SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board
A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Improving small area estimates of households' share of food consumption expenditure in Italy by means of Twitter data

Migliorare la precisione delle stime per piccola area della quota di spesa dei generi alimentari tramite i dati del social network Twitter

Marchetti S., Pratesi M., Giusti C.

Abstract In this work we use emotional data coming from Twitter as auxiliary variable in a small area model to estimate Italian households' share of food consumption expenditure (the proportion of food consumption expenditure on the total consumption expenditure) at the provincial level. We show that the use of Twitter data has a potential in predicting our target variable, reducing the estimated mean squared error with respect to what obtained by the same working model without the Twitter data.

Abstract *In questo lavoro si mostra come l'uso di dati ricavati da Twitter possa migliorare in termini di efficienza le stime per piccole aree della quota di spesa per generi alimentari a livello provinciale in Italia.*

Key words: Big Data, Area level model, Emotional data

1 Introduction

Recently, an increasing number of researchers have investigated the value of using big data (huge amounts of digital information about human activities) in socio-economic studies, see for example Eagle et al (2010); Blumenstock et al (2015); Decuyper et al (2014). Marchetti et al (2015) suggested three approaches to use big data in sinergy with small area estimation methods. Another approach to use big data in small area estimation was suggested by Porter et al (2014).

In this paper we focus on the use of data coming from the social network Twitter to investigate their potential in predicting the share of food consumption expenditure of Italian households at the province level. The paper has the following structure: the description of the data used in the analysis is in section 2; the small area estimation

Stefano Marchetti, Monica Pratesi, Caterina Giusti
University of Pisa, Via Ridolfi, 10, 56124 Pisa (PI), e-mail: stefano.marchetti@unipi.it

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

model is presented in section 3; the results of the application are detailed in section 4. Finally, we draw some concluding remarks in section 5.

2 Data used in the application

The primary source of data on households' expenditure in Italy is the Household Budget Survey (HBS) carried out annually by ISTAT. In 2012 the sample of the HBS was composed by approximately 28000 households. Data were collected on the basis of a two-stage sample design where the first stage were the municipalities and the second stage were the households. The regions (NUTS 2 level according to Eurostat) are the finest geographical level for which direct estimates of the target indicators are reliable. However, the knowledge of measures able to assess households' living conditions and well-being at a more detailed geographical level is often crucial, since this knowledge can for example enable policy makers in planning local polices aiming at reducing poverty and social exclusion (Giusti et al, 2016).

The households' consumption expenditure can be classified into food (and beverages) and non food expenditure. The share of total expenditure that an household dedicate to food items is an important indicator of the household living conditions: at risk of poverty households usually spend an higher share of their total expenditure on food with respect to the other households, with a lower impact of the share of expenditure dedicated to other resources and commodities.

To estimate the target at the province level we resort to model-based area-level small area methods, since direct estimates are unreliable. As possible sources of auxiliary variables – needed in model-based estimation – we use data coming from the Population and Housing Census 2011 and from the Survey[1] on Social Actions and Services on Single and Associates Municipalities 2012.

From the Population Census we collected information at provincial level such as the number of households, the average households' size, the tenure status, the female-headed households quota. As the target variable of our analysis can be considered as a proxy of the households' living conditions, we also considered as valuable source of auxiliary information the expenditure that Italian municipalities made in 2012 for interventions of social protection. These interventions includes the costs information on local welfare policies, such as services, benefits and transfers directed to households with children, old-age persons, poor and social excluded persons, immigrants.

Besides these sources of official statistics, we also considered as a potential source of auxiliary information big data from Twitter. In particular, we considered here as potential covariate for our small area working models the iHappy indicator referring to the year 2012. The iHappy indicator is made available every year since 2012 for all the 110 Italian provinces on the Opinion Analytics platform *Voices from the Blogs*. The iHappy indicator referring to the year 2012 was computed by

[1] This survey is a census survey, although some nonresponses can occur. Here we ignore the nonresponses and we use these data as census data.

collecting and coding more than 43 millions of tweets posted on a daily basis in all the Italian provinces. The words and emoticons of the tweets were classified using a training set in two categories: "happy" and "unhappy", together with a residual class "other". Then, Curini et al (2015) derived the frequency distribution of the happy and unhappy tweets in the entire population. The iHappy indicator was then computed for each Italian province as the percentage ratio of the number of happy tweets to the sum of happy and unhappy tweets. The overall average of the iHappy indicator in 2012 was equal to 44.5%, with a minimum value of 35.1% for Oristano and a maximum value of 56.6% for Sassari, both provinces of the Sardinia region. Indeed, the spatial variability of the iHappy values was rather high, as it is evident from the "emotional map" of Figure 1 (right).

3 Short review of the Fay-Herriot model for small area estimation

Data obtained from surveys are often used to estimate characteristics for subsets of the survey population. If the sample from a subset is small, then a traditional design-based survey estimator can have unacceptably large variance. These subsets has been defined as *small areas* (Rao and Molina, 2015).

In this study the available data allow us to rely only on area-level models (relate small area direct estimates to area-specific auxiliary variables). In addition, we do not have time-series data and the spatial correlation of the target direct estimates is low. So our choice falls on the Fay and Herriot (1979) estimator (FH). In what follows a summary description of the method is given.

Let m be the number of small areas and θ_i be the target parameter of the area i (mean or proportion). A survey provides a direct estimator $\hat{\theta}_i^{dir}$ of θ_i, $E[\hat{\theta}_i^{dir}] = \theta_i$ under the sampling design. A p-vector $\mathbf{X}_i$ contains the auxiliary data sources – exactly known – of population characteristics for area i. The FH model is as follows:

$$\hat{\theta}_i^{dir} = \mathbf{X}_i^T \beta + u_i + e_i \quad i = 1, \ldots, m, \tag{1}$$

where $u_i \overset{iid}{\sim} N(0, \sigma_u^2), i = 1, \ldots, m$ are the model errors and $e_i \overset{ind}{\sim} N(0, \psi_i^2), i = 1 \ldots, m$ are the design errors, with e_i independent from u_j for all i and j. It is assumed that the quantity of interest in area i is $\theta_i = \mathbf{X}_i^T \beta + u_i$.

Under the assumption of normality of both the errors (model and sampling design), the best linear unbiased predictor (BLUP) of θ_i is $\tilde{\theta}_i^{FH} = \gamma_i \hat{\theta}_i^{dir} + (1 - \gamma_i)\mathbf{X}_i^T \tilde{\beta}$, $\gamma_i = \sigma_u^2/(\sigma_u^2 + \psi_i^2)$, where $\tilde{\beta}$ is the Best Linear Unbiased Estimator of β. According to the theory of small area estimation (Rao and Molina, 2015), the parameters β and σ_u^2 are unknown and must be estimated, while ψ_i^2 is assumed to be known.

Estimators of β and σ_u^2 can be obtained using the restricted maximum likelihood from the marginal distribution $\hat{\theta}_i^{dir} \sim N(\mathbf{X}_i^T \beta, \sigma_u^2 + \psi_i^2)$. By plugging in the estimates of β and σ_u^2 into the BLUP we obtain the empirical best linear unbiased predictor

$$\hat{\theta}_i^{FH} = \hat{\gamma}_i \hat{\theta}_i^{dir} + (1-\hat{\gamma}_i)\mathbf{X}_i^T \hat{\beta}, \quad \hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}. \tag{2}$$

The estimator (2) has the following $MSE(\hat{\theta}_i^{FH}) = \gamma_i \psi_i^2 + (1-\gamma_i)^2 \mathbf{X}_i^T V(\hat{\beta}) \mathbf{X}_i + \psi_i^4 (\psi_i^2 + \sigma_u^2)^{-3} V(\hat{\sigma}_u^2) = g_{1i} + g_{2i} + g_{3i}$, where g_{2i} is the contribution to the MSE from estimating β and g_{3i} is the contribution to the MSE from estimating σ_u^2; $V(\hat{\beta})$ and $V(\hat{\sigma}_u^2)$ are the asymptotic variances of an estimator $\hat{\beta}$ of β and an estimator $\hat{\sigma}_u^2$ of σ_u^2, respectively. An estimator of the MSE is as follows

$$mse(\hat{\theta}_i^{FH}) = \hat{g}_{1i} + \hat{g}_{2i} + 2\hat{g}_{3i}, \tag{3}$$

where $\hat{g}_{1i} = \hat{\gamma}_i \psi_i^2$, $\hat{g}_{2i} = (1-\hat{\gamma}_i)^2 \mathbf{X}_i^T [\sum_{i=1}^m \mathbf{X}_i \mathbf{X}_i^T / (\psi_i^2 + \hat{\sigma}_u^2)]^{-1} \mathbf{X}_i$, $\hat{g}_{3i} = \psi_i^4 (\psi_i^2 + \hat{\sigma}_u^2)^{-3} 2[\sum_{i=1}^m 1/(\hat{\sigma}_i^2 + \psi_i^2)^2]^{-1}$.

4 Area-level small area model *with* and *without* Twitter data to estimate the share of food consumption expenditure in the Italian provinces

In this section we show that the use of Twitter data can improve the precision of the Share of Food Consumption Expenditure (SFCE) estimates in the Italian provinces, obtained using small area methods.

First, we estimated the SFCE at provincial level using the FH model (1) selecting the more predictive variables among the data described in section 2 without considering the iHappy variable, the one computed using Twitter 2012 data. In this way we obtained a reduction in MSE in all the provinces. Second, we added the iHappy variable to the other auxiliary variables and we estimated the SFCE again. If the iHappy variable is linearly correlated with the SFCE and this relation is not yet explained by the other auxiliary variables, then we expect a better performance in terms of MSE when using iHappy. We will show that the results obtained support this expectation.

The target variable, the SFCE, was obtained from the HBS 2012 survey as the ratio between the consumption expenditure for food (including beverages) and the total consumption expenditure. Its direct estimate at provincial level was obtained using the Horvitz and Thompson (1952) expansion estimator, $\hat{\theta}_i^{dir}$.

In 2012 the Italian provinces were 110 in total. However, in 2012 no HBS sample data were available for the province of Enna (Sicily) therefore it was not possible to obtain a direct estimate for this province, so we computed a synthetic estimator given that we know the auxiliary data for this province.

The selected auxiliary variables for the model without the iHappy variable are: the share of owners of the house x_1, the share of households lead by a female x_2, the per-household local government expenses to support several categories of citizens, households with children (x_3), old-aged persons (x_4), immigrants (x_5), at risk of poverty persons (x_6), services to families (x_7). So let $\mathbf{X}_i = [1, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}]^T$ be the design p-vector for model (1) for the area i,

where x_{ki}, $k = 0,\ldots,p = 7$, $i = 1,\ldots,m$, is the value of the kth auxiliary variable in area i (with $x_{0i} = 1$).

The FH model without the iHappy variable is then $\hat{\theta}_i^{dir} = \mathbf{X}_i^T\beta + u_i + e_i$. Estimates of β and σ_u^2 were obtained under the Normality assumptions made in section 3 using the restricted maximum likelihood (REML). From the analysis of $\hat{u}_i = \hat{\gamma}(\hat{\theta}_i^{Dir} - \mathbf{X}^T\hat{\beta})$, the Normality assumption seems reasonable. Indeed, the Shapiro and Wilk (1965) Normality test is equal to 0.978 with a p-value of 0.063.

To check the hypothesis that big data can help to increase the precision of the small area estimates - if used as auxiliary variables - we added the iHappy variable (x_8), obtained from the analysis of Twitter data as explained in section 2, to the set of the selected auxiliary variables ($x_1, x_2, \ldots, x_7$). Let $\mathbf{Z}_i = [\mathbf{X}_i, x_{8i}]^T$, where x_{8i} is the iHappy value for area i. The FH model is $\hat{\theta}_i^{dir} = \mathbf{Z}_i^T\beta^{BD} + u_i^{BD} + e_i^{BD}$, where the superscript ^{BD} refers to parameters under the model that makes use of big data (the Twitter data). Point and *mse* estimates are then obtained according to the methodology described in section 3 (replacing $\mathbf{X}_i$ by $\mathbf{Z}_i$).

In both the models - with and without iHappy variable - we selected the auxiliary variables using a step-wise procedure based on AIC (Hastie and Pregibon, 1992). The selected variables show a negative linear correlation with the target that range from -0.130 to -0.509. The negative correlations were expected for all the variable, but the share of households lead by a female. In general, in Italy, households lead by a female are positively correlated with poverty indexes and deprivation variables. However, we can suppose that the households lead by a female are associated with a reduction of the household size, so the expenses in food and beverages decreases so that to increase the SFCE. This hypothesis is supported by a linear correlation between the share of the households lead by a female and the household size equal to -0.857. As done for the model without iHappy variable, we estimated β^{BD} and σ_u^{BD} under the Normality assumptions made in section 3 using the REML. The Shapiro and Wilk (1965) Normality test for $\hat{u}_i^{BD}$s is equal to 0.980 with a p-value of 0.107.

The regression parameters estimated for both the models - with and without iHappy - are showed in table 1. The βs obtained under the two models are similar, the introduction of the iHappy variable in the FH model does not change significantly the model, it just add predictive power to it. The parameter σ_u is estimated

Table 1: Regression parameters of the FH model with/without the iHappy variable.

	$\hat{\beta}^{BD}$	p-value^{BD}	$\hat{\beta}$	p-value
Intercept	0.7165	0.0000	0.6446	0.0000
iHappy2012	−0.0019	0.0067	–	–
Share of owners of the house	−0.0038	0.0000	−0.0039	0.0000
Share of household lead by female	−0.3164	0.0009	−0.3222	0.0012
Expenses for household with children	−0.0001	0.2121	−0.0002	0.0513
Expenses for old-aged persons	−0.0001	0.0123	−0.0001	0.0280
Expenses for immigrants	−0.0013	0.0003	−0.0013	0.0009
Expenses for at risk of poverty persons	0.0006	0.0009	0.0007	0.0006
Expenses for services to families	−0.0005	0.0460	−0.0006	0.0246

equal to 0.020 for the model without iHappy and to 0.019 for the model with iHappy. To verify the null hypothesis that $\sigma_u^2 = 0$, we used the test proposed by Datta et al (2011) and we reject the null hypothesis $\sigma_u^2 = 0$ for both the models.

It is important to highlight that the iHappy indicator is based on self-selected data, the Twitter data. However, in this application we are not able to treat the self-selection bias due to lack of information. Thus, we assume that the self-selection is negligible. Moreover, the iHappy indicator can be affected by measurement error, since not any happy tweet corresponds to a happy person. In our application the MSE of the iHappy is very small, due to the very large sample size (43 millions of tweets), therefore model that account the measurement error, such as the one proposed by Ybarra and Lohr (2008), approximately corresponds to the traditional FH model.

Results on the SFCE estimates are summarized in table 2. Using the FH estimator (2) with the set of auxiliary variables $\mathbf{X}_i$s the *rmse* is reduced in all the provinces. The average reduction is about 30% with a 25% of provinces where the reduction is at least about 40%(table 2). Moreover, using also the iHappy variable the reduction of the *rmse* results in an average gain of 2%. A clearer picture of the gain in precision due to the introduction of the iHappy variable in the FH model can be see in the last line of table 2, which shows the efficiency of $\hat{\theta}_i^{FH,BD}$ against $\hat{\theta}_i^{FH}$. There is a gain in all the areas, but one where we observe a loss of 0.5%. The gain goes from about 2% up to about 7%. Given that the small area estimates obtained without the use of the iHappy variable show a remarkable gain in terms of reduction of *mse*, the further reduction of the *mse* due to the introduction of the iHappy variable in the model is a very good result. This is particularly important also because the iHappy variable can be computed every year, while updated census information on the population is not always available.

Table 2: Summary of point estimates of SFCE and their efficiency.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{\theta}_i^{Dir}(\%)$	15.38	19.44	21.34	22.45	25.56	35.42
$\hat{\theta}_i^{FH}(\%)$	15.37	19.70	21.60	22.19	24.47	29.91
$\hat{\theta}_i^{FH.BD}(\%)$	15.44	19.68	21.64	22.17	24.65	29.55
$rmse(\hat{\theta}_i^{FH})/rmse(\hat{\theta}_i^{Dir})(\%)$	19.79	60.66	74.90	70.38	82.16	99.39
$rmse(\hat{\theta}_i^{FH.BD})/rmse(\hat{\theta}_i^{Dir})(\%)$	18.44	58.29	72.37	68.49	80.35	99.43
$rmse(\hat{\theta}_i^{FH.BD})/rmse(\hat{\theta}_i^{FH})(\%)$	93.18	95.73	97.33	97.02	98.22	100.50

In order to obtain a clearer picture of the estimates across the country, we mapped them out in figure 1. In the same figure we contrast our estimates with the map of the iHappy variable to show the relationship between the two variables. The SFCE point estimate for the out of sample province of Enna has been computed using the regression synthetic estimator (see Rao and Molina, 2015). In particular, the estimated SFCE for the province of Enna is 25.29% with an rmse of 1.98%. These results seem plausible according to the estimates obtained for the neighbors provinces.

In Italy the SFCE is 22.2% at national level, showing that in average the consumption of food does not represent a large amount on total expenses for con-

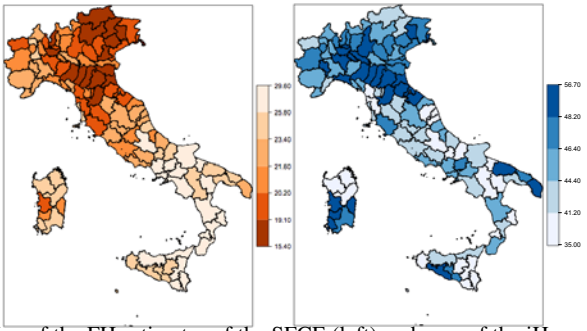

Fig. 1: Map of the FH estimates of the SFCE (left) and map of the iHappy variable for 110 provinces in Italy (right).

sumption. At provincial level (table 2) the SFCE varies between 15.44% (Ravenna, central Italy) and 29.55% (Caserta, southern Italy), so there is evidence of spatial heterogeneity. About a quarter of the provinces have an SFCE $\geq$ 25%. All these provinces are in the southern part of Italy. Nine provinces have an estimated SFCE that is below 18%, five of these provinces are in the central part and the other four are in the northern part of Italy, confirming the well known Italian north-south divide.

For a more detailed description of this application see Marchetti et al (2016)

5 Conclusions

In this paper we focused on the iHappy indicator obtained from the analysis of Twitter data. The data consist of all the geo-referenced tweets posted in 2012 in the Italian provinces, classified by Curini et al (2015) as the percentage of happy tweets to the total of tweets at provincial level.

In our analysis the iHappy indicator resulted a good additional covariate to predict households' SFCE, given the net influence of other covariates characterizing the provinces, such as the tenure status of the house, the gender of the head of the households, the level of the expenses of the local government to support vulnerable groups.

In Italy the SFCE shows a territorial variability that mimics that of many socio-economic indicators: in 2014 the north-eastern and north-western part of Italy had the lowest level of SFCE (respectively 15.7% and 15.5%) while the southern part (islands included) had the highest (21%). This north-south divide is evident also from the territorial distribution of the iHappy indicator, with few exceptions (some provinces of Sardinia, Puglia and Sicily).

Concluding, the iHappy indicator on happiness can provide useful covariates on yearly bases, free of charge and broken by provinces. It comes affected by self-

selection bias and measurement error. In this application we assumed that the self-selection is negligible and that the measurement error appears to be a minor issue.

References

Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350:1073–1076

Curini L, Iacus S, Canova L (2015) Measuring idiosyncratic happiness through the analysis of twitter: An application to the italian case. Social Indicators Research 121(2):525–542

Datta G, Hall P, Mandal A (2011) Model selection and testing for the presence of small area effects, and application to area-level data. Journal of the American Statistical Association 106:362–374

Decuyper A, Rutherford A, Wadhwa A, Bauer J, Krings G, Gutierrez T, Blondel V, Luengo-Oroz M (2014) Estimating food consumption and poverty indices with mobile phone data. Tech. rep., UNITED NATIONS GLOBAL PULSE

Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328:1029–1031

Fay R, Herriot R (1979) Estimation of icome from small places: An application of james-stein procedures to census data. Journal of the American Statistical Association 74:269–77

Giusti C, Masserini L, Pratesi M (2016) Local comparisons of small area estimates of poverty: an application within the tuscany region in italy. Social Indicators Research

Hastie T, Pregibon D (1992) Generalized linear models, Wadsworth and Brooks/Cole, chap 6

Horvitz D, Thompson D (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47:663–85

Marchetti S, Giusti C, Pratesi M, Salvati N, Giannotti F, Pedreschi D, Rinzivillo S, Pappalardo L, Gabrielli L (2015) Small area model-based estimators using big data sources. Journal of Official Statistics 31:263–281

Marchetti S, Giusti C, Pratesi M (2016) The use of twitter data to improve small area estimates of households' share of food consumption expenditure in italy. AStA Wirtsch Sozialstat Arch 10(79)

Porter A, Holan S, Wikle C, Cressie N (2014) Spatial fay-herriot models for small area estimation with functional covariates. Spatial Statistics 10:27–42

Rao J, Molina I (2015) Small Area Estimation. Wiley Series in Survey Methodology, Wiley, URL `https://books.google.it/books?id=i1B_BwAAQBAJ`

Shapiro S, Wilk M (1965) An analysis of variance test for normality (complete samples). Biometrika 67:215–216

Ybarra L, Lohr S (2008) Small area estimation when auxiliary information is measured with error. Biometrika (95):919–931