under some regularity conditions, it can be shown that the LR test follows asymptotically a Chi-square ( $\chi^2$ ) distribution. In our context, since under the null hypothesis in (5) the parameters  $\omega_1$  and  $\omega_2$  in (4) are not identified, the distribution of  $LR$  in (6) is no more a  $\chi^2$  distribution. For this reason, here we propose a bootstrap procedure to simulate the distribution of  $LR$  test (6) under the null (5). The proposed BLR procedure is as follows:

1. Estimate the unrestricted and restricted models. Compute the LR statistic as in Eq. (6).
2. Let  $\widehat{\sigma}_{i,t}$  be the estimated volatility obtained from the restricted model. Compute the standardized residuals  $\widehat{\varepsilon}_{i,t}$  under the null, for  $i = 1, \dots, N_t$  and  $t = 1, \dots, T$ , that is:

$$\widehat{\varepsilon}_{i,t} = \frac{r_{i,t}}{\widehat{\sigma}_{i,t}}.$$

Let  $\widehat{\varepsilon}_{i,t}^*$ , be the bootstrap residual, obtained from resampling with replacement from the standardized residual series  $\widehat{\varepsilon}_{i,t}$ .

3. Compute the bootstrap replicates of  $r_{i,t}$ , denoted by  $r_{i,t}^*$ , through:

$$r_{i,t}^* = \widehat{\sigma}_{i,t}^* \widehat{\varepsilon}_{i,t}^*, \quad \text{for } i = 1, \dots, N_t \quad \text{and } t = 1, \dots, T,$$

where  $\widehat{\sigma}_{i,t}^*$  is the bootstrap volatility:

$$\widehat{\sigma}_{i,t}^* = \sqrt{\widehat{\tau}_t^* \times \widehat{g}_{i,t}^*},$$

with the long-run term under the null identified as  $\widehat{\tau}_t^* = \exp(\widehat{m})$  and the short-run term as

$$\widehat{g}_{i,t}^* = \left(1 - \widehat{\alpha} - \widehat{\gamma}/2 - \widehat{\beta}\right) + \left(\widehat{\alpha} + \widehat{\gamma} \cdot \mathbb{1}_{(r_{i-1,t}^* < 0)}\right) \frac{(r_{i-1,t}^*)^2}{\widehat{\tau}_t^*} + \widehat{\beta} \widehat{g}_{i-1,t}^*,$$

where  $\widehat{\alpha}, \widehat{\gamma}, \widehat{\beta}$  and  $\widehat{m}$  are the ML estimates of the restricted model. In order to obtain (recursively) the bootstrap realizations of  $r_{i,t}^*$  for the  $N$  days, start the procedure with  $\widehat{\sigma}_{1,t}^* = \widehat{\sigma}_{1,t}$ . Finally, estimate the restricted and unrestricted models on the series  $r_{i,t}^*$ . Hence, calculate the LR on the bootstrap returns  $r_{i,t}^*$ , denoted by  $LR^*$ .

4. Repeat the previous step  $B$  times, obtaining  $(LR^{*(1)}, \dots, LR^{*(B)})$ , which is the bootstrap distribution of LR.
5. The estimate of (the bootstrap)  $CV_\alpha$ , based on  $(LR^{*(1)}, \dots, LR^{*(B)})$  and labelled as  $\widehat{CV}_\alpha$ , is obtained as the  $1 - \alpha$  quantile of the bootstrap distribution of LR.

Finally, the null in (5) is rejected through the BLR test if  $LR > \widehat{CV}_\alpha$ .



### 3 Monte Carlo Experiment

In this section, we consider a Monte Carlo experiment to learn about the profitability of using the BLR test when testing MIDAS components. For this goal, we generate  $R$  samples of data from the following data generating process (DGP):

$$r_{i,t} = \sqrt{\tau_t} \times g_{i,t} \varepsilon_{i,t}, \quad \text{with } i = 1, \dots, N_t, \quad \text{and } t = 1, \dots, T, \quad (8)$$

where:

$$\varepsilon_{i,t} \sim \text{iid } t_{(7)}, \quad (9)$$

$$\tau_t = \exp \left( m_0 + \theta_0 \sum_{k=1}^K \delta_k(\omega) MV_{t-k} \right), \quad (10)$$

$$g_{i,t} = (1 - \alpha_0 - \gamma_0/2 - \beta_0) + \left( \alpha_0 + \gamma_0 \cdot \mathbb{1}_{(r_{i-1,t} < 0)} \right) \frac{(r_{i-1,t})^2}{\tau_t} + \beta_0 g_{i-1,t}. \quad (11)$$

In Eq. (9), the error term  $\varepsilon_{i,t}$  follows a standardized Student's  $t$  distribution with 7 degrees of freedom which allows for fat tails of real financial asset returns. We assume that the simulated stationary variable  $MV_t$  follows an AR(1):  $MV_t = \phi MV_{t-1} + e_t$ , with  $\phi = 0.7$ .

Using the R package *rumidas* (Candila, 2021), the DGP in (9) is simulated  $R = 250$  times, according to two sample sizes:  $N = \{500, 1000\}$ . The true values of the parameters are:

$$\{\alpha_0 = 0.01, \gamma_0 = 0.1, \beta_0 = 0.9, m_0 = -1, \omega_{2,0} = 1.1\}.$$

The parameter of interest  $\theta$  has instead the following values:  $\theta_0 = \{0, 0.5, 1\}$ .

The results of our experiment are illustrated in Table 1, where the estimated probabilities of rejecting the null across the  $R$  replicates are reported. More in detail, Panel A shows the empirical sizes for the BLR and the LR tests that is the occurrence of null rejection when the null is true. The results of the LR test are evaluated using the  $\chi^2$  distribution. Independently of the significance level adopted (0.01, 0.05, and 0.1) and of the sample length, the empirical size of the BLR appears much more in line with the actual size. When the null is false, as in Panels B and C, both the tests appear to have reasonable powers. These results support the use of the proposed BLR test instead of the LR one when mixed frequency models are employed.

### References

Amado, C., A. Silvennoinen, and T. Teräsvirta (2019). Models with multiplicative decomposition of conditional variances and correlations. In J. Chevallier, S. Goutte, D. Guerreiro, S. Saglio, and B. Sanhaji (Eds.), *Financial Mathematics, Volatility and Covariance Modelling*, Volume 2, pp. 217–260. London: Routledge.

**Table 1** BLR and LR empirical sizes and powers

Signif. level	0.01	0.05	0.10	0.01	0.05	0.10
Panel A: $\theta = 0$		$N = 500$			$N = 1000$	
BLR	0.012	0.056	0.136	0.016	0.040	0.108
LR	0.060	0.132	0.200	0.060	0.160	0.220
-----						
Panel B: $\theta = 0.5$		$N = 500$			$N = 1000$	
BLR	0.980	0.980	1.000	0.980	1.000	1.000
LR	1.000	1.000	1.000	1.000	1.000	1.000
-----						
Panel C: $\theta = 1$		$N = 500$			$N = 1000$	
BLR	0.980	1.000	1.000	1.000	1.000	1.000
LR	1.000	1.000	1.000	1.000	1.000	1.000

**Notes:** Panel A reports the empirical sizes of the BLR and LR tests, for the null in (5), that is the number of times (across the  $R = 250$  replications) that the null is rejected (given that the null is true). The other panels report the empirical powers, that is the number of times (across the  $R = 250$  replications) that the null is rejected (given that the null is false).

- Andrews, D. W. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica* 69(3), 683–734.
- Busetti, F. and S. Di Sanzo (2012). Bootstrap LR tests of stationarity, common trends and cointegration. *Journal of Statistical Computation and Simulation* 82(9), 1343–1355.
- Candila, V. (2021). *rumidas: Univariate GARCH-MIDAS, Double-Asymmetric GARCH-MIDAS and MEM-MIDAS models*. R package version 0.1.1.
- Candila, V., G. M. Gallo, and L. Petrella (2020). Using mixed-frequency and realized measures in quantile regression. Technical report, SSRN.
- Conrad, C. and O. Kleen (2020). Two are better than one: Volatility forecasting using multiplicative component GARCH-MIDAS models. *Journal of Applied Econometrics* 35(1), 19–45.
- Di Sanzo, S. (2009). Testing for linearity in Markov switching models: a bootstrap approach. *Statistical Methods and Applications* 18(2), 153–168.
- Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics* 95(3), 776–797.
- Ghysels, E., A. Sinko, and R. Valkanov (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews* 26(1), 53–90.
- Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64(2), 413–430.

# Quantile Regression Forest with mixed-frequency data

## *Random Forest Quantilica per dati a frequenza mista*

Mila Andreani, Vincenzo Candila, Lea Petrella

**Abstract** Recent contributions to the scientific literature have exploited Random Forests and mixed-frequency data to perform quantile regression, a popular technique largely used to forecast Value-at-Risk (VaR) when financial data are concerned. However, the potential of merging these two approaches in one model has not yet been investigated. Therefore, in this paper we propose to combine the Quantile Regression Forests (QRF) approach with Mixed Data Sampling (MIDAS) one building the MIDAS-QRF in order to forecast daily VaR measures through the additional information coming from low-frequency covariates. The empirical results show that the MIDAS-QRF delivers consistent VaR predictions and outperforms popular existing models.

**Abstract** *Recenti contributi in letteratura hanno applicato le Random Forests e i dati a frequenza mista alla regressione quantile, una tecnica ampiamente utilizzata per la previsione del Value-at-Risk (VaR) applicata a dati finanziari. Tuttavia, il potenziale di modelli in grado di unire questi due approcci non è stato ancora studiato. In questo articolo la Quantile Regression Forest (QRF) e i modelli Mixed Data Sampling (MIDAS) verranno combinati per costruire un modello MIDAS-QRF al fine di prevedere il VaR tenendo conto delle informazioni aggiuntive provenienti da variabili osservate a frequenza diversa rispetto alla variabile dipendente. I risultati empirici mostrano che il MIDAS-QRF fornisce previsioni del VaR ottimali e supera i comuni modelli usati in questi contesti in termini di accuratezza previsionale.*

**Key words:** Random Forest, Quantile Regression, MIDAS term, Value-at-Risk

---

Mila Andreani  
Scuola Normale Superiore, Pisa, Italy, e-mail: mila.andreani@sns.it

Vincenzo Candila  
MEMOTEF Depart., Sapienza University of Rome, Italy, e-mail: vincenzo.candila@uniroma1.it

Lea Petrella  
MEMOTEF Depart., Sapienza University of Rome, Italy, e-mail: lea.petrella@uniroma1.it

## 1 Introduction

The recent financial crises have highlighted the role of Value-at-Risk (VaR) as the first-choice market risk measure, computed as the quantile of the conditional distribution of assets returns. Given its relevance, several econometric models have been developed in order to improve VaR predictions. The most popular one is the quantile regression technique proposed in Koenker and Bassett (1978), aimed at offering a more complete picture of the conditional distribution of the response variable with respect to standard linear regression. In this sense, quantile regression results to be particularly useful in modelling the tails of conditional distributions, and, consequently, for computing VaR.

During the last years, the statistical literature has proposed different models aimed at extending the quantile regression approach. In this direction, accordingly with the Random Forest framework considered in Breiman (2001), Meinshausen (2006) proposed to extend this approach in a quantile regression framework building the Quantile Regression Forest (QRF) models. This model is particularly appealing since it mitigates the parametric assumptions presented in the quantile regression modelling. When working with financial data, it is quite common to handle variables observed at different frequency (often lower) than the dependent one i.e the financial returns. In this context, several contributions to the financial literature have focussed on the application of the mixed-data sampling model (MIDAS) methods proposed by Ghysels et al. (2004). The MIDAS approach allows the inclusion of mixed-frequency variable(s) to models where the dependent variable is observed usually at daily frequency. Recently, the mixed frequency methods have been also extended in the quantile regression framework by Candila et al. (2020), where the final goal was to estimate the VaR risk measure.

In this paper, we propose to combine the QRF approach and the MIDAS technique. In particular, we develop the MIDAS-Quantile Regression Forest (MIDAS-QRF) model, aimed at computing conditional quantiles via QRF with the additional MIDAS components used to train the QRF.

The benefits of the proposed method are twofold: on one hand, differently from the linear quantile regression model, it allows to detect non-linear relations among variables without making any particular parametric assumption. On the other hand, it allows to consider one or more covariates sampled at different frequencies without using a, a feature that has never been considered in machine learning algorithms. From an empirical point of view, we show how our approach outperforms existing models in forecasting daily VaRs of the S&P 500 index.

The rest of the paper is organised as follows: Section 2 concerns the methodology used to develop our model and in Section 3 considers the empirical application of the proposed MIDAS-QRF to evaluate the VaR risk measure.

## 2 Methodology

In this section we describe the methodology applied to develop the MIDAS-QRF. Our model takes inspiration from Ghysels et al. (2004) and Candila et al. (2020) for the MIDAS approach and from Meinshausen (2006) for the QRF one.

Let  $Y_{i,t}$  be a high frequency response variable sampled at time  $i$  of the  $t$ -th period (for instance, a daily variable sampled in the  $t$ -th month), where  $i = 1, \dots, T$ . We define  $Z_t$  as the generic low-frequency covariate observed in the  $t$ -th time of the year. The simplest MIDAS model is specified as follows:

$$Y_{i,t} = \beta_0 + \beta_1 \sum_{j=1}^K \phi_k(\omega) Z_{t-j} + \varepsilon_{i,t} \quad (1)$$

where  $MC_{i-1,t} = \sum_{j=1}^K \phi_k(\omega) Z_{t-j}$  is the MIDAS component, obtained by considering the last  $K$  observations of the low frequency covariate up to time  $i-1$  of the  $t$ -th period. As can be seen, the  $MC$  term, is characterised by the weighting function  $\phi(\omega)$ , which can be defined by several specifications, such as for example the Beta function, as proposed in Candila et al. (2020):

$$\phi_k(\omega) = \frac{(k/K)^{\omega_1-1} (1-k/K)^{\omega_2-1}}{\sum_{j=1}^K (j/K)^{\omega_1-1} (1-j/K)^{\omega_2-1}}. \quad (2)$$

This function allows to impute a greater weight to more recent observations by setting  $\omega_1 = 1$  and  $\omega_2 > 1$ . Thus, the only parameter to be estimated in (2) is  $\omega_2$ . As mentioned in the introduction, the object of the present work is to use QRF approach in order to calculate the VaR of assets' returns when low-frequency variables are observed.

As described in Breiman (2001), Random Forests are an ensemble machine learning algorithm for classification and regression tasks operating by constructing a multitude of decision trees during the training phase. In a regression context, this model computes the conditional expected value of the response variable by averaging the predictions of each individual tree. Moreover, when the interest is in the quantile regression framework, the natural extension of Random Forests proposed by Meinshausen (2006) estimates the whole conditional distribution and then computes the quantile at level  $\tau \in [0, 1]$ .

In this paper, in order to exploit information coming from variables sampled at different frequencies than the dependent one, we introduce the low-frequency component  $MC_{i-1,t}$  in the QRF model. The resulting MIDAS-QRF is defined as follows. Let  $\{(Y_{i,t}, \mathbf{X}_{i,t})\}_{i=1}^T \in \mathbb{R} \times \mathbb{R}^P$  the sample of *i.i.d* random variables drawn from the unknown joint distribution of the random variables  $(Y, \mathbf{X})$  where  $\mathbf{X}_{i,t} = (X_{i,t}^1, \dots, X_{i,t}^P)'$  is the  $P$ -vector at time  $i$  of covariates sampled at the same frequency of  $Y_{i,t}$ . with realisation  $\{(y_{i,t}, \mathbf{x}_{i,t})\}_{i=1}^T$ .

Moreover, we define with  $\mathbf{Z}_t = (Z_t^1, \dots, Z_t^N)'$  the vector of  $N$  low frequency variables at time  $t$ . The main idea of our approach is to use both the variables  $\{(Y_{i,t}, \mathbf{X}_{i,t})\}_{i=1}^T$  and the  $MC_{i-1,t}$  ones of each low frequency variables  $\mathbf{Z}_t$  as additional covariates to train the MIDAS-QRF using an expanding window approach. More specifically, the MIDAS-QRF is trained using the first  $n$  observations, with  $n < T$ , to predict the  $m$ -steps ahead VaR measures. Afterwards, the  $m$  subsequent observations are added to the training sample and the whole set of observations is used to re-estimate the MIDAS-QRF. Then, the subsequent  $m$ -steps ahead quantile is computed. This procedure is recursively performed until the end of the series. The resulting vector of quantile forecasts are used to compute the accuracy of the model in terms of quantile loss as in González-Rivera et al. (2004). In order to find the optimal set  $\omega_2^*$  used to compute the MIDAS components of the  $N$  low-frequency covariates, a grid search value is performed by training a different MIDAS-QRF at each iteration. The optimal  $\omega_2^*$  is identified as the one delivering the lowest quantile.

### 3 Empirical Application

In this section we show the performances of our model on real data. In order to evaluate the VaR forecasting procedure using the proposed MIDAS-QRF model, we consider the S&P 500 index (SPX) from March 2002 to February 2020, for a total of 4558 daily log returns of  $Y_{i,t}$  observations. The covariates observed at the same daily frequency are the  $Y_{i-1,t}$  and the log-difference of VIX index considered at time  $i - 1, t$ . We also consider as low-frequency variables the Industrial Production (INDPRO) and the Consumer Price Index (CPI) observed at monthly time to be included in the MIDAS part of the model. Daily data and monthly data have been collected from both the OxfordMan Institute and from the Federal Reserve Economic Data (FRED). Data are differentiated when necessary. The model is trained on a dataset of 4253 observations and for computational reasons we chose  $m = 50$  to predict VaRs at level  $\tau = 0.05$ . In order to evaluate the performances of the MIDAS-QRF we compare the results with the ones obtained using the following models and trained using the same covariates: standard Quantile Regression Forest, linear Quantile Regression (QR), GARCH with skew Normal distribution (G-sm), GARCH with skew Student's t-distribution (G-st), GARCH-MIDAS and Double Asymmetric GARCH-MIDAS with INDPRO (GM-INDPRO and DAGM-INDPRO) as low-frequency variable, GARCH-MIDAS and Double Asymmetric GARCH-MIDAS with CPI as low-frequency variable (GM-CPI and DAGM-CPI) Engle et al. (2013); Amendola et al. (2019).

In order to verify the validity of our model, the VaR forecasts are evaluated by means of three backtesting procedures: the unconditional (UC) and conditional (CC) coverage tests Kupiec (1995); Christoffersen (1998) and the Actual Exceedances ratio (AE) where the AE ratio represents the actual number of VaR violations over the expected one. The closer to 100% the ratio is, the better the model performances. Moreover, we evaluate the performances of the models in terms of quantile loss of

González-Rivera et al. (2004) and by computing the increase of accuracy delivered by our model with respect to the competing models. Let  $j$  be a competing model, then the increase of accuracy with respect to  $j$ ,  $ACC_j$ , is:

$$ACC_j = 100 \cdot \left( 1 - \frac{Loss}{Loss_j} \right), \quad (3)$$

where  $Loss$  represents the quantile loss obtained from the MIDAS-QRF and  $Loss_j$  from the model  $j$ . The results for VaR at 5% level are presented in Table 1,

**Table 1:** Backtesting for VaR at 5% and accuracy

$\tau = 0.05$	MIDAS-QRF	QRF	QR	G-snorm	G-sstd	GM-INDPRO	GM-CPI	DAGM-INDPRO	DAGM-CPI
<b>AE</b>	98.04%	39.21%	71.89%	117.65%	144.79%	98.04%	104.57%	98.04%	111.11%
<b>UC</b>	0.937	0.006*	0.236	0.490	0.098	0.937	0.855	0.937	0.661
<b>CC</b>	0.426	0.019*	0.347	0.538	0.240	0.459	0.968	0.459	0.907
<b>Loss</b>	8.415	8.948	10.359	8.886	9.013	9.368	9.325	9.342	9.336
<b>ACC<sub>j</sub></b>	-	5.96%	18.77%	5.31%	6.64%	9.93%	10.18%	9.93%	9.87%

where the percentages of the AE are reported together with the p-values for the UC and the CC test. Values marked by a star denote the rejection of the test at level 5%. We also report the value of the quantile loss for each model and the increase of accuracy obtained by comparing the MIDAS-QRF with another model.

The empirical results highlight that the MIDAS-QRF model outperforms the existing ones in forecasting daily VaRs at 5% level. In particular, the backtesting procedures show that the MIDAS-QRF model delivers adequate VaR forecasts and achieves the highest AE ratio, equal to 98%.

In terms of predictive accuracy, the MIDAS-QRF model produces the lowest quantile loss. In particular, the increase of accuracy obtained by comparing the MIDAS-QRF with another model ranges from 5.96% to 18.77%.

In conclusion, the proposed model is the first attempt to include in a Quantile Random Forest variables observed at different frequencies without making any particular parametric assumption. From an empirical point of view, the MIDAS-QRF outperforms existing models in terms of prediction accuracy at 5% VaR level for the S&P 500 index. These results highlight the importance of information coming from low-frequency variables, which contribute to successfully modelling the tail risk of financial variables.

## References

- Amendola, A., V. Candila, and G. M. Gallo (2019). On the asymmetric impact of macro-variables on volatility. *Economic Modelling* 76, 135–152.
- Breiman, L. (2001). Random forests. *Machine learning* 45(1), 5–32.
- Candila, V., G. M. Gallo, and L. Petrella (2020). Using mixed-frequency and realized measures in quantile regression. *Available at SSRN 3722927*.
- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International economic review*, 841–862.
- Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics* 95(3), 776–797.
- Ghysels, E., P. Santa-Clara, and R. Valkanov (2004). The midas touch: Mixed data sampling regression models.
- González-Rivera, G., T.-H. Lee, and S. Mishra (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of forecasting* 20(4), 629–645.
- Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica*, 33–50.
- Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *The J. of Derivatives* 3(2).
- Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research* 7(Jun), 983–999.



# Higher order moments in Capital Asset Pricing Model betas

## *Il Capital Asset Pricing Model: momenti di ordine superiore nella stima del beta*

Giuseppe Arbia, Riccardo Bramante and Silvia Facchinetti

**Abstract** The traditional theory of Capital Asset Pricing Model uses a Least Square linear regression strategy to evaluate the systematic risk of an asset. In this context the consequences of non-normality are particularly relevant and may affect dramatically the investors' decisions. In this paper we propose a new regression interpolation criterion, the Least Quartic criterion, which provides an evaluation of market risk by taking into account also third and fourth moments characteristics in non-normal situations. We apply the proposed procedure to the top 300 market capitalization components of the STOXX Europe 600.

**Abstract** *La teoria tradizionale del Capital Asset Pricing Model utilizza una regressione lineare ai Minimi Quadrati per la valutazione del rischio sistematico di un asset. In questo contesto le conseguenze della non-normalità sono particolarmente rilevanti e possono influenzare notevolmente le decisioni degli investitori. In questo articolo proponiamo un nuovo criterio di interpolazione, il criterio dei Minimi Quartici, che fornisce una valutazione del rischio di mercato che tiene conto anche dei momenti del terzo e quarto ordine tipici in situazioni di non-normalità. La procedura proposta viene applicata alle prime 300 società per capitalizzazione di mercato incluse nell'indice STOXX Europe 600.*

**Key words:** Least Quartic criterion, non-gaussian distribution, systematic risk

---

Giuseppe Arbia

Department of Statistical sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1 - Milano,  
e-mail: giuseppe.arbia@unicatt.it

Riccardo Bramante

Department of Statistical sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1 - Milano,  
e-mail: riccardo.bramante@unicatt.it

Silvia Facchinetti

Department of Statistical sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1 - Milano,  
e-mail: silvia.facchinetti@unicatt.it

## 1 Introduction

The Capital Asset Pricing Model (CAPM) was introduced in the early 1960s independently by W. Sharpe, J. Lintner and J. Mossin [16, 17, 19]. This model assumes that investors construct their portfolio on the basis of a trade-off between the expected return and the variance of the returns of the market portfolio. Typically, it uses a Least Squares (LS) regression to measure the relationship between returns of an asset in relation to the market, thus providing a measure of the so-called systematic risk. In fact, the regression slope reflects the exposure of an asset to such risk, indicating how fluctuations in the returns are related to the market movements.

According to CAPM, let us consider the simple linear regression model where the  $i$ -th asset return, for  $i = 1, 2, \dots, N$ , is modeled by<sup>1</sup>

$$r_{i,t} = b_i \cdot r_{M,t} + \varepsilon_{i,t} \quad (1)$$

where  $b_i$  is the slope coefficient, while  $r_{i,t}$  and  $r_{M,t}$  are the historical returns – in excess to the risk free – of the  $i$ -th asset and the market, respectively, at a given point in time  $t \in T$ .

We remark that in this analysis, only the first two moments of the joint distribution between the asset and the market are relevant. Many authors have recognized the shortcomings associated with such an approach, and highlight that the use of a pricing model limited to the first two moments may be misleading and may wrongly indicate insufficient compensation for the investment [12, 18]. A higher-moment approach is more appropriate to detect non-linear relationships between assets and portfolio returns while accommodating for the specific risk–return payoffs. For this reason, financial literature is very rich in contributions that include considerations related to the higher moments, see [3, 5, 13, 15], among others. Our proposal is located in this area.

## 2 The Least Quartic Criterion

The Least Quartic (LQ) criterion is an optimization procedure that represents an extension of the ordinary LS strategy to provide a closed form for the slope regression coefficient estimator for situations where the phenomenon is characterized by strong non-Gaussian distribution (outliers, multimodality, skewness and kurtosis) [2]. An economic-theoretical motivation for the choice of a least quartic criterion may be found in the papers [7, 10].

Assume that the return of the market in model (1) is non-stochastic and that the error term obeys some non-normal distribution characterized by excess kurtosis. We define a quartic loss function that can be seen as a particular case of the general mul-

---

<sup>1</sup> For the sake of simplicity we consider the intercept of the model  $a_i$  equal to zero. The results are also valid for the general CAPM model  $r_{i,t} = a_i + b_i \cdot r_{M,t} + \varepsilon_{i,t}$ , for  $i = 1, 2, \dots, N$ .

tivariate loss function proposed by [1] to model skewness, fat tails, non-ellipticity and tail dependence of financial data:

$$l(b_i) = \sum_t \varepsilon_{i,t}^4 = \sum_t (r_{i,t} - b_i \cdot r_{M,t})^4 = \mu_{4,0} b_i^4 - 4\mu_{3,1} b_i^3 + 6\mu_{2,2} b_i^2 - 4\mu_{1,3} b_i + \mu_{0,4} \quad (2)$$

where,  $\mu_{4,0}$  and  $\mu_{0,4}$  represent the kurtosis of the market and the  $i$ -th asset returns, whereas,  $\mu_{3,1}$ ,  $\mu_{1,3}$  and  $\mu_{2,2}$  represent the measures of co-kurtosis<sup>2</sup> [14].

The LQ criterion is based on the minimization of the loss function in (2), by setting to zero its first derivative. This leads to the Least Quartic estimator of the CAPM regression slope:

$$b_{i,LQ} = \frac{\mu_{3,1}}{\mu_{4,0}} - \left[ \frac{\sqrt[3]{2c_1}}{3\mu_{4,0} \left( c_2 + \sqrt{4c_1^3 + c_2^2} \right)^{1/3}} \right] + \frac{(c_2 + 4c_1^3 + c_2^2)^{1/3}}{3\sqrt[3]{2}\mu_{4,0}} \quad (3)$$

where

$$c_1 = 9\mu_{3,1}^2 + 9\mu_{2,2}\mu_{4,0}$$

$$c_2 = 54\mu_{3,1}^2 - 81\mu_{2,2}\mu_{3,1}\mu_{4,0} + 27\mu_{1,3}\mu_{4,0}^2$$

The second-order condition ensure us that the real solution reported in Equation (3) is a minimum.

In case of normal distribution of stock returns, the obtained estimator reduces to the ordinary LS solution, while in non-normal cases it outperform the ordinary LS estimators in terms of out-of-sample risk-adjusted performance, as we will show in the simulation study.

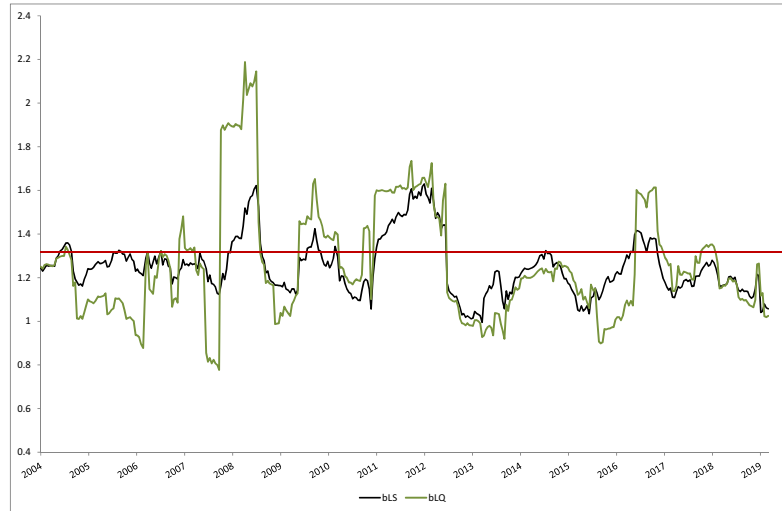
In order to compare the LS and LQ estimators, we refer to the top 300 market capitalization components of the STOXX Europe 600, covering the period from January 2004 to December 2019. The distribution of the 300 considered daily returns series and the benchmark one, deviate from the normal distribution, and there is a marked prevalence of negative skew and of positive excess kurtosis. This confirms the need of a higher-moment approach, like the proposed LQ criterion, which is expected to be more appropriate to detect non-linear relationships.

To test the significance of the proposed estimator, we employ the Monte Carlo procedure setting the shape parameter equal to that observed for the STOXX Europe 600 index over the whole sample period. The results are estimates significantly different from zero with less than 5% significance level.

Figure 1 reports an example of the slope coefficients estimates evaluated using the traditional CAPM expression based on the Least Squares method ( $b_{LS}$ ) and on

<sup>2</sup> The co-kurtosis of a bivariate distribution is defined by the mixed moments of orders  $r$  and  $s$   $\mu_{r,s} = E\{[X - E(X)]^r [Y - E(Y)]^s\}$  such that  $r + s = 4$ .

the alternative Least Quartic technique ( $b_{LQ}$ ) for the same sample asset, together with the  $\beta$  global estimate (red line) that is close to 1.3.



**Fig. 1** Comparison between LS and LQ slope coefficients estimates for a sample asset

Looking at the two patterns showed in Figure 1, it emerges that the perception of market risk exposure is best captured by the LQ estimator during market turmoil (see as an example the 2007–2008 financial crisis and the 2012 Sovereign Debt Crisis). The behaviour is typical of most of the 300 assets examined.

To compare empirically the LS and LQ estimators, we assume an investment strategy that passively follows the market, and we consider a set of 1000 simulated portfolios each containing a random sub-set of  $N$  stocks out of the top 300 components of the STOXX Europe 600<sup>3</sup>, with a historical period of length  $T - M$  ( $M$  is the data point where the out-of-sample analysis starts). All the random portfolios are formed with randomly assigned weights given some predefined constraints.

Computing the out-of-sample mean return  $\hat{r}$ , standard deviation  $\hat{\sigma}_r$ , negative semi deviation  $\hat{\sigma}_r^-$ , Sharpe ratio  $\hat{SR}$  and correlation with the benchmark STOXX Europe 600 Index  $\rho_{\hat{r},M}$ , we observe that, on average, the LQ optimization strategies outperform the LS ones with consistently higher mean returns, lower variability and higher values both of the Sharpe ratio and of the correlation with the STOXX Europe 600 index. Moreover, the raw frequency counts reported in Table 1 show values

<sup>3</sup> For the purpose of this empirical investigation  $N = 30$ , e.g. 10% of the considered stocks.

that are better in the LQ framework for each indicator, providing a heuristic indication of LQ relative advantage since benefits from lower risk are achieved while not reducing returns.

**Table 1** Percentages of cases where the LQ estimates are better than the LS ones.

Indicator	$\hat{\beta}$	$\hat{\sigma}_r$	$\hat{\sigma}_r^-$	$\hat{SR}$	$\rho_{\hat{\beta},M}$
%	60	74	72	55	58

Finally, our empirical results shows that the risk estimation using the Sharpe ratio based on the LQ estimator is better than the one obtained using the LS procedure in most of the analyzed time series. Since many financial time series incorporate the leptokurtic and asymmetric features we can argue that a better measurement of risk can be obtained if in the regression coefficient estimates we also take into account third and fourth moments of returns.

### 3 Conclusions and future developments

In this paper we presented a new criterion to estimate a linear regression model parameter, based on the minimization of the fourth power of the regression errors, to evaluate market risk within the CAPM framework by taking into account third and fourth moments characteristics of the asset price distribution. The potential of the method is illustrated with reference to a case study focused on the top 300 market capitalization components of the STOXX Europe 600. The empirical analysis, based on the Least Quartic estimation of the slope coefficient, adds insights in market analysis and helps in identifying more precisely potentially risky assets whose extreme behavior is strongly dependent on the market behavior.

A number of generalizations could be taken into consideration in the future to expand the results presented here. A first generalization would involve to extend our approach to conditional, rather than unconditional, moments (see, e.g. [4, 8] for recent examples) and to dynamic conditional joint moments (see [9]). A second extension would involve the notion of risk-neutral moments introduced by [6]. Extending our approach to both conditional and risk-neutral moments certainly represents an interesting area of development of the present contribution.

Furthermore, a further possible extension goes in the direction of the statistical estimation. In this respect, future studies in this field could be directed towards the analysis of regressions estimation within a Maximum Likelihood framework. The importance of higher-order considerations in likelihood-based estimation was recently pointed out by [11] who considered fourth-order maximum likelihood techniques based on an exponential family specification of the likelihood function. Following their suggestion one could specify a bivariate joint distribution, say  $f_{(r_i, r_M)}(r_i, r_M)$  in the CAPM application considered in this paper, by using a bivari-

ate Pearson's curve or a bivariate exponential family curve or a mixture distribution. The full likelihood could then be derived as the product of the bivariate marginals  $l(\theta) = \prod_{i=1}^n f_{(r_i, r_M)}(r_i, r_M)$  ( $\theta$  being a set of parameters). This approach would lead to the estimation of parameters that express more thoroughly the complex relationships of dependence between each asset and the market thus representing alternative measurements of the corresponding market risk.

## References

1. Alp, T., Demetrescu, M.: Joint forecasts of Dow Jones stocks under general multivariate loss function. *Computational Statistics and Data Analysis* **54**, 2360–2371 (2010)
2. Arbia, G., Bramante, R., Facchinetti, S.: Least Quartic Regression Criterion to Evaluate Systematic Risk in the Presence of Co-Skewness and Co-Kurtosis. *Risks* **8**, 1–14 (2020)
3. Bakshi, G., Kapadia, N., Madan, D.: Stock return characteristics, skew laws, and the differential pricing individual equity options. *Review of Financial Studies* **16**, 101–43 (2003)
4. Brooks, C., Burke, S.P., Heravi, S., Persaud, G.: Autoregressive conditional kurtosis. *Journal of Financial Econometrics* **3**, 399–421 (2005)
5. Byun, S.J., Kim, D.H.: Gambling preference and individual equity option returns. *Journal of Financial Economics* **122**, 155–74 (2016)
6. Conrad, J., Dittmar, R.F., Ghysels, E.: Ex ante skewness and expected stock returns. *Journal of Finance* **68**, 85–124 (2013)
7. Christie-David, R., Chaudhry, M.: Coskewness and cokurtosis in futures markets. *Journal of Empirical Finance* **8**, 55–58 (2001)
8. Dubauskas, G., Teresienė, D.: Autoregressive conditional skewness and kurtosis and Jarque Bera statistics in Lithuanian stock market measurement. *ISSN Engineering Economics* **5**, 19–24 (2005)
9. Engle, R.F.: Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **50**, 987–1007 (1982)
10. Fang, H., Lai, T.Y.: Cokurtosis and capital asset pricing. *The Financial Review* **32**, 293–307 (1997)
11. Holly, A., Monfort, A., Rockinger, M.: Fourth order pseudo maximum likelihood methods. *Journal of Econometrics* **162**, 278–293 (2011)
12. Jondeau, E., Rockinger, M.: Optimal portfolio allocation under higher moments. *European Financial Management* **12**, 29–55 (2006)
13. Jondeau, E., Zhang, Q., Zhu, X.: Average skewness matters. *Journal of Financial Economics* **134**, 29–47 (2019)
14. Kotz, S., Balakrishnan, N., Johnson, N.L.: *Continuous Multivariate Distributions. Models and Applications*. Hoboken: John Wiley and Sons (2000)
15. Kramer, W., Runde, R.: Peaks or tails: what distinguishes financial data? *Empirical Economics* **25**, 665–671 (2000)
16. Lintner, J.: The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics* **47**, 13–37 (1965)
17. Mossin, J.: Equilibrium in a Capital Asset Market. *Econometrica* **35**, 768–83 (1966)
18. Rinaldo, A., Favre, L.: Hedge Fund Performance & Higher-Moment Market Models. *Journal of Alternative Investments* **8**, 37–51 (2005)
19. Sharpe, W.: Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *Journal of Finance* **19**, 425–42 (1964)

# When Does Sentiment Matter in Predicting Cryptocurrency Bubbles?

## *Il ruolo del sentiment nel prevedere bolle nelle criptovalute*

Arianna Agosto and Paolo Pagnottoni

**Abstract** The lack of fundamental values in the cryptocurrency market paves the way for the rise of unprecedented speculative bubble phenomena, which are often associated with alternating phases of investors' fear and greed. We propose to exploit the information derived from a large set of cryptocurrency news to detect and, possibly, anticipate the presence of speculative bubbles in cryptocurrency prices. This is done by means of a Covariate Augmented Dickey-Fuller (CADF) test, which allows us to explicitly account for market sentiment when testing the presence of a unit root in cryptocurrency prices. Our results show that the covariate test statistics diverges significantly from the ADF test statistics in concomitance of price surges, highlighting its ability to foresee speculative bubble occurrences.

**Abstract** *La mancanza di fondamentali nel mercato delle criptovalute apre la strada al verificarsi di fenomeni di bolle speculative senza precedenti. Proponiamo di sfruttare l'informazione di derivante da un grande insieme di news sulle criptovalute per individuare e, possibilmente, anticipare la presenza di bolle speculative nei prezzi delle criptovalute. Operiamo per mezzo del Covariate Augmented Dickey-Fuller (CADF) test, che ci permette di considerare esplicitamente il sentiment di mercato per testare la presenza di una radice unitaria nel prezzo delle criptovalute. I nostri risultati mostrano che la statistica test che include la covariata diverge in modo significativo con l'aumento dei prezzi, sottolineando la sua abilità nel prevedere l'occorrenza di bolle speculative.*

**Key words:** Bitcoin; Cryptocurrencies; Sentiment; Speculative Bubbles

---

Arianna Agosto, Paolo Pagnottoni  
University of Pavia, Via S.Felice 5, 27100, Pavia (PV), e-mail: arianna.agosto@unipv.it, paola.cerchiello@unipv.it, paolo.pagnottoni@unipv.it

## 1 Introduction

Cryptocurrencies were conceived in first place under the advent of Bitcoin (Nakamoto *et al.*, 2008). Since then, research in this area has been highly multidisciplinary. Several authors have dealt with the description and functioning of cryptocurrencies - see Segendorf (2014). Legal concerns that have arisen through cryptocurrencies are discussed in Murphy *et al.* (2015). A growing stream of literature concentrated on studying the presence of speculative bubbles in the cryptocurrency price dynamics - see, for instance, Fry & Cheah (2016) and Corbet & Yarovya (2018) - and their interaction with investor sentiment - see e.g. Chen & Hafner (2019).

Against this background, we propose to exploit the information derived from a large set of cryptocurrency news to detect and foresee the presence of speculative bubble phenomena occurring in the price of four major cryptocurrencies, i.e. Bitcoin, Ethereum, Litecoin and Ripple. This is done by means of the Covariate Augmented Dickey-Fuller (CADF) test, which allows us to take into account for market sentiment in the context of unit root testing in cryptocurrency prices. Our results point to the informativeness of sentiment indicators towards cryptocurrency price dynamics during high-volume news regimes, thus indicating the capability of the CADF approach to detect and anticipate price bubble behaviors.

## 2 Methodology

From an econometric point of view, one of the main research questions related to cryptocurrencies concerns the possible presence of bubbles in their price. An asset bubble is defined in literature as an extreme price acceleration that cannot be driven by the underlying fundamental economic variables (Case & Shiller (2003); Dreger & Zhang (2013)). The end of this phase, often referred to as bubble burst, leads to drastic price drops, causing severe losses to investors. For example, the Bitcoin price, after one year of sharp increase, crashed at the end of 2017 with a loss of nearly 65% with respect to the peak on 5 February 2018. Several recent works provided empirical evidence of the presence of bubbles in the cryptocurrency prices (Fry & Cheah (2016); Corbet & Yarovya (2018)). From a methodological viewpoint, most of them resorted to the right-tailed unit root testing approach based on Augmented Dickey Fuller (Dickey & Fuller (1979)) regression. Indeed, the extremely rapid price increase, to which the definition of financial bubble refers to, can be described by an exponential growth, whose occurrence can be detected through unit root tests. Specifically, we perform a recursive estimation of the Augmented Dickey Fuller (ADF) regression:



$$y_t = \mu + \phi y_{t-1} + \sum_{j=1}^J \psi_j \Delta y_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

where  $y_t$  is the asset price,  $\mu$ ,  $\phi$  and  $\psi$  are estimated through ordinary least squares (OLS), and  $J$  is the maximum number of lags, which is chosen based on some model selection procedure or information criterion. When  $\phi > 1$ , the price grows exponentially.

Inspired by the recursive testing approach of Phillips & Yu (2011), we estimate Model 1 on 100-day rolling windows, in order to timely catch possible changes, from unit root to explosive, of the cryptocurrency price dynamics.

The traditional ADF specification does not consider any potential effect of exogenous covariates, whose inclusion in the regression equation could change the estimated autoregressive dependence and, consequently, the conclusions drawn from the tests. Therefore, in order to consider the role of sentiment and news volume in anticipating the cryptocurrency price dynamics, we recursively repeat, again on 100-day time windows, the Covariate ADF test (CADF) by Hansen (1995), based on the following specification:

$$y_t = \mu + \phi y_{t-1} + \sum_{j=1}^J \psi_j \Delta y_{t-j} + \sum_{k=1}^K \xi_j \Delta x_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \quad (2)$$

where  $x_t$  is a stationary covariate, which, in our application, is a variable reflecting market sentiment or news volume.

### 3 Data and Empirical Findings

To test our proposal, we collect daily closing price of Bitcoin, Ethereum, Litecoin and Ripple. As a market sentiment indicator, we consider a Sentiment Indicator produced by Brain<sup>1</sup> a research company specialized in the production of alternative datasets and in the development of proprietary algorithms for investment strategies on financial markets, which monitors public financial news on cryptocurrencies from about 2000 financial media sources. The sentiment scoring technology is based on a combination of various natural language processing techniques. The sentiment score assigned to each cryptocurrency is a value ranging from -1 (most negative) to +1 (most positive) that is updated with a daily frequency.

We first provide a graphical representation of the daily news volumes for the four selected cryptocurrencies in figure 1. Evidence shows a clear difference in the volume of news across cryptocurrencies. Not surprisingly, Bitcoin is the cryptocurrency showing the largest portion of news during our sample period, followed by Ethereum, Ripple and, lastly, Litecoin. After an initial oscillating phase, we notice a significant raise in the number of news from November 2020

<sup>1</sup> <https://braincompany.co>

onwards. This regards in first place the number of Bitcoin news which, during the month of December, even double with respect to "normal" business periods. The surge in Bitcoin news is then rapidly followed by a raise in the number of news related to the other cryptocurrencies and, particularly, Ethereum.

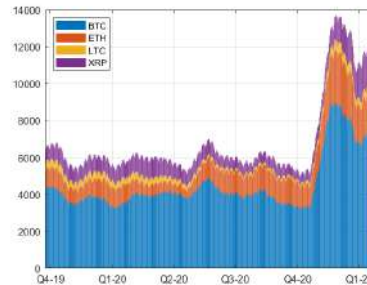


Fig. 1: News volume. The figure shows the dynamics of the daily news volumes for the four selected cryptocurrencies over the period 15 June 2019 - 19 January 2021.

Within our framework, we perform ADF and CADF tests for the selected cryptocurrencies using a rolling window of 100 observations and determine the lag order of the exogenous covariate through the Bayes-Schwarz information criterion. Figure 2 shows the dynamics of the ADF and CADF test statistics along with their difference and the price dynamics of the considered cryptocurrencies. On the one hand, evidence shows that during period of tranquil market dynamics the ADF and CADF test statistics tend to co-move strongly, giving raise to low-magnitude deviations between the two. On the other hand, from the beginning of September onwards we observe that the two test statistics start to diverge and, in most cases, the observed ADF test statistics is larger than the corresponding statistics with additional covariate. Furthermore, notice that the rising difference between the two is associated to a consequent surge in the cryptocurrency prices. This is likely due to the influence of the lagged sentiment indicator, which is able to explain a large portion of the cryptocurrency price variations, and thus lowers the value of the test statistics inducing different outcomes with respect to those observed by the traditional ADF test. This difference arises, in general, well before the cryptocurrency price surge, indicating that the misalignment between the two test statistics can be informative and act as an early-warning indicator for bubble detection purposes. To get a further insight into the strengthening relationship between the sentiment indicator and the cryptocurrency price behaviour, we perform a rolling linear regression exercise, over 100-day windows, where the response variable is the cryptocurrency return and the regressor is the lagged sentiment indicator. Figure 3 (top panels) shows that, starting from September 2020, the estimated coefficients associated to the sentiment indicators grow sharply. It can be noticed from Figure 3 (bottom panels) that, repeating the same

## When Does Sentiment Matter in Predicting Cryptocurrency Bubbles?

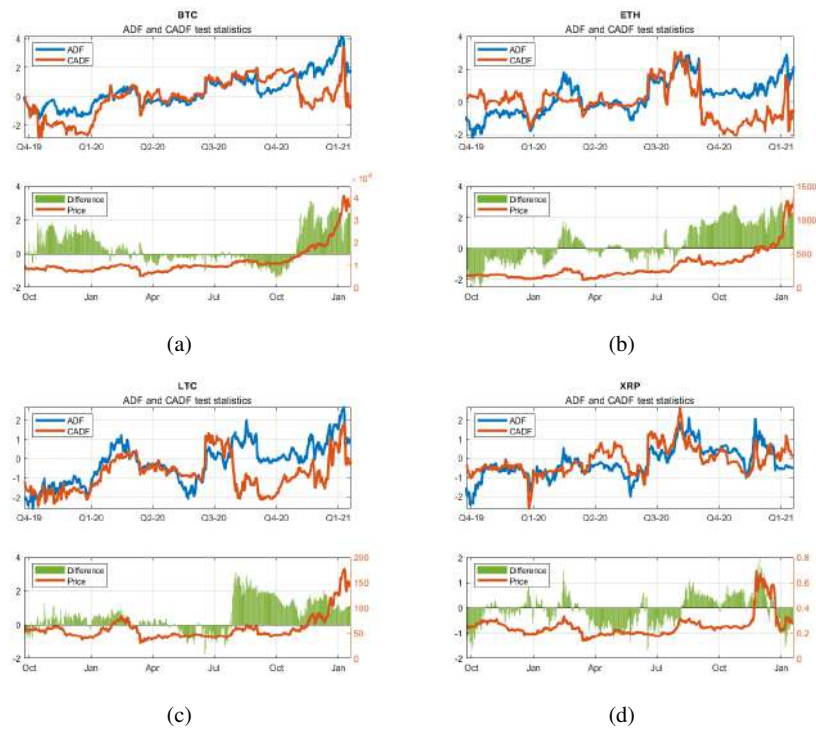


Fig. 2: ADF and CADF test statistics. The figure shows the ADF and CADF test statistics (top panels), their difference and closing price dynamics (bottom panels) of the four cryptocurrencies over the period 23 September 2019 - 19 January 2021.

exercise using lagged Google Search Indices - as a news volume proxy - as a regressor, the estimated coefficients are relatively more flat. Therefore, in the most recent period, the sentiment indicators turn out to be more informative than the news volume in predicting explosive behaviours in cryptocurrency prices.

## References

- CASE, K.E., & SHILLER, R.J. 2003. Is There a Bubble in the Housing Market? *Brookings Papers on Economic Activity*, **2**, 299–362.
- CHEN, CATHY YI-HSUAN, & HAFNER, CHRISTIAN M. 2019. Sentiment-induced bubbles in the cryptocurrency market. *Journal of Risk and Financial Management*, **12**(2), 53.
- CORBET, S., LUCEY-B., & YAROVYA, L. 2018. Datestamping the Bitcoin and Ethereum bubbles. *Finance Research Letters*, **26**, 81–88.
- DICKEY, D. A., & FULLER, W. A. 1979. Distribution of the Estimators for

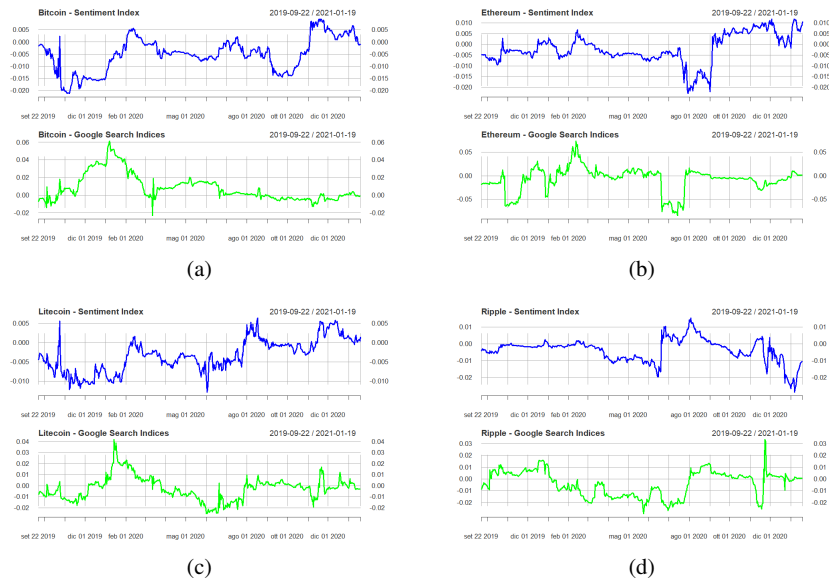


Fig. 3: Coefficients associated to the sentiment/news volume indices. The figure shows the estimated coefficients associated to the scaled lagged sentiment/news volume indices for the selected cryptocurrencies in a rolling linear regression exercise where the response variable is the cryptocurrency return, over the period 23 September 2019 - 19 January 2021.

Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, **74**, 427–431.

DREGER, C., & ZHANG, R.J. 2013. Is there a Bubble in the Chinese Housing Market? *Urban Policy and Research*, **31**(1), 27–39.

FRY, J., & CHEAH, E.T. 2016. Negative bubbles and shocks in cryptocurrency markets. *International Review of Financial Analysis*, **47**, 343–352.

HANSEN, E. B. 1995. Rethinking the Univariate Approach to Unit Root Testing; Using Covariates to Increase Power. *Econometric Theory*, **11**, 1148–1171.

MURPHY, EDWARD, MURPHY, MAUREEN, & SEITZINGER, MICHAEL. 2015. Bitcoin: Questions, answers, and analysis of legal issues. *Congressional Research Service*.

NAKAMOTO, SATOSHI, *et al.* 2008. Bitcoin: A peer-to-peer electronic cash system.

PHILLIPS, P.C.B., WU-Y., & YU, J. 2011. Explosive behavior in the 1990s NASDAQ: when did exuberance escalate asset values. *International Economic Review*, **52**(1), 201–226.

SEGENDORF, BJÖRN. 2014. What is bitcoin. *Sveriges Riksbank Economic Review*, **2**, 71–87.

## 4.32 Statistical methods for high dimensional data

# Virtual biopsy in action: a radiomic-based model for CALI prediction

## *Biopsia virtuale basata su analisi radiomica per la previsione di CALI*

Francesca Ieva, Giulia Baroni, Lara Cavinato, Chiara Masci, Guido Costa, Francesco Fiz, Arturo Chiti, Luca Viganó

**Abstract** Chemotherapy-associated liver injuries (CALI) have a major clinical impact, but their non-invasive diagnosis is still an unmet need. The present work aims at presenting a web-app for personalized risk prediction of developing CALI, elucidating the contribution of radiomic analysis. Patients undergoing liver resection for colorectal metastases after oxaliplatin-based or irinotecan-based chemotherapy between January 2018 and February 2020 were retrospectively analyzed. Radiomic features were extracted from a standardized volume of non-tumoral liver parenchyma. Multivariate logistic regression models and CART were applied to identify predictors and were internally validated. Results show that radiomic analysis of liver parenchyma may provide a signature that, in combination with clinical and laboratory data, improves diagnosis of CALI.

**Abstract** *Le lesioni epatiche associate a chemioterapia (CALI) hanno un impatto clinico molto elevato nella successiva prognosi del paziente, ma la loro diagnosi e previsione non invasiva è ancora oggetto di dibattito. In questo lavoro, viene presentata una web-app per la previsione personalizzata del rischio di sviluppare CALI a seguito di resezione epatica post trattamento chemioterapico. L'analisi di classificazione basata sulle caratteristiche radiomiche e cliniche di tali pazienti mostra incoraggianti risultati su come la biopsia virtuale possa migliorare la diagnosi di CALI.*

**Key words:** Radiomics, Machine Learning, Personalized Medicine, Variable Selection, Virtual Biopsy

---

Francesca Ieva, Giulia Baroni, Lara Cavinato, Chiara Masci  
MOX laboratory, Department of Mathematics, Politecnico di Milano - Milan, Italy  
e-mail: francesca.ieva@polimi.it

Francesco Fiz,  
Department of Nuclear Medicine, Humanitas Clinical and Research Center – IRCCS, Rozzano - Milan, Italy

Arturo Chiti, Guido Costa, Luca Viganó  
Department of Biomedical Sciences, Humanitas University, Pieve Emanuele - Milan, Italy

## 1 Background and motivations

The combination of chemotherapy and surgery is the standard treatment of patients with Colorectal Liver Metastases (CLM) [9]. Systemic chemotherapy prolongs progression-free survival, allows to select the candidates to surgery, and may convert some initially unresectable patients to secondary resectability [3]. However, beside these benefits, preoperative chemotherapy has some drawbacks, first and foremost chemotherapy-associated liver injuries (CALI), namely sinusoidal dilatation, nodular regenerative hyperplasia, and non-alcoholic steatohepatitis (NASH) [6]. CALI and NASH have been associated with the risk of intraoperative bleeding and of postoperative morbidity and liver dysfunction [10], but their preoperative diagnosis is still an unmet need. CALI can be predicted with limited accuracy by some risk factors or by some laboratory tests and scores [8]. Imaging modalities may show some signs of sinusoidal injury. However, these signs are not specific and do not allow a conclusive diagnosis. Even liver biopsy has low effectiveness because of the heterogeneous distribution of injuries and insufficient sample size [11]. In the last few years, a new approach to medical imaging has gained interest. It is driven by the hypothesis that tissue features could be expressed on the radiological images as voxel patterns, that are invisible to the human eye. To identify these patterns, mathematical functions analyzing the spatial relation and the frequency distribution of gray levels in the voxels were developed, providing modern and specific image biomarkers [7]. This texture-based approach has been termed “radiomics”. Texture analysis has shown high accuracy in the identification of liver fibrosis [1], while no study focused on radiomics for CALI. The present analysis aims to introduce a web-app for supporting personalized prediction of NASH and CALI development, elucidating the capability of radiomic features extracted from preoperative computed tomography imaging in patients undergoing liver resection for CLM after preoperative chemotherapy. A defined volume of non-tumoral liver parenchyma was analyzed, thus performing a “virtual biopsy”.

## 2 Methods

All consecutive patients that underwent liver resection for CLM between January 2018 and February 2020 were retrospectively considered. The following inclusion criteria were adopted: preoperative chemotherapy for at least two months; oxaliplatin-based or irinotecan-based chemotherapy regimen; availability of preoperative CT for imaging review and texture analysis; preoperative imaging performed less than two months before liver resection.

A multivariate logistic regression model was performed to estimate the adjusted association between each candidate predictor and the presence of different CALI (grade 2-3 sinusoidal dilatation, NRH or steatohepatitis). Clinical rationale associated with a backward stepwise regression approach was used to retain only relevant associations. In particular: a principal component analysis (PCA) of second order ra-

diomic features, i.e. textural features quantifying tumor heterogeneity by analyzing spatial distribution of pixel/voxel intensities, was performed in order to obtain effective predictors (Fingerprint in the following) for the model. PCA was performed on the following matrices: gray level co-occurrence matrices (GLCM), gray level run length matrices (GLRLM), neighboring gray level difference matrices (NGLDM), and gray level zone-length matrices (GLZLM). For each of them, we retained components of the PCA that explain at least 95% of original features variability. Clinical and laboratory variables were selected according to a priori knowledge; then a stepwise regression was run, and all the variables retained by this procedure were used for predictive purposes. Finally, correlation matrix of continuous variables and correlation heat-map were generated. Correlation between features was analyzed and, whenever higher than 0.85, one of the two features involved in the correlation was removed. The final predictive model underwent internal cross-validation by splitting the series into a training set (90% of the population) and a validation set (10%). The validation procedure was repeated 100 times over 100 different samples. Results are reported in terms of mean (Std Dev) accuracy. A decision tree was built with the variables retained by the backward stepwise selection of the multivariate model, in order to highlight and exploit the possible nonlinear association with the outcome.

R software [4] were used for all the analyses. The work schedule of the entire work is available at <https://giuliabaroni94.shinyapps.io/LiverApp/> in a dynamic user interface shiny app already in use. The web-app is intended to support the advanced analysis of data in the current context, providing a support for personalized predictions of NASH and CALI development for a new patients. Researchers may provide their own dataset, and select among the analytics available.

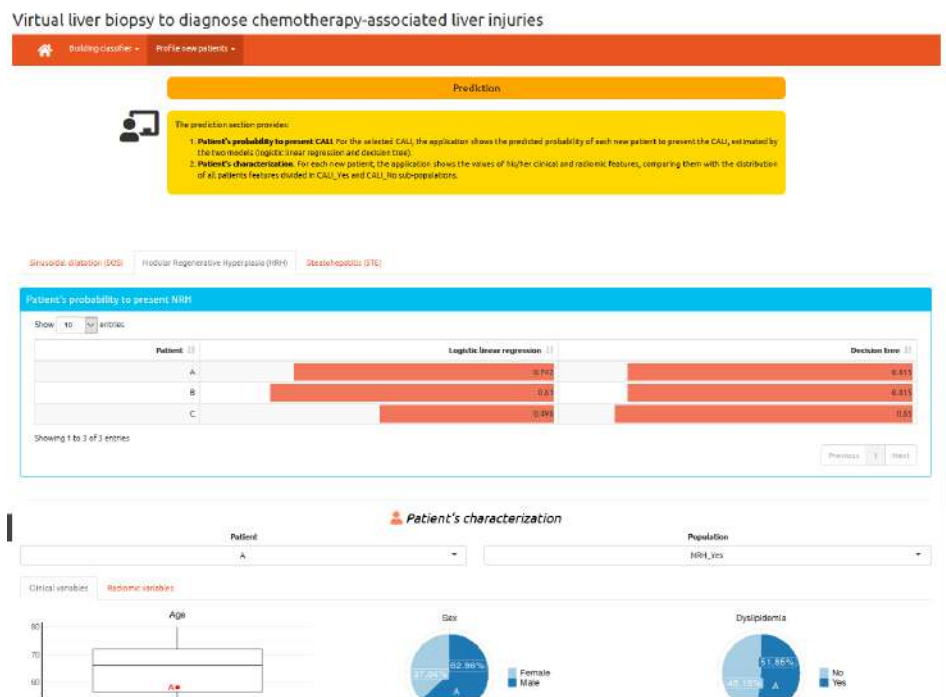
### 3 Results

At the end of the observational period, CALI were evident in 61 (78%) patients. In details, grade 2-3 sinusoidal dilatation was present in 25 (32%) patients, NRH in 27 (35%), and NASH in 14 (18%).

For sake of simplicity, we will report here the results of the NRH multivariate analysis only. Complete results about the study are available on the shiny app.

With respect to the NRH CALI, the following clinical and laboratory variables were associated with NRH: age ( $OR = 1.10$ ,  $CI_{95\%} = 1.01 - 1.20$ ,  $p = 0.027$ ), BMI ( $OR = 0.68$ ,  $CI_{95\%} = 0.49 - 0.94$ ,  $p = 0.021$ ), Irinotecan ( $OR = 28.71$ ,  $CI_{95\%} = 1.8 - 459.04$ ,  $p = 0.018$ ), number of cycles of chemotherapy ( $OR = 1.15$ ,  $CI_{95\%} = 1.01 - 1.32$ ,  $p = 0.031$ ), anti-VEGF therapy associated with chemotherapy ( $OR = 0.05$ ,  $CI_{95\%} = 0.01 - 0.49$ ,  $p = 0.010$ ), and APRI score ( $OR = 275.08$ ,  $CI_{95\%} = 4.75 - 15937.97$ ,  $p = 0.007$ ). In addition, three radiomic predictors of NRH were identified: conventional-HUQ2 ( $OR = 0.76$ ,  $CI_{95\%} = 0.62 - 0.92$ ,  $p = 0.005$ ),  $GLZLM_{f2}$  ( $OR = 0.05$ ,  $CI_{95\%} = 0.01 - 0.43$ ,  $p = 0.007$ ), and  $GLZLM_{f3}$  ( $OR = 7.97$ ,  $CI_{95\%} = 1.52 - 41.85$ ,  $p = 0.014$ ). The combined clinical, laboratory and radiomic model had 85% accuracy, 81% sensitivity, and 86% specificity





**Fig. 1** Example of panel prediction of the Virtual Liver biopsy webapp.

( $AUC = 0.91$ ). The model without radiomic features had  $AUC = 0.85$ . The decision tree based on the results of multivariate logistic regression had the following knots:  $APRI \text{ score} < 0.28$ ,  $BMI \geq 24$ ,  $GLZLM_{f3} < -0.3$ ,  $GLZLM_{f3} \leq 0.5$ ,  $GLCM_{f2} < 0.094$ ,  $GLCM_{f2} < 0.094$ . It achieved 83% accuracy, 89% sensitivity, and 80% specificity ( $AUC = 0.88$ ). In the validation setting, the multivariate logistic regression had an average accuracy of 71% (Std Dev 12%).

Fig. 1 shows the dynamic interface for computing the risk of developing NRH CALI for a new patient entering the study. Such a tool may support clinical decisions about the treatment a patient should undergo.

## 4 Discussion

An accurate non-invasive prediction of NASH and CALI is a relevant unmet need for clinicians. To date, CALI prediction relies on patients' history, i.e. chemotherapy regimen and the number of administered cycles, and on some liver function tests, such as APRI score or ICG test, but they are misleading in up to one-third

of patients. Image mining and analysis have opened new perspectives. Above all, the employment of grey level co-occurrence and higher-order matrices has rooted in clinical research as texture descriptors, although their use has generally turned in automatic feature extraction tools, namely radiomics [12]. Radiomics involves the definition of mathematical features able to capture data about the grey-scale patterns, interpixel relationships, shape, and spectral properties within regions of interest on radiological images. This technique allows researchers to access standardized texture information about images and to carry out informed inference, aiding traditional clinical investigations. To date, radiomics demonstrated a good capability to predict biological characteristics and outcomes of several diseases [2]. However, no studies analyzed the association of textural features with CALI. Radiomics is expected to detect CALI-related tissue heterogeneity and alterations, and our explorative analysis confirmed this hypothesis.

The present study tries to assess the association between virtual biopsy of the non-tumoral liver and CALIs, providing an automatic tool for personalized prediction. The virtual biopsy is highly reproducible. The adoption of software with automatic extraction of radiomic features (LifeX®), jointly with a structured working schedule for the analysis of resulting data increases the potential diffusion of this approach, even if interpretability and explainability of radiomic data are still debated.

We could argue some limitations of present analysis. It is a retrospective study collecting a limited number of patients, but CALI had a standardized and prospective evaluation, and patients were treated in a short period (two years) with homogeneous schedules. Even if the present data are preliminary and need for more robust external validation, the standard predictors of CALI were confirmed together with the new contribution of radiomic signatures. Further, internal validation provided encouraging confirmation of good performances. Finally, the usability of radiomic features remains an issue, even if present proposal concretely shows their potential translation into clinical practice. Radiomics suffers from close-source nature, unharmonized acquisition settings, discordant reconstruction parameters, lack of interpretability, redundancy and methodological bias [5]. A wide and active research area is growing around grey level quantization and pre-processing, aiming at informative rather than descriptive statistics from images. Such studies could open new perspectives in clinical applications of medical imaging analysis.

In conclusion, the present study demonstrated that texture analysis of liver parenchyma might provide a radiomic signature that, in combination with clinical and laboratory data, improves diagnosis of sinusoidal dilatation, NRH and steatohepatitis. Even if the application of radiomics to clinical practice is still to accomplish, our preliminary data can be consistently the basis for an innovative precision medicine approach to patients at risk for liver injuries.

## References

1. Lubner, M.G., Malecki, K., Kloke, J. et al.: Texture analysis of the liver at MDCT for assessing hepatic fibrosis. *Abdom Radiol.* **42**(8), 2069—2078 (2017)
2. Magalhaes Santos, M., Oliveira, C.B., Araujo-Filho et al.: State-of-the-art in radiomics of hepatocellular carcinoma: a review of basic principles, applications, and limitations. *Abdom. Rad.* **45**, 342–353 (2020)
3. Nordlinger, B., Sorbye, H., Glimelius, B., et al.: Perioperative FOLFOX4 chemotherapy and surgery versus surgery alone for resectable liver metastases from colorectal cancer (EORTC 40983): Long-term results of a randomised, controlled, phase 3 trial. *Lancet Oncol.* **14**(12), 1208—15 (2013)
4. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
5. Rizzo, S., Botta, F., Raimondi, S. et al.: Radiomics: the facts and the challenges of image analysis. *Eur. Rad. Exp.* **12** (2018)
6. Rubbia-Brandt, L., Audard, V., Sartoretti, P. et al.: Severe hepatic sinusoidal obstruction associated with oxaliplatin-based chemotherapy in patients with metastatic colorectal cancer. *Ann Oncol.* **15**(3), 460–466 (2004)
7. Sollini, M., Bandera, F., Kirienko, M.: Quantitative imaging biomarkers in nuclear medicine: from SUV to image mining studies. *EJNMMI*, **16**, 2737–2748 (2018)
8. Takamoto, T., Hashimoto, T., Sano, K. et al.: Recovery of liver function after the cessation of preoperative chemotherapy for colorectal liver metastasis. *Ann. Surg. Oncol.* **17**(10), 2747–2755 (2010)
9. Van Cutsem, E., Cervantes, A., Adam, R. et al.: ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. *Ann Oncol.* **27**(8), 1386—422 (2016)
10. Vauthey, J.N., Pawlik, T.M., Ribero, D. et al.: Chemotherapy regimen predicts steatohepatitis and an increase in 90-day mortality after surgery for hepatic colorectal metastases. *J Clin Oncol.* **24**(13), 2065–2072 (2006)
11. Viganó, L., Ravarino, N., Ferrero, A. et al.: Prospective evaluation 475 of accuracy of liver biopsy findings in the identification of chemotherapy-associated liver injuries. *Arch. Surg.* **147**(12), 1085—1091 (2012)
12. Yip, S.S.F., Aerts, H.J.W.L.: Applications and limitations of radiomics. *Physics in Medicine and Biology.* **61**, 50–66 (2016)

# Functional alignment by the “light” approach of the von Mises-Fisher-Procrustes model

## *Allineamento funzionale tramite l’approccio “light” del modello von Mises-Fisher-Procrustes*

Angela Andreella, and Livio Finos

**Abstract** Procrustes-based methods involve the singular value decomposition of a square matrix, leading to polynomial time complexity, and requiring a considerable memory for large-scale problems. Procrustes-based methods are used as functional alignment for fMRI data in multi-subjects analysis. A high-dimensional matrix expresses the subject’s neural activation, and Procrustes-based methods are infeasible (computationally). The alignment can be conducted only on regions of interest of the brain. We proposed a “light” version of the Procrustes-based methods. A semi-orthogonal transformation reduces the matrices’ dimension before applying the Procrustes alignment, maintaining the variability of the matrix that enters in the decomposition step. fMRI application shows a low decrease in predictive performance.

**Abstract** *I metodi di Procuste prevedono la decomposizione ai valori singolari di matrici quadrate, portando ad una complessità temporale polinomiale e richiedendo una memoria considerevole per problemi su larga scala. I metodi di Procuste sono utilizzati come allineamento funzionale per i dati fMRI nell’analisi multi-soggetto. Una matrice ad alta densità descrive l’attivazione neurale del singolo soggetto. L’allineamento può dunque essere effettuato solo su determinate zone del cervello. Si propone un approccio “light” del metodo di Procuste. Una trasformazione semi-ortogonale riduce la dimensione delle matrici prima di applicare l’allineamento funzionale, mantenendo la variabilità della matrice che entra nella fase della decomposizione ai valori singolari e mantenendo dunque le prestazioni predittive.*

**Key words:** Procrustes method, von Mises-Fisher-Procrustes model, semi-orthogonal matrix, fMRI data

---

Angela Andreella  
Department of Statistical Sciences, University of Padua, Italy  
e-mail: [angela.andreella@unipd.it](mailto:angela.andreella@unipd.it)

Livio Finos  
Department of Developmental Psychology and Socialization, University of Padua, Italy  
e-mail: [livio.finos@unipd.it](mailto:livio.finos@unipd.it)

## 1 Introduction

Procrustes methods are common in various fields such as neuroimaging [4]. However, dealing with high-dimensional data is critical since the Procrustes transformation must perform the Singular Value Decomposition (SVD), which is hugely time-consuming and employs a sizeable storing memory.

This paper proposes a “light” approach to Procrustes methods using the thin SVD. The Procrustes transformation is computed in a lower-dimensional manifold extracted by a semi-orthogonal transformation from the thin SVD of the reference matrix used in the Procrustes algorithm. The original fat matrix is reduced to a lower-dimensional square matrix having dimension equals the rank. Procrustes methods are then applied to these lower-dimensional matrices. Finally, the semi-orthogonal transformation’s inverse is used on the aligned matrices to project the objects in the original high-dimensional space.

In practice, the “light” approach is useful in functional Magnetic Resonance Imaging (fMRI) data analysis. High-dimensional matrix, e.g., with dimensions  $200 \times 200,000$ , represents the neural activation of a subject during some stimuli, where the rows represent the time points and the columns the units of the fMRI image, i.e., the voxels. The Procrustes-based functional alignment is applied to perform multi-subjects fMRI data analysis since the matrices’ columns are not in correspondence across subjects. It requires the SVD of a large square matrix with dimension equal to the number of voxels, e.g., roughly 200,000. Since the runtime is inadmissible, fMRI data’s functional alignment can be performed only in Region Of Interest (ROI) of the brain, instead of the whole brain [4]. Thanks to the “light” approach, the time complexity becomes equal to  $O(n^3)$ , where  $n$  is the number of time points. It speeds up the ROI analysis and permits the whole-brain analysis.

The paper is organized as follows. Section 2 analyzes the choice of the semi-orthogonal transformation. Section 3 applies the “light” approach to the von Mises-Fisher-Procrustes (vMFP) model proposed in [2]. Finally, the method is applied to fMRI data and evaluated by multi-subjects inference analysis in Section 4. We used the programming language Python, and in particular the `PYMVPA` package [3].

## 2 Semi-orthogonal transformation

Let  $\{X_i \in \mathbb{R}^{n \times m}\}_{i=1, \dots, N}$  be a set of rank  $n$  matrices, and  $M = \sum_{i=1}^N X_i / N$ . We have  $N$  independent observations to be aligned, e.g., subjects, taking values in  $\mathbb{R}^{n \times m}$ , where  $m$  is the number of variables, e.g., voxels, and  $n$  the observations, e.g., time points. The matrices  $X_i$  are projected into a lower-dimensional space by a semi-orthogonal transformation [1] defined below.

**Definition 1.** We call  $Q \in \mathbb{R}^{m \times n}$  a semi-orthogonal matrix if  $Q$  is a non-square matrix having orthonormal columns. So,  $Q^\top Q = I_n$ , it is a partial isometry of the Euclidean space, i.e., rotation or reflection applied from the left.

Functional alignment by the “light” approach of the von Mises-Fisher-Procrustes model

Instead of the matrices  $X_i$ , and  $M$  in the Procrustes analysis, we consider the transformations  $X_i^* = X_i Q$ , and  $M^* = M Q$ , taking values in  $\mathbb{R}^{n \times n}$ . Definition 1 could be rephrased considering the thin SVD with  $Q \in \mathbb{R}^{m \times k}$ , where  $k \leq n$ . We choose  $k = n$  to have a minimal loss of information, and to have unique solutions of the Procrustes-based problem [2, Lemma 1]. The semi-orthogonal matrix  $Q$  rotates  $X_i$  into a new coordinates system having  $n$  uncorrelated dimensions.

The next Theorem is the main result of the paper.

**Theorem 1.** *Let  $X_1, X_2$  be matrices in  $\mathbb{R}^{n \times m}$  with rank equals  $n$ ,  $LSQ^\top$  be the thin SVD of  $X_2$ , where  $Q \in \mathbb{R}^{m \times n}$ , then:*

$$\text{tr}(X_2^\top X_1) = \text{tr}(Q^\top X_2^\top X_1 Q). \quad (1)$$

Thus, the variability of  $X_2^\top X_1$  is also maintained after semi-orthogonal transformation. The “light” approach concentrates the similarity transformation around the first  $n$  eigenvectors instead of the full set of data.

### 3 “Light” von Mises-Fisher-Procrustes model

The vMFP model under the “light” approach is defined as:

$$X_i Q = \alpha_i (M Q + E_i) R_i^{\top*}, \quad (2)$$

where  $E_i \sim \mathcal{M}\mathcal{N}_{n,n}(0, \sigma^2 I_n, I_m)$ , and  $R_i^*$  distributed as the von Mises-Fisher distribution with location parameter  $F^* \in \mathbb{R}^{n \times n}$  and concentration parameter  $k \in \mathbb{R}^+$ . For further details about the vMFP model’ assumptions, please refers to [2].

W.l.o.g., let  $\alpha_i = 1$ , and consider the following maximization:

$$\hat{R}_i^* = \arg \max_{R_i^* \in \mathcal{O}(n)} \left\{ -\|(X_i Q)^\top - R_i^* Q^\top M^\top\|_F^2 + k \sigma^2 \text{tr}(F^{\top*} R_i^*) \right\}. \quad (3)$$

Following the idea of [2],  $R_i^*$  must combine the columns of  $X_i Q$  by exploiting some data prior feature, e.g., spatial closeness.  $F^*$  can be defined as the identity matrix, or by some lower rank approximation of the similarity euclidean distance.

The trace difference between the vMFP model and the “light” one equals  $m - n$ :

$$\arg \max_{R_i^* \in \mathcal{O}(n)} \left( \langle Q^\top X_i^\top M Q + k F^* + \frac{m-n}{n} I_n, R_i^* \rangle_F \right), \quad (4)$$

since  $F^*$  is a matrix with 1 on the diagonal. The following theorem expresses Theorem 1 in the vMFP model framework.

**Theorem 2.** *Let consider  $X_i, M \in \mathbb{R}^{n \times m}$  with rank  $n$ , where  $i = 1, \dots, N$ , and the thin SVD of  $M$  be  $LSQ^\top$ , where  $Q \in \mathbb{R}^{m \times n}$ , then:*

$$\text{tr}(X_i^\top M + F) = \text{tr}(Q^\top X_i^\top M Q + F^* + \frac{m-n}{n} I_n). \quad (5)$$

The algorithm presented in [2] must be modified:  $Q$  is applied on the data before the functional alignment by the vMFP model decomposing  $X_i^{*\top} M^* + k^* F^* + \frac{m-n}{n} I_n$ , instead of  $X_i^\top M + k^* F$ , where the term  $\frac{m-n}{n} I_n$  enters in the tuning parameter  $k\sigma^2$ .

## 4 Functional Magnetic Resonance Imaging data application

Procrustes-based functional alignment methods require the eigendecomposition of a square matrix. In the case of fMRI data, this square matrix has dimensions equals to the number of voxels, i.e., roughly 200,000. Procrustes-based methods are then unsuitable for aligning the whole brain, it can be applied only on ROIs. In contrast, the “light” version of the vMFP model permits to align the whole subjects’ brains and then perform the subsequent analysis on the entire dataset.

The vMFP model and its “light” approach are applied to the *Auditory* data collected by [6]. The neural activations of 18 subjects passively listening to vocal, i.e., speech, and non-vocal sounds are analyzed. The data are preprocessed by a standard procedure using the FMRIB Software Library (FSL) [5]. The prior location matrix is defined as the euclidean similarity distance of the matrix of the voxels’ three-dimensional coordinates multiplied by  $Q$ .

The aim is to test the group-level activation for each voxel under the null hypothesis of no activation. Let consider the model  $\hat{\beta}_{ij} = \mu_j + \varepsilon_{ij}$ , where  $\hat{\beta}_{ij}$  are the parameter estimates involving brain activation differences under the two stimuli, for each subject  $i$  and each voxel  $j = 1, \dots, m$ ,  $\mu_j$  is the unknown parameter of interest representing the between-subject mean activation, and  $\varepsilon_{ij}$  are the error terms  $\sim (0, \Sigma)$ . The one-sample t-test is performed to make inference on  $\mu_j$ :

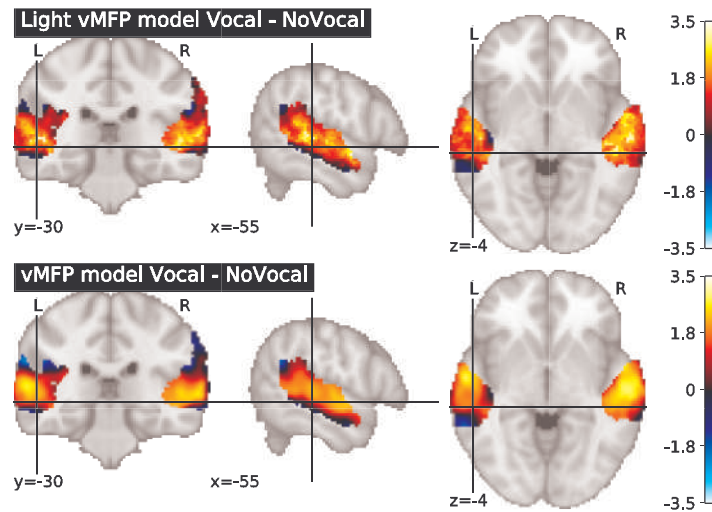
$$T_j = \frac{\hat{\mu}_j}{\sqrt{\hat{\sigma}_j^2/18}}, \quad (6)$$

where  $\hat{\mu}_j = \sum_{i=1}^{18} \hat{\beta}_{ij}/J$  and  $\hat{\sigma}_j^2 = \sum_{i=1}^{18} (\hat{\beta}_{ij} - \hat{\mu}_j)^2/17$ . So, we have  $m$  statistical tests, i.e.,  $H_0^j : \mu_j = 0$ , that create a statistical parametric mapping (SPM).

### 4.1 Region of interest analysis

We perform the group-level activation analysis by considering the Superior Temporal Gyrus (STG) as ROI being responsible for the sensation of sound. The neural activations are expressed by  $310 \times 10233$  matrices, one for each subject. Figures 1 represent the SPM (6) having data aligned by the vMFP model with and without the “light” approach. The anatomical structure is also maintained if the  $Q$  transformation is used. The “light” approach returns a value of  $|T_j|$  46.63% higher than those computed by the original vMFP model, with baseline 50%.

Functional alignment by the “light” approach of the von Mises-Fisher-Procrustes model



**Fig. 1** SPM using STG images aligned by the vMFP model with and without  $Q$  transformation.

## 4.2 Whole-brain analysis

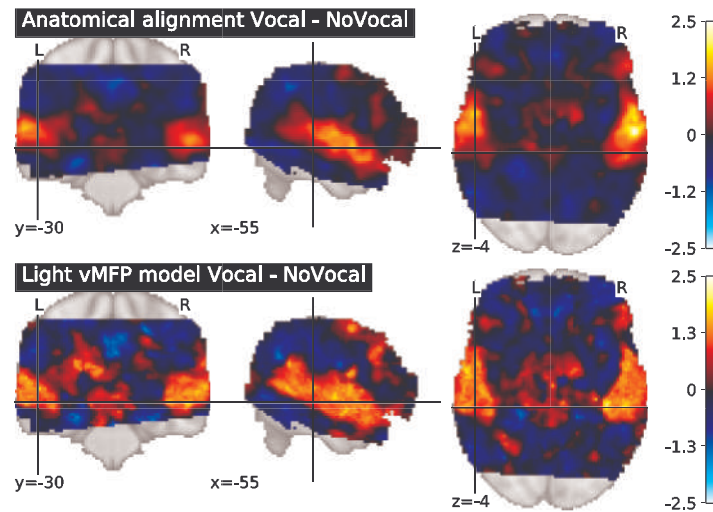
The inferential analysis is performed on the whole brain. The “light” vMFP model is compared with the anatomical alignment, being the only method applicable to the entire brain. Figures 2 show the  $T_j$ 's map using data aligned by the vMFP model and the anatomical alignment. The “light” approach returns brain maps with delineated boundaries between positive and negative t-tests preserving the anatomical structure. The functional region of the STG considering the top of Figure 2 seems more blurred than the one calculated using the vMFP model (bottom of Figure 2). The “light” version returns a value of  $|T_j|$  65.67% higher than those returned by the anatomical alignment, with again baseline 50%.

## 5 Discussion

The “light” version to the vMFP model permits to speed up the computation time in performing the SVD step of the estimation process, and at the same time, permits to apply the functional alignment on high-dimensional data.

The loss of information appears to be negligible in fMRI applications, since the trace of the data does not change if the semi-orthogonal transformation is applied. In the ROI's analysis, we found a minimal loss of power with respect to the vMFP model [2]. In addition, the alignment using the “light” approach takes approximately 5 minutes while one hour was required to run the original vMFP model. In the whole-brain analysis, the improvement with respect to the anatomical alignment is





**Fig. 2** SPM using brain images aligned by the anatomical alignment and “light” vMFP model.

noticeable, and the computational effort remains affordable (approximately 2 hours on a 1.8 GHz CPU processor with 16 GB of RAM).

The fMRI analysis is focused on understanding the neural activity in ROIs or whole brain. The hypothesis tested in the ROI analysis regards a particular region. In contrast, the whole-brain analysis tests which brain areas show task-related brain activity. Thanks to the “light” approach presented, both analyses can be performed after the functional alignment pre-processing step.

## References

1. Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra*. Cambridge University Press.
2. Andreella, A. and Finos, L. (2020). The von mises fisher procrustes model in functional magnetic resonance imaging data. *submitted*.
3. Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., and Pollmann, S. (2009). PyMVPA: A Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, (7):37–53.
4. Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., and Ramadge, P. (2011). A common high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(1):404–416.
5. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., and Woolrich, M. W., Smith, S. M. (2012). Fsl. *Neuroimage*, 62(2):782–790.
6. Pernet, C. R., McAleer, P. M., L., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., Watson, R. H., Fleming, D., Crabbe, F., Valdes-Sosa, M., and Belin, P. (2015). The Human Voice Areas: Spatial Organization and Inter-Individual Variability in Temporal and Extra-Temporal Cortices. *Neuroimage*, 119:164–74.
7. Talairach, J. J. and Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain. 3-Dimensional Proportional System: An Approach to Cerebral Imaging*. Atlante.

# A screening procedure for high-dimensional autologistic models

## *Una procedura di screening per modelli autologistici ad alta dimensionalità*

Rodolfo Metulini and Francesco Giordano

**Abstract** Logistic regression is adopted in failure prediction when the response is binary. We focus on variable selection assuming high-dimensionality and spatial autocorrelation (SAR). Pseudo maximum likelihood is asymptotically consistent for the autologistic when SAR is moderate but, in this framework, there is a shortage of variable selection methods. Robust screening procedures exist for generalized linear models with logit link function, but not for autologistic. We aim, by mean of a computational strategy, to identify the extent in which our screening procedure based on a marginal approach is valid for the autologistic model. We find a good performance, even for large SAR and moderate sample dimension.

**Abstract** Il modello logistico è usato per la previsione del fallimento quando la risposta è binaria. Ci focalizziamo sulla selezione delle variabili rilevanti sotto ipotesi di alta dimensionalità e autocorrelazione spaziale (SAR). Lo stimatore di pseudo massima verosimiglianza è asintoticamente consistente per modelli autologistici con SAR moderata, tuttavia, i metodi di selezione delle variabili in questo contesto scarseggiano. Esistono procedure di screening robuste per modelli lineari generalizzati con link logit, ma non per l'autologistico. Mediante strategia computazionale, si mira a testare la validità della procedura proposta, basata sulla selezione delle covariate rilevanti con un approccio marginale. Otteniamo una buona performance anche quando la SAR è alta e il campione è di dimensioni moderate.

**Key words:** Autologistic, High-dimensional data, Sure Independence Screening, Generalized Linear Models, Sure Screening Property

---

Rodolfo Metulini

Department of Economics and Statistics (DISES) - University of Salerno, Via Giovanni Paolo II, 132, Fisciano SA 84084, Italy, e-mail: rmetulini@unisa.it

Francesco Giordano

Department of Economics and Statistics (DISES) - University of Salerno, Via Giovanni Paolo II, 132, Fisciano SA 84084, Italy e-mail: giordano@unisa.it

## 1 Introduction

Bank failure prediction has been diffusely employed with a statistical modelling approach. Seminal works using discriminant analysis [1] make way to logistic models [2] and to machine learning techniques, such as support vector machines [3]. In this paper we focus on generalized linear models (GLM) with logit link function [4] with the inclusion of a spatial component to account for spatial autocorrelation (SAR). Firm's performance is not independent from the performance of other firms located in space, due to the presence of geographical proximity. SAR may emerge when the response at location  $i$  is dependent with the response at location  $j$ , for  $j$  neighbour of  $i$ . Ignoring this type of SAR leads to bias on model's parameters. Similarly to Andreano et al. [5] we account for SAR by resorting on the autologistic model [6]. Another fundamental aspect is that of correctly measuring and interpreting the effect of a large number of covariates [7]. The problem arises of selecting the set of relevant features. We deal this problem by resorting on screening procedures based on selecting important covariates by means of a marginal approach for ultra-high dimensional data [8] because it is proved that penalized variable selection methods suffer for noise accumulation in this framework [9]. The key point for every screening procedure is the sure screening property (SSP), which means that the estimated set of relevant covariates contains the true relevant ones with a probability that tends to 1, when the sample size grows [8]. In the case of GLM, Fan and Song [10] demonstrated such a property, but nothing has been done to prove SSP for autologistic. The aim of this paper is to evaluate SSP of our screening procedure. By mean of a computational experiment we evaluate the performance of high-dimensional autologistic models in term of their SSP, with reference to pseudo maximum likelihood (PML). Section 2 outlines the methodological framework; Section 3 introduces to the proposed screening procedure; Section 4 describes the simulation experiment. Finally, results and some concluding remarks are reported in Section 5.

## 2 Modelling framework

Autologistic models for binary response in regular lattice data set-up [11] has been firstly proposed by Besag [12] by directly imposing a joint Markov random field and reminds the formulation of the logistic regression derived by McCullagh & Nelder [4]. Let  $Y_i \in \{0, 1\}$ ,  $i = 1, \dots, n$  be the  $i$ -th binary element of the vector  $\mathbf{Y}$ ,  $\mathbf{X}_i$  be the vector column corresponding to the  $i$ -th row of the design matrix  $\mathbf{X}$  with  $n$  rows and  $p$  columns,  $\boldsymbol{\beta}$  be the vector containing the  $p$  regression parameters to be estimated. The full conditional distribution of  $\mathbf{Y}$  according to autologistic considering the assumption of stationary and isotropic processes along with Cressie's clique  $n$ . 2 [11] is given by:

$$\log \frac{P(Y_i = 1 | \mathbf{X}, \mathbf{Y})}{P(Y_i = 0 | \mathbf{X}, \mathbf{Y})} = \mathbf{X}_i' \boldsymbol{\beta} + \eta \sum_{j \neq i} w_{ij} Y_j, \quad (1)$$

where  $\eta$  is a scalar and  $w_{ij}$  is the  $(i, j)$  element of the  $n \times n$  matrix  $\mathbf{W}$ , with  $w_{ij} = 1$  if  $i$  is neighbour of  $j$ , 0 otherwise. Caragea and Kaiser [13] proposed a centred reparametrization to provide meaningful interpretations of the parameters. According to this variant,  $Y_j$  is replaced by  $Y_j - \mu_j$  in eq. 1, where  $\mu_j$  is the unconditional expectation of  $Y_j$ . By assuming positivity condition (i.e., if  $P(Y_i) > 0, i = 1, \dots, n$ , then  $P(Y_1, \dots, Y_n) > 0$ ) and Brook's Lemma (Besag [14], pag. 195) it is possible to generate the following joint distribution:

$$\pi(\mathbf{Y} | \boldsymbol{\theta}) = c(\boldsymbol{\theta})^{-1} \exp\left(\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \eta\mathbf{Y}'\mathbf{W}\boldsymbol{\mu} + \frac{\eta}{2}\mathbf{Y}'\mathbf{W}\mathbf{Y}\right), \quad (2)$$

where  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$  is the vector of expectations,  $\boldsymbol{\theta} = [\boldsymbol{\beta}', \eta]'$  and  $c(\boldsymbol{\theta})$  is the normalizing constant. We place in the framework of the maximum pseudo-likelihood estimation (MPLE) method [12], which circumvent the issue of computational intractability of the normalizing constant  $c(\boldsymbol{\theta})$  by maximizing the pseudo-likelihood with respect to the parameters as if it were a standard maximum likelihood. Related literature focused on methods for obtaining this normalizing constant [15]. However, despite MPLE is not efficient, with a loss of efficiency positively related to the absolute value of  $\eta$ , asymptotical consistency and normality are guaranteed [12]. Overall, the advantage of MPLE compared to alternative methods is its computational simplicity in exchange of a very little sacrifice of precision.

### 3 Screening procedure setup

Recently, methods based on using a penalty for both fitting and penalization of the model coefficients has been proposed, such as least absolute shrinkage and selection operator (LASSO) [16] and its generalizations. However, variable selection methods specifically proposed for autologistic model (e.g., [17]) have never been proved in high dimensional setup. In high-dimension, data require sophisticated variable selection methods accounting for i) noise accumulation, ii) spurious correlation, and iii) incidental endogeneity [9] which makes the aforementioned penalty-based methods inappropriate. To the best of our knowledge, no variable selection procedure has been developed for high-dimensional autologistic models. In this work we focus on screening procedures to select important covariates with a marginal approach. Among screening procedures adopting marginal maximum likelihood, the one proposed by Fan and Song [10] is proved, under some general conditions, to be consistent and efficient in GLM with logit link function, and to enjoy the SSP for the case of NP-Dimensionality. Thinking to the centered autologistic we have, for the  $i$ -th observation in space, the spatial component  $\sum_{j \neq i} w_{ij}(Y_j - \mu_j)$ , that introduces dependence in the model ( $Y_i \not\perp Y_j$ ). We have that marginal maximum pseudo likelihood estimates (MMPLE) reads as:

$$\tilde{\boldsymbol{\beta}}_h^{MMPLE} = \operatorname{argmax}_{\boldsymbol{\beta}_h} \prod_{i=1}^n P(Y_i | X_{ih}, \mathbf{Y}_{-i}), \quad (3)$$

where  $\mathbf{Y}_{-i} = [Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n]$  and  $X_{ih}$  is the  $i$ -th observation of the  $h$ -th covariate.

SSP for the autologistic can be written as:

$$\mathbb{P}_{n \rightarrow \infty} \{ \mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n} \} \rightarrow 1 \tag{4}$$

where the set of the true important variables with associated coefficients  $\beta_*$  is  $\mathcal{M}_* = \{1 \leq h \leq p_n : \beta_h \neq 0\}$ . Moreover,  $m = |\mathcal{M}_*|$ . Given a pre-specified threshold  $\gamma_n$ , the estimated set is  $\hat{\mathcal{M}}_{\gamma_n} = \{1 \leq h \leq p_n : |\hat{\beta}_h^{MMPLE}| \geq \gamma_n\}$ . The aim of this paper is to find at what extent SSP is valid when performing our screening procedure based on MMPLE in high-dimensional autologistic models.

### 4 Computational strategy

We employ an algorithm (Table 1) to evaluate SSP of MMPLE by mean of the median of the Minimum Model Size (MMS) of the selected marginals models, along with its associated Robust Standard Deviation (RSD), as done by Fan & Song [10]. Actually, we do not specify parameter  $\gamma_n$ . Instead, we replace  $\hat{\mathcal{M}}_{\gamma_n}$  with  $\hat{\mathcal{M}}$  being the smallest set including ordered (descending) estimated coefficients such that the set  $\mathcal{M}_*$  is a subset of it.

---

**Input:** Chosen values of  $n, \eta, \beta_1, \dots, \beta_p, \mu$  and  $\sigma, s, m$  and  $\mathbf{X}$  determined from the design of the experiment.

1. For  $k = 1$ :
  - a. simulate  $\mathbf{Y}$  with CFTP
  - b. estimate  $\hat{\beta}_1^{MMPLE}, \dots, \hat{\beta}_p^{MMPLE}$  as in eq. 3
  - c. order (descending)  $\hat{\beta}_1^{MMPLE}, \dots, \hat{\beta}_p^{MMPLE}$  in terms of their absolute value
  - d. find the minimum model size such that important variables  $\mathbf{X}_1, \dots, \mathbf{X}_m$  are all included in the estimated set  $\hat{\mathcal{M}}$
  - e. update  $k = k + 1$
2. **if**  $k \leq 200$  **then**
  - | repeat points 1(a) – 1(d)
  - else**
  - | compute MMMS with associated RSD

---

**Table 1** Algorithm adopted to evaluate SSP with MMMS using MMPLE and autologistic models.

About the design of experiment, for correctly generating the sample values for  $\mathbf{Y}$  we rely on perfect sampling coupling from the past (CFTP) [18], which better accounts for the dependence in  $\mathbf{Y}$  compared to traditional MCMC methods [6]. We generate the element  $w_{ij}$  as a realization of a *Bern*( $s$ ) process, by specifying  $s$  (support  $[0, 1]$ ) to be a parameter for the density of  $\mathbf{W}$ . All the  $\mathbf{X}$ 's are realizations of an *i.i.d.* process  $N(0, 1)$ . We generate a large number of covariates,  $p = 1000$ . We

choose  $m = 3$  and  $m = 6$  non-zero coefficients (more precisely we let  $\beta_*$  to be a vector of ones) related to relevant covariates. We also use different values for the level of spatial dependence  $\eta = \{0, 0.1, 0.2, 0.3, 0.5\}$ , just considering the case of “positive” SAR. We perform  $k = 200$  iterations on a sample dimension of  $n = 200$  and  $n = 500$ . For the simulation experiment we use package `ngspatial` in R [19] to estimate the parameters of the autologistic model.

## 5 Results and concluding remarks

Results from the computational exercise are shown in Table 2. We find that, for a moderate size of important covariates ( $m = 3$ ), SSP is guaranteed even for large levels of SAR and moderate sample dimension ( $n = 200$ ), since MMMS is 3 for all values of  $\eta$  in the design of experiment and RSD is even really small. SSP performance becomes poor when the number of relevant covariate increases ( $m = 6$ ) and the sample dimension is small ( $n = 200$ ). Under this setting, MMMS is way larger than  $m$  and RSD is also high. However, by increasing the sample to  $n = 500$ , SSP is again guaranteed, when  $m = 6$ .

n	$\eta$	MMMS (RSD)	n	$\eta$	MMMS (RSD)
m = 3, $\beta_* = [1, 1, 1]^T$			m = 6, $\beta_* = [1, 1, 1, 1, 1, 1]^T$		
200	0.0	3(1)	200	0.0	10(9)
200	0.1	3(1)	200	0.1	44(47)
200	0.2	3(0)	200	0.2	29(36)
200	0.3	3(0)	200	0.3	30(36)
200	0.5	3(1)	200	0.5	38(42)
500	0.0	3(0)	500	0.0	6(0)
500	0.1	3(0)	500	0.1	6(0)
500	0.2	3(0)	500	0.2	6(0)
500	0.3	3(0)	500	0.3	6(0)
500	0.5	3(0)	500	0.5	6(0)

**Table 2** MMMS and the associated RSD (in parenthesis) of the experiment for the MMPLE autologistic,  $k = 200$  and  $p = 1000$ .

These results may be useful for practitioners in the context of bank failure prediction because we restrict the extent in which the use of a screening procedure based on a “pseudo” marginal approach for selecting relevant covariates in autologistic is appropriate. As a further development we may think of deriving a methodological strategy in order to increase the performance of the proposed screening procedure in high-dimensional autologistic models when a large number of relevant covariates and a small sample size is assumed.

## References

1. Altman, E. I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), 589–609 (1968)
2. Ohlson, J. A.: Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131 (1980)
3. Shin, K. S., Lee, T. S., & Kim, H. J.: An application of support vector machines in bankruptcy prediction model. *Expert systems with applications*, 28(1), 127–135 (2005)
4. McCullagh, P.: *Generalized linear models*, Routledge (2018)
5. Andreano, M. S., Benedetti, R., Mazzitelli, A., & Piersimoni, F.: Spatial autocorrelation and clusters in modelling corporate bankruptcy of manufacturing firms. *Economia e Politica Industriale*, 45(4), 475–491 (2018)
6. Hughes, J., Haran, M., & Caragea, P. C.: Autologistic models for binary data on a lattice. *Environmetrics*, 22(7), 857–871 (2011)
7. Amendola, A., Giordano, F., Parrella, M. L., & Restaino, M.: Variable selection in highdimensional regression: a nonparametric procedure for business failure prediction. *Applied Stochastic Models in Business and Industry*, 33(4), 355–368 (2017)
8. Fan, J., & Lv, J.: Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911 (2008)
9. Fan, J., Feng, Y., & Tong, X.: A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4), 745–771 (2012)
10. Fan, J., & Song, R.: Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6), 3567–3604 (2010)
11. Cressie, N.: *Statistics for spatial data*. John Wiley & Sons (2015)
12. Besag, J.: Statistical analysis of nonlattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3), 179–195 (1975)
13. Caragea, P. C., & Kaiser, M. S.: Autologistic models with interpretable parameters. *Journal of agricultural, biological, and environmental statistics*, 14(3), 281–300 (2009)
14. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 192–225 (1974)
15. Ogata, Y., & Tanemura, M.: Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3), 496–518 (1984)
16. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288 (1996)
17. Fu, R., Thurman, A. L., Chu, T., Steen-Adams, M. M., & Zhu, J.: On estimation and selection of autologistic regression models via penalized pseudolikelihood. *Journal of agricultural, biological, and environmental statistics*, 18(3), 429–449 (2013)
18. Propp, J. G., & Wilson, D. B.: Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(12), 223–252 (1996)
19. Hughes, J.: *ngspatial: A Package for Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data*. *R Journal*, 6(2) (2014)

# Covariate adjusted censored gaussian lasso estimator

## *Un'estensione dello stimatore cglasso*

Luigi Augugliaro and Gianluca Sottile and Veronica Vinciotti

**Abstract** The covariate adjusted glasso is one of the most used estimators for inferring genetic networks. Despite its diffusion, there are several fields in applied research where the limits of detection of modern measurement technologies make the use of this estimator theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. In this paper we propose an extension to censored data.

**Abstract** *Il graphical lasso è uno degli stimatori più utilizzati per fare inferenza sulle reti genetiche. Nonostante la sua elevata diffusione, esistono parecchi campi applicativi dove i limiti degli strumenti di misurazione ne rendono teoricamente ingiustificato l'utilizzo, anche quando l'assunzione relativa alla distribuzione normale multivariata è soddisfatta.*

**Key words:** Censored data, Censored glasso estimator, Gaussian graphical model, glasso estimator.

## 1 Introduction

An important aim in genomics is to understand interactions among genes, characterized by the regulation and synthesis of proteins under internal and external signals. These relationships can be represented by a genetic network, i.e., a graph where

---

Luigi Augugliaro  
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: luigi.augugliaro@unipa.it

Gianluca Sottile  
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Veronica Vinciotti  
Department of Mathematics, Brunel University, UK, e-mail: veronica.vinciotti@brunel.ac.uk



nodes represent genes and edges describe the interactions among them. Gaussian graphical models [2] have been widely used for reconstructing a genetic network from expression data. The reason of such diffusion relies on the statistical properties of the multivariate Gaussian distribution which allow the topological structure of a network to be related with the non-zero elements of the concentration matrix, i.e., the inverse of the covariance matrix. Thus, the problem of network inference can be recast as the problem of estimating a concentration matrix. The covariate adjusted glasso estimator [5] is a popular method for estimating a sparse concentration matrix, based on the idea of adding two specific  $\ell_1$ -penalty function to the likelihood function of the multivariate Gaussian distribution.

Despite the widespread literature on the covariate adjusted glasso estimator, there is a great number of fields in applied research where modern measurement technologies make the use of this model theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. A first example of this is Reverse Transcription quantitative Polymerase Chain Reaction (RT-qPCR), a popular technology for gene expression profiling. This technique relies on fluorescence-based detection of amplicon DNA and allows the kinetics of PCR amplification to be monitored in real time. The analysis of the raw RT-qPCR profiles is based on the cycle-threshold, defined as the fractional cycle number in the log-linear region of PCR amplification in which the reaction reaches fixed amounts of amplicon DNA. If a target is not expressed or the amplification step fails, the threshold is not reached after the maximum number of cycles and the corresponding cycle-threshold is undetermined and the resulting data is naturally right-censored data. Another example is given by the flow cytometer, which is an essential tool in the diagnosis of diseases such as acute leukemias and malignant lymphomas. A flow cytometer measures a limited range of signal strength and records each marker value within a fixed range, such as between 0 and 1023. If a measurement falls outside this range, then the value is replaced by the nearest legitimate value; that is, a value smaller than 0 is censored to 0 and a value larger than 1023 is censored to 1023. In all these cases, a direct application of the covariate adjusted glasso for network inference is theoretically unfounded since it does not consider the effects of the censoring mechanism on the estimator of the concentration matrix. In order to overcome this problem, we propose an extension of the covariate adjusted glasso estimator that takes into account the censoring mechanism of the data explicitly.

## 2 The covariate adjusted censored Gaussian graphical model

Let  $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$  be a  $p$ -dimensional random vector. Graphical models allow to represent the set of conditional independencies among these random variables by a graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ , where  $\mathcal{V}$  is the set of nodes associated to  $\mathbf{Y}$  and  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  is the set of ordered pairs, called edges, representing the conditional dependencies among the  $p$  random variables [2]. The covariate adjusted Gaussian graphical model is an extension of the classical Gaussian graphical model based on the assumption

that the conditional distribution of  $\mathbf{Y}$  given a  $q$ -dimensional vector of predictors, say  $\mathbf{X} = (X_1, \dots, X_q)^\top$ , follows a multivariate Gaussian distribution with expected value:  $\boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{x}$ , where  $\boldsymbol{\beta} = (\beta_{hk})$  is a matrix  $q \times p$  coefficient matrix, and covariance matrix denoted by  $\boldsymbol{\Sigma} = (\sigma_{hk})$ . Denoting with  $\boldsymbol{\Theta} = (\theta_{hk})$  the concentration matrix, i.e., the inverse of the covariance matrix, the conditional density function of  $\mathbf{Y}$  can be written as follows:

$$\phi(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta}) = (2\pi)^{-p/2} |\boldsymbol{\Theta}|^{1/2} \exp[-1/2 \{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}^\top \boldsymbol{\Theta} \{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}]. \quad (1)$$

As shown in [2], the off-diagonal elements of the concentration matrix are the parametric tools relating the pairwise Markov property to the factorization of the density (1). Formally, two random variables, say  $Y_h$  and  $Y_k$ , are conditionally independent given all the remaining variables if and only if  $\theta_{hk}$  is equal to zero. This result provides a simple way to relate the topological structure of the graph  $\mathcal{G}$  to the pairwise Markov property, i.e., the undirected edge  $(h, k)$  is an element of the edge set  $\mathcal{E}$  if and only if  $\theta_{hk} \neq 0$ ,

As done in [1], we assume that  $\mathbf{Y}$  is a (partially) latent random vector with density function (1). In order to include the censoring mechanism inside our framework, let us denote by  $\mathbf{l} = (l_1, \dots, l_p)^\top$  and  $\mathbf{u} = (u_1, \dots, u_p)^\top$ , with  $l_h < u_h$  for  $h = 1, \dots, p$ , the vectors of known left and right censoring values. Thus,  $Y_h$  is observed only if it is inside the interval  $[l_h, u_h]$  otherwise it is censored from below if  $Y_h < l_h$  or censored from above if  $Y_h > u_h$ . Under this setting, a rigorous definition of the joint distribution of the observed data can be obtained using the approach for missing data with nonignorable mechanism [3]. This requires the specification of the distribution of a  $p$ -dimensional random vector, denoted by  $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$ , used to encode the censoring patterns. Formally, the  $h$ th element of  $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$  is defined as  $R(Y_h; l_h, u_h) = I(Y_h > u_h) - I(Y_h < l_h)$ , where  $I(\cdot)$  denotes the indicator function. By construction  $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$  is a discrete random vector with support the set  $\{-1, 0, 1\}^p$  and probability function  $\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r}\} = \int_{D_{\mathbf{r}}} \phi(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta}) d\mathbf{y}$ , where  $D_{\mathbf{r}} = \{\mathbf{y} \in \mathbb{R}^p : R(\mathbf{y}; \mathbf{l}, \mathbf{u}) = \mathbf{r}\}$ .

Given a censoring pattern, we can simplify our notation by partitioning the set  $\mathcal{S} = \{1, \dots, p\}$  into  $o = \{h \in \mathcal{S} : r_h = 0\}$ ,  $c^- = \{h \in \mathcal{S} : r_h = -1\}$  and  $c^+ = \{h \in \mathcal{S} : r_h = +1\}$  and, in the following of this paper, we shall use the convention that a vector indexed by a set of indices denotes the corresponding subvector. For example, the subvector of observed elements in  $\mathbf{y}$  is denoted by  $\mathbf{y}_o = (y_h)_{h \in o}$  and, consequently, the observed data is the vector  $(\mathbf{y}_o^\top, \mathbf{x}^\top, \mathbf{r}^\top)^\top$ . As done in [1], the probability distribution of the observed data, denoted by  $\varphi(\{\mathbf{y}_o, \mathbf{r}\} \mid \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta})$ , can be defined as follows:

$$\varphi(\{\mathbf{y}_o, \mathbf{r}\} \mid \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta}) = \int \phi(\{\mathbf{y}_o, \mathbf{y}_c\} \mid \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta}) \Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} \mid \mathbf{Y} = \mathbf{y}\} d\mathbf{y}_c, \quad (2)$$

where  $c = c^- \cup c^+$ .

Density (2) can be simplified by observing that  $\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} \mid \mathbf{Y} = \mathbf{y}\}$  is equal to one if the censoring pattern encoded in  $\mathbf{r}$  is equal to the pattern observed in  $\mathbf{y}$ , otherwise it is equal to zero, i.e.,

$$\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} \mid \mathbf{Y} = \mathbf{y}\} = I(\mathbf{y}_{c^-} < \mathbf{l}_{c^-})I(\mathbf{l}_o \leq \mathbf{y}_o \leq \mathbf{u}_o)I(\mathbf{u}_{c^+} < \mathbf{y}_{c^+}),$$

where the inequalities in the previous expressions are intended elementwise. From this,  $\varphi(\{\mathbf{y}_o, \mathbf{r}\} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta)$  can be rewritten as

$$\varphi(\{\mathbf{y}_o, \mathbf{r}\} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta) = \int_{D_c} \phi(\{\mathbf{y}_o, \mathbf{y}_c\} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta) d\mathbf{y}_c I(\mathbf{l}_o \leq \mathbf{y}_o \leq \mathbf{u}_o), \quad (3)$$

where  $D_c = (-\infty, \mathbf{l}_{c^-}) \times (\mathbf{u}_{c^+}, +\infty)$ . Using density (3), the covariate adjusted censored Gaussian graphical model is defined as the set  $\{\mathbf{Y}, R(\mathbf{Y}; \mathbf{l}, \mathbf{u}), \varphi(\{\mathbf{y}_o, \mathbf{r}\} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta), \mathcal{G}\}$ , where  $\varphi(\{\mathbf{y}_o, \mathbf{r}\} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta)$  factorizes according to the undirected graph  $\mathcal{G}$ .

### 3 The covariate adjusted censored glasso estimator

Suppose we have a sample of size  $n$  independent observations drawn from a covariate adjusted censored Gaussian graphical model. For ease of exposition, we shall assume that  $\mathbf{l}$  and  $\mathbf{u}$  are fixed across the  $n$  observations, but the extension to the cases where the censoring vectors are specific to each observation is straightforward and does not require a specific treatment. To simplify our notation the set of indices of the variables observed in the  $i$ th observation is denoted by  $o_i = \{h \in \mathcal{S} : r_{ih} = 0\}$ , while  $c_i^- = \{h \in \mathcal{S} : r_{ih} = -1\}$  and  $c_i^+ = \{h \in \mathcal{S} : r_{ih} = +1\}$  denote the sets of indices associated to the left and right-censored data, respectively. Denoting by  $\mathbf{r}_i$  the realization of the random vector  $R(\mathbf{Y}_i; \mathbf{l}, \mathbf{u})$ , the  $i$ th observed data is the vector  $(\mathbf{y}_{i o_i}^\top, \mathbf{x}_i^\top, \mathbf{r}_i^\top)^\top$ . Using the density function (3), the observed log-likelihood function can be written as

$$\ell(\boldsymbol{\beta}, \Theta) = \sum_{i=1}^n \log \int_{D_{c_i}} \phi(\{\mathbf{y}_{i o_i}, \mathbf{y}_{i c_i}\} \mid \mathbf{x}_i; \boldsymbol{\beta}, \Theta) d\mathbf{y}_{i c_i} = \sum_{i=1}^n \log \varphi(\{\mathbf{y}_{i o_i}, \mathbf{r}_i\} \mid \mathbf{x}_i; \boldsymbol{\beta}, \Theta), \quad (4)$$

where  $D_{c_i} = (-\infty, \mathbf{l}_{c_i^-}) \times (\mathbf{u}_{c_i^+}, +\infty)$  and  $c_i = c_i^- \cup c_i^+$ . Although inference about the parameters of this model can be carried out via the maximum likelihood method, the application of this inferential procedure to real datasets is limited for three main reasons. Firstly, the number of measured variables is often larger than the sample size and this implies the non-existence of the maximum likelihood estimator even when the dataset is fully observed. Secondly, even when the sample size is large enough, the maximum likelihood estimator will exhibit a very high variance. Thirdly, empirical evidence suggests that gene networks or more general biochemical networks are not fully connected. In terms of covariance adjusted Gaussian graphical models this evidence translates in the assumption that  $\boldsymbol{\beta}$  and  $\Theta$  have a sparse structure, i.e., only few regression coefficients and few  $\theta_{hk}$  are different from zero.

All that considered, we propose to estimate the parameters of the covariate adjusted censored Gaussian graphical model by generalizing the approach proposed in [5], i.e., by maximizing a new objective function defined by adding two lasso-

type penalty functions to the observed log-likelihood (4). The resulting estimator, called covariate adjusted censored glasso estimator, is formally defined as

$$\{\hat{\boldsymbol{\beta}}^\lambda, \hat{\boldsymbol{\Theta}}^\rho\} = \arg \max_{\boldsymbol{\beta}, \boldsymbol{\Theta} \succ 0} \frac{1}{n} \sum_{i=1}^n \log \varphi(\{\mathbf{y}_{i0_i}, \mathbf{r}_i\} | \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\Theta}) - \lambda \sum_{h,k} |\beta_{hk}| - \rho \sum_{h \neq k} |\theta_{hk}|, \quad (5)$$

where  $\lambda$  and  $\rho$  are two non-negative tuning parameters. The lasso penalty on  $\boldsymbol{\beta}$  introduces sparsity in  $\hat{\boldsymbol{\beta}}^\lambda$ , in other words by varying  $\lambda$  we can select the relevant predictors for  $\mathbf{Y}$ . Like in the standard glasso estimator, the tuning parameter  $\rho$  controls the amount of sparsity in the estimated concentration matrix  $\hat{\boldsymbol{\Theta}}^\rho = (\hat{\theta}_{hk}^\rho)$  and, consequently, in the corresponding estimated graph  $\hat{\mathcal{G}}^\rho = \{\mathcal{V}, \hat{\mathcal{E}}^\rho\}$ , where  $\hat{\mathcal{E}}^\rho = \{(h, k) : \hat{\theta}_{hk}^\rho \neq 0\}$ . When  $\rho$  is large enough, some  $\hat{\theta}_{hk}^\rho$  are shrunk to zero resulting in the removal of the corresponding link in  $\hat{\mathcal{G}}^\rho$ ; on the other hand, when  $\rho$  is equal to zero and the sample size is large enough the estimator  $\hat{\boldsymbol{\Theta}}^\rho$  coincides with the maximum likelihood estimator of the concentration matrix, which implies a fully connected estimated concentration graph.

## 4 Simulation study

In this section, we compare our proposed estimator with MissGlasso [4], which performs  $\ell_1$ -penalized estimation under the assumption that the censored data are missing at random, and with the covariate adjusted glasso estimator [5], where the empirical covariance matrix is calculated by imputing the missing values with the censoring values. These estimators are evaluated in terms of both recovering the structure of the true graph and the mean squared error. We use the method implemented in the R package `huge` [6], to simulate a sparse concentration matrix with a random structure for  $\mathbf{Y}$ . In particular, we set the probability of observing a link between two nodes to  $k/p$ , where  $p$  is the number of responses and  $k$  is used to control the amount of sparsity in  $\boldsymbol{\Theta}$ . Moreover, we set the right censoring value to 40 for any variable and the sample size  $n$  to 100. The predictors matrix  $\mathbf{X}$  is sampled from a multivariate gaussian distribution with zero expected value and sparse covariance matrix simulated as done for  $\mathbf{Y}$ . Each column of  $\boldsymbol{\beta}$  contains only two non-zero regression coefficients sampled from a continuous uniform distribution on the interval  $[0.3, 0.7]$ . The values of the intercepts are chosen in such a way that  $H$  response variables are right censored with probability equal to 0.40. The quantities  $k$ ,  $p$ ,  $q$  and  $H$  are used to specify the different scenarios used to analyze the behavior of the considered estimators. In particular, we consider the following cases:

- **Scenario 1:**  $k = 3$ ,  $p = 50$ ,  $q = 10$  and  $H = 25$ . This setting is used to evaluate the effects of the number of censored variables on the behavior of the proposed estimators when  $n > p$ .
- **Scenario 2:**  $k = 3$ ,  $p = 150$ ,  $q = 10$  and  $H = 75$ . This setting is used to evaluate the impact of the high dimensionality on the estimators ( $p > n$ ).

For each scenario, we simulate 50 samples and in each simulation, we compute the coefficients path using cglasso, MissGlasso, and glasso. Each path is computed using an equally spaced sequence of  $\rho$  and  $\lambda$ -values. However, the two scenarios differ also on the length of the two sequences, that is 20 for the Scenario 1 and 10 for the Scenario 2. Moreover, the precision-recall curves and the area under the curves (AUCs) are computed for each Scenarios.

The curves report the relationship between precision and recall for any  $\rho$  and  $\lambda$ -value, which are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP, and FN are quantities defined as the number correctly selected non-null items, the number of wrong selected non-null items and the number of wrong selected null item, respectively. Table 1 shows how cglasso gives a better estimate of the concentration and coefficient matrices in terms of AUCs, for any given value of the tuning parameters. We report only five evenly spaced values of  $\lambda$  and  $\rho$ .

**Table 1** Mean area under the curves across the sequence of  $\rho$  and  $\lambda$ -values under the specification of the two Scenarios. The first column block refers to the concentration matrix ( $\Theta$ ) when  $\lambda$  is fixed and the second refers to the coefficient matrix ( $\beta$ ) when  $\rho$  is fixed.

		$\lambda/\lambda_{\max}$					$\rho/\rho_{\max}$				
		0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
<b>Model 1</b>	cglasso	0.546	0.429	0.139	0.103	0.101	0.844	0.877	0.883	0.882	0.885
	MissGlasso	0.239	0.199	0.086	0.073	0.073	0.745	0.764	0.766	0.767	0.768
	glasso	0.414	0.218	0.097	0.092	0.091	0.813	0.847	0.864	0.866	0.866
<b>Model 2</b>	cglasso	0.418	0.094	0.037	0.035	0.035	0.794	0.930	0.931	0.929	0.933
	MissGlasso	0.329	0.098	0.033	0.031	0.030	0.753	0.830	0.831	0.830	0.831
	glasso	0.321	0.040	0.033	0.032	0.031	0.751	0.902	0.906	0.907	0.907

## References

1. Augugliaro L., Abbruzzo A., Vinciotti V.:  $\ell_1$ -Penalized censored Gaussian graphical model. *Biostatistics*. **21**(2), e1–e16 (2020)
2. Lauritzen S. L.: *Graphical Models*. Oxford University Press, Oxford (1996)
3. Little R. J. A., Rubin D. B.: *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken (2002)
4. Städler, N., Bühlmann, P.: Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Stat. Comput.* **22**(1), 219–235 (2012)
5. Yin J., Li H.: A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Statist.* **5**(4), 2630–2650 (2011)
6. Zhao T., Li X., Liu H., Roeder K., Lafferty J., Wasserman L.: huge: High-Dimensional Undirected Graph Estimation. R package version 1.2.7 (2015). <https://CRAN.R-project.org/package=huge>

# Ranking-Based Variable Selection for ultra-high dimensional data in GLM framework

## *Selezione delle Variabili basata sui ranghi per modelli GLM con dati ad alta dimensionalità*

Francesco Giordano, Marcella Niglio and Marialuisa Restaino

**Abstract** In this contribution we propose a procedure to identify the most relevant covariates in presence of ultra-high dimensional data, i.e. when the number of covariates is much larger than the number of observations. The proposed procedure extends the idea of ranking-based variable selection developed in linear regression setting, to the more general class of generalized linear models. Then, the performance of our proposal is compared in a simulation study with a two-step technique, obtained by combining the screening procedure and lasso variable selection.

**Abstract** *In questo contributo si propone una procedura per selezionare le variabili più rilevanti in presenza di dati ad alta dimensionalità, in particolare quando il numero delle variabili è di gran lunga superiore alla dimensione campionaria. La procedura proposta estende l'idea della selezione delle variabili basata sull'ordinamento sviluppata per il modello di regressione lineare, ad una classe di modelli più ampia, quella dei modelli lineari generalizzati. La performance della procedura proposta è confrontata in uno studio di simulazione con un'altra tecnica a due-stadi, ottenuta come combinazione della procedura di screening e della selezione con il lasso.*

**Key words:** Ranking-based approach, screening, variable selection, ultra-high dimensional data, GLM

## 1 Introduction

Due to development of the big data and rapid technological advances in data collection, it is very common to deal with datasets with plenty of variables, but only very few of them is supposed to be truly relevant to explain the phenomenon under investigation. Thus, variable selection is fundamental for high and ultra-high dimensional analysis, when the number of predictors is much larger than the number of observations and eventually may grow exponentially with the sample size. In addition, increasing the number of covariates could lead to having high correlation among predictors.

A relevant number of variable selection methods via penalized least squares or likelihood have been proposed for high-dimensional data in linear regression and in

---

Francesco Giordano, Marcella Niglio, Marialuisa Restaino  
Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy e-mail: [giordano, mniglio, mlrestaino]@unisa.it

its extension to generalized linear models [4]. Among them, some of the most popular approaches for selecting significant variables and estimating regression coefficients simultaneously are the least absolute shrinkage and selection operator (lasso) [12], the smoothly clipped absolute deviation (SCAD) [2], the elastic net [13].

However, most of these cited techniques are difficult to implement for ultra-high dimensional data, due to computational complexity and algorithm stability and efficiency [6]. A reasonable solution is variable screening, that might simplify the original ultra-high dimensional problem into a lower one. The most common screening approach is Sure Independence Screening (SIS) [3]. It is based on the idea that given a model  $Y = f(X_1, \dots, X_p, \beta) + \varepsilon$ , with  $\beta = (\beta_1, \dots, \beta_p)^T$  the vector of parameters, the variables that influence  $Y$  survive (with probability tending to 1) after the reduction of the number of covariates. This reduction is performed ranking the  $p$  covariates using the estimation of the marginal coefficients  $|\hat{\beta}_j|$ , for  $j = 1, \dots, p$ , such that the covariates included in the submodel  $\mathcal{M}_\gamma$ , are  $\mathcal{M}_\gamma = \{1 \leq j \leq p : |\hat{\beta}_j| \geq \gamma\}$ , where  $\gamma$  is a predefined threshold value.

[3] discuss the consistency of the SIS approach showing that  $P(K \subset \mathcal{M}_\gamma) \rightarrow 1$ , with  $K$  the true set of relevant covariates (*screening property*).

After screening the data, all standard variable selection methods can be easily implemented. For example, the lasso method is commonly used [12].

If penalized likelihood approaches are selected, we estimate the vector of coefficients  $\beta$  by solving a set of non-linear equation that satisfy the maximum likelihood criterion. Moreover, many theoretical results proposed in the literature mainly focus on the ordinary linear regression model, while few of them consider the case of the discrete response variables [5, 7, 8, 9], especially in presence of high correlated variables.

Therefore, our aim is to propose a variable selection procedure, based on the ranking of variables [1], when the number of predictors is much larger than the sample size, and for a generalization of linear models, i.e. generalized linear models. The procedure is based on the evaluation of the marginal utility of the covariates and then, differently from the penalization techniques, does not require the solution of high-dimensional optimization problems and the estimation of regularization parameters. Then even the collinearity can be properly managed. We test the performance of the proposed procedure by a simulation study and we compare it with a two-step procedure, based on a combination between screening and lasso techniques.

The generalized linear models are introduced in Section 2. Then, in Section 2.1 we present the Ranking-Based Variable Selection (RBVS), whose application is extended to GLM in Section 3, where the results of a simulation study are shown to evaluate the performance of the proposed procedure.

## 2 The model

Generalized linear models (GLMs) [10] are very useful to treat many extensions of a linear model in a unified way, giving a flexible framework to study the association

between a family of continuous and discrete outcomes and a set of independent variables. The most applied model is logistic regression for binary responses.

Suppose that a random sample of  $n$  subjects is observed. Let  $Y_i$  be a response and  $X_i = (X_{i1}, \dots, X_{ip})^T$  be a vector of  $p$  predictors for the  $i$ th subject. Assume  $Y_i$  follows a distribution in the exponential family with mean  $\mu_i = E(Y_i)$  and variance  $V_i = \text{var}(Y_i)$ . GLMs model the mean  $\mu_i$  of  $Y_i$  as a function of covariates through a known monotone link function  $g$ :

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip},$$

where  $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$  is a  $(p + 1)$  vector of unknown regression coefficients.

The density function of  $Y_i$  in the exponential family is

$$L(Y_i; \theta_i, \phi) = \exp \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}, \quad (1)$$

where  $a(\cdot)$ ,  $b(\cdot)$  and  $c(\cdot)$  are known functions which vary according to the distributions,  $\phi$  is a dispersion parameter,  $\theta_i$  is a canonical parameter, and  $\theta_i = \theta(X_i, \beta)$ .

In the following we will focus the attention on the logistic regression where  $Y_i \sim \text{Bernoulli}(\mu_i)$ , whereas  $g(\mu_i) = \text{logit}(\mu_i)$ , for  $i = 1, \dots, n$ .

### 2.1 Ranking-based variable selection

As clarified in the previous sections, different approaches have been developed in variable screening and variable selection, and most of them require the introduction of tuning parameters whose selection is often based on empirical arguments.

A recent contribution in the screening plus variable selection domain has been given in [1], where the subset of  $\{X_1, \dots, X_p\}$  that contributes to  $Y$  is based on the ranking of covariates that allows to evaluate their impact on the response variable.

The ranking is based on some measures that allow to define the top-ranked variables (*screening step*) that are then evaluated to select the relevant covariates for  $Y$  (*selection step*).

The algorithm of [1], shortly called RBVS, is based on the following main idea: given the set of  $p$  covariates  $\{X_1, \dots, X_p\}$ , the variables having higher influence on  $Y$  are those that even in presence of randomly selected subsamples, extracted from the dataset, exhibit consistent relationship with  $Y$ .

We here shortly describe the RBVS algorithm, whereas for all technical details see [1].

Let  $Z_i = \{Y_i, X_{i1}, X_{i2}, \dots, X_{ip}\}$ , for  $i = 1, 2, \dots, n$  and with  $p$  that grows with  $n$ , be the observed dataset that is used to select the subset of  $\{X_1, \dots, X_p\}$  which is relevant for  $Y$ . Further, let  $\mathcal{A} \subset (1, \dots, p)$  be the indices that identify a subset of covariates and let  $|\mathcal{A}| = k$  be the cardinality of  $\mathcal{A}$ , for  $k = 0, 1, \dots, p$ . Let  $R_{ni}(Z_1, \dots, Z_n)$  be the ranking of the  $i$ th covariate obtained using the full dataset and consider the probability:

$$\pi_n(\mathcal{A}) = \mathbf{P}(\{R_{n1}(Z_1, \dots, Z_n), \dots, R_{n|\mathcal{A}|}(Z_1, \dots, Z_n)\} = \mathcal{A})$$

with  $\pi_n(\mathcal{A}) = 1$ , if  $\mathcal{A} = \emptyset$ .



Correspondingly define:

$$\pi_{n,m}(\mathcal{A}) = \mathbf{P}(\{R_{n1}(Z_1, \dots, Z_m), \dots, R_{n|\mathcal{A}|}(Z_1, \dots, Z_m)\} = \mathcal{A}),$$

the probability of  $\mathcal{A}$  obtained from a subset of  $m$  observations, with  $1 \leq m \leq n$ .

It follows that if  $\mathcal{A}$  corresponds to the indices of the top-ranked covariates,  $\pi_{n,m}(\mathcal{A})$  is its probability computed on a randomly selected subset of  $m$  observations.

To estimate  $\pi_{n,m}(\mathcal{A})$ , [1] use a bootstrap approach where for each  $b = 1, \dots, B$  (with  $B$  the number of bootstrap replicates) and given  $r = \lfloor n/m \rfloor$ , extract from  $Z_i$ , for  $i = 1, \dots, n$ ,  $r$  independent subsets without replacement  $(I_{b1}, \dots, I_{br})$  and for each bootstrap replicate compute the empirical relative frequency of  $\mathcal{A}$ , given by  $r^{-1} \sum_{j=1}^r \mathbf{1}(\mathcal{A}|I_{bj})$ , with  $\mathbf{1}(\cdot)$  an indicator function. Then, the estimate of  $\pi_{n,m}(\mathcal{A})$  is obtained from:

$$\hat{\pi}_{n,m}(\mathcal{A}) = B^{-1} \sum_{b=1}^B r^{-1} \sum_{j=1}^r \mathbf{1}(\mathcal{A}|I_{bj}), \tag{2}$$

where  $\mathcal{A}|I_{bj} = \{R_{n1}(Z_i)_{i \in I_{bj}}, \dots, R_{n|\mathcal{A}|}(Z_i)_{i \in I_{bj}}\}$ .

The probability (2) allows to define the top-ranked variables (*screening step*) given by:

$$\hat{\mathcal{A}}_{k,m} = \arg \max_{\mathcal{A} \in \Omega_k} \hat{\pi}_{n,m}(\mathcal{A}),$$

with  $\Omega_k$  the set of all permutations of  $\{1, \dots, k\}$ .

Starting from  $\hat{\mathcal{A}}_{k,m}$ , we need to detect the relevant variables for  $Y$  (*selection step*). It is common at this stage of the variable selection algorithm to introduce a threshold that allows to discriminate between important and irrelevant variables. The alternative introduced in the RBVS algorithm is to estimate the ratio  $\hat{\pi}_{n,m}^\tau(\hat{\mathcal{A}}_{k+1,m}) / \hat{\pi}_{n,m}(\hat{\mathcal{A}}_{k,m})$  with  $\tau \in (0, 1]$  such that the relevant covariates are the  $s$  top-ranked variables where:

$$\hat{s} = \arg \min_{k=0, \dots, k_{\max}-1} \frac{\hat{\pi}_{n,m}^\tau(\hat{\mathcal{A}}_{k+1,m})}{\hat{\pi}_{n,m}(\hat{\mathcal{A}}_{k,m})}. \tag{3}$$

In practice, given the estimated probabilities of  $\hat{\pi}_{n,m}(\hat{\mathcal{A}}_{k,m})$ , for  $k = 0, \dots, k_{\max} - 1$ , with  $k_{\max}$  a fixed large integer, the number of relevant variables is related to the evaluation of the magnitude of the estimated probability and  $\hat{s}$  corresponds to the case where the ratio in (3) has the greatest decrease, whereas  $\mathcal{S} = \hat{\mathcal{A}}_{\hat{s},m}$  is the subset of  $\{0, 1, \dots, p\}$  that contains the indices of the relevant variables.

The consistency of the RBVS algorithm, and then  $\mathbf{P}(\hat{\mathcal{S}} = \mathcal{S}) \xrightarrow[n]{n} 1$ , is shown in [1], whereas the progress here considered is its extension of the GLM case, that in our knowledge has not been proposed until now.

Given the limited number of pages, we will not present further theoretical results, and we mainly show the results of a simulation study where the performance of

the RBVS algorithm is evaluated in the GLM context, with application to logistic regression.

### 3 Simulation study

To evaluate how the RVBS algorithm performs under logistic regressions, we have generated the covariates from a multivariate Gaussian distribution with null vector of means, unit variances and correlation matrix with  $\text{corr}(X_i, X_j) = \rho$ , for  $i \neq j$ , where  $\rho$  can assume two different values  $\rho = \{0, 0.50\}$ , whereas the coefficients  $\beta_j = 4$ , for  $j = 1, 2, 3, 4$  and  $\beta_j = 0$ , for  $j = 5, \dots, p$ . The response variable is obtained from a Bernoulli random variable,  $Be(\mu)$ , with  $\mu = E[Y|X_1, \dots, X_p]$ .

The number of covariates  $p$  and the sample size  $n$  are set to be equal  $p = \{1,000, 2,000\}$  and  $n = \{200, 500\}$ , respectively.

Thanks to this setting, we are able to evaluate the performance of the RBVS algorithm in presence of ultra-high dimensional data with independent and correlated covariates.

With this data generating process we have implemented a Monte Carlo simulation study with 100 replicates, where at each iteration we have selected the relevant variables for  $Y$  through the RVBS algorithm and a standard benchmark given by the combination of the screening [7] and lasso [11] approaches, where the number of covariates given by lasso procedure is set equal to  $\{50, 75\}$  respectively for  $n = \{200, 500\}$ .

The parameters of the RBVS algorithm are:  $B = 100$  the number of bootstrap replicates used to estimate the probabilities (2), with  $m = n/2$ , the number of observations in each random subset; the measure considered to rank the  $p$  variables in each subset is the estimate  $|\hat{\beta}_j|$  of the coefficient related to the marginal covariate  $X_j$ , for  $j = 1, \dots, p$ , in the logistic model whereas  $k_{\max} = 20$  and, as suggested in [1],  $\tau = 0.5$ .

The results of simulation study are shown in Table 1. The mean of the number of variables selected is about four, that is the number of relevant covariates when  $\rho = 0$  for the RBVS procedure, while the combination of Screening and Lasso produces a larger number of selected variables. When the correlation increases, RBVS remains enough stable, selecting about five variables (that include always the relevant variables), whilst the other selects bigger set of variables (Table 1). Also the standard deviation of screening and lasso is much larger than in RBVS, for both  $\rho = 0$  and  $\rho = 0.5$ . Thus, the results show the ability of RBVS to detect the relevant covariates, in both high-dimensional and ultra-high dimensional settings.

The RBVS algorithm has been compared to further approaches, such as the SRB algorithm of [8] and the stepwise procedure of [9], showing that it outperforms them.

Looking at the False Positive (FP), that is the number of covariates that are not chosen as relevant by the variable selection procedures, and False Negative (FN), that is the number of irrelevant covariates classified incorrectly, we note that our procedure has better performance in terms of higher FP-rate and lower FN-rate in all settings.

### 4 Concluding remarks

In this paper we have proposed, in the GLM context, a procedure to select the most relevant variables when the number of covariates is much larger than the observations number, even in presence of correlated covariates. We have evaluated the performance of the procedure through a simulation study. Moreover the application to real data, not reported here, further confirms the good performance of the RBVS algorithm with respect to other competitors (such as Screening plus Lasso or those proposed in [8] and [9]) which has been evaluated using proper indices. These results, combined with some theoretical issues, will be object of future research.

**Table 1** Mean of the number of variables selected by the procedure considered for  $n = \{200, 500\}$ ,  $p = \{1000, 2000\}$  and  $\rho = \{0, 0.50\}$ . In brackets the s.d.

	$n = 200$ $p = 1,000$	$n = 500$ $p = 1,000$	$n = 500$ $p = 2,000$
$\rho = 0$			
Screening & Lasso	26.53 (3.71)	15.71 (3.32)	19.34 (3.63)
RBVS	4.10 (0.41)	4.34 (1.30)	4.26 (1.16)
$\rho = 0.50$			
Screening & Lasso	19.08 (3.29)	19.55 (3.71)	20.61 (4.08)
RBVS	4.61 (2.14)	4.74 (1.90)	4.58 (1.65)

### References

1. Baranowski R., Chen Y., Fryzlewicz P.: Ranking-based variable selection for high-dimensional data. *Statist Sinica*, **30**, 1485–1516 (2020)
2. Fan J., Li R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, **96**, 1348–1360 (2001)
3. Fan J., Lv J.: Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B*, **70**, 849–911 (2008)
4. Fan J., Lv J.: A selective overview of variable selection in high dimensional feature space. *Stat Sinica*, **20**, 101–148 (2010)
5. Fan J., Lv J.: Non-concave penalized likelihood with NP-dimensionality, *IEEE Trans. Inf. Theory*, **57**, 5467–5484 (2011)
6. Fan J., Samworth R., Wu Y.: Ultrahigh dimensional feature selection: Beyond the linear model. *J Mach Learn Res*, **10**, 1829–1853 (2009)
7. Fan J., Song R.: Sure independence screening in generalized linear models with np-dimensionality, *Ann. Statist.*, **38**, 3567–3604 (2010)
8. Hwang J.S., Hu, T.-H.: A stepwise regression algorithm for high-dimensional variable selection. *J Stat Comput and Sim*, **85**(9), 1793–1806 (2015)
9. Li Y., Liu J.S.: Robust Variable and Interaction Selection for Logistic Regression and General Index Models. *J Am Stat Ass*, **114**(525), 271–286 (2019)
10. McCullagh P., Nelder J. A.: *Generalized Linear Models*, Chapman & Hall/CRC (1989)
11. Park M.Y., Hastie T.:  $L_1$ -regularization path algorithm for generalized linear models. *J Roy Stat Soc Ser B*, **69**, 659–677 (2007)
12. Tibshirani R.: Regression shrinkage and selection via lasso. *J Roy Stat Soc Ser B*, **58**, 267–288 (1996)
13. Zou H., Hastie T.: Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Stat Methodol)*, **67**(2), 301–320 (2005)

## 4.33 Statistical methods in higher education

# **Effects of remote teaching on students' motivation and engagement: the case of the University of Modena & Reggio Emilia**

## *Gli effetti della didattica di emergenza online: il caso degli studenti dell'Università di Modena e Reggio Emilia*

Isabella Morlini and Laura Sartori

**Abstract** The Covid-19 pandemic has had dramatic impact on many dimensions of living and studying conditions of students at University. This paper analyses student satisfaction and motivation during the lockdown period and try to understand whether different socio-economic and environmental conditions have influenced the needs and the demands of the students for the online didactics. Drawing from the results of a questionnaire administered to students enrolled in the University of Modena and Reggio Emilia, this research is aimed at describing which factors, beyond the quality and the professionalism of the lecturers and the quality of the education received, influence the satisfaction with the online learning experience and impact on students' motivations and perceived engagement.

**Riassunto** *Il lavoro presenta alcuni risultati della rilevazione online condotta dall'Università degli studi di Modena e Reggio Emilia sulle condizioni di vita e di studio degli studenti nel periodo 8 aprile – 2 maggio 2020, durante il lockdown causato dall'emergenza Covid-19. L'obiettivo è quello di analizzare i fattori che hanno influenzato la soddisfazione degli studenti per la didattica a distanza e, soprattutto, la loro motivazione e la loro percezione di interazione con docenti e compagni di corso. Solo analizzando questi fattori si potrà predisporre una didattica inclusiva e, nel caso di future emergenze, si potrà dare ad ogni studente la possibilità di procedere nel proprio percorso di studi e di vita nel migliore dei modi.*

**Key words:** multiple correspondence analysis. online survey. student motivation. student satisfaction, k-means cluster analysis.

---

1

Isabella Morlini, University of Modena & Reggio Emilia; email: [isabella.morlini@unimore.it](mailto:isabella.morlini@unimore.it)

Laura Sartori, University of Bologna; email: [laura.sartori@unibo.it](mailto:laura.sartori@unibo.it)

## 1 Introduction

The health emergency brought about by COVID-19 has produced a radical and rapid change in university life. In March 2020, remote teaching became the rule almost overnight in all Italian universities, with new implications in the landscape of educational learning. Classes, examinations and laboratories were suddenly reorganized by a collective effort that benefited from previous experimentation and innovation in teaching methods like, for example, blended learning (see, e.g., Purnomo *et al.*, 2019). At the core of the challenge was not only the technical and remote delivery of all classes, but especially the array of tools and practices concerning engagement of the students and developing of their attitudes toward online learning (like project works, etc...) The importance of understanding students' motivation and engagement in an online environment is shown by the significant and recent amount of research on this topic (see, e.g., Bolliger & Halupa, 2018; Ellis & Bliuc, 2019; Ferrer *et al.*, 2020; Kahu & Nelson, 2018; Martin & Bolliger, 2018). The goal of this work is to reach an insight on the needs of the university students from a customer-oriented perspective and to analyse the socioeconomic and environmental determinants of their satisfaction and motivations in a remote teaching system. Using data collected with an online questionnaire administered to students enrolled in the University of Modena & Reggio Emilia and analysed using cluster and multiple correspondence analysis, we try, in particular, to give an answer to the following research questions:

- Is there an association between students' satisfaction for online experience and intrinsic motivations and attitude to distance learning?
- Do socioeconomic problems influence the distance learning experience?
- Which are the determinants for satisfaction and engagement in online education?
- Do gender and working status play a role in students' satisfaction?
- Do university related characteristics like the area of the course in which the student is enrolled and the course year, influence students' satisfaction?

The paper is organized as follows: Section 2 briefly illustrates the questionnaire and the sample, Section 3 reports main results and Section 4 gives concluding remarks.

## 2 Data collection

The online survey was implemented by means of Survey Monkey, with an individual link sent to 27,792 students. The questionnaire consists of 36 questions grouped into four sections: (i) general information on home trips caused by the emergency, living conditions and ongoing problems and changes; (ii) organization of study with respect to the teaching materials available, the timing and methods of the organization of study; (iii) distance learning, with the focus on attendance and satisfaction, specific difficulties, conditions of concentration and interest, aspects that were missing and those that were appreciated, open questions on strengths and weaknesses of distance learning and suggestions and proposals; (iv) information on

Effects of remote teaching on students' motivation and engagement: the case of UNIMORE internships and working conditions. For a detailed description of the questionnaire and of the respondents we refer to Russo *et al.* (2020). For some questions the answer is dichotomous (1=yes, 0=no) while for other questions the answer is ordinal with four categories. In this work, we have considered only closed-questions related to socioeconomic and material problems, student motivation and attitude to online learning, task orientation and engagement. To both reduce the number of parameters in the multivariate models and simplify output results, we have re-coded all polytomous variables in order to have only binary variables of the type presence/absence, substituting categories "not at all" and "rarely" with the code 0 and categories "enough" and "very much" with 1. Moreover, we have dichotomized the rate of the satisfaction with the global learning experience re-coding a rate from 1 to 5 into 0 (not satisfied) and a rate from 6 to 10 into 1 (satisfied). The average rate of complete answers is 19.2% (5,341 records). The participation rate varies widely by area and department. Considering the type of degree course, the year of enrolment and the credits matured, the response percentages in the sample are quite similar to the percentages of those enrolled students in the population. Considering the gender, we note an over-representation of females: the percentage in the sample is 65.6% while the percentage of females in the population is 52.5%.

### 3 Main Results

The first two factors of the multiple correspondence analysis (Greenacre & Blasius, 2006) performed over the 28 dichotomous variables considered in this work explain the 25.6% of the total variability, a percentage higher than the threshold  $0.95^{28}=23.8\%$ . The screeplot (not reported for economy of space) indicates an optimal number of latent factors equal to 2. Analysing the principal coordinates (see Table 1) we may evaluate the first factor as an index of high dissatisfaction with the global online experience, strong demotivation and disengagement. On the contrary, the second factor may be considered as an index of engagement and moderate dissatisfaction. The first factor loads strongly on variables related to the lack of attitude to online learning (inability to organize daily studying activities effectively, disorientation due to the many channels of communications, ...) and to the lack of task orientation (having hard time concentrating, inability to extricate among recordings, inability to take good notes, accumulation of lessons, ...). The second factor loads only on few variables related to demotivation and lack of attitudes and is not associated with task orientation. Among socioeconomic and material problems, variables loading on both factors are those related to the care of the family and, of course, those related to proper electronic devices and internet connection. Economic problems seem not to be related to dissatisfaction and amotivation.

Considering all previous variables except for the satisfaction with the global learning experience, and analysing the decomposition of the deviance within and between groups obtained with the k-means cluster analysis and the Euclidean distance, we choose a partition into 4 clusters as the best partition for the trade-off between number of groups and homogeneity inside the groups.

**Table 1:** Principal coordinates of the Multiple Correspondence Analysis

Variable number	Questions	F1	F2
	1 Satisfied with the global distance learning experience: NO	<b>0.811</b>	0.163
	Satisfied with the global distance learning experience: YES	<b>-0.320</b>	-0.064
Socioeconomic and material problems	2 Economic problems: NO	-0.066	-0.107
	Economic problems: YES	0.272	<b>0.441</b>
	3 Problems due to the condition of family members or friends: NO	-0.045	-0.051
	Problems due to the condition of family members or friends: YES	<b>0.475</b>	<b>0.541</b>
	4 No time to study due to taking care of family or live-in friends: NO	-0.069	-0.107
	No time to study due to taking care of family or live-in friends: YES	<b>0.586</b>	<b>0.912</b>
Motivation	5 No proper internet connection: NO	-0.090	-0.055
	No proper internet connection: YES	<b>0.539</b>	<b>0.330</b>
	6 No proper electron device for studying activities: NO	-0.031	-0.027
	No proper electron device for studying activities: YES	<b>0.731</b>	<b>0.636</b>
	7 Progressing equally on all subjects in the semester: NO	0.127	<b>0.237</b>
	Progressing equally on all subjects in the semester: YES	<b>-0.344</b>	<b>-0.642</b>
Lack of task orientation	8 Appreciating the chance of keeping abreast of all subjects more easily: NO	<b>0.483</b>	0.107
	Appreciating the chance of keeping abreast of all subjects more easily: YES	<b>-0.504</b>	-0.111
	9 Appreciating the possibility to pause the recordings and listening again: NO	<b>0.616</b>	<b>0.399</b>
	Appreciating the possibility to pause the recordings and listening again: Yes	-0.113	-0.074
	10 Appreciating the interactions among students through many channels: NO	0.094	-0.004
	Appreciating the interactions among students through many channels: YES	<b>-0.359</b>	0.015
Lack of engagement	11 Appreciating having the resources of the course at all time everywhere: NO	<b>0.578</b>	0.246
	Appreciating having the resources of the course at all time everywhere: YES	-0.214	-0.091
	12 NOT understanding how to organize the studying activities: NO	<b>-0.465</b>	-0.200
	NOT understanding how to organize the studying activities: YES	<b>0.575</b>	0.247
	13 Accumulating lessons create difficulties: NO	<b>-0.620</b>	-0.144
	Accumulating lessons create difficulties: YES	<b>0.475</b>	0.110
Lack of attitude to online learning	14 Don't know how to extricate among the different recordings: NO	<b>-0.390</b>	-0.130
	Don't know how to extricate among the different recordings: YES	<b>0.781</b>	0.260
	15 Having a hard time concentrating: NO	<b>-0.797</b>	-0.194
	Having a hard time concentrating: YES	<b>0.456</b>	0.111
	16 Lacking the will to study: NO	<b>-0.538</b>	-0.265
	Lacking the will to study: YES	<b>0.425</b>	0.210
Lack of engagement	17 Difficulties in following the lessons and tacking good notes: NO	<b>-0.385</b>	-0.050
	Difficulties in following the lessons and tacking good notes: YES	<b>0.772</b>	0.099
	18 The absence of involvement makes it difficult to stay focused: NO	<b>-0.626</b>	<b>0.378</b>
	The absence of involvement makes it difficult to stay focused: YES	<b>0.510</b>	<b>-0.308</b>
	19 The absence of interaction does not allow for enough explanations: NO	<b>-0.412</b>	<b>0.394</b>
	The absence of interaction does not allow for enough explanations: YES	<b>0.441</b>	<b>-0.422</b>
Lack of engagement	20 Lacking the possibility to ask for explanations: NO	<b>-0.339</b>	<b>0.429</b>
	Lacking the possibility to ask for explanations: YES	<b>0.442</b>	<b>-0.560</b>
	21 Lacking the chance to interact face to face: NO	<b>-0.716</b>	<b>0.880</b>
	Lacking the chance to interact face to face: YES	<b>0.221</b>	<b>-0.271</b>
	22 Lacking the stimuli given in class: NO	<b>-0.848</b>	<b>0.554</b>
	Lacking the stimuli given in class: Yes	<b>0.407</b>	<b>-0.266</b>
Lack of attitude to online learning	23 It fatigues spending lot of time in front of the screen: NO	<b>-0.672</b>	0.295
	It fatigues spending lot of time in front of the screen: YES	<b>0.306</b>	-0.134
	24 It takes longer to follow a recording lesson: NO	<b>-0.647</b>	<b>0.389</b>
	It takes longer to follow a recording lesson: YES	<b>0.322</b>	-0.193
	25 Distracted by other family of care needs at home: NO	<b>-0.389</b>	-0.267
	Distracted by other family of care needs at home: YES	<b>0.419</b>	0.288
	26 Many channels of communications disorient: NO	<b>-0.190</b>	-0.108
	Many channels of communications disorient: YES	<b>0.650</b>	<b>0.370</b>
Lack of attitude to online learning	27 Don't able to organize daily studying activities effectively: NO	<b>-0.648</b>	<b>-0.357</b>
	Don't able to organize daily studying activities effectively: YES	<b>0.558</b>	<b>0.307</b>
	28 Lacking the weekly schedule of lessons: NO	<b>-0.819</b>	<b>0.402</b>
	Lacking the weekly schedule of lessons: YES	<b>0.408</b>	-0.200

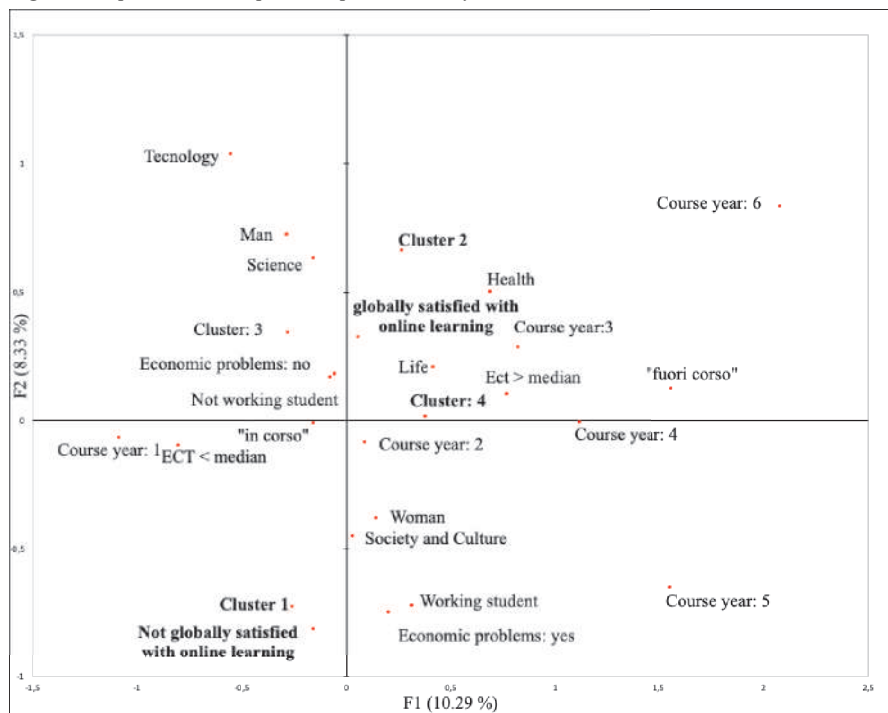


Table 2 reports the central coordinates in each cluster and Figure 1 the biplot of the multiple correspondence analysis on group membership, satisfaction and independent variables like gender, working status, course year, area of the course in which the student is enrolled (life, science, health, ...), credits acquired above or below or median. Cluster 1 identifies the group of students (31.3%) totally unsatisfied. These students are only characterized by lack of attitude, lack of engagement and orientation. Cluster 2 is the group of the globally satisfied students (21.6% of the total), mostly enrolled in the third year of a course in Health or Life and characterized by attitude to online learning, task orientation and self-engagement. Motivation seems not to discriminate between satisfied and unsatisfied students. Clusters 3 (24.1%) and 4 (23%) identify students with intermediate level of engagement and satisfaction. Socioeconomic problems do not discriminate groups.

**Table 2:** Central coordinates of the 4 clusters: in rows the clusters and in the columns the variables

Cl	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
2	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	1
4	0	0	0	0	0	0	1	1	0	1	1	0	1	1	0	0	0	0	0	1	1	1	0	1	0	1	1

**Figure 1:** Biplot of the multiple correspondence analysis



## 4 Conclusions

The different analyses show that there is a strong association between student satisfaction for online experience and attitude to distance learning, engagement and task orientation. Student intrinsic motivations seem to play a less relevant role. Socioeconomic and material problems influence the distance learning experience but are not a determinant for satisfaction, engagement and motivation. Gender, working status and factors like the area of the course in which the student is enrolled, the course year, the level of credits acquired and the status “in corso” or “fuori corso” (peculiar of the Italian university system) are not linked to satisfaction, motivation and self-engagement. The study confirms that valid predictors of student perception and engagement are, among other factors related to task orientation and attitude, the ability to organize the daily activities effectively, the ability to stay focused and not be distracted, the ability to extricate among the recorded lessons and not to be disoriented by the different channels of communications. Students without attitudes toward online learning are incapable of recognising positive aspects like the possibility to pause the recordings and listening again, the possibility of having all the resources of the course at all time and everywhere and the possibility of keeping abreast of all subjects. Results suggest that attitudes and tasks toward online can be gained across time: students enrolled in the third year of courses that are more experienced with university courses and have probably experimented some blended learning before the emergency period, are more likely to be satisfied with remote didactics and are more likely to be engaged.

## References

1. Bolliger, D.U. and Halupa, C.: Online student perceptions of engagement, transactional distance and outcomes. *Distance Education*, 39(3) 299–316 (2018)
2. Ellis, R. and Bliuc, A.: Exploring new elements of the student approaches to learning framework: the role of online learning technologies in student learning. *Active Learning in Higher Education*, 20(1) 11–24 (2019)
3. Ferrer, J., Ringer, A., Saville, K., Parris, M.A. and Kashi, K.: Student’s motivation and engagement on higher education: the importance of attitude to online learning. *Higher Education* (2020) doi: 10.1007/s10734-020-00657-5
4. Greenacre, M. and Blasius J.: *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, NY (2006)
5. Kahu, E. R. and Nelson, K.: Student engagement in the educational interface: understanding the mechanisms of student success. *Higher Education Research & Development*, 37(1) 58–71 (2018)
6. Martin, F. and Bolliger, D.H.: Engagement matters: student perceptions on the importance of engagement strategies in the online learning environment. *Online Learning*, 22(1) 205–222 (2018)
7. Purnomo, A., Kurniawan, B. and Aristin, N.: Motivation to learn independently through blended learning. *Advances in Social Science, Education and Humanities Research*, 330, 261–264 (2019)
8. Russo, M., Alboni, F., Colombini, S., Morlini, I., Pavone, P. and Sartori, L.: Covid-19 e Studenti Unimore: come l’emergenza cambia lo studio e l’esperienza universitaria. *Demb Working Paper Series*, n.173 (2020)
9. Ryan, R.M. and Deci, E.L.: Intrinsic and extrinsic motivation from a self-determination theory perspective: definitions, theory, practices and future directions. *Contemporary Educational Psychology*, 61 (2020)

# A random effects model for the impact of remote teaching on university students' performance

## *Un modello a effetti casuali per l'impatto della didattica a distanza sui risultati degli studenti universitari*

Silvia Bacci and Bruno Bertaccini and Simone Del Sarto and Leonardo Grilli and Carla Rampichini

**Abstract** The COVID-19 pandemic had a relevant impact in all aspects of the social life. In Italy, in March 2020 schools and universities suspended the activities in presence and suddenly moved to remote teaching. In this contribution we aim at analysing the effects of remote teaching on university students' careers. To this end, we consider the differences in gained credits by the freshmen cohorts of academic years 2018/2019 and 2019/2020, enrolled in two bachelor degree courses (Business Administration and Psychology) at the University of Florence. Indeed, both cohorts regularly attended courses during the first semester, while only freshmen from academic year 2019/2020 experimented remote teaching during the second semester. As outcome, we consider the proportion of gained credits in the semester over the expected credits, thus the data have a panel structure with two observations per student. We estimate the impact of remote teaching through a random effects linear model. As a main result of our preliminary analysis, we detect a significant and negative effect of remote teaching on the career progressions of academic students.

**Abstract** *La pandemia da COVID-19 ha avuto un impatto rilevante su tutti gli aspetti della vita sociale. In Italia, nel marzo 2020 le scuole e le università hanno sospeso le attività in presenza e hanno improvvisamente iniziato la didattica a distanza. In questo contributo ci proponiamo di analizzare gli effetti della didattica*

---

Silvia Bacci

Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence  
e-mail: [silvia.bacci@unifi.it](mailto:silvia.bacci@unifi.it)

Bruno Bertaccini

Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence  
e-mail: [bruno.bertaccini@unifi.it](mailto:bruno.bertaccini@unifi.it)

Simone Del Sarto

Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence  
e-mail: [simone.delsarto@unifi.it](mailto:simone.delsarto@unifi.it)

Leonardo Grilli

Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence  
e-mail: [leonardo.grilli@unifi.it](mailto:leonardo.grilli@unifi.it)

Carla Rampichini

Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence  
e-mail: [carla.rampichini@unifi.it](mailto:carla.rampichini@unifi.it)

*a distanza sulle carriere degli studenti universitari. A tal fine, consideriamo le differenze nei crediti formativi acquisiti dalle coorti di immatricolati negli anni accademici 2018/2019 e 2019/2020, iscritti a due corsi di laurea triennale (Economia Aziendale e Psicologia) presso l'Università di Firenze. Infatti, entrambe le coorti hanno frequentato regolarmente i corsi durante il primo semestre, mentre i soli immatricolati della coorte 2019/2020 hanno sperimentato la didattica a distanza durante il secondo semestre. Come variabile di risposta consideriamo la proporzione di crediti acquisiti sui crediti attesi, per cui i dati hanno una struttura panel con due osservazioni per studente. Stimiamo l'impatto della didattica a distanza tramite un modello lineare a effetti casuali. Come risultato principale della nostra analisi preliminare, abbiamo individuato un effetto significativamente negativo della didattica a distanza sulle progressioni di carriera degli studenti universitari.*

**Key words:** COVID-19, random effects model, repeated measures

## 1 Introduction

The COVID-19 pandemic manifested in Italy since February 2020, leading to disruptive effects on many aspects of people's social life. The suspension of the teaching activities in schools and universities was the first containment measure adopted by the Government to deal with the spread of the virus. Remote teaching has been the solution implemented by schools and universities to limit the damages to students' learning. In this contribution we aim at analysing the effects of remote teaching due to COVID-19 pandemic on the university students' careers. There is a growing literature on this topic (e.g., [2]), but we are not aware of systematic studies on the impact in terms of gained credits.

We compare the cohorts of freshmen of academic years 2018/2019 and 2019/2020 enrolled in the bachelor degree courses in Business Administration and Psychology at the University of Florence. Teaching activities were regular for the cohort 2018 along all its first academic year (first semester September-December 2018 and second semester February-June 2019), whereas the cohort 2019 attended regular lessons only during the first semester (September-December 2019) and experimented the remote teaching during second semester (February-June 2020). To evaluate the impact of remote teaching, we compare the number of credits (ECTS) gained during the second semester by these two cohorts, using information from the first semester to remove a possible "cohort effect" not depending on the remote teaching.

The remaining part of the paper is organised as follows. In Section 2 we describe the data and in Section 3 we illustrate results obtained by preliminary analyses. Final remarks are reported in Section 4.

**Table 1** Descriptive statistics of freshmen by degree courses (BA: Business Administration; PSY: Psychology) and year of enrolment (2018 and 2019): number of enrolled students (*N*), % of female, average high school (HS) grade (standard deviation within parentheses) and type of high school

	BA			PSY		
	2018	2019	Total	2018	2019	Total
<i>N</i>	640	668	1,308	427	429	856
% female	42.3	42.4	42.4	72.8	79.0	75.9
HS grade	78.1 (11.6)	76.6 (11.4)	77.3 (11.5)	80.6 (10.7)	79.2 (11.2)	79.9 (11.0)
<i>Type of HS (%)</i>						
Scientific	30.8	33.1	32.0	36.3	32.6	34.5
Technical	39.8	36.2	38.0	8.4	16.1	12.3
Vocational	8.0	6.4	7.2	4.7	1.9	3.3
Humanities	18.6	13.6	16.1	42.9	38.7	40.8
Other	2.8	10.7	6.7	7.7	10.7	9.2

## 2 Data

We consider data obtained from the administrative archive on students' careers, which includes some background characteristics, such as sex, high school (HS) type and grade, and information on passed exams. Specifically, we focus on two cohorts of freshmen enrolled in academic years 2018/2019 and 2019/2020 in the bachelor degree courses in Business Administration (BA) and Psychology (PSY) at the University of Florence. The dataset includes 2,164 students (about 60% in BA and 40% in PSY) whose characteristics are summarised in Table 1.

By inspecting the table, we notice a prevalence of male students in BA (57.6%), whereas female students are definitely more frequent in PSY (75.9%). Moreover, HS grade is on average slightly greater in freshmen of PSY with respect to their peers of BA. As far as the composition in terms of HS type is concerned, almost three students of PSY out of four come from scientific and humanities high schools (i.e., "licei"), while in the BA degree program we can observe a predominance of students from scientific and technical schools. Within each degree course, the two cohorts have similar characteristics, though in PSY we notice an increase of the female share (+6.2% in 2019) and the HS type composition, where a decrease in students from humanities and scientific high schools occurs in favour of technical schools or other type.

In order to study the effect of remote teaching on students' performance, we consider the proportion of credits (ECTS) gained in each semester out of the total of planned credits envisaged by the degree course. We can disentangle this effect by comparing students' performance in the two cohorts. In fact, for the exams taken in the first semester session, none of the two cohorts have experienced remote teaching, while a potential effect can be highlighted for exams taken by the cohort 2019 during the second semester session (i.e., in June, July and September 2020).

**Table 2** Average proportion of credits gained by students of BA and PSY by semester and cohort

Semester	BA		PSY	
	2018	2019	2018	2019
First	0.395	0.513	0.593	0.623
Second	0.564	0.607	0.856	0.810

The first year degree course in BA envisages three 9-credit exams in both semesters (hence 27 credits in each one), while PSY freshmen have to face three exams in the first semester (27 credits) and four in the second one (30 credits). Students' performance are summarised in Table 2 in terms of average proportion of gained credits out of the total credits envisaged by the degree course. Note that these proportions may theoretically be higher than one, whenever a student completes the exams of the first year and takes in advance exams of the second year (in practice, this is a rare instance).

Looking at Table 2, we observe that the performance of students in the second semester is higher than their performance in the first semester, in particular for PSY. Moreover, in the first semester the cohort 2019 reports a better performance with respect to cohort 2018. This is especially true for freshmen of BA, where the proportion of gained credits raises from 0.395 to 0.513. A likely reason for this trend is a structural change in the study plan of BA: in the first academic year, the cohort 2018 had Private Law, whereas the cohort 2019 had Public Law.

As concerns the second semester, results differ for the two degree courses. Credits gained by BA students tend to increase (proportions from 0.564 to 0.607 on average), whereas credits gained by PSY students decrease (proportions from 0.856 to 0.810 on average).

### 3 Estimating the impact of remote teaching

To estimate the impact of remote teaching, we fit a linear random effects model for repeated measures [3], separately for students of the two degree courses (BA and PSY). The outcome  $Y_{it}$  is the proportion of credits gained by student  $i$  in semester  $t$  over the expected credits, with  $i = 1, \dots, n_d$  (where index  $d$  refers to the degree course) and  $t = 0$  for the first semester and  $t = 1$  for the second semester. The model is formulated as

$$Y_{it} = \alpha + \mathbf{x}_i' \boldsymbol{\beta} + \gamma_1 c_i + \gamma_2 t + \delta(c_i \times t) + u_i + \varepsilon_{it},$$

where  $\mathbf{x}_i$  is a vector of student characteristics (Female, HS grade, and HS type),  $c_i$  is a dummy variable for the cohort of student  $i$  (reference: 2018) and  $t$  is a dummy variable for the semester (reference: first semester). The random effect  $u_i$  for student  $i$  collects unobserved factors at student level and is assumed to follow a Normal distribution with mean 0 and variance  $\sigma_u^2$ , whereas the residual error  $\varepsilon_{it}$  (independ-

dent of  $u_i$ ) is assumed to follow a Normal distribution with mean 0 and variance  $\sigma_\varepsilon^2$ . The parameter of main interest is  $\delta$ , namely the coefficient of the interaction between cohort  $c_i$  and semester  $t$ , which represents the effect of remote teaching after controlling for the structural differences in the cohorts and the semesters. This is a difference-in-differences approach [11] where the second semester is the post-treatment period and the cohort 2018 is the control group.

Results of fitted models are displayed in Table 3(a) for BA and in Table 3(b) for PSY. The HS grade and the HS type have similar effects in the two degree courses: students with higher HS grades and, most of all, coming from a scientific HS progress in the academic career faster than their peers. On the opposite, differences between females and males depend on the degree course: no significant difference is observed for BA, whereas male students of PSY perform significantly worse than their female colleagues.

As for the cohort effect  $\gamma_1$ , the two cohorts of BA students are significantly different, with students of cohort 2019 performing better than colleagues of cohort 2018. This result confirms the presence of structural differences between the programs of BA in the two academic years. On the opposite, no significant difference is detected between the two cohorts of PSY freshmen.

For what concerns the semester effect  $\gamma_2$ , students of the cohort 2018 perform significantly better in the second semester with respect to the first one (regression coefficient equals 0.169 for BA and 0.264 for PSY).

Finally, the estimated effect of remote teaching  $\delta$ , associated to the interaction between semester and cohort, is nearly equal for the two degree courses: the experience of remote teaching caused a statistically significant slowdown in the students' career progression, with estimated regression coefficient equal to  $-0.075$  for BA and  $-0.077$  for PSY. These values are changes in the proportion of gained credits over expected credits: in absolute terms, they correspond to a reduction of about 2 credits.

## 4 Conclusions

The preliminary results point out a negative impact of remote teaching on the productivity of students in Business Administration and Psychology. The analyses will be further developed in order to take into account the following issues. First, we intend to formulate a model that accounts in a suitable way the specific nature of the response variable, that is, a proportion with possible values greater than one and with excess of zeros (i.e., students that do not take any exam). Second, we will extend the analysis to other bachelor and master degree courses to investigate the existence of differences among academic schools in the implementation of remote teaching. In order to investigate the impact of remote teaching on specific exams, a promising route is to formulate a multilevel model with pseudo-panel data, with students as first-level units and exams as second-level units observed for two academic years. Third, we intend to incorporate teaching evaluations by students in the analysis. Since questionnaires are anonymous, the evaluations have to be aggregated at course level.

**Table 3** Regression model estimates for students of BA (a) and PSY (b). Note: the 95% confidence intervals (95% CI) for variance components are computed with 500 bootstrap replications

(a)			
	Estimate	Std. Error	<i>p</i> -value
Intercept	0.299	0.028	< 0.001 ***
Male (ref. Female)	-0.003	0.019	0.889
HS grade	0.013	0.001	< 0.001 ***
<i>HS type (ref. Scientific)</i>			
Humanities	-0.191	0.028	< 0.001 ***
Vocational	-0.408	0.037	< 0.001 ***
Technical	-0.177	0.022	< 0.001 ***
Other	-0.147	0.038	< 0.001 ***
Cohort 2018 (ref. 2019)	0.127	0.021	< 0.001 ***
Second Semester (ref. First)	0.169	0.015	< 0.001 ***
Interaction (cohort, semester)	-0.075	0.021	< 0.001 ***
$\sigma_u^2$	0.069 (95% CI: 0.061–0.079)		
$\sigma_\varepsilon^2$	0.071 (95% CI: 0.065–0.077)		

(b)			
	Estimate	Std. Error	<i>p</i> -value
Intercept	0.513	0.034	< 0.001 ***
Male (ref. Female)	-0.108	0.028	< 0.001 ***
HS grade	0.008	0.001	< 0.001 ***
<i>HS type (ref. Scientific)</i>			
Humanities	-0.127	0.042	0.003 **
Vocational	-0.083	0.027	0.002 **
Technical	-0.249	0.066	< 0.001 ***
Other	-0.082	0.038	0.034 *
Cohort 2019 (ref. 2018)	0.036	0.028	0.202
Second Semester (ref. First)	0.264	0.022	< 0.001 ***
Interaction (cohort, semester)	-0.077	0.031	0.013 *
$\sigma_u^2$	0.059 (95% CI: 0.049–0.072)		
$\sigma_\varepsilon^2$	0.102 (95% CI: 0.092–0.112)		

## References

1. Abadie A.: Difference-in-Difference Estimators. In: Palgrave Macmillan (eds.) The New Palgrave Dictionary of Economics. Palgrave Macmillan, London (2008)
2. Aucejo, E. M., French, J., Ugalde Araya, M. P., Zafar, B.: The Impact of COVID-19 on Student Experiences and Expectations: Evidence from a Survey. *J. Public Econ.* **191**, 104271 (2020)
3. Hedeker, D., Gibbons, R.D.: Longitudinal Data Analysis. Second edition. Wiley (2021)



# Multinomial semiparametric mixed-effects model for profiling engineering university students

## *Modello multinomiale a effetti misti semiparametrico per la profilazione di studenti universitari di ingegneria*

Chiara Masci, Francesca Ieva and Anna Maria Paganoni

**Abstract** Many applicative studies deal with multinomial responses and hierarchical data. In this study, we analyse Politecnico di Milano data with the aim of profiling students, modelling their probabilities of belonging to different categories, considering their nested structure within engineering degree programmes. To this end, we propose a semiparametric mixed-effects models dealing with a multinomial response and an EM algorithm to implement it. By assuming the random effects to follow a multivariate discrete distribution with an a priori unknown number of support points, that is allowed to differ across response categories, we identify a classification of degree programmes, standing on their effects on different types of student career.

**Abstract** Molti studi applicati trattano variabili multinomiali e dati con struttura gerarchica. In questo studio, analizziamo i dati del Politecnico di Milano al fine di profilare gli studenti, modellizzando la loro probabilità di appartenere a diverse categorie, considerando il loro corso di studi. A tal fine, sviluppiamo un modello semiparametrico a effetti misti per una risposta multinomiale e un algoritmo EM per implementarlo. Assumendo che gli effetti casuali abbiano una distribuzione discreta con un numero di masse non noto a priori e che differisce tra categorie della risposta, identifichiamo una classificazione dei corsi di studio, in base al loro effetto sulla carriera degli studenti.

**Key words:** Multinomial model, Mixed-effects models, Student Dropout, Semiparametric statistics.

---

Chiara Masci e-mail: chiara.masci@polimi.it · Francesca Ieva e-mail: francesca.ieva@polimi.it · Anna Maria Paganoni e-mail: anna.paganoni@polimi.it

MOX - Modelling and Scientific Computing - Politecnico di Milano, via Bonardi 9, 20133 Milano (IT)

## 1 Introduction

The Italian Higher Education (HE) system measures a high level of dropout, with many students abandoning their studies during the Bachelor. Many studies aim at individuating personal features of students who are more likely to drop out in order to partially prevent the phenomenon [3, 4]. Politecnico di Milano (PoliMI) dropout rate in engineering is around 30%, with the majority of students dropping out during the first year [5]. Concluded careers of students can be classified as *graduate*, *early dropout* (i.e. careers concluded with a dropout within the first three semesters since the enrolment) and *late dropout* (i.e. careers concluded with a dropout after more than three semesters since the enrolment). PoliMI offers about 20 different engineering degree programmes and students are nested within them. Degree programmes have heterogeneous internal dynamics, students characteristics, study plans and these aspects might lead to different dropout rates and motivations.

In this study, we aim at profiling engineering students, modelling their probabilities of belonging to different categories, standing on their personal and early career information and considering the degree programme the student is attending. In particular, we are also interested in identifying latent subpopulations of degree programmes standing on their effects on different types of student career. To this end, we develop a semiparametric multinomial mixed-effects model, whose random effects follow a discrete distribution with an unknown number of mass points. This modelling allows to identify a latent structure at the highest level of the hierarchy, where groups (i.e. degree programmes) are clustered into subpopulations, standing on their effect on the outcome, i.e. the probability of students to belong to different categories of the multinomial response variable.

Parametric mixed-effects models for a multinomial response present many issues relative to the multivariate integration over the random effects distribution [6, 11]. In this perspective, the advantage introduced by the proposed modelling is twofold: (i) the former is that, by assuming a discrete distribution at the highest level of the hierarchy, we avoid the integration issues relative to the continuous distribution of a parametric approach; (ii) the latter is that this assumption allows to identify a latent structure within the highest level of the hierarchy, i.e. the presence of subpopulations among groups. Moreover, this modelling allows to investigate how the latent structure at the highest level of the hierarchy does change across categories, with respect to the baseline. To estimate the semiparametric model parameters, we propose an Expectation-Maximization (EM) algorithm that alternates the estimates of fixed effects and random effects until the convergence is reached [2, 7].

## 2 Methodology

Let consider a multinomial logistic regression model for nested data with a two-level hierarchy [1, 6], where each observation  $j$ , for  $j = 1, \dots, n_i$ , is nested within a group  $i$ , for  $i = 1, \dots, I$ . Let  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$  be the  $n_i$ -dimensional response vector

MSPEM algorithm for profiling university students

for observations within the  $i$ -th group.  $P(Y_{ij} = k) = \pi_{ijk}$ , for  $k = 1, \dots, K$ , where  $K$  is the total number of categories assumed by  $Y$ . Mixed-effects multinomial models assume that the probability that  $Y_{ij} = k$ , i.e.  $\pi_{ijk}$ , is given by

$$\pi_{ijk} = P(Y_{ij} = k) = \frac{\exp(\eta_{ijk})}{1 + \sum_{k=2}^K \exp(\eta_{ijk})} \quad \text{for } k = 1, \dots, K, \quad (1)$$

where  $\eta_{ijk} = \mathbf{x}'_{ij}\alpha_k + \mathbf{z}'_{ij}\delta_{ik}$  is the linear predictor (we assume  $\eta_{ijk} = 0$  for  $k = 1$ ).  $\mathbf{x}_{ij}$  is the  $p \times 1$  covariates vector (includes a 1 for the intercept) of the fixed effects,  $\alpha_k$  is the  $p \times 1$  vector of regression parameters of the fixed effects,  $\mathbf{z}_{ij}$  is the  $q \times 1$  covariates vector of the random effects (includes a 1 for the intercept) and  $\delta_{ik}$  is the  $q \times 1$  vector of regression parameters of the random effects.

Considering  $\mathbf{A} = (\alpha_2, \dots, \alpha_K)$  and  $\Delta_i = (\delta_{i2}, \dots, \delta_{iK})$  and assuming that  $Y_{ij}$  and  $Y_{ij'}$  are independent for  $j \neq j'$ , the conditional distribution of  $Y_i$  takes the following form:

$$p(\mathbf{Y}_i | \mathbf{A}, \Delta_i) = \prod_{j=1}^{n_i} \prod_{k=1}^K \left( \frac{\exp(\eta_{ijk})}{1 + \sum_{l=2}^K \exp(\eta_{ijl})} \right)^{\mathbf{1}_{\{Y_{ij}=k\}}}. \quad (2)$$

In a parametric framework,  $\delta_{ik}$  are usually assumed to follow a multivariate normal distribution  $\mathcal{N}(\mathbf{0}, \Omega_k)$  [6].

Following a semiparametric framework, we assume the random effects coefficients to follow a discrete distribution with an a priori unknown number of support points:

$$\eta_{ijk} = \mathbf{x}'_{ij}\alpha_k + \mathbf{z}'_{ij}\mathbf{b}_{m_k k} \quad m_k = 1, \dots, M_k, \quad k = 2, \dots, K, \quad (3)$$

where  $M_k$  is the total number of support points of the discrete distribution of  $\mathbf{b}$  relative to the  $k$ -th category, for  $k = 2, \dots, K$ . The random effects distribution relative to each category  $k$ , for  $k = 2, \dots, K$ , can be expressed as a set of points  $(\mathbf{b}_{1k}, \dots, \mathbf{b}_{M_k k})$ , where  $M_k \leq I$  and  $\mathbf{b}_{m_k k} \in \mathcal{R}^q$  for  $m_k = 1, \dots, M_k$ , and a set of weights  $(w_{1k}, \dots, w_{M_k k})$ , where  $\sum_{m_k=1}^{M_k} w_{m_k k} = 1$  and  $w_{m_k k} \geq 0$ . Under these assumptions, the marginal likelihood can be obtained as a weighted sum of the likelihood of  $\mathbf{Y}_i$  conditioned to all the possible combinations, that are  $M_{tot} = \prod_{k=2}^K M_k$ , of the values of the  $(K - 1)$  discrete distributions of random effects:

$$h(\mathbf{Y}_i | A) = \sum_{m=1}^{M_{tot}} w_m p(\mathbf{Y}_i | A, \mathbf{B}_m), \quad (4)$$

where  $w_m$  is the weight relative to the  $m$ -th combination of the  $(K - 1)$  weights relative to the  $(K - 1)$  contrasts and, analogously,  $\mathbf{B}_m$  is the  $m$ -th combination of the random effects coefficients relative to the  $(K - 1)$  contrasts. We propose an

Expectation-Maximization algorithm (MSPEM algorithm) for the joint estimations of  $\alpha_k$ ,  $(\mathbf{b}_{1k}, \dots, \mathbf{b}_{M_k k})$  and  $(w_{1k}, \dots, w_{M_k k})$ , for  $k = 2, \dots, K$ , which is performed through the maximization of the likelihood, mixture by the discrete distribution of the random effects. During the iterations, we fix a tuning parameter  $D_k$ , for  $k = 2, \dots, K$ , and if two random effects distributions mass points are closer than  $D_k$ , they collapse to a unique point (further details in [9, 10]). In particular, we follow the procedure shown in [10]: at each iteration, given the conditional weights matrix, we alternate the estimation of fixed and random coefficients until convergence.

### 3 Simulation study

In order to reproduce the case study setting, we propose a simulation study to test the MSPEM performance considering a 3-categories response, a random intercept and two fixed-effects covariates<sup>1</sup>. We consider  $I = 100$  groups of data, where each group contains 200 observations and we induce the presence of three subpopulations regarding category  $k = 2$ , i.e.  $M_2 = 3$ , and two subpopulations regarding category  $k = 3$ , i.e.  $M_3 = 2$ . The linear predictor  $\eta_{ik} = (\eta_{i1k}, \dots, \eta_{i200k}) = \alpha_{1k}\mathbf{x}_{1i} + \alpha_{2k}\mathbf{x}_{2i} + \delta_{ik}$  is generated as follows:

$$\eta_{i2} = \begin{cases} +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 7 & i = 1, \dots, 30 \\ +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 4 & i = 31, \dots, 60 \\ +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 2 & i = 61, \dots, 100 \end{cases}$$

$$\eta_{i3} = \begin{cases} -2\mathbf{x}_{1i} + 2\mathbf{x}_{2i} - 5 & i = 1, \dots, 60 \\ -2\mathbf{x}_{1i} + 2\mathbf{x}_{2i} - 2 & i = 61, \dots, 100 \end{cases} \quad (5)$$

where variables  $\mathbf{x}_1$ ,  $\mathbf{x}_2$  and  $\mathbf{z}_1$  follow a distribution  $\mathcal{N}(0, 1)$ .

We perform 100 runs of the MSPEM algorithm, considering  $D_k = 1$  for  $k = 2, 3$ . In 94 runs out of 100 the algorithm identifies  $M_2 = 3$  and  $M_3 = 2$ ; in 91 runs out of these 94 runs, the algorithm correctly classifies groups into the identified subpopulations. Simulation results are shown in Table 1.

### 4 University student dropout across engineering degree programmes

We apply the MSPEM algorithm to data about PoliMI students, focusing on the concluded careers of students enrolled in an engineering programme of PoliMI in the a. y. between 2010/2011 and 2015/2016. The dataset considers 18,604 concluded careers of students nested within 19 engineering degree programmes. The

<sup>1</sup> Further simulations can be found in [9].

**Table 1** Fixed and random effects coefficients estimated by MSPEM algorithm in the simulation. Estimates are reported in terms of mean  $\pm$  sd, computed on the 100 runs of the simulation study for the fixed effects coefficients and on the runs in which the algorithm identifies  $M_2 = 3$  and  $M_3 = 2$  for the random effects ones.

	$\hat{\alpha}_{1k}$	$\hat{\alpha}_{2k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{12} = 4.066 \pm 0.080$	$\hat{\alpha}_{22} = -3.061 \pm 0.060$	$\hat{b}_{12} = -7.119 \pm 0.152$	$\hat{w}_{12} = 0.300$
			$\hat{b}_{22} = -4.096 \pm 0.091$	$\hat{w}_{22} = 0.300$
			$\hat{b}_{32} = -2.079 \pm 0.068$	$\hat{w}_{32} = 0.400$
k=3	$\hat{\alpha}_{13} = -2.073 \pm 0.041$	$\hat{\alpha}_{23} = 2.062 \pm 0.044$	$\hat{b}_{13} = -5.123 \pm 0.094$	$\hat{w}_{13} = 0.599$
			$\hat{b}_{23} = -2.092 \pm 0.038$	$\hat{w}_{23} = 0.401$

variable `Status` represents the multinomial response with 3 categories: *Graduate* (the reference), *Early dropout* and *Late dropout*. We consider the degree programme (variable `DegProg`, 19-levels factor) as grouping factor and two fixed-effects covariates: the gender of the student (binary variable `Gender` - Male=0, Female=1) and the number of credits obtained at the first semester of the first year of career (continuous variable `TotalCredits1.1`).

We run the MSPEM algorithm with  $D_k = 0.3^2$ , for  $k = 2, 3$ . The algorithm converges in 7 iterations and it identifies four subpopulations for both  $k = 2$  (Early Dropout) and  $k = 3$  (Late Dropout). Table 2 reports the estimated model parameters. The random intercepts associated to the four subpopulations, for each  $k$ , with their weights, are increasingly ordered. For both Early and Late Dropout, the 19 degree programmes are clustered within the identified subpopulations. The subpopulations identified suggest that the degree programmes can be divided into 4 groups standing on their effect on their student early and late dropout probability. In terms of predictive performance, the method classifies 74% of students as Graduate, 11% as Late dropout and 15% as Early dropout, with an error rate of 19%.

## 5 Conclusions

The MSPEM algorithm can be seen as an in-built clustering tool, where the identified subpopulations represent an alternative to the ranking provided by classical parametric mixed-effects models. In the general context of educational data, lower education students are nested within classes and schools, whose cardinality is often very high. In this perspective, identifying subpopulations of classes/schools instead of a ranking of hundreds or thousands of observations, whose estimates are sometimes so closed to be indistinguishable, might be easier and more effective. This

<sup>2</sup>  $D_k = 0.3$  has been chosen by measuring the uncertainty of classification of the method. By increasing (decreasing) the value of  $D_k$ , we obtain a lower (higher) number of subpopulations. The entropy of the conditional weights matrix is a good driver for the best choice of  $D_k$  (more details in [9]).

**Table 2** Fixed and random effects coefficients estimated by MSPeM algorithm for student dropout prediction.

	$\hat{\alpha}_{1k}$ (Gender)	$\hat{\alpha}_{2k}$ (TotalCredits1.1)	$\hat{b}_{m_k k}$ (random intercept DegProg)	$\hat{w}_{m_k k}$ (weight)
k=2	$\hat{\alpha}_{12} = -0.153$	$\hat{\alpha}_{22} = -2.704$	$\hat{b}_{12} = -2.841$	$\hat{w}_{12} = 0.482$
			$\hat{b}_{22} = -2.423$	$\hat{w}_{22} = 0.272$
			$\hat{b}_{32} = -2.096$	$\hat{w}_{32} = 0.193$
			$\hat{b}_{42} = -1.586$	$\hat{w}_{42} = 0.053$
k=3	$\hat{\alpha}_{13} = -0.685$	$\hat{\alpha}_{23} = -1.899$	$\hat{b}_{13} = -2.152$	$\hat{w}_{13} = 0.210$
			$\hat{b}_{23} = -1.733$	$\hat{w}_{23} = 0.421$
			$\hat{b}_{33} = -1.219$	$\hat{w}_{33} = 0.262$
			$\hat{b}_{43} = -0.880$	$\hat{w}_{43} = 0.107$

work enters in the literature about mixed-effects models with discrete random effects [2, 8, 10], proposing a novel method that deals with multinomial responses.

## References

- [1] A. Agresti. *An introduction to categorical data analysis*. Wiley, 2018.
- [2] M. Aitkin. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55(1):117–128, 1999.
- [3] O. Aljohani. A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher education studies*, 6(2):1–18, 2016.
- [4] F. Belloc, A. Maruotti, and L. Petrella. How individual characteristics affect university students drop-out: a semiparametric mixed-effects model for an italian case study. *Journal of applied Statistics*, 38(10):2225–2239, 2011.
- [5] M. Cannistrà, C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni. Not the magic algorithm: modelling and early-predicting students dropout through machine learning and multilevel approach. *MOX-report n. 41/2020*, 2020.
- [6] J. De Leeuw, E. Meijer, and H. Goldstein. *Handbook of multilevel analysis*. Springer, 2008.
- [7] J. Hartzel, A. Agresti, and B. Caffo. Multinomial logit random effects models. *Statistical Modelling*, 1(2):81–102, 2001.
- [8] J. S. Hartzel. Random effects models for nominal and ordinal data. 2000.
- [9] C. Masci, F. Ieva, and A. M. Paganoni. Semiparametric multinomial mixed-effects models: a university students profiling tool. *MOX-report num*, 2020.
- [10] C. Masci, A. M. Paganoni, and F. Ieva. Semiparametric mixed effects models for unsupervised classification of italian schools. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4):1313–1342, 2019.
- [11] G. Tutz and W. Hennevoogl. Random effects in ordinal regression models. *Computational Statistics & Data Analysis*, 22(5):537–557, 1996.

# **Evaluating Italian universities: ANVUR periodic accreditation judgment versus international rankings**

## ***La valutazione delle università italiane: il giudizio di accreditamento periodico ANVUR a confronto con i ranking internazionali***

Angela Maria D'Uggento, Nunziata Ribecco, Vito Ricci

**Abstract** Higher education ranking systems (HERSs) were developed with the goal of providing accurate information to allow global comparisons of the quality of institutions. In the last years, HERSs have widened their stakeholders and, being considered as a means for assessing the overall performance of a university, they can affect the perceived quality and the related reputation. World university rankings are one of the main factors inspiring prospective student choices and scholar mobility across domestic system, and especially abroad. Moreover, rankings can provide a steer for public policies, decision-makers and funding agencies. In this paper, a comparison among the AVA Periodic Accreditation final judgements and the results of the comparable international rankings including Italian universities is proposed.

**Abstract** *Inizialmente, i ranking internazionali sulle università (HERs) avevano l'obiettivo di fornire informazioni dettagliate per consentire confronti sulla qualità delle performance istituzionali. Negli ultimi anni, gli HERs hanno ampliato la platea degli stakeholder e, se considerate un mezzo per valutare la performance complessiva di ateneo, possono incidere sulla qualità percepita e sulla relativa reputazione, divenendo uno dei principali fattori che ispirano le scelte delle matricole e la loro mobilità all'estero. Inoltre, le classifiche possono rappresentare una guida per le politiche pubbliche, la governance universitaria e le agenzie di finanziamento. Il contributo*

---

Angela Maria D'Uggento, Department of Economics and Finance, University of Bari Aldo Moro, Italy; [angelamaria.duggento@uniba.it](mailto:angelamaria.duggento@uniba.it); corresponding author.

Nunziata Ribecco, Department of Economics and Finance, University of Bari Aldo Moro, Italy; [nunziata.ribecco@uniba.it](mailto:nunziata.ribecco@uniba.it).

Vito Ricci, Statistical analysis unit, University of Bari Aldo Moro, Italy; email: [vito.ricci@uniba.it](mailto:vito.ricci@uniba.it).

Angela Maria D'Uggento, Nunziata Ribecco and Vito Ricci  
*propone un confronto tra i giudizi finali di Accredimento Periodico-AVA e i risultati delle classifiche internazionali comparabili che includono le università italiane.*

**Key words:** world university rankings, ranking indicators, AVA periodic assessment judgement

## 1 Introduction

The evaluation of universities has gained a great impact on public opinion, research institutions and stakeholders, especially prospective students.

The stakeholders can benefit from a wide offer of university rankings, some of them specialized on research, others covering the overall performance, including also teaching, services and third mission activities.

Many scholars debate on the usefulness of rankings. Some of them state that the synthesis of much information into a single score used to rank such complex systems as universities could not be able to properly consider the variety and peculiarities of the institutions' missions and may induce distortions in their behaviours (Cremonini et al. 2010). Considering the stakeholders' perspective, Bastedo and Bowman emphasize the influence of published rankings on the perception of university performance, quality and reputation, while the last one should be an independent indicator (Bastedo and Bowman 2011). Following, HE could behave as operating in a market sector, transforming themselves from institutions "for the public good into corporations and brands, in the construction of education as a globally traded commodity" (Amsler and Bolsmann 2012). Other authors claim that rankings could legitimate and strengthen categorical inequalities (Rawlings and Bourgeois 2004; Cantwell and Taylor 2013; Pitman et al. 2020). By means of the power owned by the neutral quantitative measures, "rankings put institutions in a mould" and drive researchers to adopt conformist behaviours addressing their research interests towards the main conventional flows in scientific disciplines (Frey and Kosta 2008; Thakur 2007).

Despite all the contributes of critical literature, it could be also considered that rankings could produce positive effects, as they could stimulate universities to a fair competition and to a tension towards the continuous improvement of their quality.

Undoubtedly, apart from the opinions of supporters and deniers of rankings, some concerns have to be considered when a ranking is published. Users of rankings tend to believe that the issued scores are reliable and authoritative, without considering whether the construction of indicators had some methodological limitations or not. Furthermore, in the institutional context, rankings may lead the higher education institutions to conform with the set of indicators used by the most remarkable evaluators, preferring to lose their diversity.

In Italy, the most important effects of rankings could deal with public funding allocation and university reputation. The first is a matter of the Italian government's educational policy for financial resources to be prioritised, the second affects the students' choice of the university.



Evaluating Italian universities: ANVUR periodic accreditation judgment versus international rankings

On the international level, the most remarkable rankings are: Times Higher Education (THE), QS World University Rankings, Academic Rankings of World Universities (ARWU), Center for World University Rankings (CWUR), Round University Rankings (RUR), and U-Multirank (U).

In the Italian university system, the first large evaluation exercise (a smaller one was held in 2004) was the VQR 2004–2010 (Valutazione della Qualità della Ricerca, Research Quality Evaluation), completed in July 2013 by the Italian Agency for the Evaluation of Universities and Research Institutes (ANVUR); the third edition, always devoted to research outcomes, is still in progress and will be concluded in 2022. Although the main goal of the VQR exercise is not to create a university ranking, it however provides lots of specific information, both at university, department and scientific field level, that are used to define the allocation of portion of the university funding provided by the Ministry of Research (Blasi et al., 2017). In addition to the VQR exercises, ANVUR carries on the AVA Self-assessment, Periodic Evaluation, Accreditation, that is operational since 2013. AVA has three main aims: *“to assure that Italian HE institutions provide an appropriate and equivalent quality of their services; to support universities in the exercise of responsible and reliable autonomy in the use of public resources and in collective and individual behaviour related to education, research and knowledge and technological transfer activities; to improve the quality of education and research”*. To achieve these goals, AVA sets the standards for the self-assessment of programmes and institutions, involving their internal procedures and the outcomes of their activities, as well as for the external assessment of the quality assurance systems, based on peer review and carried out by experts entrusted by the ANVUR. The AVA process, based on experts’ onsite visits, was launched at the end of 2014. It was an important innovation for the Italian university system and aligns the country with the practices defined the Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG).

## 2 Data and methods

At the first step of our analysis, all the international rankings which include Italian universities were considered. Then, among them, only those that were comparable in terms of benchmarking matters were selected: Times Higher Education (THE), QS World University Rankings, Academic Rankings of World Universities (ARWU), Center for World University Rankings (CWUR), Round University Rankings (RUR). Some Italian universities also take part in U-Multirank (U), but it provides only ranks and not scores, then it cannot be included in this analysis. All the considered rankings were published in 2019 and refer to the previous year, like ANVUR data. Concise information about each ranking is shown in Table 1, and web sites listed in the References, while a more detailed evidence on ANVUR AVA Self-assessment and periodic accreditation of Italian Universities and their programmes is considered necessary. It started in 2014 and expected be completed within 2020, as stated by the

Angela Maria D'Ugento, Nunziata Ribecco and Vito Ricci

Ministerial Decree 6/2019, art. 3, with the aim to put in place the periodic accreditation visits of the universities and their programmes. Only for the 12 remaining universities the calendar has been redefined by ANVUR following the regulatory measures on the management of the epidemiological emergency from COVID-19, then it is expected to be concluded by 2021. During the accreditation process, the committees of experts (CEV) entrusted by the ANVUR, were sent to Italian universities to visit them, discuss with the governing board all the related AVA documents, already analysed in a preliminary remote evaluation, and make interviews to some stakeholders, including staff and students, involved in the selected programmes under examination.

**Table 1:** Dimensions and weights of the five selected HERs comparable to the Italian AVA

ARWU		CWUR		QS		RUR		THE	
Dimensions	W	Dimensions	W	Dimensions	W	Dimensions	W	Dimensions	W
Alumni winning Nobel Prizes and Fields Medals	0.10	Quality of education	0.25	Citations per Faculty	0.20	Teaching	0.40	Teaching	0.30
Staff winning Nobel Prizes and Fields Medals	0.20	Alumni employment	0.25	International Students	0.05	Research	0.40	Research	0.30
Highly Cited Researchers	0.20	Quality of Faculty	0.10	International Faculty	0.05	International Diversity	0.10	Citations	0.30
Papers published in Nature and Science	0.20	Research performance	0.40	Faculty Student	0.20	Financial Sustainability	0.10	Industry Income	0.02
Papers indexed in Science Citation Index-Expanded and Social Science Citation Index	0.20	-	-	Employer Reputation	0.10	-	-	International Outlook	0.08
Per capita academic performance of an institution	0.10	-	-	Academic Reputation	0.40	-	-	-	-

(-) related web sites are listed in the References.

One of the aims of AVA complex procedure was to obtain the evaluation reports indicating the areas of possible improvement of the universities, as a whole, and of the single programmes examined, and to highlight the presence of good practices in the Italian university system.

At the end of this process, universities receive the periodic accreditation judgment, graded according to the scale (provided for by art. 3 of Ministerial Decree 6/2019) expressed by means of a score and translated into a letter: A (final score  $\geq 7.5$ ) is the best condition and corresponds to a high positive judgement, giving a five year accreditation, while, on the lowest level, an E is given if it is “unsatisfactory” (final score  $\leq 4.0$ ) and it could lead to the closure of the evaluated object, the programme or even of the institution branch.

Evaluating Italian universities: ANVUR periodic accreditation judgment versus international rankings

After collecting data from each of the rankings under consideration, in which the same Italian universities were scored, we investigated the possible relations among the international ones and ANVUR periodic accreditations in terms of ranks by means of Spearman correlation. The data were collected in Spring 2020.

Following, the most interesting results allowed us to make a comparison with the ANVUR judgements by means of linear models.

## 2 Main results and discussion

The knowledge of the dimensions considered by each of the main international rankings let us to detect those showing comparable characteristics. It emerged that five of them could be suitably selected to investigate their possible relations with the final ANVUR score received by Italian universities participating to all the assessments under investigation. The rankings which have a positive correlation with ANVUR one are, in a decreasing order, RUR, CWUR (significant at  $\alpha=0.01$ ) and ARWU (Table 2). The rankings were obtained based on the overall scores presented in each of them. A negative correlation is only found between ANVUR and Times Higher Education. Moreover, interesting associations emerge among the five international rankings in turn, highlighting some similarities in their methodologies and results.

**Table 2:** Spearman Rho correlation matrix among the scores of the selected rankings and number of Italian HEs participating

Ranking	ANVUR	ARWU	CWUR	QS	TimesHE	RUR	Italian HEs
ANVUR	1.000	0.181	0.394*	0.023	-0.026	0.491	34
ARWU	0.181	1.000	0.799**	0.627**	0.495*	-0.257	22
CWUR	0.394*	0.799**	1.000	0.508*	0.505*	0.500	34
QS	0.023	0.627**	0.508*	1.000	0.414	0.393	18
TimesHE	-0.026	0.495*	0.505*	0.414	1.000	0.679	22
RUR	0.491	-0.257	0.500	0.393	0.679	1.000	7

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*\*' 0.01 '\*\*' 0.05 '.' 0.1 ' ' 1.

Then, a linear model is estimated to investigate the relation between the international rankings and the ANVUR one, and to assess the strength of the influence of the former on the response variable. The observations are couples of scores. The goodness of fit of the estimated linear models, assessed by means of a residual analysis, proved to be slightly significant only for CWUR (Table 3).

**Table 3:** The results of the estimated linear model for ANVUR score versus the selected international rankings

Ranking	Intercept	Coefficient	Std. Error	t value	p-value	Multiple R squared	Adjusted R-squared	Sig. code
<i>CWUR</i>	59.7691	2.0511	0.7451	2.753	0.0097	0.19150	0.16620	**
<i>Not significant models</i>								
<i>ARWU</i>	5.271	0.731	1.040	0.703	0.490	0.02413	-0.02467	
<i>QS</i>	-8.648	3.480	4.625	0.752	0.463	0.03417	-0.02619	
<i>Times HE</i>	35.9969	0.6115	1.7034	0.359	0.72336	0.00640	-0.04328	

This result is unsurprising, as the two ranks have the most similar characteristics and are based on a more complete set of indicators contributing to the final evaluation, which is obtained through a more complex methodology.

The results proposed in this paper, even in a synthetic frame, do represent a very first analysis of a wider research in progress aimed at finding those latent factors shared by the main international rankings. The most remarkable among them draw the attention of both researchers and media (newspapers and press agencies specialized in higher education, stakeholders and firms), which can communicate some messages that, once produced, could have an independent life from the intentions of the original data providers; then, this kind of information should be handled with care.

The debate over the ability of rankings in providing extensive information, mainly on Research and Education outcomes, is far from its end. Moreover, the introduction of a more complete set of indicators regarding mentoring, infrastructures and student assessments should be considered, so adding new information to build a multidimensional concept like university reputation. Giving detailed information, instead of rankings in league tables, could be a better support to a well-informed decision-making process by the stakeholders.

## References

1. Amsler, S., Bolsmann, C.: University rankings as social exclusion. *British Journal of Sociology of Education*, (2012). 33(2), 283–301.
2. Bastedo, M., Bowman, N.: College rankings as an interorganizational dependency: Establishing the foundation for strategic and institutional account. *Res. High Educ* (2011). 52(1), 3–23. doi: 10.1007/s11162-010-9185-0.
3. Blasi, B., Romagnosi, S., Bonaccorsi, A.: Playing the ranking game: media coverage of the evaluation of the quality of research in Italy. *High Ed* (2017). 73:741–757. doi: 10.1007/s10734-016-9991-1.
4. Cantwell, B., Taylor, B. J.: Global status, intra-institutional stratification and organizational Segmentation: A time-dynamic Tobit analysis of ARWU position among U.S. universities. *Minerva* (2013), 51, 195–223.
5. Cremonini, L., Benneworth, P., Westerheijden, D. F.: In the shadow of celebrity: The impact of world-class universities policies on national higher education systems. In *Annual Meeting of the Association for the Study of Higher Education*, Indianapolis (2010). <http://www.ashe.ws/?page=725>.
6. Frey, B., Kosta, R.: Do rankings reflect research quality? (CESifo Working Paper No. 2443. 2008.
7. Horstschräer, J.: University rankings in action? The importance of rankings and an excellence competition for university choice of high-ability students. *Economics of Educ. Rev.* (2012) 31, 1162– 1176.
8. Pitman, T., Edwards, D.2, Zhang, L.-C., Koshy, P., McMillan, J.: Constructing a ranking of higher education institutions based on equity: is it possible or desirable? *High Educ.* (2020). 80, 605–624. doi:10.1007/s10734-019-00487-0.
9. Rawlings, C. M., Bourgeois, M. D.: The complexity of institutional niches: Credentials and organizational differentiation in a field of American higher education. *Poetics*, (2004), 32(6), 411–446.
10. Thakur, M.: The Impact of Ranking Systems on Higher Education and its Stakeholders. *Journal of Institutional Research*, (2007). 13(1), 83–96.  
ANVUR AVA Periodic accreditation: Rapporti ANVUR di Accreditamento Periodico – ANVUR  
ARWU: <http://www.shanghairanking.com/ARWU-Methodology-2019.html>  
CWUR: <https://cwur.org/methodology/world-university-rankings.php>  
QS: <https://www.topuniversities.com/qs-world-university-rankings/methodology>  
RUR: <https://roundranking.com/methodology/academic-ranking.html>  
Times Higher Education: <https://www.timeshighereducation.com/>  
U-Multirank: <https://www.umultirank.org/>

# Women's career discrimination in the Italian Academia in the last 20 years

## *Discriminazione di genere nell'università italiana negli ultimi 20 anni*

Daniele Cuntrera, Vincenzo Falco, and Massimo Attanasio

**Abstract** We examine discrimination against women in Italian universities in terms of career advancements for the last 20 years. Data is taken from the MUR archive. Two points are examined: the transition time from assistant professor to associate professor; and the transition time from associate professor to full professor.

**Abstract** *Esaminiamo la discriminazione contro le donne nelle università italiane in termini di avanzamenti di carriera negli ultimi 20 anni. La fonte dei dati è l'archivio MUR. Due punti sono oggetto di studio: i tempi di transizione da ricercatore a professore associato e i tempi di transizioni da professore associato a professore ordinario.*

**Key words:** Italian university, gender gap, career advancement

## 1 Introduction

The recruitment and selective advancement of faculty members are essential for the medium and long-term future of academic systems. These processes are even more critical given the development of the 'world-class' university model, in which the concentration of talent is a key attribute. Academic careers can be usefully analyzed across four dimensions: participation; position; productivity; and recognition (Long and Fox, 1995). In this paper, we focus on position, and in particular on career advancement. The gender gap is notable here because contemporary academic and

---

<sup>1</sup> Daniele Cuntrera, Università degli Studi di Palermo; email: [daniele.cuntrera@unipa.it](mailto:daniele.cuntrera@unipa.it)  
Vincenzo Falco, Università degli Studi di Palermo; email: [vincenzo.falco01@unipa.it](mailto:vincenzo.falco01@unipa.it)  
Massimo Attanasio, Università degli Studi di Palermo; email: [massimo.attanasio@unipa.it](mailto:massimo.attanasio@unipa.it)

policy research on academic employment has shown how female academics throughout Europe continue to experience employment discrimination (David and Woodward, 1998).

The goal of this paper is to study: career advancement intervals in Italy; how they vary over time; and how they are influenced by covariates like gender, region and field of study. Special attention will be given to the gender gap in the transition from assistant professor to associate professor; and from associate professor to full professor.

## 2 Data and aims

This study considers Italian faculty population from 2001 to 2019, provided by the Ministry of University and Research (MUR). The dataset contains individual data for each year. Each statistical unit comprises two types of covariates:

Faculty member variables (FMVs):

- a unique identifier for each faculty member;
- first name and surname;
- gender;
- position: full professor (*Ful*), associate professor (*Ass*), or assistant professor (*Ast*);
- the field of study (or academic discipline) is classified with four digits. The academic disciplines (SSDs) are 370, which are collected into: 190 academic recruitment fields (SCs), 88 groups of academic recruitment fields (MCs), and 14 academic areas (ARs).

University variables (UVs):

- the name of the university;

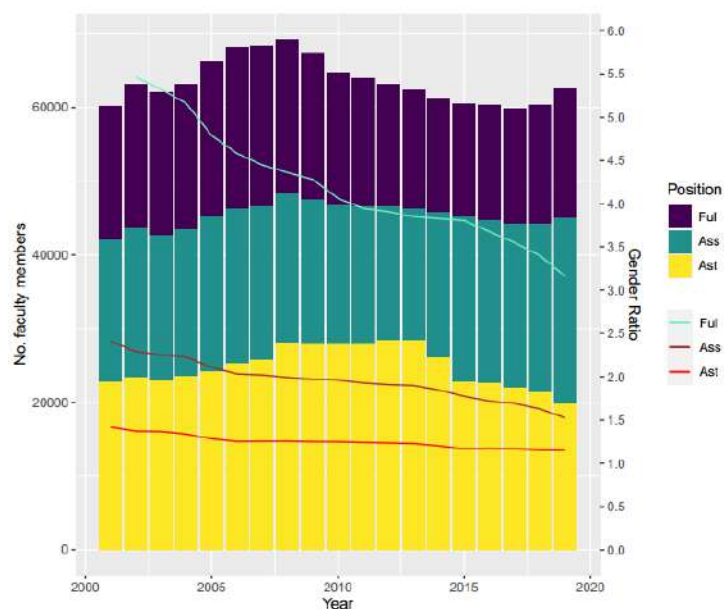
The final career record is obtained by merging the annual records with name and surname as the key and other distinguishing variables when professors have the same name. The focus is on the gender gap in academic career advancement. Here there are three levels: assistant (*Ast*), associate (*Ass*) and full (*Ful*) professors. The observation period is 20 years.

There will be two steps for data analysis: the first one is a cross-sectional analysis of the whole MUR archive, in order to describe the structure of Italian faculty positions, 2001-2019; the second is a longitudinal event-history analysis using career advancement as outcomes.

## 3 Cross-section analysis

Figure 1 and Table 1 report brief statistics on Italian professors in terms of gender. In Figure 1, the bars represent the numbers of the three types of professors in Italy

Women's career discrimination in Italian universities in the last 20 years from 2000 to 2019. There was an increase between 2005 and 2010, while in 2019 the numbers were very similar to 2001. The lines describe the gender ratio (M/F) for the three levels. The full professors' indexes fall more steeply over the years.



**Figure 1:** Number of faculty members and gender ratio by position (bars) and gender ratio by position (lines) in Italy. 2001 to 2019

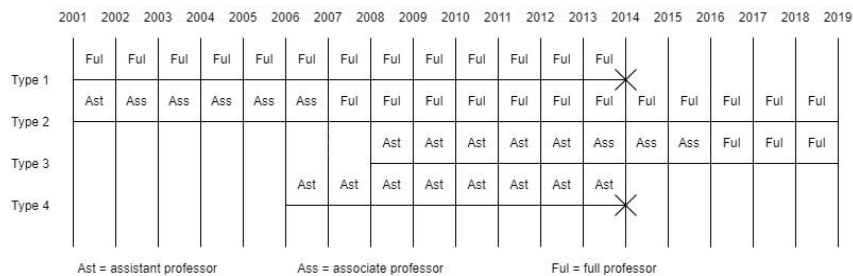
Table 1 shows the gender ratios computed for the three positions and by different academic areas. As already seen, the gender ratio decreases from 2001 to 2019 for all areas and positions, but in “mathematics and informatics” (1) the overall ratio increases. The areas “industrial and information engineering” (9) and “antiquities, philology, literary studies, art history” (10) show, respectively, maximum and minimum gender ratios in both years (and for the three levels). These values mirror the well-known gender gap in these fields. The only overall gender ratios of less than one are in “biology” (5) and “antiquities, philology, literary studies, art history” (10).

**Table 1:** Gender Ratio of university faculty members by position and academic areas in Italy (2001 and 2019)

Area	2001					2019				
	M/F				Total	M/F				Total
	Ast	Ass	Ful	Overall		Ast	Ass	Ful	Overall	
1	1,22	1,65	5,59	1,98	3351	1,89	1,82	4,29	2,28	3541
2	2,84	5,32	16,74	5,32	2599	2,72	3,66	6,22	3,73	2500
3	0,86	2,29	7,53	2,08	3228	0,65	1,04	2,23	1,03	3003
4	1,87	3,34	10,29	3,42	1327	1,92	2,35	4,41	2,49	1106
5	0,72	1,19	3,32	1,28	5101	0,59	0,84	1,98	0,88	5117
6	2,15	4,38	12,47	3,56	11376	1,17	2,20	5,11	2,01	9781
7	1,41	2,72	9,18	2,72	3105	1,05	1,31	3,71	1,51	3259
8	2,64	4,57	8,90	4,17	3864	1,37	1,93	3,41	1,98	3898
9	5,02	8,55	25,82	9,04	4547	3,19	4,26	8,13	4,55	6231
10	0,57	0,80	1,89	0,90	6003	0,74	0,74	1,32	0,86	5430
11	0,92	1,56	3,34	1,57	4818	0,86	1,09	1,82	1,17	5091
12	1,49	2,68	6,77	2,67	4915	1,15	1,36	3,05	1,74	5636
13	1,33	2,59	6,44	2,53	4428	1,09	1,34	3,21	1,68	5965
14	1,42	2,4	4,86	2,35	1519	1,09	1,40	2,75	1,51	1978
Overall	1,41	2,41	5,91	2,39	60181	1,15	1,52	3,16	1,67	62536

#### 4 Event-history analysis

The two events of interest are the transitions *Ast* to *Ass* and *Ass* to *Ful*. To better explain the structure of the dataset, we broke down the faculty career paths into four types including all possible hiring and retiring timing (Figure 2):



**Figure 2:** Observation period (2001-2009) of some career's path

- Career type 1: hired before 2001 and retired at any time within 2019,  $n = 33988$ ;
- Career type 2: hired before 2001 and on duty in 2019,  $n = 25791$ ;
- Career type 3: hired after 2001 and on duty in 2019,  $n = 36444$ ;
- Career type 4: hired after 2001 and retired at any time within 2019,  $n = 4023$ .



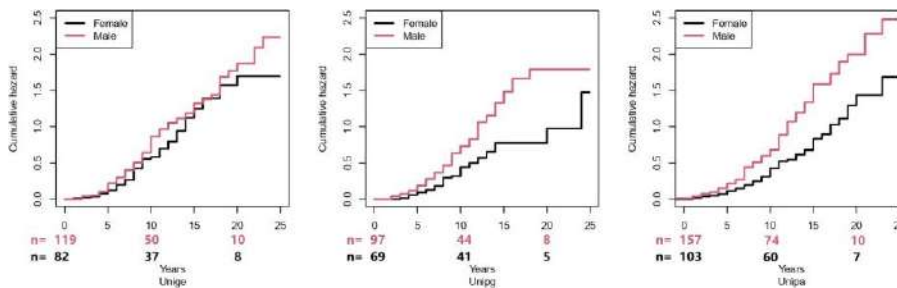
Women's career discrimination in Italian universities in the last 20 years

To carry out a longitudinal career advancement analysis with our data, we need to know the year of birth and the timing of advancements. To obtain the information mentioned above, we get web information for a sample. We draw a sample from the MUR archive, for types 2 and 3, as information for types 1 and 4 are not available online. In this way, the target population is restricted to the professors on duty in 2019. For the sake of brevity, the sampling technique is not described in detail. We restrict our study to three universities representing the three different Italian areas: Unipa (Palermo, South), Unipg (Perugia, Center), and Unige (Genoa, North) (Table 2):

**Table 2:** Faculty members population and sample by gender. Unige, Unipg, and Unipa. 2001 to 2019

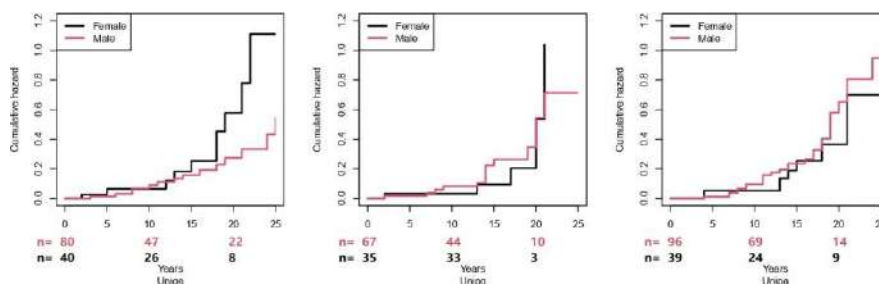
University	Type	Total Faculty Members		Total Faculty Members Sample	
		M	F	M	F
Unige	2	296	152	59	30
	3	463	326	93	65
Unipg	2	247	121	49	24
	3	315	269	63	54
Unipa	2	340	156	68	31
	3	554	392	111	78
Total		2215	1416	443	282

To describe the transitions for event 1 ( $Ast \rightarrow Ass$ ) and event 2 ( $Ass \rightarrow Ful$ ), we construct, respectively, the cumulative incidence functions  $H1(t)$  and  $H2(t)$ . The time origin is the year of recruitment as  $Ast$  for event 1 and  $Ass$  for event 2 and the length of follow-up is fixed to 25 years. The transition from  $Ast$  to  $Ful$  are not included because they are negligible. Figure 3 considers event 1 for the three universities: the male curves are always above the female ones, as the probability of becoming  $Ass$  is more likely for males. At Unige the two curves are close to 20 years covering almost all the sample, while at Unipa and Unipg the differences are much clearer.



**Figure 3:** Cumulative hazard  $H1(t)$  by university (event 1:  $Ast$  to  $Ass$ ). The numbers reported in the figures are the Assistant professors' risk sets of Males and Females

Figure 4 considers event 2 for the three universities: the male curves are always above the female ones, but at Unige there is a jump for females after seventeen years. For the other two universities, the cumulative functions are almost always in favour of the males.



**Figure 4:** Cumulative hazard  $H_2(t)$  by university (event 2: *Ass to Full*). The numbers reported in the figures are the Associate professors' risk sets of Males and Females

## 5 Conclusions

This study is a first explanatory analysis of career advancements for Italian academia and for the gender gap. Our results are in line with the wide literature on gender gap in the recruitment process (Abramo et al., 2016; Filandri and Pasqua, 2019). The novelty of our paper is the introduction of the gender gap given by the “velocity” in the career advancement in the last 20 years. The study could be enriched by applying a survival regression model including: birth cohort (in order to get information over the generations); field of study (as there is a well known discrimination gap between male and female fields of study); and geographical areas (as the gender gap in the North-Center of Italy is less pronounced). This type of model could detect interesting interactions among the above mentioned covariates.

## References

1. Abramo, G., D'Angelo, C.A., Rosati, F.: Gender bias in academic recruitment. *Scientometrics*. **106**, 199–141 (2016)
2. David, M., Woodward, D.: *Negotiating the glass ceiling careers of senior women in the academic world*. Falmer Press, London (1998)
3. Filandri, M., Pasqua, S.: *Gender discrimination in academic careers in Italy*. Department of Economics and Statistics Cognetti de Martiis. Working papers 201921, University of Turin (2019)
4. Long, J. S., Fox, M. F.: Scientific careers: universalism and particularism. *Review of Sociology*. **21**, 45--71 (1995)

## 4.34 Statistical methods with Bayesian networks

# Statistical Micro Matching Using Bayesian Networks

## *Matching Statistico con approccio micro mediante Reti Bayesiane*

Pier Luigi Conti and Daniela Marella and Paola Vicard and Vincenzina Vitale

**Abstract** The goal of statistical matching, at a *micro* level, is the construction of a synthetic data source where all the variables of interest are available. In this paper we propose the use of Bayesian Networks to deal with the statistical matching for multivariate categorical variables in the *micro* approach. Its performance is evaluated by an application to a real data set.

**Abstract** *L'obiettivo del matching statistico a livello micro è la creazione di un data set sintetico in cui tutte le variabili di interesse siano disponibili. In questo lavoro proponiamo l'uso delle reti bayesiane nell'ambito del matching statistico multivariato per variabili categoriali, secondo l'approccio micro. La sua performance è valutata mediante un'applicazione a dati reali.*

**Key words:** Bayesian networks; collapsibility; statistical matching; uncertainty

## 1 Introduction

Let  $\mathbf{X} = (X_1, \dots, X_H)$ ,  $\mathbf{Y} = (Y_1, \dots, Y_K)$  and  $\mathbf{Z} = (Z_1, \dots, Z_T)$  be vectors of random variables of size  $H$ ,  $K$ ,  $T$ , respectively. Furthermore, let  $A$  and  $B$  be two independent samples of  $n_A$  and  $n_B$  independent and identically distributed records from

---

Pier Luigi Conti

Dipartimento di Scienze Statistiche, Sapienza Università di Roma e-mail: pierluigi.conti@uniroma1.it

Daniela Marella

Dipartimento di Scienze della Formazione, Università Roma Tre e-mail: daniela.marella@uniroma3.it

Paola Vicard

Dipartimento di Economia, Università Roma Tre e-mail: paola.vicard@uniroma3.it

Vincenzina Vitale

Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma e-mail: vincenzina.vitale@uniroma1.it

$(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ . Assume that  $(\mathbf{X}, \mathbf{Y})$  are observed in sample  $A$ , while  $(\mathbf{X}, \mathbf{Z})$  are observed in sample  $B$ . In the statistical matching it is possible to consider the macro and the micro approaches. The micro approach consists in constructing a data set containing complete “observations”  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ . The macro approach consists in estimating the joint probability mass function (*pmf*) of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ . Such a distribution is generally not identifiable, since the parameters regarding the statistical relationship between  $\mathbf{Y}$  and  $\mathbf{Z}$  are not estimable due to the lack of joint information on  $\mathbf{Z}$  and  $\mathbf{Y}$ . In order to overcome the identification problem, alternative techniques have been proposed in the literature. A first group of techniques is based on the conditional independence assumption between  $\mathbf{Y}$  and  $\mathbf{Z}$  given  $\mathbf{X}$  (CIA) [6]. A second group of techniques uses external auxiliary information regarding the statistical relationship between  $\mathbf{Y}$  and  $\mathbf{Z}$  [9]. In [1] the use of Bayesian networks (BNs) to deal with the statistical *macro* matching in multivariate categorical data is proposed. The use of BNs allows: (i) to introduce extra sample information on qualitative dependencies between the components of  $\mathbf{Y}$  and  $\mathbf{Z}$ ; (ii) to use such an information to factorize the joint *pmf* by decomposing a “global” dependence into “local” dependencies. Such a representation of the joint *pmf*, taking advantage of local relationships, allows to simplify both parameters estimation and statistical matching uncertainty evaluation in a multivariate context since a smaller number of lower dimension parameters needs to be estimated. The first attempt to use BNs for statistical matching of multivariate discrete data is in [4] where the CIA is assumed thanks to its connection with  $d$ -separation criterion. Under CIA, both the dependence structure and the BN parameters are estimable from the sample data, and there is no uncertainty at all. When the CIA model is not adequate, the application of standard inferential procedures may produce highly misleading results.

In this paper we propose the use of BNs to deal with statistical matching *micro* approach. The paper is organized as follows. In Section 2 we briefly review the results obtained by [1] in the statistical *macro* matching. After having estimated a plausible *pmf* for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , BNs allow a straightforward extension to a *micro* approach. In Section 3 its performance is evaluated by an application to a real data.

## 2 Statistical Macro Matching using BNs

BNs are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG) [7]. A graph is a pair  $G = (V, E)$  consisting in a set of vertices  $V$  and a set of directed edges  $E$  between pairs of nodes. Each node corresponds to a random variable, and missing arrows between nodes imply (conditional) independence between the corresponding variables. In a BN each node  $x_i$  say, is associated with the distribution of the corresponding variable given its parents,  $\text{pa}(x_i)$ . Hence a BN consists of two components: the DAG and the set of the distributions parameters. In the statistical matching context, the non identifiability of the *pmf* for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  implies that both the components of the BN (*i.e.* the DAG and its parameters) can not be estimated from  $A$  and  $B$ . In fact, when BNs are used to deal with statistical matching, two issues must be addressed: (i) specify the dependence structure of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ ; (ii) estimate the parameters, *i. e.*

the local probability distributions associated to the edges between the components of  $\mathbf{Y}$  and  $\mathbf{Z}$ . As a consequence, two kinds of uncertainty have to be accounted for: (i) uncertainty regarding the DAG, *i.e.* the dependence structure among the variables of interest; (ii) uncertainty regarding the parameters of the statistical relationships between  $\mathbf{Y}$  and  $\mathbf{Z}$  (the conditional probability tables entries) given the DAG, *i.e.* given the factorization of the joint *pmf* for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ . [1] deals with the use of BNs in the statistical *macro* matching context. The authors propose to estimate the distribution function of variables which are not jointly observed and show how to evaluate its reliability by computing a measure of total uncertainty. Its computation requires the following steps.

- Step 1 Estimate the DAGs of  $\mathbf{X}$ ,  $(\mathbf{X}, \mathbf{Y})$  and  $(\mathbf{X}, \mathbf{Z})$  from  $A \cup B$ ,  $A$  and  $B$ , respectively. Let  $G_X = (V_X, E_X)$ ,  $G_{XY} = (V_{XY}, E_{XY})$  and  $G_{XZ} = (V_{XZ}, E_{XZ})$  be the DAGs estimated *via* samples  $A$  and  $B$  as follows. First of all, the DAG  $G_X$  is estimated on the basis of the overall sample  $A \cup B$ . Secondly, given  $G_X$ , the association structure for  $(\mathbf{X}, \mathbf{Y})$  and  $(\mathbf{X}, \mathbf{Z})$  is estimated through the sample data in  $A$  and  $B$ , respectively.
- Step 2 Insert extra sample information on the association structure between  $\mathbf{Y}$  and  $\mathbf{Z}$  in the DAGs. For instance, if variable  $Y_k$  is associated to  $Z_t$  then a link between the vertices  $Y_k$  and  $Z_t$  must be added.
- Step 3 Define the class of plausible DAGs for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  and compute the uncertainty due to the dependence structure, as described in [1]. The class of plausible DAGs has been defined through the concept of collapsibility, as defined in [5]. Given a vertex  $Z_t$  in the set  $\mathbf{Z}$ , the collapsibility requires that the estimate  $\hat{\mathbf{P}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z} \setminus \{Z_t\})$  of  $\mathbf{P}(\mathbf{X}, \mathbf{Y}, \mathbf{Z} \setminus \{Z_t\})$ , obtained by marginalizing the maximum likelihood estimate (MLE) of  $\hat{\mathbf{P}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  under the original DAG  $G_{XYZ}$ , coincides with the MLE under the DAG  $G_{XYZ \setminus \{Z_t\}}$  obtained from  $G_{XYZ}$  removing the vertex  $Z_t$ .
- Step 4 Select a DAG  $G_{XYZ}^*$  from the class of plausible models defined in Step 3 and let  $P_{XYZ}^*$  be the joint *pmf* associated to  $G_{XYZ}^*$ . Clearly, according to  $G_{XYZ}^*$ ,  $P_{XYZ}^*$  can be factorized into local probability distributions, some of which can be directly estimated from the available sample information, while others cannot. For the latter distributions, the set of all plausible MLEs is computed as in [2]. Finally, the parameters uncertainty measure proposed in [8] has been computed.
- Step 5 Compute the total uncertainty by adding the dependence structure uncertainty to the uncertainty in the parameters estimation.

After having estimated a plausible *pmf* for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , BNs allow a straightforward extension to a *micro* approach. In fact, missing  $\mathbf{Z}$  values in sample  $A$  and missing  $\mathbf{Y}$  values in sample  $B$  can be imputed from the given BN by efficient evidence propagation algorithms. The performance of such an approach has been evaluated through an application to a real dataset performed in Section 3.

### 3 Application to the 2017 Aspects of Daily Life Survey

To simulate the statistical matching scenario, we used a subset of variables belonging to the 2017 Aspects of Daily Life (ADL) survey, annually carried out by Istat.

In particular, 14 variables have been selected pertaining to the promoting factors of sharing mobility in Italy with reference to 34249 sample units, *i.e.* individuals aged 18 years and over. As shown in Table 1, the 14 variables are grouped according to

**Table 1** The ADL variables used for Statistical Matching

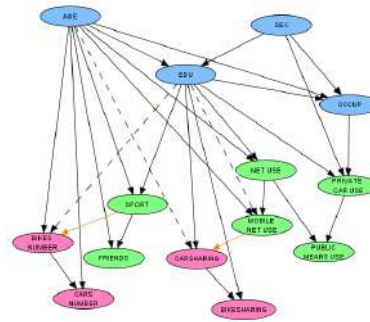
	Label	Description
<b>X variables</b>	AGE	Age in years (18-24; 25-34; 35-44; 45-54; 55-64; 65-74; 75+)
	SEX	Gender (male; female)
	EDU	Educational level (lower than high school; high school; bachelor's degree or higher)
	OCCUP	Professional activity (employed; in search of employment; inactive (excluding students); student)
<b>Y variables</b>	NET USE	the frequency of use of Internet during the last 12 months (daily; not daily)
	MOBILE NET USE	the frequency of use of a mobile connection (less than 3 months; over 3 months)
	PUBLIC MEANS USE	the frequency of use of public means (daily or a few times a week; a few times a month/year; never or the service doesn't exist)
	SPORT	the frequency of playing sport activities (no; occasionally; regularly)
	FRIENDS	the frequency of hanging out with friends (less than once a week; up to once a week)
	PRIVATE CAR USE	the frequency of use of own private car (never; a few times a week/month/year; daily)
<b>Z variables</b>	CARS NUMBER	the number of cars per family (no cars; one; two or more)
	BIKES NUMBER	the number of bikes per family (no bikes; one; two; three or more)
	BIKESHARING	the use of a bikesharing in the last 12 months (no; yes)
	CARSHARING	the use of a carsharing in the last 12 months (no; yes)

the **X**, **Y** and **Z** components. In addition, the data set has been randomly split into two subsets, named *sample A* and *sample B*, the former composed by  $n_A = 17125$  units, the latter by  $n_B = 17124$  units. The **X** variables are the only observed in both samples, being the so called matching variables, while the **Y** variables have been removed from *sample B* and the **Z** variables from *sample A*. As described in Section 2, the first step involved the estimation of  $G_X^1$  based on all the 34249 observations, the estimation of  $G_{XY}$ , based only on *sample A*, and that of  $G_{XZ}$ , based only on *sample B*. In both graphs, the dependence structure estimated in  $G_X$  has been preserved with the further constraint that the **X** variables cannot belong to the set of children of **Y** in  $G_{XY}$  and to that of **Z** in  $G_{XZ}$ . In this work, all the graphs have been estimated by means of the *hill climbing* algorithm (with *BIC* score) as implemented in the R package `bnlearn`. The most parsimonious matching graph  $G_{XYZ}$  arises from the union of the arcs belonging to  $G_{XY}$  and  $G_{XZ}$  that, properly, corresponds to the BN under the CIA (*i.e.*  $BN_{CIA}$ ). To overcome the latter assumption embedding extra sample information on the relationships between the **Y** and **Z** components, two edges have been added to the graph  $G_{XYZ}$ :  $MOBILE NET USE \rightarrow CARSHARING$ , since the use of internet on the mobile device is considered an important enabling factor<sup>2</sup>[3], and  $SPORT \rightarrow BIKES NUMBER$ , since the propensity to play one or more sports could influence the number of bikes owned. The resulting BN (*i.e.*  $BN_U$ ) is shown in Figure 1 with the two added edges coloured in orange. The three dashed arcs in  $BN_U$  have been added to ensure the sequential *c-removability*<sup>3</sup>. This means that all the nodes in *sample A* (*sample B*) can be

<sup>1</sup> The following forbidden arcs' direction have been imposed:  $SEX \rightarrow AGE$  and  $AGE \rightarrow SEX$  being sex and age logically independent;  $EDU \rightarrow AGE$ ,  $OCCUP \rightarrow AGE$ ,  $OCCUP \rightarrow SEX$ ,  $EDU \rightarrow SEX$  since education and professional status of an individual can not influence his sex and age;  $OCCUP \rightarrow EDU$  since it is more likely that an individual with a higher level of education has a better job position.

<sup>2</sup> Carsharing users must have a mobile connection to rent a car.

<sup>3</sup> For nodes belonging to *sample A*, the sequential *c-removability* is ensured according to the order: FRIENDS - SPORT - PUBLIC MEANS USE - PRIVATE CAR USE - NET USE - MOBILE NET USE. For nodes belonging to *sample B*, it is ensured according to the order: CARS NUMBER - BIKE NUMBER - BIKESHARING - CARSHARING.



**Fig. 1** Graph  $G_{XYZ}$  with extra sample information. Pink, green and blue nodes denote  $Y$ ,  $Z$  and  $X$  variables, respectively.

removed from the graph one after another according to some order. From the factorization of the joint  $pmf$  induced by  $BN_U$ , the only two local  $pmfs$  that cannot be estimated by data in samples  $A$  and  $B$  are  $P(\text{BIKES NUMBER}|\text{AGE, EDU, SPORT})$  and  $P(\text{CARSHARING}|\text{AGE, MOBILE NET USE, EDU})$ . To this purpose, EM algorithm with 100000 different starting points has been run to explore the likelihood ridge of each cell of the above two local  $pmfs$ . According to [1], the uncertainty in the dependence structure, defined as the maximum number of edges between  $Y$  and  $Z$  that can be added to the graph without generating directed cycles, is  $(K - 2)(T - 2) = 8$  and that referred to parameters estimation is 1.84 computed as in [2].

In the *micro* approach, the final goal is to get a complete synthetic dataset by imputing the missing values in  $A$  and  $B$ ; to this purpose, the two unavailable conditional probability tables have been randomly selected from the 100000 EM estimates. Therefore, the joint  $pmf$  associated to  $BN_U$  has been computed and the BN has been used to impute the missing  $Z$  values in  $A$  and the missing  $Y$  values in  $B$ . The `impute()` function in the R package `bnlearn` has been used to impute the values by averaging likelihood weighting simulations performed using all the available nodes as evidence. The number of random samples which are averaged for each observation has been fixed to 1000. To provide the extent of accuracy degree of the statistical matching procedure, we compared the performances of our proposal (*i.e.*  $BN_U$ ) with those under the CIA assumption (*i.e.*  $BN_{CIA}$ ) considering the marginal distributions and the recovered association structure. Table 2 shows the Jensen-Shannon divergence, using the base 2  $\log$ , between the true marginal distribution and that arisen from the synthetic data set, for all nodes. Even if, except

**Table 2** Jensen-shannon distance between the true marginal distribution and that arisen from the synthetic data set, for all nodes, with reference to the  $BN_U$  model and  $BN_{CIA}$  model

	SPORT	FRIENDS	NET USE	MOBILE NET USE	PUBLIC MEANS USE	BIKE-SHARING	CAR-SHARING	PRIVATE CAR USE	BIKES NUMBER	CARS NUMBER
$BN_U$	0.007	0.083	0.001	0.003	0.608	0.000	0.000	0.111	0.015	0.007
$BN_{CIA}$	0.020	0.084	0.001	0.002	0.608	0.000	0.000	0.112	0.063	0.013

for node "PUBLIC MEANS USE", these values are near zero in both cases, we argue that  $BN_U$  is better to reduce the uncertainty for the node *BIKES NUMBER* and for its Markov Blanket. Table 3<sup>4</sup> shows some well known measures of associ-

<sup>4</sup> The Goodman-Kruskal Gamma, the Kendall's tau-b and the Stuart's tau-c coefficients are suitable for ordinal variables.



ation between BIKES NUMBER and SPORT and CARSHARING and MOBILE NET USE, respectively. As expected,  $BN_U$  outperforms  $BN_{CIA}$ . In addition, under the CIA assumption, when the variables are ordinal the sign of the association is uncorrectly estimated. To provide an analysis of robustness, the imputation procedure

**Table 3** Association measures between BIKES NUMBER - SPORT and CARSHARING - MOBILE NET USE. The last four columns show the minimum, the maximum, the median and the standard deviation values of the same measures over 100 replications

	True	$BN_U$	$BN_{CIA}$	Min	Max	Median	SD	
BIKES NUMBER - SPORT	Phi ( $\phi$ )	0.241	0.264	0.218	0.074	0.482	0.265	0.083
	Contingency coefficient ( $C$ )	0.234	0.255	0.213	0.073	0.434	0.256	0.074
	Cramer's ( $V$ )	0.170	0.186	0.154	0.052	0.186	0.188	0.059
	Goodman-Kruskal Gamma ( $G$ )	0.353	0.324	-0.681	-0.223	0.590	0.279	0.174
	Kendall's tau-b ( $\tau_b$ )	0.213	0.171	-0.193	-0.100	0.353	0.142	0.097
	Stuart's tau-c ( $\tau_c$ )	0.184	0.123	-0.106	-0.069	0.248	0.102	0.068
CARSHARING - MOBILE NET USE	Phi ( $\phi$ )	0.090	0.054	0.037	0.000	0.117	0.046	0.030
	Contingency coefficient ( $C$ )	0.090	0.054	0.037	0.000	0.116	0.046	0.030
	Cramer's ( $V$ )	0.090	0.054	0.037	0.000	0.117	0.046	0.030

has been replicated 100 times by random sampling 100 different probability tables from those belonging to the likelihood ridge. The main results are shown in the last four columns of Table 3, reporting the minimum, the maximum, the median value and the standard deviations of the proposed association measures based on the 100 imputed data set. As one can observe, the median value of the interval is near the true one for almost all coefficients under consideration.

## References

- Conti P L., Marella, D., Vicard, P., Vitale, V. (2020). Multivariate Statistical Matching using Graphical Modeling. *Inter J Approx Reas* 130: 150–169.
- D’Orazio M, Di Zio M, Scanu M (2006a) Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *J Off Stat* 22: 137–157
- Dias F F, Lavieri P S, Garikapati V M, Astroza S, Pendyala R M, Bhat C R: A behavioral choice model of the use of car-sharing and ride-sourcing services. *Transportation* 44, 6: 1307–1323 (2017)
- Endres E, Augustin T (2016) Statistical matching of Discrete Data by Bayesian Networks. *JMLR: Workshop and Conference Proceedings* 52: 159–170
- Kim Sung-Ho, Kim Seong-Ho (2006) A Note on Collapsibility in DAG Models of Contingency Tables. *Scand J Stat* 33: 575–590
- Okner B (1972) Constructing a new data base from existing microdata sets: the 1966 merge file. *Ann Econ Soc Meas* 1: 325–342
- Pearl J (1998) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
- Rässler S (2002) Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Springer, New York
- Singh AC, Mantel H, Kinack M, Rowe G (1993) Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. *Surv Methodol* 19: 59–79

# Modeling school managers challenges in the pandemic era with Bayesian networks

## *Un modello per le sfide dei dirigenti scolastici nell'era pandemica con le reti bayesiane*

Maria Chiara De Angelis and Flaminia Musella and Paola Vicard

**Abstract** Due to the dramatic health situation caused by the COVID-19 pandemic, the emergency remote teaching during lockdown lasted longer in Italy than in other countries. The scholastic pattern has been deeply shocked by the mandatory teaching modalities requiring digital transformation processes in a framework where the digital divide is deep. Organizational processes have been reshaped with an amazing effort of school managers, often moved by personal initiatives. The paper shows results of a multitarget research carried out during the Italian lockdown aiming at animating the debate around school from multi-actors perspectives and at supporting policies. Here, we focus on the managerial challenges that may have been useful for improving the setting up of internal processes.

**Abstract** *A causa della drammatica situazione sanitaria causata dalla pandemia COVID-19, l'insegnamento a distanza di emergenza in Italia durante il confinamento è durato di più rispetto ad altri paesi. Il modello scolastico è rimasto profondamente scioccato dalle modalità didattiche obbligatorie che hanno richiesto un processo di digitalizzazione in una realtà ove il divario digitale è profondo. I processi organizzativi sono stati rimodellati con uno straordinario impegno dei dirigenti scolastici, spesso mossi da iniziative personali. Il presente articolo presenta i risultati di una ricerca multitarget condotta durante il lockdown italiano con l'obiettivo sia di stimolare il dibattito sulla scuola da prospettive multi-attori sia di sostenere le politiche. Nel paper, ci concentriamo solo sulle sfide manageriali che potrebbero essere state utili per migliorare l'impostazione dei processi interni.*

**Key words:** digital-divide, Bayesian networks, scenarios analysis

---

Maria Chiara De Angelis

Link Campus University, Via del Casal di S. Pio V, 44 - 00165 Roma e-mail: mc.deangelis@unilink.it

Flaminia Musella

Link Campus University, Via del Casal di S. Pio V, 44 - 00165 Roma e-mail: f.musella@unilink.it

Paola Vicard

Università Roma Tre, Via S. D'Amico, 77 - 00145 Roma e-mail: paola.vicard@uniroma3.it

## 1 Research framework and motivation

The starting raising of COVID-19 pandemic in Italy, like in most world countries, caused lockdown with the related introduction of the emergency remote teaching for schools and higher education. At the beginning, distance learning has been the unique solution for keeping students on learning for a while, and it has been still chosen in some time frames during the second phase of the emergency. It's a matter of fact that education, together with health and economy, are the main areas that globally have been suffering the most. We believe the effect of remote teaching has been amplified in Italy due to the well-known digital divide and the contrast between teachers' digital skills criticism vs teachers' promotion digital educational experimentation [4], splitting the reality in a double-speed country. As a consequence, due to the relevant issue, a national survey about the school remote experience during the lockdown has been carried out with the aim of understanding practices and emotional impacts of the experience on the society. The motivation of the research arises from the intent of informing the debate on post-emergency schools and, consequently, the education policies. Due to the different stakeholders involved, the study has been conducted in a multitarget perspective and has reached different actors such as school managers, students, teachers and parents. The web-survey has been carried out between May and July 2020 involving 474 school managers, 3444 teachers, 787 students, 2116 parents, for a total of 6821 interviews. This paper focuses on the results of remote school teaching from the school managers point of view. The main research question addressed is "what have been the challenges of school managers for improving organizational process in the school". The paper is organized as follows. Section 2 addresses the statistical modeling methodology. Section 3 shows the survey and preliminary results. Conclusions are addressed in Section 4.

## 2 Basics on Bayesian networks and structural learning

A Bayesian network (BN) represents a multivariate probability distribution of a set of variables by means of a directed acyclic graph (DAG). A directed graph is a mathematical object made of a finite set of *vertices*, also called *nodes*, representing random variables, and a set of directed *edges* displaying direct relevance of one variable to another. A directed graph is acyclic if it does not contain directed cycles. Independence relations in the joint distribution can be read off the DAG by using the *d*-separation criterion [14]. A BN can be built manually by experts if the dependence structure is known, otherwise the network has to be learnt directly from data by means of efficient algorithms [1, 2, 13]. The *structural learning* can be mainly supported through two approaches: *scoring and searching* [6, 3, 5] spanning the space of all possible models and choosing the one maximising a given score function, and *constraint-based* iteratively checking (conditional) independences by performing statistical tests on the data. Among the latter group, the most popular is

the *PC algorithm* [17] that is a stepwise backward method starting from a database and providing in output an equivalence class of models whose pictorial representation is a hybrid graph (*partially* DAG -PDAG) depicting directed edges in presence of v-structures and undirected edges, otherwise. The v-structure is a special 3-nodes  $(X_i, X_j, X_k)$  configuration such that  $X_i$  and  $X_j$  are not independent given  $X_k$ ; it is graphically represented as  $X_i \rightarrow X_k \leftarrow X_j$ . The main steps of PC algorithm are: (i) skeleton identification by a set of recursively performed independence tests based on conditional-cross entropy, *CCE*; (ii) v-structures identification on the basis of test results (PDAG as output); (iii) the PDAG extension to a DAG by orienting the rest of undirected links without producing additional v-structures and cycles.

During the last years, the PC algorithm has become a reference point and a benchmark for developing new constraint-based strategies [15, 7, 9, 11]. In particular, since in many observational studies variables are mixed (nominal-ordinal) categorical variables, the *Nominal Ordinal PC* (NOPC) algorithm has been proposed [12]. The NOPC procedure requires the following four steps: (i) set the variables type, nominal or ordinal; (ii) the skeleton of the graph is found by properly checking marginal and conditional independencies between pair of nodes, according to the variable typology as listed in Table 1; (iii) PDAG identification and (iv) DAG extension.

**Table 1** Test automatically selected by the the NOPC procedure according to the variable pair.

Class X	Class Y	CI Test
Nominal	Nominal	<i>CCE</i>
Nominal	Ordinal	<i>Kruskal – Wallis</i>
Binary	Ordinal	<i>Wilcoxon</i>
Ordinal	Ordinal	<i>Jonckheere – Terpstra</i>

### 3 The survey

The web-survey was carried out by a questionnaire arranged in 5 sections that are: profiling, organizational aspects, teaching methodology, tools and access, satisfaction. Many items have been observed for every section but a selection, eventually based on multivariate techniques of data reduction, brings to consider fewer variables in the model. Table 2 provides a synthesis of variables and corresponding modalities.

Only fully observed units were considered (238). Respondents are mainly female (70%) between 51 and 60 (50%) and mostly working in the public and comprehensive schools of Lombardia (17%) and Lazio Regions (11%). This bias may be linked both to the greater coverage of the territory of central Italy by the research unit and the high level of compliance on the part of Lombard school principals due to the

**Table 2** Variables/Nodes in the model

Section	Item (Variable/Node)	Modalities
Profiling	Gender	M, F
Profiling	Age class	<40; 40-50; 50-60; >60
Profiling	Geographical area	North, Center, Island, South
Profiling	School level	comprehensive school, other
Organizational aspects	Timeliness	less than a week, 1 weeks, 2 weeks, more than 2 weeks
Organizational aspects	All Matters	Yes, No
Organizational aspects	Involving teaching capability problems	none, few, quite, a lot
Organizational aspects	IT project	Yes, No
Organizational aspects	Guidelines	Yes, No
Teaching methodology	Interaction promotion	5-point Likert
Teaching methodology	Staff-manager coordinator	5-point Likert
Teaching methodology	Teachers' appreciation technologies	5-point Likert
Teaching methodology	Internal communication organization	5-point Likert
Teaching methodology	Remote teaching evaluation	Yes, No
Teaching methodology	Teachers' tutoring	Yes, No
Tools and access	Teachers' access problems	Never, Sometimes, Always
Tools and access	Students' access problem	Never, Sometimes, Always
Satisfaction	Priority redesign	5-point Likert
Satisfaction	Decision process redesign	5-point Likert
Satisfaction	Internal communication improvement	5-point Likert
Satisfaction	School-family communication improvement	5-point Likert
Satisfaction	Emerging teachers' needs	5-point Likert
Satisfaction	Teachers' willing	5-point Likert
Satisfaction	Families' willing	5-point Likert
Satisfaction	Teachers' cooperation	5-point Likert
Satisfaction	Organizing process improvement	5-point Likert

impact of pandemic on their territory. The e-research reached 3% of the population (considering 8094 school principals employed in the school in the scholastic year 2019-2020 - Open Data MIUR). This can be considered a good response rate since it is significantly larger than that usually characterizing e-research, generally being around 1% [8].

### 3.1 The model results

The resulting model, learnt by NOPC algorithm in R and managed in Hugin, is shown in Figure 1.

To verify the accuracy of the model, a sample of 250 observations has been generated according to the network dependencies. Then this sample has been used to validate the model in terms of predicting the *organizational processes improvement* on the basis of the receiver operating characteristic (ROC - [10]). The curve, plotting the true positive rate (sensitivity) against the true negative rate (1-specificity), describes an area (AUC) interpretable as an overall measurement of model perfor-

Modeling school managers challenges in the pandemic era with Bayesian networks

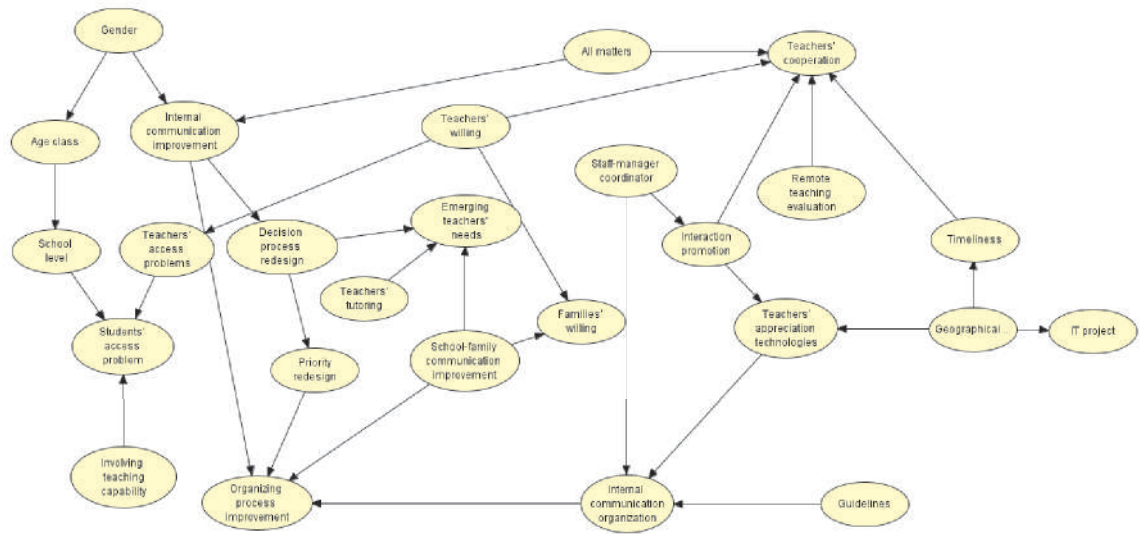


Fig. 1 Model for managing the improvement of organizational processes.

mance and varying between 1 (a model with a perfect discriminant capacity) and 0.5 (a model with a null discriminant capacity). The AUC for the learnt model with respect to *organizational processes improvement* variable is 0.87 that, according to [16] interpretation, stands for a good accuracy of the model.

#### 4 Conclusion and discussion

Research question may be addressed by performing a value of information analysis. Given a BN model and a hypothesis variable, this analysis allows the user to identify variables more informative (in terms of mutual information) with respect to the hypothesis variable. The analysis has been carried out to highlight the most sensitive factors for improving organizational processes (*Organizing Process Improvement*). The most significant challenges are related to communication management and teachers' needs listening as discussed below. A scenario instantiating the level of *internal communication organization* to its maximum, makes the probability of the highest level of *organizational processes improvement* increase by 80%; a scenario setting the highest level of *priority redesign* at 100% produces an improvement of about 66% in the probability of the maximum level of *organizing processes improvement*; a scenario developed with the aim to maximize the probability linked to the highest level of *emerging teachers' needs* allows to increase of 23% the probability of the highest level of *organizing processes improvement*.

## Acknowledgment

This research has been supported by the project "Didattica a distanza ai tempi del Covid-19" by DiTES research center of Link Campus University in cooperation with University of Roma Tre, ANP, Forum Associazioni Familiari and AIDR.

## References

1. Buntine, W.: Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, **2**, 159–225 (1994)
2. Buntine, W.: A guide to the literature on learning probabilistic networks from data. *IEEE Transaction on Knowledge and Data Engineering*, **8** (2), 195–210 (1996)
3. Heckerman, D.: A tutorial on Learning with Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research (1995)
4. Capogna, S. De Angelis, M.C., Musella, F.: Exploring Practices, Problems and Cultural Challenges of Italian Teachers in the Digital Era. *Scuola democratica, Learning for Democracy*, **2**, 259–284 (2020)
5. Chickering, D. M., Heckerman, D., Meek, C. & Madigan, D.: Learning Bayesian networks is NP-hard. Technical Report MSR-TR-94-17, Microsoft Research, Microsoft Corporation, Redmond, Washington (1994)
6. Cooper, G. & Herskovits, E.: A Bayesian method for constructing Bayesian belief networks from databases. *Machine Learning*, **9**(4), 309–47 (1992)
7. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P.: Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, **47** (11), 1-26 (2012)
8. MacElroy, B.: Measuring response rates in online surveys. *Quirk's Marketing Research* **583**, (2000)
9. Marella, D., Musella, F., & Vicard, P.: Learning Bayesian networks in complex survey sampling. In *Proceedings of SIS conference* (2014)
10. Metz, C.E.: Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8** (4), pp. 283-298 (1978)
11. Musella, F.: A PC algorithm variation for ordinal variables. *Computational Statistics*, **28** (6), 2749–2759 (2013)
12. Musella, F.: A proposal for learning Bayesian networks from categorical variables. In *SIS 2015 meeting Statistics and Demography: the Legacy of Corrado Gini* (2015)
13. Neapolitan, R. E.: *Learning Bayesian Networks*. Pearson Prentice Hall, NewHaven, Connecticut (2003)
14. Pearl, J.: A constraint-propagation approach to probabilistic reasoning. In *Uncertainty in Artificial Intelligence*. (ed.L.N. Kanal and J.F. Lemmer), North Holland, Amsterdam, The Netherlands (1986)
15. Steck, H.: *Constraint-Based Structural Learning in Bayesian Networks using Finite Data*. PhD thesis, Institut für Informatik der Technischen Universität München (2001)
16. Swets K.A.: Measuring the accuracy of diagnostic systems. *Science* **240** 1285–1293 (1988)
17. Spirtes, P., Glymour, C., & Scheines, R.: *Causation, Prediction, and Search* (2nd edn). MIT Press, Cambridge, Massachusetts (2000)

# Structural learning of mixed directed acyclic graphs: a copula-based approach

## *Apprendimento di grafi direzionati con variabili miste: un approccio basato su copule*

Federico Castelletti

**Abstract** We consider a system of random quantities, including continuous, discrete, ordinal and binary variables, giving rise to multivariate data. We adopt Directed Acyclic Graphs (DAGs) to represent dependence relations between variables that need to be inferred from the data. To accommodate different types of variables, we assume that each has been generated by a latent counterpart and that the joint distribution of the latent variables belongs to a Gaussian DAG family. We adopt a copula-based approach to effectively model dependence parameters (covariance matrix) separately from the parameters of the marginal distributions and build on a Bayesian methodology for DAG determination.

**Abstract** *Si considera una collezione di variabili aleatorie, tra cui variabili continue, discrete, ordinali e binarie, ed un dataset di osservazioni multivariate. Si adotta un grafo aciclico direzionato (DAG) per rappresentare le relazioni di dipendenza tra variabili che in quanto ignote è necessario stimare attraverso i dati disponibili. Per modellare congiuntamente variabili di diversa tipologia, si assume che ciascuna sia stata generata da una variabile latente e che la distribuzione congiunta delle latenti appartenga ad una famiglia normale che rispetta i vincoli di dipendenza imposti da un DAG. Si adotta un modello basato su copule per modellare i parametri di dipendenza tra variabili (ossia la matrice di covarianza) separatamente dai parametri delle distribuzioni marginali attraverso lo sviluppo di una metodologia bayesiana per l'apprendimento del DAG.*

**Key words:** Graphical model; Mixed variables, Gaussian copula, Well-being data

---

Federico Castelletti  
Università Cattolica del Sacro Cuore, Dipartimento di Scienze Statistiche, Largo Gemelli 1, Milan,  
e-mail: federico.castelletti@unicatt.it



## 1 Introduction

Graphical models represent a well established tool for modelling and inferring from the data dependence relations among variables; see for instance [8] and [5]. Most of the literature has focused on parametric graphical models, specifically tailored to Gaussian or categorical (multinomial) data. While the former are widely employed in biology and in particular genomics, the latter are more common in social sciences. In both cases, the underlying graphical structure, which encodes conditional independencies between variables, imposes specific constraints to the sampling distribution of the data and the allied model parameters. From a Bayesian viewpoint, this requires the adoption of appropriate prior distributions defined on the space of constrained parameters. In the Gaussian framework, G-Wishart and DAG Wishart distributions were designed specifically for covariance matrices Markov w.r.t. undirected graphs (UGs) and directed acyclic graphs (DAGs) respectively; see in particular [10] and [1].

Mixed variables, that is variables belonging to different parametric families, introduce some further complications because of the complex support of the joint distribution. In this setting, few attempts have been focused on jointly modelling categorical and continuous variables by means of conditional Gaussian distributions coupled with undirected graphical models; see for instance [4]. However, the parameterization adopted by these methods is not flexible enough to include other types of data, for instance discrete and ordinal, the latter being quite common in social sciences and psychology. If interest lies in estimating the *association parameters* of the joint density rather than the parameters of the marginal distributions, *copula models*, which allows to model the two sets of parameters separately, can provide an effective solution. In addition, *semiparametric* copula models lacks any parametric assumption on the marginal c.d.f.'s which are estimated through their empirical distributions [7]. Few contributions on copula Gaussian graphical models based on UGs are available in the literature; see for instance [6].

In this work we rely on DAGs which are particularly tailored for modeling dependencies between variables and also flexible enough to accommodate several types of dependence structures, including regression, covariate-adjusted models as limiting cases. In addition, they provide a powerful tool for causal reasoning; see [9]. Specifically, we consider a *Gaussian copula* model where the association parameter (covariance matrix) reflects the conditional independencies imposed by the DAG, leading to our *Gaussian copula DAG model*. We then develop a Bayesian strategy for structural learning of DAGs and parameter estimation and propose an MCMC scheme for posterior inference. An application to the analysis of lifestyle and well-being data is finally presented.

## 2 Model formulation

Let  $X_1, \dots, X_q$  be a collection of continuous, discrete, binary or ordinal variables,  $\mathcal{D} = (V, E)$  a DAG, where  $V = \{1, \dots, q\}$  is a set of nodes (each associated to one of the  $q$  variables),  $E \subseteq V \times V$  a set of directed edges. Let also  $Z_1, \dots, Z_q$  be  $q$  latent random variables. We assume

$$Z_1, \dots, Z_q | \boldsymbol{\Omega}, \mathcal{D} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Omega}^{-1}), \quad \boldsymbol{\Omega} \in \mathcal{P}_{\mathcal{D}}, \quad (1)$$

where  $\boldsymbol{\Omega}$  denotes the precision matrix (inverse of the covariance matrix  $\boldsymbol{\Sigma}$ ) and  $\mathcal{P}_{\mathcal{D}}$  is the set of all s.p.d. precision matrices Markov w.r.t. DAG  $\mathcal{D}$ . A link between each observed variable  $X_j$  and its latent counterpart  $Z_j$  is introduced by assuming

$$X_j = F_j^{-1} \{ \Phi(Z_j) \}, \quad (2)$$

where  $F_j^{-1}$  is the pseudo inverse c.d.f. of  $X_j$ ,  $\Phi(\cdot)$  the c.d.f. of a standard normal distribution. The joint c.d.f. of  $X_1, \dots, X_q$  can be written as

$$P(X_1 \leq x_1, \dots, X_q \leq x_q | \boldsymbol{\Omega}, F_1, \dots, F_q) = \Phi_q(\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_q(x_q)) | \boldsymbol{\Omega}), \quad (3)$$

where  $\Phi_q(\cdot | \boldsymbol{\Omega})$  denotes the c.d.f. of  $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Omega}^{-1})$  in (1). Equation (3) depends on the marginal distributions  $F_1, \dots, F_q$ . A semiparametric estimation strategy replaces  $F_j$  with the corresponding empirical estimates  $\hat{F}_j(k_j) = n^{-1} \sum_{i=1}^n \mathbb{1}(x_{i,j} < k_j)$ , where  $k_j \in \text{unique}\{x_{1,j}, \dots, x_{n,j}\}$ . Let now  $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,q})^\top$ ,  $i = 1, \dots, n$ , be  $n$  i.i.d. samples from (3) and  $\mathbf{X}$  the  $(n, q)$  data matrix (row-binding of the  $\mathbf{x}_i$ 's). Since the  $F_j$ 's are non decreasing, for each pair of distinct observations  $x_{i,j}$  and  $x_{l,j}$ , if  $x_{i,j} < x_{l,j}$  then  $z_{i,j} < z_{l,j}$ . Therefore observing  $\mathbf{X}$  implies that the latent data  $\mathbf{Z}$  must lie in the set

$$A(\mathbf{X}) = \{ \mathbf{Z} \in \mathbb{R}^{n \times q} : \max\{z_{k,j} : x_{k,j} < x_{i,j}\} < z_{i,j} < \max\{z_{k,j} : x_{i,j} < x_{k,j}\} \} \quad (4)$$

and one can take the occurrence of such an event as the data; see also [7]. Thus, one can define the *extended rank likelihood* as

$$p(\mathbf{Z} \in A(\mathbf{X}) | \boldsymbol{\Omega}) = \int_{A(\mathbf{X})} f(\mathbf{Z} | \boldsymbol{\Omega}) d\mathbf{Z}, \quad (5)$$

where  $f(\mathbf{Z} | \boldsymbol{\Omega}) = \prod_{i=1}^n f(z_1, \dots, z_q | \boldsymbol{\Omega})$  and  $f(z_1, \dots, z_q | \boldsymbol{\Omega})$  corresponds to the Gaussian density of distribution in (1).

## 3 Bayesian inference

The likelihood function in (5) depends on the precision matrix  $\boldsymbol{\Omega}$  Markov w.r.t. DAG  $\mathcal{D}$  and implicitly on the DAG itself. Following a Bayesian perspective, we then

proceed by assigning a prior distribution to  $(\boldsymbol{\Omega}, \mathcal{D})$  that we structure as  $p(\boldsymbol{\Omega}, \mathcal{D}) = p(\boldsymbol{\Omega} | \mathcal{D})p(\mathcal{D})$ . In particular, conditionally on a given  $\mathcal{D}$  we can assign a prior to  $\boldsymbol{\Omega}$  through a DAG-Wishart distribution on the corresponding Cholesky parameters  $(\mathbf{D}, \mathbf{L})$ , i.e.  $p(\mathbf{D}, \mathbf{L} | \mathcal{D})$ , which induce the re-parameterization  $\boldsymbol{\Omega} = \mathbf{L}\mathbf{D}^{-1}\mathbf{L}^\top$ ; see also [1] for details and [2] for hyper-parameters choice. Moreover, we can assign a prior to DAG  $\mathcal{D}$  through independent Bernoulli distributions on the 0-1 elements of its adjacency matrix, where each element indicates the absence/presence of an edge in the DAG. We then focus on the joint posterior distribution

$$p(\mathbf{D}, \mathbf{L}, \mathcal{D} | \mathbf{Z} \in A(\mathbf{X})) \propto p(\mathbf{Z} \in A(\mathbf{X}) | \mathbf{D}, \mathbf{L}, \mathcal{D})p(\mathbf{D}, \mathbf{L} | \mathcal{D})p(\mathcal{D}), \quad (6)$$

where we adopt the re-parameterization  $\boldsymbol{\Omega} \mapsto (\mathbf{D}, \mathbf{L})$  to write the rank likelihood in (5) and we now emphasize the dependence on DAG  $\mathcal{D}$ .

Notice that, if the latent variables  $Z_1, \dots, Z_q$  were observed, one could perform inference on  $(\mathbf{D}, \mathbf{L}, \mathcal{D})$  using a standard MCMC scheme such as the one presented in [3] for Gaussian DAG model selection and based on a Partial Analytic Structure (PAS) algorithm. However, because the latent data are known only relative to the event  $A(\mathbf{X})$ , a suitable adaptation of the MCMC scheme is required to sample them from their full conditional distributions which correspond to truncated Normal distributions; we omit details. Output of the MCMC algorithm is an approximated posterior distribution over the space of DAGs and Cholesky parameters (equivalently, covariance/precision matrices). This can be used to compute summaries of interest, such as posterior probabilities of inclusion for specific edges or Bayesian model averaging (BMA) estimates of correlation coefficients between variables.

## 4 Application to lifestyle and well-being data

We apply the proposed methodology to well-being data collected from the Global Work Life balance survey promoted by United Nations (data and further information can be found at <https://www.kaggle.com/ydalat/lifestyle-and-wellbeing-data>). The scope of this survey is to evaluate how people thrive in both professional and personal lives. To this end, several dimensions related to healthy habits, emotions and social relations are measured, together with the perceived stress level. The complete dataset includes measurements across years 2015-2020 of 20 ordinal variables (with levels ranging in 1-5 or 1-10) each measuring closeness of a subject w.r.t. to one perceived dimension, besides age and gender. We include in our analysis the  $n = 459$  observations available for year 2020.

As a summary of the MCMC output, we report in Table 1 the estimated (Bayesian model averaging) correlation matrix of twelve selected variables. For simplicity, only coefficients whose absolute value is larger than 0.2 are reported. Most variables are positively correlated, with the exception of variable “to do list” (*how well do you complete your weekly to-do list in a range 1-5?*) which as expected is negatively correlated to the perceived level of stress. A graph estimate summarizing the

posterior distribution over DAGs and obtained by including edges whose posterior probability exceeds 0.5 is finally reported in Figure 1.

	stress	vacation	friends	help	social	achieve	donation	to do list	flow	life vision	awards	passion
stress												
vacation												
friends		0.26										
help			0.30									
social			0.29	0.38								
achieve				0.38								
donation				0.31								
to do list	-0.37											
flow						0.27		0.29				
life vision												
awards						0.45	0.30					
passion					0.26	0.32			0.45	0.28		

Table 1: Estimated correlation matrix of twelve selected variables; only coefficients such that  $|\rho_{i,j}| > 0.2$  are reported.

## References

1. Cao, X., Khare, K., Ghosh, M.: Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *The Annals of Statistics* **1**(47), 319–348 (2019)
2. Castelletti, F., Consonni, G.: Bayesian inference of causal effects from observational data in Gaussian graphical models. *Biometrics* **77**(1), 136–149 (2020)
3. Castelletti, F., Consonni, G.: Bayesian causal inference in probit graphical models. arXiv preprint (2020)
4. Cheng, J., Li, T., Levina, E., Zhu, J.: High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics* **26**(2), 367–378 (2017)
5. Cowell, R. G., Dawid, P. A., Lauritzen, S. L., Spiegelhalter, D. J.: *Probabilistic Networks and Expert Systems*. New York: Springer (1999)
6. Dobra, A., Lenkoski, A.: Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics* **5**(2A), 969–993 (2011)
7. Hoff, P.D.: Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics* **1**(1), 265–283 (2007)
8. Lauritzen, S. L.: *Graphical Models*. Oxford University Press, Oxford (1996)
9. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge (2000)
10. Roverato, A.: Hyper Inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* **29**(3) (2002)

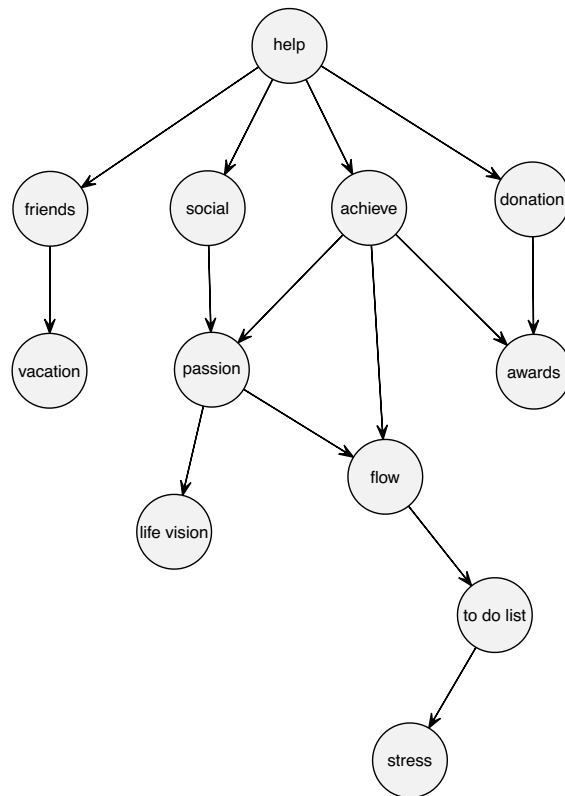


Fig. 1: Estimated graph of twelve selected variables, obtained by including edges whose posterior probability is larger than 0.5.

# Inference on Markov chains parameters via Large Deviations ABC

## *Inferenza sui parametri di Catene di Markov mediante Large Deviations ABC*

Cecilia Viscardi, Fabio Corradi, Michele Boreale, Antonietta Mira

**Abstract** We propose a method for Bayesian inference on the parameters governing the transition probabilities of finite state Markov chains. We address the difficulty of deriving the parameters' posterior distribution when the likelihood function is unavailable or computationally demanding to evaluate. The approach is an extension of the Large Deviations Approximate Bayesian Computation already proposed for i.i.d random variables. The method is developed by accommodating an information theoretic formulation of the Large Deviations Theory into Approximate Bayesian Computation (ABC). By contrast to the customary ABC, this approach avoids discarding parameter values having an (exponentially) small probability of producing simulation outcomes close to the observed data. We experimentally evaluate our method through a toy example.

**Abstract** *Proponiamo un metodo di inferenza Bayesiana per l'apprendimento dei parametri che governano le probabilità di transizione in catene di Markov a stati finiti qualora la funzione di verosimiglianza non è derivabile analiticamente ed una sua valutazione è computazionalmente costosa. In particolare, estendiamo alle catene di Markov un metodo di Approximate Bayesian Computation (ABC) basato sulla teoria delle Grandi Deviazioni proposto per variabili discrete i.i.d.. Il risultato è ottenuto integrando la teoria delle grandi deviazioni entro ABC. Questo metodo consente di non scartare le proposte di parametri che hanno una (esponenzialmente) piccola probabilità di produrre dati simulati simili a quelli osservati. Il metodo è illustrato attraverso un semplice esempio.*

**Key words:** ABC, Large deviations, Parametric Markov chains, Sample degeneracy, Method of Types.

## 1 Introduction and preliminary concepts

Parametric Markov chains (pMC) are discrete time Markov chains whose transitions probabilities are expressed as polynomials of real-valued parameters [4, 6]. Statistical methods for inferring the parameters governing such transition probabilities have been proposed, both from a classical and Bayesian viewpoint. Here we propose a method for deriving the parameters' posterior distributions when the evaluation of the joint probability function

of the Markovian sequence given the parameters is infeasible. In particular, we extend to finite state pMC the Large Deviations Approximate Bayesian Computation (LD-ABC) proposed in [13, 11] for i.i.d. discrete data. Generally speaking, Approximate Bayesian Computation (ABC) [9] is a class of likelihood-free methods allowing Bayesian inference when the likelihood function is intractable and only requiring the ability of simulating pseudo-data from a *simulator*, i.e., a probabilistic program reproducing the stochastic data generating process. The LD-ABC method represents a novel proposal for improving the ABC performances by mitigating the *sample degeneracy* problem in ABC. The method enhance the ABC likelihood resorting to an information theoretic formulation of Large Deviations Theory (LDT) based on the Method of Types [3].

### Preliminary concepts

Let  $\{X_t\}$  be a *stationary* parametric Markov process taking values in a finite set  $\mathbb{A} \triangleq \{a_1, \dots, a_k\}$  with cardinality  $k$ . For simplicity the elements of  $\mathbb{A}$  will be hereafter denoted by their labels  $\{1, \dots, k\}$ . The Markov process can be characterized by its *doublet probability distribution* (dpd),  $P_\theta$ , defined as a non-negative matrix of order  $k \times k$  inducing a probability measure  $\Pr\{\cdot, \cdot | \theta\}$  over  $\mathbb{A}^2 \triangleq \mathbb{A} \times \mathbb{A}$ . Thus, denoted by  $P_\theta(ij)$ , the entries of  $P_\theta$  are

$$P_\theta(ij) \triangleq \Pr\{X_t = i, X_{t+1} = j | \theta\} \quad \forall (i, j) \in \mathbb{A}^2$$

and sum to 1. The subscript  $\theta$  indicates the dependence from the parameter (or vector of parameters)  $\theta$ , object of our inference.

Let us denote by  $\Delta^{k^2-1}$  the  $(k^2 - 1)$ -simplex, i.e., the set of possible dpd over  $\mathbb{A}^2$ , and by  $\mathcal{M}(\mathbb{A}^2) \subset \Delta^{k^2-1}$  the set of the stationary dpd. Each  $P_\theta \in \mathcal{M}(\mathbb{A}^2)$  is characterized by entries such that  $\sum_{j \in \mathbb{A}} P_\theta(ij) = \sum_{j \in \mathbb{A}} P_\theta(ji)$ ,  $\forall i \in \mathbb{A}$ . This implies that the probability distribution over  $\mathbb{A}$ ,  $p_\theta \triangleq \{p_\theta(i) = \sum_{j \in \mathbb{A}} P_\theta(ij), \forall j \in \mathbb{A}\}$ , is invariant along the process and  $P_\theta$  captures all the relevant information about it. In fact, the *state transition matrix* of the pMC,  $Q_\theta$ , is the stochastic matrix of order  $k \times k$  composed by entries retrieved from  $P_\theta$

$$q_\theta(ij) \triangleq \Pr\{X_{t+1} = j | X_t = i, \theta\} = \frac{P_\theta(ij)}{p_\theta(i)} \quad \forall (i, j) \in \mathbb{A}^2$$

and  $p_\theta$  is a (normalized) row eigenvector of  $Q_\theta$  corresponding to eigenvalue 1:

$$(p_\theta Q_\theta)_j = \sum_{i \in \mathbb{A}} p_\theta(i) q_\theta(ij) = \sum_{i \in \mathbb{A}} P_\theta(ij) = p_\theta(j) \quad \forall j \in \mathbb{A}.$$

## 2 The Method of Types for Markov chains and LDT

The Method of Types (MoT)[3] is a powerful tool shifting the focus from a vector of random variables to a lower dimensional vector: the *type*. Originally, MoT has been proposed for i.i.d. random variables and the 1st order type was defined as the empirical distribution of a sequence of random variables. Here we consider an extension: the 2nd order type which is suitable for observations modelled as Markov chains.

Given a Markov process  $\{X_t\}$  with dpd  $P_\theta$  and an observed sample path  $x^n = x_1, \dots, x_n$  from the Markov process, the 2nd order type [3] is defined by

$$T_{x^n}^{(2)}(i, j) \triangleq \frac{1}{n-1} \sum_{t=1}^{n-1} \mathbb{1}\{x_t = i, x_{t+1} = j\} \quad \forall (i, j) \in \mathbb{A}^2.$$

This type can be thought as a matrix of order  $k \times k$  representing an empirical estimate of  $P_\theta$ . An alternative definition is based on the *cyclic convention* that poses the  $(n+1)$ -th element of the path equal to  $x_1$  and ensures the stationarity of the 2nd order type obtained from the  $n$  terms. This definition allows for establishing the MoT formulation of the Large Deviations principle for Markov chains.

LDT is concerned with probabilities of *rare events* going to zero with an exponential decay. A well-known result is the Sanov's Theorem (see [2, Th. 11.4.1]) which establishes the *rate function*, i.e., the function quantifying the probability of rare events, for sequences of i.i.d. random variables. Its analog for Markov chains can be found in the Donsker and Varadhan Theorem [5] and has been presented as an application of the MoT by Csiszár [3]. In what follows we let  $D_c(\cdot|\cdot)$  be the *conditional relative entropy* (see [2] for a definition) and  $\log$  be the logarithm to base 2.

**Theorem 1.** *Let  $\{X_t\}$  be a Markov process taking values in the finite set  $\mathbb{A}$ , with stationary doublet probability distribution  $P_\theta \in \mathcal{M}(\mathbb{A}^2)$  and let  $X^n = X_1, \dots, X_n$ . If  $E \subseteq \mathcal{M}(\mathbb{A}^2)$ , then for each  $\theta \in \Theta$*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \Pr\{T_{X^n}^{(2)} \in E | \theta\} = - \inf_{P \in E} D_c(P|P_\theta) = -D_c(E|P_\theta). \quad (1)$$

*Proof.* See [8] for a proof based on an easy counting approach.

### 3 ABC for finite state Markov Chains

Let  $x^n$  be an observed sequence from a Markov process  $\{X_t\}$  taking values in  $\mathbb{A}$  with stationary dpd  $P_\theta$ . In Bayesian framework one is interested in computing the posterior distribution of  $\theta \in \Theta$  given the data  $x^n$  and a prior distribution  $\pi(\cdot)$  over  $\Theta$ :

$$\pi(\theta|x^n) \propto \pi(\theta) \Pr\{X^n = x^n | \theta\}$$

where  $\Pr\{X^n = x^n | \theta\} = p_\theta(x_1) \prod_{t=1}^{n-1} P_\theta(x_{t+1}, x_t) / p_\theta(x_t)$ .

In many statistical applications (e.g. network analysis, epidemiological or genetic models) this probability is analytically intractable or computationally demanding to evaluate. In such cases one should resort to ABC whose key idea is to provide a conversion of samples from the prior distribution into samples from the posterior by rejecting those parameters that, given as input to the *simulator*, produce simulated observations,  $y^n$ , different from the observed data. Rejection ABC (R-ABC) displayed in Algorithm 1 produces samples from an *approximate posterior distribution* introducing three sources of approximation by 1) resorting to an arbitrary distance function  $d(\cdot, \cdot)$ ; 2) introducing a positive tolerance parameter  $\varepsilon$ ; 3) summarizing the observed and the simulated data through summary statistics  $s_x = s(x^n)$  and  $s_y = s(y^n)$  with  $s : \mathbb{A}^n \rightarrow \mathcal{S}$ . The output of the algorithm is a sample of pairs  $(\theta^{(s)}, s_y^{(s)})$  from the following ABC joint posterior distribution

$$\tilde{\pi}(\theta, s_y | s_x) \propto \pi(\theta) \Pr\{s_Y = s_y | \theta\} \mathbb{1}\{d(s_y, s_x) \leq \varepsilon\} \quad (2)$$

which, marginalising out  $s_y$ , i.e., simply discarding the simulated summaries, leads to the *marginal approximate posterior distribution*:



---

**Algorithm 1** Rejection ABC (R-ABC)
 

---

**for**  $s = 1, \dots, S$  **do**  
 Draw  $\theta^{(s)} \sim \pi$   
 Generate  $y \sim P(\cdot | \theta^{(s)})$  from the simulator  
 Accept the pair  $(\theta^{(s)}, s_y^{(s)})$  if  $d(s_y^{(s)}, s_x) \leq \varepsilon$   
**end for**

---

$$\tilde{\pi}(\theta | s_x) \propto \pi(\theta) \sum_{\mathcal{Y}} \Pr\{s_Y = s_y | \theta\} \mathbb{1}\{d(s_y, s_x) \leq \varepsilon\} ds_y = \pi(\theta) \cdot \Pr\{d(s_y, s_x) \leq \varepsilon | \theta\}. \quad (3)$$

The indicator function in (3) does not enable to discriminate between pseudo-data equal to simulated data and pseudo-data just close enough. Thus, it is often replaced by a kernel function, which is a positive function of the distance  $d(s_y, s_x)$ , defined on a compact support and decaying continuously from 1 to 0.

Looking at (3), it is apparent that the probability  $\Pr\{d(s_y, s_x) \leq \varepsilon | \theta\}$  represents the *approximate likelihood*. At each iteration  $s$ ,  $\Pr\{d(s_y, s_x) \leq \varepsilon | \theta\}$  is approximated pointwise by the indicator function or another kernel function defined on a compact support. This crude approximation causes a very large number of rejections leading to one of the major drawbacks of the ABC methods: the sample degeneracy (see [10, Ch. 4] for a discussion of the problem of sample degeneracy in ABC). This typically implies that ABC sampling schemes require a very large number of iterations to get a good approximation of the posterior distribution, especially in the tail area where  $\Pr\{d(s_y, s_x) \leq \varepsilon | \theta\}$  is exponentially small. Here, we speculate that an improvement can be achieved employing a kernel function based on LDT, thus taking into account the exponential decay of  $\Pr\{d(s_y, s_x) \leq \varepsilon | \theta\}$ .

#### 4 LD-ABC for Markov Chains

Let us consider the set  $\Gamma_\varepsilon \triangleq \{P \in \Delta^{k^2-1} : D_c(P || T_x^{(2)}) \leq \varepsilon\}$ . Letting  $y^m = y_1, \dots, y_m$  be a sample path from a pMC with dpd  $P_\theta$ , from Theorem 1 follows that

$$\Pr\{T_{y^m}^{(2)} \in \Gamma_\varepsilon | \theta\} \approx 2^{-mD_c(\Gamma_\varepsilon || P_\theta)} \cdot c. \quad (4)$$

The following Theorem proves that  $D_c(\Gamma_\varepsilon || T_{y^m}^{(2)}) \approx D_c(\Gamma_\varepsilon || P_\theta)$ , as  $m \rightarrow \infty$ .

**Theorem 2.** *Let  $\{Y_t\}$  be a Markov process taking values in the finite set  $\mathbb{A}$  whose stationary dpd is  $P_\theta \in \mathcal{M}(\mathbb{A}^2)$  and let  $Y^m = Y_1, \dots, Y_m$ . Then, under the measure induced by  $P_\theta$*

$$\lim_{m \rightarrow \infty} D_c(\Gamma_\varepsilon || T_{y^m}^{(2)}) = D_c(\Gamma_\varepsilon || P_\theta) \quad a.s. \quad (5)$$

*Proof.* See [11, Appendix D].

From (4) and Th. 2 follows that by setting the 2nd order type as summary statistics and the conditional relative entropy as divergence measure, the probability  $\Pr\{d(s_y, s_x) \leq \varepsilon | \theta\}$  can be approximated by  $2^{-mD_c(\Gamma_\varepsilon || P_\theta)}$ . Meaning that, the indicator function in (2) may be replaced by the following kernel:

$$K_\varepsilon(T_{y^m}^{(2)}) \triangleq \begin{cases} 1 & \text{if } D_c(T_{y^m}^{(2)} || T_x^{(2)}) \leq \varepsilon \\ 2^{-mD_c(\Gamma_\varepsilon || T_{y^m}^{(2)})} & \text{if } D_c(T_{y^m}^{(2)} || T_x^{(2)}) > \varepsilon \end{cases}. \quad (6)$$

Hence, the joint and the marginal ABC approximate posterior distributions become:

$$\tilde{\pi}(\theta, T_{y^m}^{(2)} | T_{x^m}^{(2)}) \propto \pi(\theta) P_{\theta}(T_y^{(2)}) K_{\varepsilon}(T_{y^m}^{(2)}) \tag{7}$$

$$\tilde{\pi}(\theta | T_{x^m}^{(2)}) \propto \pi(\theta) \sum_{T_{y^m}^{(2)} \in \mathcal{T}(m,2)} P_{\theta}(T_{y^m}^{(2)}) K_{\varepsilon}(T_{y^m}^{(2)}) \tag{8}$$

where  $\mathcal{T}(m, 2)$  is the set of the 2nd order types of sequences of length  $m$  from Markov processes taking values in  $\mathbb{A}$ .

In order to sample from (7) we present both an Importance Sampling (IS) and a MCMC scheme displayed in Alg.2 and Alg.3, respectively. Both the algorithms draw parameter values from a proposal distribution on the parametric space,  $q(\cdot)$ , and avoid implicit rejections involving the proposed kernel in the evaluation of the importance weights or of the acceptance ratio. We refer the reader to [10, Ch. 4] for the a description of the standard IS-ABC and MCMC-ABC algorithms.

---

**Algorithm 2** LD-IS-ABC

---

```

for  $s = 1, \dots, S$  do
  Draw  $\theta^{(s)} \sim q$ 
  Draw  $y^{(s)} \sim P(\cdot | \theta^{(s)})$  and compute  $T_{y^{(s)}}^{(2)}$ 
  if  $D_c(T_{y^{(s)}}^{(2)} || T_x^{(2)}) \leq \varepsilon$  then
    Set  $\omega_s = \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$ 
  else
     $\omega_s = 2^{-nD(I_{\varepsilon} || T_{y^{(s)}}^{(2)})} \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$ 
  end if
end for

```

---



---

**Algorithm 3** LD-MCMC-ABC

---

```

for  $s = 1, \dots, S$  do
  Draw  $\theta^* \sim q(\theta^{(s-1)}, \theta^*)$ 
  Draw  $y^* \sim P(\cdot | \theta^*)$  and compute  $T_{y^*}^{(2)}$ 
  Draw  $u \sim \text{Unif}[0, 1]$ 
  if  $u < \min \left\{ 1, \frac{\pi(\theta^*) K_{\varepsilon}(T_{y^*}^{(2)}) q(\theta^*, \theta^{(s-1)})}{\pi(\theta^{(s-1)}) K_{\varepsilon}(T_{y^{(s-1)}}^{(2)}) q(\theta^{(s-1)}, \theta^*)} \right\}$ 
  then
    Assign  $(\theta^{(s)}, T_{y^{(s)}}^{(2)}) \leftarrow (\theta^*, T_{y^*}^{(2)})$ 
  else
    Assign  $(\theta^{(s)}, T_{y^{(s)}}^{(2)}) \leftarrow (\theta^{(s-1)}, T_{y^{(s-1)}}^{(2)})$ 
  end if
end for

```

---

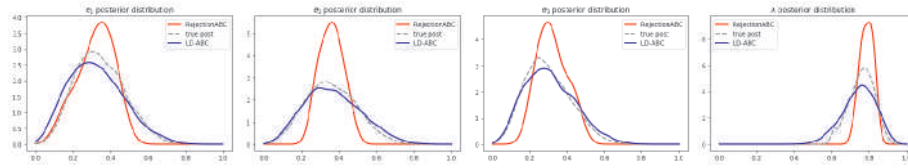
## 5 Toy example

We consider a time series  $X^{60} = X_1, \dots, X_{60}$  from an AR(1) process taking values in  $\mathbb{A} = \{1, 2, 3\}$ . Specifically, we consider the AR(1) process dealt with in [1], where

$$X_t = \begin{cases} X_{t-1} & \text{with probability } \lambda \\ \delta_t & \text{with probability } 1 - \lambda \end{cases}$$

with mixing weight  $\lambda \in [0, 1]$ .  $\delta_t$  is a discrete random variable taking values in  $\mathbb{A}$  with probabilities  $\theta \triangleq (\theta_1, \theta_2, \theta_3) \in \Delta^2$ . Our aim is approximating the posterior distributions of the four parameters  $\theta_1, \theta_2, \theta_3$  and  $\lambda$ . We assume that  $(\theta_1, \theta_2, \theta_3)$  is a priori distributed as a *Dirichlet*(1, 1, 1) and  $\lambda$  as a *Beta*(1, 1). In such a case, despite the complexity of the likelihood function, samples from the true posterior distributions can be obtained through the Importance Sampling scheme, here taken as a benchmark (see [1] for a detailed discussion of the likelihood evaluation and sampling schemes). We ran both R-ABC and LD-ABC with  $m = 120$ ,  $\varepsilon = 0.005$  and  $S = 100,000$ . Figure 1 shows the posterior distributions approximated by the two algorithms and Table 1 displays the  $\widehat{MSE}$  and  $\widehat{MISE}$ , computed by averaging the squared errors for the posterior mean and the integrated squared errors over 100 reruns for both the algorithms. We can see that the LD-ABC outperforms the standard ABC both in terms of point estimates and of posterior distributions approximation. Finally, we evaluate the effects on sample degeneracy looking at the Effective Sample Size (ESS)

(see e.g. [7]): LD-ABC achieves an ESS of 4619 versus the 11 values accepted by R-ABC.



**Fig. 1** Posterior distributions with  $m = 120$  and  $\varepsilon = 0.005$ . The red lines are the posterior densities approximated via R-ABC and the blue lines via LD-ABC. The dashed grey lines are benchmarks obtained by IS.

**Table 1** Squared errors and integrated squared errors averaged over 100 runs.

		$m = 120, \varepsilon = 0.005$			
		$\theta_1$	$\theta_2$	$\theta_3$	$\lambda$
$\widehat{MSE}$	LD	$4.56 \cdot 10^{-4}$	$0.76 \cdot 10^{-4}$	$1.66 \cdot 10^{-4}$	$1.54 \cdot 10^{-4}$
	R	$13.59 \cdot 10^{-4}$	$16.35 \cdot 10^{-4}$	$8.99 \cdot 10^{-4}$	$6.63 \cdot 10^{-4}$
$\widehat{MISE}$	LD	0.0780	0.028	0.0274	0.1922
	R	0.2575	0.3162	0.3679	1.0681

## References

1. Angers, J.F., Biswas, A. & Maiti, R. (2016). Bayesian Forecasting for time series of categorical data. *The Journal of Forecasting*, 36(3), 217-229.
2. Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons.
3. Csiszár, I. (1998). The method of types [information theory]. *IEEE Transactions on Information Theory*, 44(6), 2505-2523.
4. Daws, C. (2004). Symbolic and parametric model checking of discrete-time Markov chains. In *International Colloquium on Theoretical Aspects of Computing* (pp. 280-294). Springer.
5. Donsker, M. D., Varadhan, S. R. S. Asymptotic evaluation of certain Markov process expectations for large time I-III (1975-76). *Comm. Proc. App. Math.*, (28), 1-47, 279-301.
6. Lanotte, R., Maggiolo-Schettini, A., & Troina, A. (2007). Parametric probabilistic transition systems for system design and analysis. *Formal Aspects of Computing*, 19(1), 93-109.
7. Liu JS (2008) *Monte Carlo strategies in scientific computing*. Springer Science & Business Media
8. Natarajan, S. (1985). Large deviations, hypotheses testing, and source coding for finite Markov chains. *IEEE Transactions on Information Theory*, 31(3), 360-365.
9. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12), 1791-1798.
10. Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. Chapman and Hall\CRC.
11. Viscardi, C. (2021). *Approximate Bayesian Computation and Statistical Applications to Anonymized Data: an Information Theoretic Perspective*. PhD thesis.
12. Viscardi, C., Boreale, M. & Corradi, F. (2020). Improving ABC via Large Deviations Theory. *Book of Short Papers SIS 2020*, 673-678.
13. Viscardi, C., Boreale, M. & Corradi, F. (2021). Weighted Approximate Bayesian Computation via Sanov's Theorem . *Computational Statistics*, accepted for publication.

# A propensity score approach for treatment evaluation based on Bayesian Networks

## *Valutazione di un trattamento con un approccio sul propensity score basato su reti Bayesiane*

Federica Cugnata, Paola M.V. Rancoita, Pier Luigi Conti, Alberto Briganti, Clelia Di Serio, Fulvia Mecatti and Paola Vicard

**Abstract** In observational studies evaluating the treatment effect on a given outcome, the treated and untreated subjects may be highly unbalanced in their observed covariates, and these differences can lead to biased estimates of treatment effects. Propensity score is popular tool to reduce this bias. In this work we propose to estimate the propensity score by using Bayesian Networks as alternative to conventional logistic regression. Based on it, we develop an inferential methodology to evaluate the treatment effect. In simulation study, our proposed approach resulted in the best performance.

**Abstract** *Negli studi osservazionali che valutano l'effetto di un trattamento, i soggetti trattati e non trattati possono presentare caratteristiche molto sbilanciate e queste differenze possono portare a stime distorte degli effetti del trattamento. Il propensity score è uno strumento largamente utilizzato per ridurre questa distorsione. In questo lavoro si propone di stimare il propensity score utilizzando le reti bayesiane come alternativa alla regressione logistica. Sulla base di ciò, proponiamo una metodologia inferenziale per valutare l'effetto del trattamento. Nello studio di simulazione, l'approccio da noi proposto presenta la performance migliore.*

---

Federica Cugnata, Paola M.V. Rancoita and Clelia Di Serio  
University Centre for Statistics in the Biomedical Sciences (CUSBS), Vita-Salute San Raffaele,  
e-mail: cugnata.federica@univr.it, rancoita.paolamaria@univr.it and diserio.clelia@univr.it

Pier Luigi Conti  
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, e-mail: pier-  
luigi.conti@uniroma1.it

Alberto Briganti  
Department of Oncology/Unit of Urology, Urological Research Institute, IRCCS Ospedale San  
Raffaele, Milan, e-mail: briganti.alberto@hsr.it

Fulvia Mecatti  
Dipartimento di Sociologia e Ricerca Sociale, Università di Milano-Bicocca, e-mail: ful-  
via.mecatti@unimib.it

Paola Vicard  
Dipartimento di Economia, Università Roma Tre, e-mail: paola.vicard@uniroma3.it

**Key words:** Potential outcomes, propensity score, covariate balance, observational study, ATE estimation

## 1 Introduction

In medical research there is a growing interest in evaluating the treatment effect on a given outcome using data from observational studies that are more easily available. However, the characteristics of treated subjects may differ from those of untreated subjects, for example, due to bias in treatment decision. In this setting, methods based on propensity score are widely adopted to adjust for differences in the distributions of the characteristics of the subjects between the treated and the untreated groups. Usually, the propensity score is estimated through logistic regression and then employed in different types of approaches. In this work, we propose to estimate the propensity score by using Bayesian Networks and, based on it, we develop a methodology to evaluate the treatment effect. Two alternative estimators of the treatment effect are here considered and compared in a simulation study.

## 2 Method

Consider a random sample of  $n$  independent subjects. Let  $T$  represent the treatment assignment taking value 1 and 0, if a subject receive or does not receive the treatment, respectively. Furthermore, let  $Y_{(1)}$  and  $Y_{(0)}$  denote the potential outcomes of a subject in the presence or absence of the treatment, respectively. The observed outcome is:  $Y = Y_{(1)}I_{(T=1)} + Y_{(0)}I_{(T=0)}$ , where  $I_{(\cdot)}$  denotes the indicator function. We define that the treatment has no effect when  $Y_{(0)}$  and  $Y_{(1)}$  have the same probability distribution. As it may occur in observational studies, we assume that the assignment-to-treatment mechanism is not a “purely random” mechanism, which happens in the experimental framework. On the contrary, because of the presence of confounding covariates, there could be considerable differences among subjects receiving different treatments. In particular, we assume here that the assignment-to-treatment mechanism only depends on observed covariates, denoted by  $\mathbf{X} = (X_1 \cdots X_L)$ . The *propensity score* is defined as the probability of receiving treatment conditionally on  $\mathbf{X} = \mathbf{x}$ ,  $p_1(\mathbf{x}) = P(T = 1 | \mathbf{X} = \mathbf{x})$ . In the following,  $p_0(\mathbf{x}) = P(T = 0 | \mathbf{X} = \mathbf{x}) = 1 - p_1(\mathbf{x})$ .

In the present paper, we focus on the case where covariates are discrete, finite r.v.s, and the potential outcomes are dichotomic variables. The assumptions of the analysis are listed below.

H1. *Discreteness*. The potential outcomes  $Y_{(k)}$  are dichotomic r.v.s., taking values 0 and 1 (without loss of generality) and with (marginal) probability  $\theta_k = P(Y_{(k)} = 1)$ ,  $k=0, 1$

H2. *Unconfoundedness*.  $T \perp\!\!\!\perp (Y_{(0)}, Y_{(1)}) | \mathbf{X}$ .

A propensity score approach for treatment evaluation based on Bayesian Networks

H3. *Common support.* There exists a positive real  $\delta$  for which  $\delta \leq p_k(\mathbf{x}) \leq 1 - \delta$  for each  $\mathbf{x}$  and  $k = 0, 1$ .

In the case under examination, the absence of treatment effect is equivalent to say that  $\theta_0 = \theta_1$ . The treatment effect can be also expressed in terms of Average Treatment Effect (ATE), defined in general as  $ATE = E[Y_{(1)}] - E[Y_{(0)}]$ . Since in the present case  $E[Y_{(k)}] = \theta_k$ ,  $k = 0, 1$ , the absence of treatment effect is equivalent to  $ATE = 0$ .

#### *Estimation of propensity scores by Bayesian Network models*

In this work we proposed to estimate propensity scores  $p_1(\mathbf{x})$  by using a Bayesian Network (BN) model for  $(T, \mathbf{X})$  [1] and we will denote with  $\hat{p}_1(x)$  this estimate ( $\hat{p}_0(x) = 1 - \hat{p}_1(x)$ ). A BN is essentially a multivariate statistical model satisfying a set of conditional independence statements contained in a directed acyclic graph (DAG)  $G = (V, D)$ , consisting in a set of *vertices*  $V$ , representing random variables, and a set  $D \subseteq V \times V$  of *directed arcs*  $D$  connecting pairs of nodes. BNs enable an effective representation and computation of a joint probability distribution over a set of random variables. When covariates are discrete, the use of BNs has several positive theoretical features and practical advantages, as well, if compared to more traditional semiparametric estimation of propensity scores based on semiparametric logistic regression, as in [2], [3]. On the theoretical side, BNs allow to consider Maximum Likelihood Estimators (MLEs) of propensity scores  $p_1(\mathbf{x})$ , that possess the “usual” properties of MLEs. On the practical side, BNs allow flexibility in modeling the dependence relationships among covariates, without requiring the identification, for each treatment level, of polynomial terms and interactions terms for the covariates in  $\mathbf{X}$  to be included in the model.

#### *Estimation of potential outcomes probabilities*

We consider two estimators for estimating the marginal probabilities  $\theta_k$  based on: the BN estimate of  $\hat{p}_k(x)$  the Horvitz-Thompson-type estimator and the Hájek-type estimator. The Horvitz-Thompson-type estimator for  $\theta_k$  is defined as:

$$\hat{\theta}_k^{HT} = \frac{1}{n} \sum_{i=1}^n I_{(Y_i=1)} I_{(T_i=k)} \hat{p}_k(\mathbf{x}_i)^{-1}; \quad k = 0, 1. \quad (1)$$

The Hájek-type estimator for  $\theta_k$  considers normalized weights and it is defined as:

$$\hat{\theta}_k^H = \frac{1}{\sum_{i=1}^n I_{(T_i=k)} \hat{p}_k(\mathbf{x}_i)^{-1}} \sum_{i=1}^n I_{(Y_i=1)} I_{(T_i=k)} \hat{p}_k(\mathbf{x}_i)^{-1}; \quad k = 0, 1. \quad (2)$$

Our first result is that both (1) and (2) are consistent estimators of  $\theta_k$ . Formally, under assumptions H1-H3, as  $n \rightarrow \infty$ :  $|\hat{\theta}_k^H - \theta_k| \xrightarrow{P} 0$  and  $|\hat{\theta}_k^{HT} - \theta_k| \xrightarrow{P} 0$  with  $k = 0, 1$ . For the sake of brevity, in this paper we do not report the proof of this result.

### Testing for the presence of treatment effect

The primary goal of the present section is to construct a test for the absence of treatment effect. If we define  $\Delta = \theta_1 - \theta_0$ , testing for the absence of treatment effect reduces to the following hypothesis problem:

$$\begin{cases} H_0 : \Delta = 0 \\ H_1 : \Delta \neq 0 \end{cases} \quad (3)$$

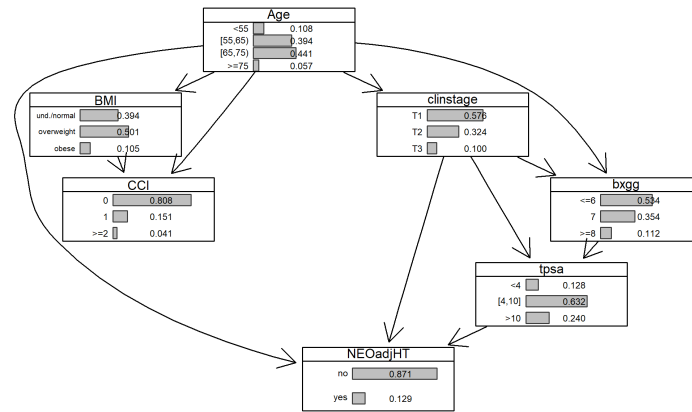
A “natural” test-statistic for the above hypotheses problem is  $D_n = \hat{\theta}_1 - \hat{\theta}_0$ . It can be shown that  $\sqrt{n}(D_n - \Delta)$  is asymptotically normally distributed with variance  $\sigma^2$ , for both types of estimators (proof not reported for the sake of brevity), and in order to estimate the asymptotic variance,  $\sigma^2$ , we developed an approach by exploiting ideas of Hirano et al. (2003) developed in a different context. As a consequence, in order to test for the presence of treatment effect, a simple procedure consists in constructing a confidence interval at level  $1 - \alpha$  for  $\Delta$

$$\left[ D_n - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, D_n + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right], \quad (4)$$

where  $\hat{z}_p$  is the  $(1 - p)$ -quantile of the Standard Normal distribution, and in rejecting  $H_0$  whenever the interval (4) does not contain 0.

## 3 Simulation study

In order to assess the performance of the proposed methods we constructed artificial data inspired by real data of patients with prostate cancer who underwent radical prostatectomy. In particular, we considered a database including 6478 patients of the San Raffaele Hospital (Milan) to evaluate of the effect of neo-adjuvant hormonal therapy on whether lymphadenectomy was performed during the surgery. In addition, the data included the following covariates: age at surgery, BMI, Charlson Comorbidity Index (CCI), Biopsy Gleason score, clinical stage and total PSA. Continuous variables were categorized according to clinical criteria.  $N = 1000$  replications with samples sizes  $n = 500, 1000, 2500, 5000$  have been generated by Monte Carlo simulation. In each replication of the simulation study, the treatment variable and the other covariates were generated from the BN estimated from the real data (Fig. 1). The binary potential outcomes  $Y_{(0)}$  and  $Y_{(1)}$  were then simulated through logistic models under two scenarios. This was done in a realistic way by setting the coefficients in the models (generating the outcome) equal to the observed coefficients estimated on the real data. More formally, they were generated in the following way:  $Y_{(k)} | \mathbf{X}_c \sim Be(P(Y_{(k)} | \mathbf{X}_c))$  for  $k = 0, 1$ , with  $\text{logit}P(Y_{(0)} = 1 | \mathbf{X}_c) = \alpha_0 + \boldsymbol{\beta}^T \mathbf{X}_c$  and  $\text{logit}P(Y_{(1)} = 1 | \mathbf{X}_c) = \alpha_0 + \alpha_1 + \boldsymbol{\beta}^T \mathbf{X}_c$ , where  $\mathbf{X}_c$  denotes the set of covariates considered in the specific scenario. In order to compute the true ATE of the considered scenario, the true  $\theta_k = P(Y_{(k)} = 1)$  were computed by marginalizing  $P(Y_{(k)} = 1 | \mathbf{X}_c)$



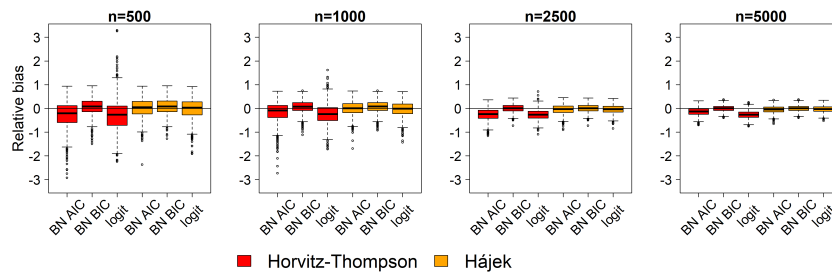
**Fig. 1** Bayesian network obtained with the Tabu search (TABU) greedy search algorithm with the AIC score functions.

over  $\mathbf{X}_c$ . Besides the treatment, we assumed that the potential outcomes  $Y_{(k)}$  were affected, in scenario (i), only by the the total PSA, whereas in scenario (ii) by all considered variables. Fixing the parameters  $\alpha_0$ ,  $\alpha_1$  and  $\beta^T$  based on the real data, in scenario (i)  $ATE = 0.114$  and in the scenario (ii)  $ATE = 0.099$ . The outcome,  $Y$ , was finally generated depending on the simulated value of  $T$  for each subject:  $Y = Y_{(1)}I_{(T=1)} + Y_{(0)}I_{(T=0)}$ .

### Results of simulations

In each simulated dataset, the propensity score was estimated with the following methods: logistic regression (logit), Bayesian Network with the Tabu search greedy search algorithm with the AIC score functions (BN AIC) or with the BIC score functions (BN BIC). The estimated propensity score was then used to obtain the Hájek-type and the Horvitz-Thompson-type estimator for  $\theta_k$ , estimation of ATE and the related confidence interval. Considering the Horvitz-Thompson-type estimator the bias of ATE was generally closer to zero for BN BIC than BN AIC and logit, which instead led to an underestimation of the ATE. Considering Hájek-type estimator, all techniques seem always to perform rather well. Increasing  $n$ , the variability of bias became lower but the performance of the methods maintained the same pattern. As example, in Fig. 2, the box-plots of the relative bias of the estimated ATE are reported for scenario (ii). Table 1 shows the proportion of times the true ATE falls within the estimated 95% confidence interval and the proportion of times the 0 does not fall within a 95% the estimated confidence interval for scenario (ii), by varying the sample size  $n$ . For BN BIC the proportion of times the true ATE falls within the 95% confidence interval is greater than for other approaches and it is close to the nominal level 95%. BN BIC also shows a higher proportion of rejection of the null hypothesis than BN AIC and logit.





**Fig. 2** Relative Bias of the estimated ATE for the scenario (ii).

**Table 1** Simulation results for scenario (ii).

n	PS method	Horvitz-Thompson		Hájek	
		prop. 95%CI includes true ATE	prop. 95%CI does not include 0	prop. 95%CI includes true ATE	prop. 95%CI does not include 0
500	BN AIC	0.781	0.594	0.919	0.782
500	BN BIC	0.940	0.866	0.948	0.873
500	logit	0.667	0.554	0.906	0.762
1000	BN AIC	0.775	0.804	0.902	0.920
1000	BN BIC	0.921	0.962	0.923	0.962
1000	logit	0.718	0.726	0.900	0.887
2500	BN AIC	0.659	0.928	0.920	0.997
2500	BN BIC	0.962	0.999	0.964	0.999
2500	logit	0.648	0.956	0.948	0.996
5000	BN AIC	0.762	1.000	0.885	1.000
5000	BN BIC	0.935	1.000	0.934	1.000
5000	logit	0.426	0.997	0.918	1.000

In this work we implemented a simulation study inspired by a medical case study, in order to provide more general results, we will implement a second simulation study using more general data generating mechanisms allowing to evaluate the methodology for different values of ATE (e.g., zero or bigger values than used currently).

## References

- [1] R G. Cowell, A P. Dawid, S L. Lauritzen, and D J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer Verlag, New York, 1999.
- [2] K. Hirano, G W. Imbens, and G. Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71:1161–1189, 2003.
- [3] D F. McCaffrey, B A. Griffin, D. Almirall, M E. Slaughter, R. Ramchand, and L F. Burgette. A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. *Statistics in Medicine*, 32:3388–3414, 2013.

## 4.35 Statistical modelling for the analysis of contemporary societies

# **Social Network Analysis to analyse the relationship between ‘victim-author’ and ‘motivation’ of violence against women in Italy**

## ***L’analisi delle reti sociali per analizzare la relazione ‘vittima-autore’ e ‘movente’ della violenza contro le donne in Italia***

Alessia Forciniti

### **Abstract**

The paper aims to analyse the phenomenon of Violence against women in the Italian context during 2020. It proposes to study the relationship between ‘victim-author’ and ‘motivation’ in femicides committed in domestic environment. By means of the properties of the Social Network Analysis on bimodal data, the study detected main actors and motivations that generated the homicides with female victims. At the same time, the structural relationships allowed to investigate the existence of motivations that better characterized the action of the various actors. The bipartite graph visualization and centrality scores calculated have demonstrated the effectiveness of the methodology for the pursued objectives.

### **Abstract**

*Il lavoro mira ad analizzare il fenomeno della Violenza contro le donne nel contesto italiano durante il 2020. Si propone di studiare la relazione tra ‘vittima-autore’ e ‘movente’ nei femminicidi commessi in ambito domestico. Attraverso le proprietà della Social Network Analysis su dati bimodali, lo studio ha rilevato i principali attori e moventi che hanno generato gli omicidi con vittime femminili. Al contempo, le relazioni strutturali hanno permesso di indagare l’esistenza di moventi che hanno maggiormente caratterizzato l’azione dei vari attori. La visualizzazione del grafo bipartito e i punteggi di centralità calcolati hanno dimostrato l’efficacia della metodologia per gli obiettivi perseguiti.*

---

<sup>1</sup>

Alessia Forciniti, University of Naples Federico II; [alessia.forciniti@unina.it](mailto:alessia.forciniti@unina.it)

**Key words:** Social Network Analysis, Two-mode network, Bipartite graph, Violence against women

## 1 Introduction

The Council of Europe, Treaty N.210 (2011) defines Violence against women (VAW) as *'any act of gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life'*. The phenomenon represents a widespread humanitarian emergency that occurs at alarming rates around the world, and it is one of most discussed topics during the last years. According to data processed by the Eures Research Institute, thanks to the collaboration of the Ministry of the Interior, criminal analysis service, in Italy, in the last year there were 91 femicide cases, of which 81 attributable to the family environment. The year 2020 was profoundly marked by the world health emergency caused by Covid-19 which indissolubly conditioned every aspect of civil life in sociability, especially the women life through the contraction of social contacts due to the restrictive measures. As a matter of fact, with respect to 2019, the official statistics (Eures Ricerche Economiche e Sociali, 2020) show a decrease in homicides with female victims (from 99 to 91, equal to - 8.1%) for occasional crimes, while more contained is the reduction in family ones (from 85 to 81, - 4.7%).

This work focuses on analysing violence against the women in the domestic context, by investigating the main actors and motivations that generated the family femicides in Italy during 2020. In particular, in order to study the existence of motivations that better characterize the homicide events of various actors, the paper aims to detect structural relations among victim-author and motivations by using the properties of the Social Network Analysis.

In the following, section 2 introduces the data and statistical approach used to reach the objective; section 3 presents main findings of the analysis and finally, section 4 indicates the conclusion and future work.

## 2 Data and Statistical approach

Data are taken from the Report on Femicide in Italy 2020, institutional investigation realized by Eures Ricerche Economiche e Sociali. It is seventh report concerning the intentional homicide in Italy that explains features and trends about the murders involving female victims. Eures data permitted to select the predominant femicide motivations in the first 10 months of 2020 in the family environment by taking into account, at the same time, the actors of each homicide event and therefore, the

Social Network Analysis to analyse the relationship between 'victim-author' and 'motivation' of violence against women in Italy

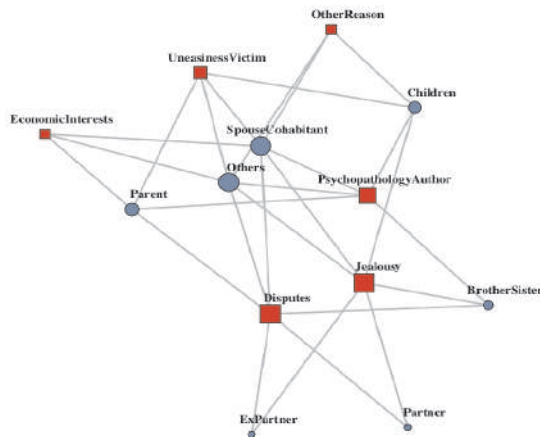
relationship between victim and author. The information was encoded by building the 'victim-author' by 'motivation' matrix, where two-mode data are represented by means of the affiliation matrix (**A**) that indicates the binary relationship between each *i*-th actor and each *j*-th femicide event motivation. The generic element  $a_{ij}$  is equal to 1 if the actor is affiliated with the homicide event motive, 0 otherwise. The linkage among actors shows their joint association to homicide events motives in the domestic context. The matrix consists in 7 rows of actors: *brother-sister, children, ex-partner, others, parent, partner, spouse-cohabitant*; and 6 columns of motivations: *disputes, economic interest, jealousy, other reason, psychopathology of the author, uneasiness of the victim*. To investigate the relationships among these actors and motivations connected to femicide behaviour in the family, the statistical approach used is the Social Network Analysis (SNA; Wasserman & Faust, 1994). The methodology considers two fundamental structural elements: the nodes that represent the actors and motivations and their ties or relations. SNA is frequently used in social sciences to understand the interactions of the individuals in a specific social context and quantifies the affiliations within a network. The table **A**, recording the relational information between the actors and homicide events motives is represented as mathematical graph (Harary, 1969) in which the nodes correspond to two entities: actors and motivations, and the lines or edges correspond to ties of affiliation among the entities. Formally, the two-mode network is represented by means of a bipartite graph  $G_B(V_1, V_2, E)$ , where  $V_1$  and  $V_2$  are the nodes corresponding to two classes of entities and  $E$  is the affiliation that draws the elements of  $V_1$  to  $V_2$  (Borgatti, 2009). One of key issue in network analysis is to identify which nodes are more 'central' than others (Freeman, 1978) so-called 'focal points' of the whole network, which can explain the interactions among actors and motivations. In order to determinate the structural features and the node's position in the network, the most frequent measures of centrality for SNA (Faust, 1997) were used: degree, betweenness, closeness and eigenvector. Degree centrality identifies the importance of one or more nodes based on the number of ties that a node has with other nodes. The higher the number of ties, the greater the level of centrality of a particular actor or motivation in the network. For this case of analysis, actor degree centrality corresponds to the number of homicide events characterized by a given actor, and motivation degree centrality is the number of actors acting based on that motivation. Betweenness centrality refers to the number of times a node lies on the shortest path between other nodes. This measure shows which nodes are 'intermediary' among nodes in a network. In this case, betweenness determines of paths from actors to actors, motivations to motivations, actors to motivations and vice versa. Closeness centrality measures the importance of a node based on how close a node is to all the other nodes. Eigenvector centrality represents the extension of degree centrality and measures the quality of nodes connected. It is the centrality of a node as proportional to its neighbour's importance. This means the actor centrality is proportional to the sum of centralities of the motivations of homicide events; analogously the centrality of homicide motivations is proportional to the sum of centralities of the homicide events motivations which actors take part in.

### 3 Main findings of the analysis

Bimodal analysis on data 'victim-author' by 'motivation' returns an affiliation network formed by 13 nodes, actors and motivations, and 54 ties among them.

With regard to descriptive statistics about the network cohesion, the density - as proportion of pairs of nodes that have ties - denotes 64% of all possible ties. Therefore, density equal to 0.643 shows a satisfying cohesion degree of network. The standard deviation of 0.485 confirms high variability within ties. To measure the cohesion the distance concept is also examined, such as the average number of edges between any two nodes in the network. The mean distance is equal to 1.846, and its low value depends on to be inversely proportional to density.

The visual representation for affiliation network by means of bipartite graph (Figure 1) presents the actors as blue circles and motivations as red squares, while the ties as conjunction lines among nodes. It shows how some actors shared common motivations in family femicide and others connected only to few specific motivations. It follows that there is a set of actors that assassinated for a set of motivations exclusive to them and others that behaved for a set of motivations inclusive to almost all actors.



**Figure 1:** Bipartite graph for affiliation network 'victim-author' by 'motivation' with degree centrality

At the top left, it is possible to identify *economic interests* for homicide events committed only by *parent*, *spouse-cohabitant* and *others*. The *uneasiness of victim* and *psychopathology of the author* are instead attributable to the aforementioned actors, however by also introducing the figure of the *children*. Similarly, *spouse-cohabitant*, *others* and *children* are connected to *other reason* as homicide events, except of the *parent*. On the lower right, the visualization makes clear the presence of *partner* and *ex-partner*, both moved only by *disputes* and *jealousy*. These two motivations allow these two actors to be part of the network and therefore, to connect the upper and lower parts of the graph. *Brother-sister* is connected to

Social Network Analysis to analyse the relationship between ‘victim-author’ and ‘motivation’ of violence against women in Italy

*disputes* and *jealousy* on the lower of graph, and to *psychopathology of the author* at the top. In addition to descriptive measures of global network, one of most relevant statistics for empirical study is the centrality of the nodes, both actors and motivations. The centrality based on degree as total number of ties - including indegree and outdegree - presents highest normalized values (0.50) for *spouse-cohabitant*, *others*, *jealousy* and *disputes* that are predominant actors and motives of network (refers to *Figure 1* and *Table 1*). The lowest degree was recorded by *partner* (non-cohabitant, 0.16) and *ex-partner* (0.16) and this means they are the actors less involved in the femicide events for the analysed motives. Betweenness indicates that 5% of ties depends on presence of intermediaries’ nodes. The highest normalized values were recorded by *disputes* (0.19) and *jealousy* (0.19) which main motivations of connecting among the examined actors. *Partner* and *ex-partner* have not instead any intermediary function in the network and therefore, they are peripheral. According to centrality based on closeness, *spouse-cohabitant* (0.67) and *others* (0.67) are the most central nodes: they are the closest to other nodes. The centrality measured by the eigenvector confirms the importance of nodes *spouse-cohabitant* and *others* (same highest values equal to 0.99) followed by *disputes* and *jealousy* (both 0.89).

**Table 1:** Normalized centrality scores for affiliation network ‘victim-author’ by ‘motivation’

<i>Node</i>	<i>N. Ties</i>	<i>Normalized degree</i>	<i>Normalized betweenness</i>	<i>Normalized closeness</i>	<i>Normalized eigenvector</i>
SpouseCohabitant	6	0.50	0.16	0.67	0.99
Jealousy	6	0.50	0.19	0.63	0.89
Disputes	6	0.50	0.19	0.63	0.89
Others	6	0.50	0.16	0.67	0.99
PsychopathologyAuthor	5	0.42	0.08	0.57	0.86
Children	4	0.33	0.05	0.55	0.67
Parent	4	0.17	0.05	0.55	0.67
UneasinessVictim	4	0.32	0.04	0.52	0.74
BrotherSister	3	0.25	0.02	0.50	0.58
EconomicInterests	3	0.25	0.01	0.48	0.59
OtherReason	3	0.25	0.01	0.48	0.59
Partner	2	0.16	0.00	0.46	0.39
ExPartner	2	0.16	0.00	0.46	0.39

## 4 Conclusion and future work

The bimodal analysis by means of bipartite graph and the centrality scores calculated allowed global understanding of the whole dataset of actors and motivations studied, by demonstrating the effectiveness of SNA to analyse the relationships among family members and motivations in domestic femicide events.

Indeed, the methodological approach was able to answer to research objectives introduced in the first section of the paper: to individuate main actors and

motivations that generated the homicides with female victims in Italy during 2020 and, to study the motivations that better characterized the homicide events of the various actors. According to results the predominant nodes of the whole affiliation network are *spouse-cohabitant*, *others* relatives, *jealousy* and *disputes*. The lowest predominance was recorded instead by the intimate *partner* without cohabitation or *ex-partner*. This means that during 2020, in Italy, domestic femicides were committed mainly by the *spouses* (husband or wife) or *cohabitant partner* for passionate or conflictual motivations. Probably, these results depend on the continuative and forced cohabitation due to the restrictive measures anti-Covid 19 that exacerbated the family conflicts. The femicide relationship between *parents* and *children* was characterized mainly by mental disorders of the author or the victim; while that between *brothers* and *sisters* by *disputes* and *jealousy*.

Some results of analysis can find the confirm for example in the SNA application proposed by Leone *et al.* (2019) aimed at detecting diagnostic patterns associated to health consequences recorded by the Emergency Department, where it emerges the psychoses as common effect of abuse by intimate partner; thus, the psychological uneasiness of the victim as motive of femicide associated to cohabitant may be configured. By comparing distributions of femicides based on motivations between 2020 and the previous years (Eures, 2006; 2020), interesting considerations are allowed. The spouse-cohabitant actor remained one of main ones, but in the past the motivation was mainly disagreement, while in the years the passionate reason is emerging. The relationships between parents and children, brothers and sisters also evolved; currently they are moved by dispute and jealousy while in the past were characterized mainly by psychopathologies or economic interests.

The future directions of work refer to extend the analysis to a unimodal approach (Borgatti, 2009) for analysing actors and motivations separately. In addition, the interest is to replicate the analyses on previous years datasets for comparing the results and to understand changes in the context of domestic femicide.

## References

1. Borgatti, S.P.: Two-Mode Concepts in Social Network Analysis. In: Meyers R. (eds.) Encyclopedia of Complexity and Systems Science. Springer, New York, NY (2009) doi: 10.1007/978-0-387-30440-3\_491
2. Council of Europe, Treaty N.210: Convention on preventing and combating violence against women and domestic violence (2011) Available via <https://rm.coe.int/168008482e>
3. Eures Ricerche Economiche e Sociali: L'omicidio volontario in Italia. Rapporto EURES-ANSA 2006 (2006)
4. Eures Ricerche Economiche e Sociali: Settimo rapporto sul femminicidio in Italia. Caratteristiche e tendenze del 2020 (2020)
5. Faust, K.: Centrality in affiliation networks. *Social Networks* 19, 157--191 (1997)
6. Freeman, L.C.: Centrality in social networks: Conceptual clarification. *Social Networks* 1, 215--239 (1978)
7. Harary, F.: *Graph Theory*. Reading, MA. Addison-Wesley (1969)
8. Leone, M., Lapucci, E., *et al.*: Social network analysis to characterize women victims of violence. *BMC Public Health* (2019)
9. Wasserman, S., Faust, K.: *Social Network Analysis*. Cambridge University Press (1994)



# Satisfaction and sustainability propensity among elderly bike-sharing users

## *Soddisfazione e propensione alla sostenibilità tra gli utenti anziani dei sistemi di bike-sharing*

Paolo Maranzano, Roberto Ascari, Paola Maddalena Chiodini, and Giancarlo Manzi

**Abstract** The eleventh United Nations' sustainable development goal (SDG) is about sustainable cities and communities, with a particular focus on increasing the use of bikes and public transportation. Bike-sharing systems might help to improve the awareness of the importance of sustainability in urban areas. A major challenge for the success of the SDGs in general, and the spread of sustainable cities in particular, is to involve all segments of the population, including the age group of the elderly (65 years and more). In this context, bike-sharing systems play a fundamental role, as they can concur to improve the awareness of the importance of sustainability in the population. Using data from a satisfaction survey conducted among the bike-sharing system "BikeMi" in Milan, Italy, we detect the best determinants of satisfaction with the service among the aged population, highlighting its relationship with sustainability.

**Abstract** *L'undicesimo obiettivo di sviluppo sostenibile (SDG) delle Nazioni Unite riguarda città e comunità sostenibili, con un'attenzione particolare all'aumento dell'uso delle biciclette e dei trasporti pubblici. I sistemi di bike-sharing possono aiutare a migliorare la consapevolezza dell'importanza della sostenibilità nelle aree urbane. Una delle principali sfide per il successo degli SDG in generale, e della diffusione delle città sostenibili in particolare, è coinvolgere tutti i segmenti della*

---

Paolo Maranzano

Department of Management, Information and Production Engineering, Università degli Studi di Bergamo, Bergamo, Italy. e-mail: paolo.maranzano@unibg.it

Roberto Ascari

Department of Economics, Management and Statistics, Università degli Studi di Milano-Bicocca, Milan, Italy. e-mail: roberto.ascari@unimib.it

Paola Maddalena Chiodini

Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Milan, Italy. e-mail: paola.chiodini@unimib.it

Giancarlo Manzi

Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Milan, Italy. e-mail: giancarlo.manzi@unimi.it

*popolazione, comprese la fascia d'età dei più anziani ( $\geq 65$  anni). In questo contesto i sistemi di bike-sharing giocano un ruolo fondamentale, in quanto possono concorrere a migliorare la consapevolezza dell'importanza della sostenibilità nella popolazione. Utilizzando i dati di un'indagine di soddisfazione condotta presso il sistema di bike-sharing "BikeMi" di Milano, rileviamo le migliori determinanti di soddisfazione per il servizio tra gli anziani, evidenziando il rapporto con la sostenibilità.*

**Key words:** Sustainability, Bike-sharing, Smart cities, Elderly

## 1 Introduction

One of the most important dimensions for the development of the so-called *smart cities* is the search for alternative urban mobility, needing new ideas and more innovation [2]. This is receiving increasing attention also because mobility is considered one of the most important areas for which a radical change with respect to the past should be introduced for the implementation of sustainable cities and communities, according to the United Nations' sustainable developing goal. In the smart city context, if mobility must be sustainable (i.e., if it contributes to economic growth and helps to improve the quality of life), bike-sharing should be implemented as a possible bridge between public well-being and private economic development. Recent developments in urban planning management have led bike-sharing systems (BSSs) to be a viable complement to traditional public transportation systems. However, there are some important quandaries in organising a BSS that is successful for the citizens. For this reason, continuous monitoring of the satisfaction for the service among users should be implemented to promptly detect possible issues and to quickly deal with them. A traditional BSS (having docking stations as terminal nodes used to lock and unlock bikes) is mainly used by workers who want to travel the last mile of their journey to work or their return home after work as fast as possible [4]. Less is known about the use of BSS by those who do not use it for commuting to workplaces, such as retirees. One would expect different behaviour in using the service and also a different level of satisfaction, dictated by factors other than those typical of the workers. Older people also have a different attitude to new technologies than youngsters. The 'free-floating' bike-sharing systems are now all organised with the most modern technologies (apps for the route, apps to find out stations with free docks, etc.). This can discourage its use by the elderly and can lead to greater dissatisfaction. Traditional BSSs have an average level of technology that could instead incentivise their use by the elderly. In addition to not using BSSs for commuting to workplaces, in the BSSs literature (see [1] for an extensive literature review on BSSs) elderly are considered as a demographic category having more problems with respect to average population in the cycling exercise. This means that their main reason for using BSSs could be different from cycling for health purposes [5]. There should be a stronger motivation for them to bike, which

could be linked, in our opinion, to being well aware of the necessity of more sustainability in the transport sector. There should be a self-selection process among elderly using BSSs which comes from being more *green-driven* than the average population. All this motivates our paper: we want to study satisfaction among the elderly with a traditional BSS in the context of a smart city reality, in an increasingly technological environment, and see if they are more satisfied and more prone to sustainability than other people [3]. For this task we analyse a subset of data coming from a 2016 survey questionnaire delivered to the subscribers of the ‘BikeMi’ BSS in Milan, Italy, with the use of regression analysis to explore major determinants of satisfaction ranks obtained through POSet-based analysis (see [3] for details).

## 2 POSets

A Partially Ordered Set (POSet)  $P = (X, \leq_P)$  is defined as a set  $X$  endowed with a partial order relationship  $\leq_P$  such that its elements satisfy axioms of reflexivity, antisymmetry, and transitivity. [3]

Starting from a set of  $K$  ordinal features, each of them taking  $h_k$  ordered scores, their product order is the POSet generated by all the possible score configurations of the  $K$  attributes. Thus, a product order POSet contains an overall number of  $H = \prod_{k=1}^K h_k$  configurations. Although the product order is composed by ordinal features, some pairs of its elements (*profile*) can be comparable (i.e., orderable) and others not. Considering all the profiles, it follows that there exist more than one ranking. The average rank of a generic profile  $x_i$  is defined as its average position in the rankings with respect to all the other elements.

## 3 Satisfaction items and predictors considered

We computed the POSet-based satisfaction index as in [3], considering five satisfaction items defined on a 1-5 Likert scale (from 1 = low satisfaction to 5 = very high satisfaction). The items were the following:

- Satisfaction with customer service;
- Satisfaction with the distribution of docking stations across the city;
- Satisfaction with cost (subscription, fares for overuse, etc.);
- Satisfaction with bike maintenance;
- Satisfaction with bike comfort (weight, manageability, etc.).

Due to the definition of ranking, lower values of rank are associated with higher user satisfaction. In the regression analysis, we considered five predictors related to the “green propensity” (i.e., the propensity of being more aware of sustainability/ecological issues). They were the following:

- Declared averaged distance covered when using the service (*Distance* - four classes: ' $\leq 1$  km' (reference class), '1-2 km', and '2-4 km', and ' $\geq 4$  km');
- Frequency of use (*Frequency* - five classes: 'occasional', 'weekend', 'working days', 'every day' (reference class), and 'other');
- Reasons for being a subscriber (*Reason* - three classes: 'Health-care', 'Sustainability' (reference class), and 'Other');
- Use of multiple transportation means together with the service (*Sharing* - three classes: 'Only individual', 'Only shared' (reference class), and 'Mixed');
- Considering the bike as a proper alternative to car (*Alternative* - Likert scale from 1='Not very much an alternative to car', to 5='Very much an alternative to car').

## 4 Results

The sub-sample of the elderly (defined as those subjects declaring an age greater or equal to 65 years) was composed of 322 subjects (almost 5% of the total respondents), the 82.9% of them being men. The mean and median age was 69.86 and 69 years, respectively, whereas education level was high, as 65.7% of the respondents had at least a bachelor's degree. The 44.4% of the respondents were still a worker and the 77.3% of them were married.

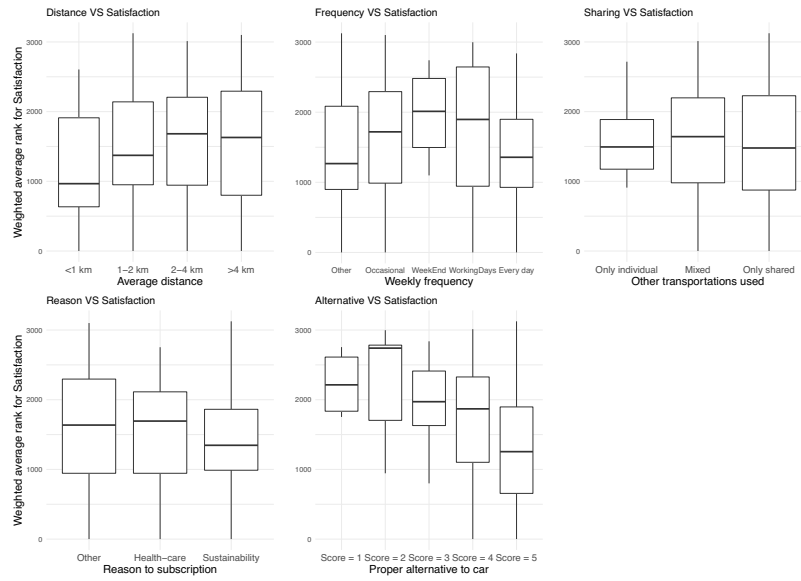
Boxplots of satisfaction rank by predictors' levels (see Figure 1) highlight some patterns. For example, as average distance increases, satisfaction decreases (satisfaction is expressed in ranks). The more the user evaluates bike-sharing as a valid alternative to private means (increasing *Alternative* score), the more the rank decreases, thus increasing satisfaction. *Reason* and *Sharing* do not seem to show patterns with satisfaction. Daily use (*Frequency* = 'Every day') is associated with a lower rank than the other identified users (other than 'Other') and therefore greater satisfaction.

### 4.1 Regression analysis

In a regression framework where the response variable  $Y$  is an average rank (i.e., it is not an integer), the interpretation of the coefficients can be misleading. Thus, we prefer to log-transform the response, obtaining a log-linear model where the coefficients are the percentage of variation on  $Y$  due to an unit-increase of the covariate.

The regression analysis (Table 1) confirms what already observed with the boxplots. Indeed, every regression coefficient concerning *Distance* and *Alternative* is significant, while those related to *Sharing* and *Reason* are generally not significant (the only exception being *Reason* = 'Other'). The patterns identified in the boxplots are reflected in the regression coefficients which are decreasing in the case of *Distance* and *Alternative* predictors.

## Satisfaction and sustainability propensity among elderly bike-sharing users



**Fig. 1** Boxplots of predictors for satisfaction

**Table 1** Regression analysis on satisfaction among over 65-year old people in the 2016 BikeMi Survey

	<i>log(Satisfaction rank)</i>		
	Estimate	Std. Error	Sign.
Constant	5.472	0.373	***
Distance (1-2 km)	0.738	0.296	**
Distance (2-4 km)	0.792	0.286	***
Distance ( $\geq$ 4 km)	0.668	0.365	*
Frequency (Other)	-0.069	0.232	
Frequency (Occasional)	0.334	0.216	
Frequency (Week-end)	0.148	0.673	
Frequency (Working days)	-0.304	0.411	
Reason (Other)	0.424	0.198	**
Reason (Health-care)	0.209	0.263	
Sharing (Only individual)	0.340	0.314	
Sharing (Mixed)	0.089	0.146	
Alternative (Score = 1)	1.188	0.644	*
Alternative (Score = 2)	1.073	0.451	**
Alternative (Score = 3)	0.873	0.273	***
Alternative (Score = 4)	0.553	0.157	***
$R^2$	0.125		
F Statistic (df = 15; 306)	2.913***		
Note:	* $p < 0.1$ ; ** $p < 0.05$ ; *** $p < 0.01$		

## 5 Discussion

The significance of the regression coefficients of Alternative and the non-significance of those of Reason (being ‘Sustainability’ the reference class) could

suggest that the elderly look at this service as something to be considered as a practical surrogate of the car in cities only, not for a general mean to obtain a more sustainable city. This could be due to the *Alternative* question being more “practical”, in the sense that it asks something about the everyday life, while the second is very general and gives little space to a dry answer. The elderly considered in this surveys are still very active (indeed, 44.4% of them is still working), but they are “young” and physically healthy users having no longer urgent needs related to quick and agile travel for work/study or family reasons (e.g., reaching children at school). However, they may still be partially interested in a health aspect (i.e., movement is good) and not yet looking for a comfortable means of transport (private car) as they still have a good dose of energy. They are aware of the problem of having a more sustainable world but are focused on everyday problems (e.g., the comfort of the bike, its weight, etc.).

Future works will be devoted to a more comprehensive analysis with more surveys to be analyzed and some aspects to be detected with a text mining analysis on the open question present in the questionnaire.

## References

- [1] Fishman, E., Washington, S., Haworth, N. (2013). Bike Share: A Synthesis of the Literature. *Transport Reviews*, 33(2): 148-165
- [2] Giffinger, R. (2021). Smart City: The Importance of Innovation and Planning. *Proceedings of the International Conference on Smart Cities and Green ICT Systems - SMARTGREENS 2019*, Springer.
- [3] Maranzano, P., Ascari, R., Chiodini, P.M., Manzi, G. (2020). Analysis of Sustainability Propensity of Bike-Sharing Customers Using Partially Ordered Sets Methodology. *Social Indicator Research*, <https://doi.org/10.1007/s11205-020-02333-8>.
- [4] Qiu, L.Y., He, L.Y. (2018). Bike Sharing and the Economy, the Environment, and Health-Related Externalities. *Sustainability*, 10(4): 1145-1155.
- [5] Woodcock, J., Tainio, M., Cheshire, J., O’Brien, O., Goodman, A. (2014). Health effects of the London bicycle sharing system: health impact modelling study. *British Medical Journal*, 348: 1-14.

# **Media and Investors' Attention. Estimating analysts' ratings and sentiment of a financial column to predict abnormal returns**

*L'attenzione dei media e degli investitori. Stima del voto degli analisti e del sentiment di una rubrica finanziaria per predire i ritorni anomali*

Riccardo Ferretti and Andrea Sciandra

**Abstract** In line with the attention grabbing theory, the publication of articles dealing with the profile of single listed companies along with financial analysts' recommendations is followed by significant increases when the recommendation is positive. In this paper, we tried to understand what happens when analysts' recommendations are missing. We estimated analysts' recommendations with a classification based on the terms in the articles and through a sentiment analysis of the same texts. Results showed that investors transform the articles' content into implicit recommendations that guide their buying decisions when the sentiment is highly positive.

**Abstract** *In linea con la teoria dell'attention grabbing, la pubblicazione di articoli che trattano il profilo di singole società quotate e le raccomandazioni degli analisti finanziari è seguita da incrementi significativi quando la raccomandazione è positiva. In questo paper abbiamo cercato di capire cosa succede in assenza delle raccomandazioni degli analisti, che abbiamo cercato di stimare sulla base dei termini presenti negli articoli o di approssimare tramite sentiment analysis degli articoli stessi. I risultati indicano che gli investitori trasformano il contenuto degli articoli in raccomandazioni implicite che orientano le loro decisioni di acquisto se il sentiment risulta molto positivo.*

**Key words:** text mining, sentiment, classification, event study, abnormal returns

---

<sup>1</sup> Riccardo Ferretti, Department of Communication and Economics, University of Modena and Reggio Emilia; email: [riccardo.ferretti@unimore.it](mailto:riccardo.ferretti@unimore.it)

Andrea Sciandra, Department of Communication and Economics, University of Modena and Reggio Emilia; email: [andrea.sciandra@unimore.it](mailto:andrea.sciandra@unimore.it)

## 1 Introduction

This paper deals with market reaction to second-hand news published on the Sunday editions of an Italian financial newspaper. Stale information published in print media can lead retail investors to buy stocks that grab their attention (Barber and Odean, 2008) such that past analysts' recommendations induce abnormal movements in stock prices and returns (Cervellati et al., 2014). Previous research (Ferretti and Sciandra, 2020), analyzed the column 'Letter to investor', which appeared in the top Italian financial newspaper (*Il Sole 24 Ore*) from 2005 to 2010. From 2011 until now, the column has changed the author of these articles, keeping practically unchanged contents, structure, and layout. The only exception is that this new version no longer publishes the analysts' rating (positive or not). Specifically, Cervellati et al. (2014) and Ferretti and Sciandra (2020) showed that the publication of articles concerning the profile of single listed companies and the recommendations of financial analysts are followed by an asymmetric reaction of stock prices. In particular, they find a statistically significant increase when the recommendation is positive (overweight or buy) and a substantial stationarity when the recommendation is not positive (hold or underweight or sell). This is precisely what is suggested by Barber and Odean's theory of attention grabbing (2008). The Attention-Grabbing Hypothesis (AGH) assumes that naïve investors' behavior affects the market. AGH predicts positive and significant abnormal returns for positively recommended stocks and no reaction for negative ratings. In other words, the market reaction is motivated by an attention-grabbing phenomenon, because only the publication of positive recommendations induces a significant (positive) price movement. Attention grabbing only addresses buying decisions of investors and only when their buying intention is supported by a positive recommendation. So, the research question that arises is what would be the investors' behavior when an explicit recommendation is missing. We identified three possible scenarios: (i) the attention grabbing resulting from the publication of the article affects buying intentions indistinctly, i.e., all articles are perceived as positive recommendations; (ii) the absence of explicit buying or selling indications causes attention to remain in a latent state (no action); (iii) investors transform the content of columns into implicit recommendations which, if positive, guide their buying decisions.

Our analysis strategy to identify the most likely scenario involves: i) evaluating the existence of the attention grabbing mechanism by means of an event study on the whole new sample of observations (2011-2020); ii) reproducing the analysts' ratings in the sample 2005-2010 by means of a statistical model based only on the words appearing in the articles, then estimating the analysts' rating in the new sample; iii) evaluating the existence of significant differences in abnormal returns - on the day after news publication (day 0) - according to the estimated analysts' rating. If no significant differences exist, we plan to assess whether the presence of implicit recommendations in the text of the articles, approximated by a tone (sentiment) analysis, can affect buying decisions.



## 2 Data collection and methods

We collected all the 'Letter to investor' columns published from January 2005 to December 2020 that were devoted to domestic companies listed on the Italian Stock Exchange. The final dataset consists of 870 records. For the first part of the sample (2005-2010 – panel A: 379 records<sup>1</sup>) analysts' ratings were available together with the columns' texts, while from 2011 (2011-2020 – panel B: 491 cases) only columns' texts were available.

As first step to answer our research question we conducted a standard event study (Brown and Warner, 1985) on returns in the Panel B. The event day ( $t=0$ ) is the Monday following each column. The event window went from day -3 to day +3. When an explicit analysts' recommendation is missing, if we found significant abnormal returns on the event day, this could suggest that the publication of the column affected buying intentions indistinctly (scenario (i)), otherwise the absence of explicit buying advice caused no action (scenario (ii)).

To predict the analysts' recommendations in the panel B, we first estimated the analysts' ratings ('buy' or 'don't buy') in Panel A exploiting the actual variable (analysts' rating - binary) and the columns' texts through some text mining and data mining techniques. Next, we classified the columns of Panel B with the best model found in Panel A in terms of accuracy. Based on this new classification on Panel B, we conducted another event study by separating estimated positive recommended columns and non-positive recommended columns. To confirm the event study results, we would also run a regression model, including potential confounding variables (see Table 4) along with the estimated rating. If we did not find significant differences in abnormal returns on the event day, then again, the attention grabbing would affect buying intentions indistinctly. Otherwise, we speculated about investors transforming the columns' content into implicit recommendations which can lead their buying decisions. In this case we planned to perform a sentiment analysis on Panel B texts and then split this sample (positive vs non-positive, or highly-positive vs positive) applying an event study and a regression analysis. If sentiment turned out to be a significant variable for abnormal returns, then we would be in scenario (iii). To exploit textual data by classifying analysts' ratings and extracting columns' sentiment, we first performed a text pre-processing, in particular by removing numbers and Italian 'stop words' (articles, prepositions, pronouns, etc.), and recoding some potentially interesting multi-word expressions into n-grams. Then we created a document-term matrix (about 16000 terms for Panel A) and select a limited number of potential features (81) based on their document frequency (at least 5% of all documents) and words' odds ratio from positive or non-positive analysts' ratings. Instead of a word count, we decided to use a binary coding, indicating the presence/absence of a given term in each column. This choice allowed us to limit the incidence of different text lengths and styles, usually without significantly decreasing the reliability of the statistical analysis (Ceron et al., 2017). To extract a sentiment measure for each column of Panel B, we exploited an

---

<sup>1</sup> Panel A included also 'The Stock of the Week' column, which appeared on the Saturday editions of 'Il Sole 24 ore' from 2005 to 2010 with very similar contents to the 'Letter to investor'.

ontological dictionary, the NRC lexicon (Mohammad and Turney, 2012). The sentiment score showed 49 different levels, all of them with a positive sign, so we decided to use the median value to split our sample into two groups, 'positive' and 'highly-positive', for the event study and to use the score as an independent variable in the regression model for abnormal returns (ARs).

### 3 Results

A summary of the event study for Panel B is reported in Table 1: from day from -3 to day +3 only the ARs of day 0 were statistically significant and the same went for the sign test (T2). We computed the ARs with two methods: the Market Adjusted Model and the standard one-factor Market Model. Moreover, we also computed ARs in the sample excluding cases with concurrent news. All models lead to similar results; therefore, we only show the results for the Market Adjusted Model:  $AR_{jt} = R_{jt} - R_{mt}$ ; where  $R_{jt}$  is the stock return of company  $j$  on day  $t$ ,  $R_{mt}$  is the stock market return (MILAN COMIT GLOBAL + R - PRICE INDEX) on day  $t$ , and  $AR_{jt}$  is the abnormal return of company  $j$  on day  $t$  ( $AR_{jt}$  are averaged across companies to get the mean Abnormal Return on day  $t$ ,  $ARt$ ).

**Table 1:** Panel B (491 records): Event study for daily Abnormal Returns (ARs) - Market Adjusted Model

<i>Day</i>		<i>ARt %</i>	<i>T</i>	<i>T2</i>
Wednesday	-3	-0.094	-1.04	-0.36
Thursday	-2	-0.085	-0.95	-1.35
Friday	-1	0.085	0.94	-0.81
Monday	0	0.275	3.05*	2.26*
Tuesday	1	0.095	1.05	1.00
Wednesday	2	-0.069	-0.76	-1.08
Thursday	3	-0.010	-0.12	0.27

\* denote significance at the 5% level. T= t-test. T2= nonparametric sign test (Ocana et al., 2002).

This first analysis reveals a significant market reaction to Panel B columns even without a clear analysts' recommendation, suggesting the possibility that (when a clear buy recommendation is missing) the inclusion of a stock in the column might affect buying decisions indistinctly with respect to what might be revealed by the text. Next, we tried to estimate the analyst rating in Panel A (where the rating was available) and predict it in Panel B, solely based on the words in the column texts. To achieve this goal, we classified analysts' rating with some data mining fitting techniques (Hastie et al., 2009): Probit, Logit, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Neural Networks, and Naïve Bayes. Considering the sample size of 379 for Panel A, we chose the optimistic corrected bootstrap (Efron and Tibshirani, 1993) as the resampling method for model validation, in order to obtain measures with less variance. QDA turned out to be the best classifier in terms of accuracy and Cohen's kappa (Table 2), and this model was used to predict the analysts' rating in Panel B.

**Table 2:** 200 bootstrap accuracy of analysts' rating classifiers (Panel A) – features: 81 words or n-grams

<i>Classifier</i>	<i>Accuracy</i>	<i>Kappa</i>
Probit	0.728	0.432
Logit	0.750	0.491
LDA	0.752	0.494
QDA	0.838	0.660
Neural Networks	0.802	0.597
Naïve Bayes	0.687	0.358

Despite a decent level of accuracy in Panel A, in the event study for Panel B, splitting the sample between positive and non-positive estimated stocks, the ARs are positive and significant in both groups on day 0, but of greater magnitude for non-positive ratings. The linear regression estimating ARs also shows the non-significance of the rating, both as a stand-alone variable or in interaction with the other predictors. This poor performance could also be related to the discontinuity in the column's author. Thus, to confirm scenario (i) or to imply that investors transform the content of the column into implicit recommendations, we used the results of the sentiment analysis applied to Panel B. As anticipated, we divided Panel B into two groups of stocks, positive and highly-positive, and applied an additional event study to the two groups. The results (Table 3) show an asymmetric reaction, as highly-positive stocks reported significantly higher ARs than positive stocks. Linear regression (Table 4) also shows that the sentiment variable turned out to be a significant predictor alone or in interaction with other variables.

**Table 3:** Panel B: Event study for daily Abnormal Returns (ARs) - Market Adjusted Model

<i>Day</i>	<i>Positive sentiment</i>				<i>Highly-positive sentiment</i>			
	<i>Obs.</i>	<i>ARt %</i>	<i>T</i>	<i>T2</i>	<i>Obs.</i>	<i>ARt %</i>	<i>T</i>	<i>T2</i>
-3	254	-0.262	-2.19*	-0.87	237	0.087	0.67	0.39
-2	254	-0.196	-1.64	-1.13	237	0.033	0.25	-0.78
-1	254	0.082	0.68	-1.88	237	0.088	0.67	0.78
0	254	0.191	1.59	0.38	237	0.366	2.80*	2.86*
1	254	0.084	0.70	0.38	237	0.107	0.82	1.04
2	254	0.012	0.10	-0.12	237	-0.156	-1.20	-1.43
3	254	-0.009	-0.07	-0.25	237	-0.012	-0.09	0.65

\* denote significance at the 5% level. T= t-test. T2= nonparametric sign test (Ocana et al., 2002).

**Table 4:** Linear regression for Panel B Abnormal Returns - Market Adjusted Model

<i>Variable</i>	<i>Estimate</i>	<i>Std. Error</i>	<i>T value</i>	<i>Pr(&gt; t )</i>
(intercept)	-3.88e-02	1.838e-02	-2.111	0.035*
PBV	-1.404e-03	9.656e-04	-1.454	0.147
Market Capit. (ln)	8.685e-03	2.691e-03	3.228	0.001**
Past Performance	3.111e-02	1.335e-02	2.330	0.020*
Beta	-2.513e-02	9.805e-03	-2.563	0.011*
Concurrent News	-2.177e-02	7.912e-03	-2.751	0.006**
Sentiment	1.777e-03	6.372e-04	2.789	0.005**
Sentiment*Market Capit (ln)	-3.283e-04	9.493e-05	-3.459	<0.001***
Sentiment*Beta	6.906e-04	3.126e-04	2.209	0.028*
Sentiment*Concurrent News	7.237e-04	2.685e-04	2.695	0.007**
Sentiment*PBV	3.975e-05	3.046e-05	1.305	0.192
Sentiment*Past Performance	-6.565e-04	4.639e-04	-1.415	0.158

\*\*\*, \*\* and \* denote significance at the 0.1%, 1% and 5% levels. The table reports a linear regression model for the abnormal returns (ARs). Explanatory variables include Price-to-book value (PBV), company's market capitalization, Market-adjusted company's past performance (stock return from day -215 to day -16 before the event day, net of the market return in the same period), risk (Market-model Beta), a dummy variable indicating the presence of concurrent news (the dummy variable equals 1 if the company reported in the column published press releases in day 0 or in day +1), columns' sentiment, and the interactions among the sentiment and all other explanatory variables. We show the estimated coefficients, the SEs, the t values, and the significance levels associated with each explanatory variable or interaction. The adjusted R-squared is 0.071 (F-statistics: 4.398, p-value: 2.62e-06).

## 4 Conclusion

The final results of the event study and the regression based on the sentiment of the columns without analysts' recommendations suggest that investors transform the content of the articles into implicit recommendations that, when (highly) positive, can direct their buying decisions. In future studies, we plan to investigate the use of lexicons specifically developed for financial texts since, as far as we know, there are no Italian ontological dictionaries for finance, as it happens in English, where the work of Loughran and McDonald (2011) has been widely used.

## References

1. Barber, B.M. and Odean, T.: All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors, *The Rev. of Financial Stud.*, 21,785–818 (2008) doi: 10.1093/rfs/hhm079
2. Brown, S.J. and Warner, J.B.: Using daily stock returns. The case of event studies, *Journal of Financial Economics*, 14, 3–31 (1985). [http://dx.doi.org/10.1016/0304-405X\(85\)90042-X](http://dx.doi.org/10.1016/0304-405X(85)90042-X)
3. Ceron, A., Curini, L., Iacus, S.M.: *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. Routledge, New York (2017) doi: 10.1080/23248823.2019.1619298
4. Cervellati, E.M., Ferretti, R., Pattitoni, P.: Market reaction to second-hand news: Inside the attention-grabbing hypothesis. *Appl. Econ*, 46(10), 1108-1121, (2014) doi: 10.1080/00036846.2013.866206
5. Efron, B. & Tibshirani, R.J.: *An introduction to the bootstrap*. Chapman & Hall, Boca Raton (1993).
6. Ferretti, R. and Sciandra, A.: The weight of words: textual data versus sentiment analysis in stock returns prediction, *Book of short papers - SIS 2020*, Pearson, 1099–1104, (2020).
7. Hastie, T., Tibshirani, R., Friedman, J.: *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York (2009) doi: 10.1007/978-0-387-84858-7
8. Loughran, T. and McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The J. of Finance*, 66(1), 35-65, (2011).
9. Mohammad, S. and Turney, P.: Crowdsourcing a word-emotion association lexicon, *Computational Intelligence*, vol. 29, no. 3, pp. 436-465, (2012) <https://doi.org/10.1111/j.1467-8640.2012.00460.x>.
10. Ocana, C.J., Pena, I., Robles, D.: Reactions of the International Stock Exchange to Company Employment Announcements: Redundancies and New Jobs, *J. of Business, Finance & Accounting* 29, 9-10, 1181–1208, (2002).

# Predictions of regional HCE: spatial and time patterns in an ageing population framework

## *Predizioni regionali di spesa sanitaria: analisi spazio-temporale in un contesto di invecchiamento della popolazione*

Laura Rizzi, Luca Grassetto, Divya Brundavanam, Alvisa Palese and Alessio Fornasin

**Abstract** This research aims at building reliable information about the drivers of expenditures at municipality level that could be used to guide policy makers in the public health care organization. In addition, the spatial temporal structure of the dataset (built using administrative data) is investigated considering a panel data spatial error model for per capita expenditures. The results of model estimation and the demographic projections are employed for the long-run predictions of regional health expenditures in the well-known aging population framework characterizing the Italian health system.

**Abstract** *La ricerca si pone l'obiettivo di fornire informazioni affidabili sui fattori determinanti le spese pro-capite a livello comunale, utili alla definizione delle politiche economiche regionali. La struttura spazio temporale del dataset (basato sugli archivi amministrativi) viene investigata con un modello spaziale per dati panel. I risultati del modello per la spesa individuale vengono infine utilizzati, assieme alle proiezioni demografiche della popolazione, per definire delle previsioni di spesa sanitaria regionale di medio-lungo periodo calate sulla situazione di invecchiamento della popolazione che caratterizza il sistema sanitario italiano.*

**Key words:** administrative data, HCE, panel data, spatial analysis, prediction.

## 1 Introduction

The aims and the context of the present paper are tied to central themes of modern health care policy, such as the increasing trend of health care expenditures (HCE),

---

L. Rizzi, L. Grassetto, D. Brundavanam and A. Fornasin  
Dept. of Economics and Statistics - University of Udine, Via Tomadini, 30/a Udine (Italy)  
e-mail: laura.rizzi@uniud.it

A. Palese  
Dept. of Health Sciences - University of Udine, Via Tomadini, 30/a Udine (Italy)

the role of time to death (TTD) and of population ageing on HCE patterns (see [6]), and the increasing burden of health care profiles due to medical innovations. This paper focuses on analysing time and spatial patterns of population HCE in the Italian region Friuli Venezia Giulia (FVG), to derive a potential projection of the regional health care budget in an ageing population framework. From 2002 to 2018 the demographic context in FVG shows a relevant increase in the proportion of people older than 65 and 75 years. Proportions increased from 10.58% to 13.52%, of people 75+, and from 21.45% to 25.94% of people older than 65 years.

Moreover, demographic projections show an increase in the share of the population older than 55 by 2028. The literature presents different approaches to the analysis of HCE trends. Some studies consider HCE dynamics in a macro-econometric framework ([8]); other studies are devoted to the assessment of the micro and macro determinants of HCE (see [7] and [9] for further details). In general, the increasing time patterns of HCE at the population level is widely assessed, pointing out the role of different factors, focusing on chronic pathology (see [1]) or on elderly population health burden, comparing HCE evolution in different countries as in [4] and in [3]. An open issue in the HCE literature is whether these expenditures are more concentrated in the last years before death, which is a proposition mainly assessed by all studies supporting the hypothesis of *compression of morbidity (the Time To Death causal effect hypothesis)* as, for instance, in [11]. Even if the onset of comorbidity is deferred to older age groups, HCE has grown dramatically at local and national levels. For this reason, some researchers focus on the role of factors such as longevity, scientific innovation, expansion of individual income levels on extension of long term care, known as the hypothesis of *morbidity expansion*. This analysis exploits micro data availability on the whole regional population on public 2002-2017 HCE, for out-patient care, pharmaceutical prescriptions and hospital services. This kind of data allow comparing dynamic and spatial behavior of total, per-capita and per-service HCE in all age classes. Finally, accounting for regional demographic projections and employing a model for the spatial-temporal municipality data, a forecast of future regional elderly health care budget is derived. Section 2 reports the data and the modeling approach adopted, while Section 3 is devoted to some descriptive analysis and spatial model estimation results. Finally, Section 4 presents the HCE forecasts scenario and the final discussion.

## 2 Data and Methods

The population dataset collects information on all regional patients for the period 2002 - 2017. We decide to aggregate the individual expenditures by the municipality of residence, age classes, and gender, and the analyses considers drugs prescriptions, out-patient, and in patient care services. The obtained dataset presents a distinctly panel-like structure - strongly balanced with repeated yearly observations for each age class and gender group within each municipality. First of all, a general descriptive analysis on regional time trends of total, per-capita, and per-service HCE

is performed on age macro classes (0-64; 65-74; 75+) to disentangle the role of population ageing, changes in health care profiles and increased costs of different types of services on trends of HCEs. Subsequently, focusing on elderly HCE patterns only, the impact of proximity among municipalities and distance from the main regional hospitals is studied using spatial temporal analysis. Social, demographic and economic factors are also considered at this stage of the analysis.

## 2.1 The spatial HCE model

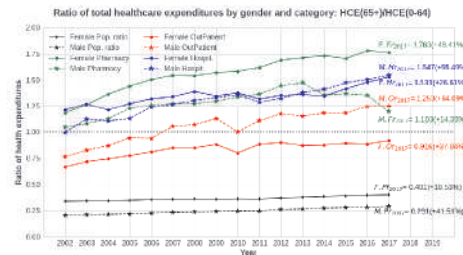
Given the geographical nature of the data, we hypothesise a spatial interaction between municipalities. We assume the existence of a sharing of social and cultural aspects between the neighboring municipalities. For this reason, to account for unobserved shared characteristics of the sub-populations, a spatial correlation in the error term is considered. We adopt a fixed-effects approach to shape our spatial panels, and a Spatial Error Model (SEM) ([2] and [10]) to capture the geographical interactions through the spatial autoregressive specification of the error term. The spatial correlation is based on an inverse distance matrix built on the city halls' geographical coordinates. We model the per-capita HCE associated with three types of services:  $pHCE_{drugs}$  - drug prescriptions,  $pHCE_{op}$  - out-patient and,  $pHCE_{hosp}$  - in-patient care. The outcome variable is measured on the elderly age class (65+) at the municipality level, and conditionally to gender (SEX, female coded as 1). The covariates consider four indicators measured at municipality level: the proportion of low-income resident population (defined as the ratio of tax-payers earning less than 15,000 euros/year to the whole taxpayers in each municipality – LIR, valued between 0 and 100%), the proximity of the municipality to a major hospital (DIST is one if the municipality is located near a hospital), altitude zone of the municipality (ALT, binary indicator with mountainous municipalities coded as one), and population density in each municipality (DENS, residents per  $km^2$ ). Moreover, to capture any substitution effect in the expenditures, in each model on a per-capita HCE type, the other two types of per-capita HCE are included as covariates. The total expenditure analysis is also performed but we omit its results for space reasons. The models are estimated on  $T = 16$  time points (years 2002-2017) and  $n = 214$  municipalities. The panel SEM model on per-capita expenditures for drugs prescriptions is:

$$pHCE_{drugs,i,t} = \beta_1(pHCE_{op,i,t}) + \beta_2(pHCE_{hosp,i,t}) + \beta_3(LIR_i) + \beta_4(SEX_i) + \beta_5(ALT_i) + \beta_6(DIST_i) + \beta_7(DENS_i) + \alpha_i + u_{i,t} \quad (1)$$

where  $u_{i,t} = \lambda \sum_{j \neq i} w_{i,j} u_{j,t} + \varepsilon_{i,t}$ , and  $\varepsilon_{i,t} \sim N(0, \sigma_\varepsilon^2)$ , index  $i = 1, \dots, 214$  refers to municipalities, and  $t = 2002, \dots, 2017$  to observational period. We similarly define the models to estimate  $pHCE_{op,it}$  and  $pHCE_{hosp,it}$ , the out-patient and in-patient per-capita expenditures, respectively.

### 3 Results

The first analysis shows the time trends for the ratios between the elderly (population 65+ according to usually adopted definition) and 0-64 population of observed HCEs. The analysis also considers gender distinction. The plot in Figure 1 compares the HCEs' patterns with the general demographic ratio identifying the ageing phenomenon.



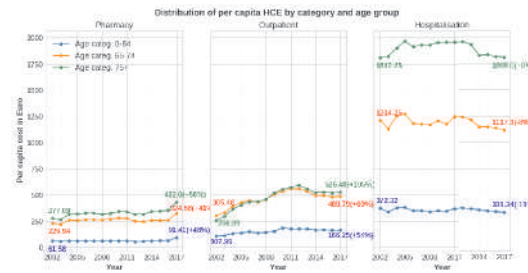
**Fig. 1** Time trends of the ratios between 65+ and 0-64 population and total HCE types, distinction by gender.

The relevance of the health care burden of individuals 65+ is also described in Figure 2, which points out the increase of HCEs burden through patterns of per-capita expenditures. The analysis of per-capita HCEs ratios reveals the the higher levels of in-patient component, with a decreasing trend, and the upwards patterns of out-patient and drugs services in all age classes. Further investigations regards the trends of per-occurrence costs and of per-capita number of services, in drug prescriptions and out-patients care (results are omitted for space reasons). These analyses on the population 65+ points out interesting insights. In particular, the increasing burden of hospital services to the whole budget growth is mainly due to the increased average cost of services. Furthermore, the drugs prescriptions and out-patient services' increases are mainly explained by the per-capita number of prescriptions and services.

Table 1 summarises the results of the SEM models estimation on elderly HCE, obtained using a quasi-maximum likelihood estimation approach through the xsmle package in STATA ([5]). The expenditures do not point out substitution effects but positive and significant correlations, as municipalities with higher elderly burden reveal higher per-capita HCE for all services' types. Gender always shows lower females' expected expenditures both in in-patient and out-patient services. The elderly living in municipalities with higher proportions of low-income people show higher in-patient services expenditures but lower HCEs for drug prescriptions and out-patient care. Similarly, alpine municipalities reveal significant lower per-capita drug prescriptions expenditures. The proximity to one of the major hospitals (Udine, Trieste, Pordenone) reduces drugs and out-patient services expenditures. Furthermore, the spatial component is relevant for all kinds of per-capita HCEs. These



Predictions of regional HCE



**Fig. 2** Yearly trend of per-capita HCE by age class and type of service: drugs (*Pharmacology*), left panel; out-patient and in-patient services (*Hospitalization*), left and right panels.

results allow us to compute the predicted HCEs at 2023 and 2028, derived considering population demographical forecasts, at municipality and gender level, computed through the cohort component approach using 2018 as a basis. The fertility rate is assumed fixed during the projection period, and derived for the year 2017 for each municipality considering the province level. Moreover, we assume that life expectancy at birth increases over the period adopting the 2018 values 80,7 and 85.5, and the 2028 values 82.3 and 86.7, for males and females, respectively. The projections do not consider the migration phenomenon. The differences between HCEs predicted through the SEM models, for year 2023 and 2028 (*HCE predictions*), and the observed 2017 HCEs are reported in Table 2.

**Table 1** Spatial fixed-effects model on per-capita HCE in population +65, estimation results on drug prescriptions, out-patient and in-patient services

Model	Drugs prescriptions		Out-patient		In-patient	
	Coef.	$p >  z $	Coef.	$p >  z $	Coef.	$p >  z $
Per-capita Out-patient	0.017	< 0.001			0.237	< 0.001
Per-capita In-patient	0.014	< 0.001	0.046	< 0.001		
Per-capita Drugs prescriptions			0.266	< 0.001	0.881	< 0.001
Low-income population %	-0.232	0.077	-3.007	< 0.001	4.482	< 0.001
Gender	-38.565	< 0.001	-111.836	< 0.001	-423.217	< 0.001
Mountainous area	-7.181	0.002	14.456	< 0.001	17.444	0.191
Hospital proximity	-10.806	< 0.001	-15.436	0.015	23.838	0.093
Population density	-0.007	0.124	0.033	0.047	0.048	0.198
Spatial parameter						
$\lambda$	1.717	< 0.001	0.379	< 0.001	0.34	< 0.001

**Table 2** Differences between predicted HCEs, at 2023 and 2028, and observed HCEs in 2017 (differences based on predicted HCEs by gender in last columns).

HCE type	2023-2017 (%)	2028-2017 (%)	(F) 2028-2017	(M) 2028-2017
Drugs prescriptions	-22,850,610 (-19.2%)	-18,457,057 (-15.5%)	-11,100,000	-7,401,043
Out-patient services	58,137 (0.04%)	7,462,866 (4.7%)	793,796	6,669,069
In-patient services	66,890,125 (14.5%)	91,379,119 (19.8%)	39,600,000	51,700,000
Total	44,097,652 (6.0%)	80,384,928 (10.9%)	29,293,796	50,968,026

## 4 Conclusions

The central insight suggested by the descriptives is the concurrent role of different drivers on global HCE increase at the regional level, with heterogeneous impacts on different types of health services. Population ageing seems to be the main driver of the increase of hospital services costs, while scientific and technological innovation and changes in care profiles seem to be predictive for the growth of average out-patient services costs. Finally, the rising impact of drug (pharmacy) costs is mostly due to the increase of the number of per-capita prescriptions. The SEM model approach points out the relevance of the spatial similarities between neighboring municipalities, due to unobserved characteristics of resident elderly populations, and the lower health care burdens of elderly living in more economically deprived and in mountainous municipalities. Moreover, male present an higher health care burden. These results allow us to compute the predicted per-capita and total elderly expenditures, at the municipality and gender level, at 2023 and 2028. The HCEs, predicted through the SEM models, reveal a general decrease of drug prescriptions expenditures, a heterogeneous pattern for out-patient care and a relevant increasing trend for in-patient services, if compared with the 2017 observed levels.

## References

1. Abegunde D.O., Mathers C.D., Adam T., Ortegón M., Strong K.: The burden and costs of chronic diseases in low-income and middle-income countries. *Lancet* **370**, 1929-38 (2007)
2. Anselin, L.: *Spatial Econometrics: Methods and Models*. Springer Science & Business Media. (2013).
3. Baltagi B.H., Lagravinese R., Moscone F., Tosetti E.: Health Care Expenditure and Income: A Global Perspective. *Health Econ.* **26**, 863-874 (2017)
4. Bech M., Christiansen T., Khoman E., Lauridsen J., Weale M.: Ageing and health care expenditure in EU-15. *Eur J Health Econ*, **12**, 469-478 (2011)
5. Bellotti, F., Huhghes, G., Piano Mortari, A.: Spatial panel-data models using Stata, *The Stata Journal*, **17**, (1), 139-180 (2017)
6. Breyer F., Costa-Font J., Felder S.: Ageing, health, and health care. *Oxford Review of Economic Policy*, **26**, Number 4, 674-690 (2010)
7. Koopmanschap, M., De Meijer, C., Wouterse, B., Polder, J.: Determinants of health care expenditure in an aging society. Panel Paper, **22**, (2010).
8. Van Baal, P.H., and Wong, A.: Time to death and the forecasting of macro-level health care expenditures: Some further considerations, *J Health Econ*, **31** (6), 876-887 (2012)
9. Wang, Z.: The determinants of health expenditures: evidence from US state-level data, *Appl Econ*, **41** (4), 429-435 (2009)
10. Ward, M.D., Gleditsch, K.S.: *Spatial Regression Models*. SAGE, (2018)
11. Werblow, A., Felder, S., and Zweifel, P.: Population ageing and health care expenditure: a school of 'red herrings'?, *Health Econ*, **16** (10), 1109-1126 (2007)

## 4.36 Surveillance methods and statistical models in the Covid-19 crisis

# The Italian Social Mood on Economy Index during the Covid-19 Crisis

## *Il Social Mood on Economy Index durante la crisi del Covid-19*

Alessandra Righi and Diego Zardetto<sup>1</sup>

**Abstract** Since 2016, Istat has investigated whether social media messages may be successfully exploited to assess the Italian mood on the country's economic situation. An experimental high-frequency sentiment index based on Twitter data, called Social Mood on Economy Index (SMEI) has been published since October 2018. This index, derived from samples of public tweets in Italian captured in real-time, has already gained a good spread among economic analysts for short-term analysis or forecasting models. This paper is aimed at studying the relationships of a new indicator derived by the SMEI, namely, the share of tweets containing the terms "Coronavirus" or "Covid" in the text out of the total tweets used for SMEI, with some monthly indicators coming from nontraditional sources (e.g., Google Trends series, the daily number of Covid-19 deaths and new positive cases reported by the Civil Protection Department) during the period of Covid-19 pandemic. This analysis allows better understanding of the actual contribution in terms of information coming from the SMEI and assessing its robustness through the comparison of series.

**Abstract** Dal 2016 l'Istat studia se i messaggi dei social media possano essere sfruttati con successo per valutare l'umore degli italiani sulla situazione economica del Paese. Dall'ottobre 2018 viene pubblicato un indice sperimentale di sentiment ad alta frequenza basato sui dati di Twitter, denominato Social Mood on Economy Index (SMEI). Derivato da campioni di tweet pubblici catturati in tempo reale, l'indice ha già guadagnato una buona diffusione tra gli analisti per analisi economiche congiunturali o modelli previsionali. Vengono qui studiate le relazioni di un nuovo indicatore derivato dallo SMEI, ovvero la quota di tweet contenenti i

---

<sup>1</sup> Alessandra Righi, Istat; email: righi@istat.it  
Diego Zardetto, Affiliation; zardetto@istat.it

*termini "Coronavirus" o "Covid" nel testo sul totale dei tweet utilizzati per lo SMEI, con alcuni indicatori mensili tratti da fonti non tradizionali (es. serie di Google Trends, numero giornaliero di decessi per Covid-19 e nuovi casi positivi segnalati dal Dipartimento della Protezione Civile) durante il periodo di pandemia di Covid-19. L'analisi permette di comprendere meglio l'effettivo contributo informativo dello SMEI e di valutarne la robustezza attraverso il confronto delle serie.*

**Key words:** sentiment analysis, Twitter, Google Trends

## 1 Introduction

In recent years, the Italian National Institute of statistics has investigated whether social media messages may be successfully exploited to assess the Italian mood on the country's economic situation. In October 2018, this effort led to the first release of the Social Mood on Economy Index (SMEI), an experimental high-frequency sentiment index based on Twitter data. Although daily SMEI values only cover the period after February 2016, the new index, derived from samples of public tweets in Italian captured in real-time, has already gained a good spread among economic analysts for short-term analysis or forecasting models.

This paper is aimed at presenting this new statistical tool for economic analysis studying its relationships with some daily and monthly macroeconomic indicators coming from traditional and nontraditional sources. This analysis, performed in the period before and during the Covid-19 pandemic, allows better understanding of the actual contribution in terms of information coming from the SMEI. Other time series are produced from nontraditional sources to relate them to a new indicator derived by the SMEI, namely, the share of tweets containing the terms "Coronavirus" or "Covid" in the text out of the total tweets used for SMEI. The daily number of Covid-19 deaths and new positive cases reported by the Civil Protection Department are compared to the new indicator to assess the robustness of the SMEI during the Covid-19 pandemic.

## 2 Background

Sentiment analysis is an increasing area of research and application, due to the enormous amount of unstructured natural language data currently available. Many studies are published presenting main algorithms to be used for text mining and sentiment analysis (Gandomi and Haider 2015; Feldman and Sanger 2007). Further methodological developments, focused on sentiment analysis on social media short texts performing the analysis in different ways. There are methods based on unsupervised learning techniques (Pak and Paroubek 2010), lexicon-based methods (Taboada et al. 2011), and combinations of two previous approaches (Kolchyna et

The Social Mood on Economy Index

al. 2015). Other methods are based on the Twitter specificities as is the case of polarization, controversy and topic tracking in time, developed through network measures and clustering techniques (Garimella et al. 2016) or based on the use of the hashtags classification developed through probabilistic models (Coletto et al. 2016b).

Many studies making use of lexicon-based methods refer to the English language (Miller 1995, Strapparava and Valitutti 2005; Esuli and Sebastiani 2006), and there are only a few studies adopting the method for the Italian language due to a minor development of the lexicon in this language. However, Basile and Nissim (2013) developed the Sentiment Italian Lexicon (Sentix), a vocabulary whose lemmas are associated with pre-computed sentiment scores, which is the result of the alignment of several sources. Istat followed the lexicon-base method and used the Sentix lexicon in developing the experimental high-frequency sentiment index on the overall economic situation from Twitter data (Fabbri et al., 2018 and Zardetto, 2018).

The research question refers to the robustness of the SMEI during the pandemic. How did the index react during the pandemic, was it sufficiently inclusive of the messages on the pandemic and the emergency?

### 3 Background

Twitter's Streaming API is used to collect samples of public tweets matching a filter made up of 60 keywords relevant for the study of the general and personal economic dimension. All the filtered tweets reported in a single day are processed as a single block to compute daily index values. Messages are cleaned and normalized and then undergo a sentiment analysis procedure, which calculates positive and negative sentiment scores for each tweet.

Using an Italian sentiment lexicon, the scores of matched words are calculated and then averaged to yield tweet-level scores. Tweets are then clustered according to their sentiment scores into three mutually exclusive classes (Positive, Negative and Neutral). The daily index value is derived as an appropriate central tendency measure of the score distribution of the tweets belonging to the Positive and Negative classes and is linearly transformed<sup>1</sup>.

Special care has been devoted to making the index robust against possible contaminations by off-topic tweets that might pass the filter. An automatic outlier detection procedure is set to discriminate truly anomalous data from proper data within the daily time series. Daily index values classified as truly anomalous are imputed via a method of multivariate interpolation (nearest-neighbor interpolation).

However, estimates can possibly be affected by bias due to different causes: 1) the Italian Twitter users cannot be considered a representative sample of the Italian population, due to different Twitter penetration rates among various sub-population

---

<sup>1</sup> The transformation makes equal to zero the SMEI long-run mean (referred to the baseline period 10 February 2016 – 30 September 2018).

(e.g., young people is overrepresented among Twitter users); 2) the free download from Twitter API did not allow full access to all the tweets generated by the nine-million Italian active Twitter users, but to a subset (a non-probability sample) of tweets.

The information coming from the SMEI was further exploited to produce one new daily Twitter series: the share of tweets containing the terms “Coronavirus” or “Covid” in the text (among those used in the measure of the SMEI) out of the total tweets used to compute the SMEI. This series, available for the period from 1 January 2020 to 15 October 2020, has been normalized by mapping to 100 its peak value in the period (31%, registered on 24 February 2020). From now on we will refer to this series as “filtered Covid SME series”.

Using Google Trends other daily series are calculated for the same period computing the weighted average of the query shares referring to the keywords “Coronavirus” and “Covid” for the Italian territory, using as weights the relative “popularity levels” of the two keywords in the period supplied by Google Trends (18 for “Coronavirus” and 5 for “Covid”). This Google Trends daily series is normalized by mapping to 100 its peak value in the period (observed on 23 February 2020).

The daily series on the number of Covid-19 newly infected people and deaths coming from the “health risk associated with Coronavirus infection database” of the Civil Protection Department of the Presidency of the Council of Ministers are also considered for further analysis.

A study on how SMEI reacted during the pandemic is carried out to try to assess the robustness of the index during the pandemic. It is particularly important to understand if the considered tweets in the measurement were sufficiently inclusive of the talks on the pandemic and the problems related to the health emergency and the resulting economic crisis. As the series are too short, standard econometric analysis would face serious hindrances and possibly turn out unsuitable. Thus, graphical analysis is used to compare the filtered Covid SME series to the Google Trends series referring to Covid/Coronavirus shares of queries and to the series of the death by Coronavirus observed in Italy.

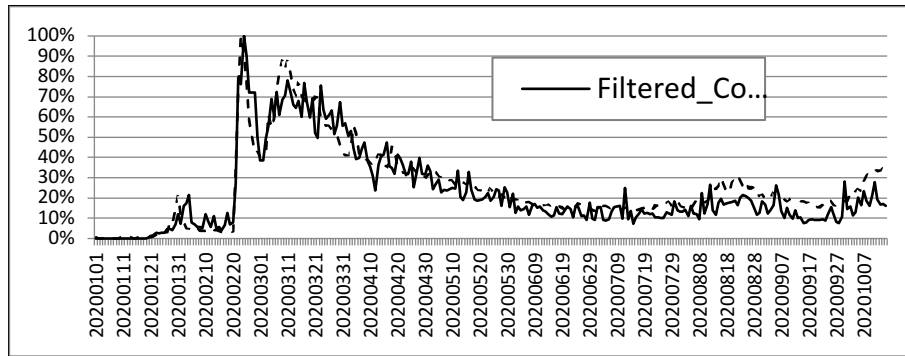
## 4 Main results

From the beginning of March 2020, SMEI showed a strong increase in the volume of tweets, which in two days more than doubled (from around 67 thousand tweets on 3 March 2020 to around 115 thousand ones on 5 March 2020) to reach 144 thousand on 28 March; then, from the end of May, the tweets returned to a volume comparable to the one before the Covid-19 crisis.

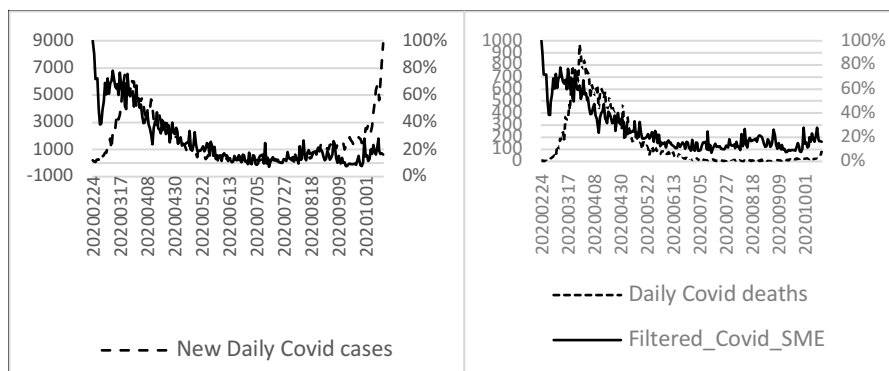
Signals extracted from the Covid SME series and Google Trends look quite similar. In both cases, series peak on 23-24 February 2020. Both present a negative peak at the beginning of March and an increase before mid-March. After March, the interest for the keywords reduced by 50% and the decrease continued until June,

The Social Mood on Economy Index

when both the shares remain between 10% and 20%. The increase of Google Trends series started again in August (with a peak registered on 22 August 2020), afterwards opening a declining phase with a following rapid increase observed in October. In this latest phase, the increase is less evident for the filtered Covid SME series, even if peaks are observed at the beginning and at the end of September (Figure 1).



**Figure 1: Comparison of the Istat filtered Covid SME and the Google Trends Covid/Coronavirus query shares series – 01 Jan 2020 – 15 Oct 2020**



**Figure 2: Comparison between the Istat filtered Covid SME series and the series of the number of daily new COVID cases and deaths in Italy - 01 Jan 2020 – 15 Oct 2020**

Comparisons with the daily series on the number of newly infected persons and deaths, coming from the Civil Protection Department health risk associated with Coronavirus database, showed that the series of filtered Covid SME and the Covid deaths behave more similarly after 24 March 2020; the series evolved similarly decreasing very rapidly and remaining to almost the same level (between 10% and 20%, in the case of Twitter SMEI) from the beginning of June 2020 onward. Whereas the series of newly positive cases register an increase after the summer period not observed in the filtered Covid SME series (Figure 2).



The interest in the Covid-19 issue among the Italians declined with the reduction of the severity of the sanitary emergency and the increase in the new cases in September and October 2020 caused a slight increase of the number of the filtered Covid SME conversations.

## References

1. Basile, V., Missim, M. Sentiment analysis on Italian tweets. 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA 2013, June 14th 2013, (pp. 100-107). Atlanta, U.S.A. (2013).
2. Coletto, M., Lucchese, C., Orlando, S., Perego, R. Polarized user and topic tracking in Twitter. SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, July, 945–948 (2016b)
3. Esuli, A., & Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In LREC, May, 6, 41--422 (2006)
4. Fabbri, C., Iannaccone, R., Righi, A., Scannapieco, M., Testa, P., Valentino, L., Zardetto, D., Zurlo, D., The Social Mood on Economy Index. Methodological note. Istat, Rome (2018)
5. Feldman, R., Sanger, J. et al. (2007). The Text mining Handbook: advanced approaches in analysing unstructured data. Cambridge U.P., Cambridge.
6. Gandomi, A., Haider, M. Beyond the hype. Big data concepts, methods and analytics. Int J Inform manage, 35(2), 137--144 (2015)
7. Garimella, K., De Francisci Morales, G., Gionis, A. & Mathioudakis, M. Quantifying controversy in social media. The 9th ACM International Conference on Web Search and Data Mining WSDM '16, (pp. 33–42) San Francisco, California (2016)
8. Hyndman, R. J., & Khandakar, Y. Automatic time series for forecasting: the forecast package for R (No. 6/07). Monash University, Department of Econometrics and Business Statistics, Clayton VIC Australia (2007)
9. Kolchyna, O., Souza, T. T., Treleaven, P. & Aste, T. Twitter sentiment analysis: lexicon method, machine-learning method and their combination. arXiv preprint arXiv:1507.00955 (2015).
10. Miller, G. A. WordNet: a lexical database for English. Communications of the ACM, 38(11), 39--41 (1995).
11. Pak, A., Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh Conference on International Language Resources and Evaluation. LREC 2010, 17-23 May 2010, Valletta, Malta, 1320--1326 (2010)
12. Strapparava, C., & Valitutti, A. Wordnet affect: an affective extension of wordnet. In LREC, May, 4, No. 1083-1086, 40 (2004)
13. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. Lexicon-based methods for sentiment analysis, Comput linguisti, 37 (2), 267--307 (2011)
14. Zardetto, D. Using Twitter Data for the Social Mood on Economy Index, Atti della XIII Conferenza nazionale di statistica, Rome, 4-6 July 2018, 385--390 (2018)

# Modeling the first wave of the COVID-19 pandemic in the Lombardy region, Italy, by using the daily number of swabs

## *Modellazione della fase iniziale della pandemia da COVID-19 in Lombardia, con l'utilizzo del numero giornaliero di tamponi*

Claudia Furlan and Cinzia Mortarino

**Abstract** The daily fluctuations in the released number of Covid-19 cases played a big role both at the beginning and in the most critical weeks of the outbreak, when local authorities in Italy had to decide whether to impose a lockdown and at which level. Public opinion was focused on this information as well, to understand how quickly the epidemic was spreading. In this work, we propose a nonlinear asymmetric diffusion model, which includes information on the daily number of swabs, to describe daily fluctuations in the number of confirmed cases in addition to the main trend of the outbreak evolution. The proposed model is compared with alternative model structures in the application to data of the Lombardy region.

**Abstract** *Le oscillazioni giornaliere nel numero di contagi diagnosticati di Covid-19 sono state al centro dell'attenzione nelle fasi iniziali della pandemia, quando le autorità avevano scarsi elementi per decidere quali restrizioni adottare. Anche l'opinione pubblica e i media erano costantemente focalizzati su questo dato quotidiano, per cercare di ricavarne elementi sull'evoluzione del contagio. In questo lavoro, proponiamo un modello di diffusione nonlineare asimmetrico per descrivere, oltre al trend dei contagi diagnosticati, le oscillazioni giornaliere. Il modello utilizza come input il numero di tamponi processati quotidianamente. Il modello proposto è sottoposto a comparazione con altri cinque modelli sui dati della regione Lombardia.*

**Key words:** nonlinear models, Generalized Bass model, logistic, diffusion

## 1 Introduction

Italy was the first nation to be affected by Covid-19 after China, and the epidemic has mainly been located in Northern Italy. On February 21st, 2020, an infected patient was detected in the small town of Codogno, which is located in the Lombardy

---

Claudia Furlan and Cinzia Mortarino  
Department of Statistical Sciences, Padova, Italy,  
e-mail: furlan@stat.unipd.it, mortarino@stat.unipd.it

region. During the *first wave*, among the Italian regions, Lombardy is the most affected by the epidemic, with a death toll three times greater than that in China [1]. It is apparent that, in Italy, the regional autonomy regarding health policy has resulted in services with different levels of quality [1], such as the number of beds and the capacity of processing swabs. With regard to the number of beds in Italy, the forecasts of hospitalisations was faced by [2] for the bordering Veneto region, while [3] modeled the intensive care unit occupancy.

The capacity of processing swabs is of particular importance for detecting the state of the epidemic, measuring the lockdown effects and, most importantly, reducing the outbreak. Our opinion is that it is necessary to include the number of swabs to describe the local fluctuations in the epidemic evolution in addition to detecting the main trend. At the beginning of the outbreak, the curve of confirmed cases was usually modeled through an exponential [4] or a logistic growth model [5]. When the data collection window became long enough, the models were usually of two types: the compartmental and ARIMA models.

We made an effort to describe the cumulative number of confirmed cases in the Lombardy region, based on the combination of a nonlinear model and the number of completed swabs. In the class of growth models, we propose a new version of the dynamic potential model [6], where the novelty consists of the formulation of a new intervention function with the number of daily swabs as an explanatory variable. The model is particularly parsimonious since the intervention function has only one additional parameter. The base of the dynamic potential model was chosen since a) it has an asymmetric shape and makes it possible to model a ‘saddle’, which is a rather common nonlinear pattern; b) it gives an estimate of the total number of confirmed cases at the end of the epidemic; and c) the total number of confirmed cases is not fixed throughout the outbreak, but it is allowed to change over time. Since the capability of processing swabs increased over time and, consequently, the meeting criteria for people for being tested were enlarged with the aim of detecting a larger number of asymptomatic positive subjects, it is sensible to suppose that the number of diagnosed cases increases with time.

The proposed model was compared with five alternative growth models described in Section 2. Three-week forecasts of the spreading dynamics were provided for each model as well. The models were compared in terms of  $R^2$  and BIC values, for the cumulative values. The squared linear correlation coefficient between observed and fitted daily values was evaluated as well.

## 2 Models

A general diffusion of innovations model can be defined through a nonlinear regression model as follows:

$$y(t) = z(t, \vartheta) + \varepsilon(t), \tag{1}$$

where  $y(t)$  are the cumulative sales of a product at time  $t$  and  $z(t, \vartheta) = z(t)$  is a specific structure to be used to describe an evolution process. Here,  $\varepsilon_t$  are assumed

Title Suppressed Due to Excessive Length

to be i.i.d. Gaussian with variance  $\sigma^2$ . The components of the parameter vector  $\vartheta$  are jointly estimated using nonlinear least squares (or, equivalently, likelihood estimation).

In this paper, we will compare the performance of alternative evolution structures. The basic model is a logistic one (LOG):

$$z(t) = m \frac{e^{\frac{t-\lambda}{\eta}}}{1 + e^{\frac{t-\lambda}{\eta}}}, \quad (2)$$

where  $\lambda$  is the mode, median and average of the distribution, while  $\eta$  is a shape parameter. Parameter  $m$  is the market potential, which is the limiting value for  $z(t)$ , as  $t$  goes to infinity.

The Generalized Bass Model corresponds to:

$$z(t) = m \frac{1 - e^{-(p+q) \int_0^t w(\tau) d\tau}}{1 + \frac{q}{p} e^{-(p+q) \int_0^t w(\tau) d\tau}}, \quad (3)$$

where  $m$  is the market potential,  $p$  is the innovation coefficient,  $q$  is the imitation coefficient and  $w(t)$  can be any integrable function. The effect of the intervention function  $w(t)$  is to accelerate or decrease diffusion with respect to a symmetric unimodal path, which would arise in (3) for  $w(t) = 1$  for all  $t$  values. For  $t$  values such as  $w(t) > 1$  diffusion is accelerated, while  $w(t) < 1$  corresponds to time periods with decreased diffusion speed. Below, we examine the model (GBM<sub>RECT</sub>) arising when  $w(t)$  is specified by the so-called *rectangular* shock:

$$w_R(t) = 1 + cI_{a \leq t \leq b}. \quad (4)$$

This allows us to describe the diffusion of a product for which we observe a constant shock with intensity  $c$ , either positive or negative, in the time interval  $[a, b]$  [7].

Due to the asymmetric path observed for almost every region, we also examine the more flexible Bemmaor model, in an extended version with a rectangular shock (BeGBM<sub>RECT</sub>):

$$z(t) = m \frac{1 - e^{-(p+q) \int_0^t w_R(\tau) d\tau}}{[1 + \frac{q}{p} e^{-(p+q) \int_0^t w_R(\tau) d\tau}]^A}, \quad (5)$$

where  $A$  is a further parameter allowing for asymmetry (positive asymmetry for  $A > 1$ , negative asymmetry for  $A < 1$ ), with function  $w_R$  specified as in (4).

A different way to provide flexibility to the evolutive structure can be obtained through a dynamic market potential model [6], eventually perturbed by shocks

$$z(t) = m \sqrt{\frac{1 - e^{-(p_c+q_c)t}}{1 + \frac{q_c}{p_c} e^{-(p_c+q_c)t}} \frac{1 - e^{-(p+q) \int_0^t w(\tau) d\tau}}{1 + \frac{q}{p} e^{-(p+q) \int_0^t w(\tau) d\tau}}}, \quad (6)$$

where  $p_c$  and  $q_c$  are two parameters to describe how fast the dynamic market potential approaches its maximum value,  $m$ , while  $w(t)$  is a general intervention function. If in model (6) we use, as proposed in [8], the following intervention function

$$w_s(t) = 1 + \alpha_1 \cos\left(\frac{2\pi t}{s}\right) + \alpha_2 \sin\left(\frac{2\pi t}{s}\right), \tag{7}$$

we allow the model to incorporate cyclic seasonal fluctuations of width  $\alpha_1$  and  $\alpha_2$  with period  $s$  (DMPseas). If  $w(t) = 1$ , we obtain dynamic market potential (DMP).

Here, we propose to assess the usefulness of a dynamic market potential model as in (6), but with an intervention function depending upon the number of swabs analyzed at day  $t$ ,  $B(t)$  (DMPsw). In particular, we suggest using

$$w_B(t) = 1 + \xi \left( \frac{B(t) - \mu_B}{\sigma_B} \right), \tag{8}$$

where  $\mu_B$  and  $\sigma_B$  are the average and the standard deviation, respectively, of the  $B(t)$  values recorded during the observation period. It is easy to appreciate that such a structure accelerates, with respect to an underlying trend described by a DMP, the number of cases whenever  $B(t)$  exceeds its average, while cases are reduced with a below-average number of swabs.

### 3 Applications

Models of Section 2 were applied to the data of Lombardy, and forecasts up to May 24th are provided (three weeks ahead for each region). The six models were fitted to the cumulative confirmed cases using NLS estimation. Table 1 summarizes the values of the determination index  $R^2$  for all models: the huge values of  $R^2$  are unsurprising, given that we are working with cumulative data and any S-shaped fitting produces high determination indexes.

Lombardy is the Italian region where COVID-19 spread in the most dramatic way. The total number of infected people on May 3rd was 77528 with more than 14000 deaths (about half of the death toll up to that date in Italy as a whole). The results for Lombardy are displayed in Table 1 ( $R^2$ , BIC and  $\rho^2$ ), and in Figure 1, where observed and fitted daily values are plotted.

For this region, the logistic (Figure 1(a)) is the less effective model in describing the asymmetrical evolution of the epidemic.

A positive ( $\hat{c} > 0$ ) rectangular shock is significantly diagnosed at the beginning of the time series, both in the  $GBM_{RECT}$  and the  $BeGBM_{RECT}$ . The  $GBM_{RECT}$  estimates the end of the shock on March 25th ( $t \simeq 34$ ), but according to Figure 1(b),

**Table 1**  $R^2$  of the nonlinear models and corresponding BIC (cumulative data as response variable) and squared linear correlation coefficient,  $\rho^2$ , between observed *instantaneous* sales and fitted *instantaneous* sales.

	LOG	$GBM_{RECT}$	$BeGBM_{RECT}$	DMP	DMPseas	DMPsw
Lombardy $R^2$	0.993010	0.999629	0.999900	0.999834	0.999860	0.999919
BIC	1143.348	941.8574	850.6930	878.7936	879.2002	830.5268
$\rho^2$	0.593900	0.690817	0.826706	0.803006	0.820098	0.902698

this is not perfectly matching with the data. This is the reason why, for this model,  $\rho^2$  is particularly small (0.690817).

Conversely, the BeGBM<sub>RECT</sub> better identifies the end of the shock three days later, on March 28th, when we observe a relevant stable decrease. For this region, the lockdown policy had a delayed effect compared to what happened in Veneto, as the decrease in the number of cases was registered 20 days after March 8th, while the incubation period is up to 14 days. One reason for such a wider interval could be possible delays in taking and processing the swabs; in fact, the health system of Lombardy experienced an unexpected overload.

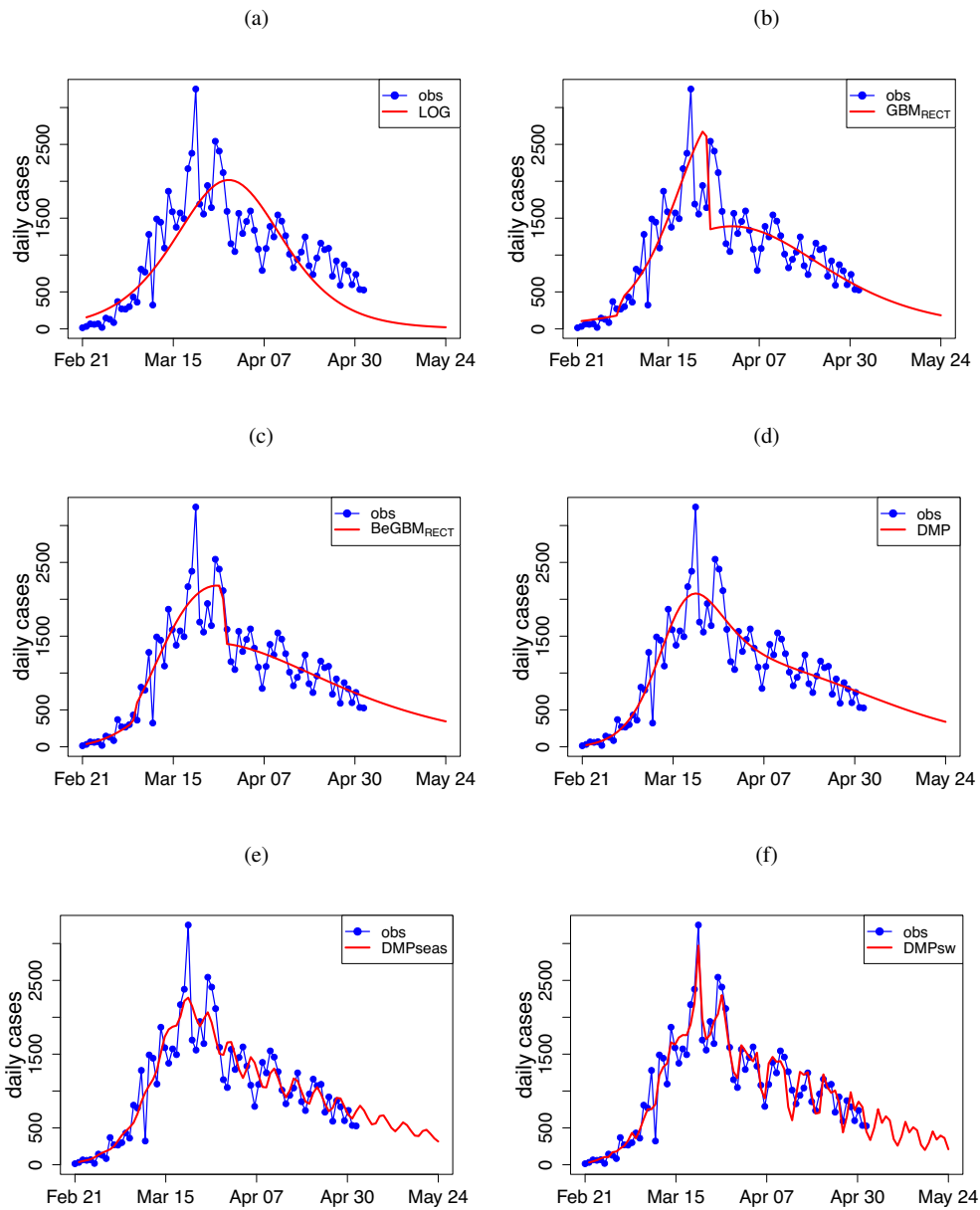
With the DMP model (Figure 1(d)), it is possible to fully appreciate the asymmetrical shape of the outbreak, especially the slow decrease in the number of cases in this region. However, its performance in terms of  $R^2$ ,  $\rho^2$  and BIC is worse than that of the BeGBM<sub>RECT</sub>.

The performance of the DMP<sub>seas</sub>, with a weekly cycle ( $\hat{s} = 7.005$  days), is not satisfactory, as it does not adequately capture the fluctuations (except for the very end of the series). Here, too, the  $R^2$ ,  $\rho^2$  and BIC values are worse than those obtained with the BeGBM<sub>RECT</sub>.

Finally, the DMP<sub>sw</sub> (Figure 1(f)) performs very well. With this model, we obtained the largest values for  $R^2$ , 0.999919, and  $\rho^2$ , 0.902698. The BIC value for this model, 830.5628, supports it with respect to the BeGBM<sub>RECT</sub> (850.6930), which was the best model up to this point. The proposed model, which is highly parsimonious, is able to describe the daily fluctuations in cases very well and proved to be the best of the models analysed here for Lombardy.

## References

1. Indolfi, C., Spaccarotella, C.: The outbreak of covid-19 in italy. *JACC: Case Reports* **2**(9), 1414–1418 (2020)
2. Gregori, D., Azzolina, D., Lanera, C., Prosepe, I., Destro, N., Lorenzoni, G., Berchiolla, P.: A first estimation of the impact of public health actions against covid-19 in Veneto (Italy). *Journal of Epidemiology and Community Health* **74** (10), 858–860 (2020)
3. Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., Lovison, G.: An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. *Biometrical Journal* (2020) doi:202000189
4. Remuzzi, A., Remuzzi, G.: Covid-19 and italy: what next? *Health policy* **395**, 1225–1228 (2020)
5. Shen, C.Y.: A logistic growth model for covid-19 proliferation: Experiences from china and international implications in infectious diseases. *International Journal of Infectious Diseases* **96**, 582–589 (2020)
6. Guseo, R., Guidolin, M.: Modelling a dynamic market potential: a class of automata networks for diffusion of innovations. *Technological Forecasting and Social Change* **76**(6), 806–820 (2009)
7. Guseo, R., Dalla Valle, A.: Oil and gas depletion: diffusion models and forecasting under strategic intervention. *Statistical Methods and Applications* **14**(3), 375–387 (2005)
8. Guidolin, M., Guseo, R.: Modelling seasonality in innovation diffusion. *Technological Forecasting and Social Change* **86**, 33–40 (2014)



**Fig. 1** Lombardy. Observed and fitted values with the alternative models. (a) Logistic (LOG); (b) GBM with rectangular shock (GBM<sub>RECT</sub>); (c) Bemmar GBM with rectangular shock (BeGBM<sub>RECT</sub>); (d) Dynamic market potential (DMP); (e) Dynamic market potential+seasonal effect (DMP<sub>seas</sub>); (f) Dynamic market potential+swabs (DMP<sub>sw</sub>).

# Analysing the Covid-19 pandemic in Italy with the SIPRO model

## *Analizzare la pandemia di Covid-19 in Italia con il modello SIPRO*

Martina Amongero, Enrico Bibbona, Gianluca Mastrantonio

**Abstract** We propose an epidemic compartmental model that extends the classical SIR, in order to allow for an estimate of the unobserved infected people that have not been tested. The model is then fitted to the epidemic curves of the 20 Italian regions using Bayesian methods for mixed-effects models. Despite the interpretability of the model, we still face identifiability issues. We explain how we can alleviate them.

**Abstract** *Abstract in Italian* Proponiamo un modello epidemico compartimentale che estende il modello SIR classico, allo scopo di permettere la stima del numero (latente) di persone infette che non sono state testate. Il modello è poi stimato sulla base delle curve epidemiche delle 20 regioni italiane, usando un metodo Bayesiano per modelli ad effetti misti. Nonostante l'interpretabilità del modello, si incontrano comunque alcuni problemi di identificabilità. Li illustriamo e indichiamo come possono essere mitigati.

**Key words:** SIR, Covid-19, MCMC, Compartmental models, Epidemic curves, mixed models

## 1 Introduction

In Italy, as all over the world, the COVID19-pandemic has had a terrible impact on people's lives, and has caused nearly 100 000 deaths by the end of February 2021. The entire scientific community is at work, and statisticians are contributing in several ways. One direction is that of formulating mathematical models that can help the surveillance of the epidemics and can be used to evaluate the effectiveness of public policies.

---

Martina Amongero, Enrico Bibbona, Gianluca Mastrantonio  
DISMA - Politecnico di Torino  
e-mail: [martina.amongero@polito.it](mailto:martina.amongero@polito.it), [enrico.bibbona@polito.it](mailto:enrico.bibbona@polito.it), [gianluca.mastrantonio@polito.it](mailto:gianluca.mastrantonio@polito.it)



Two main approaches have been proposed: mechanistic models and phenomenological ones. Mechanistic models try to interpret the data by the underlying basic mechanisms. Notable work in this direction are [4, 10, 7, 3]. On the other hand, phenomenological models focus on making the best forecast, no matter if the data generating mechanism is interpretable or realistic anyhow. Examples are [2, 5, 1].

This work is about a deterministic compartmental model (SIPRO) that falls in the first category. The main difficulty in this kind of model is that, the more it is detailed and realistic, the more difficult it becomes to fit it to the available data on epidemic curves. Indeed, most of the information is unobserved, and hardly reconstructable from the available data.

The goal of this paper is to propose an extension of the SIR model that includes a compartment of infected people that is not observed directly. When and if these people are tested they enter the positive compartment and, since quarantined, they stop spreading the disease. The motivation is that, as it was shown in [8], the asymptomatic part of the infected population gives a significant contribution to the pandemic's spread and it is often unreported. This model is applied at the level of the 20 Italian regions, with some parameter that is region-specific and some other parameters that are common to all regions. Some of the regional parameters are moreover assumed to be drawn from a common population (random effects).

Despite the model is much simpler than other proposals, the inference problem is still very challenging since the reconstruction of the latent components suffers from poor identifiability. We illustrate how this can be alleviated by adopting suitable expedients. For our analysis, we use the Italian public data collected by the Italian "Protezione Civile", which can be found at the link [6].

## 2 The SIPRO model $(\rho(t), \mu, \alpha, \nu)$

The population is divided into five compartments: Susceptibles ( $S$ ), Infected ( $I$ ), who are infectious, Positives ( $P$ ) who have been tested and quarantined (no longer source of contagion), Recovered ( $R$ ), formerly infected that recovered or died without having been tested, and, finally, Out ( $O$ ), the recovered or dead people who tested positive. The state of the system is identified by the proportion of the population in each compartment, a vector  $(i, p, r, o)$ . The proportion of susceptibles is  $1 - i - p - r - o$ . The evolution of the state variables is described by the functions  $(I(t), P(t), R(t), O(t))$  that solves the following system of ordinary differential equations (ODE)

$$\begin{cases} \dot{I}(t) = \beta(t)I(t)[1 - I(t) - P(t) - R(t) - O(t)] - \mu I(t) - \alpha I(t) \\ \dot{P}(t) = \alpha I(t) - \nu P(t) \\ \dot{R}(t) = \mu I(t) \\ \dot{O}(t) = \nu P(t). \end{cases} \quad (1)$$

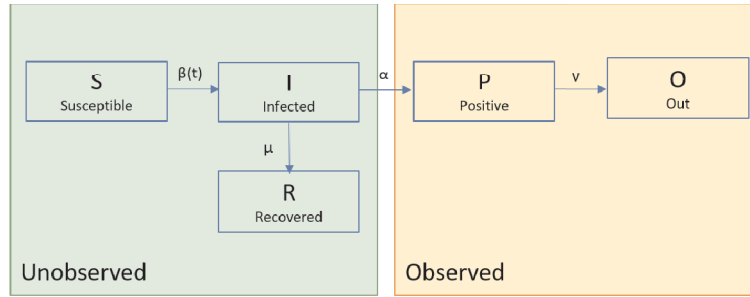


Fig. 1 SIPRO compartmental model

We can further define  $S(t) = 1 - P(t) - R(t) - O(t)$  as the proportion of susceptibles. A schematic representation is provided in Figure 1. The rate  $\beta(t)$  of infection is allowed to vary over time to model the impact of social distancing measures. The value  $1/(\mu + \alpha)$  represents the mean time an individual spends in compartment  $I$ . The quantities  $1/\mu$  and  $1/\alpha$  are respectively the time between infection and recovery and between infection and positivity. Furthermore  $1/\nu$  gives the mean time between positivity and recovery. Finally  $\rho(t) = \beta(t)/(\mu + \alpha)$  gives the reproduction number at time  $t$ , and we parameterize the model by  $(\rho(t), \mu, \alpha, \nu)$ .

### 2.1 The SIPRO model combined with a mixed approach

We build a mixed-model to describe the evolution of the Pandemic in the 20 Italian regions. We assume that within each region, the spread of the pandemic follows the SIPRO model (1), with regional reproduction numbers  $\rho_i(t)$ , testing policy  $\alpha_i$  and recovery rates from positivity  $\nu_i$ . The recovery rate  $\mu$  of the unobserved (asymptomatic) infections is common to all regions.

Let  $\{(Y_{2,i}(j), Y_{4,i}(j))\}_{j=1, \dots, T}^{i=1, \dots, 20}$  be the daily proportions of Positives and Out individuals derived from the Protezione Civile database [6]. We assume they are noisy observations from the regional SIPRO variables  $P_i(j)$  and  $O_i(j)$ , while the other variables of the model are latent. Since these variables represent proportions, a reasonable noise model is given by the Dirichlet distribution. We, therefore, assume

$$\left( Y_{2,i}(j), Y_{4,i}(j), Y_{0,i}(j) \right) \sim \text{Dirichlet} \left( \gamma P_i(j), \gamma O_i(j), \gamma (1 - P_i(j) - O_i(j)) \right)$$

where for all  $i = 1, \dots, 20$  and  $j = 1, \dots, T$ ,  $Y_{0,i}(j) = 1 - Y_{2,i}(j) - Y_{4,i}(j)$ . The parameter  $\gamma$  regulates the noise amplitude (in all regions). We estimate directly the regional reproduction number  $\rho_i(t)$  from which the infection rate is computed as  $\beta_i(t) = \rho_i(t) \cdot (\mu + \alpha_i)$ . For simplicity we assume that the function  $\rho_i(t)$  is piece-wise linear between a grid of equispaced temporal nodes  $\{\tau_k\}_{k=1, \dots, M}$  with  $\tau_1 = 0, \dots, \tau_M = T$ , whose temporal location is specified together with the priors in

the next section. Therefore it is sufficient to estimate the values  $\rho_{ik} = \rho_i(\tau_k)$  at the nodes. The logarithms of the regional SIPRO parameters  $\rho_{ik}$  and  $v_i$  are assumed to be drawn from a Gaussian population (random effects),

$$\log(v_i) \stackrel{i.i.d.}{\sim} \mathcal{N}(\log(v), \omega_v^2), \quad \log(\rho_i(k)) \stackrel{i.i.d.}{\sim} \mathcal{N}(\log(\rho_k), \omega^2)$$

since they are both affected by national and regional public policies, while the  $\alpha_i$  are considered as fixed regional effects. The parameter  $\mu$  should not be affected by any regulation, and it is therefore kept common to all regions. The initial values  $(I(0), P(0), R(0), O(0))$  of the SIPRO equations (1) need also to be estimated.

### 3 Tuning parameters and results

We focus on the data from all Italian regions from the 24th of February 2020 to the 30th of June 2020, corresponding to the so-called first wave. The parameters are estimated using a Bayesian algorithm, but since the model is poorly identifiable, any attempt to estimate the parameter solely based on uninformative priors fails. A lot of information is hidden in the unobserved components, and we need to guide the algorithm by the use of prior information. The parameters that are linked to transition between two unobserved compartments are the most difficult to estimate, namely  $\rho_i(t)$ ,  $\alpha_i$ , and  $\mu$ . We emphasize that the value of  $\rho_i(t)$  at the first and last node is particularly hard since only half of the information on it is available. This will be reflected in the amplitude of the credible intervals. Finally, the inference is more difficult for regions that had few cases.

To obtain sensible values it is necessary to

- Use informative prior on  $\alpha_i$ . In particular we impose  $\frac{1}{\alpha_i} \in [0.001, 30]$ .
- Fix different values of  $\mu \in \{\frac{1}{5k} \text{ s.t. } k = 1, \dots, 9\}$  and then choose the best values using the DIC criterion.
- Use a quite informative prior on the initial state of the system.

The prior on  $1/\alpha_i$  is set by constructing an auxiliary variable  $\varepsilon_i$  with a Gaussian prior distribution

$$\log(\varepsilon_i) \stackrel{i.i.d.}{\sim} \mathcal{N}(-2, 0.7^2),$$

and then by remapping it to the interval  $[0.001, 30]$  (in days) by the transformation

$$\frac{1}{\alpha_i} = 0.001 \cdot \mathbf{1}_{(-\infty, 0.001)} \left( \frac{1}{\varepsilon_i} \right) + \frac{1}{\varepsilon_i} \mathbf{1}_{[0.001, 30]} \left( \frac{1}{\varepsilon_i} \right) + 30 \cdot \mathbf{1}_{(30, +\infty)} \left( \frac{1}{\varepsilon_i} \right)$$

The prior for the initial conditions is specified by the vector  $d = 100 \cdot (0.96, 0.01, 0.01, 0.01, 0.01)$ , setting for all  $i$

$$(S_i(0), I_i(0), P_i(0), R_i(0), O_i(0)) \stackrel{i.i.d.}{\sim} \mathcal{D}(d)$$

The other priors are

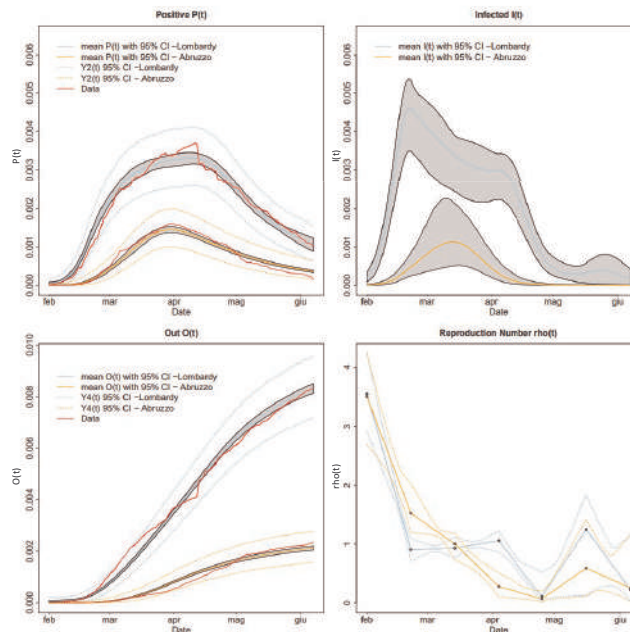
Analysing the Covid-19 pandemic in Italy with the SIPRO model

$$\begin{aligned} \log(v) &\sim \mathcal{N}(0, 100) & \text{and} & & \log(\omega) &\sim \mathcal{N}(0, 100) \\ \log(\rho_k) &\sim \mathcal{N}(0, 100), \forall k \in K & \text{and} & & \omega_v &\sim \mathcal{U}[0, 100] \\ \log(\gamma) &\sim \mathcal{N}(0, 100) \end{aligned}$$

**Table 1** SIPRO model: Global Parameters Estimated Values

Parameter	Mean	95% CI	Parameter	Mean	95%
$\rho(t_1)$	2.60	[2.12, 3.17]	$\rho(t_6)$	0.59	[0.13, 1.10]
$\rho(t_2)$	1.41	[1.15, 1.71]	$\rho(t_7)$	0.21	[0.02, 0.82]
$\rho(t_3)$	0.87	[0.70, 1.05]	$1/v$	31.17	[27.06, 35.84]
$\rho(t_4)$	0.33	[0.23, 0.45]	$1/\mu$	10	-
$\rho(t_5)$	0.06	[0.02, 0.20]	$\log \gamma$	10.04	[10.00, 10.08]

Table 1 reports the estimates of the global parameters with 95% credible intervals. As an example, we present the estimates for two different regions: Lombardy, the Italian region that counted more cases, and Abruzzo. Figure 2 graphically represents the results obtained. The functions  $\rho_i(t)$  clearly reflects the effect of social



**Fig. 2**  $P(t)$ ,  $O(t)$  and  $\rho(t)$  mean curves, with 95% CI, for Lombardy and Abruzzo. Red lines represent data. The black dots in the last panes represents the grid of time nodes of the linear interpolation. They are set every three weeks on Mondays

distancing measures. As expected, the function decreased after the lockdown. The estimates of  $\frac{1}{\alpha_i}$  for Lombardy and Abruzzo, are  $\frac{1}{\alpha_L} \in [20.55, 30]$  with posterior mean 27.67 and  $\frac{1}{\alpha_A} \in [7.25, 30]$  with posterior mean 16.19.

## 4 Conclusion

We analyzed and modeled the data collected from February to June (the so-called first phase of the pandemic) using a simple but realistic compartmental model. We highlighted the difficulties in the inference and proposed some solutions. Finally, we presented our estimates that are in agreement with what has been reported in other works. Future work will be needed to compare our model and our results to the ones obtained using other approaches, and to apply the SIPRO-mixed-model to the data collected in September-January, the so-called second phase, to evaluate the effectiveness of the new containment measures that are activated on a regional basis.

**Acknowledgements** The authors thankfully acknowledge funding from the Italian Ministry of Education, University and Research, MIUR, grant Dipartimento di Eccellenza 2018–2022 (E11G18000350001) that also provided computational resources, together with HPC@POLITO (<http://www.hpc.polito.it>).

## References

1. Agosti, A. and Giudici, P. A Poisson Autoregressive Model to Understand COVID-19 Contagion Dynamics. *Risks*, 8 (3), 2020
2. Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., Lovison, G., An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. *Biometrical Journal*, 2020
3. Gatto, M., Bertuzzo, E., Mari, L., Miccoli, S., Carraro, L., Casagrandi, R., and Rinaldo, A. (2020). Spread and dynamics of the covid-19 epidemic in italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences* **117**, 10484–10491.
4. Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., and Colaneri, M. (2020). Modelling the covid-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine* pages 1–6.
5. Griette, Q. and Demongeot, J. and Magal, P. A robust phenomenological approach to investigate COVID-19 data for France. *medRxiv*, <https://doi.org/10.1101/2021.02.10.21251500>
6. <https://github.com/pcm-dpc/COVID-19>.
7. Jia, W., Han, K., Song, Y., Cao, W., Wang, S., Yang, S., Wang, J., Kou, F., Tai, P., Li, J., et al. (2020). Extended SIR prediction of the epidemics trend of COVID-19 in italy and compared with Hunan, China. *Frontiers in Medicine*, 7
8. Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C., Rossi, L., Manganelli, R., Loregian, A., Navarin, N., et al. (2020). Suppression of a SARS-Cov-2 outbreak in the Italian municipality of Vo'. *Nature* 584, 425–429 .
9. Prague, M., Wittkop, L., Clairon, Q., Dutartre, D., Thiébaud, R., and Hejblum, B. P. (2020). Population modeling of early COVID-19 epidemic dynamics in French regions and estimation of the lockdown impact on infection rate. *medRxiv* .
10. Russo, L., Anastassopoulou, C., Tsakris, A., Bifulco, G., Campana, E., Toraldo, G., and Siettos, C. (2020). Tracing day-zero and forecasting the fade out of the COVID-19 outbreak in Lombardy, Italy: a compartmental modelling and numerical optimization approach. *Plos one*, 2020.

# **Intentions of union formation and dissolution during the COVID-19 pandemic. The role of risk aversion**

## ***Intenzioni di formazione e di dissoluzione di un'unione durante la pandemia da COVID-19. Il ruolo dell'avversione al rischio***

Bruno Arpino and Daniela Bellani

**Abstract** Socio-demographic research recently addressed the role of risk preferences in demographic behaviors. No study has analyzed the relationship between risk aversion and intentions of forming a new couple and of divorcing during the COVID-19 pandemic. We examine this relationship using data collected in April 2020 with an online survey (Intergen-Covid) in three countries, Italy, France and Spain. Results of logistic regressions indicate that risk-oriented individuals are relatively more likely to intend to make these transitions compared to their risk averse counterparts.

**Abstract** *La ricerca socio-demografica ha recentemente sottolineato l'importanza del ruolo delle preferenze per il rischio per spiegare i comportamenti demografici. Nessuno studio finora ha però analizzato la relazione tra tali preferenze e intenzioni relative alla formazione di una nuova unione e alla separazione durante la crisi pandemica COVID-19. In questo studio utilizziamo i dati raccolti ad Aprile 2020 dall'indagine online "Intergen-Covid" per esaminare se l'avversione al rischio è associata alle intenzioni menzionate in tre Paesi, l'Italia, la Francia e la Spagna. I risultati delle regressioni logistiche indicano che all'aumentare della propensione al rischio aumenta la probabilità di avere intenzione di compiere ognuna delle transizioni demografiche considerate.*

**Key words:** fertility intentions, COVID-19, risk aversion

---

<sup>1</sup> Bruno Arpino, Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze; [bruno.arpino@unifi.it](mailto:bruno.arpino@unifi.it)  
Daniela Bellani, Dipartimento di Scienze Politiche e Sociali, Università di Bologna; [daniela.bellani@unibo.it](mailto:daniela.bellani@unibo.it)

## 1 Introduction

This paper aims at contributing to the literature on the determinants of intentions of forming and dissolving a union. In particular, we argue that among the (latent) preferences that are likely to explain heterogeneity in demographic intentions over and beyond standard socio-economic characteristics, risk preferences may play an important role.

As pointed out by previous research [1, 2], individuals may feel threatened or, on the contrary, encouraged by the stochastic dimension that characterizes the consequences of demographic choices, such as forming a new union. In fact, risk-taking propensity significantly affects the expected utility of decisions characterized by uncertain outcomes, regardless of individual's socio-economic background (e.g., [3]). Following this argument, we aim at explaining differences in intentions concerned with union formation and dissolution that stem from the individual preferences about the extent one is willing to accept risk – the so-called risk tolerance (the reciprocal of risk aversion).

In particular, we make use of timely data of the online survey Intergen-Covid collected during the first lockdown implemented to fight the COVID-19 pandemic. The data come from three of the most seriously hit countries in Europe (Italy, France and Spain) and were collected in April 2020. We examine whether risk aversion has an impact in influencing union formation and dissolution in this period of characterised by a high of uncertainty. As such, this period represents an ideal context to study the relationship between risk aversion and demographic intentions.

### 1.1 Hypotheses

The rationale behind the link between risk aversion and demographic intentions is that differences in individuals' propensities for risk-taking have consequences on intentions given that a certain level of risk tolerance embodies a specific calculation of (non-)monetary costs and rewards of certain demographic choices, e.g. union formation, union dissolution. In our study, we allow the utility of potential demographic choices to depend (also) on the level of subjective inclination for risk tolerance.

Given the strong uncertainty characterising expected gains and costs from forming or dissolving a couple, both union formation and dissolution could be viewed by risk-averse individuals as a risky option – especially in a context of strong uncertainty such as the COVID-19 pandemic. If this is true, we can expect that more risk averse individuals are more likely not to intend getting married/cohabitating and to divorce/separate than their less risk averse counterparts. This leads to formulating the following hypothesis:

(H1A:) More risk averse individuals are less likely to have the intention to get married/start cohabiting or divorce/separate.

### **Intention of union formation and dissolution during the pandemic in Italy**

However, as suggested by several scholars, the family, as other forms of associations, may provide members with mutual insurance [4]. People may invest in the family as a way of being protected from negative events (health, income, etc.) and as a source of help and wellbeing. If marriage/cohabitation is perceived as a source of security, risk averse individuals may intend to "buy" this security by entering a partnership in order to reduce future risks. This leads to the partially competing hypothesis:

(H1B:) More risk averse individuals are more likely to have the intentions to getting married/cohabiting and less likely to have the intention to divorce/separate.

## **2 Data and empirical strategy**

Our analyses rely on the online survey Intergen-Covid [5]. Respondents were interviewed in April 2020 during the first national lockdowns. The questionnaire has been administered via CAWI (Computer Assisted Web Interviewing). The total sample size was 9,186 individuals, with approximately 3,000 respondents per country (Italy, France and Spain).

The questionnaire explores the core respondents' experiences during home confinement as well as their family-related prospects in the future. In particular, it includes questions devoted at detecting individuals' intentions to form/dissolve a union, as well individual's risk preferences. The survey company Lucid has collected the data, with the imposition of representative quotas at country level by gender, age, region of residence and educational attainment.

### **2.1 Variables**

The explanatory variable we consider is an indicator of the individual risk aversion, capturing respondents' willingness to take risks. More specifically, the following question was asked 'How do you see yourself: are you a person who is generally willing to take risks, or do you try to avoid taking risks? Please use a scale from 0 to 10, where a 0 means you are "completely unwilling to take risks" and a 10 means you are "very willing to take risks". You can also use the values in-between to indicate where you fall on the scale.' In the analyses, we consider this variable as continuous.

The dependent variables refer to the intentions to make certain demographic choices within the next three years from the time of the interview. The first dependent variable is the intention to getting married or cohabiting. In this case, we select male and female respondents, aged between 18 and 50, not married and not cohabitating at the time of interview. The working sample is  $N = 2077$ .

The second outcome variable is the intention to divorce/separate. In this case we select male and female respondents, aged between 18 and 50, married or cohabitating. The working sample in this case is  $N = 2513$ .



Bruno Arpino and Daniela Bellani

In all the models we control for several variables: gender; age and its square; level of education (below secondary education ‘low’, up to high-school ‘medium’, and tertiary education ‘high’, according to the ISCED classification); respondent’s employment status (if employed or not); perceived economic situation (whether the respondent reports that he/she finds difficult to cope with present income); whether the respondent have (or not) dealt with income and/or job loss during the pandemic; whether the respondent has at least one child younger than 15. We control also for time discounting preferences. This is a measure of the extent the individual is forward looking. Respondents have to answer to the following question. ‘Are you a person who is generally willing to give up something today in order to benefit from that in the future or are you not willing to do so? Please use a scale from 0 to 10, where a 0 means you are “completely unwilling to give up something today” and a 10 means you are “very willing to give up something today”. You can also use the values in-between to indicate where you fall on the scale.’ We add fixed effects by country and a measure of the regional prevalence of COVID-19 cases cumulated at the time of the interview. In the model where the dependent variable is the intention to divorce/separate, we also control for the type of partnership, i.e., whether a marriage or a cohabitation.

## **2.2 *Method and statistical analysis***

We present some descriptive findings considering the respondents’ preferences as interpretative lens. We use country-specific weights to offer nationally representative estimates. In our sample, about 30% of the respondents has the intention to form a union, 2.3% has the intention to divorce/separate. Respondents are generally risk neutral (the average of the variable risk aversion is 4.43, linearized standard error is .14).

Our multivariate analyses rely on logistic regression to model the probability of intending (versus not) to form a union or to dissolve a union (in both cases a marriage or a cohabitation).

## **3 Empirical results**

First, we report a table with the main findings from the logistic regressions. M1 refers to the intention of forming a new partnership, M2 of separating/divorcing. Second, in order to make the interpretation of the results clearer, we show predicted probabilities graphically.

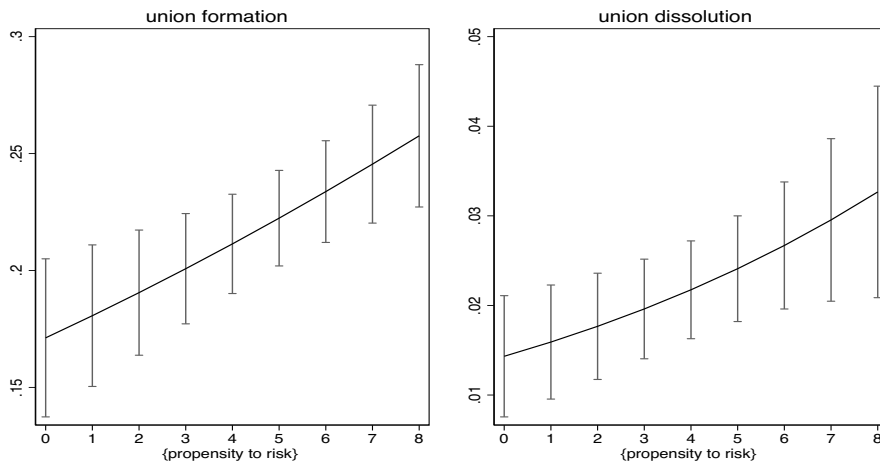
**Intention of union formation and dissolution during the pandemic in Italy**

**Table 1:** Estimates for risk aversion, log-odds (standard errors in paratheses)

	<i>M1</i>	<i>M2</i>
Risk aversion	.070** (.030)	.110* (.066)
N	2077	2513

*Note.* Controls: gender; age and its square; level of education; respondent's employment status; perceived economic situation; whether the respondent has (or not) dealt with income and/or job loss during the pandemic; whether the respondent has at least one child younger than 15; time discounting preferences, country, regional cumulated prevalence of COVID-19 cases. Model M2 additional controls for whether the respondent is married or cohabiting. Standard error in parenthesis. \*  $p < .1$ , \*\*  $p < .05$ , \*\*\*  $p < .01$

Results reported in Table 1 indicate that, during the pandemic, individuals who are more tolerant to risk are also more likely to intend forming a union and separating/divorcing. This result holds controlling for socio-economic and other standard determinants of demographic intentions. In Figure 2 we report the predicted probabilities for each intention we consider. Noticed that in each graph the y-axis has different scale reflecting the very different overall percentage of intention related to the two outcomes. We can observe that more risk tolerant individuals have a probability of about 9 percentage points higher to intend forming a union compared to those more risk-averse. Considering the intention of separating this gap is about 2 percentage points. These associations are rather strong from a substantive point of view. In fact, the predicted probability of intending to form a union is about 50% higher for the most risk-tolerant as compared to the most risk-averse. Instead, the predicted probability of intending to dissolve a union for the most risk-tolerant is about double that of the most risk-averse.



**Figure 1:** Predicted probabilities of intending to form a union (right-hand side panel) or to dissolve a union (left-hand side panel)

## 4 Conclusions

We found that the higher is the risk propensity, the higher is the probability of intending to form or to dissolve a union during the COVID-19 pandemic. Given the strong uncertainty characterising expected gains and costs of union formation and dissolution, especially in the uncertain context of the pandemic, marriage/cohabitation as well as divorce/separation could be viewed by risk-averse individuals as risky choices and they may tend to prefer maintaining the status quo.

Our findings also serve to point to the importance of considering risk aversion in demographic research to gain a more complete understanding on intentions. Adjusting for risk aversion may be also important when studying other determinants of intentions of union formation/dissolution.

## References

1. Schmidt, L. (2008). Risk preferences and the timing of marriage and childbearing. *Demography*, 45(2), 439-460.
2. Bellani, D. & Arpino, B. (2021) Risk aversion and fertility. Evidence from a lottery question in Italy (No. 2021\_02). Università degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti".
3. Breen, R., & Goldthorpe, J. H. (1997). Explaining educational differentials: Towards a formal rational action theory. *Rationality and society*, 9(3), 275-305.
4. Appelbaum, E., & Katz, E. (1991). The demand for children in the absence of capital and risk markets: A portfolio approach. *Oxford Economic Papers*, 43(2), 292-304.
5. Arpino, B., Pasqualini, M., Bordone, V., & Solé-Auró, A. (2020). Indirect consequences of COVID-19 on people's lives. Findings from an on-line survey in France, Italy and Spain. <https://doi.org/10.31235/osf.io/4sfv9>.

## 4.37 Time series methods

# Bootstrap-based score test for INAR effect

## *Un metodo bootstrap per verificare la presenza di effetti INAR*

Riccardo Ievoli and Lucio Palazzo

**Abstract** This work exploits the potential of bootstrap methods in testing for serial dependence in the class of observed driven Integer-AutoRegressive (INAR) models with Poisson innovations. The main contribution is to develop a novel restricted bootstrap algorithm to improve the performance of score-based test statistic, especially in case of moderately small series.

**Abstract** *In questo lavoro si indagano le potenzialità del bootstrap per verificare la presenza di correlazione seriale in serie storiche a valori discreti con innovazioni Poisson. Il contributo innovativo riguarda l'implementazione di un metodo bootstrap allo scopo di migliorare le performance delle statistiche test basate sulla funzione score, specialmente in serie che presentano una numerosità limitata.*

**Key words:** discrete time series, score test, semiparametric bootstrap, parametric bootstrap

## 1 Introduction

INteger AutoRegressive models (INAR) became popular to model non-negative integer time series, especially under the assumption of Poisson innovations (P-INAR). Therefore, a relevant part of the literature has been focused on testing for the presence of a (possibly unknown) serial dependence. This task may be not trivial for discrete time series, where conventional methods for continuous data may fail, especially in case of low counts. In principle, [3] proposed a test statistic based on the

---

Riccardo Ievoli

University of Ferrara, Department of Economics and Management, Via Voltapaletto 11, Ferrara, e-mail: riccardo.ievoli@unife.it

Lucio Palazzo

University of Naples Federico II, Department of Political Science, Via Rodinò 8, Naples, e-mail: lucio.palazzo@unina.it

score function for the P-INAR(1). Then, [5] compared the performance of score-based statistics with other proposals (e.g., the runs test and Portmentau-type statistics), while [10] extended this approach to a more generalized class of INAR. Under the null hypothesis of non-serial dependency, these statistics can be generally approximated by a standard normal distribution in large samples.

A possible drawback can be found in the poor performance of the score test due to asymptotic approximation issue: simulations in [5] and [7] shown that the test may be severely undersized in moderately small samples (e.g.,  $T = 50$ ), but also undersized in larger samples (e.g.,  $T = 500$ ).

Starting from this issue, we propose a novel bootstrap algorithm to improve the performance of score-based test for the P-INAR(1). Bootstrap methods in INAR are recently developed by [4] to obtain more reliable inference in point estimation and confidence bounds. To the best of our knowledge, this is the first work that exploits the potential of bootstrap in testing for the presence of INAR effect, especially in a (moderately) small samples perspective.

The paper is organized as follows: Section 2 presents the P-INAR(1) model and the score-based test statistic based on Poisson assumption. Section 3 introduces the novel bootstrap algorithm while results of a small-scale Monte Carlo simulation are shown in Section 4. Finally, Section 5 concludes with some remarks and possible advances.

## 2 The model and score test statistic

We focus on testing serial dependence in one-lagged and stable INAR model [1]:

$$X_t = \alpha \odot X_{t-1} + \varepsilon_t \quad (1)$$

where  $\alpha \in (0, 1)$  is the thinning parameter of interest. In equation (1) the binomial thinning operator  $\odot$  is defined as a random sum of i.i.d. random variables  $\{Y_i\}$ , with  $Y_i \sim \text{Ber}(\alpha)$ , independent of  $X_t$ , such that  $E(Y_i) = \alpha$  and  $\text{Var}(Y_i) = \alpha(1 - \alpha)$ .

The DGP of the marginal process varies according to the distribution of the innovations  $\{\varepsilon_t\}$ . Here we consider i.i.d.  $\varepsilon_t \sim \text{Po}(\lambda)$ , where  $E(\varepsilon_t) = \lambda$ , i.e., the so-called P-INAR(1) model. We consider the following system of hypothesis:

$$H_0 : \alpha = 0 \quad \text{vs} \quad H_1 : \alpha > 0.$$

Under this system of hypothesis, the score statistic for testing P-INAR(1) model, introduced in [5], takes the following specification:

$$\hat{S} = S^P(\hat{\lambda}) = T^{-1/2} \frac{\sum_{t=1}^T (x_{t-1} - \hat{\lambda})(x_t - \hat{\lambda})}{\hat{\lambda}} \quad (2)$$

where  $\hat{\lambda} = T^{-1} \sum_{i=1}^T x_t$ .

### 3 Bootstrap algorithm for testing INAR

In this Section, a restricted bootstrap method to obtain bootstrap p-value for the test statistic in equation (2) is developed. Non-parametric approaches, such as block bootstrap [9], or autoregressive resampling scheme [6] are not considered since they do not take into account the truly features of the data (positiveness and integer values), leading to inconsistent results. In addition, [4] have shown the infeasibility of those methods in point estimation, as also confirmed in a Monte Carlo study for hypothesis testing [8].

We consider a semiparametric bootstrap for its suitability and compare its results with a parametric bootstrap, where the bootstrap DGP is based on Poisson assumption. We employ a “restricted” procedure, imposing the null hypothesis  $\alpha = 0$  in the bootstrap DGP for the score-based test statistic, such that  $\hat{\varepsilon}_t = x_t$ . The restricted method ensures that residuals will belong to the support of innovations’ DGP.

As mentioned, pseudo residuals can be obtained by using either parametric or semiparametric method. In the parametric case, the  $\{\varepsilon_t^*\}_{t=1}^T$  are sampled from a specific probability law, i.e., bootstrap residuals are sampled from a Poisson distribution with parameter equal to the estimate of  $\lambda$ . To summarize:  $\varepsilon_t^* \sim Po(\hat{\lambda})$ . Conversely, in the semiparametric method, the pseudo residuals are sampled from the Empirical Distribution Function (EDF) of the restricted residuals  $\hat{\varepsilon}_t$ , i.e.,  $\varepsilon_t^* \sim EDF(\hat{\varepsilon}_t)$ .

Explanation of proposed method can be summarized through the following algorithm.

**Algorithm 1** (Restricted Bootstrap).

Given a random sample  $x_1, \dots, x_T$  of size  $T$

- Step 1. Estimate the parameters  $(\hat{\alpha}, \hat{\lambda})$  and the test statistic  $\hat{S}$ . Residuals can be obtained imposing  $\alpha = 0$ , i.e.,  $\hat{\varepsilon}_t = x_t$
- Step 2. Use  $\hat{\varepsilon}_t$  to obtain bootstrap pseudo residuals  $\varepsilon_1^*, \dots, \varepsilon_T^*$
- Step 3. Create  $x_1^*, \dots, x_T^*$ , plugging pseudo residuals in the bootstrap DGP.
- Step 4. Compute the bootstrapped score statistic

$$\hat{S}^* = S^*(\hat{\lambda}^*) = T^{-1/2} \frac{\sum_{t=1}^T (x_{t-1}^* - \hat{\lambda}^*)(x_t^* - \hat{\lambda}^*)}{\hat{\lambda}^*} \tag{3}$$

- Step 5. Repeat  $B$  times steps 1–4 producing  $\hat{S}_1^*, \dots, \hat{S}_B^*$ .
- Step 6. Compute the bootstrap p-value  $p^* = B^{-1} \sum_b I(|\hat{S}_b^*| > |\hat{S}|)$

### 4 Simulation study

To analyze the finite sample behavior of score bootstrap-based tests, illustrated in Section 3, we generate  $M = 100,000$  samples through the following DGP:

$$x_t = \alpha \odot x_{t-1} + \varepsilon_t,$$

where  $\varepsilon_t \sim Po(\lambda)$  considering in the parameter setting  $\lambda = \{0.5, 2, 10\}$ . Different length of the series are considered i.e.,  $T = \{50, 75, 100, 150, 250, 500\}$ . The selected nominal level for the test is equal to 0.05. Empirical size of bootstrapped statistic  $\hat{S}^*$  is reported by fixing  $\alpha = 0$ , while we choose an increasing sequence of  $\alpha$  by 0.05 (starting from  $\alpha = 0$ ) to evaluate the empirical power, stopping at  $\alpha = 0.8$  to avoid issues arising in the near-unit root situation [2]. To obtain the bootstrap p-values, the number of bootstrap replications is set equal to  $B = 999$ . As a reference, we also compute asymptotic rejection frequencies for the score-based statistic.

Regarding the empirical size, the asymptotic rejection frequencies can be far from the nominal level especially with series of moderately small length, i.e.,  $T \leq 150$  and  $\lambda = 2, 10$ . Indeed, the distribution of rejection frequencies obtained through bootstrap-based score statistic shows the successful of proposed methods. Thus, while the excellent performance of parametric bootstrap is certainly due to the imposition of the true DGP in the simulation setup, mixed with the usage of the appropriate score statistic, the semiparametric method is also quite close to the nominal level even with series of small length (i.e.,  $T \leq 100$ ).

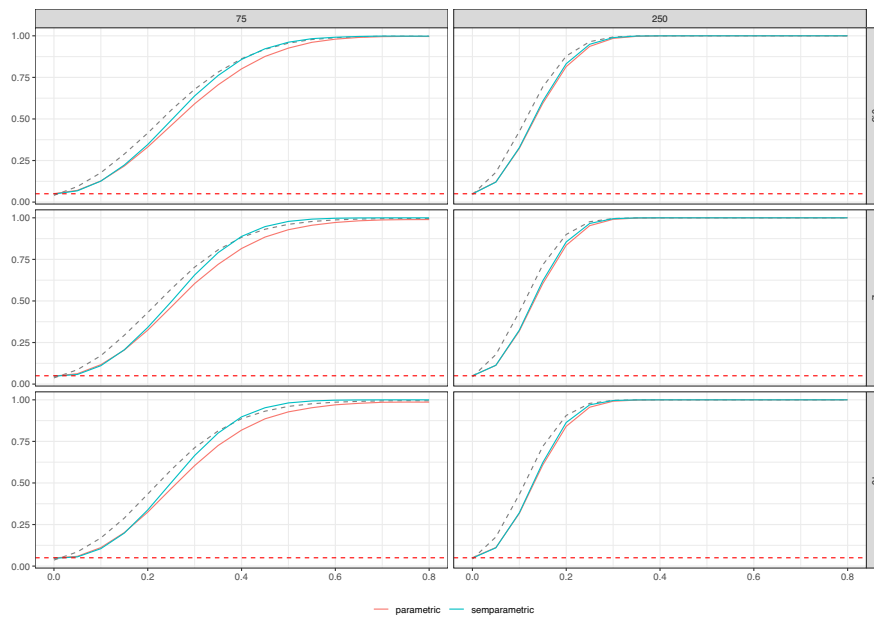
**Table 1** Empirical size of asymptotic and bootstrap-based score test for INAR effect.

$\lambda$	$T$	Asymptotic	Parametric Bootstrap	Semiparametric Bootstrap
0.5	50	0.0356	0.0502	0.0491
	75	0.0406	0.0509	0.0484
	100	0.0419	0.0507	0.0487
	150	0.0429	0.0493	0.0484
	250	0.0455	0.0506	0.0489
	500	0.0465	0.0496	0.0494
2	50	0.0331	0.0491	0.0469
	75	0.0365	0.0480	0.0475
	100	0.0389	0.0504	0.0486
	150	0.0406	0.0493	0.0492
	250	0.0434	0.0498	0.0496
	500	0.0454	0.0498	0.0498
10	50	0.0328	0.0499	0.0460
	75	0.0369	0.0492	0.0468
	100	0.0391	0.0506	0.0472
	150	0.0412	0.0503	0.0493
	250	0.0426	0.0505	0.0481
	500	0.0451	0.0491	0.0499

Figure 1 shows the performance in terms of empirical power of proposed tests, considering six different scenarios. The overall performance of bootstrap-based tests are comparable with respect to the asymptotic one. Although the two tests seem little more conservative, especially with  $\alpha \leq 0.4$  when  $T = 75$  and  $\alpha \leq 0.2$  when  $T = 250$ , the semiparametric method outperforms both the asymptotic and the parametric bootstrap in case of moderately small series ( $T = 75$ ) for a reasonably large  $\alpha$  (i.e.,



### Bootstrap-based score test for INAR effect



**Fig. 1** Empirical power plot. Dark dashed line is the asymptotic empirical power, while red dashed line is the nominal level (0.05). Two sample sizes (in columns) and three values of  $\lambda$  (in rows) are considered.

$\alpha \geq 0.4$ ). Therefore, the considered tests not appear to be particularly sensitive with respect to the amount of  $\lambda$  parameter.

## 5 Concluding remarks

The score-based statistic is a reasonable way to test for the presence of serial dependence in discrete time series. The two proposed bootstrap methods help to improve the performance of score-based statistic in case of P-INAR model in terms of empirical size, especially considering series of moderately small length, also showing a comparable (and more reliable) performance in terms of empirical power.

While the excellent results of parametric bootstrap are due to the specific features of the simulation setup, the satisfying performance of semiparametric method suggests its usefulness, especially in a more generalized context (e.g., under several possible distributions for the innovations). We also expect that the parametric bootstrap, strongly relying on Poisson assumption, may fail in case of (possibly mild) deviations from this specific probability law.

Further research will extend the proposed bootstrap algorithm to more generalized versions of the score statistic [10], even considering possible sources of mis-

specifications (e.g., overdispersion and zero inflation). Applicability of score-based bootstrap tests should be also investigated through the analysis of real integer-valued time series in many fields, such as: finance, healthcare, and environment.

## References

1. Al-Osh, M. A., Alzaid, A. A.: First-order integer-valued autoregressive (INAR (1)) process. *J. Time Ser. Anal.*, **8**(3), 261–275 (1987) doi: 10.1111/j.1467-9892.1987.tb00438.x
2. Drost, F. C., van den Akker, R., Werker, B. J. M: An Asymptotic Analysis of Nearly Unstable inar (1) Models. CentER Discussion Paper, Vol. 2006-44, *Econometrics* (2006) doi: 10.2139/ssrn.905484
3. Freeland R.K.: Statistical analysis of discrete time series with applications to the analysis of workers compensation claims data [unpublished doctoral dissertation]. Vancouver (Canada), University of British Columbia (1998)
4. Jentsch, C., Weiß, C. H.: Bootstrapping INAR models. *Bernoulli*. **25**(3), 2359–2408 (2019) doi: 10.3150/18-BEJ1057
5. Jung, R. C., Tremayne, A. R.: Testing for serial dependence in time series models of counts. *J. Time Ser. Anal.*, **24**(1), 65–84 (2003) doi: 10.1111/1467-9892.00293
6. Kreiss, J. P., Paparoditis, E., Politis, D. N.: On the range of validity of the autoregressive sieve bootstrap. *Ann. Stat.*, **39**(4), 2103–2130 (2011) doi: 10.1214/11-AOS900
7. Larsson, R.: Testing for INAR effects. *Commun. Stat. - Simul. Comput.*, **49**(10), 2745–2764 (2020) doi: 10.1080/03610918.2018.1530784
8. Palazzo, L., Ievoli, R.: Bootstrap test in Poisson-INAR models. *Book of Short Papers SIS2020*, 1351–1356. Pearson (2020) ISBN: 978-88-919-1077-6
9. Politis, D. N., Romano, J. P.: A circular block-resampling procedure for stationary data. In: Lepage, R. and Billard, L., Eds., *Exploring the Limits of Bootstrap*, Wiley, New York, 263–270 (1992)
10. Sun, J., McCabe, B. P.: Score statistics for testing serial dependence in count data. *J. Time Ser. Anal.*, **34**(3), 315–329 (2013) doi: 10.1111/jtsa.12014

# Evaluating the performance of a new picking algorithm based on the variance piecewise constant models

## *Valutazione dell'accuratezza di un nuovo algoritmo di picking basato sui modelli con varianza costante a tratti*

Nicoletta D'Angelo, Giada Adelfio, Antonino D'Alessandro and Marcello Chiodi

**Abstract** In this paper, a new picking algorithm for the automatic seismogram onset time determination is tested on a dataset of simulated waveforms. We aim at capturing the variations in the performance due to some characteristics of both the seismic event and its detection, which in turn affect some characteristics of the waveforms. We therefore simulated seismic events with different magnitude, assumed to be detected with different distances from the nearest seismic station. Our tests permit to highlight the scenarios most suitable for our algorithm.

**Abstract** *In questo articolo valutiamo la performance di un nuovo algoritmo di picking su un dataset di forme d'onda simulate. L'obiettivo è identificare differenze della performance dell'algoritmo in base a caratteristiche differenti sia dell'evento sismico che della sua rivelazione. Abbiamo dunque simulato eventi sismici con diversa magnitudo, e ipotizzando che questi siano rilevati a diverse distanze dalle stazioni sismiche. I nostri test permettono di evidenziare gli scenari più adatti per il nostro algoritmo.*

**Key words:** Earthquake Early Warning, Picking, Change-points, Earthquake

## 1 Introduction

An earthquake early warning (EEW) system, is a system of accelerometers, seismometers, communication, computers, and alarms that is devised for notifying adjoining regions of a substantial earthquake while it is in progress. The implemen-

---

Nicoletta D'Angelo, Giada Adelfio and Marcello Chiodi  
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy,  
e-mail: nicoletta.dangelo@unipa.it; giada.adelfio@unipa.it; marcello.chiodi@unipa.it

Antonino D'Alessandro  
Istituto Nazionale di Vulcanologia e Geofisica, Rome, Italy,  
e-mail: antonino.dalessandro@ingv.it

tation of efficient and computationally simple picking algorithms is necessary to produce decisive event warnings, as well as automatic picking of seismic phases for seismic surveillance and routine earthquake location for fast hypocenter determination. To be suitable for both applications, a false alarm must be avoided and time picking must be as accurate as possible.

First arrival times on seismograms coincides with the arrival of the first P-wave. The time of the phase-detection  $\hat{T}_i$  at a station  $i$  is interpreted as the first P-phase arrival time, which is, of course, affected with an error  $\varepsilon_i$ .  $\hat{T}_i$  may be written as  $\hat{T}_i = T_0 + t_i + \varepsilon_i$ , where  $T_0$  is the source time and  $t_i$  is the travel time of a P-wave to station  $i$ . The coincidence trigger detects an event, if for any combination of a minimum number of stations (typically three or four) the condition  $|\hat{T}_i - \hat{T}_j| \leq \varepsilon$  is met.  $\varepsilon$  is the maximum allowed difference between trigger times at neighbouring stations. This coincidence trigger works satisfying for local networks, where the number of stations and the aperture of the network is not large. For regional and global networks this simple event detection algorithm has to be modified. For a complete review of the most widespread automatic picking algorithms and their properties, we refer to Küperkoch et al. (2012).

In D'Angelo et al. (2020), a new automatic picking algorithm suitable for the implementation of EEW and in seismic surveillance, based on changes in variance, is proposed and tested on a set of 100 synthetic seismograms, showing that the model is always able to correctly detect the arrival of the first P-wave, as well as other relevant phases of the seismic event, such as the arrival of the first S-wave and the end of the seismic event. The simulated waveforms in D'Angelo et al. (2020) all presented the same true values of arrival times but different underlying noise.

The aim of this paper is therefore to test the performance of the proposed algorithm on a set of waveforms simulated as generated by seismic events with different characteristics, such as the magnitude, and with different scenarios of detection, namely with different distances from the nearest seismic station that first recorded the event. This allows us to assess the performance of the algorithm with respect to the different characteristics of both the seismic event and the detection scenario, to identify the most suitable scenario for the application of our algorithm. Our guess is that the picking algorithm should work best when detecting seismic events with a high magnitude, and that occurred near the seismic station that first recorded the event. Therefore, the performance of the algorithm is expected to worsen as the magnitude decreases and the distance from the station increases.

The structure of the paper is as follow. Section 2 reviews the new picking algorithm. Section 3 reports the testing of the algorithm on a dataset of simulated waveforms. Finally, Section 4 contains the conclusions and future works.

## 2 The new picking algorithm based on segmented models

We refer to the new automatic picking algorithm proposed in D'Angelo et al. (2020). For this new proposal, the authors considered the model introduced in Adelfio

(2012), suitable for the case of changepoint detection for changes in variation, and applied it to the analysis of the seismic waveforms. The considered approach can be seen as a wider version of the *cumSeg* models proposed in Muggeo and Adelfio (2011). Let  $y_i$  be the outcome and  $x_i$  be the observed sample with  $i = 1, 2, \dots, n$ , that denotes time in this context. We refer to the framework in which  $y_i = \mu_i + \varepsilon_i$ , where  $\mu_i$  is a function representing the observed signal,  $\varepsilon_i \sim N(0, \sigma_i^2)$  is an error term, and  $\sigma_i^2$  is a variance function approximated by a piecewise constant regression function with  $K_0 + 1$  segments. The estimation of the mean signal  $\hat{\mu}$  can be carried out by using a smoothing procedure, e.g., fitting a cubic smoothing spline to the data. Following Smyth et al. (2001), a gamma generalized linear model (GLM) is fitted with a log-link function, with response given by the squared studentized residuals  $s_i = (y_i - \hat{y}_i)^2 / w_i$ , (with  $\hat{y} = \hat{\mu}$  and weights  $w_i = 1 - h_i$ , where  $h_i$  is the  $i$ th diagonal element of the hat matrix  $H$ ). According to this approach, it means that we are looking for a change in the mean of the squared residuals from a fitted linear model, rather than directly looking for a change in the variance of the signal. Therefore, the following model is specified  $g(\theta_i) = \beta_1 x_i + \delta_1 (x_i - \psi_1)_+ + \dots + \delta_{K_0} (x_i - \psi_{K_0})_+$  where  $\theta_i = E[\sum_j^i s_j]$ . The term  $(x_i - \psi_k)_+$  is defined as  $\sum_j^i I(x_j > \psi_k) = (x_i - \psi_k) I(x_i > \psi_k)$ , where  $I(\cdot)$  is the indicator function. The vector parameter  $\psi$  represents the  $K_0$  locations of the changes, while  $\beta_1$  is the mean level for  $x_i < \psi_1$  and  $\delta$  is the vector of the differences in the mean levels.

This model typically estimates  $K^* \geq 2$  changepoints. Since in the context of the picking of the seismic phases we are usually interested in detecting specific phases, such as the arrival times of the first P- and S-waves, as well as the end of the seismic event, we need an algorithm for keeping the relevant changepoints, denoted by  $\hat{K}$ , among the  $K^*$  estimated ones. At this aim, in D'Angelo et al. (2020), the authors propose a further post-selection algorithm that compares the ratio between the variances of subsequent phases, individuated by the application of the model to the signal, selecting only the  $\hat{K}$  biggest ones. Moreover, the authors also report the results of the application of this model on a set of 100 simulated waveforms over 60 seconds (with sampling step 0.004 s). By analysing the estimated arrival times of the first P- and S-waves, as well as the end of the seismic event (i.e.  $\hat{K}=3$ ), the authors conclude that the algorithm succeeds most of the times in correctly picking the arrival time of the first P-wave. Therefore, to test the performance of the algorithm with respect to different characteristics of both the event and its detection, a further dataset of simulated waveforms is simulated and analysed in the next Section.

### 3 Simulations

This section is devoted to the analysis of the simulated dataset of waveforms. The waveforms are simulated as coming from seismic events with different characteristics, namely different levels of magnitude, and as coming from different detection scenarios, namely reporting different distances from the nearest seismic station. The dataset consists of 1000 waveforms of 300 seconds each (with sampling step 0.004

s). On the basis of their characteristics, we define 16 different scenarios of waveforms, given by the combination of the categories obtained from the quantiles of the univariate distributions of the two variables:

- 4 ranges of distances from the nearest station that recorded the seismic event: (0-62], (62-125], (125-187] and (187-250] km;
- 4 levels of magnitude: (2-2.75], (2.75-3.5], (3.5-4.25] and (4.25-5].

Consequently, each scenario contains a different number of waveforms each, ranging from 50 to 84. For each seismic event we generate all three components of the waveform, i.e North-South, East-West and Vertical. In our analysis, we only report the results for the Vertical component for the sake of brevity. Also, we define the relevant changepoints to be identified as the true arrival times of the first P- and S-waves, denoted by  $\psi_1$  and  $\psi_2$ , respectively (i.e.  $\hat{K} = 2$ ). Finally, we only retain the first third of observed seconds, as the analysis of the whole waveform would not contain any other relevant information and only make the computation heavier. In Figure 1, a single waveform for each of the considered scenarios, and its corresponding true arrival times, are shown.

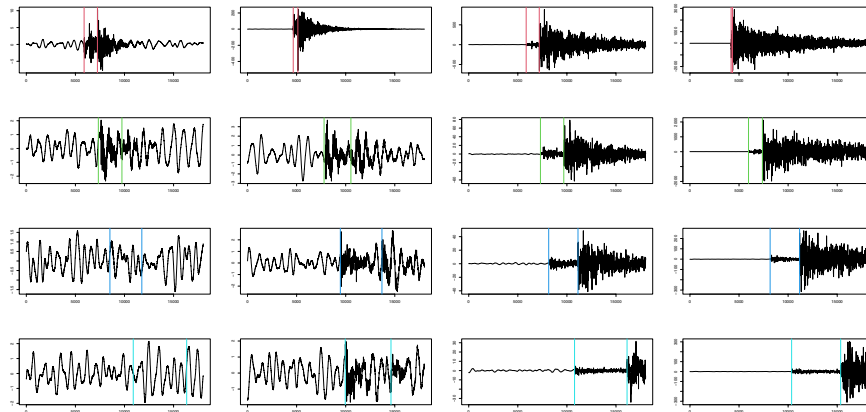
Table 1 reports the empirical means (m) and Mean Squared Error values (MSE) of the two relevant changepoints over the Vertical component of each simulated scenarios. Percentages of NAs are reported alongside, denoting when no arrival times are detected in the given waveform. The left columns of the table report the results concerning the first steps of the algorithm, estimating a  $K^* \geq 2$  number of changepoints, while the right columns of the table report the results with the addition of a post-selection algorithm, that further selects only the  $\hat{K} = 2$  relevant changepoints, among the  $K^*$  estimated ones. The post-selection algorithm is proposed in D'Angelo et al. (2020), and it considers the ratio between the variances of the subsequent phases identified by the  $K^*$  changepoints estimated by the main algorithm. The relevant changepoints  $\hat{K}$  are selected as the ones in correspondence to the biggest variance ratios. Of course, the post-selection adds uncertainty to the identification of the changepoints obtained through the application of the algorithm. For this reason the MSE values obtained before the post-selection are always larger, therefore it is important to look at both results.

Overall we may notice that, as expected, the algorithm performs the best as the distance from the nearest seismic station that recorded the event decreases, and as the magnitude of the seismic event increases. Indeed, NAs are most likely to occur when the magnitude is small and the distance is large, that is basically when the P- and S-waves are indiscernible from the background noise. In such cases the arrival times cannot be estimated. Furthermore, by comparing the results of the main algorithm with the results in which also the post-selection is applied, it appears evident that the performance of the post-selection depends on the scenario analysed, and therefore on the specific shape of the seismic waveform. The scenarios in which the seismic event is generated with a magnitude belonging to the range (3.5 – 4.25] is always the one that reports the lower number of NAs, regardless of the assumed distance from the seismic station. Nevertheless, this does not represent the best picking scenario as the post-selection algorithm worsens the final performance.

Evaluating the performance of a new picking algorithm

**Table 1** Empirical means (m) and Mean Squared Error values (MSE) of P- and S-waves arrival times over the Vertical components of each simulated scenario. Percentages of NAs are alongside.

M		Main Algorithm				+ Post-selection				NA (%)	
		$\psi_1$	$\psi_2$	$\psi_1$	$\psi_2$	$\psi_1$	$\psi_2$	$\psi_1$	$\psi_2$		
		(0-62] km					(62-125] km				
(2-2.75]	true	24.66	28.01	-	-		34.29	44.57	-	-	
	m	25.13	28.59	26.18	38.82	24	34.67	42.08	34.43	48.81	71
	MSE	0.32	6.93	6.84	148.48		3.05	41.48	3.31	68.84	
(2.75-3.5]	true	25.21	28.97	-	-		34.63	45.17	-	-	
	m	25.17	29.03	25.20	41.24	7	35.16	44.79	36.00	50.11	3
	MSE	0.48	2.80	0.67	472.07		1.02	16.57	2.55	133.59	
(3.5-4.25]	true	25.68	29.77	-	-		34.41	44.79	-	-	
	m	25.28	28.83	25.10	37.34	7	34.47	44.86	35.62	46.16	0
	MSE	0.68	8.67	1.03	366.12		0.29	0.41	2.03	75.07	
(4.25-5]	true	24.61	27.94	-	-		34.19	44.41	-	-	
	m	24.58	27.59	23.86	29.25	48	33.94	44.33	34.40	44.86	13
	MSE	0.41	2.75	5.83	101.12		0.24	0.02	3.50	13.55	
		(125-187] km					(187-250] km				
(2-2.75]	true	43.63	61.11	-	-		51.33	75.93	-	-	
	m	-	-	-	-	100	-	-	-	-	100
	MSE	-	-	-	-		-	-	-	-	
(2.75-3.5]	true	43.62	61.16	-	-		51.58	76.38	-	-	
	m	43.28	58.62	45.74	61.84	32	50.61	61.17	57.57	64.53	85
	MSE	0.17	118.47	36.37	107.38		0.04	342.21	158.57	373.40	
(3.5-4.25]	true	43.96	61.74	-	-		51.32	75.91	-	-	
	m	44.00	61.65	44.52	61.75	0	51.09	75.71	55.41	76.23	4
	MSE	0.07	0.05	7.68	0.54		0.09	0.06	89.31	2.24	
(4.25-5]	true	43.63	61.08	-	-		51.62	76.45	-	-	
	m	43.52	60.88	43.56	60.88	9	51.53	76.54	51.53	76.61	22
	MSE	0.05	0.03	0.08	0.03		0.09	0.08	0.09	0.12	



**Fig. 1** A simulated waveform for each scenario and true arrival times. *From left to right*: increasing magnitude levels. *From top to bottom*: increasing distance from the nearest seismic station.

## 4 Conclusions

In this paper we have tested a new picking algorithm on a dataset of simulated waveforms to assess its performance, according to scenarios, ascribable to the characteristics of the seismic event, i.e. the magnitude, and its detection, i.e. the distance from the nearest seismic station that first recorded the event. Our preliminar experiments show that the algorithm performs well in identifying the arrival times of the first P- and S-waves. In particular, the arrival time of the first P-waves is detected more easily than the arrival time of the first S-waves. This is a relevant result because the arrival time of the first P-wave represents the beginning of the seismic event. Overall, we notice that the post-selection algorithm is not always able to correctly identify the relevant changepoints among the first estimated subset of possible values. Therefore, the first issue to be addressed in future work is the improvement of the post-selection algorithm to deal with the shape of the waveform in the specific scenario analysed.

Moreover, we have only analysed the vertical component of the recorded seismic event. In future, we wish to analyse also the two other horizontal available components, both for comparing the results with the ones of the vertical one, and also for developing a multivariate version of the proposed algorithm, able to simultaneously detect the arrival times of the two phases of the different components of the same seismic event, accounting also for the correlation among them.

Other hints for future work regard the possibility exploiting the available information given by the other detected components of the seismic event, to improve the fitting of the model and therefore the picking of the seismic phases. This could be done using one of the components as external covariate, as well as employing the functional principal component analysis. Finally, we also aim to make comparisons with other already existing algorithms in the literature.

## References

- Adelfio, G. (2012). Change-point detection for variance piecewise constant models. *Communications in Statistics-Simulation and Computation*, 41(4):437–448.
- D'Angelo, N., Adelfio, G., D'Alessandro, A., and Chiodi, M. (2020). A fast and efficient picking algorithm for earthquake early warning application based on the variance piecewise constant models. In *International Conference on Computational Science and Its Applications*, pages 903–913. Springer.
- Küperkoch, L., Meier, T., and Diehl, T. (2012). Automated event and phase identification. *New Manual of Seismological Observatory Practice 2 (NMSOP-2)*, pages 1–52.
- Muggeo, V. M. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 27(2):161–166.
- Smyth, G. K., Huele, A. F., and Verbyla, A. P. (2001). Exact and approximate reml for heteroscedastic regression. *Statistical modelling*, 1(3):161–175.



# Conditional moments based time series cluster analysis

## *Cluster analysis basata sui momenti condizionali di serie storiche*

Raffaele Mattera and Germana Scepi

**Abstract** In this paper, we propose a new approach for clustering time series showing similar time-varying moments. At this aim, we compute a dissimilarity measure assuming that the estimated conditional moments are continuous functions indexed by time. Conditional moments based clustering allows to obtain different classifications according to the data distribution's parameters. We show the usefulness of the proposed clustering procedure with an application to the financial time series in the DAX30 index.

**Abstract** *In questo articolo si propone un nuovo approccio di clustering con l'obiettivo di classificare serie storiche sulla base dei loro momenti condizionali. A questo scopo, viene calcolata una misura di dissimilarità assumendo che i momenti condizionali stimati siano funzioni continue nel tempo. Considerare la similarità dei momenti condizionali permette di ottenere classificazioni diverse sulla base dei parametri della distribuzione dei dati. Si mostra l'utilità pratica dell'approccio proposto attraverso un'applicazione alle serie storiche finanziarie contenute nell'indice DAX30.*

**Key words:** Generalized Autoregressive Score (GAS), Conditional moments, Time series clustering, Spectral density, Functional Data Analysis

## 1 Introduction

Clustering is one of the most important data mining algorithm, usually implemented for exploratory purposes, but also for more complex tasks like anomaly detection or classification. Despite the clustering techniques for time series have been exten-

---

Raffaele Mattera

Department of Economics and Statistics - University of Naples "Federico II", e-mail: raffaele.mattera@unina.it

Germana Scepi

Department of Economics and Statistics - University of Naples "Federico II", e-mail: germana.scepi@unina.it

sively studied, an approach based on their conditional moments is not very explored yet.

The time series clustering approaches can be divided into three main groups: observation-based, feature-based and model-based. The first uses raw data and the distances are directly computed on the observed time series. While these methods are particularly useful with short time series, they could also deal with time series of different length, for example by the Dynamic Time Warping. Differently, the second approach, aims to group time series by taking into account the autocorrelation function (ACF) or the partial autocorrelation function (PACF). Some of the methods belonging to this class, are based on the frequency domain features like the periodogram with its transformations [6] or the cepstral [2]. The last approach is based on the assumption that the time series are generated by the same statistical model. Therefore, this approach aims to group dynamic objects according to the estimated parameters of the underlying statistical model (e.g. ARIMA models [8] or the GARCH processes [7]).

The above clustering approaches have several drawbacks in presence of noisy or missing data [5]. Indeed they usually require pre-processing data or the application of robust procedures. Moreover, if the considered sample size is very large, most of the discussed approaches become computationally challenging.

In order to overcome these issues, clustering approaches based on Functional Data Analysis (FDA) are proposed in literature [5]. The two main approaches in functional cluster analysis are the *filtering methods* and the *nonparametric methods* [5]. The filtering approaches involve two steps: in the first, the data's dimension is reduced and, in the second, the classification is performed. The dimensionality reduction step usually consists in approximating the curves with a finite basis of functions (e.g. the spline basis [3, 4]). The nonparametric clustering methods, on the other hand, show only one step and is based on the choice of a specific distance or dissimilarity measure among curves (e.g. [9, 10]). The nonparametric clustering approaches take the advantage of a quickly implementation.

In this paper, we put ourself in a FDA perspective and we propose to cluster time series according to their estimated conditional moments. We estimate the conditional moments by the Generalized Autoregressive Score (GAS) model [1] and, following a nonparametric approach, we choose a dissimilarity measure based on the estimated conditional moments' log-spectra.

In the next section, the new approach is presented, while in the Section 3 we show an application of the clustering procedure to the financial time series in the DAX30 index. Final remarks are in the Section 4.

## 2 The proposed strategy

Our approach is based on two steps. The first step consists in the estimation of the conditional moments of time series by applying the Generalized Autoregressive

Score (GAS) model [1]. Let be  $y_{n,t}$  ( $n = 1, \dots, N; t = 1, \dots, T$ ) the  $n$ -th time series generated by the following observation density  $p(\cdot)$ :

$$y_{n,t} \sim p(y_{n,t} | f_{n,t}, \mathcal{F}_{n,t}; \theta_n), \tag{1}$$

where  $\theta_n$  is a vector of static parameters,  $\mathcal{F}_{n,t}$  is the information set at time  $t$ ,  $f_{n,t}$  is a vector of length  $J$  ( $j = 1, \dots, J$ ) of time-varying parameters depending by the probability distribution. For example, if the distribution shows a Gaussian density, we have that  $f_{n,t} = (\mu_{n,t}, \sigma_{n,t}^2)$  with  $J = 2$ , where  $\mu_{n,t}$  and  $\sigma_{n,t}^2$  represent respectively the conditional mean and the conditional variance for the  $n$ -th time series. By assuming different density functions, we get more (or less)  $J$ -th time varying parameters.

The model's information set at a given point in time  $t$ ,  $\mathcal{F}_{n,t}$ , is obtained by the previous realizations of the time series  $y_{n,t}$  and the time varying parameters  $f_{n,t}$ .

The Generalized Autoregressive Score of order one, the  $GAS(1, 1)$ , can be written as:

$$f_{n,j,t} = \omega_{n,j} + \mathbf{A}_{n,j,1} s_{n,j,t-1} + \mathbf{B}_{n,j,1} f_{n,j,t-1} \tag{2}$$

where  $\omega_{n,j}$  is a real vector and  $\mathbf{A}_{n,j,1}$  and  $\mathbf{B}_{n,j,1}$  are diagonal matrices. All the scalar parameters  $\omega_{n,j}, \mathbf{A}_{n,j,1}, \mathbf{B}_{n,j,1}$  are collected in the vector  $\theta_n$ . Moreover,  $s_{n,j,t}$  is the *scaled* score of the conditional density (1) in a time  $t$  with respect to a  $j$ -th parameter of the  $n$ -th time series.

In other words, in the GAS model we suppose that the evolution of the time-varying parameter vector  $f_{n,t}$  depends both by a vector  $s_{n,t}$ , proportional to the score of the density, and by an autoregressive component.

Another useful feature of the GAS model is that the vector of parameters  $\theta_n$  is estimated by maximum likelihood (for the details see [1]). Once the parameters in (2) are estimated, the conditional moments could be obtained by the in-sample predictions  $\hat{f}_{n,j,t}$ .

In the second step of our procedure, we assume that the  $j$ -th estimated conditional moment, for a given  $n$ -th time series,  $\hat{f}_{n,j,t}$  is a continuous function:

$$\hat{f}_{n,j,t} = g_j(t) + \varepsilon_j(t) \tag{3}$$

where  $g_j(t)$  is the unknown smooth function representation of the  $j$ -th estimated conditional moment  $\hat{f}_{n,j,t}$ .

Therefore, our approach consists in firstly converting the estimated conditional moments into a functional representation and in successively using these functions as inputs of the clustering procedure.

Following [6, 11], we consider the  $j$ -th conditional moments' *spectral density function* (also called the *spectrum*) as functional representation. To compute the dissimilarity between two spectra we use the canonical  $L^2$  distance of functional analysis:

$$d_j(\hat{f}_{j,n}, \hat{f}_{j,m}) = \left\{ \int [g_n(\lambda) - g_m(\lambda)]^2 d\lambda \right\}^{1/2} \quad (4)$$

where  $g_n$  and  $g_m$  are the spectral densities at the frequency  $\lambda$  of the  $j$ -th conditional moment for the time series  $n$  and  $m$ , respectively. The two spectra  $g_n$  and  $g_m$  are usually unknown and have to be estimated. In this paper, we take advantage of the *smoothed log-periodogram* estimator of [11].

We choose a Partitioning Around Medoids (PAM) clustering approach using the (4) as dissimilarity measure. Note that the number of clusters is a priori fixed. We select the optimal number of clusters with the Silhouette criterion (e.g. [6, 2]).

An interesting property of the proposed approach lies in the use of conditional moments for clustering time series. This choice allows to obtain different classifications according to the data distribution's parameters. For example, a researcher can potentially classify dynamic object with similar time varying skewness, if needed. This can be particularly useful for financial time series, where the investors have the preferences for positive skewed assets.

A second advantage lies on the use of the spectrum-based dissimilarity measure. The spectral density allows a functional representation of the time series and, at the same time, accurately describes the data temporal variability in the frequency domain.

### 3 Application to financial time series

We apply the proposed clustering approach to the daily time series included in the DAX30 Index from the 1<sup>th</sup> January 2015 to the 1<sup>th</sup> January 2020. We exclude time series showing missing values. The resulting sample contains 10 time series (Fig. 1) with a length  $T = 1258$ .

For each time series we estimate the conditional mean and the conditional variance by a Gaussian-GAS(1,1) model.

First, we select the optimal number of cluster with the Silhouette criterion (see Tab. 1). In the case of conditional mean, we found that  $C = 2$  clusters is the best choice with an average silhouette width of 0.997 for the stocks placed in the first group and an average width equal to 0.984 for those in the second group.

Number of clusters	$C = 2$	$C = 3$	$C = 4$
Conditional mean	0.997 - 0.984	0.995 - 0 - 0	0.868 - 0 - 0 - 0.785
Conditional variance	0.950 - 0.384	0.999 - 0.983 - 0.795	0.999 - 0.937 - 0 - 0

**Table 1** Conditional mean and conditional variance clustering: average silhouette widths

Conditional moments based time series cluster analysis

In the case of conditional variance, instead, we found that the highest average silhouette width is reached with a partitioning with  $C = 3$  clusters. In other cases, e.g.  $C = 2$  or  $C = 4$ , for some clusters we found a Silhouette average silhouette width close to zero. Instead, in the case of  $C = 3$  we get a value of 0.999 for the first group, 0.983 for the second and 0.795 for the third one.

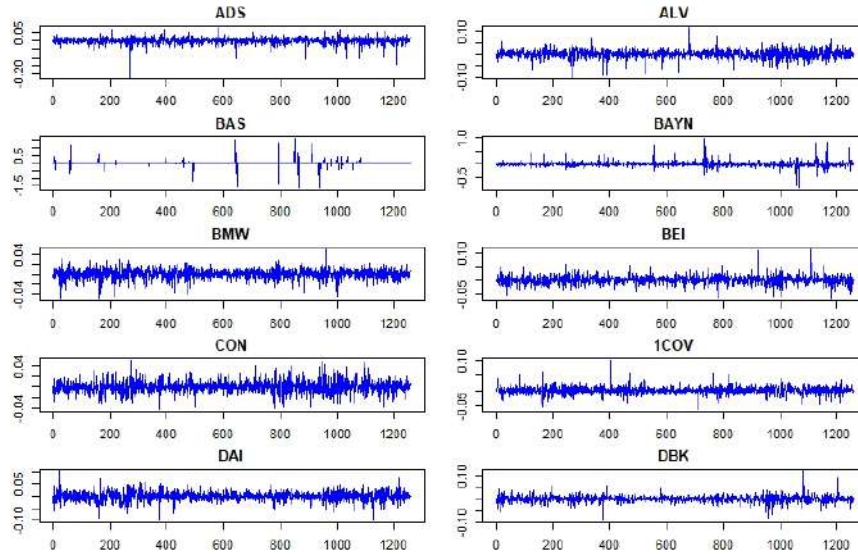


Fig. 1 Stock returns' time series

The Tab. 2 shows the final classification for both conditional mean and conditional variance, where the dissimilarity has been computed with respect the estimated conditional moments' log-spectra in (4).

	ADS	ALV	BAYN	DPW	DTE	HEI	LIN	MRK	MTX	SAP
Cluster (mean)	1	1	2	2	1	1	1	1	1	1
Cluster (variance)	1	1	2	3	2	1	2	1	2	3

Table 2 Conditional mean and conditional variance clustering: group assignment

Some stocks show similar dynamic trajectories for both the conditional mean and the conditional variance and therefore are always placed in the same groups (e.g. ADS, ALV or HEI for the group 1 and BAY for the group 2). Nevertheless we observe that some stocks have similar conditional means but dissimilar conditional variances or vice versa (e.g. DTE, LIN, MTX and SAP). In these cases it is possible to: a) consider only the conditional mean classification or b) consider only the conditional variance classification. In general, in the case of uncertain classification, a

simulation study with the aim of evaluating the miss-classification rate derived by the choice between the two approaches has to be conducted. For financial time series it is reasonable to consider the conditional variance classification rather than the conditional mean since the stock returns show heteroskedasticity. Moreover we implicitly account for the *volatility clustering* phenomenon.

#### 4 Final remarks

In this paper, we show the advantages of using a clustering approach based on the estimated conditional moments' log-spectra. Future works on this direction could be the development of a fuzzy extension of the proposed procedure and to apply our method to a portfolio selection problem (e.g. [4]).

#### References

- [1] D. Creal, S. J. Koopman, and A. Lucas. Generalized autoregressive score models with applications. *J APPL ECONOM*, 28(5):777–795, 2013.
- [2] P. D'Urso, L. De Giovanni, R. Massari, R. L. D'Ecclesia, and E. A. Maharaj. Cepstral-based clustering of financial time series. *EXPERT SYST APPL*, 161:113705, 2020.
- [3] C. Iorio, G. Frasso, A. D'Ambrosio, and R. Siciliano. Parsimonious time series clustering using p-splines. *EXPERT SYST APPL*, 52:26–38, 2016.
- [4] C. Iorio, G. Frasso, A. D'Ambrosio, and R. Siciliano. A p-spline based clustering approach for portfolio selection. *Expert Systems with Applications*, 95:88–103, 2018.
- [5] J. Jacques and C. Preda. Functional data clustering: a survey. *ADV DATA ANAL CLASSI*, 8(3):231–255, 2014.
- [6] E. A. Maharaj and P. D'Urso. Fuzzy clustering of time series in the frequency domain. *INFORM SCIENCES*, 181(7):1187–1211, 2011.
- [7] E. Otranto. Clustering heteroskedastic time series by model-based procedures. *COMPUT STAT DATA AN*, 52(10):4685–4698, 2008.
- [8] D. Piccolo. A distance measure for classifying arima models. *J TIME SER ANAL*, 11(2):153–164, 1990.
- [9] T. Tarpey and K. K. Kinatader. Clustering functional data. *J CLASSIF*, 20(1), 2003.
- [10] S. Tokushige, H. Yadohisa, and K. Inada. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *COMPUTATION STAT*, 22(1):1–16, 2007.
- [11] J. M. Vilar, J. A. Vilar, and S. Pértega. Classifying time series data: A non-parametric approach. *J CLASSIF*, 26(1):3–28, 2009.

# On the asymptotic mean-squared prediction error for multivariate time series

## *Errore quadratico medio asintotico di previsione per serie storiche multivariate*

Gery Andrés Díaz Rubio, Simone Giannerini, and Greta Goracci

**Abstract** The aim of the article is to extend the Misspecification Resistant Information Criterion (MRIC) proposed in [H.-L. Hsu, C.-K. Ing, H. Tong: On model selection from a finite family of possibly misspecified time series models. *The Annals of Statistics*. 47 (2), 1061–1087 (2019)] to the bivariate case. We obtain an asymptotic expression for the mean squared prediction error matrix in the context of multivariate time series models and present the example of a bivariate time series model with a single regressor. This decomposition features the same structure as in the scalar case and paves the way to the complete extension of the MRIC to multivariate time series models.

**Abstract** *L'obiettivo principale dell'articolo è estendere al caso bivariato il Misspecification Resistant Information Criterion (MRIC) proposto in (H.-L. Hsu, C.-K. Ing, H. Tong: On model selection from a finite family of possibly misspecified time series models. The Annals of Statistics. 47 (2), 1061-1087 (2019)). A tal fine si deriva un'espressione asintotica per la matrice dell'errore quadratico medio di previsione per serie temporali multivariate. Presentiamo l'esempio di un modello di serie temporale bivariata con un singolo regressore. Questa scomposizione presenta la stessa struttura del caso scalare e rappresenta la base di partenza per l'estensione del MRIC a modelli di serie temporali multivariate.*

**Key words:** multivariate time series, MSPE matrix, information criteria, vectorial MRIC

---

Gery Andrés Díaz Rubio

Dipartimento di Scienze Statistiche, Università di Bologna, via delle belle arti 41, Bologna, e-mail: [geryandre.diazrubio2@unibo.it](mailto:geryandre.diazrubio2@unibo.it)

Simone Giannerini

Dipartimento di Scienze Statistiche, Università di Bologna, via delle belle arti 41, Bologna, e-mail: [simone.giannerini@unibo.it](mailto:simone.giannerini@unibo.it)

Greta Goracci

Dipartimento di Scienze Statistiche, Università di Bologna, via delle belle arti 41, Bologna, e-mail: [greta.goracci@unibo.it](mailto:greta.goracci@unibo.it)

## 1 Introduction

The model selection step is a fundamental task in statistical modelling and its implementation typically depends upon the objective of the exercise. In the time series framework the focus is on either forecasting future values or describing/controlling the process that has generated the data (DGP). A good model selection criterion must feature a good ability to identify the model with the ‘best’ fit to future values, in a specified sense. In particular, in the parametric time series framework, we can identify two main properties. The first is consistency, namely, the ability to select the true DGP with probability one as the sample size diverges. This assumes that a true model exists and is among the set of candidate models. If either the set of candidate models does not contain the true DGP, or, for some reason, a true model cannot be postulated, then a selection criterion should be efficient in the mean square sense, i.e. it minimizes the mean squared prediction error. Starting from the seminal work of Akaike, [2] a plethora of model selection criteria has been proposed. These include Akaike’s AIC [2, 1], Schwarz’s Bayesian Information Criterion (BIC) [7], and Rissanen’s Minimum Description Length (MDL) [6]. Such criteria paved the way for various extensions dealing with different unsolved issues. For instance, the AIC is efficient but not consistent (i.e. it leads to select overfitting models), whereas the BIC is consistent but not efficient, see [3] for a discussion.

A recent development for model selection in possibly misspecified parametric time series models in the fixed-dimensionality setting is given by the Misspecification Resistant Information Criterion (MRIC) [3]. By fixed-dimensionality it is intended that the number of observations goes to infinity while the number of ‘true’ parameters is finite. Among other features, the MRIC enjoys both consistency and asymptotic efficiency so that, in this respect, it provides a solution to the original research question of Akaike. In this work we present an extension to the bivariate case of the decomposition of the mean square prediction error, which is a necessary step towards the generalization of the MRIC to the multivariate time series case. In Section 2 we present the approach based upon the MRIC for parametric univariate time series models, whereas in Section 3 we present the main theorem and the sketch of the proof.

## 2 The MRIC approach

Let  $\{\mathbf{y}_t\} \equiv \{(y_{t,1}, \dots, y_{t,w})^\top\}$  and  $\{\mathbf{x}_t\} \equiv \{(x_{t,1}, \dots, x_{t,m})^\top\}$  be two weakly stationary stochastic processes defined over the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  with  $w$  and  $m$  positive integer numbers. We observe the multivariate time series  $\mathbf{y}_t$ , and  $\mathbf{x}_t$ , with  $t = 1, 2, \dots, N, N+1, \dots, N+h = n$ . Also, define the sample means of  $\mathbf{y}_t$  and  $\mathbf{x}_t$  respectively as  $\bar{\mathbf{y}} = n^{-1} \sum_{t=1}^n \mathbf{y}_t$ , and  $\bar{\mathbf{x}} = n^{-1} \sum_{t=1}^n \mathbf{x}_t$ . The stationarity assumptions allows us to replace the unconditional expectations,  $E[\mathbf{y}_{t+h}]$  and  $E[\mathbf{x}_t]$ , with their respective sample counterparts. Without loss of generality, assume:  $E[\mathbf{y}_t] = \mathbf{0}$ , and  $E[\mathbf{x}_t] = \mathbf{0}$ . We adopt the following  $h$ -step ahead forecasting model:



$$\mathbf{y}_{t+h} = \beta_h \mathbf{x}_t + \boldsymbol{\varepsilon}_{t,h}, \quad (1)$$

where  $\boldsymbol{\varepsilon}_{t,h}$  is the vector of  $h$ -step ahead forecasting errors having the same dimension as  $\mathbf{y}_t$ . Note that since the model can possibly be misspecified, the prediction error  $\boldsymbol{\varepsilon}_{t,h}$  vector can be both serially and cross-correlated and also correlated with  $\mathbf{x}_s$ , with  $s \neq t$ . The OLS estimators of  $\beta_h$  is denoted with  $\hat{\beta}_n(h)$ . In [3] the asymptotic decomposition of the MSPE for the  $h$ -step ahead prediction is derived in the univariate case, i. e.  $w = 1$  and  $m \geq 1$ :

$$\text{MSPE}_h = \text{MI}_h + n^{-1}(\text{VI}_h + o(1)). \quad (2)$$

The MSPE is decomposed in two parts, the first one being the Misspecification Index (MI), which is linked to the goodness-of-fit of the model and is equal to the variance of the  $h$ -step ahead prediction error, i. e.

$$\text{MI}_h = E[\boldsymbol{\varepsilon}_{1,h}^2]. \quad (3)$$

The second component is the Variability Index (VI), which depends upon the variance of the  $h$ -step ahead predictor,  $\hat{y}_{n+h} = \hat{\beta}_n^T(h)\mathbf{x}_n$ , and which is also connected to the estimation error of  $\hat{\beta}_n(h)$ :

$$\text{VI}_h = L_h = \text{tr}\{\mathbf{R}^{-1}\mathbf{C}_{h,0}\} + 2 \sum_{s=1}^{h-1} \text{tr}\{\mathbf{R}^{-1}\mathbf{C}_{h,s}\}.$$

Here,  $\mathbf{R} = E[\mathbf{x}_1\mathbf{x}_1^T]$  is the (non-singular) variance-covariance matrix of the regressors, whereas  $\mathbf{C}_{h,s} = E[\mathbf{x}_1\mathbf{x}_{1+s}^T\boldsymbol{\varepsilon}_{1,h}\boldsymbol{\varepsilon}_{1+s,h}^T]$  represents the cross-covariance matrix between the regressors and the  $h$ -step ahead prediction error. Based upon such asymptotic decomposition of the MSPE it is possible to derive the MRIC as follows:

$$\text{MRIC}_h = \hat{\text{MI}}_h + \frac{\alpha_n}{n} \hat{\text{VI}}_h \quad (4)$$

where  $\alpha_n$  is a penalization term such that  $\alpha_n/n^{1/2} \rightarrow +\infty$ ,  $\alpha_n/n \rightarrow 0$  and MI and VI are estimated through the method of moments. The asymptotic efficiency of the MRIC is proved in [3], Theorem 3.1 and 3.2. The MRIC selects the model that minimises the MSPE by selecting the model with the smallest VI among those with the smallest MI.

### 3 Main results

In this work we extend the decomposition that is at the basis of the MRIC to the case of bivariate time series models. In particular, we present the case with a bivariate response ( $w = 2$ ) and a single regressor ( $m = 1$ ). The difference with the scalar case is not just algebraic and requires a different approach since  $\boldsymbol{\varepsilon}_{t,h}$  is a vector pro-

cess with non-negligible cross-dependence. Consider the multivariate  $h$ -step ahead forecasting model (where the variables are centred):

$$\mathbf{y}_{t+h} = \beta_h x_t + \varepsilon_{t,h}, \tag{5}$$

where  $\varepsilon_{t,h} = (\varepsilon_{t,1,h}, \varepsilon_{t,2,h})^\top$  is the vector of  $h$ -step ahead forecasting errors,

$$\beta_h = (E[x_t^2])^{-1} E[x_t \mathbf{y}_{t+h}] = R^{-1} E[x_t \mathbf{y}_{t+h}].$$

The MSPE matrix is given by:

$$\text{MSPE} = E[(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})^\top],$$

where  $\hat{\mathbf{y}}_{n+h} = \hat{\beta}_h x_n$ , with  $\hat{\beta}_n(h) = \hat{\beta}_h$  being the OLS estimator of  $\beta$ :

$$\hat{\beta}_h = \hat{R}^{-1} \left( \frac{1}{N} \sum_{t=1}^N x_t \mathbf{y}_{t+h} \right) \quad \text{and} \quad \hat{R} = N^{-1} \sum_{t=1}^N x_t^2.$$

It is possible to show that

$$\text{MSPE} = E[\mathbf{A}] + E[\mathbf{B}] + E[\mathbf{C}],$$

where

$$\begin{aligned} E[\mathbf{A}] &= E \left[ \left( (\hat{\beta}_h - \beta_h) x_n \right) \left( (\hat{\beta}_h - \beta_h) x_n \right)^\top \right], \\ E[\mathbf{B}] &= -E \left[ \left[ (\hat{\beta}_h - \beta_h) x_n \right] \varepsilon_{n,h}^\top + \varepsilon_{n,h} \left[ (\hat{\beta}_h - \beta_h) x_n \right]^\top \right], \\ E[\mathbf{C}] &= E \left[ \varepsilon_{n,h} \varepsilon_{n,h}^\top \right]. \end{aligned}$$

Different versions of this MSPE matrix are present in the literature; for instance, in [5, pp. 15-17, 47] the emphasis is on vectorial time series whereas for the multivariate regression context see [4, pp. 662-663]. We assume the following conditions

C1)  $\exists q_1 > 5, 0 < C_1 < \infty$ : for any  $1 \leq n_1 < n_2 \leq n$

$$E \left[ \left| (n_2 - n_1 + 1)^{-1/2} \sum_{t=n_1}^{n_2} x_t^2 - E[x_t^2] \right|^{q_1} \right] \leq C_1.$$

C2) 1.  $\mathbf{C}_{h,s} = E[\varepsilon_{t,h} x_t (\varepsilon_{t+s,h} x_{t+s})^\top] \perp t$ ,  
 2.  $E[x_1 x_n \varepsilon_{1,i,h} \varepsilon_{n,j,h}] = o(n^{-1}) \forall i, j = 1, 2$ .

C3)  $\sup_{-\infty < t < \infty} E[|x_t|^{10}] < \infty$ ,  $\sup_{-\infty < t < \infty} E[\|\varepsilon_{t,h}\|^6] < \infty$ .

C4)  $\exists 0 < C_2 < \infty$ : for  $1 \leq n_1 < n_2 \leq n$ ,

MSPE in multivariate time series

$$E \left[ \left\| (n_2 - n_1 + 1)^{-\frac{1}{2}} \sum_{t=n_1}^{n_2} \varepsilon_{t,h} x_t \right\|^5 \right] < C_2.$$

C5) For any  $q > 0$ ,  $E [\|\hat{R}^{-1}\|^q] = O(1)$ .

C6)  $\exists \mathcal{F}_t \subseteq \mathcal{F}$ ,  $\mathcal{F}_t$  an increasing sequence of  $\sigma$ -fields such that:

1.  $x_t$  is  $\mathcal{F}_t$ -measurable;
2.  $\sup_{-\infty < t < \infty} E \left[ \|E[x_t^2 | \mathcal{F}_{t-k}] - R\|^3 \right] = o(1)$ , as  $k \rightarrow \infty$ ;
3.  $\sup_{-\infty < t < \infty} E \left[ \|E[\varepsilon_{t,h} x_t | \mathcal{F}_{t-k}] - \mathbf{0}\|^3 \right] = o(1)$ , as  $k \rightarrow \infty$ .

Note that these are the natural vectorial extension of those in [3]. We present the main result, which extends Theorem 2.1 [3], i. e. derives the asymptotic decomposition of the MSPE in the bivariate case.

**Theorem 1.** *Under the regularity conditions C1) – C6), the asymptotic expression of the MSPE matrix for the case  $w = 2$ , and  $m = 1$  results*

$$N \{ E [(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})^\top - E[\varepsilon_{n,h} \varepsilon_{n,h}^\top]] \} = R^{-1} E [(\varepsilon_{1,h} x_1)(\varepsilon_{1,h} x_1)^\top] + R^{-1} E \left[ \left( \sum_{s=1}^{h-1} \{ (\varepsilon_{1,h} x_1)(\varepsilon_{s+1,h} x_{s+1})^\top + (\varepsilon_{s+1,h} x_{s+1})(\varepsilon_{1,h} x_1)^\top \} \right) \right] + o(1). \quad (6)$$

### Sketch of the proof

With routine algebra we can write the MSPE matrix as:

$$E [(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})^\top] - E [\varepsilon_{n,h} \varepsilon_{n,h}^\top] = (\text{I}) + (\text{II}) + (1) + (2) + (3),$$

where

$$\begin{aligned} (\text{I}) &= -E \left[ x_n \hat{R}^{-1} \left[ \hat{\Sigma} \varepsilon_{n,h}^\top + \varepsilon_{n,h} \hat{\Sigma}^\top \right] \right]; & (\text{II}) &= E \left[ \hat{R}^{-1} \hat{\Sigma} x_n x_n \hat{\Sigma}^\top \hat{R}^{-1} \right]; \\ (1) &= E \left[ R^{-1} \Sigma x_n x_n \Sigma^\top R^{-1} \right]; & (2) &= -E \left[ R^{-1} \hat{R}^{-1} x_n x_n \left[ \Sigma \hat{\Sigma}^\top + \hat{\Sigma} \Sigma^\top \right] \right]; \\ (3) &= E \left[ x_n R^{-1} \left[ \Sigma \varepsilon_{n,h}^\top + \varepsilon_{n,h} \Sigma^\top \right] \right], \end{aligned}$$

with  $\Sigma = E[x_t \varepsilon_{t,h}]$ , and  $\hat{\Sigma} = (N^{-1} \sum_{t=1}^N x_t \varepsilon_{t,h})$ . Now, it can be shown that under conditions C1) - C6), (1), (2), (3) vanish and

$$N(\text{I}) = (\text{III}) + o(1); \quad N(\text{II}) = (\text{IV}) + o(1),$$

with

$$\begin{aligned} \text{(III)} &= -E \left[ x_n R^{-1} \left[ \hat{\Sigma}_A \varepsilon_{n,h}^\top + \varepsilon_{n,h} \hat{\Sigma}_A^\top \right] \right]; \\ \text{(IV)} &= E \left[ \hat{\Sigma}_B R^{-1} \hat{\Sigma}_B^\top \right], \end{aligned}$$

where  $\hat{\Sigma}_A = \sum_{t=1}^N \varepsilon_t x_t$  and  $\hat{\Sigma}_B = N^{-\frac{1}{2}} \sum_{t=1}^N \varepsilon_t x_t$ . Therefore, the asymptotic expression for the MSPE matrix in the bivariate case results:

$$\begin{aligned} N \{ E [ (\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h}) (\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})^\top ] - E [ \varepsilon_{n,h} \varepsilon_{n,h}^\top ] \} &= N [ \text{(I)} + \text{(II)} ] \\ &= \text{(III)} + \text{(IV)} + o(1). \end{aligned}$$

and the result of Eq. (6) follows from routine algebra and conditions C1) - C6).

Similarly to the scalar case (Eq. 4), the vectorial MRIC can be derived from the method of moments estimators. Note that now the MRIC is a vector with 2 components. An overall criterion can be obtained by considering a suitable norm. The next step is to prove the asymptotic efficiency and the misspecification resistance and this is the subject of current further investigations.

## References

1. H. Akaike.: A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19** (6), 716–723 (1974).
2. H. Akaike.: Information Theory and an Extension of the Maximum Likelihood Principle. In: S. Kotz, N.L. Johnson (eds.), *Breakthroughs in Statistics: foundations and basic theory*, pp. 610–624. Springer Science+Business Media, New York (1992).
3. H.-L. Hsu, C.-K. Ing, H. Tong: On model selection from a finite family of possibly misspecified time series models. *The Annals of Statistics*. **47** (2), 1061–1087 (2019).
4. C.L. Mallows.: Some remarks of  $C_p$ . *Technometrics*. **15**, 661–675 (1973).
5. G.C. Reinsel.: *Multivariate Time Series Analysis*. Springer-Verlag, New York (1993).
6. J. Rissanen.: Modeling by shortest data description. *Automatica*, **14** (5):465–471 (1978).
7. G. Schwarz.: Estimating the dimension of a model. *The Annals of Statistics*. **6** (2), 461–464 (1978).

# Spherical autoregressive change-point detection with applications

## *Stima del punto di cambio mediante modelli autoregressivi sulla sfera e sue applicazioni*

Federica Spoto, Alessia Caponera and Pierpaolo Brutti

**Abstract** Spatio-temporal processes arise very naturally in a number of different applied fields, like Cosmology, Astrophysics, Geophysics, Climate and Atmospheric Science. In most of these areas, the detection of structural breaks or regime shifts in the data stream is key. To this end, in the present work, we aim at generalizing the recently introduced SPHAR(p) process by allowing for temporal changes in its functional parameters and variability structure. Our approach, which intrinsically integrates the spatial and temporal dimensions, could give multiscale insights into both the global and local behavior of changes, and its performance will be tested on a real dataset of global surface temperature anomalies.

**Abstract** *I processi spatio-temporali sorgono naturalmente in numerosi campi applicativi, come la Cosmologia, l'Astrofisica, la Geofisica, le Scienze del Clima e dell'Atmosfera. In molti di questi ambiti, l'individuazione di break strutturali nella serie dei dati è fondamentale. A tal fine, nel presente lavoro, ci proponiamo di generalizzare i processi SPHAR(p) introducendo cambiamenti temporali nei parametri funzionali e nella loro struttura di variabilità. Il nostro approccio, oltre ad integrare esplicitamente sia la dimensione spaziale che quella temporale del fenomeno in studio, permette al contempo di estrarre informazioni multiscala che meglio qualificano e caratterizzano i punti di cambio individuati. Le prestazioni della modellistica proposta saranno testate su un dataset reale relativo ad anomalie della temperatura superficiale globale.*

**Key words:** Spherical Functional Autoregressions, Change-point model, Spatio-temporal model, Global climate change

---

Federica Spoto  
Sapienza University of Rome, e-mail: federica.spoto@uniroma1.it

Alessia Caponera  
École Polytechnique Fédérale de Lausanne, e-mail: alessia.caponera@epfl.ch

Pierpaolo Brutti  
Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

## 1 Introduction

Over the last few decades, the study of random fields on the sphere has received increasing attention because of their real-life applications in a variety of different areas like Cosmology, Astrophysics, Climatology and many more. In most of these areas, the detection of structural breaks or regime shifts in the data stream is key. In Climate Sciences, for example, variations in the rate at which global surface temperatures evolve is the most prominent and widely studied footprint of global warming. Despite this, the vast majority of such analyses are purely temporal or do not take into account the spatial dependence. A few notable exceptions are [2], [1] and [9].

In the present work, we aim at generalizing the recently introduced SPHAR( $p$ ) process by allowing for temporal changes in its functional parameters and variability structure. Our approach, which intrinsically integrates the spatial and temporal dimensions, could give multiscale insights into both the global and local behavior of changes.

## 2 Materials and Methods

### 2.1 Spherical functional autoregressions

Working in a functional time series setup, we focus on time-varying spherical random field  $\{T(x, t) : (x, t) \in \mathbb{S}^2 \times \mathbb{Z}\}$  which exhibits a discrete temporal dynamics over the unit sphere  $\mathbb{S}^2$  so that, for every fixed  $t \in \mathbb{Z}$ , the field  $T_t \equiv T(\cdot, t)$  is a random element of  $L^2(\mathbb{S}^2)$  (the space of square-integrable functions on the unit sphere), and admits a characterization in terms of spherical functional autoregressive models as described in [6, 7].

Specifically, sphere-cross-time random fields belonging to the class of spherical functional autoregressions of order  $p$  (SPHAR( $p$ )) satisfy

$$T(x, t) = \sum_{i=1}^p (\Phi_i T_{t-i})(x) + Z(x, t), \quad \forall (x, t) \in \mathbb{S}^2 \times \mathbb{Z}, \quad (1)$$

where  $\{Z(x, t) : (x, t) \in \mathbb{S}^2 \times \mathbb{Z}\}$  is a Gaussian isotropic spherical white noise and  $\{\Phi_i : i = 1, \dots, p\}$  are integral operators on  $L^2(\mathbb{S}^2)$  associated with  $p$  continuous isotropic kernels  $\{k_i : i = 1, \dots, p\}$ ; see [7] for more formal and detailed definitions. Such processes can be interpreted as a generalization of autoregressive (AR( $p$ )) processes, taking values on  $L^2(\mathbb{S}^2)$ , rather than on the real line (see also [3]).

The existence of a unique spatially isotropic and temporally stationary solution for Equation (1) is guaranteed by assuming some conditions on the  $\Phi_i$ 's. For instance, when  $p = 1$ , a necessary and sufficient condition is given by  $\|\Phi\|_{\text{op}} =$

$\max_{\ell \in \mathbb{N}} |\phi_\ell| < 1$ , where the  $\phi_\ell$ 's are the eigenvalues of  $\Phi$ . See [5] for an in-depth discussion.

It is well known that the following spectral representation holds in  $L^2(\Omega)$

$$T(x, t) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell,m}(t) Y_{\ell,m}(x), \quad \forall (x, t) \in \mathbb{S}^2 \times \mathbb{Z},$$

where the set  $\{Y_{\ell,m}(\cdot) : \ell \geq 0, m = -\ell, \dots, \ell\}$  is a standard basis for  $L^2(\mathbb{S}^2)$  of real-valued spherical harmonics (see, e.g., [11]), and  $\{a_{\ell,m}(\cdot) : \ell \geq 0, m = -\ell, \dots, \ell\}$  are the (random) generalized Fourier coefficients defined as  $a_{\ell,m}(t) = \langle T_t, Y_{\ell,m} \rangle_{L^2}$  and satisfying

$$\mathbb{E}[a_{\ell,m}(t) a_{\ell',m'}(s)] = 0, \quad \text{for } \ell \neq \ell', m \neq m'.$$

Moreover, for every fixed  $(\ell, m)$ , it is possible to show that

$$a_{\ell,m}(t) = \sum_{i=1}^p \phi_{\ell,i} a_{\ell,m}(t-i) + \varepsilon_{\ell,m}(t), \quad \forall t \in \mathbb{Z},$$

where  $\{\varepsilon_{\ell,m}(t) = \langle Z_t, Y_{\ell,m} \rangle_{L^2} : t \in \mathbb{Z}\}$  is a Gaussian white noise with variance  $\sigma_\ell^2 > 0$ . Note that the  $\phi_{\ell,i}$ 's and  $\sigma_\ell^2$  do not depend on  $m$ , as a consequence of isotropy.

## 2.2 Spherical change-point detection

Under the assumptions described in the previous section, we introduce the spherical autoregressive change-point model and the methodology to detect possible change-points in the data. For the sake of simplicity, our arguments are presented for a SPHAR(1) model, allowing a single change-point; however, the analysis can be generalized to higher autoregressive orders and multiple change-points. In this setting, the model is written as the composition of two stationary SPHAR segments and takes the form

$$T(x, t) = \begin{cases} (\Phi_1 T_{t-1})(x) + Z_1(x, t) & t < \tau \\ (\Phi_2 T_{t-1})(x) + Z_2(x, t) & t \geq \tau \end{cases},$$

that, given  $\tau$ , are assumed to be independent; equivalently, thanks to the spectral representation, one can jointly look at

$$a_{\ell,m}(t) = \begin{cases} \phi_{\ell;1} a_{\ell,m}(t-1) + \varepsilon_{\ell,m;1}(t) & t < \tau \\ \phi_{\ell;2} a_{\ell,m}(t-1) + \varepsilon_{\ell,m;2}(t) & t \geq \tau \end{cases}, \quad \ell \geq 0, m = -\ell, \dots, \ell.$$

The task consists in detecting the time-stamp  $\tau$  at which the model parameters have a variation in value. The optimal change-point is selected through a model choice criteria based on information theory.

Throughout this paper, we shall assume to be able to observe a finite set of Fourier coefficients

$$\alpha = \{a_{\ell,m}(t) : t = 1, \dots, n, \ell = 0, \dots, L, m = -\ell, \dots, \ell\}.$$

Given  $\tau$  and for fixed  $(\ell, m)$ , one can define the vectors

$$\alpha_{\ell,m;1} = (a_{\ell,m}(1), \dots, a_{\ell,m}(\tau - 1))^T, \quad \alpha_{\ell,m;2} = (a_{\ell,m}(\tau), \dots, a_{\ell,m}(n))^T,$$

of dimensions  $n_1$  and  $n_2$ , respectively. Thus, for  $j = 1, 2$ ,

$$\mathbb{E}[\alpha_{\ell,m;j} \alpha_{\ell,m;j}^T] = \sigma_{\ell,j}^2 V_{\ell;j},$$

where  $\sigma_{\ell,j}^2$  is the noise variance and  $V_{\ell;j}$  is a  $n_j \times n_j$  symmetric and positive definite matrix depending on  $\phi_{\ell,j}$ . The likelihood function for the parameters  $\theta = \{\phi_{\ell,j}, \sigma_{\ell,j}^2, \ell = 0, \dots, L, j = 1, 2\}$  and  $\tau$  is then

$$L(\theta, \tau; \alpha) = \prod_{\ell=0}^L \prod_{m=-\ell}^{\ell} \prod_{j=1}^2 (2\pi\sigma_{\ell,j}^2)^{-n_j/2} |V_{\ell;j}|^{-1/2} \exp \left\{ -\frac{1}{2\sigma_{\ell,j}^2} \alpha_{\ell,m;j}^T V_{\ell;j}^{-1} \alpha_{\ell,m;j} \right\};$$

moreover, using the standard approximation to the log-likelihood for AR models (see [4]), one gets

$$-\frac{2}{n} \log L(\hat{\theta}, \tau; \alpha) = \frac{1}{n} \sum_{\ell=0}^L (2\ell + 1) \sum_{j=1}^2 n_j \log(2\pi\hat{\sigma}_{\ell,j}^2) + (L + 1)^2 + o_L(1),$$

with  $\hat{\theta} = (\hat{\phi}_{\ell,j}, \hat{\sigma}_{\ell,j}^2, \ell = 0, \dots, L, j = 1, 2)^T$  being the corresponding maximum likelihood estimate (MLE) of  $\theta$ . Hence,  $\hat{\tau}$  can be defined as the value that minimizes

$$R(\tau) = \sum_{\ell=0}^L (2\ell + 1) \sum_{j=1}^2 n_j \log(2\pi\hat{\sigma}_{\ell,j}^2). \tag{2}$$

Note that this is equivalent to minimize the AIC score, since the number of parameters is constant. In addition, for computational reasons, one may replace the MLE estimate of  $\sigma_{\ell,j}^2$  with the Yule-Walker or least squares estimates, due to their equivalence in large sample size regimes.

### 3 Results

The methodology presented above was applied to *global (land and ocean) surface temperature anomalies*. More in detail, the dataset is built starting from the NCEP/NCAR monthly averages of the surface air temperature (in degrees Celsius) from 1948 to 2020, over a global grid with 2.5° spacing for latitude and longi-



tude, see [10]. Following the World Meteorological Organization policy, temperature anomalies are obtained by subtracting the long-term monthly means relative to the 1981–2010 base period. They are then averaged over months to switch from a monthly scale to an annual scale.

By means of the `healpix` package (see [8] and the official `healpix` website), we converted the gridded data into spherical maps with a resolution of  $12 \cdot \text{NSIDE}^2$  pixels ( $\text{NSIDE} = 16$ ) and then we computed the Fourier coefficients up to  $L = 2 \cdot \text{NSIDE}$ .

In order to handle possible anisotropies in the mean, for each segment  $j = 1, 2$ , we introduced an intercept  $\mu_j \in L^2(\mathbb{S}^2)$ , which has a representation in terms of spherical harmonics

$$\mu_j = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \mu_{\ell,m;j} Y_{\ell,m}, \quad \text{in } L^2(\mathbb{S}^2),$$

with  $\mu_{\ell,m;j} = \langle \mu_j, Y_{\ell,m} \rangle_{L^2}$ .

By minimizing (2) over  $\tau \in \{1953, \dots, 2016\}$ , the best change point results to be  $\hat{\tau} = 1982$ . Then, we can estimate the functional parameters  $(\mu_j, \Phi_j)$  by solving the following least-squares minimization problem, see [6, 7],

$$(\hat{\mu}_j, \hat{\Phi}_j) := \operatorname{argmin} \sum_{t \in \mathcal{T}_j} \|T_t - \mu_{j:L} - \Phi_{j:L} T_{t-1}\|_{L^2(\mathbb{S}^2)}^2,$$

where  $\mu_{j:L}$  and  $\Phi_{j:L}$  are the truncated version of  $\mu_j$  and  $\Phi_j$ , respectively, and  $\mathcal{T}_j$  is the set of time-stamps belonging to each segment, i.e.  $\mathcal{T}_1 = \{1949, \dots, \hat{\tau} - 1\}$  and  $\mathcal{T}_2 = \{\hat{\tau}, \dots, 2020\}$ .

The comparison between the two periods can be carried out by computing the two mean surfaces

$$(I_L - \hat{\Phi}_j)^{-1} \hat{\mu}_j = \sum_{\ell=0}^L \sum_{m=-\ell}^{\ell} \frac{\hat{\mu}_{\ell,m;j}}{1 - \hat{\Phi}_{\ell;j}} Y_{\ell,m}, \quad j = 1, 2.$$

Figure 1 shows the estimated mean surfaces pre and post  $\hat{\tau} = 1982$  (on the same color scale) and their difference. A positive anomaly indicates that the observed temperature was warmer than the reference value, while a negative anomaly indicates that the observed temperature was cooler than the reference value.

The analysis suggests an overall increase in the mean surface temperature anomalies, which is particularly evident for the North and South poles.

## References

- Altieri, L., Cocchi, D., Greco, F., Illian, J., Scott, E.: Bayesian P-splines and advanced computing in R for a changepoint analysis on spatio-temporal point processes. *Journal of Statistical Computation and Simulation* **86**(13), 2531–2545 (2016)

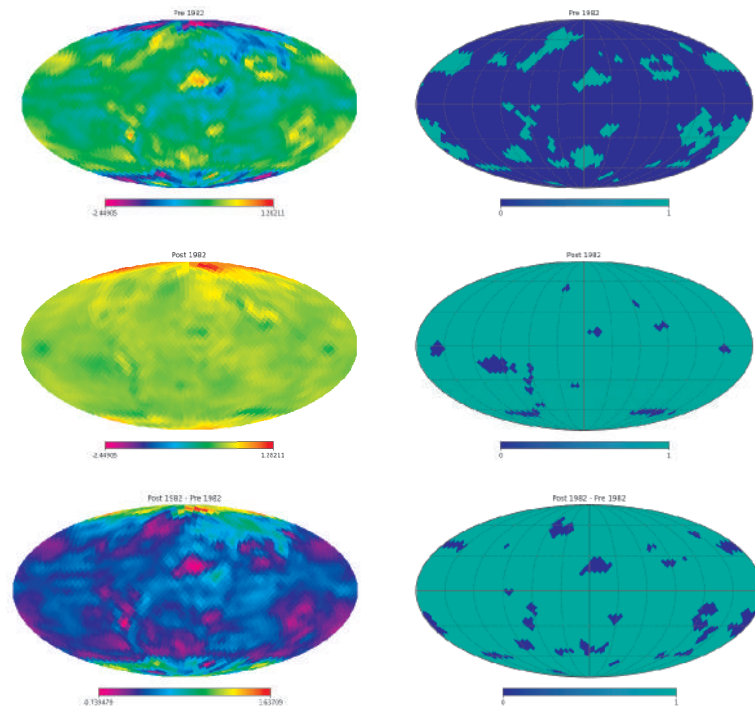


Fig. 1: Left: estimated mean surfaces pre and post  $\hat{\tau} = 1982$  (on the same color scale) and their difference. Right: corresponding negative (0) and positive (1) pixels.

2. Altieri, L., Scott, E.M., Cocchi, D., Illian, J.B.: A changepoint analysis of spatio-temporal point processes. *Spatial Statistics* **14**, 197–207 (2015). *Spatio-Temporal Stochastic Modelling of Environmental Hazards*
3. Bosq, D.: *Linear Processes in Function Spaces: Theory and Applications*. Springer (2000)
4. Brockwell, P., Davis, R.: *Time Series: Theory and Methods: Theory and Methods*. Springer Series in Statistics. Springer New York (1991)
5. Caponera, A.: SPHARMA approximations for stationary functional time series on the sphere (2020). arXiv:2009.13189
6. Caponera, A.: Statistical inference for spherical functional autoregressions. PhD Thesis, Sapienza University of Rome (2020)
7. Caponera, A., Marinucci, D.: Asymptotics for spherical functional autoregressions. *Annals of Statistics* **49**(1), 346–369 (2021)
8. Gorski, K.M., Hivon, E., Banday, A.J., Wandelt, B.D., Hansen, F.K., Reinecke, M., Bartelmann, M.: HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal* **622**(2), 759–771 (2005)
9. Jun, M.: Matérn-based nonstationary cross-covariance models for global processes. *Journal of Multivariate Analysis* **128**, 134–146 (2014)
10. Kalnay, E., et al.: The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**(3), 437–472 (1996)
11. Marinucci, D., Peccati, G.: *Random Fields on the Sphere: Representation, Limit Theorems and Cosmological Applications*. London Mathematical Society Lecture Note Series. Cambridge University Press (2011)



# 5 Posters

# A method for incorporating historical information in non-inferiority trials

## *Un metodo per l'uso di informazione storica nelle prove cliniche di non-inferiorità*

Fulvio De Santis and Stefania Gubbiotti

**Abstract** In non-inferiority trials a new experimental treatment is compared to an active control therapy. Often information on the current control therapy is available from previous studies. Exploiting historical evidence is useful both for reserving more resources to the arm of the new therapy and for improving the accuracy of estimates, as long as current and historical control data are sufficiently homogeneous. In this article we propose a Bayesian method for incorporating past information based on *dynamic power priors* that tunes the degree of borrowing according to a posterior measure of compatibility between current and historical control data. Frequentist Type-I error and power properties of the methods are also discussed.

**Abstract** *Nelle prove cliniche di non-inferiorità, il trattamento sperimentale è confrontato con una terapia di controllo, per la quale spesso è disponibile informazione proveniente da studi precedenti. Sfruttare tale informazione è utile per poter dedicare maggiori risorse al gruppo sperimentale e per ottenere stime più accurate delle quantità di interesse, purché i dati storici sul controllo siano sufficientemente omogenei a quelli correnti. In questo articolo proponiamo un metodo bayesiano basato su “power prior dinamiche” che consente di calibrare il livello di integrazione dell’informazione passata in funzione di una misura di compatibilità tra dati storici e correnti. Si discutono inoltre le proprietà frequentiste del metodo proposto.*

**Key words:** borrowing information, clinical trials, historical data, power prior, Type-I error.

---

Fulvio De Santis

Department of Statistics, Sapienza University of Rome, e-mail: fulvio.desantis@uniroma1.it

Stefania Gubbiotti

Department of Statistics, Sapienza University of Rome, e-mail: stefania.gubbiotti@uniroma1.it

## 1 Introduction and methodology

Let us consider a two-arms trial where an experimental drug ( $e$ ) is compared to a standard therapy, here used as control ( $c$ ). Let  $\theta_e$  and  $\theta_c$  denote the corresponding unknown probabilities of success and let  $X_e$  and  $X_c$  denote the random number of positive responses out of  $n_e$  and  $n_c$  observations in the two arms. We assume that  $X_j | \theta_j \sim \text{Bin}(n_j, \theta_j)$ ,  $j = e, c$  and that  $X_e \perp X_c | \theta_e, \theta_c$ . Non-inferiority of drug  $e$  with respect to drug  $c$  is assessed if the null hypothesis

$$H_0 : \theta_e - \theta_c \leq -\delta \quad \text{vs.} \quad H_1 : \theta_e - \theta_c > -\delta \quad (1)$$

is rejected, where  $\delta > 0$  is a selected margin. Adopting the Bayesian paradigm, we proceed as follows. We determine a credible interval  $C = [L, U]$  for  $\theta = \theta_e - \theta_c$  and we reject  $H_0$  if  $L > -\delta$ . Determination of  $C$  requires the posterior distributions of  $\theta_e$  and  $\theta_c$ . We assume that no information on  $\theta_e$  is available, whereas historical data regarding  $\theta_c$  can be retrieved. The overall procedure consists of two main steps: (i) prior construction; (ii) NI test.

(i) *Prior construction.* Recall that if  $\theta_j \sim \text{Beta}(\alpha_j, \beta_j)$ ,  $j = e, c$ , the corresponding posterior is the  $\text{Beta}(\bar{\alpha}_j, \bar{\beta}_j)$  density, where  $\bar{\alpha}_j = \alpha_j + x_j$  and  $\bar{\beta}_j = \beta_j + n_j - x_j$ ,  $j = e, c$ . Let us denote by  $\pi_0^j(\cdot)$  the non-informative  $\text{Beta}(1, 1)$  density prior for  $\theta_j$ . Given the experimental data  $x_e$  and using  $\pi_0^e(\cdot)$  we obtain the posterior  $\pi_0^e(\cdot | x_e)$  for  $\theta_e$  that is a  $\text{Beta}(1 + x_e, 1 + n_e - x_e)$  density. Conversely, we assume that a previous study provides historical data  $(n_h, x_h)$  that yield information on the control parameter  $\theta_c$ , where  $n_h$  and  $x_h$  are the size and the number of successes. As prior for  $\theta_c$  in the current experiment we then consider its posterior density given  $(n_h, x_h)$ . In order to take into account potential heterogeneity between current and historical information on  $\theta_c$ , we consider the power prior originally defined by [10] as

$$\pi_P(\theta_c | x_h) \propto \pi_0(\theta_c) \times [f(x_h | \theta_c)]^{a_0}, \quad a_0 \in [0, 1] \quad (2)$$

where  $\pi_0(\theta_c)$  is a starting prior (typically a non-informative prior),  $f(x_h | \theta_c)$  the likelihood function of  $\theta_c$  given the historical data  $x_h$  and  $a_0 \in [0, 1]$  a discount parameter. The smaller  $a_0$ , the lighter the degree of incorporation of historical information;  $a_0 = 0$  corresponds to no borrowing, whereas  $a_0 = 1$  implies full borrowing. Noting that  $[f(x_h | \theta_c)]^{a_0} \propto \theta_c^{a_0 x_h} (1 - \theta_c)^{a_0(n_h - x_h)}$  and assuming  $\pi_0(\cdot)$  to be a  $\text{Beta}(1, 1)$  density, from (2) we have that  $\pi_P(\theta_c | x_h, x_c)$  is the  $\text{Beta}(1 + a_0 x_h + x_c, 1 + a_0(n_h - x_h) + n_c - x_c)$  density. The choice of the fraction  $a_0$  is crucial in determining the impact of historical data on the analysis. In the basic definition of power priors, the tuning parameter  $a_0$  is either fixed or random, but it does not depend on the available data. The *dynamic* power prior, on the contrary, is characterized by having  $a_0$  dependent on a measure of homogeneity between historical and current control data. In a Bayesian perspective it is natural to consider a measure of the agreement between  $\pi_0^e(\cdot | x_c)$  and  $\pi_0^h(\cdot | x_h)$ , where  $\pi_0^j(\cdot | x_j)$  is the posterior density for the control parameter obtained by updating  $\pi_0^j(\cdot)$  with  $x_j$ ,  $j = h, c$ . The stronger the consensus, the larger the value of  $a_0$  in the power prior. More specifically, we construct  $a_0$  as

follows. First, let  $S(x_h)$  be a  $(1 - \gamma)$ -credible set for the control parameter built using  $\pi_0^h(\cdot|x_h)$  (historical posterior). Then, we define the dynamic fraction as

$$a_0(x_c, x_h) = \int_{S(x_h)} \pi_0^c(y|x_c) dy \tag{3}$$

i.e. the posterior probability of  $S(x_h)$  with respect the posterior density of  $\theta_c$  given the current control data. Note that the more compatible information provided by  $\pi_0^c(\cdot|x_c)$  and  $\pi_0^h(\cdot|x_h)$ , the larger  $a_0(x_c, x_h)$ . The use of a dynamic power prior has been previously considered for instance by [11] who proposes a hybrid frequentist-Bayesian approach with a p-value based adjustment of the amount of information borrowed from historical data.

(ii) *NI test.* Given  $\pi_0^e(\theta_e|x_e)$  and  $\pi_P(\theta_c|x_h, x_c)$  the lower limit  $L$  of the  $(1 - \gamma)$ -equal tails interval for  $\theta = \theta_e - \theta_c$  is simply computed via Monte Carlo. Then, if  $L > -\delta$  the null hypothesis of (1) is rejected.

In Section 2 we consider an application to assess how historical information borrowing affects the posterior probability of the hypothesis that the new treatment is non-inferior to the control. In compliance with regulatory agencies recommendations (see [1]), in Section 3 we explore the basic operating characteristics (Type-I error and power) of the proposed methodology for some selected scenarios.

The topic of our paper is related to the following research areas in Bayesian analysis of clinical trials: borrowing of historical data ([12], [14]); power priors ([4], [10], [9]); NI studies ([7], [8], [11]).

## 2 Motivating example

In this section we consider an example described in [11], where a NI study is conducted to compare a pentavalent vaccine (RotaTeq) with a placebo against Rotavirus, both administered together with routine pediatric vaccines. The data are the number of subjects in the two groups who give a positive response to vaccination. Table 1 reports current data for both experimental and control arms and historical data on the control, that are obtained by combining four different studies using a meta-analytic model (see [11] for details). Let  $\hat{\theta}_j = x_j/n_j$ ,  $j = e, c, h$  denote the response rates. Real data show that  $\hat{\theta}_c < \hat{\theta}_h$  (scenario (a)). For comparison, we also consider two fictitious historical studies (keeping the same study dimension  $n_h = 483$ ) such that (b)  $\hat{\theta}_c = \hat{\theta}_h = 0.72$ , (c)  $\hat{\theta}_c > \hat{\theta}_h = 0.68$ . In these alternative

**Table 1** Historical and current data for the RotaTeq vaccine example.

Study	Arm	$j$	$n_j$	$x_j$	$\hat{\theta}_j$
<i>Current</i>	<i>Experimental</i>	$e$	558	415	0.74
	<i>Control</i>	$c$	592	426	0.72
<i>Historical</i>	<i>Control</i>	$h$	483	367	0.76

**Table 2** Credible intervals (bounds and length) for  $\theta = \theta_e - \theta_c$  and posterior probability of  $H_1 : \theta > -\delta = -0.03$  for different choices of  $a_0$  and historical data.

		$a_0$	$L$	$U$	$U - L$	$P(H_1 x_c, x_e)$
(a) <i>Real historical data</i>	$\hat{\theta}_h > \hat{\theta}_c$	1	<b>-0.039</b>	0.050	0.089	0.944
		0.486	<b>-0.035</b>	0.058	0.093	0.964
		0	-0.028	0.075	0.103	0.982
(b) <i>Fictitious historical data</i>	$\hat{\theta}_h = \hat{\theta}_c$	1	-0.024	0.068	0.092	0.989
		0.970	-0.024	0.068	0.092	0.988
		0	-0.027	0.076	0.103	0.981
(c) <i>Fictitious historical data</i>	$\hat{\theta}_h < \hat{\theta}_c$	1	-0.006	0.085	0.091	0.999
		0.587	-0.012	0.083	0.095	0.996
		0	-0.027	0.076	0.103	0.981

setups, we obtain three different values of  $a_0$  that reflect three different levels of compatibility between historical and current control data. Table 2 illustrates the impact of the dynamic choice of  $a_0$  on the posterior probability of  $H_1$   $P(H_1|x_c, x_e)$ , in contrast with full borrowing ( $a_0 = 1$ ) and no borrowing ( $a_0 = 0$ ) of historical information (NI margin  $\delta = 0.03$ ). In case (a) historical data strengthen  $H_0$  and the larger the degree of borrowing the smaller  $P(H_1|x_c, x_e)$ . Conversely, in case (c) historical data favour  $H_1$  and  $P(H_1|x_c, x_e)$  increases with  $a_0$ . In the intermediate case (b) the high compatibility between current and historical data ( $a_0 = 0.97$ ) implies a limited effect of borrowing. Hence, the use of  $a_0(x_c, x_h)$  allows a dynamic downweighting of historical data according to their heterogeneity with current control data.

### 3 Frequentist Type-I error and power

Regulatory agencies (see for instance [1]) require that new statistical methodologies for clinical trials analysis are evaluated in terms of their frequentist properties, such as Type-I error rate  $\alpha$  and power  $\eta(\theta)$ . Table 3 reports the empirical values of  $\alpha$  and  $\eta(\theta)$  under the three scenarios (a), (b) and (c) described in the previous section. The simulation study is organized as follows:

1. Specify  $x_h, n_h, n_e, n_c, \delta, 1 - \gamma$ .
2. Fix a design value for  $\theta_c^*$  for  $\theta_c$  and generate  $M$  values  $\tilde{x}_c$  from  $Binom(n_c, \theta_c^*)$ .
3. For each  $\tilde{x}_c$  compute  $a_0(\tilde{x}_c, x_h)$  according to (3).
4. Draw  $M$  values  $\tilde{x}_e$  from  $Binom(n_e, \theta_e^*)$ , where  $\theta_e^* = \theta_c^* - \delta + \xi$ , with  $\xi = 0$  under  $H_0$  and  $\xi > 0$  under  $H_1$ .
5. Draw  $B$  values  $\tilde{\theta}_e$  and  $\tilde{\theta}_c$  from  $\pi_0^e(\cdot|x_e)$  and  $\pi_P(\cdot|x_h, x_c)$  and set  $\tilde{\theta} = \tilde{\theta}_e - \tilde{\theta}_c$ .
6. Compute  $\tilde{L}$  as the empirical  $(1 - \gamma/2)$ -quantile of the  $B$  values  $\tilde{\theta}$ .
7. Compute the fraction of  $\tilde{L} > -\delta$  and obtain the empirical Type-I error (if  $\xi = 0$ ) or the empirical power (if  $\xi > 0$ ).

**Table 3** Empirical Type-I error rates and powers for different values of  $\xi$  under scenarios (a), (b) and (c), given  $\delta = 0.03$ ,  $n_e = 558$ ,  $n_c = 592$ ,  $\theta_c^* = 0.72$ , for different levels of  $a_0$ .

(a)				(b)				(c)			
$a_0$	$\alpha$	$\xi$	$\eta(\theta)$	$a_0$	$\alpha$	$\xi$	$\eta(\theta)$	$a_0$	$\alpha$	$\xi$	$\eta(\theta)$
1	0.002	0.05	0.268	1	0.015	0.05	0.571	1	0.081	0.05	0.825
		0.1	0.967			0.1	0.996			0.1	1.000
0.486	0.019	0.05	0.392	0.931	0.020	0.05	0.562	0.587	0.065	0.05	0.673
		0.1	0.974			0.1	0.993			0.1	0.986
0	0.025	0.05	0.480	0	0.025	0.05	0.480	0	0.025	0.05	0.480
		0.1	0.975			0.1	0.975			0.1	0.975

Note that, in the partial borrowing case,  $a_0$  is computed according to step 3 and the values reported in Table 3 are the empirical medians over the  $M$  simulations. For comparison we also consider  $a_0 = 0$  and  $a_0 = 1$ . For each given combination of scenario and borrowing level, we consider two different values of  $\xi$  to show the increasing trend of the power with respect to  $\xi$ . As in the previous section, scenario (a), that is based on the pooled real data, is favorable to  $H_0$ : hence a larger  $a_0$  corresponds to smaller  $\alpha$ , but also to a lower power for each given value of  $\xi$ . Conversely, when the contrast between historical and current data goes in the opposite direction, as in scenario (c),  $H_1$  is strengthened and, therefore, both  $\alpha$  and  $\eta(\theta)$  increase with the level of discount  $a_0$ . Finally in case (b) the almost perfect compatibility between historical and current control data (i.e., the empirical median of  $a_0(\tilde{x}_c, x_h)$  is as large as 0.931) implies that partial and full borrowing are substantially equivalent in terms of Type-I error and power. Not surprisingly, they are both preferable with respect to the no borrowing case, which yields a larger empirical value of  $\alpha$  and a lower power.

## 4 Conclusions

The present article is a wholly Bayesian conversion of the hybrid frequentist-Bayes method proposed by [11]. The main features of Liu’s approach are: (i) implementation of a frequentist test for selecting among hypotheses (1); (ii) instrumental use of a dynamic power prior *only* for the selection of the amount of borrowing from historical data (no posterior analysis is considered); (iii) definition of the fraction  $a_0$  as an arbitrary function of the p-value for testing the hypothesis of equivalence between the current and historical control true response rates. Features (ii) and (iii) present some controversial aspects. Specifically, for (ii) one can object that an instrumental use of the power prior does not have a clear justification outside a Bayesian context; with respect to (iii), one can call into question Liu’s arbitrary choice of the p-value function that may yield any value of  $a_0$  in  $[0, 1]$ . For these reasons, in this paper we propose: (i) to make use of a Bayesian test of NI, based on a credible interval for  $\theta = \theta_e - \theta_c$ ; (ii) to consider a power prior to build the posterior distributions of  $\theta$  necessary for feature (i); (iii) to define a new dynamic fraction  $a_0(x_c, x_h)$  based on a



sensible measure of compatibility between historical and current data obtained from the posterior densities  $\pi_0^c(\cdot|x_c)$  and  $\pi_0^h(\cdot|x_h)$  of the control parameter.

Basic posterior and frequentist properties of our proposal have been illustrated in Sections 2 and 3, via numerical examples and simulations based on vaccine data. Preliminary results are encouraging. On the Bayesian hand, our dynamic approach allows for a sensible tuning of the amount of historical information to be incorporated in posterior analysis; on the frequentist hand, Type-I error and power seem to be adequately controlled. However, our methodology needs further investigation that we hope to deepen in the future. First of all, we intend to explore the effects of more thoughtful selections of the NI margin  $\delta$  (see, for instance, in [8], [6] and [2]). Secondly, we want to extend simulation studies in order to improve insight into Type-I error and power performances of the method. Finally, we would like to introduce sample size determination criteria to take advantage of the use of historical data in designing the experiment in order to reduce the required number of current control (see [5], [3], [13]).

## References

1. CDHR/FDA Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Guidance for industry and FDA staff. (2010)
2. CDER-CDBR/FDA Non-Inferiority clinical trials to establish effectiveness. Guidance for industry. (2016)
3. Chen, M.H., Ibrahim, J.G., Lam, P., Yu, A., Zhang, Y. Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics*, **67**, 1163–1170 (2011)
4. De Santis, F. Power priors and their use in clinical trials, *The American Statistician*, **60**(2), 122–129 (2006)
5. De Santis, F. Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society Series A*. **170**, 95–113 (2007)
6. EMA Guideline on the choice of non-inferiority margin. EMEA/CPMP/EWP/2158/99 (2005)
7. Gamalo, M.A., Wu, R., Tiwari, R.C.: Bayesian approach to noninferiority trials for proportions. *Journal of Biopharmaceutical Statistics* **21**(5), 902–919 (2011)
8. Gamalo-Siebers M., Gao A., Lakshminarayanan M., Liu G., Natanegara F., Railkar R., Schmidli H., Song G. Bayesian methods for the design and analysis of noninferiority trials, *Journal of Biopharmaceutical Statistics*. **26**(5), 823–841 (2016)
9. Gravestock, I., Held, L. Power priors based on multiple historical studies for binary outcomes. *Biometrical Journal*. **61**, 1201–1218 (2019)
10. Ibrahim J.G., Chen M.H.: Power prior distributions for regression models. *Statistical Science*. **15**, 46–60 (2000)
11. Liu, G.F.: A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint. *Pharm Stat*. **17**, 61–73 (2018)
12. Neuenschwander, B., Capkun-Niggli, G., Branson, M., Spiegelhalter, D.J. Summarizing historical information on controls in clinical trials. *Clinical Trials* **7**, 5–18 (2010)
13. Psioda, M.A., Ibrahim, J.G. Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, **20**(3), 400–415 (2019)
14. Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J.G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., Thompson, L.: Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. **13**(1), 41–54 (2014)

# Optimal credible intervals under alternative loss functions

## *Intervalli di credibilità ottimi per diverse funzioni di perdita*

Fulvio De Santis and Stefania Gubbiotti

**Abstract** This article deals with Bayesian interval estimation of the parameter of a statistical model from a decision-theoretic perspective. We consider the class of monotone loss functions, that take into account both size and posterior probability of the sets and that, under general conditions, guarantees the optimality of highest posterior probability sets. More specifically, we focus on three families of loss functions: linear, rational and exponential. Resorting to numerical examples and simulations, we examine both posterior and pre-posterior features of these choices for the Poisson-Gamma model.

**Abstract** *Questo articolo riguarda la stima intervallare bayesiana del un parametro incognito di un modello statistico in un'ottica decisionale. La classe delle funzioni di perdita monotone coinvolge sia la dimensione degli insiemi sia la loro probabilità a posteriori. Sotto condizioni generali questa classe garantisce l'ottimalità degli insiemi HPD (highest posterior density). In particolare, questo lavoro si concentra su tre famiglie di funzioni di perdita: lineare, razionale ed esponenziale. Mediante esempi numerici e simulazioni viene condotta un'analisi a posteriori e un'analisi predittiva delle caratteristiche delle diverse funzioni di perdita per il modello Poisson-Gamma.*

**Key words:** Bayesian inference, credible sets, decision theory, interval estimation, loss functions, sample size determination, set estimation.

---

Fulvio De Santis  
Department of Statistics, Sapienza University of Rome, e-mail: fulvio.desantis@uniroma1.it

Stefania Gubbiotti  
Department of Statistics, Sapienza University of Rome, e-mail: stefania.gubbiotti@uniroma1.it

### 1 Introduction

Given a parametric model,  $\{f_n(\cdot|\theta), \theta \in \Theta\}$ , let  $\pi(\theta)$  denote the prior distribution of  $\theta$ ,  $\mathbf{x}_n$  an observed sample of size  $n$  and  $\pi(\theta|\mathbf{x}_n)$  the corresponding posterior distribution. For simplicity, suppose that  $\Theta \subseteq \mathbb{R}^1$  and that  $\pi(\theta)$  is a probability density function. Assuming to be interested in set estimation of  $\theta$  from a decision theoretic perspective (see [1] and [4]), let  $\mathcal{C}$  be a class of subsets of  $\Theta$  and  $\mathbb{L}(\theta, C)$  the loss function for a generic set  $C \in \mathcal{C}$ . This approach prescribes one to select a set  $C^*$  that minimizes the posterior expected loss  $\rho(C, \mathbf{x}_n)$  as  $C$  varies in  $\mathcal{C}$ , i.e:

$$C^* = \arg \min_{C \in \mathcal{C}} \rho(C, \mathbf{x}_n), \quad \text{where} \quad \rho(C, \mathbf{x}_n) = \int_{\Theta} \mathbb{L}(\theta, C) \pi(\theta|\mathbf{x}_n) d\theta.$$

The most widely used family of losses for set estimation is defined by setting

$$\mathbb{L}(\theta, C) = \mathbb{S}[\mathcal{L}(C)] + \mathbb{I}_{\bar{C}}(\theta), \tag{1}$$

where the *size*  $\mathbb{S}(\cdot)$  is an increasing function of  $\mathcal{L}(C)$  - the Lebesgue measure of  $C$  - and  $\mathbb{I}_{\bar{C}}(\cdot)$  is the indicator function of the set  $\bar{C} = \Theta \setminus C$ . The resulting posterior expected loss of  $C \in \mathcal{C}$  is

$$\rho(C, \mathbf{x}_n) = \mathbb{S}[\mathcal{L}(C)] + 1 - \mathbb{P}(C|\mathbf{x}_n),$$

which embodies a compromise between the size of  $C$  and its posterior probability of containing  $\theta$ , denoted as  $\mathbb{P}(C|\mathbf{x}_n)$ . One important property of the class of monotone functions (see, for instance, [3]) is that, if  $\theta$  is an absolutely continuous random variable (as we assume here), optimal actions are HPD sets defined as  $C^* = \{\theta \in \Theta : \pi(\theta|\mathbf{x}_n) \geq k\}, k \geq 0\}$ . More specifically, we here assume that HPD sets are intervals  $C = [L, U]$  and that  $\mathcal{L}(C) = U - L$  is the length of  $C$ . The simplest form of loss (1) is obtained by selecting

$$\mathbb{S}_\ell[\mathcal{L}(C)] = a\mathcal{L}(C), \quad a > 0, \tag{2}$$

as size function, which yields the class of *linear* loss functions,  $a\mathcal{L}(C) + \mathbb{I}_{\bar{C}}(\theta)$ . Casella, Hwang and Robert in [2] and [3] show that, in the case of unbounded parameter space, optimal sets under the linear loss function may be dominated by unreasonable sets. For instance, in the case of the normal model  $N(\theta, \sigma^2)$  with unknown variance, the standard Student's t-interval for  $\theta$  is dominated by a set that is empty as the sample variance is sufficiently large. They also show that (under mild conditions) these kinds of problems are avoided if both the components of (1) assume values in  $[0, 1]$  or, more specifically, if  $\mathbb{S}(\cdot)$  is a nonlinear and increasing function that ranges monotonically in the unit interval and  $\lim_{A \rightarrow \emptyset} \mathbb{S}(A) = 0$  and  $\lim_{A \rightarrow \Theta} \mathbb{S}(A) = 1$ . To resolve the paradox observed in the normal model, the authors propose some nonlinear functions  $\mathbb{S}(\cdot)$ . Among these, they consider

$$\mathbb{S}_e[\mathcal{L}(C)] = 1 - e^{-\frac{a\mathcal{L}(C)^2}{2}} \quad \text{and} \quad \mathbb{S}_r[\mathcal{L}(C)] = \frac{a\mathcal{L}(C)}{a\mathcal{L}(C) + 1}, \quad a > 0 \tag{3}$$

that result in the class of *exponential* and *rational* loss functions. The posterior expected losses corresponding to the three size functions under examination in this article are then given by:

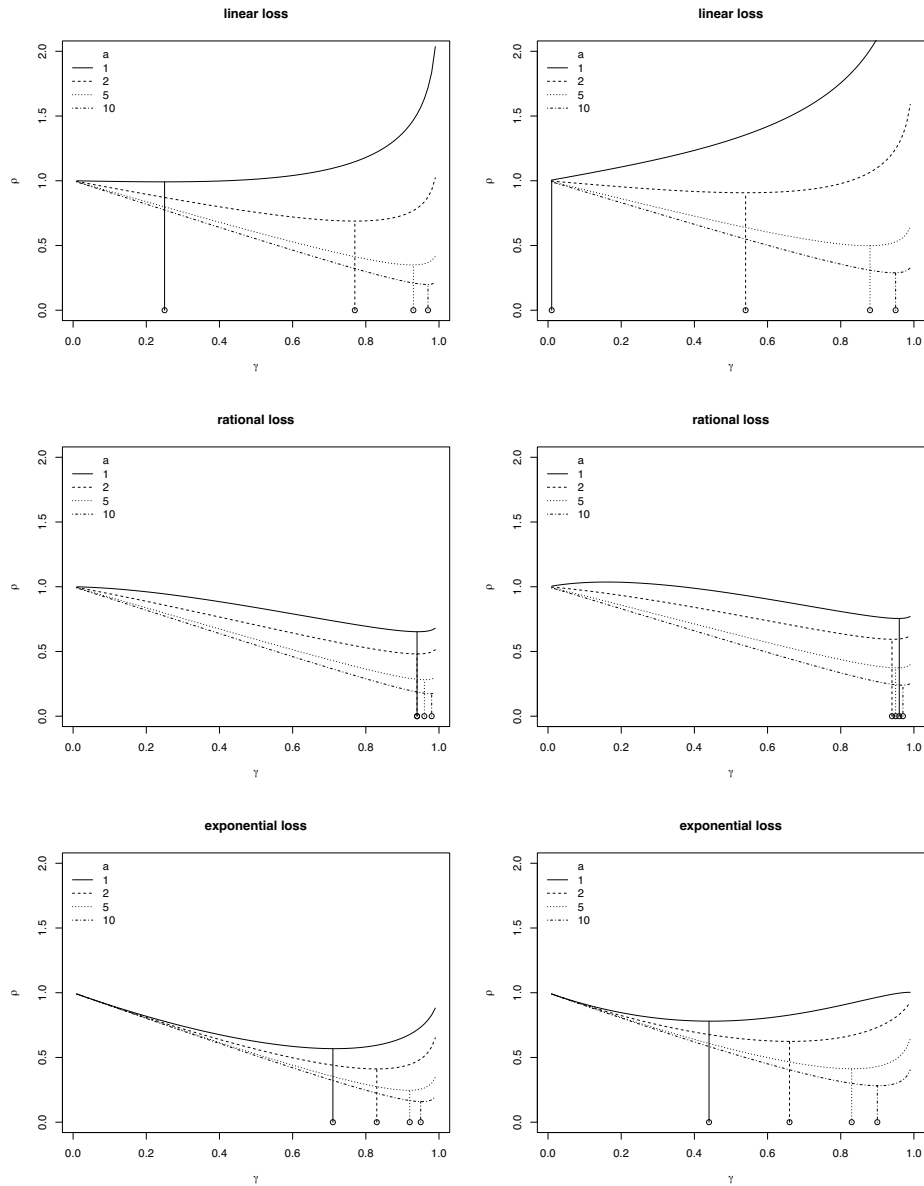
$$\rho_j(C, \mathbf{x}_n) = \mathbb{S}_j[\mathcal{L}(C)] + 1 - \mathbb{P}(C|\mathbf{x}_n), \quad j = \ell, e, r \quad (4)$$

In [2] and [3]  $\rho_e$  and  $\rho_r$  were originally introduced and motivated for the normal model. We here explore their behavior for the Poisson-Gamma model.

The article is organized as follows. In Section 2 we consider numerical examples to investigate the impact on optimal actions of the choice of the size function and of the coefficient  $a$ , which controls the degree of penalization of intervals length. In Section 3, we adopt a pre-posterior point of view and compare optimal actions to usual intervals with fixed credibility, using the three different loss functions. Section 4 contains some concluding remarks.

## 2 Posterior comparison of loss functions

Let  $X_i|\theta \sim \text{Pois}(\theta)$ ,  $i = 1, \dots, n$  (i.i.d.),  $\theta > 0$  and  $\theta \sim \text{Ga}(\alpha, \beta)$ ,  $\alpha, \beta > 0$ . Then, from standard results  $\theta|\mathbf{x}_n \sim \text{Ga}(\bar{\alpha}, \bar{\beta})$ , where  $\bar{\alpha} = \alpha + s_n$ ,  $\bar{\beta} = \beta + n$  and  $s_n = \sum_{i=1}^n x_i$ . For each loss function and for selected values of  $a$  we determine the optimal sets  $C^*$  using the following numerical procedure. We consider a grid of values for  $\gamma = \mathbb{P}(C|\mathbf{x}_n) \in (0, 1)$ . For each value of  $\gamma$  we determine  $C_\gamma$ , the HPD interval for  $\theta$  using the R function `HDInterval::hdi` and compute  $\rho_j(C_\gamma, \mathbf{x}_n)$ . Then, we select  $\gamma^*$  as the minimizer of  $\rho_j(C_\gamma, \mathbf{x}_n)$ . Figure 1 shows the plots of  $\rho_j(C_\gamma, \mathbf{x}_n)$  as functions of  $\gamma$  for Gamma posteriors of parameters  $(\bar{\alpha}, \bar{\beta}) = (6, 2)$  (left column) and  $(\bar{\alpha}, \bar{\beta}) = (14, 2)$  (right column). For each value of  $a$  the selected  $\gamma^*$  is circled. As a consequence of the mathematical structure of  $\mathbb{S}_j$ , for  $j = \ell, e$ , the larger  $a$  the smaller  $\gamma^*$  for both the linear and the exponential loss functions, whereas this is not true for the rational loss. As expected  $\rho_\ell(C_\gamma, \mathbf{x}_n)$  is highly sensitive to the values of  $a$ . Hence the range of  $\gamma^*$  is the highest among the three loss functions. Conversely, values  $\gamma^*$  for  $\rho_r(C_\gamma, \mathbf{x}_n)$  are substantially unaffected by the choice of  $a$ , that however influences the minimum values  $\rho_r(C^*, \mathbf{x}_n)$ . They invariably bring  $C^*$  with posterior probability close to the conventional level 0.95, thus revealing an excessive robustness with respect to  $a$ . Finally, the exponential loss seems to represent a sensible trade-off between the two other competitor loss functions. Table 1 reports the values of  $\mathcal{L}(C^*)$ ,  $\mathbb{P}(C^*|\mathbf{x}_n)$  and  $\rho_j(C^*, \mathbf{x}_n)$  for optimal intervals  $C^*$ . Even though the role of  $a$ , namely the coefficient that penalizes the length of the intervals, is not equivalent in the different size functions  $\mathbb{S}_j$ , a look at their effect on the resulting  $C^*$  and related quantities is still informative: the larger  $a$ , the smaller  $\mathcal{L}(C^*)$  and  $\mathbb{P}(C^*|\mathbf{x}_n)$ , the larger the corresponding  $\rho_j(C^*, \mathbf{x}_n)$ . This effect is mostly remarkable in the linear loss case. The variations in the values of  $\rho_r(C^*, \mathbf{x}_n)$  depend almost entirely on the values of  $a$ , whereas  $\rho_\ell(C^*, \mathbf{x}_n)$  and  $\rho_e(C^*, \mathbf{x}_n)$  change according to  $\mathcal{L}(C^*)$  and  $\mathbb{P}(C^*|\mathbf{x}_n)$ .



**Fig. 1** Posterior expected losses  $\rho_j(C, \mathbf{x}_n)$ ,  $j = \ell, r, e$ , as functions of  $\mathbb{P}(C|\mathbf{x}_n)$  for different values of  $a$  for Gamma posteriors of parameters  $(\bar{\alpha}, \bar{\beta}) = (6, 6)$  (left column) and  $(\bar{\alpha}, \bar{\beta}) = (14, 6)$  (right column). For each  $\rho_j$ ,  $j = \ell, r, e$ , circles denote  $\mathbb{P}(C^*|\mathbf{x}_n)$ , i.e. the posterior probabilities of optimal sets.

Optimal credible intervals under alternative loss functions

Loss	$a$	$(\bar{\alpha}, \bar{\beta}) = (6, 6)$			$(\bar{\alpha}, \bar{\beta}) = (14, 6)$		
		$\mathcal{L}(C^*)$	$\mathbb{P}(C^* \mathbf{x}_n)$	$\rho_j(C^*, \mathbf{x}_n)$	$\mathcal{L}(C^*)$	$\mathbb{P}(C^* \mathbf{x}_n)$	$\rho_j(C^*, \mathbf{x}_n)$
<i>linear</i>	1.0	0.242	0.250	0.992	0.015	0.010	1.005
	0.5	0.917	0.770	0.689	0.895	0.540	0.907
	0.2	1.399	0.930	0.350	1.890	0.880	0.498
	0.1	1.689	0.970	0.199	2.390	0.950	0.289
<i>rational</i>	1.0	1.454	0.940	0.652	2.507	0.960	0.755
	0.5	1.454	0.940	0.481	2.292	0.940	0.594
	0.2	1.594	0.960	0.282	2.390	0.950	0.373
	0.1	1.817	0.980	0.174	2.652	0.970	0.240
<i>exponential</i>	1.0	0.807	0.710	0.568	0.706	0.440	0.780
	0.5	1.051	0.830	0.411	1.156	0.660	0.624
	0.2	1.350	0.920	0.247	1.667	0.830	0.412
	0.1	1.518	0.950	0.159	2.001	0.900	0.281

**Table 1** Bounds, length, posterior probability and posterior expected loss for  $C^*$  under the three loss functions for selected values of  $a$ .

$\theta_d$	$n$	$\mathbb{E}(\mathcal{L})$	$\mathbb{E}(\mathbb{S})$	$\mathbb{E}(\rho_e)$	$\mathbb{E}[\mathbb{P}(\cdot \mathbf{x}_n)]$
(i)					
2	4	1.142	0.278	0.575	0.703
	10	1.072	0.250	0.430	0.820
	100	0.606	0.088	0.118	0.970
10	4	0.807	0.151	0.860	0.291
	10	1.019	0.229	0.775	0.454
	100	0.966	0.208	0.328	0.880
(ii)					
2	4	2.178	0.687	0.737	0.950
	10	1.574	0.461	0.511	0.950
	100	0.548	0.072	0.122	0.950
10	4	4.290	0.988	1.038	0.950
	10	3.319	0.934	0.984	0.950
	100	1.217	0.310	0.360	0.950
(iii)					
2	4	2.594	0.787	0.837	0.950
	10	1.716	0.519	0.569	0.950
	100	0.553	0.074	0.124	0.950
10	4	6.128	1.000	1.050	0.950
	10	3.904	0.976	1.026	0.950
	100	1.239	0.319	0.369	0.950

**Table 2** Monte Carlo approximations of predictive expectations of length, size, posterior expected loss  $\rho_e$  and probability of (i)  $C^*$  and (ii)  $C_\gamma^*$  with an informative prior  $\text{Gamma}(4, 2)$  and (iii)  $C_\gamma^*$  with a non-informative prior  $\text{Gamma}(0, 0)$ , for different values of  $n$  and  $\theta_d$ .

### 3 Predictive comparison of optimal sets under exponential loss

In this section we focus on credible intervals optimal that are under the exponential loss, which has shown the most promising results in the explorative analysis of Section 2. For the sake of brevity we select the case  $a = 0.5$ . We consider a pre-posterior comparison between (i) optimal credible sets  $C^*$  and two conventional optimal cred-

ible intervals: HPD intervals  $C_\gamma^*$  of fixed credibility  $\gamma = 0.95$ , respectively obtained assuming (ii) the same prior  $\text{Gamma}(4, 2)$  which yields  $C^*$ ; (iii) the non-informative  $\text{Gamma}(0, 0)$ . In Table 2 the three intervals are compared in terms of the predictive expected values  $\mathbb{E}(\cdot)$  of their length, size function, posterior expected loss and probability. For simplicity, as predictive distribution we assume the sampling distribution  $f_n(\cdot|\theta_d)$ , where  $\theta_d$  is a design value. The simulation steps are the following: draw  $M$  samples of size  $n$  from  $f_n(\cdot|\theta_d)$ ; for each sample repeat the minimization described in Section 2 to derive  $C^*$  and compute  $C_\gamma^*$ ; for each of the three intervals determine  $\mathcal{L}$ ,  $\mathbb{S}$ ,  $\rho_e$  and  $\mathbb{P}(\cdot|\mathbf{x}_n)$ ; compute the Monte Carlo means of  $\mathcal{L}$ ,  $\mathbb{S}$ ,  $\rho_e$  and  $\mathbb{P}(\cdot|\mathbf{x}_n)$ .

As a first comment note that, by construction, the values of  $\mathbb{E}[\mathbb{P}(\cdot|\mathbf{x}_n)]$  are variable for (i) and fixed for (ii) and (iii). As expected,  $C^*$  outperforms  $C_\gamma^*$  (ii) and (iii) in terms of  $\rho_e$ . In the cases of low values for  $\mathbb{E}[\mathbb{P}(C^*|\mathbf{x}_n)]$  a gain in terms of expected length (and size function) is observed. In addition, whereas for the smallest value of  $\theta_d$  the optimal set  $C^*$  guarantees a sufficiently large expected posterior probability for all sample sizes, for  $\theta_d = 10$  the posterior probability of  $C^*$  may be excessively small (e.g. 0.291 for  $n = 4$ ), unless the sample size is sufficiently large (e.g. 0.888 for  $n = 100$ ). Finally, note the uniformly better performance of  $C_\gamma^*$  (ii) with respect to  $C_\gamma^*$  (iii) and their substantial equivalence for  $n = 100$ .

## 4 Conclusions

In this work in addition to the most widely used linear loss, we examine two alternative monotone loss functions for set estimation, previously proposed for the normal model. Monotonicity guarantees that, under mild conditions, optimal credible intervals are HPD sets. In the preliminary numerical comparisons of Section 2, the exponential loss seems to have an intermediate behavior with respect to the linear loss (highly sensitive to the choice of  $a$ ) and the rational loss (excessively robust with respect to  $a$ ). Pre-posterior analysis of  $\rho_e$  suggests that, in order to obtain optimal sets with sensible posterior probability, attention has to be paid to the selection of the sample size.

## References

1. Berger J. O. Statistical decision theory and Bayesian analysis. New York: Springer-Verlag. Chicago (1985)
2. Casella G., Hwang J.T.G., Robert C. A paradox in decision-theoretic interval estimation. *Statistica Sinica*, **3**(1), 141–155 (1993a)
3. Casella G., Hwang J.T.G., Robert C. Loss function for set estimation. In *Statistical Decision Theory and related topics V*, J.O. Berger and S.S. Gupta (Eds.), 237–252. Springer-Verlag, New York (1993b)
4. Robert C. The Bayesian choice: from decision-theoretic foundations to computational implementation Springer Science & Business Media (2007)

# Statistical Learning for Credit Risk Modelling

## *Statistical Learning per Modelli di Rischio del credito*

Veronica Bacino and Alessio Zoccarato and Caterina Liberati and Matteo Borrotti

**Abstract** The objective of credit scoring is to develop accurate rule of classification that aids to distinguish between good and bad clients. In this context, also Statistical Learning (SL) techniques have been explored, for building models that estimate the clients' probability of insolvency. Although there are some encouraging results in literature, two main issues makes this classification task very hard: (i) high imbalance ratio between the two groups in the target variable and (ii) the effect of hyperparameter settings on overall performance. In this work, Bayesian Optimization (BO) is used to optimize the hyperparameters of a cost sensitive eXtreme Gradient Boosting (XGBoost) model. Experimental results reveal that the proposed solution is a promising starting point for future development.

**Abstract** *L'obiettivo del credit scoring è quello di sviluppare regole di classificazione accurate che aiutino a distinguere tra clienti buoni e cattivi. In questo contesto sono state esplorate anche tecniche di Statistical Learning (SL), per costruire modelli che stimino la probabilità di insolvenza dei clienti. Sebbene ci siano alcuni risultati incoraggianti in letteratura, due questioni principali rendono questo compito di classificazione molto difficile: (i) un elevato rapporto di squilibrio tra i due gruppi nella variabile target e (ii) l'effetto delle impostazioni degli iperparametri sulle prestazioni complessive. In questo lavoro, l'ottimizzazione bayesiana (BO)*

---

Veronica Bacino

University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano (MI), e-mail: v.bacino1@campus.unimib.it

Alessio Zoccarato

University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano (MI), e-mail: a.zoccarato2@campus.unimib.it

Caterina Liberati

University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano (MI), e-mail: caterina.liberati@unimib.it

Matteo Borrotti

University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano (MI), e-mail: matteo.borrotti@unimib.it



*viene utilizzata per ottimizzare gli iperparametri di un modello eXtreme Gradient Boosting (XGBoost) sensibile ai costi. I risultati sperimentali rivelano che la soluzione proposta è un punto di partenza promettente per lo sviluppo futuro.*

**Key words:** Credit risk modeling, Bayesian optimization, XGBoost, cost sensitive learning

## 1 Introduction

The objective of quantitative Credit Scoring (CS) is to develop accurate models that can distinguish between good and bad applicants [2]. In particular CS aims to estimate the probability that an applicant will be able to pay off the debit taken out with the bank. CS models, especially those for screening individuals, are primarily based on financial/economic ratios computed on the client banking account which are considered relevant information and are used to predict delinquency behaviour. We refer to customer demographics (as job position, profession, location) or personal credit histories including past borrowing and repaying actions as *input variables*.

Based on such data, a variety of techniques have been applied in such predictive learning [2, 6, 8, 9, 10] with different success. As pointed out by Xia et al. (2017) [13], ensemble classifiers perform better compared to single classifiers. This is also justified in accordance with the “no free lunch theorem” [12]. So in the context of CS it means: information derived from different banks have their own properties (i.e. sizes, data structures, and features), a single classification algorithm cannot solve all problems effectively. Consequently, applying ensemble methods is necessary in credit risk field. However, modeling is not the only matter: some issues are hidden in the type of data itself. The majority of the credit applicants are good users, the residuals (a very tiny percentage of the instances) are bad, that leads to a severe data imbalance ratio. In addition, imbalanced datasets come with certain challenges for the construction of a classification model. One common approach is to oversampling or undersampling the target variable [7]. Another issue is related to the number of hyper-parameters that could be tuned on recent ensemble algorithms [13]. Performance can be influenced from the initial hyper-parameters settings. This problem has been called “...search space odyssey...” by Greff et al. (2017) [5].

In this work, we investigate the performance of Bayesian Optimization (BO) [1] and eXtreme Gradient Boosting (XGBoost) algorithm [3] together with a cost sensitive learning approach [4] for imbalanced data for developing a credit scoring model.

## 2 Methodology

eXtreme Gradient Boosting (XGBoost) algorithm was recently proposed by Chen et al. (2016) [3]. XGBoost optimizes the objective function and its estimation. The fast, efficient, and scalable system achieves promising results on numerous standard classification benchmarks. XGBoost combines a series of weak base learners, which are normally regression trees, into a strong one. The weak learner herein refers to a model that only performs slightly better than a random guess. Boosting fits additive base learners to minimize the loss function provided. Loss function measures how well the model fits the current data. The process of boosting continues until the loss function reduction becomes limited. For a more detailed description see Chen et al. (2016) [3].

XGBoost is characterized by several hyper-parameters which dramatically influence the performance of models. Therefore, careful tuning of these hyper-parameters, *i.e.*, hyper-parameter optimization, is important. Bayesian Optimization is a sample-efficient strategy for global optimization of black-box, expensive and multi-extremal functions, traditionally constrained to over a box-bounded search space  $\Omega: \min_{\theta \in \Omega} g(\theta)$ .

BO is based on two key components: a *probabilistic surrogate model* (*i.e.* Gaussian Process [11]) of the objective function  $g(\theta)$  in order to provide an estimate of  $g(\theta), \forall \theta \in \Omega$ , along with a measure of uncertainty about such an estimate and an *acquisition function* that is based on the current approximation of  $g(\theta)$ . The optimization of the acquisition function allows to select the next promising  $\theta'$  where to evaluate the objective function. The observed value,  $g(\theta')$  (or  $g(\theta') + \varepsilon$  in the case that the objective function is also *noisy*), is then used to update the probabilistic model approximating  $g(\theta)$ , and the process is iterated until a given termination criteria is reached (*e.g.*, a maximum number of function evaluations). One of the most widely used acquisition functions is Upper Confidence Bound (UCB) that manages exploration—exploitation by being optimistic in the face of uncertainty. Several acquisition functions have proposed - an overview is provided in Archetti et al. (2019) [1] - each one offering a different mechanism to balance the exploitation-exploration trade-off.

XGBoost classifier is typically evaluated by estimating its error rate (or equivalently, the classification accuracy) on the test data. Usually, classifiers are designed to minimize the number of errors (incorrect classifications) made. When misclassification costs vary between classes as in credit scoring, this approach is not suitable. A possible solution is to balance the classes according to their costs re-weighting the training examples in proportion to their costs [4].

## 3 Data

One real world dataset is adopted herein to evaluate the predictive performance of the proposed solution. The sample is composed by 7500 individuals that applied for

a bank loan since June 2015 until February 2020. Instead of using standard economic/financial variables referred to the clients, we monitored their transactional data <sup>1</sup>. According to the consumer credit regulation, we computed our target variable (Y) as a dummy, checking the clients behaviour at the end of 12 months after the loan acquisition. Specifically, we labelled an applicant not creditworthy (Y=1) if she/he had at least three installments to repay still, otherwise we labelled her/him creditworthy (Y=0). As expected, just the 2% of the total instances belong to the first group, most of the client (98% ) are creditworthy, given our definition.

The input variables of our model are 83 and have different metrics: they are dummy (28), counting (16), numerical (39). They have been computed in order to investigate different aspects of the financial behaviour of the customers. More in detail, 27 variables are related to the capacity of the client to have positive cash flow (Capacity), 23 to the client reliability (Reliability), 13 to the variety of banking payments different from cash (Bank intensity), 2 to presence of life insurances (Protection seek) and 16 related to the planning behavior respects to the expenditures (Lending behavior). No missing values is present. To protect the confidentiality of data, the input variable names and values have been changed to meaningless symbolic data.

## 4 Results

On our credit risk model, Bayesian Optimization is used to optimize a cost sensitive learning version of the XGBoost algorithm, from now on BO\_costXGBoost, on which the balance of creditworthy and not creditworthy is adjust by a specific weight sets as the ratio between the two class labels. The hyper-parameters involved in the optimization process are summarized in Table 1.

**Table 1** XGBoost hyper-parameters involved in the Bayesian Optimization. For a detailed description of hyper-parameters meaning please refers to [www.xgboost.ai](http://www.xgboost.ai).

Hyperparameter	Values
eta	{0.01, 0.05, 0.1, 0.3}
max.depth	{1, 3, 5, 7}
min_child_weight	{1, 3, 5, 7}
subsample	{0.5, 0.8, 1 }

The acquisition function is set as the Upper Confidence Bound (UCB). A 5-folds cross validation technique is used to ensure robust results. The evaluation metrics for validation data is *Area Under the Curve* (AUC). The best configuration is eta = 0.05, max.depth = 2, min\_child\_weight = 3 and subsample = 0.51 and it reaches an

<sup>1</sup> The monitoring of the clients transactional data has been realized according to Open Banking regulation

AUC value of 0.833. Best configuration is then retrained on all data available on the training set.

Two approaches are compared against the proposed solution: a standard version of the XGBoost (defaultXGBoost) and a XGBoost with a cost sensitive learning approach (costXGBoost). Both approach are deployed with default hyper-parameter settings (see `xgboost` R package). Both solutions are trained on the training set. Table 2 shows the final comparison made on the test set. The best accuracy (*i.e.* number of correct predictions divided by the total number of predictions) is reached by defaultXGBoost. However, this result can be dangerously misleading on imbalanced classification modeling problems. In fact, on credit risk model the cost of misclassifying ‘bad’ objects as ‘good’ is much larger than the cost of misclassifying ‘good’ objects as ‘bad. More formally, credit risk models should minimize the error of type II. For this reason, a better performance indicator is *recall* or *sensitivity*. Considering this indicator, the best approach is BO\_costXGBoost, that reaches a value of 0.605.

**Table 2** Performance metrics.

	<b>Recall</b>	<b>F1-score</b>	<b>Type II error</b>	<b>Accuracy</b>
defaultXGBoost	0.105	<b>0.178</b>	0.895	<b>0.975</b>
costXGBoost	0.105	0.131	0.895	0.965
BO_costXGBoost	<b>0.605</b>	0.163	<b>0.395</b>	0.843

## 5 Conclusions

This paper explores the usage of supervised machine learning techniques in the context of credit risk. In our study, we compared three approaches: a default version of the eXtreme Gradient Boosting (XGBoost) model, a cost sensitive XGBoost with default hyper-parameter settings and a cost sensitive XGBoost with hyper-parameters tuned by means of Bayesian Optimization (BO). Therefore, the proposed XGBoost-based credit scoring model comprises two main elements: the Bayesian Optimization (BO) approach used to select the best hyper-parameter settings in order to reduce the effect of random settings and a cost-sensitive loss function that it is use to be more close to the real world cost distribution in credit scoring. Experimental results show that the proposed solution is the best approach while considering error of type II and sensitivity as the main performance indicators. This is justify by the fact that in credit risk model the cost of misclassifying ‘bad’ objects as ‘good’ is much larger than the cost of misclassifying ‘good’ objects as ‘bad. The two other solutions suffer of a common problem if unbalanced classification task is considered. New observations are mainly predicted as the majority class on the training set. In fact, in both cases the number of observations predicted as creditworthy al-

most corresponds to the cardinality of the test set.

Despite the promising results, some future works are needed to improve the proposed solution. First of all, a comparison of acquisition functions should be done while using the Bayesian Optimization. Two possible acquisition functions are Expected Improvement (EI) and Maximum Probability of Improvement (MPI). A more carefully calibration of the probability cut-off used to classify a new applicant as not creditworthy should be done in order to improve the tradeoff between sensitivity and specificity. Last point, it is the investigation of hyper-parameter optimization and cost sensitive learning on statistical learning approaches.

**Acknowledgements** We greatly acknowledge the DEMS Data Science Lab for supporting this work by providing computational resources.

## References

1. Archetti, K.O., Candelieri, A.: Bayesian Optimization and Data Science. Springer (2019).
2. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: Benchmarking state-of-the-art classification algorithms for credit scoring. *J. Oper. Res. Soc.* **54**, 627–635 (2003).
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. arXiv arXiv:1603.02754, 1–13 (2016).
4. Elkan, C.: The Foundations of cost-sensitive learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pp. 973–978 (2001).
5. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., Schmidhuber, J.: LSTM: A Search Space Odyssey. *IEEE Trans. Neural Networks Learn. Syst.* **28**(10), 2222–2232 (2017).
6. Hand, D. J., Henley, W. E.: Statistical classification methods in consumer credit scoring. *J. R. Stat. Soc.* **160**(3), 523–541 (1997).
7. He, H., Zhang, W., Zhang, S.: A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Syst. Appl.* **98**(2018), 105–117 (2018).
8. Huang, C. L., Chen, M. C., Wang, C. J.: Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl.* **33**(4), 847–856 (2007).
9. Li, X., Ying, W., Tuo, J., Li, B.: Applications of classification trees to consumer credit scoring methods in commercial banks. In: Proceedings of IEEE international conference on systems, man and cybernetics, pp. 4112–4117. IEEE, New Jersey (2004).
10. West, D.: Neural network credit scoring models. *Comp. Oper. Res.*, **27**(11), 1131–1152 (2000).
11. Williams, C.K.I. and Rasmussen, C.E.: Gaussian processes for machine learning. MIT Press (2006).
12. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, **1**(1), 67–82 (1997).
13. Xiaa, Y., Liua, C., Lib, Y., Liua, N.: A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring, *Expert Syst. Appl.* **78**(2017), 225–241 (2017).

# Evaluating Heterogeneity of Agreement with Strong Prior Information

## *Valutare l'Eterogeneità di Concordanza in Presenza di Forte Informazione a Priori*

Federico M. Stefanini

**Abstract** Fleiss' Kappa is a statistic typically calculated to quantify the degree of agreement among raters. Starting from the reformulation of Kappa in terms of data generating process and from plausible assumptions in the considered medical context, the number of reports that a Bayesian Network has to classify in order to obtain reasonable statistical power is found. Estimates obtained by Monte Carlo simulation suggest that samples should be at least of size 40, and that 80 or more case reports for each disease class are needed to reach enough statistical power on the interval  $\tilde{\pi}_z \in [0.5, 0.99]$  with a difference as small as  $\delta = 0.15$  or less.

**Abstract** *Il Kappa di Fleiss è una statistica tipicamente calcolata per quantificare la concordanza tra valutatori. Partendo dalla riformulazione in termini di processo generatore dei dati di concordanza e da alcune plausibili assunzioni nel contesto medico considerato, viene valutato il numero di report che una rete Bayesiana deve classificare per ottenere una potenza statistica ragionevolmente elevata. Le stime ottenute via simulazione Monte Carlo suggeriscono che i campioni dovrebbero essere di almeno di dimensione 40, e che 80 o più report per ogni classe di malattia sono necessari per raggiungere sufficiente potenza statistica nell'intervallo  $\tilde{\pi}_z \in [0.5, 0.99]$  con una differenza piccola quanto  $\delta = 0.15$  o inferiore.*

**Key words:** Fleiss' Kappa, Bayesian test, raters

## 1 A Generative Model Behind Fleiss' Kappa

Fleiss' Kappa [2] is a statistic often calculated to assess the agreement when several nominal classes are considered by a fixed number of raters. It is an improvement upon the joint probability of agreement because it takes into account the expected

---

Dipartimento di Statistica, Informatica, Applicazioni, Università di Firenze, viale Morgagni 59, I-50134 Firenze, Italy, e-mail: federico.stefanini@unifi.it  
orcid: 0000-0003-4248-6275

amount of agreement due to chance. A null hypothesis is often formulated to test if agreement between raters is significantly better than what is expected by chance. In this work,  $n_I$  physicians are considered as raters, but one further artificial rater is also included: a Bayesian Network for Auscultation (hereafter BN4A). While some degree of agreement is expected between pairs of physicians, less clear are expectations on the pair physician-BN4A.

Versions of Kappa statistics have been studied by several authors, but here just two references belonging to different statistical paradigms are mentioned. In the frequentist work [3], the coverage probability of the asymptotic confidence interval for Fleiss'  $K$  was compared with the coverage probability based on bootstrap re-sampling. The Bayesian Monte Carlo approach described in [1] is computationally simple and it may take into account prior information using Dirichlet distributions to represent the starting belief of an expert, as elaborated below.

Let  $i$  be the index for physicians,  $i \in \{1, 2, \dots, n_I\}$ , stating the disease class  $k \in \{1, 2, \dots, n_K\}$  of case report  $j \in \{1, 2, \dots, n_J\}$ . Let  $Z_j$  be the actual disease class of case report  $j$ , with sample space  $\Omega_Z = \{1, 2, \dots, n_K\}$  and  $Y_{i,j}$  the class assigned by physician  $i$ , with  $\Omega_Y = \Omega_Z$ . Conditionally on  $Z_j = z$ , scores assigned by physicians are assumed to be independent and with the same vector of probability values  $\pi_z = (\pi_{z,1}, \pi_{z,2}, \dots, \pi_{z,n_K})$ , because they all share advanced training in the problem domain, thus  $\pi_{z,y} = P[Y_{i,j} = y | Z_j = z] \quad \forall j$ .

The random variable  $B_j$ , with  $\Omega_B = \Omega_Z$ , represents the disease class assigned to case report  $j$  by BN4A, and the vector of parameters  $\theta_z = (\theta_{z,1}, \dots, \theta_{z,n_K})$  is made by elements  $\theta_{z,b} = P[B_j = b | Z_j = z]$ . In Figure 1, two DAGs are shown representing the reference pairs to be compared, that is agreement between two physicians (left) and agreement between one physician and BN4A (right).

The probability of agreement between two physicians is obtained by exploiting conditional independence relationships in Figure 1, right panel:

$$P[Y_1 = Y_2] = \sum_y \sum_z P[Y_1 = y | Z = z] \cdot P[Y_2 = y | Z = z] \cdot P[Z = z]$$

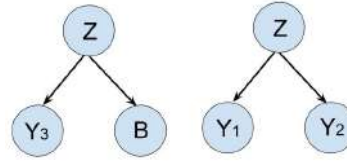
and for physician-BN4A pair (Figure 1, left panel) the probability value is:

$$P[Y = B] = \sum_y \sum_z P[Y = y | Z = z] \cdot P[B = y | Z = z] \cdot P[Z = z].$$

Under the assumption that  $P[Z = z] = 1/n_K$ , values of Fleiss'  $K$  are obtained by correcting the above probability values for the amount of agreement expected by chance:  $K_{P,P} = (P[Y_1 = Y_2] - \sum_z (1/n_K)^2) / (1 - \sum_z (1/n_K)^2)$  and  $K_{B,P} = (P[Y = B] - \sum_z (1/n_K)^2) / (1 - \sum_z (1/n_K)^2)$ , for BN4A-physician pair.

The Multinomial-Dirichlet conjugate model is a natural choice for the two types of raters,  $\pi_z \sim \text{Dirichlet}(\alpha_{P,z})$ ,  $\theta_z \sim \text{Dirichlet}(\alpha_{B,z})$ , with  $\alpha_{P,z}, \alpha_{B,z}$ ,  $\forall z \in \Omega_Z$  the hyperparameters. Final distributions of parameters, given class  $z$ , are defined by hyperparameters  $\alpha_{P,z}^* = \alpha_{P,z} + (n_{P,z,1}, \dots, n_{P,z,n_K})$  and  $\alpha_{B,z}^* = \alpha_{B,z} + (n_{B,z,1}, \dots, n_{B,z,n_K})$ , where the vector of counts  $(n_{P,z,1}, \dots, n_{P,z,r}, \dots, n_{P,z,n_K})$  collects assessments over case reports whose actual disease class is  $z$ , with all physicians included in the study;

**Fig. 1** Generative models with unknown disease class  $Z$ : disease classes  $Y_1, Y_2$  and  $Y_3$  are assigned by physicians, while class  $B$  is assigned by the BN4A rater (left).



the vector of counts generated by BN4A is  $(n_{B,z,1}, \dots, n_{B,z,r}, \dots, n_{B,z,n_K})$  when the true class is  $z$ .

The statistical hypothesis of interest  $H_0 : K_{B,P} \geq K_{P,P}$  states that the agreement between physician and BN4A,  $K_{B,P}$ , is equal or larger than the agreement between two physicians,  $K_{P,P}$ . After collecting data  $\mathcal{D}$ , the final probability of the null hypothesis  $P[K_{B,P} - K_{P,P} \geq 0 \mid \mathcal{D}]$  may be calculated from the posterior distribution of parameters in the two classification models, one for each type of pair (Figure 1), thus the null hypothesis is rejected if the final probability is small, say if  $P[H_0 \mid \mathcal{D}] \leq 0.05$ .

The above approach has a drawback due to the difficulty of eliciting the degree of belief about model parameters [3] with relation to plausible values of  $K_{B,P}$  and  $K_{P,P}$ . Moreover, in the considered medical context, further assumptions may be exploited because they stem from strong belief of experts, as described in the next section.

## 2 Scoring Synthetic Case Reports in Pulmonology: Which Sample Size?

We consider a classification task in pulmonology where  $n_K = 4$  disease classes are considered. A collection of synthetic case reports is created by top pulmonologists not included in the rating experiment, therefore the actual class  $z_j^*$  of case report  $j$  is known; furthermore, the number of case reports belonging to each disease class is constant; for example, if  $n_J = 20$  case reports then 5 case reports belong to each disease class. Lastly, no case report belongs to the "Doctor House" challenging class of case reports, where peculiar patient's configurations, joint with very unlikely expositions and pathogens, typically make the work of physicians very hard. Under this setup, there is no need to learn the distribution of  $Z$ , and, more important, the analysis may be directed towards comparisons of  $\pi_z$  and  $\theta_z$ , that is to classification probabilities, instead of comparing  $K_{P,P}$  with  $K_{B,P}$ :

It is worth noting that the comparison may be focused on the probability of picking the right disease class, say  $\tilde{\pi}_z$  for physicians and  $\tilde{\theta}_z$  for BN4A, because differences of allocation to wrong disease classes are not of primary interest in our context; a reduction in the number of parameters follows from the aggregation of disease class into "right" and "wrong" for each actual disease class  $z$ .

A total of  $n_I = 20$  physicians are considered and we assume that in our study such sample size can't be increased further. Exchangeability of physicians is also assumed after considering that they share long training and years of practice in this



**Table 1** Estimated probability of rejecting the homogeneity hypothesis for an increasing number of case reports processed by BN4A (first five columns right); on the left, the most interesting probability values of correct classification (physicians and BN4A) are shown. Labels ss5, ss10, ..., ss80, indicate the number of case report classified by BN4A in each disease class.

$\tilde{\pi}_z$	$-\delta$	$\tilde{\theta}_z$	ss5	ss10	ss20	ss40	ss80
0.99	0	0.99	0.048	0.038	0.007	0.003	0.001
0.90	0	0.90	0.082	0.059	0.049	0.044	0.036
0.80	0	0.80	0.099	0.113	0.088	0.066	0.051
0.70	0	0.70	0.211	0.133	0.102	0.075	0.059
0.60	0	0.60	0.189	0.157	0.107	0.079	0.065
0.50	0	0.50	0.232	0.162	0.111	0.084	0.064
0.99	-0.2	0.79	0.686	0.744	0.862	0.969	0.997
0.90	-0.2	0.70	0.468	0.545	0.681	0.807	0.930
0.80	-0.2	0.60	0.383	0.481	0.595	0.730	0.856
0.70	-0.2	0.50	0.444	0.440	0.549	0.680	0.811
0.60	-0.2	0.40	0.374	0.444	0.522	0.657	0.795
0.50	-0.2	0.30	0.385	0.435	0.527	0.667	0.820

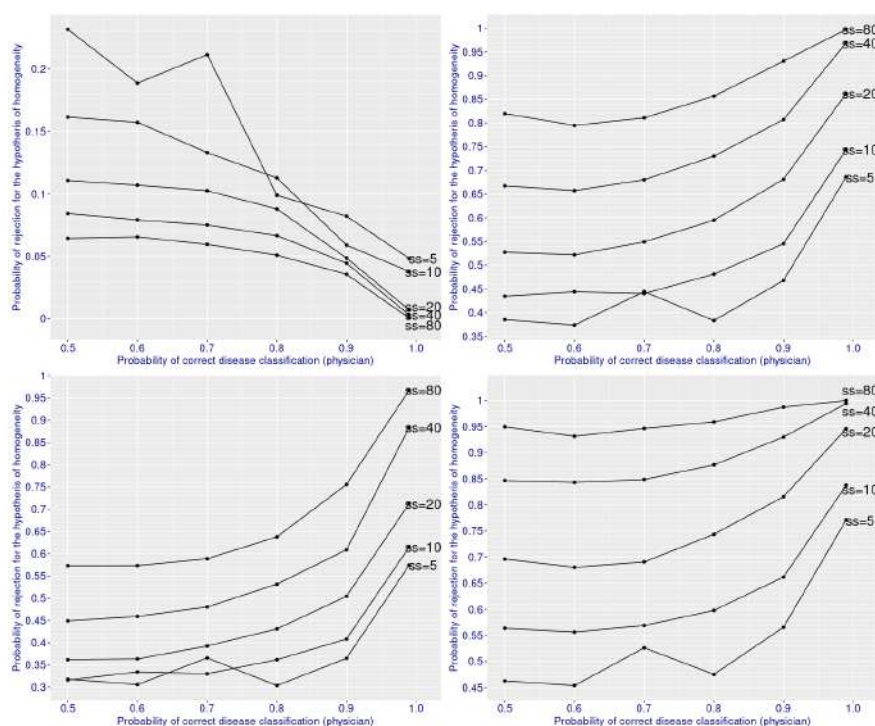
field, thus potential differences are immaterial. The key question deals with BN4A, the artificial rater: how many case reports are required in order to detect (meaningful) differences of classification probability with respect to those shown by physicians?

A Monte Carlo study was performed by simulating  $3 \times 10^6$  datasets for each pair of hypothesized parameters, the first made by  $\tilde{\pi}_z \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$  and the second  $\tilde{\theta}_z = \tilde{\pi}_z - \delta$ , where  $\delta \in \{0, 0.15, 0.2, 0.25\}$ . In Table 1, the most interesting pairs of parameters are shown, and columns on the right differ for an increasing numbers of case reports (5, 10, 20, 40, 80) processed by BN4A. The Table shows the probability of rejecting the homogeneity hypothesis of BN4A and physicians for an increasing number of case reports processed by BN4A.

The Dirichlet conjugate model presented in Section 1 was exploited, with  $n_K = 2$ , so that after collecting data  $\mathcal{D}_P$  from 20 physicians on  $n_J = 20$  reports, initial hyperparameters  $\alpha_{P,z} = (1, 1)$  for each disease class  $z$  were updated into the final value  $\alpha_{P,z}^*$ . Bayes Factor (BF) of two models was calculated, where the first states homogeneity and the second heterogeneity of BN4A. It is a ratio between the BetaBinomial integrated likelihood with hyperparameters taken from the posterior distribution of physicians and the BetaBinomial integrated likelihood with hyperparameters  $\alpha_{B,z} = (1, 1)$ , i.e. constant-non informative on the parameter space.

Results of the simulation when the difference of parameters is null (homogeneity) show that BN4A is wrongly considered heterogeneous with respect to physicians with estimated probability less than 0.10 only if 40 or more reports are processed by BN4A (Table 1, first 6 rows, column SS40; Figure 2, top left) for each disease class. When the difference  $\delta = 0.20$  (Table 1, last 6 rows, Figure 2, top right) then 40 reports suffice if the probability  $\tilde{\pi}_z = 0.90$  or above, but if the correct classification happens with smaller probability for physicians then the quite large difference

## Evaluating the heterogeneity of agreement

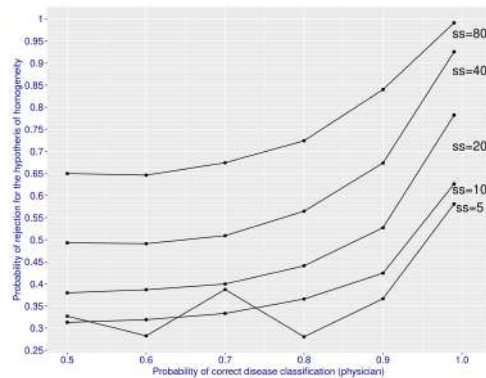


**Fig. 2** Probability of rejecting the hypothesis of homogeneity if it holds ( $\delta = 0$ , top left), if  $\delta = 0.20$  (top right), if  $\delta = 0.15$  (bottom left), if  $\delta = 0.25$  (bottom right). A label like  $ss = 40$  indicates that a sample of 40 case reports is rated by BN4A in each actual disease class.

is detected with decreasing probability towards 0.66. In Figure 2, bottom left, the assumed  $\delta = 0.15$  makes detection of heterogeneity harder. The curve of a sample size equal to 80 closely resembles to the curve for 40 case reports if  $\delta = 0.20$ . If the target difference is  $\delta = 0.15$  then more than 80 case reports should be processed by BN4A to reach reasonable values of testing power. In Figure 2, bottom right, The curve for a sample size equal to 40 stays above 0.80 whatever the value of  $\tilde{\pi}_c \in [0.5, 0.99]$  if  $\delta = 0.25$ : this is the size recommended for such quite large difference in the probability of a correct classification. A general feature of all curves with  $ss = 5$  in Figure 2 is the lack of monotonicity which is likely due to role played by prior distributions on BF when the sample size is small.

In Figure 3 curves show an increase of about 0.05 when  $n_j = 40$ , i.e. after doubling the number of case reports rated by physicians, with  $\delta = 0.15$ .

**Fig. 3** Probability of rejecting the hypothesis of homogeneity if  $n_I = 40$  case reports are rated by physicians and  $\delta = 0.15$ . A label like  $ss = 40$  indicates that a sample of 40 case reports is rated by BN4A in each actual disease class.



### 3 Conclusions

A problem of heterogeneity of one rater, BN4A, with respect to a group of physicians was restructured given favorable assumptions that stem from the nature of synthetic case reports to be assessed. Indeed 160 synthetic case reports to formulate, or more than 320 in the worst case, are a challenge for a pulmonologist but not impossible, especially by restructuring actual reports of hospitalization.

Some reduction in sample size is expected under the assumption of equal probability of correct classification for physicians,  $\tilde{\pi}_1 = \tilde{\pi}_2 = \tilde{\pi}_3 = \tilde{\pi}_4$ , and with BN4A,  $\theta_1 = \theta_2 = \theta_3 = \theta_4$ , but indeed this assumption requires quite a strong belief in face of differences existing among considered diseases.

The exchangeability of physicians might not hold if pulmonologists and cardiologist were both included as raters: then we might expect differences in  $\tilde{\pi}_z$ , at least for some subset of case reports. This is an issue that deserves attention and investigation in the future.

**Acknowledgements** The author would like to thank Shaun Bender for providing constructive criticism on the manuscript. The author acknowledges the financial support provided by Boehringer Ingelheim International GmbH and by the ‘Dipartimenti Eccellenti 2018-2022’, ministerial funds, Italy.

### References

1. Calle-Alonso, F., Pérez Sánchez, C.J.: A Monte Carlo-based Bayesian approach for measuring agreement in a qualitative scale. *Appl Psychol Meas*, **39**, 189–207 (2015) doi: 10.1177/0146621614554080
2. Fleiss, J.L.: *Statistical Methods for Rates and Proportions* (2nd Edition). John Wiley & Sons, New York (1981)
3. Zapf, A., Castell, S., Morawietz, L., Karch, A.: Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate?. *BMC Med Res Methodol*, **16**, 93 (2016) doi: 10.1186/s12874-016-0200-9

# Analysis of the spatial interdependence of nuclear size in confocal microscopy images of plant roots

## *Analisi della interdipendenza spaziale della dimensione dei nuclei in radici vegetali osservate in microscopia confocale*

Ivan Sciascia<sup>1</sup>, Andrea Crosino<sup>1</sup>, Gennaro Carotenuto<sup>1</sup>, Andrea Genre<sup>1</sup>

**Abstract** The research consists in the construction of the experimental and theoretical variogram for nuclear size in a sample of frame series from images acquired in 3D confocal microscopy. The analysis compares the estimated variograms for mycorrhizal and non-mycorrhizal roots.

**Abstract** *Il lavoro consiste nella costruzione del variogramma sperimentale e teorico della dimensione nucleare per un campione di frame selezionato da una serie di immagini 3D acquisite al microscopio confocale. L'analisi confronta i variogrammi stimati per le radici micorizzate e non micorizzate.*

**Key words:** mycorrhized roots, estimated variogram, nuclei size

## 1. Introduction

Arbuscular mycorrhizal (AM) symbiosis, a beneficial interaction between the majority of plants and a small group of soil fungi, culminates with the development of arbuscules, the structures devoted to the nutrient exchange, inside living root cells. Recent evidence from our research group shed light on the activation of plant cell cycle-related mechanisms during AM colonization using confocal microscopy observations, gene expression and flow cytometry analyses, and highlighted the occurrence of endoreduplication in AM colonized areas of the root, with the appearance of polyploid nuclei in arbusculated and

---

<sup>1</sup> Ivan Sciascia, DBIOS, Università degli Studi di Torino; email: [ivan.sciascia@unito.it](mailto:ivan.sciascia@unito.it)

Andrea Crosino, DBIOS, Università degli Studi di Torino; email: [andrea.crosino@unito.it](mailto:andrea.crosino@unito.it)

Gennaro Carotenuto, DBIOS, Università degli Studi di Torino; [gennaro.carotenuto@unito.it](mailto:gennaro.carotenuto@unito.it)

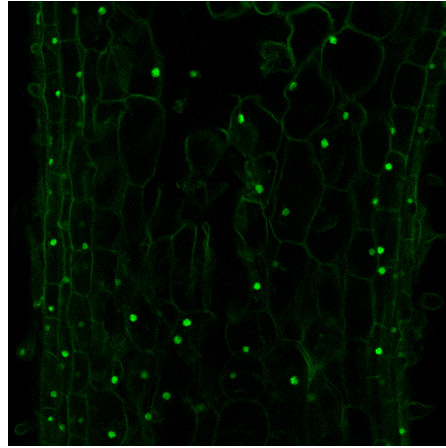
Andrea Genre, DBIOS, Università degli Studi di Torino; [andrea.genre@unito.it](mailto:andrea.genre@unito.it)

neighboring cells of the root cortex [1], [2]. Importantly, due to the correlation between DNA content and nuclear size, the mapping of nuclear ploidy was possible in images of sectioned roots [1]. The aim of the present research is to describe the spatial variability of the observed increase in nuclear size in mycorrhizal compared to non-mycorrhizal roots. The analysis was based on a marked point process, where the attribute of the point is the equatorial section area of each nucleus expressed in  $\mu\text{m}^2$ , and was applied to a large set of 3D confocal z-stack images to define nuclear size variability with the R package gstat.

## 2. Materials and methods

Semivariance is a function of distance between pairs of objects [3] and the plot experimental function is called variogram:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(p_i, p_{(i+h)}) \in S} [y(p_i) - y(p_{(i+h)})]^2 \quad (1)$$



**Figure 1.** Confocal image of a root

If we consider digital images from confocal microscopy with resolution  $\delta = 2,7307 \frac{\text{pixel}}{\mu\text{m}}$  as in Figure 1 the elements of the formula (1) are:

$\gamma(h)$  is the semivariance in function of the distance  $h$

$N(h)$  are the root nuclei pairs at distance  $h$

$S$  is the frame sample

$y(p_i)$  is the nucleus area in  $\mu\text{m}^2$  at the coordinates  $p_i$

$y(p_{(i+h)})$  is the nucleus area in  $\mu\text{m}^2$  at the coordinates  $p_{(i+h)}$

The sampling design of a series of control and inoculated roots prevised two phase: selection of a root, and in the second phase selection of nuclei in a frame.

The point cloud of the variogram and the theoretical estimation function of the variogram are then drawn [3] according to one of the estimation models.

## 3. Estimated variograms

Spatial plot of the sampled nuclei and estimated variogram are described in Figure 1 and Figure 2. The value of the sill of the theoretical graph was set on the basis of the variance of the sample measurements while the range was visually evaluated and set equal for the two data series. The nugget was conventionally set to zero.

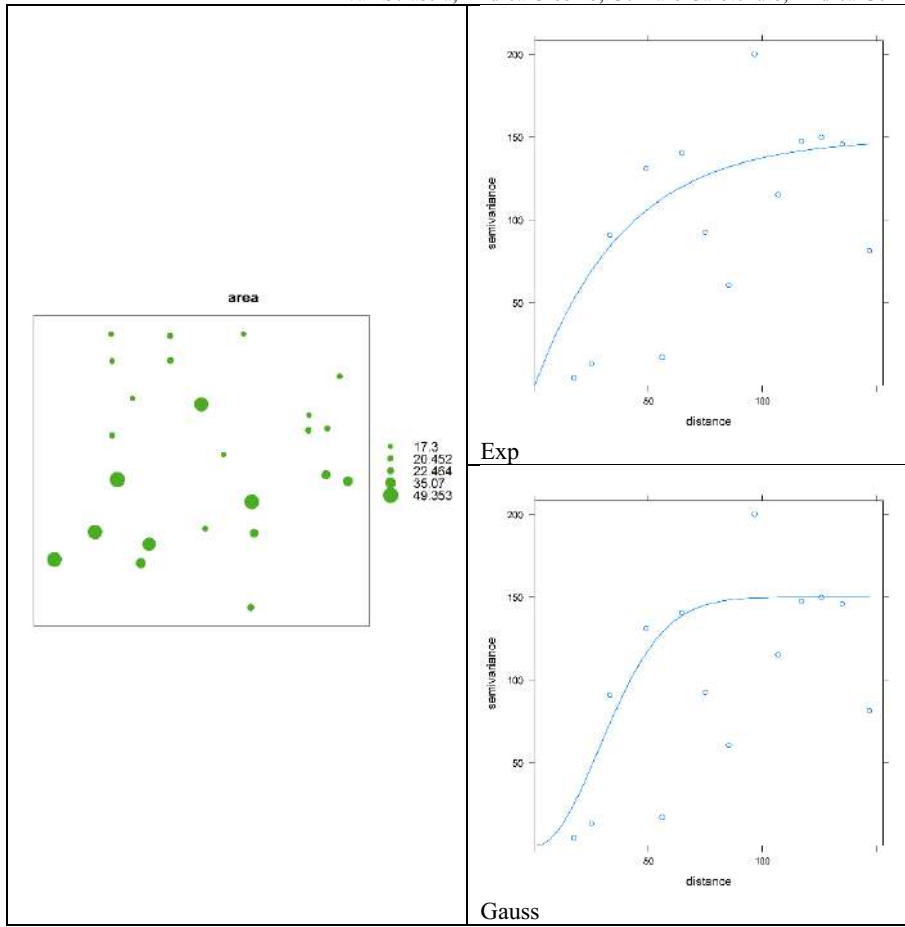
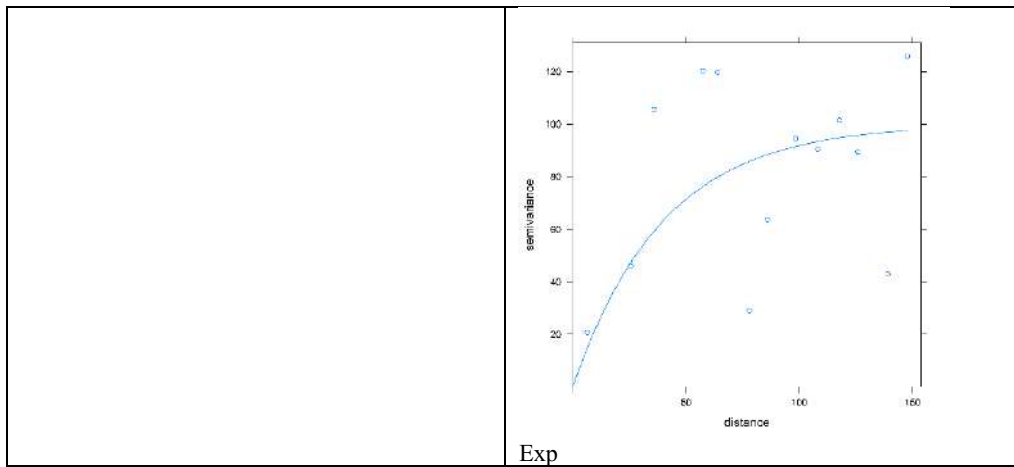
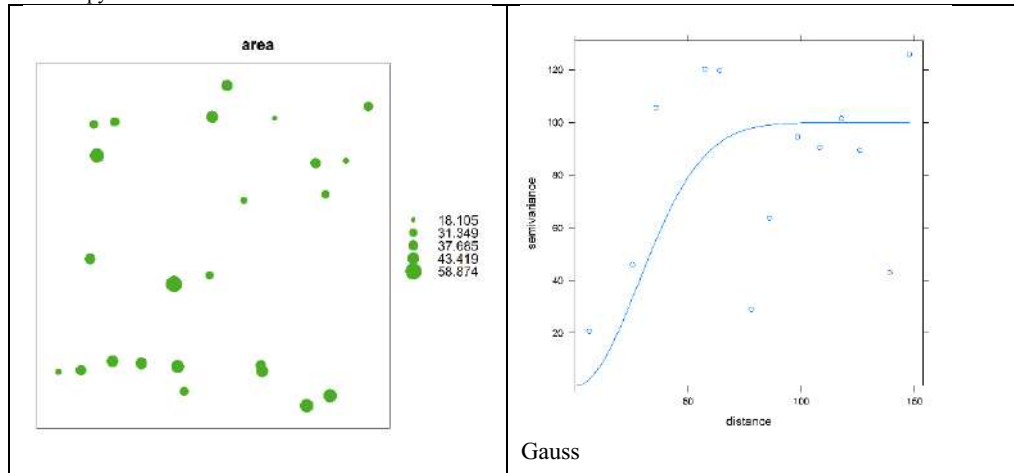


Figure 2. Area plot of control nuclei and estimated exponential and Gaussian variograms



Analysis of the spatial interdependence of the size of endoreduplicated nuclei observed in confocal microscopy



**Figure 3.** Area plot of mycorrhized nuclei and estimated exponential and Gaussian variograms

#### 4. Performance indexes

To compare the experimental and the estimated semivariance in function of the distance, the computation of the error of estimates is

$$e_i = Y_i - M_i$$

where

$e_i$  is the error of the estimates for the nucleus  $i$

$Y_i$  is the experimental semivariance

$M_i$  is the model (exponential or Gaussian) estimated semivariance

Considering the errors for each point  $i$  the mean absolute deviation (MAD) is:

$$MAD = \frac{1}{n} \sum_{i=1}^n |e_i|$$

And the root mean square error (RMSE) is:  $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2}$

The Table 1 describes the indexes for control and mycorrhized digital images groups

**Table 1:** Performance indexes for the estimation models

<i>Model</i>	<i>MAD</i>	<i>RMSE</i>
Control Exponential	3,784341	4,735795
Control Gaussian	3,616071	4,948705
Mycorrhized Exponential	2,459242	3,201468
Mycorrhized Gaussian	2,702641	3,368925

This exploratory research with digital images from confocal microscopy took into consideration spatial statistical analysis techniques used in geostatistics that allow the observation of the variability of study characteristics as a function of spatial coordinates. We have drawn the point cloud of the variograms for a sample of



Ivan Sciascia, Andrea Crosino, Gennaro Carotenuto, Andrea Genre uninoculated and mycorrhized roots observing any differences. From this first study using theoretical semivariance estimation models we observed that the estimation models perform better for the frame sample of mycorrhized roots. This may suggest minor variability in the distribution of nuclei sizes, to be investigated with other research in confocal microscopy.

## References

1. Carotenuto G., Sciascia I., Oddi L., Volpe V., Genre A. Size matters: Three methods for estimating nuclear size in mycorrhizal roots of *Medicago truncatula* by image analysis, *Bmc Plant Biology*, 19(1), 180-193 (2019a)
2. Carotenuto G., Volpe V., Russo G., Politi M., Sciascia I., de Almeida-Engler J., Genre A. Local endoreduplication as a feature of intracellular fungal accommodation in arbuscular mycorrhizas, *New Phytologist*, 223, 430-446 (2019b)
3. Cressie N.A.C., 1993, *Statistics for Spatial Data*, Wiley

# A Density-Peak Approach to Clustering Graph-Structured Data

## *Un approccio density-peak per il clustering tra grafi*

Riccardo Giubilei

**Abstract** Mode-based clustering is a nonparametric method that defines clusters using the basins of attraction of a kernel density estimator's modes. A successful way to identify modes and their basins of attraction is the density-peak approach, which is based on the idea that cluster centers have higher density than their neighbors, and that they are quite distant from points with higher densities. The resulting clustering procedure has many advantages, the most important in our perspective being that it does not require embedding the data in a vector space. This paves the way for analyzing structured data, an ever growing necessity in modern data analysis. In this work, we adapt the density-peak approach to the important but not trivial task of clustering graph-structured data. After investigating the changes to be made, we identify an appropriate distance between graphs. Then, we test our method in a simple simulation scenario, obtaining promising results. Finally, we discuss our findings and outline some ideas for future work.

**Abstract** *Il mode-based clustering è un metodo non parametrico che definisce i cluster utilizzando i bacini di attrazione delle mode di uno stimatore kernel di densità. Un modo popolare di identificare le mode e i loro bacini di attrazione è l'approccio density-peak, che è basato sull'idea che il centro dei cluster ha densità più alta rispetto ai propri vicini, e che esso è abbastanza lontano da punti con densità più alta. La procedura di clustering che ne deriva presenta numerosi vantaggi, tra cui il più importante qui è che i dati non devono necessariamente appartenere a uno spazio vettoriale. Ciò apre la strada all'analisi di dati di tipo strutturato, che è una necessità crescente al giorno d'oggi. In questo lavoro, adattiamo l'approccio density-peak all'importante ma non banale problema del clustering tra grafi. Dopo aver studiato le modifiche da apportare, identifichiamo una distanza tra grafi appropriata. Dopodichè, testiamo il nostro metodo in un semplice scenario di simulazione, ottenendo risultati incoraggianti. Infine, discutiamo i nostri risultati e tracciamo alcune idee per sviluppi futuri.*

**Key words:** Density-peak clustering, Mode-based clustering, Graphs, Structured data, Unsupervised Learning, Object-Oriented Data Analysis.

---

Riccardo Giubilei  
Sapienza University of Rome, Rome, e-mail: riccardo.giubilei@uniroma1.it

## 1 Introduction

Clustering is the task of grouping elements from a set in such a way that elements in the same group, also defined as *cluster*, are in some sense similar to each other, and dissimilar to those belonging to other groups. Even though no consensus has been reached even on the definition of a cluster [12], clustering methods have been applied to many different fields, are classified into several categories, and a variety of different clustering strategies have been proposed.

Mode-based clustering is a nonparametric distance-based method that was formally introduced in the statistical literature by [9, 1]. The related idea that *clusters may be thought of as regions of high density separated from other such regions by regions of low density* was first proposed in [8]. Mode-based clustering is commonly based on the mean-shift algorithm [3, 4] for finding cluster centers, i.e. modes, and their basins of attraction. However, more recent methods such as the density-peak algorithm [12] have been proved to outperform mean-shift, at least in terms of time complexity.

Density-peak mode-based clustering has drawn a lot of attention since its introduction in 2014, the main reason being its favorable properties. During recent years, several extensions, improvements and speedups have been proposed; see e.g. [13] for a review. However, the algorithm's propensity to be easily generalized in order to cluster structured data objects – such as functional data or graphs – has not been adequately explored.

In this work, we propose a method to perform density-peak clustering on graph-structured data. The idea falls and finds its motivation within Object-Oriented Data Analysis [14], a stream of research concerned with the analysis of data objects in their most natural form, however complex may it be. Specifically, the problem of clustering graphs has many important potential applications, such as clustering protein-protein interaction networks for disease evaluation, or clustering brain networks for neurological disorder analysis. Despite many previously proposed methods (e.g., [10, 11]), so far none of them has been acknowledged as well-established in the literature.

## 2 Mode-based clustering

Mode-based clustering is a nonparametric method that works by first estimating the density, usually via kernel density estimators, and then identifying in some way the modes and the corresponding clusters, also defined as *basins of attraction*. Formally, let  $f$  be the density function of a random vector  $X \in \mathbb{R}^d$ . Suppose that  $f$  has  $r$  local maxima  $\mathcal{M} = \{m_1, \dots, m_r\}$ , and its Hessian is non-degenerate at each critical point. The number of local maxima  $r$  is not assumed to be known. Given an observed point  $x \in \mathbb{R}^d$  and excluding sets of measure zero, there is a unique gradient ascent path from  $x$  to one of the  $r$  modes. The sets of points whose ascent paths arrive at the same mode is the cluster defined by that mode, and has been equivalently defined *basin of attraction* starting from [2].

Descending more into detail, define an *integral curve* through  $x$  as a path  $\pi_x : \mathbb{R} \rightarrow \mathbb{R}^d$  such that:

- i.  $\pi_x(0) = x$ ;
- ii.  $\pi_x'(t) = \nabla f(\pi_x(t))$ .

These curves partition the space and have their destination  $\varphi_\pi(\cdot)$ , function of the starting point  $x$ , defined as:

$$\varphi_\pi(x) = \lim_{t \rightarrow \infty} \pi_x(t). \quad (1)$$

Then, by property ii. of  $\pi_x$ , necessarily  $\varphi_\pi(x) = m_j$  for some mode  $m_j$ ,  $j = 1, \dots, r$ , and for all  $x$ , again excluding sets of measure zero. In light of this, the *basin of attraction* of the mode  $m_j$  is defined by:

$$C_j = \{x : \varphi_\pi(x) = m_j\}, \quad j = 1, \dots, r. \quad (2)$$

The resulting partition  $\mathcal{C} = \{C_1, \dots, C_r\}$  is formed of the  $r$  *population* clusters.

However, since  $f(x)$  is usually unknown in practice, it needs to be estimated. A very common way to do so is using the kernel density estimator (KDE). Let  $X_1, \dots, X_n$  be a random sample from  $f$ , and let  $K$  be a smooth and symmetric kernel. The KDE with bandwidth  $h > 0$  is defined by:

$$\hat{f}_h(x) = \frac{1}{nh^d} \sum_i K\left(\frac{\|x - X_i\|}{h}\right). \quad (3)$$

The modes  $\hat{\mathcal{M}} = \{\hat{m}_1, \dots, \hat{m}_r\}$  of  $\hat{f}_h(x)$  and the integral curve destinations under  $\hat{f}_h(x)$  of any point  $x$ , i.e.  $\hat{\varphi}_\pi(x)$ , can be found using a variety of techniques; an example is given by the density-peak algorithm described in Section 3. Using any technique, the resulting basins of attraction are:

$$\hat{C}_j = \{x : \hat{\varphi}_\pi(x) = \hat{m}_j\}, \quad j = 1, \dots, r, \quad (4)$$

and the *sample* clusters are defined by:

$$\mathcal{X}_j = \{X_i : X_i \in \hat{C}_j\} = \{X_i : \hat{\varphi}_\pi(X_i) = \hat{m}_j\}. \quad (5)$$

### 3 Density-peak algorithm

The density peak algorithm [12] is brilliantly based on a simple idea: cluster centers, i.e. the modes, are surrounded by neighbors with lower local density, and, on the other hand, they are at a relatively large distance from any points with a higher local density. This idea is formalized with two quantities that are computed for each observation  $x_i$ : the local density  $\rho_i$ , and the minimum distance  $\delta_i$  from other data points with higher density. Specifically, let  $d_{ij} = d(x_i, x_j)$  be the distance between data points  $x_i$  and  $x_j$ . The local density  $\rho_i$  of data point  $x_i$  is defined as:

$$\rho_i = \sum_j I_{(d_{ij} - d_c)} \quad (6)$$

where  $I_{(\cdot)}$  is the indicator function, and  $d_c$  is a cutoff distance. In simple terms,  $\rho_i$  is the number of points that are closer than  $d_c$  to  $x_i$ . Since the algorithm is sensitive only to the relative magnitude of  $\rho_i$ , it is robust with respect to the choice of  $d_c$ , at least when dealing with large datasets.

Based on equation (6), it is now possible to build  $\delta_i$  as the minimum distance between point  $x_i$  and any other point  $x_j$  with higher density, i.e. for which  $\rho_j > \rho_i$ :

$$\delta_i = \min_{j: \rho_j > \rho_i} (d_{ij}) \quad (7)$$

The point with the highest density has its *minimum distance* conventionally set to  $\delta_i = \max_j(d_{ij})$ . The interpretation of  $\delta_i$  is where the algorithm's core idea becomes explicitly involved: for data points that are not local or global maxima,  $\delta_i$  is constrained by other data

points within the same cluster, following the definition in (7); thus, cluster centers are identified as points for which the value of  $\delta_i$  is anomalously large.

Comparing the density-peak approach with the formerly more established mean-shift algorithm, both have the ability to detect nonspherical clusters and to automatically find the correct number of clusters [12]. However, unlike the mean-shift algorithm, the density-peak clustering does not demand for data to be embedded in vector spaces, nor to maximize explicitly the density field for each data point [12]. This inevitably has consequences from a computational point of view, making the density-peak approach much faster. Additionally, since the two fundamental quantities  $\rho$  and  $\delta$  are entirely based on the distances between the data points and not also on their observed values, density-peak clustering may be easily adapted to deal with structured data.

## 4 Graphs and density-peak adaptation

A *graph* is mathematical object consisting in a collection of *vertices* linked by *arcs* or *edges* between them. Graphs are studied in graph theory, but also in network science, where they are usually referred to as *networks* made of *nodes* and *links*. Whatever the terminology, they are a key theoretical instrument when studying complex systems characterized by a set of objects or people and where some pairs have some type of relationship; notable examples include the Internet, social networks and brain networks. Formally, a graph can be denoted with  $G = (V, E)$ , where  $V$  is the set of vertices and  $E$  is the set of edges. If  $e \in E$  joins vertices  $u, v \in V$ , i.e.  $e = \{u, v\}$ , then  $u$  and  $v$  are said to be *adjacent* or *neighbors*. The number of edges incident with any vertex  $v$  is called *degree* of  $v$ . If edge  $e = \{u, v\}$  is equivalent to edge  $e = \{v, u\}$ , the graph  $G$  is said to be *undirected*; otherwise, it is *directed*.  $G$  may be equipped with auxiliary numerical values on its vertices, edges, or both; a *weighted graph*  $G$  is a graph for which each edge  $e$  has an associated real-valued number  $w_e$  called *edge weight*. In the simple case when  $G$  is not weighted, all the information about its connectivity may be stored in an  $|V| \times |V|$  binary matrix  $\mathbf{A}$  with entries:

$$A_{pq} = \begin{cases} 1 & \text{if } \{p, q\} \in E \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where the integers  $p, q \in \{1, \dots, |V|\}$  are used to denote the elements of  $V$ . The matrix  $\mathbf{A}$  is defined *adjacency matrix*, and is often used as an alternative representation of  $G$ .

In order to adapt the density-peak algorithm for mode-based clustering to the case of graph-structured data, an appropriate distance between graphs must be identified. While the latter is an ongoing trend of research [6, 5], finding an optimal distance of such a kind is still an open problem. In this work, we choose the Edge Difference Distance [7], that is defined as the Frobenius norm of the difference between the two graphs' adjacency matrices. The reasons for this choice are the suitability of its definition also for signed and weighted graphs, the reasonable results it yields, and its limited computational time complexity. Formally, the Edge Difference Distance between two graphs  $x_i$  and  $x_j$  is defined as:

$$d_{ij}^{ED} = \|\mathbf{A}^i - \mathbf{A}^j\|_F := \sqrt{\sum_p \sum_q |A_{pq}^i - A_{pq}^j|^2} \quad (9)$$

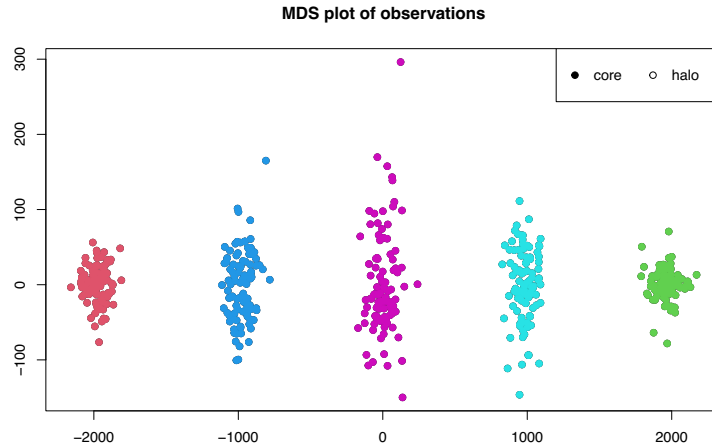
where  $\mathbf{A}^i$  and  $\mathbf{A}^j$  are the adjacency matrices of  $x_i$  and  $x_j$  respectively, and  $\|\cdot\|_F$  denotes the Frobenius norm. Once the Edge Difference Distance defined in (9) is used in place of  $d_{ij}$  in Equations (6) and (7) respectively, the density-peak algorithm is ready to be used for clustering graph-structured data points.

## 5 Preliminary results

Our approach is tested on a simple simulation scenario. We generate  $N = 500$  graphs that do belong to  $k = 5$  different equally-sized classes, in that they come from the same model but with 5 different parameter configurations. Specifically, we use the classical  $G(n, p)$  Erdős–Rényi model to simulate random graphs with  $n = 100$  nodes and a number of edges determined by a predefined connection probability  $p$  that is different for each class. We set  $p = \{0.3, 0.4, 0.5, 0.6, 0.7\}$  respectively for the 5 classes, so that the differences between the probabilities are neither large enough to make the task too easy, nor too close to prevent from detecting any structural difference between the graphs.

Once the graphs are generated, we calculate the Edge Difference Distance  $d_{ij}^{ED}$  defined by (9) for each couple of graphs  $(x_i, x_j)$ . Then, the two quantities  $\rho$  and  $\delta$  are computed for each graph following Equations (6) and (7) respectively. Finally, all the graphs that show a very large value for  $\delta$  are selected as cluster centers. This allows to identify 5 different clusters.

The results of the clustering task are shown in Figure 1, which is a Multidimensional scaling plot that includes each simulated graph as a single dot. The 5 distinct point clouds represent each a different class of graphs; on the other hand, dots are colored based on the cluster they have been assigned to by our approach. Hence, our method manages to perfectly rebuild the 5 different clusters, correctly identifying the affiliation of all the 500 graphs. Additionally, in general the method may place some observations in the cluster halo, which would be represented as hollow circles in the plot and interpreted as noise; however, all the circles are filled, meaning that all the graphs are assigned directly to the clusters' cores.



**Fig. 1** Multidimensional scaling plot for the 500 simulated graphs. Data points are colored differently based on their cluster affiliation. Circles are either filled, if the observations belong to the cluster core, or hollow, if they belong to the cluster halo.

Therefore, our density-peak approach to clustering graph-structured data produces encouraging results, even though in a simulation scenario as simple as the one presented here. Interestingly, we have conducted similar experiments with a different number  $k$  of classes. For  $k \in \{2, 3, 4\}$  and comparable differences in the connection probabilities  $p$ , the performance is also perfect. However, for growing  $k > 5$ , the method yields increasingly worse results, partly because of the more overlapping structure of the graphs. In these cases, the performance depends more heavily on the chosen connection probabilities for the classes.

## 6 Conclusions and future work

In this work, we have proposed a density-peak approach to graph-structured data. After giving proper introduction, background and motivation for the problem, we have described our methodology. Then, we have tested our approach in a simple simulation scenario with 5 clusters obtained by generating random graphs using the Erdős–Rényi model and different connection probabilities.

Event though they are only preliminary, the results are promising. We have briefly discussed them in the previous section, together with those obtained through little variations on the theme. The next step will be to test our methods in more complex simulation scenarios, as well as using them with real-world data. Especially in the last case, we are aware that it could be necessary to revise the chosen distance between graphs in order to meet the specific needs of the application.

Other ideas for future work include exploiting some of the variants that have been proposed over the years and that speedup the standard density-peak algorithm. Besides, other types of structured objects, such as functional data, persistence diagrams and shapes, could be considered for further extending the applicability of this density-peak approach. Last but not least, studying the statistical properties of the density-peak clustering both with traditional and structured data would be of great interest and utility.

## References

1. Azzalini, A., Torelli, N.: Clustering via nonparametric density estimation. *Statistics and Computing* **17**(1), 71–80 (2007)
2. Chacón, J.E.: Clusters and water flows: a novel approach to modal clustering through morse theory. arXiv preprint arXiv:1212.1384 (2012)
3. Cheng, Y.: Mean shift, mode seeking, and clustering. *IEEE PAMI* **17**(8), 790–799 (1995)
4. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. *IEEE PAMI* **24**(5), 603–619 (2002)
5. Donnat, C., Holmes, S., et al.: Tracking network dynamics: A survey using graph distances. *The Annals of Applied Statistics* **12**(2), 971–1012 (2018)
6. Emmert-Streib, F., Dehmer, M., Shi, Y.: Fifty years of graph matching, network alignment and network comparison. *Information sciences* **346**, 180–197 (2016)
7. Hammond, D.K., Gur, Y., Johnson, C.R.: Graph diffusion distance: A difference measure for weighted graphs based on the graph laplacian exponential kernel. In: *IEEE GlobalSIP 2013*, pp. 419–422. IEEE (2013)
8. Hartigan, J.A.: *Clustering algorithms*. John Wiley & Sons, Inc. (1975)
9. Li, J., Ray, S., Lindsay, B.G.: A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research* **8**(Aug), 1687–1723 (2007)
10. Ma, G., He, L., Cao, B., Zhang, J., Philip, S.Y., Ragin, A.B.: Multi-graph clustering based on interior-node topology with applications to brain networks. In: *ECML PKDD*, pp. 476–492. Springer (2016)
11. Mukherjee, S.S., Sarkar, P., Lin, L.: On clustering network-valued data. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, p. 7074–7084. Curran Associates Inc. (2017)
12. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. *Science* **344**(6191), 1492–1496 (2014)
13. Sieranoja, S., Fränti, P.: Fast and general density peaks clustering. *Pattern Recognition Letters* **128**, 551–558 (2019)
14. Wang, H., Marron, J.S.: Object oriented data analysis: Sets of trees. *The Annals of Statistics* **35**(5), 1849–1873 (2007)

# The employment situation of people with disabilities in Tuscany

## *La situazione occupazionale delle persone con disabilità in Toscana.*

Paolo Addis, Alessandra Coli and Gianfranco Francese

**Abstract** Italian legislation recognizes the right of people with disabilities to be able to support themselves through a freely chosen job, in an open work environment that favors inclusion. Ires Toscana and the Dirpolis Institute of the Sant'Anna School of Advanced Studies conducted a survey in Tuscan companies to investigate the condition of disabled people in the workplace. The survey was designed to capture the point of view of both the disabled and their colleagues. This article describes the main features of the survey, focusing on questionnaires.

**Abstract** *La legislazione italiana riconosce il diritto delle persone con disabilità di potersi mantenere attraverso un lavoro liberamente scelto, in un ambiente di lavoro aperto che favorisca l'inclusione. Ires Toscana e l'Istituto Dirpolis della Scuola Superiore Sant'Anna hanno condotto un'indagine presso le aziende toscane per indagare la condizione dei disabili sul posto di lavoro. L'indagine è stata progettata per cogliere l'opinione sia dei disabili che dei loro colleghi. Questo articolo descrive le caratteristiche principali dell'indagine, soffermandosi in particolare sui questionari.*

**Key words:** Disability, Work Environment, Inclusion

---

Paolo Addis  
Istituto Dirpolis della Scuola Superiore S. Anna, Via Vernagalli, 22R, 26R 56127 Pisa, Italy, e-mail: paolo.addis@santannapisa.it

Alessandra Coli  
Dip. Economia e Management, via C. Ridolfi 16, 56124 Pisa, e-mail: alessandra.coli1@unipi.it, Scientific Committee of Ires Toscana

Gianfranco Francese  
Ires Toscana, Via Stradella 15, 50127 Firenze, e-mail: gianfranco.francese@irestoscana.it, President of Ires Toscana



## **1 Introduction**

Work is a fundamental part of people's lives not just for pay. In fact, work takes up a large part of everyday life and thoughts. It allows you to express your creativity, fulfill desires and satisfy needs and is an opportunity for social relations and comparison. Work is also the strongest tool against any kind of discrimination.

Today, work is in crisis due to the unfavorable economic situation but also due to economic policy choices that have diminished its social value. Work precariousness was favored, especially for the most fragile categories of the market, namely young people, women and people with disabilities.

The economic crisis has not spared Tuscany in recent years, violently hitting the various production sectors and leading to a negative employment balance. Due to the economic prices, people with disabilities have paid a high price in terms of reduced job placements.

Unfortunately, although they are able to work, many people with disabilities are often victims of clichés and prejudices, due to the misinformation of many employers, public or private. In addition, people with disabilities often encounter difficulties associated with reaching and accessing the workplace. A very critical situation, therefore, that requires great attention.

For these reasons, with the 2008 financial crisis behind us and even before the outbreak of the Covid 19 epidemic, Ires Toscana and the Dirpolis Institute of the Scuola Superiore S. Anna deemed it necessary to carry out a sample survey to investigate the employment situation of people with disabilities in Tuscany. The Coordinamento Disabilità CGIL Toscana provided practically for the administration of the questionnaires in the workplace, from June to September 2018.

This research represents a pilot experiment for the realization of a wider survey project on the theme of disability and employment. Indeed, we believe that the job placement of people with disabilities represents respect for a fundamental right, an objective of social justice and, last but not least, an important indicator of the degree of civilization of a society.

The paper is structured as follows. Section 2 recalls the main Italian legislation concerning the employment of people with disabilities, while Section 3 describes the main characteristics of the surveys.

## **2 Legal provisions on the employment of people with disabilities in the Italian legal framework**

As is well known, Italy is a "Republic founded on work" (Article 1 of the Constitution, see [1] ) and "The Republic recognizes the right to work for all citizens and promotes the conditions that make it effective. Every citizen has the duty to bring outside, according to their possibilities and their choice, an activity or function that contributes to the material or spiritual progress of society."(art. 4). On the other

hand, it is precisely the inability to work that triggers the protections provided for by art. 38 of the Constitution, where it is stated that “I. Every citizen unable to work and without the necessary means to live has the right to maintenance and social assistance. II. Workers have the right to provide and guarantee adequate means for their life needs in case of accident, illness, invalidity and old age, involuntary unemployment. III. Disabled and handicapped people have the right to education and vocational training [...]”.

The right to work of people with disabilities has been the subject of various legislative interventions, starting in the years following the First World War. It was in this period that the interventions “in favor of people with work disabilities” saw the light, “the war had in fact left the sad legacy of a not negligible mass of mutilated ex combatants and therefore more or less disabled” [2]. Alongside a system of pensions and benefits for those who had completely lost the opportunity to work, mechanisms were put in place to allow people with residual skills to enter the labor market. In times closest to us, law 462/1968 provided for compulsory hiring mechanisms from private employers and public administrations for “war invalids, military and civilians, service invalids, work invalids, civil invalids, the blind, deaf and dumb, orphans and widows of those who died in war or for service or at work, former tuberculosis and refugees”(therefore art. 1 of the law in question).

Finally, the law 68/1999, once the l. 462/1968, imposed a change of course in the discipline of the placement of people with disabilities, passing from the idea of an almost charitable placement to a system in which the right to work of the person with disabilities meets the needs of the employer. The national discipline is flanked by the regional one, within the limits of competence defined by art. 117 of the Constitution. As regards Tuscany, for example, the recent provisions of article 21 of the regional law 60 of 2017 states that “The Region promotes the involvement of people with disabilities in the definition of active territorial employment policies through the participation of the most representative associations at the regional level”.

The protection provided to workers with disabilities by the anti-discrimination law should also be borne in mind. In the Italian legal system there is both an anti-discrimination instrument of general application (Law 67 of 2006), and a discipline that expressly aims at the elimination of discrimination in the workplace (Legislative Decree 216 of 2003).

The aforementioned legislative decree receives a source from the European Union, Directive 2000/78 / EC. In addition to what is established by European Union law, it is worth reminding the Convention C-159 of the International Labor Organization, entered into force in 1985 but ratified by Italy with Law 189 of 1989 and the United Nations Convention on the Rights of Persons with Disabilities, which Italy ratified and executed with Law 18 of 2009. Art. 27 states that “States Parties recognize the right of persons with disabilities to work, on an equal basis with others; in particular the right to be able to support oneself through a job freely chosen or accepted in a labor market and in an open work environment, which favors inclusion and accessibility for people with disabilities”.

### **3 A Survey on the employment situation of people with disabilities in Tuscany**

The International Labor Organization (ILO) argues that useful data on the employment situation of people with disabilities is rarely available at the required level of detail and frequency. In a number of countries there is even no data on the employment situation associated with disability. This despite the strong interest in the availability of data capable of monitoring the effects of the legislation aimed at promoting job opportunities for people with disabilities [3].

In Italy, the Italian National Institute of Statistics (ISTAT) and the Ministry of Labor and Social Policy limit to publish data on the distribution of employed people by severity of limitation, self-declared activity status and sex of the employed at the national level [4]. In Tuscany, the annual report on disability provides some other details on the characteristics and the distribution of employed people with disabilities by province [5]. However, unfortunately, not even such report investigates the situation of people with disabilities at the workplace.

The survey planned by Ires Toscana and the Dirpolis institution aims to fill the gap in the knowledge of the phenomenon, trying to outline a comprehensive picture of the situation of employees with disabilities in the Tuscan region.

To outline as clear a picture as possible, it was decided to capture multiple points of view, therefore both people with disabilities and their colleagues were interviewed. Various aspects were considered to evaluate the inclusiveness of the workplace, from the legal status of disabled people, to the accessibility and functionality of the workstation, to the relationship with colleague and trade unions.

As a first step, we run a sort of pilot survey, aimed primarily to test the questionnaire on a wide selection of observation units. Sample was selected based on non-probabilistic criteria, taking into account the availability of local CGIL delegates to administrate the questionnaires on the field.

The development of the questionnaire took a long time because it involved various subjects, namely the researchers of Ires Toscana and of the Dirpolis Institute of the Sant'Anna School of Advanced Studies, as well as representatives of the Coordinamento Disabilità CGIL Toscana.

Two separate questionnaires were created, the first, more detailed, for people with disabilities and a shorter one for colleagues. The two questionnaires share a first common part in which details about the company are requested, i.e. the province of residence, the type of economic activity, the number of employees, whether the company is public or private. All this information refers to the specific local unit where disabled people (and their colleagues) are employed. Then some personal questions are asked, namely the year of birth, the sex and the level of education achieved.

The questionnaire for people with disabilities includes eight further sections, each focused on one of the following topics: personal data, work history, entry into the world of work, work environment, reaching the workplace, training and professional and personal development paths, disability and safety at work, trade union

and disability. The interviewee is asked to indicate the type of disability, distinguishing between sensory, mental or physical disability. Furthermore, a question is aimed at assessing whether it is a serious disability or not (art. 3, Law 104 of 1992). The questionnaire is mainly composed of closed questions. Open questions are inserted whenever it is considered complex to identify possible answers a priori. For example, the interviewee is asked to specify the defects of the workstation or to provide some examples of the difficulties encountered in the relationship with colleagues or superiors. Table 1 shows the distribution of questionnaires for people with disabilities by economic sector and size of the firm where the person is employed, and gender of the interviewee.

**Table 1** Distribution of the questionnaires for people with disabilities by economic activity and size (number of workers) of the firm and by gender of the worker.

Economic activity	10-49		50-249		> 250		Total
	Male	Female	Male	Female	Male	Female	
Agriculture, forestry and fishing					1	1	2
Mining and quarrying			1	1			2
Manufacturing	4	4	6	7		3	24
Water supply, sewerage, etc.			1	1			2
Wholesale and retail trade	3	1	1		1		6
Transportation and storage						1	1
Accommodation and food service activities		1					1
Information and communication			2	2	1	6	11
Financial and insurance activities	2	5			3	1	11
Public administration and defence			6	16	2	5	29
Education	1	4				1	6
Human health and social work activities	2	2	1	1		1	7
Total	12	17	18	28	8	19	102

*Notes:* 4 questionnaires are not included in the table due to missing information.

Questionnaires for colleagues are much simpler. In addition to the common part mentioned above, the interviewee is asked to express an opinion on the impact of people with disabilities in the workplace. The following possible answers are given: completely negative impact, mainly negative impact but with some positive aspects, mainly positive impact but with some negative aspects and completely positive impact. The interviewee can then provide a justification for the answer provided if they deem it appropriate.

The questionnaires were administered through face-to-face interviews during the year 2019. In total, 106 people with disabilities and 979 colleagues were interviewed. The CGIL trade union delegates conducted the interviews on the territory. The transcription and coding of the answers provided through the questionnaires is still in progress.

## 4 Conclusions and way forward

The survey described in this work represents a first contribution to the development of a stable survey on the employment situation of people with disabilities in Tuscany. The analysis of the results, still in progress, will allow us to improve the questionnaire but also to give some preliminary results useful to start analyzing the phenomenon.

## References

1. Luciani M.: Radici e conseguenze della scelta costituzionale di fondare la Repubblica democratica sul lavoro, ADL, pp. 628-652, 2010
2. Pera G: Invalidi e mutilati, Enciclopedia del diritto, XXII, Giuffrè, Milano, 1972
3. ILO in collaboration with the InFocus Programme on Skills Knowledge and Employability: Statistics on the employment situation of people with disabilities. A Compendium of national methodologies. ILO working papers, 2003
4. Istat : Disability in figures. Available at <http://www.disabilitaincifre.it/>. Cited 25 Feb 2021
5. Regione Toscana: Tutele, difficoltà, vita quotidiana delle persone con disabilità. Quinto Rapporto sulle Disabilità in Toscana 2020-2021. Available at <http://https://www.regione.toscana.it/-/quinto-rapporto-sulle-disabilita-in-toscana-2020>. Cited 25 Feb 2021

**Acknowledgements** We sincerely thank Adriano Turi, head of the Coordinamento Disabilità CGIL Toscana, for his commitment and strength.

# Robustness of statistical methods for modeling paired count data using bivariate discrete distributions with general dependence structures

*Robustezza dei metodi statistici per modellare dati di conteggio appaiati utilizzando distribuzioni discrete bivariate con struttura di dipendenza generale*

Marta Nai Ruscone and Dimitris Karlis

**Abstract** Bivariate Poisson models are appropriate for modeling paired count data. However the bivariate Poisson model does not allow for negative dependence structure, therefore it is necessary to consider alternatives, which can produce both positive and negative dependence. A natural way is to consider copulas to generate various bivariate discrete distributions. While such models exist in the literature, the issue of choosing a suitable copula has been overlooked so far. Different copulas lead to different structure, any copula misspecification can render the inference useless. In this work, we consider bivariate Poisson models generated with a copula and investigate its robustness under outliers contamination and model misspecification. Particular focus is given on the robustness of copula related parameters.

**Abstract** *I modelli di Poisson bivariati sono appropriati per modellare dati di conteggio appaiati, tuttavia il modello standard di Poisson bivariato non consente di considerare una struttura di dipendenza negativa, quindi è necessario considerare modelli alternativi che possono considerare sia strutture di dipendenza positive che negative. Un modo naturale è quello di considerare le funzioni copula per generare varie distribuzioni discrete bivariate. Sebbene tale modello esista in letteratura, è sempre stato trascurato il problema della scelta dell'uso delle copule. Copulae differenti portano a strutture differenti ed è importante capire se l'errata specificazione della copula può causare problemi. In questo lavoro, consideriamo modelli di Poisson bivariati generati con una copula e ne indaghiamo la robustezza nel caso di contaminazione con valori anomali e l'errata specificazione del modello. Particolare attenzione è data alla robustezza dei parametri relativi alla copula.*

---

Marta Nai Ruscone

University of Genoa, Department of Mathematics - DIMA, Via Dodecaneso, 35, 16146 Genoa, Italy e-mail: marta.nairuscone@unige.it

Dimitris Karlis

Athens University of Economics and Business, 76 Patission str, 10434 Athens, Greece e-mail: karlis@aueb.gr

**Key words:** copula, dependence, outliers, robustness

## 1 Introduction

Bivariate Poisson models are appropriate for modeling paired count data exhibiting correlation. Paired count data arise in a wide context including marketing (e.g. number of purchases of different products), epidemiology (e.g. incidents of different diseases in a series of districts), accident analysis (e.g. number of accidents in a site before and after infrastructure changes), medical research (e.g. the number of seizures before and after treatment), sports (e.g. the number of goals scored by each one of the two opponent teams in soccer), econometrics (e.g. number of voluntary and involuntary job changes), just to name a few. Several models are available that can incorporate different structures and marginal properties, see for example Kocherlakota and Kocherlakota[8] and Karlis and Ntzoufras [6]. See also the work in Nikoloulopoulos[7] for defining models with copulas. While several extensions and models have been proposed up to our knowledge issues of robustness have been overlooked. Following Grunert da Fonseca and Fieller [4], there are two kinds of achieved robustness that one should consider. The first one refers to contamination from outlier observations or, better, from observations that are unexpected under a certain model. The second one refers to model deviation, i.e. a researcher would like to fit the model with such a method that even if the model is not correct the method would protect from deriving inconsistent results.

In this work, we consider a copula based bivariate Poisson distributions. We apply a minimum distance estimation methodology using Hellinger distance. We investigate its robustness under outliers contamination and model misspecification. Particular focus is given on the robustness of copula related parameters that measure the association exhibited by paired count data.

## 2 Copulas

Copula are functions that join multivariate distribution functions to their marginal distribution functions [1]. They describe the dependence structure existing across marginal random variables. In this way we can consider bivariate distributions with dependency structures different from the linear one that characterizes the multivariate Gaussian distribution.

A bivariate copula  $C : I^2 \rightarrow I$ , with  $I^2 = [0, 1] \times [0, 1]$  and  $I = [0, 1]$ , is the cumulative bivariate distribution function of the random variables  $(U, V)$  with uniform marginal distributions in  $[0, 1]$  and it is given as

$$C(u, v; \theta) = P(U \leq u, V \leq v; \theta), \quad 0 \leq u \leq 1 \quad 0 \leq v \leq 1 \quad (1)$$

where  $\theta$  is a parameter measuring the dependence between  $U$  and  $V$ .

The following theorem by Sklar [1] explains the use of the copula in the characterization of a joint distribution. Let  $(Y_1, Y_2)$  be a bivariate random vector with marginal cdfs  $F_{Y_1}(y_1)$  and  $F_{Y_2}(y_2)$  and joint cdf  $F_{Y_1, Y_2}(y_1, y_2; \theta)$ , then there always exists a copula function  $C(\cdot, \cdot; \theta)$  with  $C : I^2 \rightarrow I$  such that

$$F_{Y_1, Y_2}(y_1, y_2; \theta) = C(F_{Y_1}(y_1), F_{Y_2}(y_2); \theta), \quad y_1, y_2 \in \mathbb{R}. \quad (2)$$

Conversely, if  $C(\cdot, \cdot; \theta)$  is a copula function and  $F_{Y_1}(y_1)$  and  $F_{Y_2}(y_2)$  are marginal cdfs, then  $F_{Y_1, Y_2}(y_1, y_2; \theta)$  is a joint cdf. If  $F_{Y_1}(y_1)$  and  $F_{Y_2}(y_2)$  are continuous functions then the copula  $C(\cdot, \cdot; \theta)$  is unique. Moreover, if  $F_{Y_1}(y_1)$  and  $F_{Y_2}(y_2)$  are continuous the copula can be found by the inverse of Eq. (2):

$$C(u, v) = F_{Y_1, Y_2}(F_{Y_1}^{-1}(u), F_{Y_2}^{-1}(v)) \quad (3)$$

with  $u = F_{Y_1}(y_1)$  and  $v = F_{Y_2}(y_2)$ . This theorem states that each joint distribution can be expressed in term of two separate but related issues, the marginal distributions and the dependence structures between them. The dependence structure is explained by the copula function  $C(\cdot, \cdot; \theta)$ . Moreover the Eq. (2) provides a general mechanism to construct new multivariate models in a straightforward manner. By changing the copula function we can construct new bivariate distributions with different dependence structures, with the association parameter indicating the strength of the dependence, also different from the linear one that characterizes the normal distribution. When  $Y_1$  and  $Y_2$  are discrete random variables taking values on some lattice,  $\Omega$ , the copula  $C$ , is unique in  $(y_1, y_2) \in \Omega$  but not elsewhere. Thus, in the discrete case the mapping from two marginals and a copula  $\{F_1, F_2, C\}$  to a bivariate distribution  $F(Y_1, Y_2)$  is not one-to-one. However, this is not-uniqueness is of no consequence as the region outside  $\Omega$  is not of interest in the discrete case [1]. The representation Eq. 2 and uniqueness follows essentially from a multivariate extension to the probability integral transformation [2].

### 3 Bivariate count models based on copulas

For count data, a common starting point is to use the Poisson distribution for the marginals:

$$f(y; \mu_j) = \mu_j^y e^{-\mu_j} / y!, \quad j = 1, 2 \quad y = 0, 1, \dots \quad (4)$$

where  $\mu_j > 0$ .

The (cumulative) distribution functions are given by:

$$F(y; \mu_j) = \sum_{m=0}^y f(m; \mu_j), \quad j = 1, 2 \quad (5)$$

In case of independence, the joint distribution function, is given by



$$F(y_1; y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2) = F(y_1; \mu_1)F(y_2; \mu_2). \tag{6}$$

Let's couple the marginals to add a dependence structure. Models based on copulas in the case of bivariate counts offer the advantage of allowing easy generalization to several different models which is not easy in general. Take, for instance, the Frank copula:

$$C(u, v; \gamma) = -\gamma^{-1} \log \left[ 1 + \frac{(\exp^{-\gamma u} - 1)(\exp^{-\gamma v} - 1)}{\exp(-\gamma) - 1} \right], \quad \gamma \in R - \{0\}, \quad u, v \in [0, 1]. \tag{7}$$

Then

$$F(y_1, y_2; \mu_1, \mu_2, \gamma) \equiv C(F(y_1; \mu_1), F(y_2; \mu_2); \gamma), \tag{8}$$

is a well defined distribution function with a dependence structure. However for describing the joint probability mass function we need to take differences from the cdf above, and hence the probability mass function (pmf) has to be written as

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2; \mu_1, \mu_2, \gamma) &= F(y_1, y_2; \mu_1, \mu_2, \gamma) \\ &\quad - F(y_1 - 1, y_2; \mu_1, \mu_2, \gamma) \\ &\quad - F(y_1, y_2 - 1; \mu_1, \mu_2, \gamma) \\ &\quad + F(y_1 - 1, y_2 - 1; \mu_1, \mu_2, \gamma) \end{aligned} \tag{9}$$

In the present paper we focus on bivariate models. Copula defined models in higher dimensions are more challenging because generalization of copulas in higher dimensions are not straightforward and estimation can be cumbersome. Several other copulas have been proposed in the literature. For a review of discrete valued models based on copulas see [7].

### 4 Minimum distance estimation

Lindsay [5] has shown that in discrete data, model robustness and efficiency can be achieved almost at the same time, i.e by appropriately defining distances that in some sense downweight some observations. It is shown that minimum distance (MD) estimators can be interpreted (and they are) weighted likelihood estimators, the weights are determined by some kind of distance between observed and expected frequencies. For example, such an estimator can be based on Minimum Hellinger (MH) distance of the form

$$\sum_x \left( d(x)^{1/2} - m_\beta(x)^{1/2} \right)^2$$

where  $d(x)$  is the observed relative frequency (or some other simple estimate of the probability at  $x$ ) and  $m_\beta(x)$  is the assumed model with parameters of interest  $\beta$ . We extend the approach to bivariate count data. Now  $x$  implies a pair of observations. Also, in our case the parameters  $\beta$  to estimate are those of the marginal distribution plus the copula parameter(s). It turns out that this quantity leads to estimating equations of the form

$$\sum_x \left( \frac{d(x)}{m_\beta(x)} \right)^{1/2} \frac{\partial m_\beta(x)}{\partial \beta} = 0$$

directly comparable to the ML estimating equations

Title Suppressed Due to Excessive Length

$$\sum_x \frac{d(x)}{m_\beta(x)} \frac{\partial m_\beta(x)}{\partial \beta} = 0$$

which actually implies that we weight the observations differently (see Lindsay [5]). Here we extend the approach for bivariate count models defined by copulas aiming at deriving robust estimators for both the marginal and the copula parameters. We have also developed an iterative algorithm that facilitates the estimation.

**Remark** In the bivariate case we are interested in, the relative frequencies are still reasonable estimators of the underlying probabilities but we need larger sample sizes for that. As we move on higher dimensions, problems similar to that of the regression setting may occur.

## 5 Simulations

In this work, we investigate the robustness under outliers contamination and model misspecification. Particular focus is on the robustness of copula related parameters that measure the association exhibited by paired count data. The first experiment considers contamination from outliers observations in a bivariate Poisson with a Frank copula (with  $n = 100, 500, 1000$  observations) located in a different regions of the copula support. These observations are unexpected observations under the model.

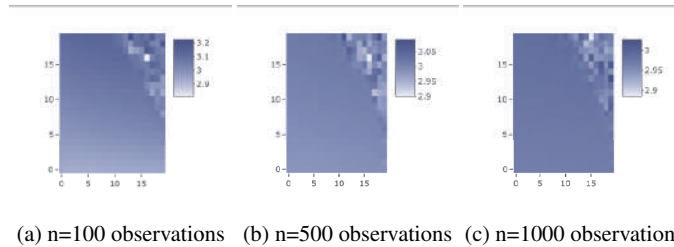


Fig. 1: ML estimator with contamination from outliers observation in bivariate Poisson with Frank copula with parameter  $\theta = 1$  and for the marginal Poisson model with means  $\lambda_1 = \lambda_2 = 3$ . The plots show the value of  $\lambda$  when contaminating the data with one observation located at  $x, y$ .

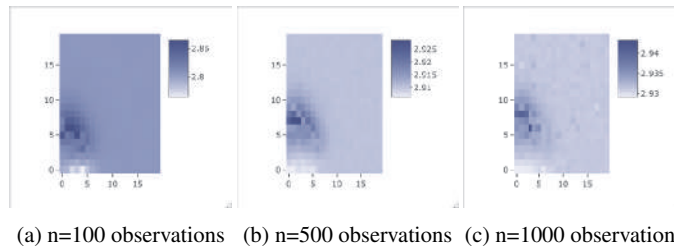


Fig. 2: Same a Figure 1 but now for the MHD estimator.

We compare MHD with Maximum Likelihood (ML) estimator. The simulation results indicate that ML is robust when outliers are closed to the observations. MHD is more robust when the contamination is located quite far from the observed values, this is due to the different way to downweight the observations of MHD.

The second experiment considers  $\varepsilon$ -contaminated bivariate Poisson distribution with Frank copula containing varying proportions  $\varepsilon$  of contaminating points located at different regions at the copula support. The simulation results indicate that MHD always underestimate the copula parameter in all the different settings.

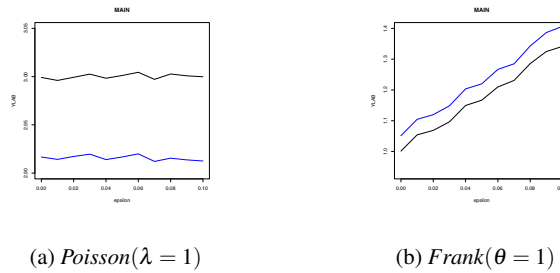


Fig. 3: Bivariate Poisson with Frank copula, the copula parameter is  $\theta = 1$  and the marginal distributions are Poisson with  $\lambda = 1$  with  $n=500$  observations  $\varepsilon$ -contaminated with a Gumbel copula with parameter  $\gamma = 2$ . Black:ML; Blue= MHD.

## References

1. Nelsen, R. B.: An introduction to copulas, 2nd edn, Springer, New York (2006)
2. Joe, H.: Multivariate models and dependence concepts, Chapman and Hall, London (1997)
3. Nikoloulopoulos, A. K.: Copula-based models for multivariate discrete response data. Copulae in Mathematical and Quantitative Finance, pp
4. Grunert Da Fonseca, V., Fieller, N. R. J.: Distortion in statistical inference: the distinction between data contamination and model deviation. *Metrika*, **63**, 169–190 (2006)
5. Lindsay, B. G.: Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Stat.*, **22**, 1018–1114 (1994)
6. Karlis, D., Ntzoufras, I.: Analysis of sports data using bivariate Poisson models. *J. R. Stat. Soc. D*, **52**, 381–393 (2003)
7. Nikoloulopoulos, A. K.: Copula-based models for multivariate discrete response data. Copulae in Mathematical and Quantitative Finance, 231–249 (2013)
8. Kocherlakota, S., Kocherlakota, K. : Bivariate Discrete Distributions (1st ed.). CRC Press. (1992)



## 6 Satellite events

## 6.1 Measuring uncertainty in key official economic statistics

# Uncertainty in production and communication of statistics: challenges in the new data ecosystem

## *L'incertezza nella produzione e comunicazione delle statistiche: le sfide nel nuovo ecosistema dei dati*

Giorgio Alleva<sup>1</sup>, Piero Demetrio Falorsi<sup>2</sup>

**Abstract.** In the paper we focus on how to measure and communicate to users the accuracy in the new data ecosystem, with estimates based on an integrated system of statistical registers, in order to explicitly consider the main sources of uncertainty that can influence the estimates. We propose feasible computational strategies and discuss several approaches useful for communicating accuracy to users.

**Abstract.** Nel paper ci concentriamo su come misurare e comunicare agli utenti l'accuratezza di stime basate su un sistema integrato di registri statistici, considerando esplicitamente le principali fonti di incertezza che possono influenzare le stime. Sono proposte strategie computazionali fattibili e discussi diversi approcci utili per comunicare l'accuratezza agli utenti.

**Key words:** Accuracy, Global Mean Squared Error, Data Integration.

## 1. Introduction

Statistics is the science of data and uncertainty. Uncertainty is the natural environment in which we operate as statisticians. In this paper we emphasize that Official statistics (OS) should make any effort to support users on how to evaluate the accuracy of the estimates on the parameters of interest also making them able to calculate the accuracy of parameters specified by themselves.

---

<sup>1</sup> Sapienza University of Rome, Piazzale Aldo Moro 5, IT-00185 Rome, Italy. Email for contact: [giorgio.alleva@uniroma1.it](mailto:giorgio.alleva@uniroma1.it).

<sup>2</sup> Former Director of Methodology at Istat and international consultant, [piero.falorsi@gmail.com](mailto:piero.falorsi@gmail.com)

Why is it crucial to communicate uncertainty in OS? The users should take their decisions fully aware of the uncertainty of estimates. The quality of decisions may suffer if decision makers incorrectly take reported statistics at face value or incorrectly conjecture error magnitudes.

We should be aware of the importance to accompany systematically our estimates with measure of accuracy: not only point estimates, but as much as possible intervals, associated with a predetermined level of confidence/credibility. We should overcome any fear of creating confusion among users or losing credibility. Of course, tailoring the communication on uncertainty in different ways and languages for different targets. Several studies have shown how good communication of uncertainty can be understood and appreciated by users (Van der Laan et al., (2015)).

Regardless of the source of the statistics, some principles for communicating uncertainty and change apply. Overall, OS should provide sufficient and appropriate information to allow users to judge the goodness of fit for their purpose; and to maintain and increase users' confidence in the estimates. For change estimates, the point is to provide the direction and size, level of uncertainty and the estimated trend. OS has long been equipped in communicating uncertainty of statistics and the Code of Practice include recommendations on this issue, in particular on the necessary transparency in communicating estimates and processes to calculate them. In the sample context, the accuracy and communication measures are more consolidated; also for statistics based on administrative sources OS developed guidelines and indicators. In the last review of the Code of Practice (Eurostat, (2018)) reference is also made to the accuracy in estimates based on the integration of different sources.

The remainder of the paper is organized as follows. In Section 2 the main taxonomies of uncertainty are presented and discussed. The measures of accuracy in different statistical and informative contexts are presented and discussed in Section 3. Conclusions are highlighted in Section 4.

## **2. Does the traditional taxonomies of uncertainty still make sense in OS?**

In the literature there are interesting taxonomies of uncertainty, from which measurement and communication strategies have been developed.

A three-level categorizations have been provided over time by different authors for focusing on statistical modelling of risks (Diebold et al., (2010); Spiegelhalter, (2017)). *Aleatory uncertainty*: the natural randomness in a process, fully expressed by classical probabilities. *Epistemic uncertainty*: the scientific uncertainty about the structure and parameters of our statistical model of a process, expressed, for example, through Bayesian probability distributions, default parameter values, safety factors, and sensitivity analyses to assumptions (Morgan et al., (2009)). *Ontological uncertainty*: unrecognised ignorance about the entire modeling process as a description of reality, or failure to comprehend unprecedented circumstances.

The paper of Manski (2015) provided a new significant boost to the measurement and communication of uncertainty. Starting from the traditional classification of errors in sampling and non-sampling ones, Manski identified three main sources of uncertainty for economic statistics, namely: transitory, permanent and definitional.

*Transitory statistical uncertainty* arises because data collection takes time. Agencies sometimes release a preliminary estimate of an OS in an early stage of data collection and revise the estimate as new data arrive (a typical example is GDP). *Permanent statistical uncertainty* derives from incompleteness or inadequacy of data collection that does not diminish with time. In survey research, considerable permanent uncertainty may stem from non-response and from the possibility that some respondents may provide inaccurate data. *Definitional uncertainty* arises from incomplete understanding of the information that OS provide about well-defined concepts or from lack of clarity in the concepts themselves. Thus, conceptual uncertainty concerns the interpretation of statistics rather than their magnitudes.

When communicating uncertainty, it is interesting to distinguish two fundamental levels of uncertainty (Van der Bles *et al.*, (2019)). *Direct uncertainty* about the fact, number or hypothesis. This can be communicated either in absolute quantitative terms, say a probability distribution or confidence interval, or expressed relative to alternatives, such as likelihood ratios, or given an approximate quantitative form, verbal summary and so on. *Indirect uncertainty* in terms of the quality of the underlying knowledge that forms the basis for any claim about the event, number or hypothesis. This will generally be communicated as a list of warnings about the underlying sources of evidence, possibly blended into a qualitative or ordered categorical scale.

In the new ecosystem of OS, estimates are based on an integrated system of statistical registers fed in a systematic and continuous way by surveys, administrative archives and new sources. Does the distinction between transitional and permanent uncertainty still make sense? Doesn't the revision process typically planned for national accounts during the progressive consolidation of sources now represent the way forward for any statistics produced through an integrated system of registers? How to communicate this sort of continuous review to users? Does it still make sense to distinguish between direct and indirect uncertainty? We have some doubts. In anyhow, an integral part of the challenge of producing OS based on an integrated system of statistical registers is the measurement and communication of uncertainty. The approaches for communicating uncertainty in OS are very different depending on the type of data and information contexts of the dissemination. Leaving out uncertainty is current practice for most of the contexts. De Jonge (2020), underlines that communicating uncertainty in *statistical visualizations*, even if it is not the most widespread practice, adds value to users and clarifies that statistical offices produce statistics. The main approaches for tabular data deriving from survey sampling is that of avoiding dissemination of very unreliable figures, or making the users aware of the uncertainty with specific graphic signals (e.g. an asterisk). Special studies on non-coverage or measurement errors are used for Census data or register-based statistics. Alleva *et al.* (2021) suggest a feasible calculation strategy for register-based statistics allowing a dynamic calculation of the Global Mean Squared Error (GMSE) which



could allow to release the statistics along with the related GMSE, thus improving the relevance, transparency and confidence of official statistics.

### 3. Measuring the uncertainty

Given the need to compute the statistical errors in disseminated data, it is necessary to determine the measure of accuracy to be calculated and communicated to users. To make it simple, we introduce this topic for the total  $Y = \sum_{k \in U} y_k$ , of the variable  $y$  within the population  $U$ , where  $y_k$  is the true value of the variable  $y$  for unit  $k$ . Let  $\hat{Y}$  be the estimation of  $Y$ . There are multiple sources of error (ranging from sampling errors to coverage errors, etc.). Each specific approach to inference focuses on different sources of variability and bias in the definition of the measure of accuracy; these are related to what is treated as fixed or random in the specific inferential approach. For instance, the *design based* (Cochran, (1977)) or the *model assisted* approaches (Särndal *et al.*, (1992)) treat the population values  $y_k$  as unknown constants and the sample selected, with the sample design  $P$ , is the only source of randomness; therefore, they develop their inference considering only the variability of the sampling design. The *model-based* approach (Chambers and Clark, (2015)) considers the sample as *fixed* and the  $y_k$  values as random variables generated according to the model,  $M$ .

If the methodology embedded in the estimator is transparent and does not introduce bias in the estimates, the main advice is to compute at least the leading components of the errors: sampling variance, model variance, or both. *Sampling variance*, which measures the uncertainty deriving from the randomness of the observed set of data, is an adequate measure of accuracy when the construction of statistical indicators is based on the inferential properties of repeated sampling. It may be defined as  $V_P(\hat{Y}) = E_P[\hat{Y} - E_P(Y)]^2$ , where  $E_P$  and  $V_P$  denote the operators of expectation and variance under repeated sampling. *Model variance* is a suitable measure of accuracy when the construction of statistical indicators is based on models using  $x$ -auxiliary variables, generating the value of the target variable  $y$  for the units in the population. Model variance may be defined as  $V_M(\hat{Y}) = E_M[\hat{Y} - E_M(Y)]^2$ , where  $E_M$  and  $V_M$  denote the operators of expectation and variance under the model  $M$  generating the data. However, some statistical indicators can be obtained via statistical procedures that utilize model-based approaches jointly with inference based on sampling design. For these cases, it is suggested to consider *global variance*,  $GV_M(\hat{Y}) = E_P E_M[\hat{Y} - E_P E_M(Y)]^2$  (Wolter, (1985)) as the measure of accuracy.

When statistical data are produced from the census or administrative records, it is also necessary to consider the bias in measuring accuracy. Bias generally derives from the measurement error (based on statistical models) and the coverage error, the latter deriving from erroneous inclusion in the observation of elements extraneous to the population of interest (over-coverage) or from incorrectly excluding certain units from the target population (under-coverage). These types of error can be detected with

special observational techniques (based on double and independent measurements), which may however be costly. In the case of OS, the techniques are implemented only in certain specific cases. Alleva et al. (2021) propose the GMSE as a more general measure of accuracy:  $GMSE(\hat{Y}) = E_P E_M (\hat{Y} - Y)^2$ .

The GMSE includes, as particular cases, the bias and the measure discussed above (the GV, sampling and model variance). Moreover, the GMSE is an extension of the well-known Mean Squared Error (Biemer, (2010)) taking into consideration all the random components involved in the inferential process for computing the statistics. For instance, we may consider the non-response by defining GMSE as:  $GMSE(\hat{Y}) = E_P E_M E_{NR} (\hat{Y} - \hat{Y})^2$ , in which  $E_{NR}$  indicates the expectation under the models adopted for imputing the non-response in survey data. The GMSE could be accepted as a measure of precision by the main professional families of methodologists within the National Statistical Offices (NSOs): at least those who base their inference only on statistical models and those who use the statistical models as a support for inference which continues to be based essentially on sampling design. The global measure has a number of advantageous qualities, including the following: generality, stability over time and robustness in the case of model failures. GMSE is simple to use and to communicate to users. It is based on the first and second moments of the random distributions of the specific source of uncertainty. Its calculus does not imply the full knowledge of the underlying distributions.

As far regards the measurability of the accuracy, we note that Statistical data may be the result of different statistical surveys where, according to the statistical quality framework followed by Statistics Canada (2009), the term survey includes the following components: (i) a census, which attempts to collect data from all members of a population; (ii) a sample survey, in which data is collected from a (usually random) sample of population members; (iii) a collection of data from administrative records, in which data is derived from records originally kept for non-statistical purposes; (iv) a derived statistical activity, in which data is estimated, modelled, or otherwise derived even integrating a multiplicity of existing statistical data sources. Each of the previous components introduces different sources of uncertainty that should be considered when informing the users on the accuracy.

For instance, the component (i) introduces a possibility of coverage errors, which we can deal with specific statistical models. In this case, the adequate measure of accuracy is the GMSE, which incorporates the coverage-bias. An interesting example of how measuring the components of errors in the GMSE is given in Daddi, *et al.* (2021) for the Italian Census Population Coverage Survey (PCS) carried out each year as a component of the Permanent Census Survey System.

The component (ii) includes the sampling errors. According to the inferential process adopted for the predictions, either the sampling variance or the model variance may be adequate. These can be computed with the standard statistical techniques.

The components (iii) and (iv) comprise the uncertainty derived by models adopted for building the predictions at the unit level. Alleva *et al.* (2021) propose computing the GMSE by adopting an approach based on linearization techniques. Scholtus (2019) proposes an approach for computing the GMSE based on replication methods.

## 4. Conclusions

Official statistics should make any effort to support users on how to evaluate the accuracy of the estimates on the parameters of interest. We should overcome any fear of creating confusion among users or losing credibility. In the new ecosystem of OS, with estimates based on an integrated system of statistical registers (ISSR) fed in a systematic and continuous way by surveys, administrative archives and new sources, we propose the GMSE to take into account a plurality of sources of uncertainty. A strategic choice is whether to make the use of ISSR limited and allow the dissemination of only planned outputs having a certified accuracy or make the system more flexible for the users allowing different users to produce their own statistics from the ISSR. We suggest to opt for the second option which makes OS more relevant for its users but which obliges the NSOs to impose a policy for reducing the risk of inappropriate use of the data.

## References

1. Alleva, G., Falorsi, P. D., Petrarca, F., Righi, P. Measuring The Accuracy Of Aggregates Computed From A Statistical Register, *Journal of Official Statistics*. Accepted for publication (2021)
2. Biemer, P.P.: Total Survey Error Design, implementation, and evaluation. *Public Opinion Quarterly*, Vol. 4, No. 5, pp. 817-848. (2010)
3. Chambers, R.L., Clark, R.G.: *An Introduction to Model-Based Sampling with Applications*. Oxford Statistical Science. 37. (2015)
4. Cochran, W. G.: *Sampling techniques*. Third Edition. New York Wiley. (1977)
5. Daddi S., Falorsi P.D., Fiorella E, Massoli P., Righi P., Terribili. M.D. Optimal sampling for the Population Coverage Survey of the new Italian Register Based Census. *Journal of Official Statistics*, Accepted for publication (2021)
6. De Jonge, E.: *Communicating uncertainties in official statistics. A review of communication methods*, European Commission. (2020)
7. Diebold, F.X., Doherty, N.A., Herring, R.J.: *The Known, the Unknown, and the Unknowable in Financial Risk Management: Measurement and Theory Advancing Practice*. Princeton, NJ: Princeton Univ. Press. (2010)
8. Eurostat: *European statistics Code of Practice*, Publ. Office of the EU (2018)
9. Manski, C.: *Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern*. *J Econ Lit*, 53:631–653. (2015)
10. Morgan, G., Dowlatabadi, H., Henrion, M., Keith, D., Lempert, R.: *Best Practice Approaches for Characterizing, Communicating, and Incorporating Scientific Uncertainty in Climate Decision Making*. Silver Spring, MD: Natl. Ocean. Atmos. Organ. (2009)
11. Morgan, G., Dowlatabadi, H., Henrion, M., Keith, D., Lempert, R.: *Best Practice Approaches for Characterizing, Communicating, and Incorporating Scientific Uncertainty in Climate Decision Making*. Silver Spring, MD: Natl. Ocean. Atmos. Organ (2009)
12. Särndal, C.E., Swensson, B., Wretman, J.: *Model Assisted Survey Sampling*. Springer-Verlag (1992).
13. Scholtus, S.: *A bootstrap method for estimators based on combined administrative and survey data*. NTTS Conference 2019 (2019)
14. Spiegelhalter, D.: *Risk and Uncertainty Communication*, *Annu. Rev. Stat. Appl.* 2017.4:31-60 (2017)
15. *Statistics Canada Statistics Canada Quality Guidelines*, 5th edn. (2009)
16. Van der Bles, A.M., Van der Linden, S., Freeman, A.L.J., Mitchell, J., Galvao, A.B., Zaval, L., Spiegelhalter, D.J.: *Communicating uncertainty about facts, numbers and science*, *R. Soc. open sci.* 6: 181870. <http://dx.doi.org/10.1098/rsos.181870> (2019)
17. Wolter, K.M.: *Some Coverage Error Models for Census Data*. *Journal of the American Statistical Association*, 81, 338 - 346 (1986)

# Uncertainty and variance estimation techniques for poverty and inequality measures from complex surveys: a simulation study

*Tecniche di stima della varianza e dell'incertezza per le misure di povertà e disuguaglianza da indagini complesse: uno studio di simulazione*

De Santis Riccardo, Barabesi Lucio, Betti Gianni

**Abstract** Variance estimation is one of the most complex issues in socio-economic surveys, where we face with complex designs and complex statistics. Two main approaches can be found in the literature, the linearization methods and the resampling methods, and both have advantages and drawbacks. In this paper we conduct a simulation study based on a complete population available. We focus on some official poverty measures considered by Eurostat.

**Abstract** *Il tema della stima della varianza è una delle attività più complesse nelle indagini socio-economiche. Nella letteratura possiamo trovare due approcci principali, i metodi basati sulla linearizzazione e quelli basati sul ricampionamento, ed entrambi presentano vantaggi e svantaggi. In questo lavoro abbiamo condotto uno studio di simulazione basato su una popolazione completa disponibile, concentrandoci su alcune misure di povertà considerate da Eurostat.*

**Key words:** Poverty measures, Variance estimation, Simulation study

## 1 Introduction

In the paper we consider variance estimation of poverty and inequality measures in population-based surveys of households and persons, whose main reference is the

---

De Santis Riccardo

Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, Italy, e-mail: [riccardo.desantis.1@phd.unipd.it](mailto:riccardo.desantis.1@phd.unipd.it)

Barabesi Lucio

Department of Economics and Statistics, University of Siena, Piazza San Francesco 7, Siena, Italy, e-mail: [lucio.barabesi@unisi.it](mailto:lucio.barabesi@unisi.it)

Betti Gianni

Department of Economics and Statistics, University of Siena, Piazza San Francesco 7, Siena, Italy, e-mail: [gianni.betti@unisi.it](mailto:gianni.betti@unisi.it)

European Union - Statistics on Income and Living Conditions (EU-SILC) survey. Let  $\theta := \theta(y_1, \dots, y_N)$  be the population parameter; if  $S$  is a random sample of fixed size  $n$ ,  $\theta$  may be estimated as  $\hat{\theta} := \hat{\theta}(\{y_i : i \in S\})$ . Actually, variance estimation for that parameter may be cumbersome for two main reasons: the use of complex designs, which does not allow to know the second-order probabilities, and  $\theta$  may be non-linear. The procedures to estimate the variance of complex statistics can be subdivided into two main approaches, based on resampling or linearization techniques. Both methods present advantages and drawbacks: resampling methods are introduced only in a model-based approach, they may need a massive computational burden, even if the same procedure can be applied for  $\theta$  of any complexity, and standardized routines for the common statistical software are available or may be easily implemented, without the necessity of computing specific quantities for each statistic, a point which can be helpful for researchers. Linearization techniques are well defined in a design-based approach, they need a smaller computational burden, even if they require to compute the linear form for each statistic, which may be a difficult task for researchers, and it may not be unique. Besides, a useful approximation in the case of surveys with large sample and large population size is provided by the “ultimate cluster approach” [14], which consists in a simplification in computing the variance estimation by taking account solely of the variation among Primary Selection Units (PSUs) totals.

Section 2 describes the two main approaches, the Jackknife and the linearization methods. The results of a simulation study are shown in Section 3. Finally, Section 4 contains the conclusions.

## 2 Jackknife and linearization methods

There are many resampling methods presented in the literature [7], [5], mainly the Jackknife Repeated Replication (JRR), the Bootstrap and the Grouped Balanced Method. The concept is to estimate the variance through comparisons among replications generated by repeated re-sampling of the same parent sample. Here we introduce the “Jackknife Delete One PSU” version [15]. For the application we need a sampling procedure where two or more PSUs are selected from each stratum of the population independently, at the first stage, while subsampling of any complexity is allowed within each PSU. Each JRR replication consists in deleting one sample PSU from one particular stratum, increasing the weights of the remaining primary units in that stratum appropriately, and computing the parameter estimate. Consequently, there are as many replications as the amount of PSUs which are present in the sample.

About the linearization methods, the most known method is the Taylor linearization. However, it requires that the statistic is a regular function of estimated totals, continuously differentiable up to order two. As explained by [12] and [11], for many complex statistics - as many poverty measures - this request is not satisfied, therefore a different way to derive the variance estimator has to be found. The main concept

is to obtain a linearized variable  $z_i$  for each observation  $y_i$ ; the consequence is that the variance of the estimator may be approximated by the variance of the linearized variable. There are several methods about the linearization approach [11], but probably the most relevant in the literature is the influence function's approach [6], [12], [3]. This is based on the concept of influence function, which was first introduced in robust statistics by [10].

### 3 Simulation study

We conduct a simulation study to evaluate the empirical properties of the proposed methods, since the exact ones are not known. We apply the simple procedure described in 2 and the naive Bootstrap described in [2], taking account of the presence of stratification and clustering in the sample design.

The data used has been obtained combining the 2011 census of Albania, which contains a limited amount of information for the whole population, with the Albanian Living Standard Measurement Survey (LSMS) of 2012, a multi-purpose survey which collects information to measure poverty and living conditions, in order to simulate the consumption of each household by the use of a methodology named Poverty Mapping [8], [4]. In this way 100 simulations of the population consumption distribution have been obtained, thus the expected value for each household has been taken to obtain the per-capita consumption for each population unit - our variable of interest. We use the consumption as a proxy variable for the income. The population consists of 722,262 households for a total of 2,784,539 individuals, which are subdivided in 24 strata and 11,579 PSUs, defined by the Albanian Institute of Statistics. The strata are obtained by joining the 12 prefectures of Albania with the dummy variable Urban, which indicates whether the household lives in an Urban or a Rural context, while the PSUs correspond to the Census Enumeration Areas, where each includes on average 100-120 households. We have drawn a total of 1,000 samples, following the instructions of the survey LSMS 2012. A two-stage design has been adopted: at the first stage, 834 PSUs have been drawn with a systematic stratified sampling, where the sample size for each stratum has been decided by the Albanian Institute to represent the whole country, while within each stratum the inclusion probability of each PSU is proportional to its number of households contained. At the second stage 8 households are selected within each PSU previously selected, with Simple Random Sampling without replacement. Finally, the sample of 6,672 households is obtained. All the individuals of the households selected are included.

We apply the analysis to four well-known inequality indices. The Poverty Rate with a fixed poverty line (*PR-fix*) represents the proportion of people whose income is below a pre-fixed value, the Poverty rate adopted by Eurostat (*PR-60*) stands for the proportion of people whose income is below the 60% of the median, the Quintile Share Ratio (*Qsr*) represents the proportion between the total income received by the richest 20% of the population, and the total income received by the poorest

20% of the population, while (*Gini*) represents the well-known Gini Coefficient. See [9] for the details of their computation. Note that all measures are considered at individual level, and that the Poverty Rates and the Gini Index are represented with the percentage values. Concerning the measure of accuracy, we consider the square root of the variance - known as Standard Error ( $Se[\hat{\theta}]$ ) - which has the advantage of having the same unit of measure of the point estimator, in such a way that it may give clearer results. In order to comprehend the performance of the standard error estimator ( $\hat{Se}[\hat{\theta}]$ ), we take account of the expected value ( $E[\hat{Se}[\hat{\theta}]])$  and the relative bias ( $RB[\hat{Se}[\hat{\theta}]])$ .

**Table 1** Simulation results for the first population

$\theta$	$Se[\hat{\theta}]$	Jackknife		Bootstrap		Linearization	
		$E[\hat{Se}[\hat{\theta}]]$	$RB[\hat{Se}[\hat{\theta}]]$	$E[\hat{Se}[\hat{\theta}]]$	$RB[\hat{Se}[\hat{\theta}]]$	$E[\hat{Se}[\hat{\theta}]]$	$RB[\hat{Se}[\hat{\theta}]]$
<i>PR-fix</i> 14.300	0.704	0.740	0.051	0.729	0.036	0.736	0.045
<i>PR-60</i> 7.113	0.496	0.687	0.385	0.536	0.081	0.639	0.288
<i>Qsr</i> 3.076	0.068	0.088	0.294	0.069	0.015	0.071	0.044
<i>Gini</i> 22.500	0.523	0.524	0.002	0.506	-0.033	0.522	-0.002

Table 1 reports the outcome of the simulation, where we do not find univocal results. We observe greater differences between measures rather than methods, as we can see for the nearly unbiasedness for the Gini index, opposed to the high upward bias for the measures depending on the median. Generally speaking, the Jackknife seems to give the most conservative estimates.

Again, we decide to apply the simulation study also to a different consumption distribution. This is because, by construction, we have artificially reduced the tails of the distribution above. Therefore, we adopt a Log-normal model [1] with a two-steps procedure. Firstly, the parameters are estimated on each simulation of consumption. Secondly, the consumption for the 722,262 households is generated from the model, whose parameters are equal to the expected value of their estimates over the 100 simulated distribution.

**Table 2** Simulation results for the second population

$\theta$	$Se[\hat{\theta}]$	Jackknife		Bootstrap		Linearization	
		$E[\hat{Se}[\hat{\theta}]]$	$RB[\hat{Se}[\hat{\theta}]]$	$E[\hat{Se}[\hat{\theta}]]$	$RB[\hat{Se}[\hat{\theta}]]$	$E[\hat{Se}[\hat{\theta}]]$	$RB[\hat{Se}[\hat{\theta}]]$
<i>PR-fix</i> 14.300	0.704	0.740	0.051	0.729	0.036	0.736	0.045
<i>PR-60</i> 16.454	0.620	0.865	0.395	0.655	0.056	0.642	0.035
<i>Qsr</i> 4.552	0.104	0.145	0.394	0.117	0.125	0.119	0.144
<i>Gini</i> 29.598	0.479	0.537	0.121	0.524	0.094	0.536	0.119

Table 2 reports the simulation results related to the distribution obtained from the Log-normal model. Here, we observe always conservative estimates, and in almost all the cases the upward bias is greater than the first population. The Jackknife seems sometimes to be a bit unsatisfactory, while the other two methods gives similar result.

## 4 Conclusion

In the ambit of sampling from finite population, we see that the theme of variance estimation can be faced safely with different approaches, for many measures, even if we meet some problematic. In presence of complex surveys, and complex measures, some approximations are required. The purpose is to get an unbiased variance estimator, if it exists. Otherwise, we look for getting an estimator which is not downward biased. Therefore, after having shown the main methods, we decide to focus on some well-known and easily implementable techniques, Jackknife Repeated Replication, Linearization and Bootstrap, making an interesting comparison.

The results say that we do not have a method which has always a major reliability. We see different behavior for each statistic, and also a different bias between the two populations. The Jackknife seems to be more conservative and sometimes more unstable, while - in the first population - the Bootstrap and the Linearization gives sometimes a systematic underestimation. Finally, we can conclude that there is not a clear superiority of any approach over the others, and the preference for one method may be influenced also by practical considerations.

## References

1. Aitchinson, J. and Brown, J.A.C. : The Lognormal distribution with special reference to its uses in economics. Cambridge University Press (1969)
2. Alfons, A. and Templ, M.: Estimation of social exclusion indicators from complex surveys: the R package Laeken. *Journal of Statistical Software*, **54(15)** 1–25 (2013)
3. Barabesi, L., Diana G. and Perri P. F.: Linearization of inequality indices in the design-based framework. *Statistics*, **50(5)**, 1161–1172 (2016)
4. Betti, G., Bici, R., Neri, L., Sohnesen, T.P. and Thomo, L.: Local poverty and inequality in Albania. *Eastern European Economics*, **56**, 223–245 (2018)
5. Davison, A.C. and Hinkley, D.V.: *Bootstrap methods and their application*. Cambridge University Press (1997)
6. Deville, J. C.: Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, **25(2)**, 193–203 (1999)
7. Efron, B.: *The Jackknife, the Bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics, Philadelphia (1982)
8. Elbers, C., Lanjouw J. O. and Lanjouw, P.: Micro-level estimation of poverty and inequality. *Econometrica*, **71(1)** 355–364 (2003)
9. European Commission, Eurostat: *Laeken indicators. Detailed calculation methodology*. DOC. E2/IPSE/2003 (2003)
10. Hampel, F. R.: The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383–393 (1974)
11. Langel, M. and Tillé, Y.: Variance estimation of the Gini index: revisiting results several times published. *Journal of the Royal Statistical Society*, **176**, 521–540 (2013)
12. Osier, G.: Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, **3(3)**, 167–195 (2009)
13. Osier, G., Berger, Y. G. and Goedemé, T.: Standard error estimation for the EU-SILC indicators of poverty and social exclusion. *Eurostat Methodologies and Working Papers*, Eurostat, Luxembourg, (2013)



14. Särndal, C. E., Swensson, B. and Wretman, J.: Model assisted survey sampling, Springer Series in Statistics, New York (1992)
15. Verma, V. and Betti, G.: Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics*, **38**, 1549–1576 (2011)

# Pandemics and uncertainty in business cycle analysis

## *La pandemia l'incertezza nell'analisi congiunturale*

Jacques Anas, Monica Billio, Leonardo Carati, Gian Luigi Mazzi, Hionia Vlachou

**Abstract:** We analyse the impact of pandemic induced uncertainty on business cycle analysis and we describe actions undertaken to mitigate its effects. Results confirm that mitigating measures have preserved the stability of business cycle signals.

*Abstract:* Analizziamo gli effetti dell'incertezza indotta dalla pandemia sulle valutazioni congiunturali nonché le azioni intraprese per mitigarne gli effetti. I risultati confermano che le misure intraprese hanno preservato la stabilità dei segnali congiunturali.

**Key words:** Business cycle analysis, Uncertainty, Pandemics

## 1 Introduction

This paper originates from our experience in providing Eurostat with an integrated and comprehensive statistical framework for business cycle analysis. This framework, based only on official statistics, provides a simultaneous monitoring of three main cycles: the business, the growth and the acceleration ones. It is split into two exercises: the dating (quarterly update of historical turning points chronologies based on a non-parametric dating rules; and the detecting (a monthly real-time detection of turning points based on probabilistic models). The simultaneous monitoring of the three cycles is based on the so called  $\alpha AB\beta CD$  approach [1] and [2] where  $\alpha, \beta$ , A, D and B, C represent peaks and throats respectively of the acceleration, growth and business cycles. A detailed description of our system is presented in [2]. This system constitutes

---

<sup>1</sup> Jacques Anas, Senior consultant; email: [Jacques.Anas@free.fr](mailto:Jacques.Anas@free.fr)  
Monica Billio, University Ca' Foscari of Venice; email: [billio@unive.it](mailto:billio@unive.it)  
Leonardo Carati, Consultant, email: [leonardigno89@yahoo.it](mailto:leonardigno89@yahoo.it)  
Gian Luigi Mazzi, Senior Consultant, email: [glmazzi@pt.lu](mailto:glmazzi@pt.lu)  
Hionia Vlachou, GOPA Luxembourg, email: [hionia.vlachou@gopa.lu](mailto:hionia.vlachou@gopa.lu)

the main engine of the business cycle clock, available on the [Eurostat website](#) and described in [3], [6] and [7]. As for all statistics, uncertainty is inherent to business cycle indicators and affects the proper interpretation and communication of cyclical signals. Together with traditional uncertainty sources such as data uncertainty (E.G. transitory uncertainty) and model uncertainty (E.G. specification, estimation, etc.), an additional one characterizes business cycle analysis related to the difficulties in interpreting business cycle signals. Past recessions have shown that the uncertainty tends to increase in proximity of turning points challenging their provisional dating and detecting. The pandemic crisis confronted analysts with a completely new scenario with extraordinary high level of uncertainty, a never observed before cyclical configuration and further difficulties in interpreting the current situation due to exogeneity of shocks. This paper shows actions undertaken to mitigate uncertainty effects to preserve a high level of accuracy of our indicators.

## ***2 Effect of uncertainty in the dating and detecting exercises***

In this section, we analyze the impact of the high uncertainty on the dating and the detecting exercises.

### ***2.1 Dating exercise***

Dating can be performed either by parametric or non-parametric tools and it is characterized by uncertainty which is difficult to quantify. We could complement point estimates with confidence intervals or density, even if this implies to abandon the non-parametric dating approach we are privileging. Furthermore, in presence of very short cycles and phases the possibility of partially overlapping intervals could increase confusion among user. In our dating strategy, we make a difference between the final and the provisional dating, which refers to the most recent period (around 3 to 4 years). Our past experience indicates that provisional business cycle dating is generally not revised a lot while growth and acceleration cycles are. Additional details on the complexity of the dating are provided in [4]. However, due to the very high data volatility and the strategy adopted for seasonal adjustment we cannot exclude to revise also the business cycle provisional dating. With the pandemics, another type of uncertainty of a more economic nature appears, i.e., the superimposition of economic and non-economic shocks. The issue of the determination of the start of the recession is a straightforward question. For example, GDP growth started to decrease already in the second half of 2019 in some countries of the euro area, but we are not aware if this was the start of a recession without the pandemics. On the contrary, an industrial recession started before the pandemic. Three main features adding complexity to the dating: the sudden, brief but dramatic impact of a pandemic wave through restrictive lockdown measures, the succession of waves provoking a series of economic effects,

Pandemics and uncertainty in business cycle analysis

the tough worldwide nature of the pandemic. We could consider the substantial atypical brief decrease in activity as a pure outlier, but the repeated waves and the worldwide contamination made this shock feel like an actual cycle. In such a complex situation with several competing source of uncertainty, the role of expert judgment has been crucial to identify a sound economic and statistical interpretation of facts. Finally, the need of providing reasonable estimates of the growth cycle before dating represented another important problem to face with. At the beginning of 2020 there was a consensus on treating the pandemics as an outlier and a conservative approach was generally followed because mainly of its uncertain evolution. After several months, it is still unclear which is the best strategy to be followed like removing any kind of outlier, considering the pandemics as a level shift, etc.

## 2.2 *The detecting exercise*

Our detecting system is based on two set of turning points coincident indicators: a multivariate indicator for the simultaneous detection of growth and business cycles turning points based on Markov switching (MS) VAR; and a univariate MS model for detecting acceleration cycle turning points. More details on recent improvements in our detecting system are provided in [5]. The latter indicator has been the first one suffering of the pandemics effect. The acceleration cycle coincident indicator (ACCI) is based on a simple univariate MS model, where the only involved variable is the [Economic Sentiment Indicator \(ESI\)](#). Recently, the procedure used to estimate the model underpinning the ACCI did not converge. This is due to the sharp decline observed in the ESI starting in April 2020 reflecting the very high level of uncertainty perceived by managers and households. This decline resulted first in a great instability of the ACCI and then in the lack of convergence of the model. In the ACCI model the ESI entered after a transformation (1-month change of the 6-month difference) that experienced a drop corresponding to more than 10 times its standard deviations in April 2020. This affected significantly the stability of the signal delivered by the ACCI. In order to cope with this limitation we decided to re-specify the MSI(3)-AR(0) model used so and move to a state-dependent heteroscedastic model of the form: MSIH(3)-AR(0). Further, to match as closely as possible the ACCI released so far, the definition of the ACCI has been changed assuming that the deceleration probability is equal to the sum of filtered probabilities of the first two regimes, instead of the one of the first regime alone as considered before. With this specification, the convergence problem has been fixed without adversely affecting the performance of the ACCI in detecting the historical turning points.

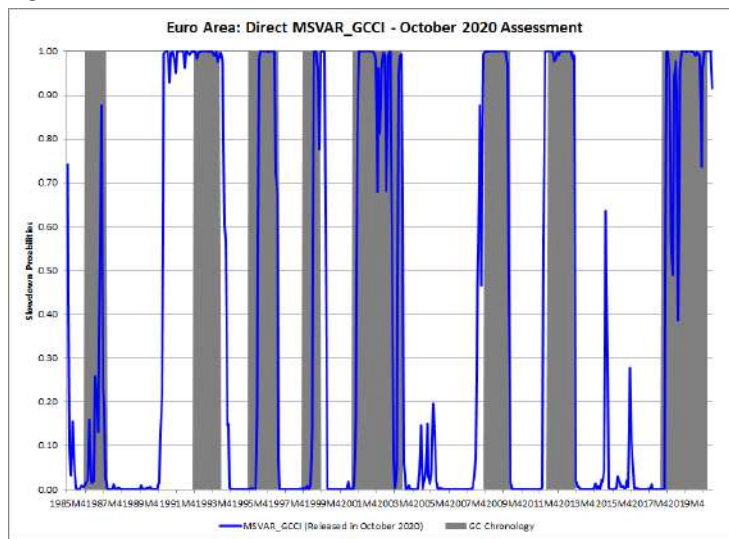
A similar case concerned also the multivariate coincident indicators for growth and business cycles, based on a MS-VAR model, due to the extreme volatility of one of the component variables, namely the [Industrial production index \(IPI\)](#). The IPI for the euro area has shown large drawdown in April 2020 followed by a similar rebound in August 2020 so that its volatility was significantly raising. In such a situation, the Expectation Maximization (EM) algorithm used to estimate the MS-VAR model did

not converge anymore. To fix this problem, which was compromising the reliability of recession signals returned by the model, we decided to trim the IPI in order to reduce its volatility. In such a way, without any model re-specification, the EM algorithm converged again returning realistic recession signals. An alternative approach could have been adding one or more outliers to the MS-VAR model however, since the two approaches look very similar and after a small comparative exercise, we have decided that the former was preferable because easier to be implemented in a short time period without changing the model specification. Summarizing our coincident indicators, even in very extreme situations, have shown, to be very resilient as also demonstrated by their concordance indices with the corresponding dating chronologies.

### 3 Empirical results

Figure 1-3 show respectively the filtered probabilities of the MS-VAR GCCI before the convergence problem, with trimmed IPI and with outlier utilization. The comparison shows the pattern in figures 2 and 3 they are very similar which confirms the equivalence of the two approaches. Finally, the signals compared with figure 1 look very stable and reliable.

Figure 1: MS-VAR GCCI as in October



Pandemics and uncertainty in business cycle analysis

Figure 2: MS-VAR GCCI as in November 2020 with trimmed IPI

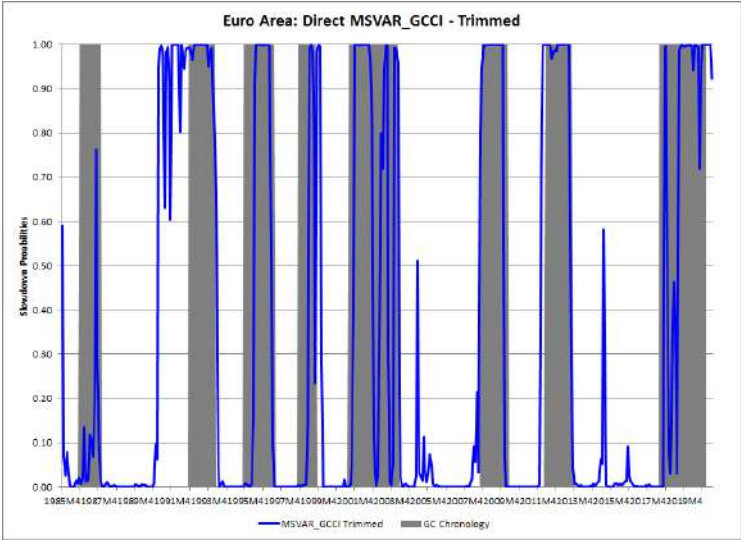
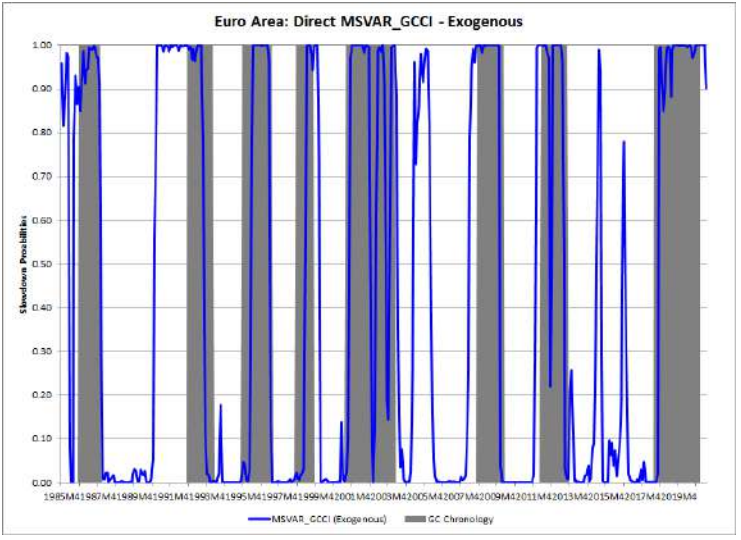


Figure 3: MS-VAR GCCI as in November 2020 with outlier



## 4 Conclusions

In this paper, we have discussed the effects of uncertainty on business cycle analysis with special attention to the pandemics. We have also presented some actions undertaken to mitigate such effects. Nevertheless, besides mitigating the uncertainty effect, it is clear that users should be informed about the inherent uncertainty also in our analysis. How far we should go still remains an open question. Although providing numerical measures of uncertainty (confidence intervals, densities, etc.) is a totally reasonable approach in almost all statistics, we have some perplexities in this field. Especially if cycles are very short and the uncertainty is pretty high the risk of partially overlapping confidence intervals is real with the associated risk of further confusing users. By contrast, we consider that developing a strategy of verbal communication for business cycle analysis could be particularly beneficial for users.

## References

1. Anas J., Billio M., Carati L., Ferrara L. and Mazzi G.L. Cyclical composite indicators detecting turning points in: Handbook of cyclical composite indicators Ed. By G.L. Mazzi and A. Ozyildirim – Eurostat (2017)
2. Anas J., Billio M., Ferrara L. and Mazzi G.L. A System for dating and detecting cycles in the euroarea in: The Manchester School bulletin (2008)
3. Anas J., Mazzi G.L., Paracho-Soares C., Rieser D. and Ruggeri-Cannata R. An overview of existing business cycle clock applications in: Handbook of cyclical composite indicators ed. By G.L. Mazzi and A. Ozyildirim – Eurostat (2017)
4. Anas J. and Mazzi G.L. The complexity of the dating exercise in: proceedings of the NTTS 2021 Eurostat (2021)
5. Billio M., Carati L., Mazzi G.L. and Vlachou H. Towards the improvement of Eurostat's turning points coincident indicators in: Proceedings of the NTTS 2021 Conference – Eurostat (2021)
6. Mazzi G.L. Complementing scoreboard with composite indicators: the new Eurostat business cycle clock in: eurona – Eurostat (2015)
7. Ruggeri-Cannata R. The Eurostat Business Cycle Clock: A Complete Overview of the Tool' in: Statistical Journal of the IAOS, Vol 37 (2021)

## 6.2 Covid-19: the urgent call for a unified statistical and demographic challenge



## 6.2.1 Environmental epidemiology and the Covid-19 pandemics

# The Covid-19 outbreaks and their environment: The Valencian human behaviour

## *I focolai di Covid-19 e il loro ambiente*

Xavier Barber, Elisa Espín, Lucia Guevara, Aurora Mula, Kristina Polotskaya and Alejandro Rabasa

**Abstract** One of the forgotten methods to study the spread of Covid-19 is the behaviour of the outbreaks along the time. From a road bar to a 24 hour convenience store, from a hospital to a bus stop, from industrial states to teleworking, without forgetting Christmas, new year eve, the three wise men, and other special events. This set of factors calls into question the randomness of Covid-19 outbreaks in the Valencian region. We study different factors as temperature or sun hours in order to find the correlation between the outbreaks and their spread. We observe some patterns of spread depending on temperature and social behaviour.

**Key words:** Covid-19, Temperature, Human behaviour, Social contacts, Generalized additive models

## 1 The Covid-19 outbreaks

We define an outbreak when in the trace of a case the healthcare system detects two or more cases which are directly related (at the end 3 or more cases from the same origin). We categorize 3 types of outbreaks: occupational (people who work together), social (family or friends in a social event at home or restaurant) and other (specially elder people in nursing homes).

There is a large list of papers that try to study the relationship between the Covid-19 and temperature<sup>3</sup>, but in some Mediterranean areas this pattern could not be found clearly<sup>1</sup>.

Another point that is studied is the human interaction, and this is a big point for Mediterranean people, and for a country as Spain with a high ratio of bars per person (1/175 habitants, with more than 277.539 establishments). This point is important in the social outbreak spread, because people don't use mask cover while they consume

---

Xavier Barber, Elisa Espín, Lucia Guevara, Aurora Mula, Kristina Polotskaya, Alejandro Rabasa  
Center of Operation Research Institute, Miguel Hernandez University e-mail: xbarber@umh.es

and do not stop talking, laughing and shouting. It is well known that the aerosol spread of the virus is the main cause of coinfection, making social events even more dangerous than occupational risks<sup>4</sup>.

We try to analyse the daily and weekly outbreaks (of type social, occupational and other) in all Spanish regions and especially in the most relevant (inland or coastal) cities. In order to study this we collect daily information from September 2020 to February 2021 from regional government and newspapers (sometimes the latter have better information than government web pages) including: city (name, latitude and longitude), number of cases, temperature, wind velocity, % of humidity and sun hours (from the closest meteorological station).

## 2 Methods and models

In order to study the total amount of cases we use different models, some simpler than others, where the response variables was 'Number of Cases'.

We use generalized linear model, generalized linear mixed model and generalized additive mixed model (GAMM) with autocorrelated error, always using the family Poisson for the response variable distribution in a frequentist approach<sup>5</sup>.

In the mixed models we try to use City and Week (using number of epidemiological week, in 2020 there were 53 weeks). Unfortunately we can't prove the interaction City:Week in the mixed model because this approach produced infeasible models.

Finally we tried a Bayesian approach to study the same dataset, in order to check the robustness of findings, by using a temporal model with spatial dependence. The spatial dependence was studied in different ways. The easy way was the correlation between the number of cases in the cities that shared the industrial estates. Another method was a test that studied if the outbreaks locations were randomly distributed or not, using a point processes approach as Log Gaussian Cox Processes or a spatio-temporal model when studying the evolution with the presence or absence (or abundance) of outbreaks as a Species Distribution Model.

## 3 Results

In all the models there was a main pattern that was that the minimum temperature was relevant, but not the minimum temperature the same day or the day before if not the temperature of three or more days before the outbreak.

The hours of sun was another variable that appears as relevant in all the models, and this variable confirms that the human behaviour is probably the main cause of the spread.

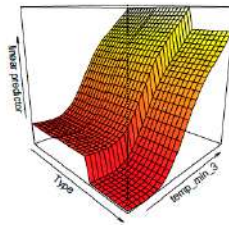
At the end, the non-pharmacological intervention (NPI) imposed by the regional government, has been decisive in stopping the third wave by reducing public and private social events as much as possible.

Model	$\phi_{GAMM}$	$\phi_{INLA}$
All cities	0.28	0.84
Elx	0.01	0.16
Valencia	0.23	0.37
Sagunt	0.01	0.26
Castelló	0.14	0.09
Alcoi	0.22	0.07

**Table 1** Value for the correlated errors with structure AR(1) for the days using generalized additive mixed models and Bayesian temporal models with INLA<sup>2</sup>

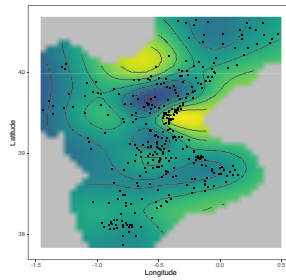
Regarding the individual effect of co-variables when we studied the cities with more outbreaks the results are very similar to the complete model but in some cities with a few differences (see Table 1 to show the AR(1) coefficients for the autocorrelated residuals). It seems that the Bayesian model gives more importance to the autoregressive component than the GAMM.

In Figure 1 we show the perspective plot views of GAMM model predictions, fixing all but the values in view to the values supplied in the co-variables selected. In this figure we can observe the relationship between the lag three days minimum temperature and the type of outbreak and how a greater values of temperature number of cases, again social live style can explain this pattern.



**Fig. 1** The representation of the GAMM model. The linear predictor in the Z axis and the 3 days lag minimum temperature and the type of outbreak in the X,Y axis.

Concerning the spatial effect, Figure 2 shows some relevant zones where the outbreaks were more recurrent.



**Fig. 2** The representation of the GAMM model with spatial information. In yellow two important cities, first Valencia the greatest city which had more outbreak, and the second yellow zone is Requena a place with a great amount of outbreaks in nursing homes.

## 4 Conclusion

Although it seems contradictory that at higher temperatures we observe a higher number of outbreaks, we must remember that this is associated with a higher level of social contacts and longer time without a mask, thus increasing the risk of spread. To all this, we have to add that there have been days that are traditionally very familiar in Spain, that means that the outbreaks have increased so much in the last weeks of 2020 and this explains why the beginning of the third wave in Spain has been so virulent.

Regarding spatial patterns, a clear relationship has been seen with areas with a better climate (coast) compared with the center of the region with a harsher climate in order to have closer social or work contacts.

## References

- [1] Álvaro Briz-Redón and Ángel Serrano-Aroca. A spatio-temporal analysis for exploring the effect of temperature on covid-19 early evolution in spain. *Science of the total environment*, 728:138811, 2020.
- [2] Virgilio Gómez-Rubio. *Bayesian inference with INLA*. CRC Press, 2020.
- [3] Narges Nazari Harmooshi, Kiarash Shirbandi, and Fakher Rahim. Environmental concern regarding the effect of humidity and temperature on 2019-ncov survival: fact or fiction. *Environmental Science and Pollution Research*, pages 1–10, 2020.
- [4] Leonardo Setti, Fabrizio Passarini, Gianluigi De Gennaro, Pierluigi Barbieri, Maria Grazia Perrone, Massimo Borelli, Jolanda Palmisani, Alessia Di Gilio, Prisco Piscitelli, and Alessandro Miani. Airborne transmission route of covid-19: why 2 meters/6 feet of inter-personal distance could not be enough, 2020.
- [5] Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

## 6.2.2 Estimation of Covid 19 prevalence

# Optimal spatial sampling for estimating the SARS-Cov-2 crucial parameters

## *Campionamento spaziale ottimale per la stima dei parametri cruciali SARS-Cov-2*

Piero Demetrio Falorsi<sup>1</sup>, Vincenzo Nardelli<sup>2</sup>

**Abstract.** The data monitoring SARS-CoV-2 pandemic evolution has generally been collected without a statistical design. In a recent paper, Alleva *et al.* (2021a) proposed a two-stage sample design to build a surveillance system designed to correctly quantify the number of infected people in the recent SARS-CoV-2 pandemic. The proposed method exploits the indirect sampling. In this paper, we extend the proposal to include a spatial sampling mechanism in the process of data collection to achieve the same level of precision with fewer sample units, thereby facilitating the process of data collection in a situation where timeliness and costs are crucial elements.

**Abstract.** *I dati che monitorano l'evoluzione della pandemia SARS-CoV-2 sono stati generalmente raccolti senza un disegno statistico. In un recente articolo, Alleva et al. (2021a) hanno proposto un disegno campione in due fasi per costruire un sistema di sorveglianza progettato per quantificare correttamente il numero di persone infette nella recente pandemia di SARS-CoV-2. Il metodo proposto sfrutta il campionamento indiretto. In questo documento, estendiamo la proposta per includere un meccanismo di campionamento spaziale nel processo di raccolta dei dati per ottenere lo stesso livello di precisione con un minor numero di unità di campionamento, facilitando così il processo di raccolta dei dati in una situazione in cui tempestività e costi sono elementi cruciali.*

**Key words:** balanced sampling, spatial sampling, SARS-CoV-2.

---

<sup>1</sup> Piero Demetrio Falorsi, Former Director of Methodology at Istat and International Consultant, piero.falorsi@gmail.com

<sup>2</sup> Vincenzo Nardelli, University of Milan-Bicocca, Milan, v.nardelli2@campus.unimib.it

## 1. Introduction

The SARS-CoV-2 pandemic has affected in Western countries in a sudden and devastating way. In most cases, the data describing its evolution were generally collected without an organic design. To tackle the pandemic, non-pharmaceutical interventions (NPIs) (such as lockdown and social distancing) have been successfully adopted and, in most cases, generalized for the entire population. In a recent paper (Alleva *et al.*, 2021a), we proposed a sampling design for building a continuous-time surveillance system to assess the prevalence of infected people in the population. The quoted methodology does not contain any reference to the spatial correlation. The present work builds upon the contribution of Alleva *et al.* (2021a) and extends it by considering two-stage spatial sampling designs. Indeed, geographically distributed data are typically characterized by significant spatial correlation, which is especially true if they refer to contagious diseases (Cliff *et al.*, 1981). Many examples of sampling plans can be found in the literature based on the use of contiguous units (see Stevens and Olsen, 2004; Arbia *et al.*, 2007; Tillé *et al.*, 2018; Fattorini *et al.*, 2020). In particular, we expand the results of Alleva *et al.* (2020) using local pivotal methods (Grafström *et al.*, 2012; Deville and Tille, 1998) and local cube methods (Chauvet and Tillé, 2006) to produce more parsimonious sample plans in terms of the sample size. The proposed sampling scheme is a general two-stage mechanism. In the first stage, the clusters are drawn by following an explicit spatial sampling design. In particular, we consider the cube algorithm (Chauvet and Tillé, 2006) with different balancing elements and spatial sampling (Tillé, 2020). The remainder of the paper is organized as follows. A review of the previous proposals on Covid 19 monitoring is presented in Section 2. The basic sampling framework together with the specific first-stage sampling design are described in Section 3. Conclusions are highlighted in Section 4.

## 2. Previous proposals for monitoring Covid 19 and the importance of spatial sampling

Starting from the observation that official medical swabs are directed mainly to infected people and therefore they largely underestimate the number of infected people, Alleva *et al.* (2021a) proposed a sampling strategy aiming at estimating the total population of infected, including those that show no symptoms so as to produce a better estimation of mortality and lethality rates. The proposed procedure preliminarily involves dividing the population into two groups. The first group (say Group A) consists of all the individuals who have a verified state of infection together with those who had contact with them. The second group (say Group B), in contrast, contains both healthy people and those who are still in a phase of incubation with no symptoms. Obviously, the proportion of the infected people is much larger in the first than in the second group so that focusing on their contacts maximizes the number of infected people included in the sample. However, also the second group is necessary to correctly estimate the epidemic parameters. For both groups, Alleva *et al.* (2021a) proposed two distinct methodologies based on the idea of *indirect sampling*, (Lavalle, 2007; Kiesl, 2016). In particular, a sample from Group A is drawn without replacement partly from the whole population of the infected and partly without replacement from all their contacts. Furthermore, the design also included a traditional population panel survey with sample rotation aimed at inspecting group B and associated with an indirect sampling mechanism so as to trace and sample the individuals who came into contact with the infected people. The competitive



Optimal spatial sampling for estimating the SARS-Cov-2 crucial parameters

advantage of Alleva et al. (2021a) proposal with respect to other sampling plans (e. g. those based exclusively on indirect sampling or only on the panel sample), relies mainly on the combination of the two sampling strategies which produce a flexible tool designed to closely monitor the dynamic of the epidemics in their different phases. In the quoted paper the authors proved formally the unbiasedness of their proposed design and its greater efficiency with respect to the simple random sampling and reinforced their conclusions with a set of simulations.

The present work builds upon the contribution of Alleva *et al.* (2021a) and extends it considering a spatial sample design both in the first and in the second group. Our aim is to further increase the estimator's efficiency so as to achieve the same level of precision with fewer sample units and to reduce the number of the tests and their associated costs which is of important concern in the phase of epidemic monitoring and surveillance.

In the statistical literature, we refer to *spatial sampling* design as to the process through which observations are collected in a two-dimensional framework with a specific attention to their location (Müller, 2007; Wang et al., 2012). Generally speaking, a spatial sampling scheme is designed to maximize the probability of capturing the spatial variability of the various empirical phenomena. Indeed, geographically distributed data are usually characterized by significant spatial correlation especially if they refer to contagious diseases (Cliff et al, 1983). In this respect, spatial sampling is distinct from conventional sampling where data are assumed to be independent and identically distributed. The uncertainty for spatial sample estimation propagates in the stochastic field to sample distribution and to the statistical tools used to obtain an estimate and judge their quality (Tillé et al., 2018). A rather popular spatial sampling strategy, initiated by Arbia (1993), consists in exploiting spatial correlation to maximize the information content while minimizing the sample size thus reducing the overall costs. This strategy was termed DUST after the acronym of Dependent Areal Units Sequential Technique and it was inspired by model-based assumptions on the dependence of the random field generating the data. The method was characterized by variable inclusion probabilities at each step (Brewer & Hanif, 1983) and extended the idea of the Balanced Sampling design Excluding Contiguous Units (BSEC) presented in Hedeyat *et al.* (1988) for unidimensional data. More recently, some new methodologies have been proposed that explicitly use the distances between the point-level locations of units in the selection procedure such as the Local Pivotal Methods (Grafström et al., 2012) and the Spatially Correlated Poisson Sampling (Grafström, 2012). These methods are implemented in the R packages 'BalancedSampling' (Grafström, 2014) and 'sampling' (Matei & Tillé, 2005). Two alternative procedures to select samples with fixed probability of inclusion and correlated inclusion probabilities derived as an extension of the Pivotal methods introduced by Deville and Tille (1998). They are based on an updating rule of the probabilities of inclusion that, at each step, locally keep the sum of the probabilities as constant as possible. In our proposal which will be presented in the next section, we will expand the results of Alleva et al. (2021a) by introducing these spatial sample designs in the target population group.

## 2. The basic sampling framework

Before presenting our proposed sample design, let us define  $U$  as the population of interest of size  $N$  and suppose that it can be partitioned into  $M$  subpopulations, called clusters (or areas) denoted as  $U_1, \dots, U_i, \dots, U_M$ . The set of clusters is symbolically represented as  $U_1 = \{1, \dots, i, \dots, M\}$ . Cluster  $U_i$  has  $N_i$  units, with  $N = \sum_{i=1}^M N_i$ .

Let  $v_{ij}$  ( $i = 1, \dots, M; j = 1, \dots, N_i$ ) be a dichotomous variable  $v$  that assumes a value of 1 if person  $j$  of cluster  $U_i$  has a verified state of infection (either hospitalized or in compulsory quarantine) and a value of 0 otherwise. Let

$$(1) \mathcal{V}_i = \sum_{j=1}^{N_i} v_{ij} \text{ and } \mathcal{V} = \sum_{j=1}^M \mathcal{V}_i$$

be the known totals of the verified infected in cluster  $U_i$  and in population  $U$ . The quantities  $\mathcal{V}_i$  and  $\mathcal{V}$  are generally known

Let  $y_{ij}$  be the value of variable  $y$  for person  $ij$ , where  $y$  is equal to 1 if the person is infected and 0 otherwise. If  $v_{ij} = 1$ , then  $y_{ij} = 1$ ; however, if  $v_{ij} = 0$ , then it is possible that either  $y_{ij} = 1$  (an infected person for whom the infection has not yet been verified) or  $y_{ij} = 0$  (a healthy person). The target parameter,  $Y$ , is the total number of infected people, that is

$$(2) Y = \sum_{i=1}^M \sum_{j=1}^{N_i} y_{ij} = \sum_{i=1}^M Y_i,$$

where  $Y_i$  indicates the number of infected people of cluster  $i$ .

We select a sample  $S$  by a general two-stage sampling design without replacement. A first-stage sample,  $S_1$ , of fixed size  $m$  is selected without replacement from  $U_1$ , with inclusion probabilities  $\pi_{1i}$  ( $i = 1, 2, \dots, M$ ). A standard solution is to select cluster  $i$  with a probability proportional to its size (PPS):

$$(3) \pi_{1i} = m \frac{N_i}{N}.$$

Clusters can be sampled with different algorithms, thus leading to specific first-stage sampling designs. The basic techniques considered here are the cube algorithm (Chauvet and Tillé, 2006) and spatial sampling (Tillé, 2019). The cube algorithm ensures that the first-stage Horvitz-Thompson (1952) estimates reproduce (at least approximately) the known characteristics of some auxiliary variables available for population  $U_1$ . Spatial sampling avoids the joint selection of units that are positively correlated. As the correlated units are generally geographically close, the sample must be well spread in the territory.

The first-stage sampling designs we consider are:

- (i) **The fixed-size probability proportional to size (FPPS) sampling design.** We can implement this design by the cube algorithm, collapsing balancing variables  $\mathbf{d}_i$  to scalar value  $\pi_{1i}$ .
- (ii) **The cube method based on the verified infected (CBV).** This method ensures that the first-stage sampling size is fixed and that first-stage HT estimates of the total number of verified infected individuals reproduce the known totals. The balancing variables are  $\mathbf{d}_i = (\pi_{1i}, \mathcal{V}_i)'$ .
- (iii) **The local pivotal (LP) method.** This method (Grafström, 2012) uses the distance between units to create small joint inclusion probabilities for nearby units, forcing the samples to be well dispersed. The sample is fixed in size, and the first-stage inclusion probabilities are equal to the planned  $\pi_{1i}$  ( $i = 1, \dots, M$ ).
- (iv) **The local cube method based on verified infected (LCBV).** The local cube method (Grafström and Tillé, 2013) selects a sample that is both spread out geographically and balanced in terms of auxiliary variables. The balancing equations of the LCBV method are the same as those of the CBV method.
- (v) **The local cube method based on the geography of the phenomenon (LCBG).** In this case, the balancing variables are given by a transformation of the geographical coordinates of the clusters. Therefore, the sample is well spread and reproduces the known spatial distribution of  $U_1$ .

A second stage sample,  $S_{1i}$ , of fixed size  $\bar{n}$  is selected from sample cluster  $i$  by *drawing the units* without replacement via a simple random sampling procedure. The second-stage inclusion probability  $\pi_{11i}$  of people in the sampled cluster  $i$  is

Optimal spatial sampling for estimating the SARS-Cov-2 crucial parameters

$$(5) \pi_{1i} = \frac{\bar{n}}{N_i}.$$

The final inclusion probability of person  $j$  being selected from cluster  $i$  is

$$(6) \pi_{ij} = \pi_i \pi_{1i} = m \frac{N_i \bar{n}}{N N_i} = m \frac{\bar{n}}{M}.$$

The sampling process is *self-weighting* in the sense that all the units in  $U$  have an equal probability of being selected, irrespective of their cluster. The *self-weighting* property avoids the negative impact of the variability of the sampling weights on the sampling variances.

We gather the information on the status of infection (variable  $y_{ij}$ ) on each of the  $m \times \bar{n}$  people selected in the sample, we carry out a COVID test on the non-verified infected people and collect this information from informative health care systems for the verified infected people.

The Horvitz Thompson estimator of  $Y$  is

$$(7) \hat{Y} = \sum_{i=1}^m \sum_{j=1}^{\bar{n}} y_{ij} \frac{1}{\pi_{ij}} = \sum_{i=1}^m \hat{Y}_i \frac{1}{\pi_i}, \text{ where } \hat{Y}_i = \sum_{j=1}^{\bar{n}} y_{ij} \frac{1}{\pi_{1i}}.$$

The spatial correlation of the units may be formalized by the following working model (WM)  $M$ , according to which

$$(8) y_{ij} = \hat{y}_{ij} + u_{ij},$$

where  $\hat{y}_{ij} = m(\mathbf{x}_{ij}; \boldsymbol{\beta})$  is a known function applied to the column vector of auxiliary variable  $\mathbf{x}_{ij}$ ,  $u_{ij}$  is a random residual, and  $\boldsymbol{\beta}$  is the unknown column vector of the model parameters. Let  $E_M(\cdot)$ ,  $V_M(\cdot)$ , and  $Cov_M(y_{ij}, y_{\ell k})$  denote the model expectation, variance and covariance, respectively. The spatial correlation of the units may be formalized as

$$(9) E_M(y_{ij}) = m(\mathbf{x}_{ij}; \boldsymbol{\beta}), V_M(y_{ij}) = \sigma_u^2, Cov_M(y_{ij}, y_{\ell k}) = \sigma_u^2 \rho_{ij, \ell k},$$

where  $\rho_{ij, \ell k}$  is the model correlation between units  $ij$  and  $\ell k$ , and  $\sigma_u^2$  is a variance scalar factor  $\rho_{ij, \ell k}$  is defined as  $\rho_{ij, \ell k} = \rho[d(ij, \ell k)]$  where  $|\rho(\cdot)| \leq 1$  is a decreasing function of the distance,  $d(ij, \ell k)$ , between units  $ij$  and  $\ell k$ . Alleva *et al.* (2021b), using model (8), with model expectations (9), report the full derivation of the anticipated variance of  $\hat{Y}$  (see Appendix 1, in Alleva *et al.* 2021b). A monte carlo validation with simulated data of the procedure presented is in Alleva *et al.* (2021b).

## 4. Conclusions

Understanding the mechanism of space propagation is fundamental for the evaluation of the spread of the epidemic underway. This is a complex objective that must consider a plurality of aspects. To improve the current practice in epidemic data collection we suggest to introduce a sampling design that exploits the intrinsic peculiarity of data being positively spatially correlated. The advantages of introducing the spatial dimension appear to be highly significant. The results obtained encourage us to extend the research in several directions. Some developments represent a natural extension of the present proposal. In fact, the focus is on the selection of sample units in the first stage. Future developments could also concern the introduction of capture-recapture methods for the second stage. Other possible future developments may concern the adaptation of the proposed method for the selection of sample units on which to administer diagnostic tests to trace the diffusion of the virus. One good example is the tracing of the variants of COVID-19 observed in 2020, with a specific focus on their diffusion in a territory.

## References

- Alleva G., Arbia, G., Falorsi P. D., Nardelli, V. and Zuliani, A. (2021a), A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemic: an operational design, *Journal of Official Statistics*, (forthcoming).
- Alleva G., Arbia, G., Falorsi P. D., Nardelli, V. and Zuliani, A. (2021b) Spatial sampling design to improve the efficiency of the estimation of the critical parameters of the SARS-CoV-2 epidemic, Arxiv.
- Arbia, G. (1993) The use of GIS in spatial surveys", *International Statistical Review*, 61, 2, 1993, 339-359.
- Arbia, G. Lafratta G., Simeoni C. (2007) Spatial Sampling Plans to Monitor the 3-D Spatial Distribution of Extremes in Soil Pollution Surveys, *Computational Statistics and Data Analysis*, 51 ,8, 4069- 4082.
- Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) *Hierarchical modeling and analysis for spatial data*, Chapman & Hall/CRC
- Brewer K.R.W., Hanif M. (1983) *An Introduction to Sampling with Unequal Probabilities*. In: Sampling With Unequal Probabilities. Lecture Notes in Statistics, vol 15. Springer, New York, NY.
- Chauvet G, Tillé Y. (2006) A fast algorithm of balanced sampling. *Journal of Computational Statistics* 21: 9–31
- Cliff, A. D., Haggett, P., Ord, J. K. and Verfey, F. R. (1981) *Spatial Diffusion: an Historical Geography of Epidemics in an Island Community*. Cambridge University Press.
- Deville, J.-C. and Tillé, Y. (1998) Unequal probability sampling without replacement through a splitting method, *Biometrika*, 85, 1, 89-101, DOI: 10.1093/biomet/85.1.89
- Fattorini, L., Marcheselli, M. Pisani, C. and Pratelli, L. (2020) Design-based consistency of the Horvitz–Thompson estimator under spatial sampling with applications to environmental surveys, *Spatial Statistics*, 35.
- Grafström A (2012) Spatially correlated Poisson sampling, *Journal of Statistical Planning and Inference* 142, 1, 139–147, DOI: 10.1016/j.jspi.2011.07.003
- Grafström A (2014) <http://www.antongrafstrom.se/balancedsampling/>
- Grafström A, Lundström NL, Schelin L. (2012) Spatially balanced sampling through the pivotal method. *Biometrics*, 68, 2, 514-20. doi: 10.1111/j.1541-0420.2011.01699.x.
- Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals". *Environmetrics*, 24, 120–131.
- Hedayat, A.S., Rao, C.R. and Stufken, J. (1988) Sampling Plans Excluding Contiguous Units. *Journal of Statistical Planning and Inference*, 19, 159–170.
- Horvitz, D.G. and D.L. Thompson (1952) A generalisation of sampling without replacement from finite-universe. *J. Amer. Statist. Assoc.*, 47,663-685.
- Kermack, W.O. and McKendrick, A. G. (1927) A contributions to the mathematical theory of epidemics" *Proceedings of the Royal society London, series A*, 115, 700-721.
- Kiesl, H. (2016). Indirect sampling: a review of theory and recent applications. *ASTA Wirtschafts-und Sozialstatistisches Archiv*, 10(4), 289-303
- Lavallée, P. (2007) *Indirect Sampling*, springer series in statistics.
- Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17-23.
- Müller W. G. (2007) *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, Springer-Verlag
- Stevens D. L. Jr. and Olsen, A. R. (2004) Spatially Balanced Sampling of Natural Resources, *Journal of the American Statistical Association*, 99, 465, 262-278, DOI: 10.1198/016214504000000250
- Tillé, Y. (2020). *Sampling and estimation from finite populations*. John Wiley & Sons.
- Tillé, Y. Dickson, M. M., Espa, G and Giuliani, D. (2018) Measuring the spatial balance of a sample: A new measure based on Moran's I index, *Spatial Statistics*, 23, 182-192.
- Wang, J.-F., Stein, A. ,Gao, B-B and Ge, Y. (2012) A review of spatial sampling, *Spatial Statistics*, 2, 1-14,
- Xu, Z., & Harriss, R. (2010). A Spatial and Temporal Autocorrelated Growth Model for City Rank-Size Distribution. *Urban Studies*, 47(2), 321-335.

# Survey aimed to estimate the seroprevalence of SARS-CoV-2 infection in Italian population at national and regional level

## *Indagine finalizzata a stimare il tasso di sieroprevalenza da SARS-CoV-2 nella popolazione italiana a livello nazionale e regionale*

Stefano Falorsi, Michele D'Alò, Claudia De Vitiis, Andrea Fasulo, Danila Filipponi, Alessio Guandalini, Francesca Inglese, Orietta Luzi, Enrico Orsini, Roberta Radini

**Abstract** This paper illustrates the main statistical methodological choices adopted in defining the sample strategy of the survey. The work starts from the description of the main objectives of the survey to highlight the close link that exists between the objectives themselves and the methodological choices that derive from them.

**Abstract** *In questo lavoro si illustrano le principali scelte statistiche metodologiche adottate nella definizione della strategia campionaria della rilevazione. Il lavoro parte dalla descrizione degli obiettivi conoscitivi dell'indagine per mettere in evidenza lo stretto legame che esiste tra gli obiettivi stessi e le scelte metodologiche che da questi discendono.*

**Key words:** Complex sampling design, Tracing.

## 1 Introduction

In Spring 2020 several serological surveys of SARS-CoV-2 have been done and others were ongoing. However, many of them were small or based on non-random sampling of participants (e.g., focusing on health-care workers or blood donors) and thus could not provide precise estimates of seroprevalence by age groups in the general population. Additionally, some of these studies have used antibody tests with low sensitivity or specificity or have not reported the characteristics of the test

---

<sup>1</sup> Stefano Falorsi, Istat; email: [stfalors@istat.it](mailto:stfalors@istat.it)

seroprevalence by age group in the general population. For the above reasons, in April 2020, the Italian Ministry of Health and the Italian National Statistical Institute, Istat, in collaboration with the Italian Red Cross that carried out the field operations with the help of the Regions, launched a nationwide, population-based, seroepidemiological survey, aimed to estimate the extent of SARS-CoV-2 diffusion in the country. In particular, the survey aimed to evaluate together with the serum prevalence rate for SARS-CoV-2 in the population, the fraction of asymptomatic and subclinical infections. It was planned a nationwide sample of 150.000 individuals randomly selected from Istat's Population Register. In order to deal with expected high non-response rates - in a context in which sample substitution mechanisms were not recommendable - an oversampling rate of 25% was applied leading the final sample size to 195.000 individuals. The survey was aimed to produce a detailed snapshot of the phenomenon of interest in spring 2020 being representative of Italian population by region, age group, sex and of working people by economic activity groups as well as other risk's factors. The survey was conceived as part of a more articulated study able to update the cross-sectional estimates giving account of the strong evolutionary dynamics of the investigated phenomenon. For this reason, an *anticipatory sub-sample* of 20.000 individuals from the overall sample was designed and randomly selected to take precedence during field survey operations. This could be eventually observed longitudinally over time to the extent that the founded budget of the study permits. It is worth adding that, the anticipatory sample could have been an excellent ground for testing in the field different techniques of *tracing* – aimed to detect all the people who had contact with each infected individual in the previous 14 days – to be applied in subsequent editions of the longitudinal survey. Indeed, technical literature on the epidemiological studies on SARS-COV2 epidemics shows how sampling selection procedures based on tracing rules may result effective in improving the efficiency of the final estimates. See, for example, the recent works by [1] in which a continuous monitoring system over time is proposed, based on indirect sampling techniques and the use of *tracing rules*, to estimate the prevalence of the number of people infected with SARS-COV-2.

Here, we describe the - design of the first wave of the study, conducted between May 25 and July 15, 2020. To the sample individuals, in addition to being subjected to a blood sample to carry out the serological test, were administered a short questionnaire aimed at detecting the presence of symptoms and risks factors.

The main parameters of interest of the cross-sectional survey concern the rate of individuals according to their epidemiological status with reference to different sub-populations related to territorial and/or structural characteristics of the investigated population referred as domains of interest. As far as domains are concerned, a subdivision into *primary* and *secondary* domains was considered for the study. For the former - relating to the main administrative and structural subdivisions of the population – we should ensure pre-established levels of accuracy of the estimates, compatibly with the overall budget constraints. For the latter - relating to important territorial subdivisions of the population of statistical or administrative nature - on the other hand, we would ensure that the selected sample has a good coverage as to guarantee, at least for part of them, acceptable levels of precision. For large-scale

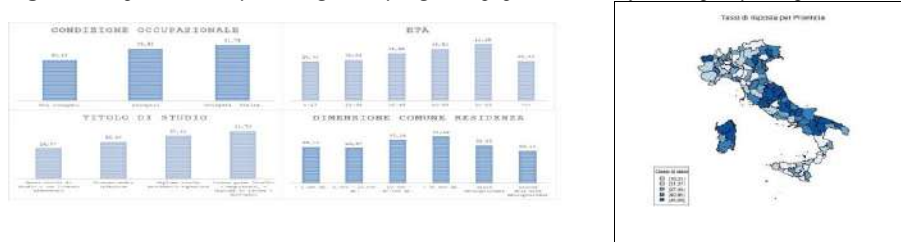
Survey aimed to estimate the seroprevalence of SARS-CoV-2 infection in Italian population at national and regional level

surveys, percentage Coefficients of Variation, CV%, below the threshold of 15-10% are considered as levels of medium-high reliability, while CV% around at the 33% threshold are considered as low but still acceptable levels of reliability. Estimates whose corresponding CVs exceed the 33% threshold are classified as having unacceptable levels of reliability. In particular, for the survey the primary territorial domains of interest are the Italian Geographical Regions and Autonomous Provinces of Bolzano and Trento. While, the structural primary domains, within each geographical region, for the general population consist of: age groups, 0-17; 18-34; 35-49; 50-59; 60-69; 70 and more; sex by large age groups. Furthermore, for working people sub-populations within each region, four economic activity macro-classes are considered. On the other hand, the secondary territorial domains which are related to the distribution of the phenomenon of interest, are made up of the Italian Provinces, the Local Labor Market Areas (LLMA) and the Local Health Authorities (LHA).

Briefly, the sample individuals were selected from Istat's Population Register, PR (excluding care-home residents and other collective residences), through a stratified two-stage sampling design, with municipalities as Primary Stage Units, PSU, and individuals as Secondary Stage Units, SSUs. The final sample size was: 1915 PSUs, out of 7904 Italian municipalities, and 150.000 SSUs with an oversampling of 25% correspondent to a final sample of 195.000 individuals. The strata were formed by the 110 Italian provinces with municipalities ordered by population size and grouped in Self-Representing, *SR*, and Non Self Representing, *NSR*, strata; the latter, formed by approximately equal population sizes. In *NSR* strata, the municipalities were selected with probability proportional to their size and 50 individuals were randomly sampled within each selected PSU. In this respect, we noted that there was a need, in the sample selection phase of SSUs, to ensure to get prefixed final sample sizes of individuals with respect to regional domains by sex, age and economic activity of working people. For this, the 50 SSUs within the PSUs were randomly selected by means of a simple balancing technique. The base information needed to stratify the individuals was available from Istat's System of Statistical Registers. In particular, PR provided information on gender, age and municipality of usual residence while Istat's Labour Statistical Register, LSR, provided the information on working activity and related economic activity class of each individual living in Italy the 1st January 2020. At the end of the selection phase, it was necessary to enclose in the list of sampled individuals their addresses and telephone contacts for contacting them by telephone, carry out a brief interview and finally invite them to take the blood sample for the test in a collection centre identified by the Italian Red Cross. The link of each sampled individual to his mobile telephone contacts was provided by mobile phone operator companies. For this operation, taking into account privacy preserving needs a special law act was required. In comparison with other Istat's large scale sampling surveys on population, the survey was characterized by a very low response rate. The percentage of respondents to the survey at national level is approximately 38% of the initial sample. Furthermore, not all respondents to the survey underwent the

serological test, the target variable of the epidemiological study, but only about 34% of the entire sample. This result was determined by the combination of several causes: the unavailability of individuals due to lack of telephone contact, the refusal to collaborate in the investigation expressed by the individuals contacted and the refusal of a part of the individuals interviewed to carry out the serological test. Furthermore, differential non-response rates were observed among the different investigated sub-populations of interest.

**Figure 1.** Response rates by working activity, age and population size of municipality and provinces



For this reason, the procedures for the construction of sampling weights were particularly accurate in order to try to mitigate the potential bias in the final estimates. Then, a first weighting step was carried out to correct the final estimates for non-response, by means of competing models for total non-response. Once the model with the best fit was identified, different strategies of weighting - based on quantiles classes of non-response weights - were applied to the data. In this phase, the statistical properties of different weighting strategies, in terms of variability of the weights and relative impact on final estimates, were compared. Because, all of the weighting strategy resulted robust in producing estimates of the same level, the procedure with lowest variability of the weights was chosen. A final calibration step was applied to non-response weights in order to take account of known population totals by demographic and educational level. Once the final estimates have been produced, in order to be confident on results to disseminate, in presence of high non response rates, a validation phase with thematic experts has been necessary. To this aim, other than the overall rate, the different distribution of final estimates has been compared with external data available from different geographical regions and for specific municipalities included in the sample. The validation step has not evidenced particular systematic differences with known external data. From the provisional data published by Istat in August 2020 [6] it was found that the people who resulted positive for the seroprevalence test in the period 25 May-15 July 2020 - that is, who have developed antibodies for SARS-CoV-2 - are equal to 2.5 % of resident population in family (excluding people living in collective households). Those that have come into contact with the virus are therefore 6 times more than the total of cases officially intercepted during the pandemic, through the identification of the RNA viral, as produced by the National Institute of Health. In the following pages are given more details for the phase of sampling design, par. 2, and non response weighting procedures in par. 3.

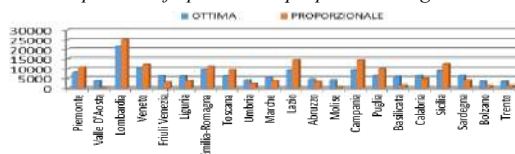


Survey aimed to estimate the seroprevalence of SARS-CoV-2 infection in Italian population at national and regional level

## 2 Allocation and sampling strategy

We limit ourselves to giving some hints on the most important aspects of the sampling strategy regarding: (1) the definition of the general sampling scheme; (2) the allocation of the sample; (3) non-response and calibration weighting. The methodological note of the survey (Istat, 2021) gives further details on the aforementioned aspects. As regards the point (1), an optimal scheme of complex sampling plan was adopted; this is applied for the most important large-scale population surveys at national and international level. Within each, territorial strata, it is based on the macro-stratification of municipalities into SR area and NSR area. This criterion allows to properly mediating the positive and negative properties of one-stage sampling design and two-stage sampling one. The sample was extracted trying to optimize territorial coverage even for unplanned domains. We defined the provinces as a minimum territorial domain of stratification and a relatively high number of municipalities was selected: 1910 out of 7904. In this way, the selected sample of municipalities showed excellent coverage both at the level of LHAs and at the LLMAs. With reference to point (2), the allocation of the sample of individuals and municipalities among the various regions was determined by adopting the optimal multivariate and multidomain allocation methodology for two-stage stratified designs implemented in the R R2BEAT package [5], generalizing the method proposed by [2] referring to the case of a stratified one-stage design and a single territorial domain of study. A critical aspect of the design concerned the planning of a survey aimed at observing a phenomenon for which only incomplete information was available, relating to particular subsets of the entire population. For these reasons, the ex-ante information was appropriately treated in order to avoid that the final sample sizes allocated in the different primary domains were strongly "disproportionate" with respect to the resident population of each of them. To reduce this effect and partially obtain the opposite effect in order to be able to provide estimates with finer territorial and structural detail in regions most affected by COVID-19, we decided to define different levels of precision depending on four groups of regions: (1) Regions with the highest prevalence (CV%<12.1); other Regions of North and Center of Italy (CV%<11.1); Regions of South and Islands (CV%<10.1); Basilicata (CV%<15.1). Furthermore, in order to preserve the proportionality of the sample as much as possible, greater importance was given to the precision constraint relating to national estimates (CV%<2.3). In Figure 1, the optimal allocation adopted with the proportional one is compared at the regional level.

**Figure 2:** Comparison of optimal and proportional regional allocation



Once the number of sample individuals for each Region was defined, its distribution at sub-regional level – for territorial strata (Provinces) and domains (Regions) as well as structural domains within each region - was purely based on *proportionality* to the population size of each sub-population. As regards the point (3), the procedure that led to the adjustment of the sample weights due to the total lack of response to the survey was developed in several steps: (a) acquisition of variables available for the definition of the response model; (b) study of the response models and choice of the working model; (c) estimation of individual response probabilities for the construction of corrective factors. The auxiliary variables available at individual level from Statistical Registers and those available at municipality level were used as covariates in the study of models for estimating the individual probability of response. The relationship between the response dependent variable and a set of independent auxiliary variables was studied by regressions of logistic type. The model fit was evaluated on the basis of some main indicators and significance tests. The predictors of the probability of response used in the model consist of: geographic regions; municipal types (metropolitan city; crown of the metropolitan area; less than 2000 inhabitants; between 2000 and 10,000 inhabitants; between 10,000 and 50,000 inhabitants; over 50,000 inhabitants); gender; age groups (0-17; 18-34; 35-49; 50-59; 60-69; 70+); classifications of the Activity class status (suspended employed, non-suspended PA + Education employees, non-suspended health workers, other non-suspended employees, non-employed); qualification modalities (Illiterate, Alphabets without a qualification, Primary school license, Lower secondary school license, Upper secondary school diploma, Bachelor's or first level academic diploma, Master's / specialist degree or Academic diploma II level, PhD); municipal positivity rate to SARS-COV2, estimated on the basis of the accumulated infections since the beginning of the pandemic in May (forecasts provided by the NHI); percentage difference in municipal mortality rates compared to the same period of the previous year; number of contact attempts for interview; anticipatory sample (1 unit belonging to the panel, 0 otherwise).

## References

1. Alleva G., Arbia G., Falorsi P.D. Nardelli V., Zuliani A, 2020. A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design, *Journal of Official Statistics*, forthcoming.
2. Bethel, J., 1989. Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57.
3. Bethlehem, J., Cobben, F. e Schouten, B., 2011, *Handbook of Nonresponse in household surveys*. New York: Wiley
4. Cochran, W.G., 1977. *Sampling techniques*, Wiley, New York.
5. Falorsi, S. Package R2Beat, R Cran, <https://cran.rproject.org/web/packages/R2BEAT/R2BEAT.pdf>
6. Hansen, M.H., Hurwitz, W.N. e Madow, W.G., 1953. *Sample survey methods and theory*, vol. I e II, Wiley, New York.
7. Istat 2020, Primi risultati dell'indagine di sieroprevalenza sul SARS-CoV-2, 2021, <https://www.istat.it/it/files//2020/08/ReportPrimiRisultatiIndagineSiero.pdf>
8. Mahalanobis, P.C. Some aspects of the design of sample surveys, 1952. *Sankhya*, 12, 1-7.

### 6.2.3 Measuring and modeling inequalities following the Covid-19 crisis

# COVID-19 impacts on young people's life courses: first results

## *Gli effetti del COVID-19 sui corsi di vita dei giovani: i primi risultati*

Antonietta Bisceglia, Concetta Scolorato, Giancarlo Ragozini

**Abstract** The pandemic, with its restrictions and economic consequences, has deeply impacted people's lives across the globe. In the case of young Italians, although they now become adults even without completing all the stages of the transition to adulthood, the partial suspension of the normal flow of life due to COVID-19 has led to a sort of forced moratorium on this path. From an initial analysis of the official data, the effects of the pandemic include the increase in the inactivity rate and the number of NEETs; a partial change of residence; the postponement or cancellation of marriages; and a further reduction in births. In this research, conducted through an online survey, we investigate these aspects and their physiological effects on the identity process.

**Abstract** *La pandemia, con le sue restrizioni e conseguenze economiche, ha inciso sui corsi di vita delle persone. Nel caso dei giovani italiani, sebbene essi diventino ormai adulti anche senza compiere tutte le tappe di transizione all'adulthood, la parziale sospensione del normale flusso di vita dovuta al COVID-19 ha indotto una sorta di moratoria forzata di tale percorso. Da una prima analisi dei dati ufficiali, gli effetti della pandemia sono: l'aumento del tasso di inattività e del numero dei NEET; un parziale cambio delle dimore; il rinvio o l'annullamento dei matrimoni; una riduzione ulteriore delle nascite. Nella presente ricerca, condotta attraverso un'indagine online, indaghiamo questi aspetti e i loro effetti fisiologici sul processo di identità.*

**Key words:** Transition to adulthood, Forced Moratorium, Official Statistics

---

<sup>1</sup> Antonietta Bisceglia, University of Naples Federico II; antonietta.bisceglia@unina.it  
Concetta Scolorato, University of Naples Federico II; concetta.scolorato@unina.it  
Giancarlo Ragozini, University of Naples Federico II; giancarlo.ragozini@unina.it

## **1 Becoming adult in Italy**

The transition to adulthood has been described mostly through a series of life milestones, including the completion of studies, labour market entry, leaving the parental home, forming a first union, and finally, entry into parenthood [8, 5, 11, 16, 9, 12]. In the past, the transition process ended when an individual had experienced all of these events [6, 1]. Until the early 1980s in Europe, these life events were standardized and followed a linear trend, and each phase was preparatory to a later stage. In this period, although the modes and times of transition varied considerably according to gender and social class, youth was thus configured as a temporally defined and socially recognizable phase [6].

However, in recent decades, the transition to adult life across various European societies has been profoundly altered. This can be attributed to the economic, institutional, demographic, and cultural transformations such as increased global competition, deregulation of the job market, structural unemployment [4] and the longer periods spent in full-time education or training. Thus, the modern transitions have inevitably become more fluid, complex, risky, uncertain and prolonged.

There are two processes that define the transformations of the times and the transition to adult life, including the formation of one's own family: the postponement transition and the partnership revolution. The first process corresponds to an extension of transition times; the second, instead, corresponds to the revolution of partnerships and the change in the ways of forming a unions, less and less in the form of marriage [15, 3].

Considering the Italian case, these processes were intensified by the economic crisis starting in 2008. In the years of the Great Recession (2008-2013), many people felt excluded from the social and economic context and were convinced that they did not have the power to change their future for the better. The health emergency due to the COVID-19 pandemic has considerably worsened the transition process, inducing a kind of "forced moratorium". For instance, the economic crisis could lead some to move back to the parents' homes for a short or long period of time; the ban on ceremonies and wedding parties could lead some to postpone or cancel their marriages; and the heightened pressure on the health care system and the related uncertainties may have led some to change their childbearing decisions.

To investigate the effects of the pandemic on the transitions to adulthood and the trajectories of life of young Italian people, in the next paragraph, we present some evidence of these changes. To do so, we incorporate data from official statistics and socio-demographic research, organizing them by stages of life. In the last paragraph, we introduce our ongoing research.

## **2 First evidence of COVID-19 effects on young people course life**

The most up-to-date official Italian statistics on socio-demographic data show that the COVID-19 pandemic has amplified the trend of the country's population decline

COVID-19 impacts on young people's life courses: first results

that began in 2015. In particular, limitations and economic consequences due to the pandemic have caused impediments, slowdowns, suspensions or remodulations of the five events of life that mark the path to adulthood. According to the data coming from the official statistics, some of the life milestones to become adult were slightly affected while others were more heavily influenced by the pandemic.

In terms of the school and university system, thanks to the introduction of large-scale distance education practices, there are evidences that the graduation rate has remained high, indicating positive educational outcomes. Indeed, the percentage of young people (aged 15-29) with tertiary education rose from 13.1% in 2019 to 14% in 2020. Moreover, the percentage of young people in Despite the public fear that the pandemic will be detrimental to learning outcomes, there is no evidence this. In addition, appears that the pandemic has not affected the Italian rate of early leavers from education and training. For instance, in the second trimester of 2020, the rate of early leavers was 13.5%, which is equal to the 2019 rate, and far from the 2020 European goal of 10%.

On the contrary, looking at the rate of young people not in education, employment or training (NEET) in the age class of 18 to 29, there is a slight growth from 26% in 2019 to 27.7% in 2020, after six years of continuous reduction. This is largely due to an increase in the inactivity rate of young people in the Central and Northern regions [10], which rose from 42.6% to 45.3% in Northern Italy and from 49.8% to 52.2% in Central Italy. On the contrary, in the South of Italy, in which the NEET rate was already higher, it remained almost stable, varying from 38.6% to 38.8%.

Considering the labour market and the employment rates, younger generations appear to be the most disadvantaged segment of the population. Indeed, after the decline in employment began in the early 2000s and youth unemployment peaked in 2013, the generalized recovery observed in the period of 2013 to 2019 was abruptly slowed down by the pandemic. In the second quarter of 2020, the employment rate fell to 38.6% (3.2 points less than the value of the second quarter of 2019) and then rose again to 40% in the third quarter of 2020 (2.2 points less than in the same period of the previous year). In addition, in the initial phases of the lockdown, the share of discouraged young people searching for jobs grew, increasing the percentage of the inactive as already mentioned [10].

Further, to the best of our knowledge, there is no available updated data that could help in understanding to what extent young people delay the decision of leaving their family's home or to what extent they have been forced by the economic crisis to their family's home. Housing choices, especially for younger workers, have also been influenced by the so-called "South working", i.e., when an individual from Southern Italy who works for a northern company and is able to telework from home in the South [18].

Moreover, in reference to family and union formation, in the first year of the pandemic, there were almost 50% fewer marriages observed than in 2019. The report on "Marriages, civil unions, separations and divorces", which was published by ISTAT in February 2021, and shows a sharp decline in marriages and civil unions, in addition to a slight decrease in divorces [13]. The analysis also included the first half of 2020, which coincided with the onset of the pandemic, specifying that the data are still provisional; in that semester, marriages, civil unions, as well as

separations and divorces literally collapsed. These trends are certainly also linked to the consolidating progressive spread of free unions (or cohabitation), both as a step before marriage and as an alternative to marriage; however, the effect of the pandemic is certainly also visible on these relationships [13], as weddings and parties were banned for many months.

Finally, looking at official statistics, the impact of the pandemic on childbearing decisions is clearly evident; the number of empty cradles in 2020 sharply increased, worsening the already declining Italian birth rate [2]. In 2020, only 404,104 births were registered, a new negative record that shows a decrease of 4% (16,000 fewer births) compared to 2019 [14]. Although the downward trend in births also preceded the onset of the pandemic, in December 2020, the first month in which children conceived during the pandemic were born, the decline proved to be drastically significant (about 21.6%).

### **3 Some elements of the ongoing survey**

As evident from the existing data, 2020 witnessed a slowdown or a halt in at least three of the five steps in the process of becoming an adult, resulting in a forced moratorium of the courses of young people's lives.

After one year of the pandemic and heightened restrictions, we are conducting a survey with the specific aim of uncovering the effects of COVID-19 on the five steps to adulthood and of reconstructing the possible life paths. The dimensions taken into consideration are the following: sociodemographic characteristics, educational status and training, working status, residential status, marriage and parenting, changes of perspective and future projects, support networks and emotions and moods. Along with this proper socio-demographic approach, we also consider the identity status dimensions [7] in order to verify if the forced standby due to pandemic restrictions is also related to a future-oriented or present-focused identity moratorium [17].

### **Acknowledgements**

The present work has been funded by the Observatory of Youth Policies of Campania Region, DD N. 244 del 04/06/2020, POR-ESF 2014-2020 CUP E64I19002390005, and by the research grant "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide" MIUR-PRIN 2017HBTk5P-CUP B78D19000180001.

### **References**

1. Barbagli, M., Castiglioni, M., Dalla Zuanna, G.: *Fare famiglia in Italia. Un secolo di cambiamenti*, Il Mulino, Bologna (2003)
2. Billari, F. C., Liefbroer, A. C., & Philipov, D.: The postponement of childbearing in Europe: Driving forces and implications. *Vienna Yearbook of Population Research*, 1-17 (2006)

COVID-19 impacts on young people's life courses: first results

3. Billari, F. C., & Liefbroer, A. C.: Towards a new pattern of transition to adulthood?. *Advances in life course research*, 15(2-3), 59-75 (2010)
4. Brückner, H., Mayer, K. U.: De-standardization of the life course: What it might mean? And if it means anything, whether it actually took place?. *Advances in life course research*, 9, 27—53 (2005)
5. Buchmann, M.: *The Script of Life in Modern Society. Entry into Adulthood in a Changing World*. Univ. Chicago Press., Chicago (1989)
6. Cavalli, A., Galland, O.: *Senza fretta di crescere. L'ingresso difficile nella vita adulta*, Liguori, Napoli (1996)
7. Crocetti, E., Sica, L. S., Schwartz, S. J., Serafini, T., Meeus, W.: Identity styles, dimensions, statuses, and functions: Making connections among identity conceptualizations. *European Review of Applied Psychology*, 63(1), 1-13 (2013)
8. Elder, G.H.: Perspectives on the life course. In Moen, P., & Elder, G. H.(eds.) *Life Course Dynamics. Trajectories and Transitions, 1968–1980*, 23–49. Cornell Univ. Press, Ithaca, NY/London (1985)
9. Elder, G.H., Shanahan, M.J.: The life course and human development. In Lerner, R.M.: *Handbook of Child Psychology*, Vol. 1, 665–715. New York John Wiley and Sons (2006)
10. Fraboni R., Rosina A., Marzilli E.: I giovani e la transizione allo stato adulto. I Billari F.C., Tommasini, C. (eds): *L'Italia e le sfide della demografia*, il Mulino, Bologna (2021)
11. George, L.K.: Sociological perspectives on life transitions. *Annual Review of Sociology*. 19(1), 353–73 (1993)
12. Gauthier, A.H.: Becoming a young adult: an international perspective on the transitions to adulthood. *European Journal of Population*. 23:217–23 (2007)
13. ISTAT: *Rapporto su Matrimoni, unioni civili, separazioni e divorzi*, Roma (2021)
14. ISTAT: *La dinamica demografica durante la pandemia Covid-19*, Roma (2021)
15. Macura, M., MacDonald, A. L., Haug, W.: *The new demographic regime*. United Nations, Geneva (2005).
16. Shanahan, M.J.: Pathways to adulthood in changing societies: variability and mechanisms in life course perspective. *Annual Review of Sociology*. 26(1), 667–92 (2000)
17. Sica, L. S., Crocetti, E., Ragozini, G., Aleni Sestito, L., Serafini, T.: Future-oriented or present-focused? The role of social support and identity styles on 'futuring' in Italian late adolescents and emerging adults. *Journal of Youth Studies*, 19(2), 183-203 (2016)
18. SVIMEZ: *Rapporto Svimez 2020 sull'economia del Mezzogiorno*, Il Mulino, Bologna (2020)



# Exploring Students' Profile and Performance Before and After Covid-19 Lock-down

## *Un'Analisi Comparativa del Profilo e del Rendimento degli Studenti Prima e Dopo il Lock-down*

Cristina Davino and Marco Gherghi

**Abstract** Universities around the world have responded to the emergency arising from the Covid-19 pandemic by moving teaching activities online. Nowadays it is important to study on the effects of this sudden change on students' life. This paper proposes some reflections on the effects that the closure of universities has had on the performance and characteristics of university students. The proposed empirical analysis is based on data from the University of Naples Federico II in Italy.

**Abstract** *Le Università di tutto il mondo hanno risposto all'emergenza derivante dalla pandemia da Covid-19 con il trasferimento online delle attività didattiche. In questo periodo in cui la pandemia ancora perdura ma anche in prospettiva di una totale ripresa, è importante riflettere sugli effetti di questo cambio repentino che ha investito gli studenti di ogni ordine e grado. Questo lavoro propone alcune riflessioni sugli effetti che la chiusura delle strutture universitarie ha avuto sulla performance e sulle caratteristiche degli studenti universitari. L'analisi empirica proposta è basata sui dati relativi all'Università di Napoli Federico II.*

**Key words:** e-learning, students' performance, multiple factorial analysis

---

Cristina Davino

Department of Economics and Statistics, University of Naples Federico II e-mail: cristina.davino@unina.it

Marco Gherghi

Department of Economics and Statistics, University of Naples Federico II e-mail: marco.gherghi@unina.it

## 1 Introduction

The Covid-19 pandemic has forced all countries around the world to suddenly adopt exceptional measures to ensure the continuity of educational processes at all levels. It is estimated that in recent months, governments in more than 190 countries closed their educational facilities fully or partially to limit the rapid spread of Covid-19 [5]. This situation has massively disrupted teaching and learning. In particular, schools and universities have been, and in some cases are still closed, for long periods and all activities (lectures, tutorials, exams) have been transferred to online mode.

The scientific community, policy-makers and civil society as a whole are now questioning the consequences that this sudden change in educational methods is having on student learning and thus on their current and future performance. The well-founded fear is that there will be an inevitable increase in educational poverty, especially to the detriment of the most socially and economically vulnerable students [2]. Further investigations are aimed at understanding the extent to which it is possible to take advantage of the current crisis to modernise education systems by exploiting, at least in part, the benefits provided by digital technologies [9]. In other words, the challenge consists in understanding whether online learning just serves as a panacea in the time of crisis or, if properly used, it can be a tool for more inclusive education [3].

This paper focuses on the effects of the pandemic in higher education where campus and university closures have been almost total, estimated at around 99% [7]. In this context, in addition to the problems linked to learning difficulties and access to technology, there is also the question of examination methods [8]. The shift to online courses has required additional solutions to measure and certify students' acquisition of knowledge and skills in an online setting. Aim of this paper is to provide a comparative view of the performance and characteristics of students in the pre-pandemic years and during 2020, year of the first educational breakdown.

In particular, results are presented for an Italian mega-university, the University of Naples Federico II, located in the south of Italy. The analysis is proposed on two levels of evaluation: the effect of lock-down and of distance learning. In the first case, the outcomes of the exam sessions over the last three years (2018 to 2020) are compared, taking 10 March 2020, as the threshold for breaking with the past, being the day on which Italy entered lockdown. The second level of analysis focuses only on the courses that were taught in an online mode, in the second semester of a.y. 2019/2020. Again, a comparative analysis with the exam sessions of previous years allows us to investigate on possible effects related to the online transfer of lectures and exams.

## **2 Data Description and Lock-down Effect on Students' Performance**

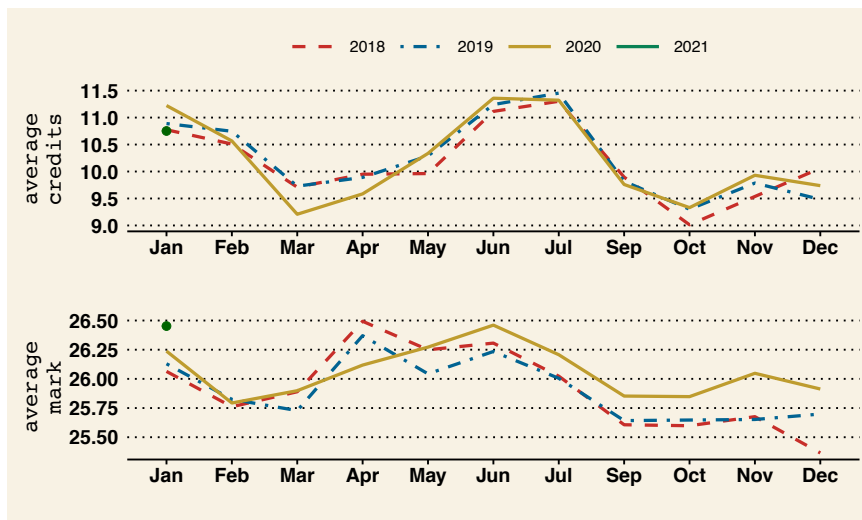
One of the concerns related to the recent and sudden development and adoption of digital and online technologies in higher education is the evaluation of pros and cons of this new learning and teaching modalities. In particular, the challenge would be to exploit the change brought about by the crisis to open up promising prospects to improve students' performance and assessment and the inclusion of students in more disadvantaged circumstances. The first step towards an evaluation of the effects of the disruption of the traditional patterns of teaching, learning and assessment requires a comparison with the years before the current crisis.

This study is proposed for exploratory purposes considering the students of the University of Naples Federico II which is located in the south of Italy and is considered one of the largest universities in Italy in terms of number of enrolled students. The reference population is all students who took exams in the years from 2018 to 2019 (for 2020, only the January exam session was available). Considering that the population of enrolled students has remained constant over the years in terms of number of enrolled students and socio-demographic characteristics, it is possible to compare examination sessions over the considered period. In total, 8709 lectures were considered for a total of 768.164 exams and 634.098 students.

Figure 1 shows the trend in the average of European Credit Transfer System Credits (upper panel) and the average mark achieved (lower panel) for each examination session in the period under consideration. Each line refers to a different year. It is worth to notice that the April session is not comparable because in 2020 the session was open to all students whereas in the past it was reserved only to not regular students. Considering the month of March 2020 as the change point with respect to the past, Figure 1 shows an improvement in the performance of the students in terms of credits acquired but also with respect to the assessment received. This result can be considered comforting because it shows a good adjustment and adaptation of the students to the post-lock-down situation, which does not seem to have affected the continuation of their studies. The study of the change in the evaluation criteria between face-to-face and online exams deserves a separate discussion with ad hoc designs for data allocation.

## **3 Students' Profile and Performance: a Comparison 'Pre' and 'Post' Lock-down**

As mentioned above, since 10 March 2020, almost all universities in Italy have transferred their teaching activities entirely online. For this reason, May, June and July sessions are worth a closer look as they include students who took distance learning courses. These examination sessions are particularly interesting as they took place immediately at the end of the first term where online courses were of-



**Fig. 1** Distribution over months, from January 2018 to January 2021, of the average number of European Credit Transfer System Credits (upper panel) and average mark (lower panel).

ferred. In this section, we present an in-depth look at the performance of exams and the characteristics of involved students.

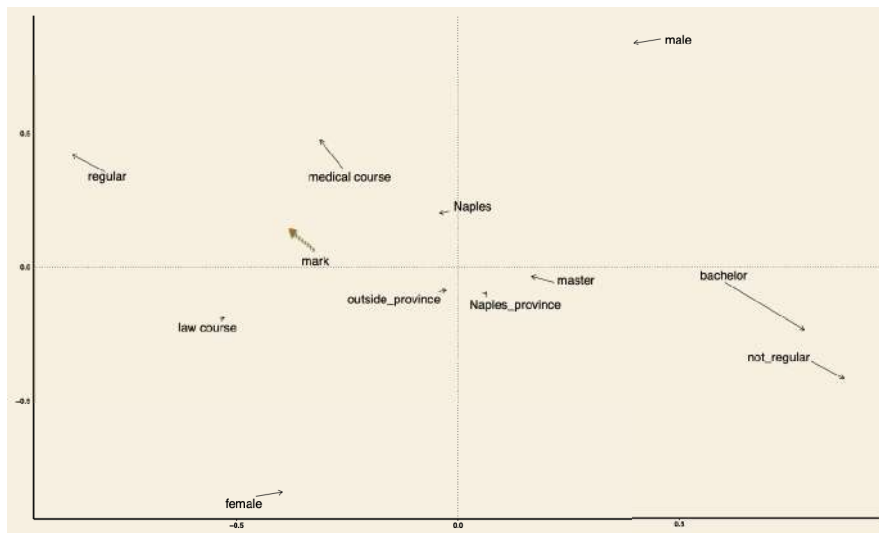
A broader view of the population of students who took exams associated with online courses must also include the socio-demographic profile of the students. This analysis was carried out considering students who took exams related to online courses but restricting the focus to the May, June and July 2019 and 2020 sessions, considering 10 March 2020 as the threshold date. From now on, each year will be referred as 'pre' and 'post' lock-down respectively. In total, 1125 lectures were considered for a total of 65474 exams and 20664 students.

The data table has the typical row partitioned structure, i.e. the same variables are observed on two groups of students ('pre' and 'post' lock-down). The considered variable are: gender (male, female), residence (Naples, Naples province, outside province), type of course (Bachelor, Master, 6-year medical course, 5-year law course), year of course (regular and not regular), average mark.

The aim is to visually compare the students' profiles in the two periods, simultaneously analysing the whole set of variables. In a geometrical framework, data related to each group of individuals could be synthesized using a factorial method. In particular, we exploit the Dual Multiple Factor Analysis (DMFA) [6], an extension of Multiple Factor Analysis [4], since it allows the comparison of these different factorial analyses conducted on the same variables in the two different groups of individuals. In DMFA, variables are centered by group and a simultaneous representation is provided to study the evolution of each variable through the groups of individuals. As the considered variables are all categorical, we refer to an extension of DMFA based on the transformation of categorical variables into properly

weighted indicator variables as it is usually done in multiple correspondence analysis [1]. The only quantitative variable, average mark, was projected on the map considering it as a supplementary variable.

Figure 2 enables a visual comparison of the association structures between categories in the two groups and allows to identify their similarities and differences. For each category, each arrow connects the points related to the two groups, 'pre' and 'post', the direction of the arrows indicating the direction from 'pre' to 'post'. While the second axis opposes mainly males and females, the first axis provides a more nuanced picture of students' characteristics: the left-hand side of the first axis defines the profile of students in track with their exams, also with higher average marks, mainly female and enrolled in Medicine and Law courses, an opposite profile to male students, enrolled in a Bachelor's degree. The behavior of 2019 and 2020 students is similar, as expected, however some minor differences appear clearly looking at the direction of the longer arrows which denotes a greater difference in the two periods. In particular, the performance of students in track with their exams, enrolled in a medical course or a Master, seems to be improving (arrows pointing to the left). On the other hand, the results of students, mainly females, enrolled in a bachelor's programme, who are still enrolled after the legal duration of the course, seem to be worsening (arrows pointing to the right).



**Fig. 2** Representation of the categories associated with the students' profiles on the factor plane (1,2) on the two groups 'pre' and 'post' lockdown. The direction of each arrow denotes the changes after 10 March 2020, the day in which the first lockdown started in Italy.

The proposed analyses show light and shade in the past year at the University of Naples. If on the one hand, the results on the whole population (Section 2) can be considered comforting because they show a good adjustment and adaptation of the

students to the post-lock-down situation, which does not seem to have affected the continuation of their studies, on the other hand DMFA reveals that students show a different spirit of reaction depending on the students' characteristics.

The results of this study can be enhanced by considering a wider study which would collect data related to several universities to give a deeper understanding of the effects of this epidemic. Other studies could focus on how social and family conditions could have affected students' performance. It is a matter of fact that the social and economic conditions of families have a major influence on the e-learning experience because less advantaged students are less likely to have access to relevant learning digital resources

## References

1. Abascal, E., Díaz de Rada, V., García Lautre, I., Landaluce, M.I.: Extending dual multiple factor analysis to categorical tables. *Journal of Applied Statistics*, **40**(2), 415–428 (2013)
2. Di Pietro, G., Biagi, F., Costa, P., Karpiski Z., Mazza, J.: The likely impact of COVID-19 on education: Reflections based on the existing literature and international datasets. EUR 30275 EN, Publications Office of the European Union, Luxembourg (2020)
3. Dhawan, S.: Online Learning: A Panacea in the Time of COVID-19 Crisis. *Journal of Educational Technology Systems*. **49**(1), 5–22 (2020)
4. Escofier, B., Pagès, J.: *Analyses Factorielles Simples et Multiples. Objectifs Méthodes et Interprétation*. Paris: Dunod (1988)
5. Giannini, S., Jenkins, S., Saavedra, J.: Reopening schools: When, where and how? UNESCO. <https://en.unesco.org/news/reopening-schools-when-where-and-how> (2020)
6. Le S., Pagès, J.: DMFA: Dual Multiple Factor Analysis. *Communications in Statistics—Theory and Methods*. **39**(3), 483–492 (2010)
7. Malee Bassett, R., Arnhold, N.: COVID-19's immense impact on equity in tertiary education. In: World Bank Blogs. <https://blogs.worldbank.org/education/covid-19s-immense-impact-equity-tertiary-education> (2020)
8. OECD: Remote online exams in higher education during the COVID-19 crisis. *OECD Education Policy Perspectives*, No. 6, OECD Publishing, Paris. <https://doi.org/10.1787/f53e2177-en> (2020)
9. Pokhrel, S., Chhetri, R.: A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning. *Higher Education for the Future*. **8**(1), 133–141 (2021)

#### 6.2.4 Nowcasting the Covid-19 outbreaks methods and applications

# Modeling subsequent waves of COVID-19 outbreak: A change point growth model

*Un modello basato su curve di crescita e punti di cambio per descrivere le ondate del contagio da COVID-19*

Luca Greco, Paolo Girardi, Laura Ventura

**Abstract** The COVID-19 pandemic has brought a remarkable amount of data to support policy makers engaged in contrasting its spread. The availability of data about infections, hospitalizations, deaths allowed statisticians to face the challenging problem of monitoring, modeling and nowcasting the evolution of the outbreak, despite the quality of the data was often unsatisfactory. Here, we propose a method to model cumulative counts of incidence data based on the five parameters log-logistic growth function. This function adapts well to describe the typical behavior of one wave of the contagion. Then, a flexible strategy to fit subsequent waves is proposed, according to a change point model in a likelihood framework.

**Abstract** *La disponibilità di dati su infezioni, ricoveri, decessi da COVID-19 ha coinvolto gli statistici nelle analisi del monitoraggio e della modellizzazione dell'evoluzione dell'epidemia. In questo contributo: 1) proponiamo un metodo per modellare i conteggi cumulati dei dati di incidenza utilizzando la funzione di crescita log-logistica a cinque parametri, che si adatta bene per descrivere il comportamento tipico di un'ondata del contagio; 2) discutiamo un modello basato su punti di cambio nell'ambito della teoria della verosimiglianza per descrivere ondate successive.*

**Key words:** Change point, COVID-19, incidence, independence loglikelihood, log-logistic

---

Luca Greco  
DEMM Department, University of Sannio, e-mail: luca.greco@unisannio.it

Paolo Girardi  
Department of Developmental and Social Psychology, Department of Statistical Sciences, University of Padova e-mail: paolo.girardi@unipd.it

Laura Ventura  
Department of Statistical Sciences, University of Padova e-mail: ventura@stat.unipd.it



## 1 Introduction

Summer 2020 has brought a renewed spread of COVID-19 outbreak all around the world. The larger mobility combined with a relaxed easing of the previously imposed restrictive measures to allow a return to *normality* and to support the economy, made numbers of new cases of infection, hospitalizations and deaths grow again, also in those countries where the epidemic stood on very low numbers for several consecutive weeks, such as in Italy. In Italy, the growth of the contagion has been massive during autumn. The number of daily cases peaked in November 13, whereas the daily hospitalizations and deaths reached the maximum value ten and twenty days later, respectively. The Italian Government introduced targeted lockdowns and other severe actions, as school closure and a prolonged stop to catering and unessential goods selling activities, in addition to the strict restrictions aimed to regulate social behaviors, transportations, sport events, and all those circumstances characterized by a high risk of gathering. The availability of counts from COVID-19 outbreak has soon represented a crucial modeling challenge for statisticians all over the world, in order to provide meaningful descriptions and predictions. Here, we propose a change point growth model to fit cumulative incidence data, such as infections and deaths, that is able to catch subsequent waves of the pandemic. The model is meant to describe the main features of the observed trends and, in particular, to give evidence about the time when different waves were more likely to originate. The latter estimate could aid the investigation of the main causes leading to different waves. This contribution is structured as follows: modeling background is given in Section 2; the change point growth model is introduced in Section 3; an application to Italian data is presented in Section 4, with a short discussion in Section 5.

## 2 Modeling one wave of cumulative counts

The evolution of the pandemic suggests that, in each wave of the outbreak, cumulative incidence data, such as new cases and deaths, exhibits an exponential growth first, whereas the growth becomes logistic from some point on and moves towards an upper bound. This behavior can be modeled by the well-known five parameters log-logistic growth function [5], given by

$$\mu(t; \theta) = c + \frac{d - c}{\left[1 + \left(\frac{t}{e}\right)^b\right]^f}, \quad \theta = (b, c, d, e, f), \quad b < 0, \quad c, d, e, f > 0 \quad (1)$$

and expressed as a function of time. The presence of five parameters in (1) allows a great modeling flexibility. The parameters  $b, e$  and  $f$  determine the shape of the growth function,  $c$  returns the lower asymptote, such that  $\lim_{t \rightarrow 0} \mu(t; \theta) = c$ ,  $d$  represents the upper asymptote, that is a direct measure of the final size of the pandemic,

with  $\lim_{t \rightarrow \infty} \mu(t; \theta) = d$ . When  $f = 1$ , the model is such that  $\mu(e; \theta) = d - \mu(e; \theta) = (d - c)/2$  and the log-logistic model is said to be symmetric. The first derivative  $\mu'(t; \theta) = \partial \mu(t; \theta) / \partial t$  allows to describe the behavior of daily incidence data.

Let  $y^c = (y_1^c, y_2^c, \dots, y_T^c)$  denote the series of cumulative counts data and assume that its expected value follows the log-logistic growth curve (1). Since the nature of the data is such that  $y_{t+1}^c \geq y_t^c, \forall t$ , the assumption of independence is questionable. In this respect, in the following we pursue an approach based on a *pseudo*-loglikelihood function defined as the sum of  $T$  contributions defined as

$$\ell_I(\theta) = \sum_{t=1}^T \log p(y_t^c; \theta), \tag{2}$$

with marginal models assumed to be of Poisson type. A related approach has been also adopted in [3]. The reader is pointed to [2] for a likelihood based nowcasting strategy well suited for count incidence data.

The function in (2) represents a *composite* loglikelihood function for  $\theta$ , based on only marginals and sometimes referred to as the *independence* loglikelihood [6]. The validity of inference about  $\theta$  using the independence loglikelihood can be justified invoking the general theory of unbiased M-estimating functions. Actually, the composite loglikelihood in (2) shares the properties of a loglikelihood from a misspecified model. In particular, the maximum composite likelihood estimate (MCLE)  $\hat{\theta}^I = \operatorname{argmax}_{\theta} \ell_I(\theta)$  can be also defined as the root of the composite score equation  $u_I(\theta) = \partial \ell_I(\theta) / \partial \theta^\top = 0$ . The corresponding estimator is asymptotically normally distributed with mean  $\theta$  and covariance matrix  $V(\theta) = G(\theta)^{-1} = H(\theta)^{-1} J(\theta) H(\theta)^{-1}$ , where  $G(\theta)$  is the Godambe information matrix with  $H(\theta) = E(-\partial u_I(\theta) / \partial \theta^\top)$  and  $J(\theta) = \operatorname{Var}(u_I(\theta)) = E(u_I(\theta) u_I(\theta)^\top)$ . Composite likelihood versions of Wald, score and suitably adjusted likelihood ratio statistics can be obtained that all share the standard asymptotic chi-squared distribution [4].

The classical sandwich estimator of  $V(\theta)$  may not be able to catch two main aspects in the data and lead to under-estimate uncertainty. First, the assumption of Poisson marginals may lead to neglect possible overdispersion in the data. Furthermore, the non-stationarity of the series of cumulative counts data suggests some adjustments to take into account both heteroskedasticity and autocorrelation. Therefore, we propose a couple of possible adjustments. One correction term comes from the general theory of regression models for counts data [1]. By paralleling the approach based on quasi-likelihood inference, one could take into account overdispersion by inflating the sandwich variance-covariance matrix by a dispersion parameter estimate obtained as  $\hat{\phi} = \frac{1}{n} \sum_{t=1}^T \frac{[y_t^c - \mu(t; \hat{\theta}^I)]^2}{\mu(t; \hat{\theta}^I)}$ . A second proposal consists in correcting the variance-covariance matrix for heteroskedasticity and autocorrelation and evaluate a HAC sandwich estimate [7].

### 3 Change point growth model

Let us consider the situation with two waves. Each wave can be modeled according to (1), that is

$$\mu(t; \tau) = \begin{cases} \mu(t; \theta_1), & t \leq t_0 \\ \mu(t - t_0; \theta_2), & t > t_0 \end{cases} \quad (3)$$

with  $\tau = (\xi, t_0)$ ,  $\xi = (\theta_1, \theta_2)$ . The function (3) is characterized by one change point at unknown time  $t_0$  where the mean switches from  $\mu(t; \theta_1)$  to  $\mu(t; \theta_2)$ . Moreover, in the second branch the lower asymptote is fixed as  $c = \mu(t_0; \theta_1)$  so that  $\mu(t; \theta_2) \geq \mu(t; \theta_1)$ ,  $\forall t$  and equality holds at  $t = t_0$ . Therefore, a four parameters log-logistic model is fitted in the second wave. The independence loglikelihood function is

$$\ell_I(\tau) = \sum_{t=1}^T [z_t \log p(y_t; \mu_1(t; \theta_1)) + (1 - z_t) \log p(y_t; \mu_2(t - t_0; \theta_2))],$$

where  $z_t = 1$  for  $t \leq t_0$  and zero otherwise. The change point can be estimated by a composite profile approach as

$$\hat{t}_0^I = \operatorname{argmax}_{t_0} \ell_{IP}(t_0), \quad (4)$$

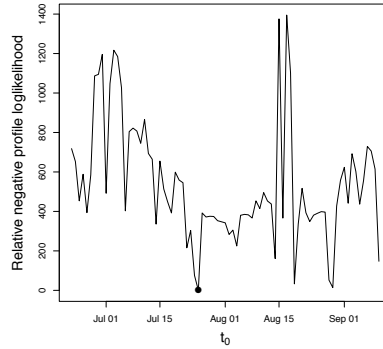
where  $\ell_{IP}(t_0) = \ell_I(\hat{\xi}_{t_0}^I, t_0)$  and  $\hat{\xi}_{t_0}^I$  is the constrained MCLE of the branches parameters for fixed change point. Then, the unconstrained MCLE of  $\xi$  is obtained as  $\hat{\xi}_{t_0}^I$ . Standard errors are properly evaluated conditionally on  $\hat{t}_0^I$ . Wald-type asymptotic confidence intervals around the mean function  $\mu(t; \theta)$  and its first derivative  $\mu'(t; \theta)$  can be obtained based on the delta method. Confidence intervals around  $\hat{t}_0^I$  can be obtained according to the inverse function below at  $\mu = \mu(\hat{t}_0^I; \hat{\theta}_1^I)$ ,

$$\mu^{-1}(\mu; \theta) = e \left[ \left( \frac{d-c}{\mu-c} \right)^{1/f} - 1 \right]^{1/b}.$$

### 4 Real data example: Italian death counts

Italian COVID-19 epidemic data are available since February 24, 2020 from a GitHub repository daily updated by the *Dipartimento della Protezione Civile*<sup>1</sup>. In this section, we consider the cumulative death counts collected from February 24, 2020, until February 16, 2021. As previously stated, in this period we have observed two waves, characterizing the trajectories of infections, hospitalizations and deaths. The change point growth model presented in Section 3 for cumulative death counts leads to locate the structural break on July 25, 2020. The criterion in (4) is displayed in Figure 2. The entries in Table 1 give parameters estimates with 99%

<sup>1</sup> <https://github.com/pcm-dpc/COVID-19>

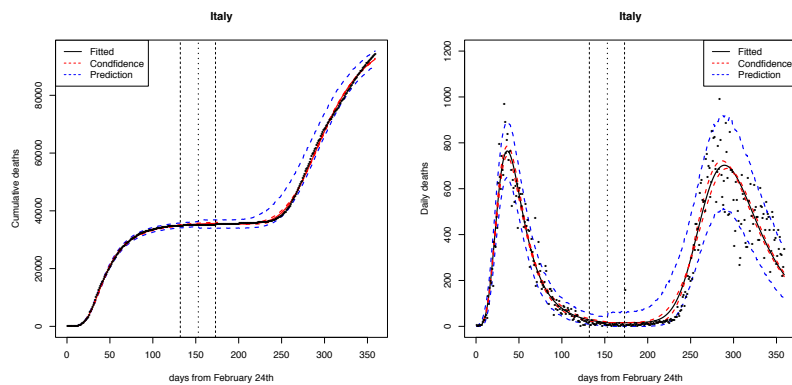


**Fig. 1** Selection of  $t_0$ . The plain black circle gives the fitted change point.

confidence intervals. The fitted curve (3) is given in the left panel of Figure 2, together with point-wise 0.99-level Wald-type asymptotic confidence intervals based on the HAC sandwich covariance matrix estimate and prediction intervals derived through a parametric double bootstrap procedure: values for  $\theta$  are simulated from its asymptotic normal distribution using the HAC sandwich estimate, then, data are generated from a Poisson-Gamma mixture with linear variance function. Prediction intervals are obtained by computing point-wise quantiles. The right panel displays the daily counts along with the first derivative of each branch of (3) and corresponding confidence and prediction intervals. In order to avoid unpleasant optimization convergence issues, we set  $c = y_1^c$  in  $\mu(t; \theta_1)$ . We also notice that the choice  $c = 0$  returned a lower likelihood.

**Table 1** Parameter estimates with 99% confidence intervals based on the overdispersed-inflated sandwich and the HAC sandwich covariance matrix estimate. The entries in the last row are dates in the form month/day

	MCLE	overdispersed	HAC
$\theta_1$ b	-2.99	-3.08 – -2.91	-3.28 – -2.71
d	36224.69	36026.15 – 36423.24	35662.82 – 36786.56
e	37.67	36.53 – 38.81	33.96 – 41.39
f	1.54	1.44 – 1.64	1.22 – 1.86
$\theta_2$ b	-9.65	-10.14 – -9.16	-10.38 – -8.92
d	101056.29	99478.29 – 102634.29	100324.58 – 101788.00
e	262.32	256.43 – 268.20	241.13 – 283.50
f	2.95	2.47 – 3.42	1.20 – 4.69
$t_0$	07/25	07/16 – 08/02	07/04 – 08/14



**Fig. 2** Left: cumulative deaths. Right: daily deaths. Fitted model with 0.99-level confidence and prediction intervals based on the HAC sandwich covariance matrix estimate. The dotted vertical lines give the fitted change point with the corresponding 0.99-level confidence interval.

## 5 Discussion

We presented a method to model cumulative counts in an epidemic characterized by two distinct waves. We combined two growth curve employing a change point model. To relax assumptions about independence and take into account overdispersion, we proposed a couple of corrections in the evaluation of standard errors. We considered an application to Italian death counts. The fitted model locates the change point during the period from the beginning of July to the first half of August and estimates a total amount of deaths in the second wave about three times the amount in the first wave, with about seven thousand COVID-19 deaths that are still to happen, assuming that a third wave does not occur.

## References

1. Agresti, A.: Foundations of linear and generalized linear models. John Wiley & Sons (2015)
2. Di Loro, P. A., Divino, F., Farcomeni, A., Jona-Lasinio, G., Lovison, G., Maruotti, A., Mingione, M.: Nowcasting COVID-19 incidence indicators during the Italian first outbreak, arXiv preprint arXiv:2010.12679 (2020)
3. Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E., Ventura, L.: Robust inference for non-linear regression models from the Tsallis score: Application to COVID-2019 contagion in Italy, *Stat* **9**, e309 (2020)
4. Pace, L., Salvan, A., Sartori, N.: Adjusting composite likelihood ratio statistics. *Stat. Sin.* **21**, 129–148 (2011)
5. Ritz C., Baty F., Streibig J.C., Gerhard D.: Dose-Response Analysis Using R. *PLoS ONE* **12**, 5, 1–22 (2015)
6. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. *Stat. Sin.* **21**, 5–42 (2011)
7. Zeileis A.: Econometric computing with HC and HAC covariance matrix estimators. *J. Stat. Soft.* **11**, 1-17 (2004)

# The second wave of SARS-CoV-2 epidemic in Italy through a SIRD model

## *Analisi della seconda ondata epidemica di SARS-CoV-2 tramite un semplice modello compartimentale*

Michela Baccini and Giulia Cereda

**Abstract** Using a SIRD model calibrated on COVID19-related deaths, we describe the dynamics of the SARS-CoV-2 epidemic in Italy at regional level from August 2020 to the end of February 2021. We estimate the time-varying reproductive number,  $R_0(t)$ , modelled as a natural cubic spline with six internal equi-spaced knots, and quantify the number of infections, included their submerged portion, under different infection fatality rate (IFR) scenarios. Comparing the observed number of infections with its prediction under different scenarios, some hints about a possible upper bound for IFR are drawn. As an example, we report the main results for three regions: Lazio, Toscana, and Valle d'Aosta.

**Abstract** Utilizzando un modello di tipo SIRD calibrato sui decessi COVID19, forniamo una descrizione della dinamica epidemica da SARS-CoV-2 a livello regionale in Italia da agosto 2020 alla fine di febbraio 2021. Oltre a stimare il numero di riproduzione dell'infezione come una funzione variabile nel tempo attraverso una spline cubica naturale, l'approccio proposto produce una stima del numero di infezioni, inclusa la loro porzione sommersa, sotto diversi scenari di letalità, intesa come tasso di letalità dell'infezione (IFR). Confrontando il numero di casi notificati con la loro stima da modello ottenuta sotto i diversi scenari di letalità, è possibile ottenere indizi circa un possibile limite superiore per l'IFR. A titolo esemplificativo, abbiamo qui riportato i principali risultati per tre regioni: Lazio, Toscana e Valle d'Aosta.

**Key words:** SIRD model, SARS-CoV-2, COVID19, reproductive number, infection fatality rate, second epidemic wave

## Introduction

With the aim of describing the SARS-CoV-2 epidemic dynamics during the second wave of autumn 2020/winter 2021 in the Italian Regions, in [4] we used a com-

partmental model of SIRD type [1, 9], calibrated on the COVID19-related deaths as routinely reported and made available by Protezione Civile [11]. In that SIRD model, that we estimated separately for each region, the infection reproductive number  $R_0(t)$  was allowed to flexibly change over time in order to capture variations in infection transmission possibly due to restrictive policies, school reopening, and other environmental factors. While  $R_0(t)$  was considered unknown and estimated from data, to assure model identification the infection fatality rate (IFR), that is the probability of dying for the (notified and not notified) infected, was considered as a known fixed parameter and results were obtained under alternative IFR values. The IFR value is obviously crucial to determine the total number of infections from which deaths derive, thus to quantify the submerged fraction of the epidemic. Taking advantage of this relationship, we get insights about the plausibility of different IFR values, by comparing the total number of infections predicted by the SIRD model with the observed notified cases and checking their mutual consistency [11].

In this work, we updated the analysis in [4] extending the study period to most recent data (August 1st 2020 - February 21st 2021) and using a different estimation algorithm which allows to constraint  $R_0(t)$  to be non-negative.

## 1 Methods

The SIRD is described by the following system of equations:

$$\begin{cases} S(t) &= S(t-1) - \beta(t) \frac{S(t-1)}{S(0)} I(t-1) \\ I(t) &= I(t-1) + \beta(t) \frac{S(t-1)}{S(0)} I(t-1) - \alpha I(t-1) - \delta I(t-1) \\ R(t) &= R(t-1) + \alpha I(t-1) \\ D(t) &= D(t-1) + \delta I(t-1) \end{cases} \quad \forall t \in \{1, \dots, T\}, \quad (1)$$

where  $S(t)$ ,  $I(t)$ ,  $R(t)$  and  $D(t)$  are the sizes of the Susceptible, Infected, Recovered and Deceased compartments at time  $t$ . We set  $S(0) = N - I(0)$ , with  $N$  regional population size,  $D(0) = 0$  and  $R(0) = 0$ , thus starting to count deaths and recoveries from August 1st. We perform analyses on each region, fixing  $I(0)$  to the number of notified circulating infections on July 31st [11]. The parameters  $\alpha$  and  $\delta$  in Eq. (1) are the transition rates from the compartment of the infected to the compartment of recovered and deceased, respectively. They depend on the IFR value  $p$ , the average times from infection to death,  $T_D$ , and from infection to recovery,  $T_R$ , that we assumed to be equal:  $T_D = T_R = T$  [4]. Under this assumption, the following relationships apply [8]:  $\alpha = \frac{1-p}{T}$ ,  $\delta = \frac{p}{T}$ . The time-varying infection rate  $\beta(t)$  is related to the reproductive number  $R_0(t)$ , as follows:

$$\beta(t) = R_0(t)(\alpha + \delta) = \frac{R_0(t)}{T}. \quad (2)$$

In order to get a flexible estimate of the time-varying basic reproductive number, we modelled it through a natural cubic regression spline, with 6 internal equi-spaced knots:  $R_0(t) = s(t; \vartheta)$ , where  $\vartheta$  is a vector of unknown coefficients [12].

We assured parameter identifiability by fixing  $T = 14$  [4, 13] and, in three separate analyses,  $p = 0.78\%$ ,  $1.14\%$  and  $1.79\%$  [14, 3].

The vector  $\vartheta$  was estimated by minimizing the following sum of squares :

$$Q(\vartheta) = \sum_{t=1}^K (D(t; \vartheta) - D^{\text{obs}}(t))^2, \quad (3)$$

subject to a positive bound constraint on  $R_0(t)$ :  $\min_t(s(t; \vartheta)) \geq 0$ . In Eq. 3,  $t = 1$  corresponds to August 1st,  $t = K$  to February 21st and  $D^{\text{obs}}(t)$  denotes the cumulative number of deaths observed starting from August 1st. The minimization was done via the algorithm implemented in the Auglag function of the nloptr package of R software [12]. The algorithm has been initialized using 100 different initial values sampled from a multivariate grid defined on  $\vartheta$ . Among the 100 estimates thus obtained, we selected as the best one the estimate  $\hat{\vartheta}$  associated to the lowest value of  $Q(\cdot)$ .

A parametric bootstrap procedure has been implemented in order to quantify uncertainty around the estimates [5, 6]. We assumed a Negative Binomial distribution on the daily increments of the estimated time series  $D(t; \hat{\vartheta})$  and generated 200 bootstrap samples to be used as observed time series in as many calibrations. The 90% confidence intervals or bands for the quantities of interest have been calculated as the 5th and 95th percentiles of the bootstrap distributions.

## 2 Results

As an example, in Figure 1 we report for three Italian regions, Lazio, Toscana and Valle d'Aosta, the estimated  $R_0(t)$  curves and the estimated number of circulating infections over time, with their 90% pointwise confidence bands, for  $p = 0.78\%$ . In Valle d'Aosta the first COVID19-related death from July 31st was observed in September, hence we show the curve starting from 14 days before the first death, being the estimate very unstable before that date. The pattern of  $R_0(t)$  appears to be heterogeneous among the three regions, even though they all show a peak around the end of October. The large confidence bands at the end of the study period are due to the fact that deaths at time  $t$  mainly refer to infections around  $t - 14$ , making information for the most recent weeks quite poor.

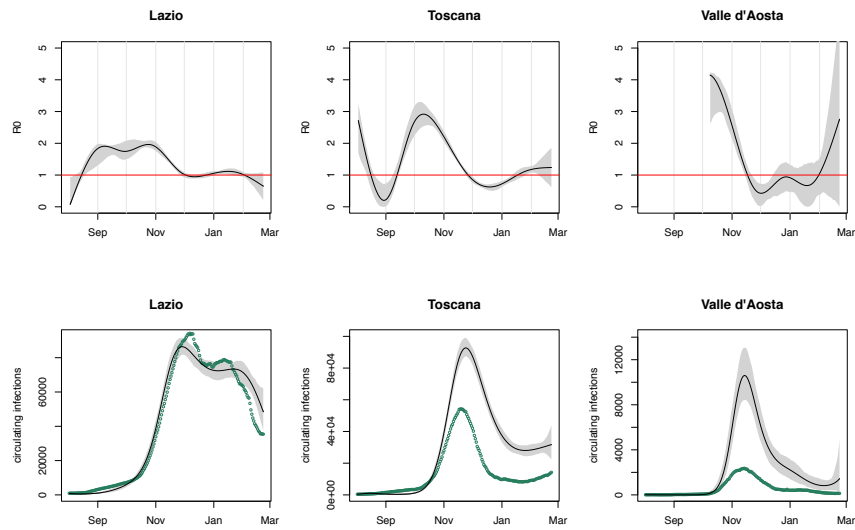
Regarding circulating infections, the peak of the observed infections in the region reported by Protezione Civile [11] (green dots) approximately corresponds to the peak of  $I(t)$  estimated by the model, despite the fact that it has not been calibrated on the notified cases. The three regions show the largest peak of prevalence reached in the first half of November, followed by a decline.



Interestingly, looking at the Lazio region in Figure 1, one can notice that the number of notified cases is often larger than the number of infected individuals predicted by the SIRD model. This result is clearly not consistent with the fact that the number of infected individuals predicted by the SIRD model should include both notified and not notified (unobserved) cases, thus exceed the observed. This is indicative that  $p = 0.78\%$  could be too high for this region during this specific study period.

In Figure 2 circulating infections predicted by the model under all three IFR scenarios are reported as well as the observed number of notified cases. As expected, for Lazio  $p = 1.79\%$  and  $p = 1.14\%$  are even less plausible than  $p = 0.78\%$ . On the contrary, for Tuscany the observed number of circulating infections seems to be compatible with  $p = 0.78\%$  and  $p = 1.14\%$ , but not with the higher  $p = 1.79\%$ . For Valle d'Aosta, observed data are coherent even with the largest IFR used in the analyses.

The estimate of  $R_0(t)$  appeared to be robust to changes of  $p$  (results not reported).

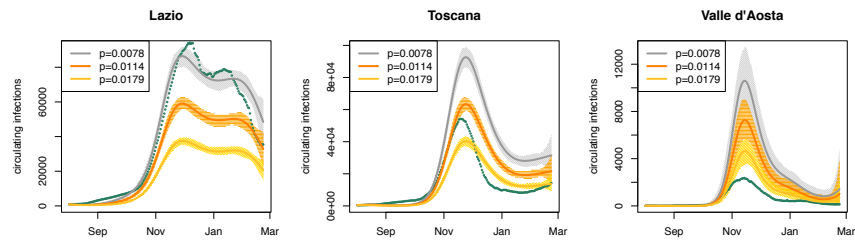


**Fig. 1** Estimates of  $R_0(t)$  (top) and number of circulating infections (bottom) in the three regions. Green dots indicate the observed number of cases in the region.

### 3 Conclusion

The approach proposed in this work is a simple and useful tool for monitoring and describing the epidemic dynamics over time (the results are weekly updated and

## The second wave of SARS-CoV-2 epidemic in Italy through a SIRD model



**Fig. 2** Estimated number of circulating infections in the three regions, under three different IFR scenarios. Green dots indicate the observed number of cases in the region.

make available on GitHub). It provides at the same time both a flexible estimate of  $R_0(t)$  and a prediction of the number of new and circulating infections in the area of interest, under different IFR scenarios. As we showed, the comparison between the observed and predicted number of infections may be used to define plausible IFR values and their possible upper bound.

In interpreting the results of our model, one should consider that it relies on strong assumptions. While some of them are well known, for example the assumption of close population, others are implicit and less discussed, as the one concerning the transition times between different compartments that, as in classical compartmental models, are assumed to be Exponentially distributed [15, 2]. As a future development of our work, we will define a SIRD model where the rate of transition from the infection status to death or recovery increases as time passes from the infection onset, by assuming an Erlang distribution on the infectivity time [10, 7].

## References

1. N. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. Mathematics in Medicine Series. Griffin, 1975.
2. F. Brauer, P. van den Driessche, and J. Wu. *Mathematical Epidemiology*, chapter Compartmental Models in Epidemiology, pages 19–79. Springer Berlin Heidelberg, 2008.
3. N. F. Brazeau, R. Verity, S. Jenk, H. Fu, C. Whittaker, P. Winskill, I. Dorigatti, P. Walker, S. Riley, R. P. Schnekenberg, H. Hoeltgebaum, T. A. Mellan, S. Mishra, J. T. Unwin, O. J. Watson, Z. M. Cucunubá, M. Baguelin, L. Whittles, S. Bhatt, A. C. Ghani, N. M. Ferguson, and L. C. Okell. Covid-19 infection fatality ratio: Estimates from seroprevalence. <https://doi.org/10.25561/83545>, 2020.
4. G. Cereda, C. Viscardi, L. Gherardini, F. Mealli, and M. Baccini. A SIRD model calibrated on deaths to investigate the second wave of the SARS-CoV-2 epidemic in Italy. <https://repo.epiprev.it/2052>, 2020.
5. G. Chowell. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*, 2:379–398, 2017.

6. B. T. Grenfell, O. N. Bjørnstad, and B. F. Finkenstädt. Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. *Ecological Monographs*, 72:185–202, 2002.
7. O. Krylova and D. J. D. Earn. Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of The Royal Society Interface*, 10:20130098, 2013.
8. J. Legrand, R. F. Grais, P.-Y. Boelle, A.-J. Valleron, and A. Flahault. Understanding the dynamics of ebola epidemics. *Epidemiology & Infection*, 135(4):610–621, 2007.
9. F. Lin, K. Muthuraman, and M. Lawley. An optimal control theory approach to non-pharmaceutical interventions. *BMC Infect Dis*, 10:1–13, 2010.
10. A. L. Lloyd. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theor Popul Biol*, 60(1):59–71, Aug 2001.
11. Protezione Civile. Repository of Covid-19 outbreak data for italy. <https://github.com/pcm-dpc/COVID-19>, 2013.
12. D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.
13. SARS-CoV-2 Surveillance Group. Characteristics of sars-cov-2 patients dying in italy. report based on available data on december 9th, 2020. [https://www.epicentro.iss.it/en/coronavirus/bollettino/Report-COVID-2019\\_9\\_december\\_2020.pdf](https://www.epicentro.iss.it/en/coronavirus/bollettino/Report-COVID-2019_9_december_2020.pdf), 2020.
14. M. Villa. Coronavirus: la letalità in italia, tra apparenza e realtà. <https://www.ispionline.it/it/pubblicazione/coronavirus-laletalita-italia-tra-apparenza-e-realta-25563>, 2020.
15. H. J. Wearing, P. Rohani, and M. J. Keeling. Appropriate models for the management of infectious diseases. *PLOS Medicine*, 2(7), 07 2005.

## 6.2.5 The impact of Covid-19 on survey methods

# Collecting cross-national survey data during the COVID-19 pandemic: Challenges and implications of data collection for the 50+ population in the Survey of Health, Ageing and Retirement in Europe (SHARE)

Michael Bergmann, Arne Bethmann, Yuri Pettinicchi, and Axel Börsch-Supan<sup>1</sup>

**Abstract** The COVID-19 pandemic started to hit Europe early 2020 and even one year later affects virtually all aspects of life – including survey research. The Survey of Health, Ageing and Retirement in Europe (SHARE) was in the middle of its Wave 8 data collection when fieldwork had to be suspended in all 28 participating countries. Responding to the evolving crisis and taking steps to resume fieldwork was especially challenging for SHARE, since the 50+ population is more severely affected by COVID-19 than younger age groups. Against this background, we describe the efforts made by SHARE to switch interview mode from face-to-face to telephone and carry out a survey on the effects of the pandemic. We will focus on the challenges for SHARE’s target population as well as the effects of the introduced changes on data collection. In addition, we summarize the implications for future methodological plans of SHARE in reaction to the pandemic.

**Key words:** SHARE, COVID-19, mode switch, CATI, mode effects

---

<sup>1</sup> Michael Bergmann, Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, Technical University of Munich (Chair for the Economics of Aging), [bergmann@mea.mpisoc.mpg.de](mailto:bergmann@mea.mpisoc.mpg.de)

Arne Bethmann, Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, Technical University of Munich (Chair for the Economics of Aging), [bethmann@mea.mpisoc.mpg.de](mailto:bethmann@mea.mpisoc.mpg.de)

Yuri Pettinicchi, Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, [pettinicchi@mea.mpisoc.mpg.de](mailto:pettinicchi@mea.mpisoc.mpg.de)

Axel Börsch-Supan, Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, [axel@boersch-supan.de](mailto:axel@boersch-supan.de)

## 1 Introduction

The COVID-19 pandemic started to hit Europe in early 2020 and even one year later affects virtually all aspects of life – including survey research. Similar to other surveys the Survey of Health, Ageing and Retirement in Europe (SHARE) had to suspend its regular face-to-face interviewing in all 28 participating countries in March 2020 (Scherpenzeel et al., 2020). The implementation of strict epidemiological control measures in nearly all participating countries made it infeasible to continue face-to-face fieldwork. Stopping the survey was particularly urgent considering another crucial aspect of SHARE: its target population of people aged 50+. This includes very old respondents as well as retirement and nursing homes residents who face the highest risks from a possible infection. Against this background, SHARE switched the interview mode to telephone interviewing, using a special “SHARE COVID-19” questionnaire. In the following, we will describe the efforts made by SHARE to carry out the survey by focusing on the challenges regarding the abrupt change of interview mode for SHARE’s target population as well as the effects of the introduced changes on data collection.

## 2 Suspension of face-to-face fieldwork and development of a new data collection strategy: The SHARE Corona Survey

At the beginning of February 2020, COVID-19 was spreading quickly across Europe, leading to a suspension of SHARE fieldwork in all participating countries between March 10 and March 23. All stakeholders involved shared the opinion that SHARE data about the health and living situation of the 50+ population in Europe were now needed more than ever to shed light on the short- and long-term implications of this global pandemic. This led to the development of the SHARE Corona Survey and the successful application for a research grant proposal with the European Commission<sup>1</sup> aimed at understanding the non-intended consequences of epidemic control decisions across Europe in terms of prevention, protection, and treatment of the population 50+ during the first phase of the pandemic. The survey was developed and successfully conducted in all 28 countries between April and August 2020. Together with the full wealth of background information on people aged 50 years and older from previous SHARE waves, this data collection offers huge potential for substantive analyses and cross-national comparisons regarding health, social, and economic developments and outcomes. The key concept of identifying cause and effect is to exploit (a) the cross-national differences in the

---

<sup>1</sup> Submitted proposal for the “Second call for an Expression of Interest for innovative and rapid health-related approaches to respond to COVID-19 and to deliver quick results for society for a higher level of preparedness of health systems” (H2020-SC1-PHE-CORONAVIRUS-2020-2C).

Collecting cross-national survey data during the COVID-19 pandemic extent of the COVID-19 pandemic, (b) the cross-national differences in the severity and stringency of the epidemic control actions, and (c) the longitudinal variation of observations before, during and after the pandemic. Therefore, in addition to the regularly planned ninth wave at the end of 2021, a second wave of the SHARE Corona Survey will be conducted in spring 2021, supplying information on longer-run health and socioeconomic outcomes caused by the pandemic. Based on these data, SHARE can add important insights across European countries and Israel allowing common responses to the short-, mid-, and long-term consequences of the pandemic by (health) policy makers and social organizations.

### **3 Switch from face-to-face to telephone interviewing during the COVID-19 pandemic**

The abrupt stop of fieldwork accelerated the process of adapting the mode of data collection in SHARE. After fieldwork was suspended, it soon became clear that a quick return to the normal face-to-face Computer Assisted Personal Interview (CAPI) was unlikely. A short-term change to web interviewing was rejected due to the variation in internet use across countries and especially across age groups in SHARE. Thus, particularly the 80+ year-olds still report a much lower rate of usage. This age group, however, is a very important target group in ageing studies like SHARE and should certainly be included in a study about COVID-19. Moreover, other studies have shown that mode effects on response behavior and measurement error tend to be larger between interviewer- and self-administered modes than between modes that are both interviewer-administered such as face-to-face and telephone (e.g. Couper, 2011). Therefore, it was decided to resume interviewing with a Computer Assisted Telephone Interview (CATI), targeted to the COVID-19 living situation of people who are 50 years and older.

When using CATI, a further benefit was that some of the existing SHARE software tools could be adapted more easily and faster (de Bruijne, Pennings, & van der Wielen, forthcoming). This holds in particular for the application to manage the respondents' contact and household information and eligibility (CaseCTRL). As this tool was already installed on the interviewers' laptops at the start of Wave 8, interviewers were able to continue using it by telephone. The most crucial change was that the newly developed SHARE Corona questionnaire had to be programmed anew, translated into the 40 SHARE languages, tested, and made accessible by the interviewers. A web-based survey tool (Quest) was used to program the SHARE Corona questionnaire. Quest easily handles online questionnaires with large numbers of respondents and multiple languages. It uses a central approach, which significantly increases the efficiency of the many steps of data collection and data delivery. Quest also makes the data and metadata available to administrators in real time and in various formats. The usage of a web-based platform (TranslationCTRL) enabled a controlled translation process that is crucial for an ex-ante harmonized study like SHARE. The continued use of the CaseCTRL prevented a disruption of

Michael Bergmann, Arne Bethmann, Yuri Pettinicchi, Axel Börsch-Supan  
the longitudinal household data and data structure. A simple patch allowed the interviewers to access the SHARE Corona questionnaire directly from the CaseCTRL. Furthermore, the separate end-of-life interview that could be conducted via telephone without any technical changes was already available via CaseCTRL.

The implementation of the SHARE Corona Survey as a web-based application brought about several changes to the usual workflows in SHARE, but at the same time introduced many benefits in terms of efficiency. Thus, the innovations in programming of the generic version, in compiling and testing of the national versions of the survey, and in data extraction procedures did shorten the time necessary to make the survey agencies operative for fieldwork. With Quest, SHARE took a first step towards a more flexible data collection infrastructure that can rapidly respond to a changing world.

#### **4 Adaptions in sample and fieldwork design due to COVID-19**

The SHARE data are unanimously based on full probability samples (Bergmann, Kneip, De Luca, & Scherpenzeel, 2019). For most countries, the SHARE Corona Survey was based on the whole national SHARE panel sample, including both panel members who had not been interviewed before the suspension of fieldwork as well as panel members who had already been interviewed face-to-face in Wave 8. Only in two countries (the Netherlands and Sweden) a sub-sample had to be selected due to funding issues. Overall, more than 58,000 panel interviews from 28 countries with respondents 50+ could be realized in summer 2020. This number also includes more than 600 end-of-life interviews with a proxy (usually, a close relative or friend of the respondent) of the deceased person as well as interviews in institutions. In contrast to many other cross-national studies, SHARE includes persons living in nursing or retirement homes. It was decided that these panel members should also be asked to participate in the SHARE Corona Survey. However, interviewers were instructed to avoid pressing refusal conversion attempts among nursing home respondents or on the caretaker or staff members of the nursing home if they are hesitant to allow the interview. Nonetheless, about 500 nursing home interviews could be realized by telephone. The preliminary average response rate based on eligible respondents in Wave 8 was 79 percent, ranging from 58 percent (Luxembourg) to 96 percent (Romania). Given the challenges in obtaining contact information of people aged 50 years and older as well as the extremely short timeframe, this was an excellent outcome.

In contrast to the longitudinal sample, the recruitment of the Wave 8 refreshment samples was not continued after the suspension of fieldwork, nor were any of the already-recruited refreshment sample members re-interviewed. The reason for this decision was twofold: First, different from panel households, telephone numbers were unavailable for most refreshment sample households, with exceptions in a few (Scandinavian) countries. Second, the exceptional value of collecting COVID-19 data in SHARE lies in the merging of these new data with what we already know about the life of panel respondents from previous waves. This



Collecting cross-national survey data during the COVID-19 pandemic provides us with crucial context information on respondent's/household's situations before the outbreak of the pandemic and enables us to thoroughly investigate how COVID-19 has changed the situation of 50+ respondents and what are the consequences thereof.

To prepare the change from CAPI to CATI fieldwork, several aspects of the normal SHARE fieldwork design, also including amendments to the existing contracts, had to be adapted. However, SHARE's principle of providing the same software tools and programmed questionnaire to all survey agencies in order to harmonize and standardize fieldwork and monitoring (see Börsch-Supan et al., 2013) was also followed for the SHARE Corona Survey. Survey agencies were asked to send a new advance letter to the Wave 8 panel members, even if they had already sent one before the suspension. The new advance letter announced the telephone interview and included the standard SHARE data protection statement as well as a reply card that respondents could use to update their telephone number if needed. A condensed read-out version of the data protection statement was prepared for countries where postal services did not work properly. Survey agencies were also asked to include a monetary incentive in the advance letter when possible. In countries where sending money by post was not allowed or possible, a link to a gift voucher or other gift was used.

Further, SHARE demanded that for the CATI fieldwork, survey agencies had to employ only interviewers who had received general interviewer training as well as the SHARE-specific interviewer training at the start of Wave 8. The reason for this was that the personal contact between interviewer and respondent is an important prerequisite for the success of panel studies like SHARE. In this respect, it can be expected that successive interview contacts with the same interviewer carrying out the survey each wave help to build trust and confidence (Kühne, 2018), in particular with respect to older respondents. Therefore, all survey agencies were instructed to maintain interviewer continuity as much as possible. In addition, all interviewers working in the CATI fieldwork received additional training with the SHARE Corona Survey via webinars. Key elements in these national training sessions (NTS) were the simulation of the actual SHARE interview, the correct use of the adapted software, as well as adapted techniques regarding contacting households, handling refusals, and conducting the actual interview. Having only SHARE-experienced interviewers made the delivery of these instructions much easier. National interviewer trainings were preceded by Train-the-Trainer (TTT) webinars for the survey agency staff, centrally conducted by the SHARE Central coordination team and conveying all technical, logistical and managerial aspects of successful fieldwork. Part of the TTT also allowed for an open discussion of all involved parties about how to handle the mode switch to telephone interviews with a focus on specific conditions regarding SHARE's target population, including the search for telephone numbers or the handling of incentives.

## **5 Implications of the introduced mode switch for data collection and analyses**

Michael Bergmann, Arne Bethmann, Yuri Pettinicchi, Axel Börsch-Supan

The resumption of SHARE Wave 8 by asking a special SHARE Corona questionnaire over the telephone was carried out in 27 European countries and Israel from June to August 2020 and yielded more than 58,000 individual interviews. Together with about the same number of face-to-face interviews in Wave 8 prior to the suspension of fieldwork, this allows to evaluate how the high-risk group of older respondents coped with the crisis, how the national healthcare and social systems responded to the pandemic, and which lessons for the future should be drawn by policy makers and social organizations. At the same time, the changes to mode and timing of data collection raises a number of important questions that have not been sufficiently answered yet: Did the suspension of Wave 8 fieldwork affect data quality with respect to the interrupted panel samples? What is the state of affairs with respect to the fielded refreshment samples? What are the effects of a mode switch during an ongoing panel survey and in particular for SHARE's specific target population of people aged 50 years and older? What about mode effects, i.e. who does and who does not respond in a certain interview mode (mode selection effects) and how is a question answered in a certain interview mode (mode measurement effects)?

To answer the questions above, we are currently analyzing potential effects of the introduced changes in data collection with a particular focus on the older population and give practical advice for researchers facing similar challenges in directing their efforts. In addition, we are investigating cross-national differences in data quality as some countries only just started fieldwork, while other countries were close to rounding up fieldwork when Wave 8 data collection was suspended.

## References

- Bergmann, M., Kneip, T., De Luca, G., & Scherpenzeel, A. (2019). *Survey participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), Wave 1-7*. SHARE Working Paper Series 41-2019, Munich Center for the Economics of Aging (MEA): Munich.
- Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaaf, B., Stuck, S., & Zuber, S. (2013). Data resource profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, 42(4), 992-1001.
- Couper, M. P. (2011). The future of modes of data collection. *Public Opinion Quarterly*, 75(5), 889-908.
- de Bruijne, M., Pennings, S., & van der Wielen, I. (forthcoming). Software innovations: Quest. In M. Bergmann & A. Börsch-Supan (Eds.), *SHARE Wave 8 Methodology*. Munich: MEA, Max Planck Institute for Social Law and Social Policy.
- Kühne, S. (2018). From strangers to acquaintances? Interviewer continuity and socially desirable responses in panel surveys. *Survey Research Methods*, 12(2), 121-146.
- Scherpenzeel, A., Axt, K., Bergmann, M., Douhou, S., Oepen, A., Sand, G., Schuller, K., Stuck, S., Wagner, M., & Börsch-Supan, A. (2020). Collecting survey data among the 50+ population during the COVID-19 pandemic: The Survey of Health, Ageing and Retirement in Europe (SHARE). *Survey Research Methods*, 14(2), 217-221.

# Adapting a Long-Term Panel Survey to Pandemic Conditions

## *Adattare un sondaggio 'panel' a lungo termine alle condizioni pandemiche*

Peter Lynn

**Abstract** This paper summarises two major adaptations to *Understanding Society*, the UK Household Longitudinal Study, resulting from the covid-19 pandemic. First, main survey fieldwork was rapidly transitioned from one that relied heavily on face-to-face interviews to one that involved only web questionnaires and telephone interviews. Second, a new rapid online monthly survey was introduced, with telephone follow-ups. We summarise the design and impact of each.

**Abstract** *Riassumiamo due importanti adattamenti a Understanding Society, lo studio longitudinale delle famiglie del Regno Unito, risultanti dalla pandemia covid-19. In primo luogo, il lavoro sul campo del sondaggio principale è stato rapidamente spostato da uno che si basava principalmente su interviste faccia a faccia a uno che comprendeva solo questionari web e interviste telefoniche. In secondo luogo, è stato introdotto un nuovo sondaggio mensile online rapido, con follow-up telefonico. Riassumiamo il design e l'impatto di ciascuno.*

**Key words:** covid-19 pandemic, data collection mode, household panel survey, re-interview rate, telephone survey, web survey

---

<sup>1</sup>

Peter Lynn; Director of ISER, University of Essex. email address: [lynn@essex.ac.uk](mailto:lynn@essex.ac.uk)

## 1 Background

*Understanding Society*: the UK Household Longitudinal Study, is a national probability-based household survey that has been collecting data continuously since January 2009 (Institute for Social and Economic Research, 2019), primarily through face-to-face interviews. This paper presents an overview of the ways in which the survey has responded to the changed circumstances brought about in 2020 by the covid-19 pandemic. An earlier paper (Burton et al, 2020) described how protocols for the main survey fieldwork were adapted in response to the pandemic situation. We extend that work by examining the effect of the changes on field work outcomes. Additionally, we present an overview of the new *Understanding Society* monthly covid surveys that began in April 2020.

*Understanding Society* aims to provide a data resource for the research community and is funded primarily by the UK Economic and Social Research Council, with co-funding from a consortium of government departments. The sample is large: around 100,000 people in 40,000 households at wave 1 in 2009-10, and around 22,400 households at wave 11 (2019-20). At each annual wave, every sample member aged 16 or over is requested to take part in an interview of around 40 minutes, as is each other adult member of their current household. One person in each household completes a household interview (around 12 minutes), and children aged 10-15 years complete a paper self-complete questionnaire. Interview topics include employment, education, health, housing, income, social and family networks, and civic engagement.

The sample represents the entire UK population (Lynn, 2009) and includes boost samples of ethnic minorities and immigrants (Lynn et al, 2018). The sample is divided into 24 monthly samples, so fieldwork for each wave takes 24 months. As interviews take place annually, waves overlap: the first year of each wave takes place concurrently with the second year of the previous wave. Fieldwork is currently carried out under contract by two survey agencies, Kantar and NatCen.

The first six waves of data collection (2009-15) were carried out almost entirely by face-to-face in-home interviewing, (98% of interviews) with a small number of telephone interviews (CATI – 2%). At wave 7, web interviewing was introduced for the first time, but was offered only to sample members in households in which no-one had responded at wave 6. Wave 8 (2016-17) saw the introduction of a mixed-mode approach including web on a large scale (Carpenter 2018): 40% of sample members were asked to participate online (the “web-first” sample), a proportion that increased to 60% at wave 9 and 70% since wave 10. Nearly all those who did not complete the web survey were subsequently approached for a face-to-face at-home interview, as were the remaining sample members who were not invited to the web survey. This latter group (“CAPI-first”) consisted of a random 20% of the total sample plus the remaining households who were predicted to have the lowest probability of completing online. CATI continued to be used for a small number of interviews. At wave 10, 20,082 individual interviews (58.5%) were completed online, 14,051 face-to-face (40.9%) and 185 by telephone (0.5%).

## **2 Effect of the Covid Pandemic on Main Fieldwork**

### **2.1 Changes to Procedures**

The first change to *Understanding Society* in response to the pandemic was made on 11 March 2020 when new strict hygiene procedures were introduced for field interviews and some text was added to the participant website to reflect this. As the mood and news across Europe changed rapidly, just four days later, on 15 March, the *Understanding Society* Executive Team agreed to suspend all face-to-face interviewing, including on the ring-fenced CAPI-first sample. As it turned out, this decision would in any case have been forced upon us the following day, 16 March, as the UK government introduced measures to “stop all essential contact with others and to stop all unnecessary travel”. A message was added to the *Understanding Society* website to let participants know about the suspension of face-to-face fieldwork and that they would have the opportunity to complete their interviews online or by telephone instead.

As noted earlier, active waves of *Understanding Society* already included CAPI, CATI and web versions of the questionnaires and the means to manage sample households in a mixed-mode environment involving any combination of these modes. The survey was therefore well-equipped for a rapid shift away from CAPI. The decision was made to implement a sequential mixed-mode approach with all panel members, requesting a web interview initially, and following up by telephone if necessary. The main procedures needed to allow the shift to go ahead were ethical approval for the new design (obtained 18 March), the transmission of the CATI script to the laptops of field interviewers, who would now take over all responsibility for all telephone interviews (completed 20 March), and the mailing of a letter to all sample members who were in the process of being contacted. Further details of the field procedures can be found in Burton et al (2020).

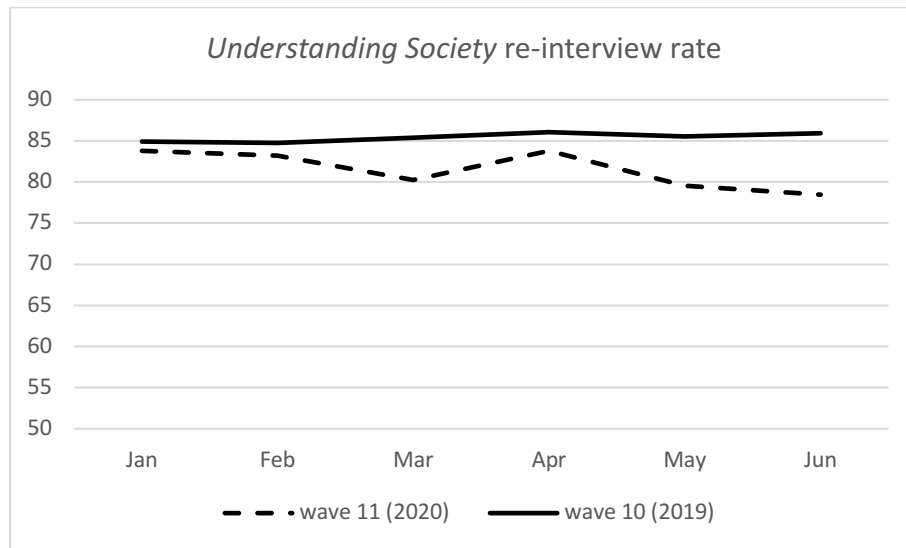
It should be noted that a large part of the March samples and small proportions of the January and February samples were active in the field at the time face-to-face interviewing was suspended. The April samples were the first ones to be attempted entirely with the new web-CATI sequential mixed-mode approach.

### **2.2 Participation**

The effect of the changed procedures on survey participation rates are not easy to estimate as the counter-factual is unknown. It is generally the case on *Understanding Society* that the re-interview rate (the proportion of persons interviewed at the previous wave who were also interviewed at the current wave) increases slightly at each wave, with the extent of the increase reducing over the waves. It may therefore be informative to compare the re-interview rates at wave 11 in 2020 to the equivalent rates for the same monthly samples one year earlier, at wave 10. We see (Figure 1)

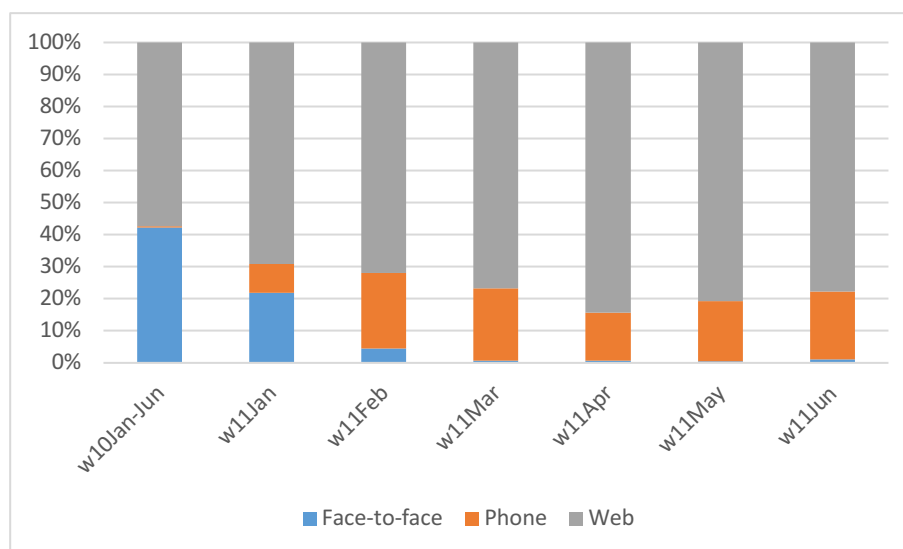
that for the March sample – for which fieldwork was briefly interrupted and then converted to a different mode – the re-interview rate appears to have suffered a little, being 80.3%, compared to 85.4% at the previous wave. But for the April sample, the first for which the new mode protocol was implemented from the start, the re-interview rate held up well at 83.8%. The fieldwork for this sample was conducted almost entirely during the first national lockdown in the UK, when having to remain at home was novel and being contacted to take part in a survey – or perhaps for any reason – was unusual. The re-interview rate then falls substantially in May and June, being fully seven percentage points lower than at the previous wave.

**Figure 1:** Re-interview rates at wave 10 (2019) and wave 11 (2020) for six monthly samples



The obvious effect of the change of protocol on the mode of data collection is much clearer (Figure 2). The proportion of interviews completed online was 69% for the January sample at wave 11 and 72% in February, but ranged between 77% and 84% for the March to June samples. The growth in CATI interviews was even more noticeable, from just 0.4% at wave 10 to 9% for the January sample at wave 11 and between 15% and 24% for the February to June samples. This great shift in the mode of completion has the potential to introduce mode measurement effects in longitudinal measures. The survey has taken measures to try to minimise these, but the need to discount mode effects will certainly receive some attention from analysts in the months to come.

**Figure 2:** Mode of interview at wave 10 (2019) and wave 11 (2020) for six monthly samples



### 3 New Monthly Surveys

The rapidly-evolving pandemic situation in early 2020 created a need for regular and rapid data on the many ways in which people's lives were being affected. While the long history of longitudinal information available for panel members and the breadth of socio-economic topics covered in *Understanding Society* provided considerable strengths, the two-year field period for each survey wave and one-year interval between interviews rendered the study unsuitable for immediate data needs. To overcome this, the *Understanding Society* COVID-19 Study was rapidly developed and implemented as a monthly web survey, starting in April 2020. After the first three waves the frequency was reduced to bi-monthly; seven waves have been conducted between April 2020 and January 2021. Additionally, two telephone follow-up waves have been carried out in an attempt to include panel members who would not be able or willing to participate online. The telephone follow-ups have been shown to improve the representativity of the samples (Benzeval et al, forthcoming).

From the second wave onwards, the research community was consulted regarding the contents of the study questionnaires and well over 100 submissions of suggested questions to include were received by the study team. All data from these new surveys have been made freely and promptly available through the UK Data

Peter Lynn

Service. The data from the April 2020 survey were released in late May, for example, and by December five waves of data were available. Over 1,200 researchers have down-loaded the data and more than 50 publications have already resulted. Findings from the study have been widely reported in the media and the study data have been used by the UK government's Scientific Advisory Group for Emergencies (SAGE) committee to help inform government policy. Further information about the *Understanding Society* COVID-19 Study can be found at <https://www.understandingsociety.ac.uk/topic/covid-19>.

## References

Benzeval, M., Burton, J., Crossley, T.F., Fisher, P., Gardiner, C., Jäckle, A., Moore, J.: High frequency online data collection in an annual household panel study: some evidence on bias prevention and bias adjustment. Under review.

Burton, J., Lynn, P., Benzeval, M: How *Understanding Society*: The UK Household Longitudinal Study adapted to the COVID-19 pandemic. *Surv. Res. Methods.* (2020) doi: 10.18148/srm/2020.v14i2.7746.



## 6.2.6 Young contributions in Covid-19 statistical modelling

# ***Statistical communication of COVID-19 epidemic using widely accessible interactive tools***

## ***Comunicazione statistica dell'epidemia di COVID-19 attraverso l'utilizzo di uno strumento interattivo***

M. Mingione and P. Alaimo Di Loro

**Abstract** *High-quality data is crucial for guiding decision making. Data quality frailties have been exposed worldwide during the current COVID-19 pandemic. The latter complicates the prediction of its evolution and the assessment of both health and economic interventions. Indeed, the process of data collection of the main pandemic variables is murky and not intended for statistical analysis, favoring convenient narratives and only apparently supporting policy-making processes. We aim at providing proper communication to the general public and inform on the daily evolution of the epidemic. That is achieved by the interactive tool here introduced, along with some alerts highlighting the fallacy of indicators as poorly informative when considered alone. We discuss the utmost importance to consider simultaneously multiple indicators, cross-verifying their behavior in order to distinguish relevant information from harmful and dangerous misinterpretations. Information are summarized through easily readable and accessible graphs and interactive maps. Predictions are based on novel approaches and models and can be used as alerts to identify at-risk situations.*

**Abstract** *Dati di alta qualità sono cruciali per guidare il processo decisionale. Lacune nella qualità dei dati sono emerse in tutto il mondo durante l'attuale epidemia di COVID-19. Queste lacune complicano la previsione dell'evoluzione dell'epidemia e la valutazione dei relativi interventi sanitari ed economici. Infatti, il processo di raccolta dati dei principali indicatori dell'epidemia è confuso e non progettato per l'analisi statistica, favorendo interpretazioni convenienti e soltanto apparentemente a supporto del processo legislativo. Il nostro scopo è quello di fornire al pubblico una corretta comunicazione statistica e di informare sull'andamento giornaliero dell'epidemia. Ciò si realizza attraverso uno strumento interattivo introdotto di seguito, in aggiunta ad alcune avvertenze che mostrano la scarsa informatività degli indicatori se considerati singolarmente. Si rileva la fondamentale importanza del considerare contemporaneamente più indicatori, mediante la verifica incrociata del loro comportamento, al fine di distinguere le informazioni rilevanti da interpretazioni errate, dannose e pericolose. Le informazioni sono riassunte in grafici e mappe interattivi. Le previsioni si basano su nuovi approcci e i modelli possono essere utilizzati come segnali per identificare le situazioni a rischio.*

**Key words:** COVID-19, Shiny, Data quality, Open data

---

Marco Mingione

University of Rome "La Sapienza", Statistical Science Department, e-mail: marco.mingione@uniroma1.it

Pierfrancesco Alaimo Di Loro

University of Rome "La Sapienza", Statistical Science Department, e-mail: pierfrancesco.alaimodiloro@uniroma1.it

## 1 Introduction

This work is the result of the joint project of a group of statisticians who share the same commitment to the social role of statistics, but are aware of the pitfalls that can stem from poor quantitative communication. In this regard, throughout the first year of the epidemic, the goal of our research group was manifold: (i) predict the evolution of the most relevant epidemic indicators and produce the forecast of the day of the peak for each curve; (ii) predict ICU occupancy by region to allow for an optimal allocation of health resources; (iii) sensitize the general public to the importance of correct statistical communication, allowing for a transparent and reproducible policy-making process.

COVID-19 public Italian data present several issues that severely affect their quality. Since the beginning of the epidemic, data have been collected for administrative and surveillance purposes mainly. Attention to the coherency, comparability and consistency of the collection process has been largely overlooked, hindering the inferential capability of any statistical analysis. To the best of our knowledge, data are and have always been gathered with very few shared standard guidelines. As a matter of fact, each regional healthcare department has its own different data collection and transmission system, which do not require compliance to any specific criteria. Measurement errors and errors in data entry are therefore expected to be often present, as well as substantial delays in reporting. Hence, any analysis of these data shall be limited to monitoring the *status quo* and produce scenarios projections rather than reliable medium to long-term predictions. In order to study and understand current and future states of the epidemic, higher quality and detailed information is of the utmost importance. Indeed, it is necessary that research groups are able to align the different indicators and follow the individual pathways of contagion and clinical evolution. Currently, the only recognized source of public data about the Italian COVID-19 pandemic is the Italian Protezione Civile (IPC) Github repository<sup>1</sup>. Data are aggregated and daily updated with the new flow of information coming from the regional system at around 6 p.m. . Despite all these limitations, StatGroup-19 believed that a more compelling and informative picture of the pandemic could be sketched using that data. This motivated the production of the *web application* described in Section 2.

## 2 A COVID-19 web app

The web application described here is built using R Shiny [8] and intends to provide the general public with a tool for accessing information about the Italian COVID-19 epidemic in an interactive and transparent way. The application is automatically updated at every user access with the most recent version available in the IPC Github repository and is accessible at <https://statgroup19.shinyapps.io/Covid19App/><sup>2</sup>. It shows both descriptive and model-based analysis, allowing the user to customize several choices. In particular, it is composed of 4 main panels: (i) "Overview", which provides a general description of the Italian epidemic; (ii) "Short-term forecast", which allows the modeling and short-term forecast of daily incidence indicators, at national and regional level; (iii) "ICU Nowcasting", which is specifically built to provide robust and trustworthy 1-day ahead intensive care unit (ICU) hospitalizations forecast; (iv) "Vaccines", which includes some useful information about the vaccination campaign in Italy. Plots, data and all source codes are public and can be freely accessed at <https://github.com/minmar94/StatGroup19>, in the spirit of a completely *Open Data* community.

<sup>1</sup> <https://github.com/pcm-dpc/COVID-19>

<sup>2</sup> English version of the app is available at <https://statgroup19.shinyapps.io/StatGroup19-Eng/>



Fig. 1: Overview - daily report.

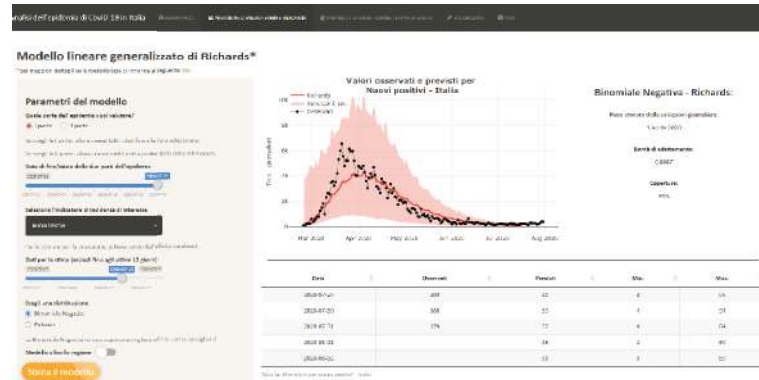
## 2.1 Overview

In the *Overview* page, the most relevant epidemic indicators are jointly recorded and visualized in order to provide an as accurate as possible picture of the Italian epidemic situation (both at national and regional level) based on data from the most recent update. The page is organized in different sections and includes: (i) the IPC daily report (see Figure 1), enriched with some more interpretable relative measures; (ii) the *decrees* timeline, allowing for the understanding and quantification of the eventual effects yielded by the measures adopted to contain the spread of the contagion; (iii) time-series and maps, so that a detailed investigation of the temporal and the spatial distribution of the available indicators and their ratios (e.g. positivity rate, fatality rate, healing rate, etc.) is possible; (iv) the table containing the daily raw data (the user can go back to the day in which the systematic data gathering process started, i.e. February 24, 2020), together with some relative indices for comparison among regions.

## 2.2 Short-term forecast of incidence indicators

This section provides short to medium term forecast of incidence indicators at both national and regional level. Incidence indicators measure the number of individuals with a particular condition, related with the epidemic, recorded during a given period. These indicators can be considered, by analogy with the terminology used in econometrics, as flow data, quantifying the daily input (e.g. positives) and output (e.g. deceased and recovered/discharged) of the system. We propose a parametric regression model for the modeling of incidence indicators based on the use of the Richard's curve [6] as response function in place of the widely used exponential or polynomial trend. Furthermore, we replace the generally entrenched Gaussian assumption for the distribution of log-counts [5; 7] by the more appropriate Poisson or Negative Binomial distributions for counts.

Further details on the specific methodology are described in [2]. The current version of the model provided robust and accurate forecasts during the first wave, but it is able to describe only one pandemic



(a)



(b)

Fig. 2: Short-term forecasting of incidence indicators (a) and nowcasting of ICUs (b).

wave at a time. The user can decide either modeling the first or the second wave, but a more comprehensive extension is currently under development and will be added in the next future. This page shows a graphical representation of the fit, predicted values and the 95% confidence intervals (with related coverage) up to the next 15 days and reports the estimated day of the *true peak* and various goodness of fit measures (see Figure 2a).

### 2.3 Nowcasting of intensive care units

The overcrowding of hospital facilities and the consequent risk of a breakdown of the National Health Care System is the greatest challenge this pandemic has put Italy through. Hence, monitoring the available ICU capacity is critical in order to act timely and prevent this from happening. We dedicated a specific section of the application to the 1-day ahead prediction of ICU occupancy for each region. Specific details of the methodology are described in [3]. The model is based on an optimal ensemble of two simple methods. The terms are a generalized linear mixed regression model [1], that pools information

over different areas, and an area-specific non-stationary integer autoregressive methodology [4]. Both regional population and ICU capacity are used as offsets in the modeling efforts.

As soon as Protezione Civile updates the Github repository, the app updates predictions for ICU occupancy for the next day. Point predictions are provided together with 99% confidence intervals. Since the beginning of the epidemic, the forecasts have always been accurate up to 3 – 6 beds at a regional level, with the confidence intervals containing the true future value in  $\approx 100\%$  of the cases. The user can compare and download predicted and observed values for each day (see Figure 2b).

## 2.4 The vaccination campaign

Italy (and the rest of the world) started seeing the light at the end of the tunnel on December 27, 2020. On this day, also called *V-Day*, the vaccination campaign started and, ever since, most of the effort put in place from the health-care systems has been dedicated to this task. The goal is to complete the vaccination of the whole Italian population (or at least the 70% of it) by the end of 2021. For this reason, we decided to dedicate a section to the monitoring of the Italian vaccination campaign. Percentages of vaccinated people are available at both national and regional level by gender, category and age class. The user can also customize a regional map in which administered vaccine doses are reported either in absolute value, either as a fraction of the delivered doses or as a fraction the residents.

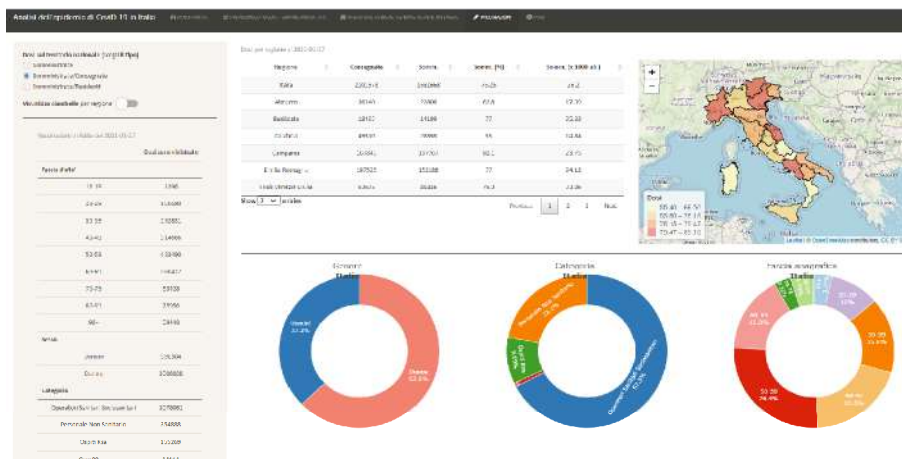


Fig. 3: Vaccination campaign in Italy.

## 3 Conclusions and further development

The web app described in Section 2 has been release on a <https://www.shinyapps.io/> server since the beginning of May 2020. Many other web apps have been devised with the same purposes during the last year, both at national and international levels. Our group wanted to give its own contribution in monitoring what we believe are the key aspects of the epidemic. Even though far from perfection, it has

drawn considerable attention since its first release. The app is particularly appreciated for its ease of usage and the interactive visualization tools that facilitates interpretation of the IPC data in a more friendly and perceptive way. The app is continuously under development, following the new possibilities and needs as well as the feedback and suggestions of the most zealous users.

Nevertheless, we must deal with the fact that the data necessary to construct more insightful and adequate information are currently in possession of government agencies and bodies, but not made available to the wide scientific community. We are perfectly aware that the guarantee of privacy and confidentiality are at stake, but we are concerned that further unknown considerations are limiting the proper pre-processing and masking that would turn the raw data into harmless accessible information. At this point in the evolution of the pandemic, the aggregated public data are no longer sufficient to make the government's decision-making mechanism transparent. More importantly, the scientific community has not been able to understand (and to replicate) some crucial quantities on which these decisions are taken.

## Acknowledgments

We would like to thank our fellow colleagues from the StatGroup-19, Professors Alessio Farcomeni, Fabio Divino, Giovanna Jona Lasinio, Gianfranco Lovison and Antonello Maruotti for getting us involved in this research adventure.

## References

- [1] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.
- [2] Pierfrancesco Alaimo Di Loro, Fabio Divino, Alessio Farcomeni, Giovanna Jona Lasinio, Gianfranco Lovison, Antonello Maruotti, and Marco Mingione. Nowcasting covid-19 incidence indicators during the italian first outbreak. *arXiv preprint arXiv:2010.12679*, 2020.
- [3] Alessio Farcomeni, Antonello Maruotti, Fabio Divino, Giovanna Jona-Lasinio, and Gianfranco Lovison. An ensemble approach to short-term forecast of covid-19 intensive care occupancy in italian regions. *Biometrical Journal*, 2020.
- [4] Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439, 2009.
- [5] G. Grasselli, A. Pesenti, and M. Cecconi. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: Early experience and forecast during an emergency response. *Journal of the American Medical Association*, 323:1545–1546, 2020.
- [6] FJ Richards. A flexible growth function for empirical use. *Journal of experimental Botany*, 10(2):290–301, 1959.
- [7] G. Sebastiani, M. Massa, and E. Riboli. COVID-19 epidemic in Italy: evolution, projections and impact of government measures. *European Journal of Epidemiology*, 35:341–345, 2020.
- [8] Hadley Wickham. *Mastering Shiny*. O' Reilly, 2020.

# Modelling COVID-19 evolution in Italy with an augmented SIRD model using open data

## *La modellizzazione dell'evoluzione del COVID-19 in Italia con un modello SIRD rivisto stimato su dati open*

Vincenzo Nardelli, Giuseppe Arbia, Andrea Palladino and Luigi Giuseppe Atzeni

**Abstract** We propose an augmented version of the traditional SIRD epidemic model and we estimate its parameters using the SaRs-Cov-2 Italian open-data. The model's parameters are estimated partly using numerical optimization and partly with ABC. Our estimation procedure provides a good fit to real data.

**Abstract** *Proponiamo una estensione del tradizionale modello epidemiologico SIRD e ne stimiamo i parametri usando i dati Italiani ufficiali relativi alla seconda ondata di diffusione del SaRs-Cov-2. I parametri sono stimati in parte tramite ottimizzazione numerica ed in parte usando il metodo ABC. Il procedimento produce un buon adattamento ai dati reali.*

**Key words:** epidemiological models, open data, Approximate Bayesian Computation, uncertainty evaluation

## 1 Introduction

The recent SARS-COV-2 epidemic is the first global pandemic in the big data era. Differently from other past epidemics, it developed even in technologically advanced countries and put the most innovative health systems in crisis. Moreover, this event brought to light different problems related to the quality of data and the

---

Vincenzo Nardelli  
University of Milan-Bicocca, e-mail: v.nardelli2@campus.unimib.it

Giuseppe Arbia  
Catholic University of Sacred Heart e-mail: giuseppe.arbia@unicatt.it

Andrea Palladino  
Apheris AI

Luigi Giuseppe Atzeni  
Boraso, e-mail: luigi.atzeni@boraso.com



related decision-making. Indeed, the public sector in most countries was not ready to collect, validate and distribute open data and the lack of statistical knowledge in the citizens and in most of the media led to the inability to clearly distinguish between “data” and “information” [1] [2]. A large number of researchers during the Covid pandemic have unsuccessfully required the access to anonymous individual data. Many active research groups, (among which e. g. [3]) developed models to predict the trend of the epidemic using all available open data, trying to mitigate the problems due to the poor data quality and to implement and estimate the model’s parameters together with its uncertainty. In the next section we will present our proposal.

## 2 Model definition

Historically, one of the first model used to predict the spread of the pandemic was the SIR model [4] based on a system of ordinary differential equations that models 3 categories of population (Susceptible, Infected, Recovered). In any given moment of time  $t$ ,  $I(t)$  and  $S(t)$  indicate respectively the number of infected people and the number of vulnerable people, while  $R(t)$  (removed) represents the total of those who develop immunity (recovered) or died. Obviously in any moment of time we have:  $S(t) + I(t) + R(t) = N$  with  $N$  the total population. The SIR model describes the variation of  $S(t)$ ,  $I(t)$ , and  $R(t)$  and the transitions from one category to the other. The original model specification does not consider population mobility in response to possible lockdown measures nor the impact of the asymptomatic. In this paper we propose an adaptation of [5] model which can be applied to model the spread of the epidemic in Italy using the available open-data diffused from Protezione Civile [6].

Our model is based on 6 categories, namely: **S**usceptible (people that can still be affected by the virus); **I**nfected (people that are currently infected); **H**ospitalized (people that need a medical treatment in hospital); **I**CU (people with severe symptoms that need to go to Intensive Care); **R**ecovered (people that recovered from the illness) and **D**eaths. We will refer to this model with the acronym “SIHCRD”.

The model is characterized by six non-linear ordinary differential equations:

$$\begin{aligned}
 \frac{dS}{dt} &= -\beta(t) \frac{S}{N} I & (1) \\
 \frac{dI}{dt} &= \beta(t) \frac{S}{N} I - \gamma I - k_1 I \\
 \frac{dH}{dt} &= k_1 I - k_4 H - k_2 H \\
 \frac{dC}{dt} &= k_2 H - k_5 C - k_3 C \\
 \frac{dD}{dt} &= k_3 C
 \end{aligned}$$

$$\frac{dR}{dt} = k_4H + k_5C + \gamma I$$

where  $\beta(t) = \beta_0 \cdot e^{-t/\tau}$ . The model is characterized by 8 free parameters. From the previous equations, we have  $S + I + H + C + D + R = \text{constant}$ . In what follows we describe the parameters in details.

- $\beta$  is related to the spread of the infection. Larger values of  $\beta$  corresponds to a fast spread of the epidemic;
- $\gamma$  is related to the (inverse of) time necessary to move from the category “infected” to the category “recovered”, without passing through hospital;
- $k_1$  is the product between the fraction of infected people that need to go to hospital (roughly 5% in the Italian experience) times the inverse of the average time required to move from “infected” to “hospitalized”;
- $k_2$  denotes the product between the fraction of hospitalized people that need to go to intensive care units (roughly 10% in Italy) times the inverse of the average time required to move from “hospitalized” to “intensive care units”;
- $k_3$  denotes the product between the fraction of patients that die times the average time that they stay in ICU (Intensive Care Unit) before the death;
- $k_4$  denotes the product between the fraction of people that do not go to ICUs (roughly 90%) times the inverse of the average time required to recover;
- $k_5$  denotes the product between the fraction of people the do not die in ICUs (roughly 70% during the second pandemic wave in Italy) times the average time required to recover;
- the parameter  $\tau$  denotes the timescale of the decreasing of the parameter  $\beta$

When  $k_1 = 0$  we go back to the original SIR model. The model contains some working hypotheses. The first is that people die only in ICUs. The second is that once a patients is recovered is removed from the susceptible, it it cannot be infected again [7].

### 3 Fitting procedure

In our study we fitted model (1) to the real Italian data during the second wave of the epidemic in the period October 1st and November 15th 2020. For the initial number of infected people we assumed the 6 million, estimated by [8]. All the others variables in Equation (1) are initialized according to the data available for the previous day. In the loss function, we assign the same weights of the errors (absolute percentage error) to hospitalized (H), patients in ICU (C) and deaths (D), while we don't use the number of infected people. This procedure allows us to make a more reliable estimate given the large uncertainties in evaluating the number of positive individuals and the irregularities in the transmission of data and in the testing procedures. A mixed approach was used to estimate the model parameters. For the parameters involving the transition between the categories of infected, the model

was fitted through numerical optimization starting from the estimates published by the Istituto Superiore di Sanità (ISS) <sup>1</sup> and other studies such as [9] and [10]. In particular, we used the optimizer algorithm Nelder-Mead [11] implemented in SciPy to tune 6 parameters of the model. In Table 1, we report the result of this optimization.

**Table 1** Parameter fit after numerical optimization

Parameter	Fitted value
$\gamma$	0.200
$k_1$	0.008
$k_2$	0.030
$k_3$	0.172
$k_4$	0.119
$k_5$	0.118

The remaining parameters ( $\beta$  and  $\tau$ ) were estimated through the Approximate Bayesian Computation - Sequential Monte Carlo (ABCSMC) [12] which allows to evaluate the uncertainty of the parameters considering the SIHCRD model as a black-box, starting from non-informative prior. In particular, the prior distribution was assumed to be Uniform between 0 and 1 for the  $\beta$  parameter and uniform varies between 0 and 600 for the  $\tau$  parameter. Fig. 1 shows the credible intervals of the estimations using PyABC [13] with a population size of 400 and the stopping rule with minimum error set to 1.5%. In Table 2, we report the estimation of the parameters from the posterior distribution.

**Table 2** Summary of the posterior distributions

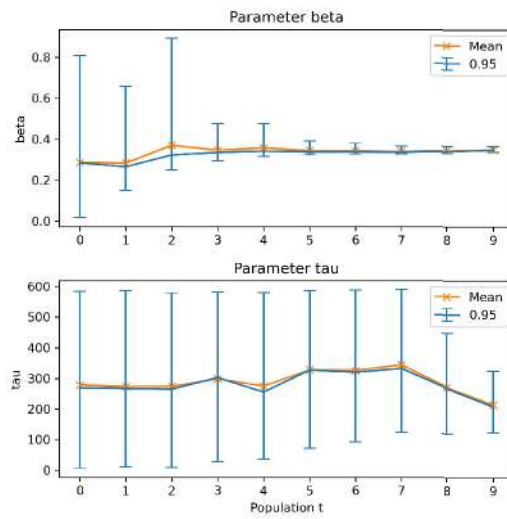
Quantile	0.025	0.5	0.975	Mean
$\beta$	0.33	0.34	0.36	0.34
$\tau$	123.73	228.66	326.67	226.20

## 4 Estimation results

Details about the proposed model are published online<sup>2</sup> where the results are constantly updated. During the second wave, the average error of the fit never exceeded 10% despite the great irregularity in the official data. The results are in agreement with other models published during the second wave.

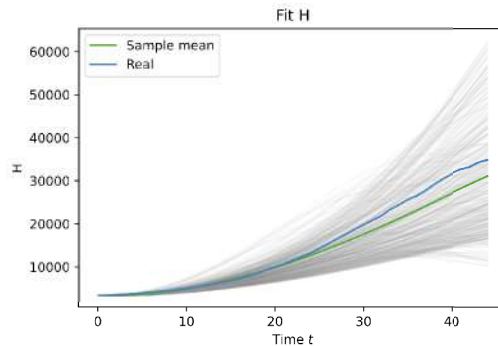
<sup>1</sup> <https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia>

<sup>2</sup> <https://dashboard.covstat.it/>



**Fig. 1** Estimates and credible interval of the parameters  $\beta$  and  $\tau$

The parameters' estimation with ABCSMC allows to model the uncertainty. An example of the fit of the hospitalized curve (H) is reported in Figure 2.



**Fig. 2** True and estimated of the posterior distribution of hospitalized patients (H)

## 5 Conclusions

In this paper we propose a mixed strategy to estimate the parameters of an augmented SIRD model combining numerical optimization and ABC procedures. In

this way we can calculate credible intervals for the crucial epidemic parameters thus helping their interpretation and their use in the monitoring and surveillance of the pandemic diffusion.

## References

- [1] G. Arbia and V. Nardelli, “I dati non parlano da soli: l’epoca del Coronavirus smaschera l’inganno dell’algoritmo-onnipotente e rivaluta il metodo statistico,” in *Giust. Insieme*, no. 923, 2020.
- [2] J. Zarocostas, “How to fight an infodemic,” *The Lancet*, vol. 395, no. 10225, p. 676, 2020.
- [3] Covstat, “Covstat - monitoraggio covid-19 in italia,” Nov 2020.
- [4] W. O. Kermack and A. G. McKendrick, “Contributions to the mathematical theory of epidemics-I,” *Bull. Math. Biol.*, vol. 53, pp. 33–55, mar 1991.
- [5] S. Khailaie, T. Mitra, *et al.*, “Development of the reproduction number from coronavirus sars-cov-2 case data in germany and implications for political measures,” *medRxiv*, 2020.
- [6] A. Palladino, V. Nardelli, L. G. Atzeni, N. Cantatore, M. Cataldo, F. Croccolo, N. Estrada, and A. Tombolini, “Modelling the spread of covid19 in italy using a revised version of the sir model,” *arXiv preprint arXiv:2005.08724*, 2020.
- [7] A. Wajnberg, F. Amanat, *et al.*, “Robust neutralizing antibodies to sars-cov-2 infection persist for months,” *Science*, vol. 370, no. 6521, pp. 1227–1230, 2020.
- [8] F. Bassi, G. Arbia, and P. Falorsi, “Observed and estimated prevalence of covid-19 in italy: how to estimate the total cases from medical swabs data,” *Science of The Total Environment*, vol. 764, p. 142799, 2021.
- [9] S. Richardson, J. S. Hirsch, *et al.*, “Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area,” *Jama*, vol. 323, no. 20, pp. 2052–2059, 2020.
- [10] G. Grasselli, A. Zangrillo, A. Zanella, *et al.*, “Baseline characteristics and outcomes of 1591 patients infected with sars-cov-2 admitted to icu of the lombardy region, italy,” *Jama*, vol. 323, no. 16, pp. 1574–1581, 2020.
- [11] F. Gao and L. Han, “Implementing the nelder-mead simplex algorithm with adaptive parameters,” *Computational Optimization and Applications*, vol. 51, no. 1, pp. 259–277, 2012.
- [12] T. Toni and M. P. Stumpf, “Simulation-based model selection for dynamical systems in systems and population biology,” *Bioinformatics*, vol. 26, no. 1, pp. 104–110, 2010.
- [13] E. Klinger, D. Rickert, and J. Hasenauer, “pyabc: distributed, likelihood-free inference,” *Bioinformatics*, vol. 34, no. 20, pp. 3591–3593, 2018.