InGRID

Integrating expertise in inclusive growth

www.inclusivegrowth.be

Deliverable 23.1

CASE STUDIES

Yves G. Berger, Jan Pablo Burgard, Adrian Byrne, Alexandru Cernat, Caterina Giusti, Pinar Koksel, Simon Lenau, Stefano Marchetti, Hariolf Merkle, Ralf Münnich, Iñaki Permanyer, Monica Pratesi, Nicola Salvati, Natalie Shlomo, Duncan Smith and Nikos Tzavidis

February 2016

This project has received funding from the European Union's Seventh Programme for Research, Technological Development and Demonstration under Grant Agreement No 312691

Preface

The InGRID workpackage 23 on high-performance statistical quality management focuses on improvements in the development and accuracy of indicators. That requires the construction of a shared knowledge on theories and best practices to judge the quality and appropriateness of indicators through an empirical analysis. This covers methodological advances as well as practical considerations of indicators for poverty, social exclusion, and related fields. Additionally to methodological advances, a simulation lab is developed to foster open and reproducible research for future developments in the InGRID research area using SILC related data.

This first deliverable covers methods in several areas:

- Multidimensional Indicators
- Non-response and Imputation
- Small Area Estimation, and
- Measuring Level and Change

As an additional asset, the AMELIA dataset, which was originally started within the FP7 project AMELI (http://ameli.surveystatistics.net), was further developed to provide a sound basis for comparable and reproducible research. The start of AMELIA was related to SILC data. Further developments will integrate time change and will allow to enhance the data with other sources of interest for the InGRID research.

Ralf Münnich, Co-ordinator of WP23

This report constitutes Deliverable 23.1 ‘Case Studies’, for Work Package 23 of the InGRID project.

February 2016

General contact: inclusive.growth@kuleuven.be
p.a. InGRID
HIVA - Research Institute for Work and Society
Parkstraat 47 box 5300, 3000 LEUVEN, Belgium

For more information lenau@uni-trier.de

Please refer to this publication as follows:
Berger, Y. G., Burgard, J. P., Byrne, A., Cernat A., Giusti, C., Koksel, P., Lenau, S., Marchetti, S., Merkle H., Münnich, R., Permanyer, I., Pratesi, M., Salvati, N., Shlomo, N., Smith D. and Tzavidis, N.: Case Studies. InGRID Deliverable 23.1.

Information may be quoted provided the source is stated accurately and clearly.
This publication is also available via http://www. Inclusivegrowth.be/project-output

This publication is part of the InGRID project, this project has received funding from the European Union’s Seventh Programme for Research, Technical Development and Demonstration under Grant Agreement No 312691.

The information and views set out in this paper are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

InGRID
Integrating expertise in inclusive growth

European policy-oriented research can and must deliver useful contributions to tackle the Europe 2020 challenges of Inclusive Growth. Key tools in this social sciences research are all types of data earning statistics, administrative social data, labour market data, surveys on quality of live or working conditions, policy indicators. The project aims to integrate and optimise these existing European data infrastructures and accompanying expertise.

Contents

1 The AMELIA Dataset - A Synthetic Universe for Reproducible Research

Hariolf Merkle **Jan Pablo Burgard** **Ralf Münnich**

Economic and Social Statistics Department

Trier University

1.1 Introduction

Open and reproducible research is one of the important steps in modern research. For survey statistics when introducing new estimators and methods, this is often accompanied with simulation studies. The InGRID research, however, focusses of SILC data and methods using these data. Poverty measurements is one of the major examples in this context. To evaluate survey statistical methodology adequately, real(istic) data shall be considered as a sound basis for performing simulation studies. The aim of AMELIA is to provide a realistic synthetic universe that is based on SILC data and its structures.

Changing simulation settings my lead to interesting behavior or to discovery of peculiarities which hardly can be found with mathematical proofs and sometimes the observations may lead to proofs. The use of careful set-ups makes it possible to learn about the applicability and gather experience of the different kinds of estimators in various situations. In practice, we have only one sample. This raises the question if our approach is adequate. But the set-up including the corresponding evaluation of a Monte Carlo study is utmost important. There should be a distinction between examples and simulation studies.

However, there is a problem regarding the availability of appropriate data especially for design-based simulations. This type of simulations requires a given universe from which samples can be drawn. Usually, a dataset comprising variables of the whole population is not obtainable due to confidentiality reasons or it does not simply exist. For this reason, an artificial dataset has to be generated making use of sample data. The artificial population should have similar properties as the sample data. Also, the confidentiality requirements have to be considered. This chapter deals with the generation of the artificial universe called AMELIA. At first, the requirements are presented. Later on, the generation process, the properties and further developments are shown.

1.2 Requirements

The AMELIA dataset is a synthetic dataset created on the basis of the *European Union Statistics on Income an Living Conditions* (EU-SILC) scientific-use-file from 2005 and originates from the project *Advanced Methodology for European Laeken Indicators* (AMELI). During this project, AMELIA was created to evaluate estimators considering different sampling designs using design-based simulations. Alfons et al. (2011a) and Kolb (2013) give a more detailed description of the AMELIA data. The Sections 1.3 and 1.4 as well as this one refer to both last-mentioned publications. The results of the AMELI simulation study are presented in Hulliger et al. (2011).

The dataset contains variables on person- and household-level which encompass variables regarding social, economic, and regional attributes. The following list shows an overview on some of the requirements that had to be met for the AMELI project:

- Heterogeneity: the heterogeneity of different entities has to be considered to evaluate the performance loss of small area estimators.
- Administrative structure: an administrative structure must be available for the implementation of different sampling designs and for the evaluation of small area and small domain methodologies.
- Micro structures: the household structure which differs across countries in the original dataset has to be retained.
- Income components: income components are necessary to calculate Laeken indicators (e.g. quintile share ratio, at risk of poverty rate) and should reflect the distributions of EU-SILC.
- Zero values of income components: the dependencies between different income components must be considered for consistency.
- Outliers: the treatment of outliers has to be regarded.
- Missing values: missing values in the original dataset have to be replaced by synthetic values.
- Editing: logical editing is necessary to ensure consistency, statistical editing is important for the data structure.

1.3 Generation of AMELIA

This section deals with the generation of the AMELIA dataset. Since this is a synthetic dataset that does not reproduce the true underlying population exactly, the size of the synthetic population characteristics have to be defined at first. These are number of persons, number of regions and the size of the regions. These regions are artificial and were generated by grouping countries from the EU-SILC dataset.

In general, the generation of the AMELIA dataset is based on sampling from distributions. The initial step of the generation of the AMELIA universe is based on resampling of households according to their respective weight from the EU-SILC dataset. Households are drawn with replacement within their particular artificial region to which they are assigned beforehand. For this purpose, the set of variables is divided into blocks because pure resampling with all desired variables simultaneously would lead to exact replications of the households. In this case, the synthetic population would contain real confidential data.

In the first block, basic variables like household ID, age class, sex and marital status are included to preserve the household structure. Households are drawn with these variables because this basic household structure is complicated to generate. Every household in the synthetic dataset is assigned a new unique ID that is independent from the original dataset. Furthermore, a latent class model is used to generate placeholders for profile clusters, e.g. the economic profile which encompasses variables like activity status and employment relationship. Further variables are generated block-wise. The blocks group variables that are dependent to each other. These blocks should be as far as possible independent from each other.

The dataset contains many different income components. The original variables are divided into categories and the respective category is drawn. Within these categories, the corresponding value is drawn from a uniform distribution in a first step. Later on, rejection sampling is applied. This leads to an approximation to the original distribution from the respective variable of the EU-SILC dataset. It is necessary to obtain similar values for the Laeken indicators from the synthetic population as from EU-SILC data. Therefore, the income distribution should reflect the original distribution.

1.4 Properties of AMELIA

An important feature of the AMELIA dataset is its hierarchical structure. There are many regional hierarchical levels which are usually not available in other micro datasets. These administrative structures are also important to draw samples using different sampling designs. The hierarchical structure of AMELIA is shown below:

1. The AMELIA country itself
2. 4 regions
3. 11 provinces
4. 40 districts
5. 1,592 cities/communities
6. 3,781,289 households
7. 10,012,600 persons

Adapting existing and creating new structures of hierarchical levels between and within the respective hierarchical components is essential to account for heterogeneity. There is also an artificial map available. It was decided to create an artificial map so there is no direct spatial relation between individuals in the AMELIA population and in reality. Figures 1.1 and 1.2 display the map of AMELIA including regional structures whereas Figure 1.1 shows the regions and provinces. Figure 1.2 depicts the districts and cities/communities.

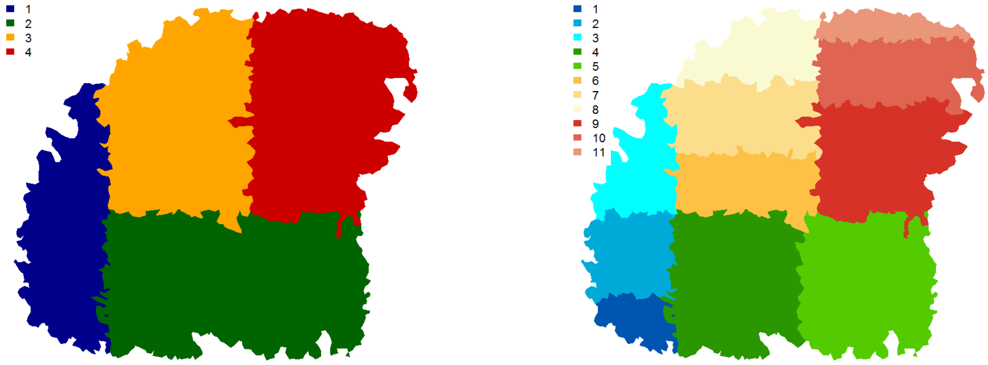


Figure 1.1: Regions (left) & provinces (right) of AMELIA

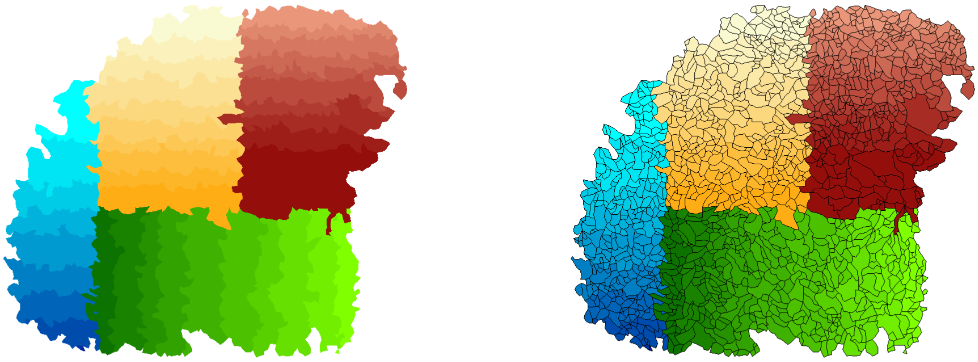


Figure 1.2: Districts (left) & cities/communities (right) of AMELIA

Figure 1.3 depicts the spatial distribution of the equivalised disposable income (EDI) by cities/communities. Region 3 has the highest equivalised disposable income on average whereas the differences between the other regions are not very large. The means by region are shown in Table 1.1.

REGION	1	2	3	4
EDI (mean)	27,085	27,342	42,526	31,424

Table 1.1: Equivalised disposable income by regions

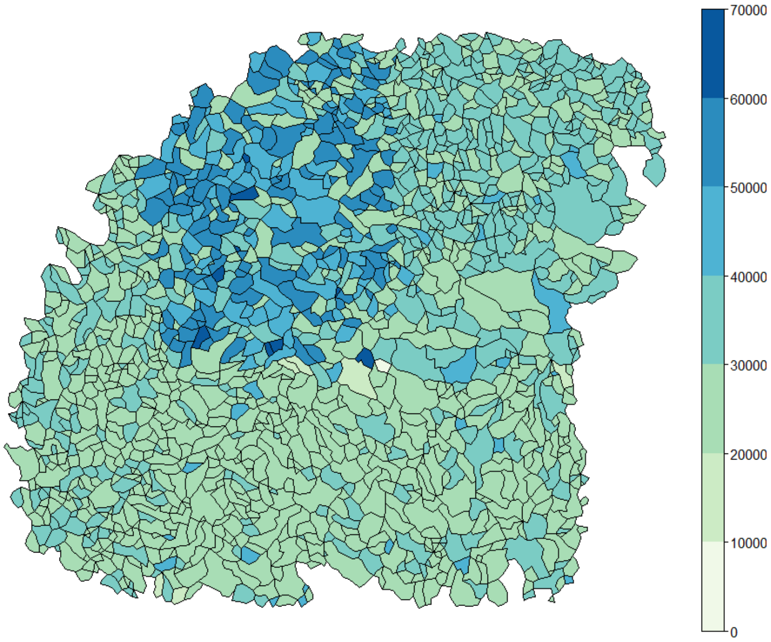


Figure 1.3: Equivalised disposable income by cities/communities

1.5 Steps Forward

The first part of Section 1.5 explains the evolution of the AMELIA dataset since the AMELI project. The second part looks ahead of the status quo.

1.5.1 Proceedings of AMELIA

For the AMELI project, AMELIA was created to provide a universe for design-based simulations to researchers. Since the original EU-SILC data has to be protected and is only a sample, this synthetic dataset was generated. There was not only the need of an artificial population for the AMELI-project, AMELIA is also used for InGRID. This dataset is used in Chapter 5 for a simulation study which deals with variance estimation for estimators where the variable of interest is a categorised continuous variable.

After the AMELI project, the artificial population is still maintained and developed at the

Economic and Social Statistics Department at the Trier University. For instance, maintaining involves restructuring and the creation of additional variables which is currently the challenge. Making use of the *Eurosystem Household Finance and Consumption Survey* from the European Central Bank, variables describing the amount of different financial assets are currently being generated using regression model approaches. The concepts of these regression model approaches considered here to create new variables are described in Alfons et al. (2011b).

For the asset variables, the two-step regression approach was used. The first step involves the estimation of a logit model. This is necessary for the treatment of semi-continuous variables. The predictions from the logistic regression indicate whether the respective outcome of a variable is zero or not. For the non-zero outcomes, a linear regression is conducted with a logarithmized dependent variable and random draws from the residuals are added. Otherwise, all individuals with the same characteristics would have identical outcomes.

In addition to this, the *Labour Force Survey* is taken as a basis for the estimation of a logit model indicating individuals with occupation as supervisors. This model is estimated only for individuals that are marked as employed. Also, the prediction of this variable is conducted on the respective individuals. The reason for this is to account for consistency. Otherwise, it might be possible that some unemployed individuals turn out as supervisors. Considering this issue in the estimation and in the generation process, the effort of editing can be set to a minimum. Once the models are estimated, the outcomes of new variables are predicted using the relevant variables from the AMELIA dataset as regressors.

Figure 1.4 shows the share of supervisors at working individuals. This share does not show a large variation across the 40 districts and is often just under 30%.

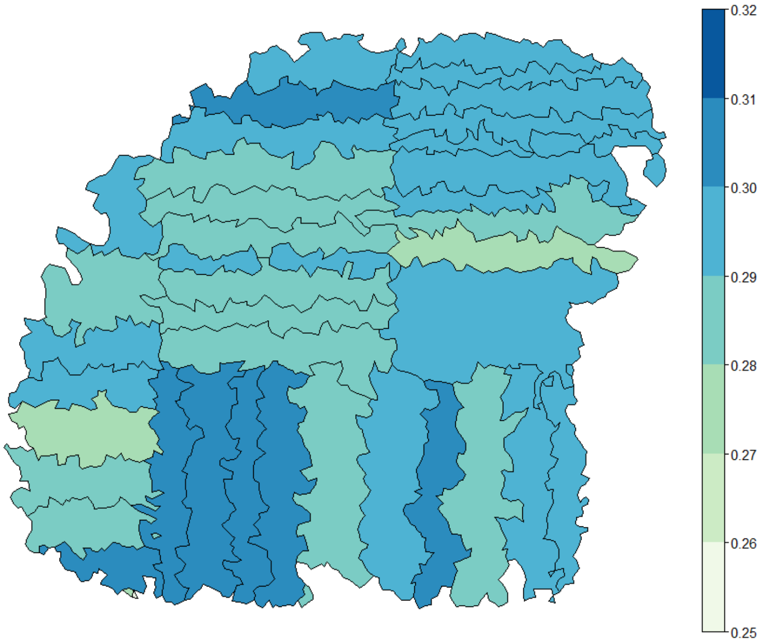


Figure 1.4: Share of supervisors at working individuals by district

1.5.2 The Future of AMELIA

There is not only the need of individual researchers for an artificial universe for design-based simulations. Moreover, there is also the issue of reproducing results from statistical studies. Often, reviewers of journals demand for the data that is used in the study to reproduce the results. Unfortunately, the data can not be delivered due to confidentiality reasons.

Amongst others, Peng (2015) describes some cases with issues concerning the reproducibility of scientific results. Also, Stodden (2015) deals with the issue of reproducing statistical results. She presents five reasons why reproducibility using statistical methods might fail. These are low power and sampling issues, misapplication of statistical tests, robustness and lack of generalization, lack of access to data, software and tools of analysis, and ineffective cultural incentives.

In the near future, AMELIA will be publicly available on a web platform. Researchers can download the variables they need as well as the already drawn samples of different sampling designs. Most of them are similar to the sampling designs that were used in Hulliger et al. (2011). These encompass one- and two-stage sampling designs considering simple random sampling, stratified sampling and sampling proportional to size. There already exist web platforms to enable reproducible statistical research, e.g. researchcompendia.org (Stodden et al., 2015), and RunMyCode.org (Hurlin et al., 2014). These platforms link code and data as well as the article so other researchers can validate the results of the respective studies. The web platform here will be slightly different. The main difference is the usage of only that one specific AMELIA dataset.

The following list shows one of the most important properties of this platform:

- English language
- freely accessible
- version control if updates are available
- link publications with the platform
- open to extensions
- url: `http://amelia.surveystatistics.net/`

Soon, a more detailed description of the available variables and sampling designs will be disseminated via the web platform. From this point on, the evolution of AMELIA will also be community-driven. Users have the possibility to develop the dataset and the supplements, i.e. the sampling designs or new variables. The development of AMELIA is a long term process. Currently, adding the longitudinal component to AMELIA is part of the work. This creates the opportunity to consider longitudinal sampling designs, e.g. rotation sampling designs.

2 Robust domain estimation of income-based inequality indicators

Stefano Marchetti
Department of Economics and Management
University of Pisa

2.1 Introduction

Social deprivation is the reduction or prevention of culturally normal interaction between an individual and the rest of society. Social deprivation is included in a broad network of correlated factors that contribute to social exclusion and include mental illness, poverty, poor education, and low socioeconomic status. The European Union has proposed a core of statistical indicators to quantify deprivation and social exclusion in each country commonly known as Laeken Indicators. This is a set of 21 indicators that includes the incidence and the intensity of poverty for a set of domains (for example, young unemployed individuals), the incidence of particular social situations (for example, low education), inequality measures and life expectancy. These indicators were agreed by the European Council in December 2001, in the Brussels suburb of Laeken (cf. Commission of the European Communities, 2003).

The Laeken indicators at national or regional levels (NUTS 1 and 2 level) are usually estimated by the Survey of Income and Living Conditions (EU-SILC). This survey is often designed to obtain accurate estimates at NUTS 1 or 2 level while policy makers require estimates of social exclusion and living conditions at finer levels of geographical/domain disaggregation for which direct estimates are often inaccurate. This is mainly due to the small sample sizes for the domains of interest. Given that oversampling is not feasible due to budget constraints, there is need to resort to small area estimation methods.

The majority of the small area literature focuses on the estimation of domain averages, but recent methodology has also focused on the estimation of non-linear statistics (incidence of income poverty, the poverty gap and percentiles of the income distribution function). More specifically, until very recently the industry standard for estimating poverty indicators was based on the so called World Bank (WB) method, proposed firstly by Elbers et al. (2003). From a small area perspective Molina and Rao (2010) proposed the Empirical

Best Predictors (EBP) that is the best estimation method for in-sample domains when the normality assumptions of the nested error regression model hold. As an alternative to the EBP Marchetti et al. (2012a) and Tzavidis et al. (2010a) proposed robust small area methods of poverty indicators when the parametric assumptions of the nested error regression model do not hold.

In this work the focus is on extending the methodology in Marchetti et al. (2012a) and Tzavidis et al. (2010a) for estimating two specific Laeken indicators namely, the income quintile share ratio (S80/S20) and the Gini coefficient (G). First, we define the two income-based inequality measures of interest. Second, we present the M-quantile-based small area methodology for estimating the two inequality measures and the related Mean Squared Error estimator. Then, we present results from Monte-Carlo simulations and from an application of the methodology to real data. The aim of this application is to obtain estimates of G and S80/S20 for unplanned domains (domains that do not feature in the design and allocation of the EU-SILC sample) in Tuscany. For the application in this chapter these domains are defined by provinces in Tuscany cross-classified by the gender of the head of the household. Finally, some final remarks and open areas for research.

2.2 Definition of S80/S20 and G

S80/S20 is the ratio of the total income received by the 20% of the country's population with the highest income (top quintile) to that received by the 20% of the country's population with the lowest income (lowest quintile).

Denote by N the population size, by $q_{0.2}$ and $q_{0.8}$ the 20th and 80th percentiles of the income distribution and by y_j the corresponding income for household j. The S80/S20 is defined as follows,

$$S80/S20 = \frac{\sum_{j=1}^{N} \left[y_j \mathbb{I}(y_j > q_{0.8}) \right]}{\sum_{j=1}^{N} \left[y_j \mathbb{I}(y_j < q_{0.2}) \right]}. \tag{2.1}$$

G measures the relationship of the cumulative shares of the population arranged according to the level of income, to the cumulative share of the equivalised total net income received by them.

Denote further by (j) the order of the income of household j. Let's assume that y_j denotes the ordered income values. Then, G can be defined as follows,

$$G = \frac{N+1}{N} - \frac{-2\sum_{j=1}^{N}[N+1-(j)]y_j}{N\sum_{j=1}^{N} y_j}. \tag{2.2}$$

2.3 M-quantile approach to Small Area Estimation of Inequality Measures

In this section we describe the methodology for estimating S80/S20 and G for small areas (domains) of interest. Our approach is based on the use of M-quantile models for small area estimation (cf. Chambers and Tzavidis, 2006) and their extension for estimating deprivation indicators (cf. Marchetti et al., 2012a, Tzavidis et al., 2010a).

Let $\mathbf{x}_{jd}$ be a vector of p auxiliary variables that is known for each population unit j in small area d and let the income variable of interest, y_{jd}, be available from a random sample, s, that includes units from all target domains. Population size, sample size, sampled part of the population and non-sampled part of the population in a given domain are denoted respectively by N_d ,n_d ,s_d and r_d. As usual in small area estimation framework we further assume that conditional on the covariates available, the sampling design is ignorable.

The data needed to use the method we propose are: i. survey data on an income variable and explanatory variables, ii. Census/administrative data on the same set of explanatory variables. Some methods further assume that the Census and survey data are linked. However, this assumption is fairly unrealistic as in most cases the link between the survey and the Census data is unknown. Having said this, the estimation methods can be modified so that the linkage assumption is not necessary.

Usually, model-base methods which work with income variable as outcome variable often use a box-cox transformation (WB and EBP). In contrast, the M-quantile approach uses the raw values of the income variable. However, before proceeding to small area estimation, it is always advisable to use model diagnostics. Depending on the results of the model diagnostics, all methods can be implemented either by using the raw or the transformed values of the income variable.

The classical regression model summarizes the behavior of the mean of a random variable at each point in a set of covariates. Instead, quantile regression summarizes the behavior of different parts of the conditional distribution $f(y|\mathbf{x})$ at each point in the set of the $\mathbf{x}$'s. Because of notational simplicity the area-specific subscript d is dropped for the moment. Let $(\mathbf{x}_j\ ,y_j\ )$, $j = 1, \ldots, n$ denotes the observed values of a random sample consisting of n units, where $\mathbf{x}_j$ are p-vectors of a known design matrix $\mathbf{x}$ and y_j is a scalar response variable corresponding to a realization of a continuous random variable with unknown continuous cumulative distribution function. A linear regression model for the q conditional quantile of $f(y|\mathbf{x})$ is

$$Q_y(q|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}_q.$$

Estimates of the quantile regression parameter $\boldsymbol{\beta}_q$ are obtained by minimising

$$\sum_{j=1}^{n} |y_j - \mathbf{x}_j^T \boldsymbol{\beta}_q| \{(1-q)\mathbb{I}(y_j - \mathbf{x}_j^T \boldsymbol{\beta}_q \leq 0) + q\mathbb{I}(y_j - \mathbf{x}_j^T \boldsymbol{\beta}_q > 0)\}\,.$$

M-quantile regression (cf. Breckling and Chambers, 1988) is a “quantile-like” generalization of regression based on influence functions (M-regression). The M-quantile q of the conditional density $f(y|\mathbf{x})$, m, is defined as the solution to the estimating equation

$$\int \psi_q(y - m) f(y|\mathbf{x})\, dy = 0,$$

where ψ_q denotes an asymmetric influence function, which is the derivative of an asymmetric loss function ρ_q. A linear M-quantile regression model is defined by

$$m_y(q|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}_{\psi,q}.$$

Estimates of $\boldsymbol{\beta}_{\psi,q}$ are obtained by minimising

$$\sum_{j=1}^{n} \rho_q(y_j - \mathbf{x}_j^T \boldsymbol{\beta}_{\psi,q}). \tag{2.3}$$

The choice for ρ_q is the Huber loss function (cf. Breckling and Chambers, 1988). The estimating equation defined by (2.3) is

$$\sum_{j=1}^{n} 2\psi(r_{jq})\{(1-q)\mathbb{I}(r_{jq} \leq 0) + q\mathbb{I}(r_{jq} > 0)\} = 0,$$

where $r_{jq} = s^{-1}(y_j - \mathbf{x}_j^T \boldsymbol{\beta}_{\psi,q})$, and s is an estimate of scale such as the Mean Absolute Deviation. Provided that the tuning constant of the influence function is strictly greater than zero, estimates of $\boldsymbol{\beta}_{\psi,q}$ are obtained by using iterative weighted least squares (IWLS).

Using the M-quantile (or even the quantile) regression it is possible to characterise the conditional variability across the population of interest by the M-quantile coefficients of the population units (cf. CHAMBERS and TZAVIDIS, 2006). For unit j with values y_j and $\mathbf{x}_j$, this coefficient is the value θ_j such that $m_y(\theta_j|\mathbf{x}_j) = y_j$. If a hierarchical structure explains part of the variability in the population data then units within domains should have similar M-quantile coefficients. An area specific semi-parametric (empirical) pseudo-random effect, θ_d, is then computed by taking the expected value of the M-quantile coefficients θ_j in area d.

After we briefly described the M-quantile small area model, we now focus on the estimation of the Laeken indicators of interest at small area level.

Let us define the S80/S20 for area d by

$$S80/S20_d = \frac{\sum\limits_{j \in s_d} y_j \mathbb{I}(y_j > q_{d,0.8}) + \sum\limits_{k \in r_d} y_k \mathbb{I}(y_k > q_{d,0.8})}{\sum\limits_{j \in s_d} y_j \mathbb{I}(y_j < q_{d,0.2}) + \sum\limits_{k \in r_d} y_k \mathbb{I}(y_k < q_{d,0.2})}. \tag{2.4}$$

As an alternatives, one can use income quantiles defined at an aggregate level instead of the area-specific income quantiles. This leads to an alternative definition of S80/S20. In (2.4) the realised household income values in the sample is denoted by y_j, $j \in s_d$ and the unobserved out of sample income values are denoted by y_k, $k \in r_d$. The y_ks should be predicted to estimate the area-specific S80/S20. Moreover, since linked Census and survey data hardly ever exist, we further replace the sample y values also by their prediction under the model. This leads to the following definition of S80/S20 at the population level

$$S80/S20_d = \frac{\sum\limits_{j \in U_d} E(y_j \mathbb{I}(y_j > q_{d,0.8}))}{\sum\limits_{j \in U_d} E(y_j \mathbb{I}(y_j < q_{d,0.2}))}, \tag{2.5}$$

where U_d is the set of population units in area d.

The estimates of $E(y_j \mathbb{I}(y_j > q_d))$ can be obtained non-parametrically using a smearing-type estimator motivated by the work of Duan (1983) leading to

$$E(y_j \mathbb{I}(y_j > q_d)) = \int (\mathbf{x}_j^T \boldsymbol{\beta}_{\psi,\theta_d} + \epsilon) \mathbb{I}((\mathbf{x}_j^T \boldsymbol{\beta}_{\psi,\theta_d} + \epsilon) > q_d) \, dF(\epsilon). \tag{2.6}$$

We estimate the unknown error term distribution $F(\epsilon)$ by using the empirical distribution function of the estimated model residuals. Further, substituting $\boldsymbol{\beta}_{\psi,\theta_d}$ by their estimates under the M-quantile small area model $\hat{\boldsymbol{\beta}}_{\psi,\hat{\theta}_d}$ leads to

$$\hat{E}(y_j \mathbb{I}(y_j > q_d)) = \int (\mathbf{x}_j^T \hat{\boldsymbol{\beta}}_{\psi,\hat{\theta}_d} + \hat{\epsilon}) \mathbb{I}((\mathbf{x}_j^T \hat{\boldsymbol{\beta}}_{\psi,\hat{\theta}_d} + \hat{\epsilon}) > q_d) \, d\hat{F}(\epsilon). \tag{2.7}$$

The evaluation of $\hat{E}(y_j \mathbb{I}(y_j > q_d))$ is achieved by means of Monte-Carlo simulation. The technique is as follows

1. estimate the M-quantile model and compute the residuals
2. draw a with replacement sample of size N_d from the empirical distribution of the residuals
3. microsimulate a population of synthetic income values for each small area
4. from micro simulated population draw a random sample and estimate $q_{0.8,d}$, $q_{0.2,d}$ and $\hat{E}(y_j \mathbb{I}(y_j > q_d))$
5. repeat step 2 to 4 H times, each time estimating $\widehat{\text{S80/S20}}_d^h$, $h = 1, \ldots, H$
6. the S80/S20 estimate in area d is obtained by taking the average of the S80/S20 values over the Monte-Carlo replications

A similar approach to the one described above is used for estimating the area-specific Gini coefficient G_d. Let us define G in area d by

$$\text{G}_d = \frac{N_d + 1}{N_d} - \frac{-2 \sum\limits_{j \in U_d} (N_d + 1 - (j)) y_j}{N_d \sum\limits_{j \in U_d} y_j}. \tag{2.8}$$

Similar to the case of S80/S20, y_j in (2.8) is replaced by its expectation under the model

$$\hat{E}(y_j) = \int (\mathbf{x}_j^T \hat{\boldsymbol{\beta}}_{\psi,\hat{\theta}_d} + \hat{\epsilon}) \, d\hat{F}(\epsilon).$$

The expectation is evaluated also by using Monte-Carlo simulation.

2.4 Mean Squared Error Estimation

Mean Squared Error (MSE) estimation for small area estimators under the M-quantile model is discussed in detail in Marchetti et al. (2012a) and is based on the use of a non-parametric bootstrap scheme. Here we recall the main steps of this bootstrap scheme. Starting from sample s, selected from a finite population U without replacement, we fit the M-quantile small area model and obtain estimates $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}_{\psi,\hat{\theta}_d}$ which are used to compute the model residuals. We then generate B bootstrap populations, U^{*b}. From each bootstrap population we select L bootstrap samples using simple random sampling within the small areas and without replacement such that $n_d^* = n_d$. Applying the methodology described in 2.2 to the bootstrap samples we obtain estimates of the inequality indicators of interest.

To generate the bootstrap populations there are two alternatives: i. sampling from the empirical distribution of the residuals or ii. sampling from a smoothed version of this distribution. For each of these alternatives the sampling can be conditional or unconditional on the small areas. So we have a total of four possibilities to generate the bootstrap population. Denoting by $\hat{\tau}_d$ the estimated small area parameter, bootstrap estimators of the bias and variance are defined respectively by

$$\hat{B}(\hat{\tau}_d) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}(\hat{\tau}_d^{*bl} - \tau_d^{*b}),$$

$$\hat{V}(\hat{\tau}_d) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}(\hat{\tau}_d^{*bl} - \hat{\bar{\tau}}_d^{*bl})^2.$$

In the expressions for the bias and the variance τ_d^{*b} is the small area parameter of the b bootstrap population, $\hat{\tau}_d^{*bl}$ is the small area parameter estimated by using the l sample from the b bootstrap population and $\hat{\bar{\tau}}_d^{*bl} = L^{-1}\sum_{l=1}^{L}\hat{\tau}_d^{*bl}$. The bootstrap MSE estimator of the estimated small area target parameter is then defined as

$$\widehat{MSE}(\hat{\tau}_d) = \hat{V}(\hat{\tau}_d) + \hat{B}(\hat{\tau}_d)^2.$$

2.5 Empirical Evaluations

In order to compare the performances of the various small-area estimators defined in the preceding sections we used a model-based simulation study. Small-area population and sample data were simulated based on a two-level linear mixed model with different parametric assumptions for the area- and unit-level random effects. As usual in model-based simulation the sample design is a stratified random sampling, where the strata correspond to the small areas and the allocation is proportional to the population size in the small area.

We generated data for $D = 30$ small areas which are partitions of a population of size $N = 6000$ with $100 \leq N_d \leq 300$, $d = 1, \ldots, D$. For each area, we selected a simple random sample (without replacement) of size $5 \leq n_d \leq 8$, leading to an overall sample size of $n = 175$.

We used a random intercept model $y_{jd} = 5000 + \beta x_{jd} + u_d + e_{jd}$ to generate target values in the population. The x_{jd}s are generated as independently and identically distributed realizations from a mixture model $(1 - \gamma)N(\mu_d, 1) + \gamma N(\eta_d, 1)$ where the weight γ was set to 0.2 and held fixed over the simulations. The small-area x-means μ_d and η_d were themselves drawn at random from the uniform distribution on the interval $[1, 10]$ and $[81, 90]$ respectively, and held fixed over the simulations.

A total of 1000 populations and related samples were generated and used to estimate the small-area quintile share ratio ($S80/S20$) and the Gini coefficient (G).

We generate the area and unit level error according to two different distributions, in this way generating two simulation scenarios. In the first simulation experiment (Scenario 1) the random effects u_d and e_{dj} were independently and identically generated as $N(0,\ 40000)$ and $N(0,\ 640000)$ realizations respectively. In this scenario β was set equal to 250. In the second simulation experiment (Scenario 2) the random effects u_d and e_{jd} were independently and identically generated as mean-corrected Singh-Maddala distribution realizations, both with parameters ($a = 2.8, b = 100^{-5/14}, q = 1.7$). The Singh-Maddala distribution has density function equal to

$$f(y) = \frac{aqy^{a-1}}{b^a \left(1 + \frac{y^a}{b^a}\right)^{1+q}} \,.$$

The mean correction was set as $15000u_d - D^{-1}\sum_{d=1}^{D} 15000u_d$ and $15000\ e_{jd} - D^{-1}n_d^{-1}\sum_{d=1}^{D}\sum_{j=1}^{n_j} 15000\ e_{jd}$. The purpose of Scenario 2 was to examine the effect of misspecification of the Gaussian assumptions of a mixed model using a density function able to mimic an income distribution.

The reason why we used a mixture distribution for generating auxiliary variables is to ensure realistic values of the target parameters in the simulated population. Indeed, in Scenario 2 the Gini coefficients vary between 0.12 and 0.5 and the S80/S20 between 1.9 and 11.8 over areas and simulations. In Scenario 1 the Gini coefficient and the S80/S20 range respectively between 0.29 and 0.39, and 3.9 and 6.2. These values are considered realistic in the sense that in the Euro Area in 2012 Eurostat reported a Gini Coefficient between 0.23 (Norway) and 0.36 (Latvia) and a Quintile Share Ratio between 3.5 (Czech Republic) and 7.2 (Spain) (`http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/data/main_tables`). Similar results are obtained in others years.

Using the simulated data, point estimation and MSE estimation is performed using the methodology described in sections 2.3 and 2.4. All computations were performed by using R (cf. R Core Team, 2015). Biases and root MSEs over these simulations, summarized over the 30 areas, are shown in Table 2.1, Table 2.2 and in Table 2.3. Moreover, in Table 2.4 and Table 2.5 we also show the bias of the root MSE estimator we presented in section 2.4.

The bias for both estimators in both scenarios is negative. In the first scenario - results in Table 2.1 - it is fairly small with the maximum absolute relative bias being 3.7% for the Gini coefficient and 7.3% for the quintile share ratio. Using a heavy tailed error distribution - scenario 2 - results in an increase, albeit small compared to Scenario 1, in the bias for both estimators - results in table 2.2. In Scenario 2 the mean of the relative bias for the Gini coefficient is equal to -3% and the mean of the absolute relative bias is equal to 5.6%. For the quintile share ratio the mean of the relative bias and the absolute relative bias are equal to -3.5% and 8% respectively. Also considering the minimum and maximum values of the relative bias and the absolute relative bias, the results obtained are still acceptable in both scenarios.

Table 2.3 shows the empirical MSE of the Gini coefficient and quintile share ratio estimators for both scenarios summarized over areas. The empirical MSE for an estimator $\hat{\theta}_d$ of

Table 2.1: Scenario 1 (Normally distributed errors). Distribution over areas and Monte Carlo simulations of the bias, absolute bias, relative bias and absolute relative bias of the Gini coefficient (G) and the quintile share ratio (S80/S20).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Bias($\widehat{\text{G}}$)	−0.005	−0.004	−0.004	−0.003	−0.003	−0.002
Abs. Bias($\widehat{\text{G}}$)	0.009	0.010	0.011	0.011	0.012	0.013
Rel. Bias($\widehat{\text{G}}$) %	−1.340	−1.130	−1.050	−1.000	−0.890	−0.630
Rel. Abs. Bias($\widehat{\text{G}}$) %	2.770	3.080	3.410	3.310	3.580	3.690
Bias($\widehat{S80/S20}$)	−0.133	−0.104	−0.090	−0.091	−0.079	−0.058
Abs. Bias($\widehat{S80/S20}$)	0.228	0.279	0.312	0.313	0.350	0.391
Rel. Bias($\widehat{S80/S20}$) %	−2.430	−1.980	−1.800	−1.770	−1.510	−1.140
Rel. Abs. Bias($\widehat{S80/S20}$) %	5.120	5.540	6.240	6.170	6.700	7.320

θ_d is computed as

$$MSE(\hat{\theta}_d) = N^{-1}\sum_{h=1}^{H}(\hat{\theta}_{dh} - \theta_{dh})^2,$$

where θ_{dh} and $\hat{\theta}_{dh}$ are respectively the true and the estimated values of the target statistics (G and S80/S20) for area d in the iteration h of the Monte Carlo simulation, and H is the total number of Monte Carlo runs (1000). The true values are computed from the corresponding Monte-Carlo population in each small area. The empirical MSEs are treated as true MSEs of the proposed estimators and are used in Tables 2.4 and 2.5 to compute the bias, absolute bias and relative bias of the root MSE bootstrap estimator (*rmse*) we proposed in section 2.4.

The *rmse* bootstrap estimator performs well in both the scenarios, Tables 2.4 and 2.5. The known tendency to underestimate the true root MSE observed by Marchetti et al. (2012a) is also confirmed in this simulation. Some improvements of the used bootstrap technique are currently under study. Moreover, given that there are no remarkable differences in the behavior of the *rmse* estimators under both scenarios, we can assume that the non-parametric *rmse* bootstrap estimator is robust to the assumed unit- and area-level error distributions. The *rmse*s estimates have been obtained sampling from the empirical distribution of the residuals conditionally to the small areas. Furthermore, we can note

Table 2.2: Scenario 2 (Singh-Maddala distributed errors). Distribution over areas and Monte Carlo simulations of the bias, absolute bias, relative bias and absolute relative bias of the Gini coefficient (G) and the quintile share ratio (S80/S20).

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Bias($\widehat{\text{G}}$)	−0.012	−0.010	−0.009	−0.009	−0.008	−0.006
Abs. Bias($\widehat{\text{G}}$)	0.014	0.016	0.017	0.017	0.018	0.020
Rel. Bias($\widehat{\text{G}}$) %	−3.800	−3.400	−3.070	−2.990	−2.660	−1.970
Rel. Abs. Bias($\widehat{\text{G}}$) %	4.690	5.230	5.840	5.640	6.070	6.360
Bias($\widehat{S80/S20}$)	−0.247	−0.188	−0.162	−0.163	−0.137	−0.096
Abs. Bias($\widehat{S80/S20}$)	0.281	0.338	0.389	0.380	0.417	0.484
Rel. Bias($\widehat{S80/S20}$) %	−4.920	−4.020	−3.560	−3.520	−2.950	−2.120
Rel. Abs. Bias($\widehat{S80/S20}$) %	6.540	7.320	8.040	8.040	8.580	9.620

that the true root MSEs are in general small values hence a small difference between the estimated value and true values can lead to a large relative bias. Examining the absolute bias the error of the *rmse* estimator for the Gini coefficient is maximum 0.007 and it is maximum 0.366 for the *rmse* estimator of the Quintile Share Ratio.

2.6 Estimating G and S80/S20 for unplanned domains (LAU 1) in Tuscany

The aim of the application presented here is to estimate G and S80/S20 for unplanned domains in the region of Tuscany in Italy. We used data from EU-SILC 2008 and from the Italian Population Census 2001. The domains of interest are defined by the cross-classification of provinces in Tuscany (LAU 1) by the gender of the head of the household. The domains are 20 in total (10 provinces × 2 gender categories).

The EU-SILC 2008 surveys the household equivalised income which is our outcome variables. For each individual the equivalised total net income is calculated as its household total net income divided by the equivalised household size according to the modified OECD scale, where the head of the household has weight equal to 1, other household members aged 14 or more have a weight of 0.5 and members aged 13 or lesser have a weight of 0.3. The set of explanatory variable must be selected within the set of common variables in EU-

Table 2.3: Distribution over areas of the empirical root MSE (RMSE) of the Gini coefficient (G) and the quintile share ratio (S80/S20).

	Scenario 1					
	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
$RMSE(\widehat{\text{G}})$	0.012	0.013	0.014	0.014	0.015	0.016
$RMSE(\widehat{S80/S20})$	0.288	0.347	0.388	0.393	0.438	0.492
	Scenario 2					
	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
$RMSE(\widehat{\text{G}})$	0.018	0.021	0.022	0.022	0.023	0.026
$RMSE(\widehat{S80/S20})$	0.386	0.460	0.532	0.530	0.593	0.676

SILC and the Italian Population Census. The explanatory variables we selected are the marital status of the head of the household (four categories, single, married, divorced and widow), the employment status of the head of the household (working/not working), the years of education of the head of the household, the mean house surface (in square meters) at municipality level (LAU 2 level) and the number of household members. It is important to underline that EU-SILC and Census datasets are confidential. The datasets were provided by ISTAT, the Italian National Institute of Statistics, to the researchers of the SAMPLE (2009) project and were analyzed by respecting the confidentiality restrictions.

We carried out some exploratory analysis on the sample data, resumed in Figure 2.1 and 2.2. In particular the Figure 2.1 shows box-plots of the household equivalised income in the 20 domains. The box-plots highlight the asymmetry of the income distribution in each domains. Furthermore, we observe differences in the income distribution between households whose head is female and those whose head is male. The only exceptions are the provinces of Massa-Carrara (MC) and Prato (PO) where this difference is less evident, but present. Income quantiles in the household leads by a female tend to be lower.

The use of the M-quantile approach instead of the linear-model-based approach is motivated when the normal hypothesis of the linear model does not hold. A graphical analysis of level one and two residuals obtained by fitting a two-level random effects model to the EU-SILC data is shown in Figure 2.2. Here, households are the level one units and the 20 domains define the level two units. The plots in Figure 2.2 suggest departures from the normality assumptions of the random effects model. Also the use of the Shapiro test

Table 2.4: Scenario 1 (Normally distributed errors). Distribution over areas and Monte Carlo simulations of the bias, absolute bias and relative bias of the bootstrap root MSE estimator of the Gini coefficient (G) and the quintile share ratio (S80/S20).

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Bias $rmse(\widehat{\mathrm{G}})$	-0.003	-0.002	-0.001	-0.001	0.000	0.001
Abs bias $rmse(\widehat{\mathrm{G}})$	0.001	0.001	0.001	0.002	0.002	0.003
Rel bias $rmse(\widehat{\mathrm{G}})$ %	-17.070	-11.620	-5.860	-6.000	-2.270	12.680
Bias $rmse(\widehat{S80/S20})$	-0.108	-0.046	-0.033	-0.032	-0.011	0.047
Abs bias $rmse(\widehat{S80/S20})$	0.043	0.058	0.068	0.071	0.077	0.125
Rel bias $rmse(\widehat{S80/S20})$ %	-26.410	-12.160	-8.040	-7.290	-2.710	19.260

statistic confirms that the hypothesis of normally distributed level one residuals, both when using the raw and log-transformed income variable, is rejected.

The results of the application are shown in Table 2.6 which reports model-based estimates of the Gini coefficient (G) and of the quintile share ratio (S80/S20). Results are obtained by applying the M-quantile-based small area methodology presented in section 2.2. Estimated root mean squared errors are reported in parentheses. Table 2.6 also reports the sample and population sizes in each domain.

Looking at the results in Table 2.6 we observe that the inequality in those domains where the head of the household is female is in the majority of cases higher than in those domains where the head of the household is male. Moreover, the coefficient G provides some evidence of higher inequality in the provinces of Grosseto (GR), Pistoia (PT) and Florence (FI) in the domains where a female as the head of the household. Furthermore, it is important to take into account the uncertainty of the estimates to make detailed comparisons. According to estimates by Eurostat (`http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&language=en&pcode=tessi190`) the Gini coefficient in Italy in 2008 was 31%, a number consistent with the estimates we present in Table 2.6.

With respect to the quintile share ratio, the results in table 2.6 indicate again that inequality in those domains where the head of the household is a female is very often higher than in those domains where the head of the household is a male. The estimates indicate higher

Table 2.5: Scenario 2 (Singh-Maddala distributed errors). Distribution over areas and Monte Carlo simulations of the bias, absolute bias and relative bias of the bootstrap root MSE estimator of the Gini coefficient (G) and the quintile share ratio (S80/S20).

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Bias $rmse(\widehat{\mathrm{G}})$	-0.006	-0.002	-0.001	-0.001	0.001	0.003
Abs bias $rmse(\widehat{\mathrm{G}})$	0.003	0.004	0.004	0.004	0.005	0.007
Rel bias $rmse(\widehat{\mathrm{G}})$ %	-20.210	-9.800	-5.790	-4.840	2.560	15.390
Bias $rmse(\widehat{S80/S20})$	-0.132	-0.049	-0.008	-0.009	0.04	0.119
Abs bias $rmse(\widehat{S80/S20})$	0.141	0.194	0.24	0.245	0.284	0.366
Rel bias $rmse(\widehat{S80/S20})$ %	-19.410	-9.940	-2.160	-0.440	7.860	28.240

inequality in the Province of Pistoia (PT), Livorno (LI) and Florence (FI) when the head of the household is a female. The quintile share ratio (S80/S20) in Italy in 2008 is estimated by Eurostat and is equal to 5.1 (`http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&language=en&pcode=tessi180&plugin=1`), which is consistent with the estimates in Table 2.6.

It is often recommended to represent the estimates on a map, when possible. Figure 2.3 and 2.4 are asymetric maps that offer a representation of the model-based estimates of the Gini coefficient and of the quintile share ratio. In both figures the map on the left refers to the provinces with a female as the head of the household, while the one on the right refers to households leads by a male. The provinces are grouped in four different classes of colors, where the darker the color the lower the Gini coefficient and the quintile share ratio. To facilitate the comparison between males and females we used a common scale for the classes of colors.

The inequality in the provinces of Tuscany, measured by the small area estimates of the Gini coefficient and quintile share ratio, is in line with national (reliable) estimates. The insight of this application is some evidence of differences in inequality between males and females heads of household. This result is sensible, given that in Italy most households where the head is a female are often cases where there is a widow living alone or a widow living with dependents. These results allow us to identify domains with potentially higher

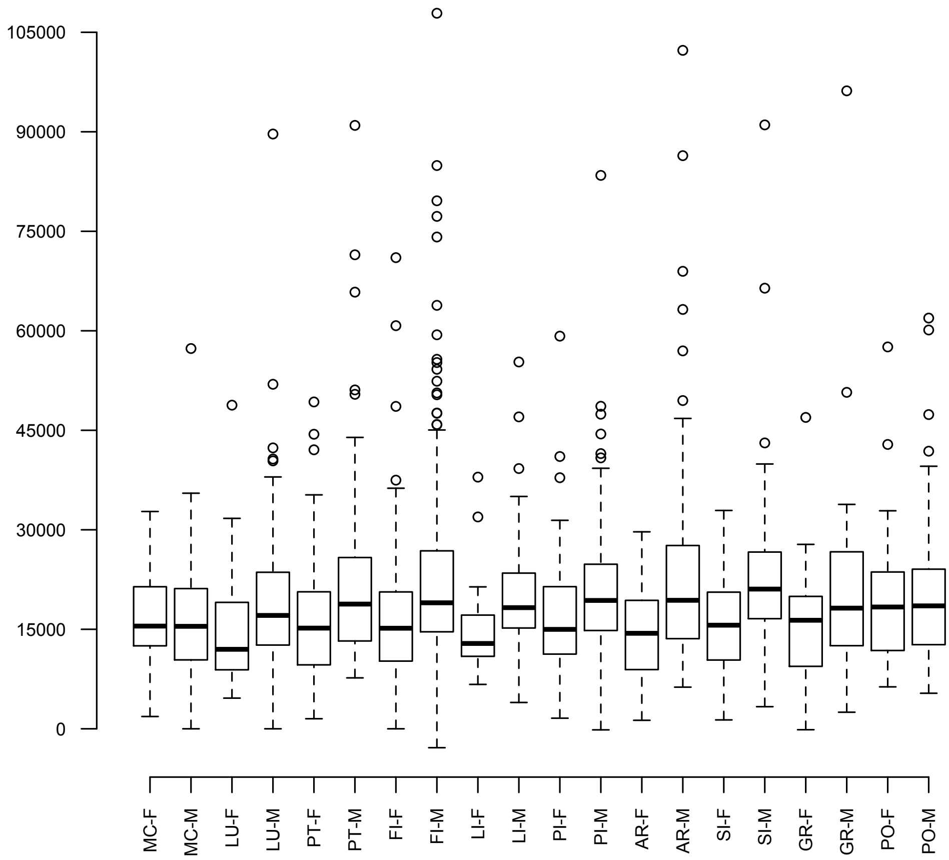


Figure 2.1: Boxplots of the equivalised household income for Tuscany Provinces by gender of the head of the household (F=Female, M=Male).

inequality, which alongside other poverty indicators can better depict the monetary-based deprivation with an high geographical resolutions. Last but not least, they can assist policy makers and stakeholders to plan and implement appropriate social policies at local level.

2.7 Conclusions

The increased demand for producing estimates of deprivation and inequality indicators at the level of unplanned domains stimulate a rapid development in model-based small area methodologies. However, the request of non-linear statistics, such as most of the Laeken indicators, together with the strong asymmetry and irregularity of income distributions is still a challenge for statisticians. In this work we presented a small area methodology that is based on the use of linear M-quantile regression model and is used for estimating inequality indicators such as the Gini coefficient and the quintile share ratio. One of the advantages of the M-quantile approach to small area estimation is that it avoids the use

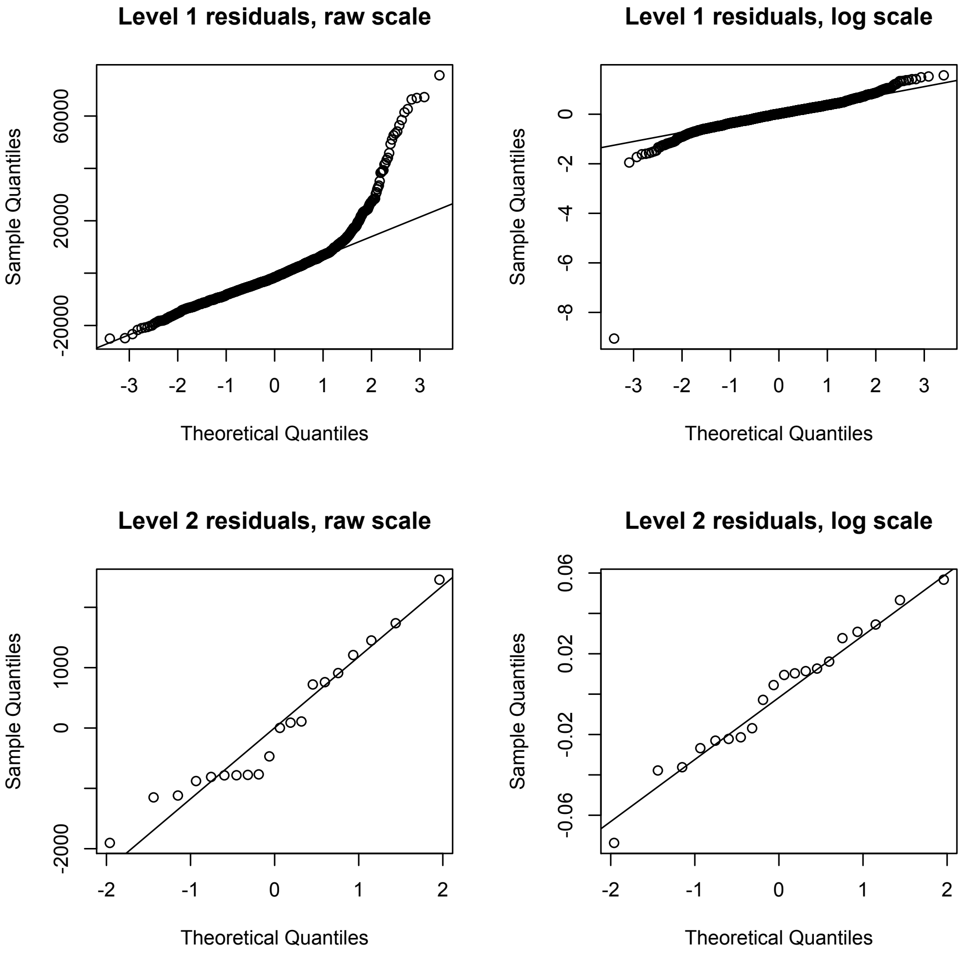


Figure 2.2: Normal probability plots of level one and level two residuals derived by fitting a two level linear mixed model to the EU-SILC data using the original income variable (left) and a log-transformed income variable (right).

of strong parametric assumptions in estimation.

Point and MSE estimation is facilitated by the availability of open source software that has been written in the statistical language *R* (cf. R CORE TEAM, 2015) and can be easily adapted for estimating other Laeken indicators.

Despite the availability of a wide range of small area methodologies for estimating income-based deprivation and inequality, there are practical problems that require the development of new methodology.

Finally, it is known that the use of only monetary-based indicators, such as the Laeken indicators, can not fulfill a real picture of deprivation and social exclusion. To fill this gap multidimensional indicators that incorporate additional dimensions such as educational, health and social security inequality have been developed. Also other approaches to mea-

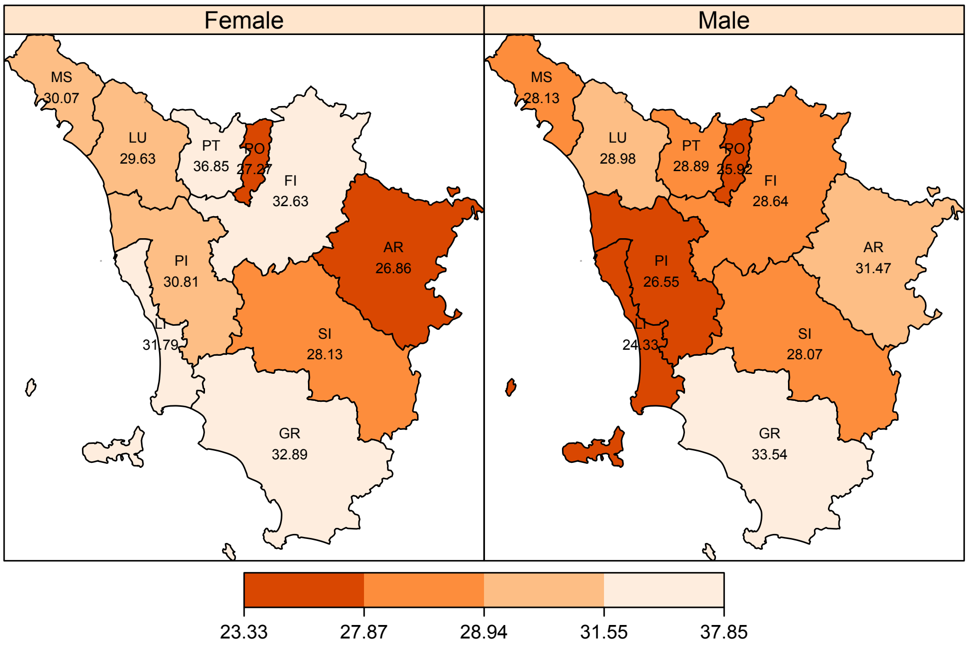


Figure 2.3: Model-based estimates of Gini coefficient for Provinces in Tuscany by gender of the head of the household: female (left) and male (right).

sure deprivation, like the capability approach, are under study. Research for developing small area methodologies appropriate to tackle this problem is currently under way.

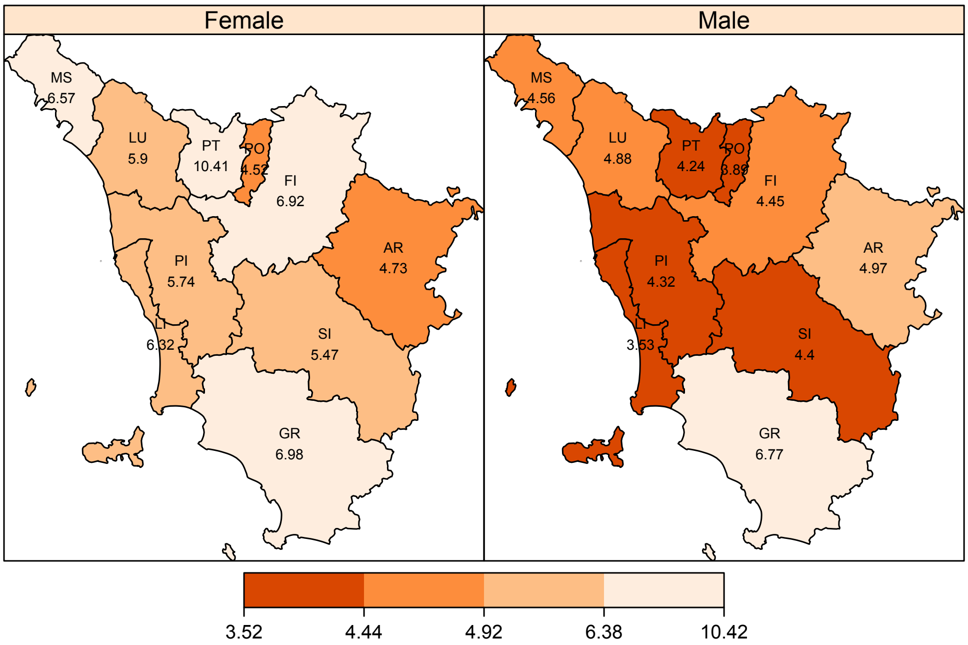


Figure 2.4: Estimates of quintile share ratio (S80S20) for Tuscany Provinces, by gender of the head of the household: female (left) and male (right).

Table 2.6: Model-based (M-quantile) estimates of the Gini coefficient (G) and the quintile share ratio (S80/S20) with corresponding estimated Root Mean Squared Errors (in parantheses). The population and sample size of households in each domain (n and N) are also reported.

Province	**HH Gender**	**n**	**N**	$\widehat{\mathbf{G}}$(%)	$\widehat{\mathbf{S80/S20}}$
MC	Female	34	24608	30.07 (3.66)	6.57 (3.96)
	Male	71	56202	28.13 (2.07)	4.56 (0.50)
LU	Female	38	41622	29.63 (3.68)	5.90 (2.14)
	Male	112	104495	28.98 (2.84)	4.88 (0.93)
PT	Female	51	27684	36.85 (3.96)	10.41 (6.35)
	Male	85	76782	28.89 (2.93)	4.24 (0.86)
FI	Female	140	110484	32.63 (2.36)	6.92 (1.45)
	Male	275	265771	28.64 (1.70)	4.45 (0.60)
LI	Female	31	37646	31.79 (3.49)	6.32 (2.28)
	Male	74	96083	24.33 (2.00)	3.53 (0.44)
PI	Female	44	37673	30.81 (3.89)	5.74 (1.90)
	Male	105	112586	26.55 (2.53)	4.32 (0.87)
AR	Female	34	30589	26.86 (2.15)	4.73 (0.76)
	Male	109	93291	31.47 (2.79)	4.97 (1.11)
SI	Female	29	25699	28.13 (3.04)	5.47 (1.54)
	Male	75	75700	28.07 (3.34)	4.40 (1.08)
GR	Female	30	24531	32.89 (3.62)	6.98 (2.41)
	Male	35	63189	33.54 (6.32)	6.77 (8.57)
PO	Female	37	19130	27.27 (3.40)	4.52 (1.24)
	Male	86	64487	25.92 (2.47)	3.89 (0.67)

3 An Application of Empirical Best Prediction to Poverty Estimation in Spain

Nikos Tzavidis

Social Statistics & Demography and Southampton Statistical Sciences Research Institute
University of Southampton

Iñaki Permanyer **Pinar Koksel**

CED
University Autonoma Barcelona

Summary

This part of the deliverable describes the cooperation between different pillars of the INGRID consortium, namely the statistical pillar represented in this report by Nikos Tzavidis (University of Southampton) and the poverty pillar represented by Iñaki Permanyer (CED, Autonoma University of Barcelona). The aim of this cooperation was to support researchers from the poverty pillar with the implementation of statistical methodology for estimating poverty in disaggregated geographic areas. Support involved organising two meetings, one for explaining the methodology and for providing software for implementing the methodology and a second meeting for testing the implementation of the code and for examining the results. For doing so, Nikos Tzavidis visited CED in November 2014 and Iñaki Permanyer visited Southampton in December 2014.

3.1 Introduction

Estimating economic indicators is crucial for achieving a targeted implementation of welfare policies. However, for such policies to be effective policy makers must have access to a detailed picture of deprivation that goes beyond aggregate estimates at the country (national) level, extending to finer geographical levels and to other domains of interest for example, specific groups of individuals. Such a picture can only be constructed by having access to survey and administrative/Census data at appropriate spatial scales that are timely and accurate. One possible solution for obtaining accurate indicators at finer spatial scales is by using small area estimation methodologies. The term ”small areas” is typically used to describe domains (e.g. geographic areas) whose sample sizes are not large enough to allow sufficiently precise direct estimation i.e. estimation that is based only on the sample data from a domain (cf. Rao, 2003). Small area-specific sample sizes often also hamper the use of conventional design-based estimators. In such cases model-based estimation procedures can be considered for improving the precision of the direct estimates. Small area estimation is conventionally concerned with the estimation of small area averages and totals. More recently some of the research effort has been shifted towards methods for estimating poverty (deprivation) indicators at the small area level, also known as poverty mapping. Poverty mapping can offer a detailed description of the spatial distribution of poverty and inequality within a country. It combines individual and household survey data with Census/administrative data with the objective of estimating welfare indicators for geographic areas or domains of interest. In recent years a range of alternative model based small area methodologies for poverty mapping have been proposed.

The seminal paper by Elbers et al. (2003) proposed a methodology for estimating poverty indicators at the small area level. The methodology consists of a nested error regression (random effects) model with cluster random effects that is estimated by using survey data. The response variable, which is not available in the Census, is the logarithm of a welfare variable, e.g. income or consumption, and the explanatory variables, used for modeling the welfare variable, are available both in the survey and in the Census datasets. Once the model has been estimated using the survey data, the estimated model parameters are combined with Census micro-data to form unit level synthetic Census predictions of the welfare variable. The synthetic values of income/consumption alongside a defined poverty line are then used for estimating deprivation indicators for example, the incidence of poverty (HCR), the poverty gap (PG) and the poverty severity (cf. Foster et al., 1984). Routinely we refer to this poverty mapping methodology as the World Bank (WB) approach.

More recently, Molina and Rao (2010) proposed an Empirical Best Prediction (EBP) approach for estimating poverty indicators at the small area level, which is similar to the WB approach but generates Census predictions of income/consumption by using the conditional predictive distribution of the out of sample data given the sample data. Molina and Rao (2010) demonstrated the superior performance of the EBP approach, when compared to the WB approach, under the nested error regression model.

The aim of this report is to review some of the recently proposed small area methodologies for poverty estimation that mainly use the nested error regression model and present a case study by applying the EBP approach using data from the EU-SILC survey in Spain. The report is organised as follows. In Section 2 we present the WB and EBP approaches to small area poverty estimation. In Section 3 we review robust methodologies and in Section 4 we apply the EBP methodology using data from the survey of income and living conditions (EU-SILC) in 2004 and 2011 from Spain and a sample of Census microdata from the Integrated Public Use Microdata Series (IPUMS) in 2001 and 2011. The target parameters include the Head Count Ratio (HCR), the Poverty Gap (PG) and the Gini coefficient (Gini). Finally, in Section 5 we provide a summary and outline current challenges.

3.2 Small Area estimation of poverty indicators using the nested error regression model

In what follows we assume that a vector of p auxiliary variables $\mathbf{x}_{ij}$ is known for each population unit i in small area $j = 1, \ldots, m$ and that values of the welfare variable of interest y_{ij} are available from a random sample, s, that includes units from all the small areas of interest. We denote the population size, sample size, sampled part of the population and non-sampled part of the population in area j respectively by N_j, n_j, s_j, r_j. We assume that the sum over the areas of N_j and n_j is equal to N and n respectively. We further assume that conditional on covariate information, for example design variables, the sampling design is ignorable.

All poverty mapping methods we describe in this deliverable assume the availability of survey data on a welfare variable (income/consumption) and explanatory variables that can be used for modelling the outcome variable. In addition, the methods assume the availability of Census/administrative data on the same set of explanatory variables. Some methods further assume that the Census and survey data are linked. However, this assumption is fairly unrealistic as in most cases the link between the survey and the Census data is not known. Finally, the methods that are based on the nested error regression model (WB and EBP) conventionally use a logarithmic transformation of the welfare variable. Nevertheless, before proceeding to small area estimation, it is always advisable to use model diagnostics because a logarithmic transformation may not be the optimal one. In this report we focus on the estimation of HCR and PG as defined by Foster et al. (1984) and on the estimation of the Gini coefficient. Denoting by t the poverty line, different poverty indicators are defined by the area-specific mean of the derived variable

$$z_{ij}\left(\alpha,t\right)=\left(\frac{t-y_{ij}}{t}\right)^{\alpha}\mathbb{I}\left(y_{ij}\leq t\right)\quad,\; i=1,\ldots,N_j\quad.$$

Setting $\alpha = 0$ defines the HCR in small area j, $z_j\left(0,t\right)$, which is the mean of $z_{ij}\left(0,t\right)$. Similarly, setting $\alpha = 1$ defines the PG in small area j, $z_j\left(1,t\right)$, which is the mean of $z_{ij}\left(1,t\right)$.

3.2.1 The WB method

The most widely used method for small area poverty mapping is the so-called WB method (cf. Elbers et al., 2003). In its simplest form and assuming a non-informative sampling design, the WB method assumes a nested error regression model on the logarithmically transformed values of y_{ij},

$$y_{ij}=\mathbf{x}_{ij}\boldsymbol{\beta}^{T}+u_j+\varepsilon_{ij},\quad u_j\sim N\left(0,\ \sigma_u^2\right)\ ,\quad \varepsilon_{ij}\sim N\left(0,\ \sigma_{\varepsilon}^2\right). \tag{3.1}$$

The WB method starts by estimating equation (3.1) using the sample data. Once estimates of the fixed effects, $\widehat{\beta}$, of the variance components, $\widehat{\sigma}_u^2\ \widehat{\sigma}_{\varepsilon}^2$, and of the area random effects $\widehat{u}_j$ have been obtained, the WB method uses the following bootstrap population model for generating L synthetic Censuses,

$$y_{ij}^{*}=\mathbf{x}_{ij}\widehat{\boldsymbol{\beta}}^{\mathbf{T}}+\mathbf{u}_{\mathbf{j}}^{*}+\varepsilon_{ij}^{*},\quad u_j^{*}\sim N\left(0,\ \widehat{\sigma}_u^2\right),\quad \varepsilon_{ij}^{*}\sim N\left(0,\ \widehat{\sigma}_{\varepsilon}^2\right). \tag{3.2}$$

The exact steps of the Monte-Carlo simulation are as follows. Start by estimating equation (3.1) using the sample data; draw L population vectors of y_{ij}^{*} using equation (3.2); Using the synthetic values of the welfare variable, y_{ij}^{*}, compute the WB estimate from the l^{th} synthetic Census, $\widehat{z}_j^{*WB\,(l)}\left(\alpha,t\right)$; average the results over L Monte Carlo simulations.

Using the bootstrap population model equation (3.2), one can further compute the MSE of the estimated poverty indicators,

$$MSE\left[\widehat{z}_j^{WB}\left(\alpha,t\right)\right]=L^{-1}\sum_{l=1}^{L}\left[\widehat{z}_j^{*WB(l)}\left(\alpha,t\right)-E\left(\widehat{z}_j^{*WB}\left(\alpha,t\right)\right)\right]^2.$$

One distinct aspect of the WB method is that the random effect is specified at the cluster (e.g. primary sampling units) level and not necessarily at the level of the target small area.

This is in contrast to the alternative methodologies we describe later on in this deliverable. To simplify things, for the purposes of this report we assume that clusters and target small areas coincide. As MOLINA and RAO (2010) pointed out, when small areas and clusters coincide and since $\left(u_j^*\right) = 0, E\left(\varepsilon_{ij}^*\right) = 0$, in the simplest case of estimating a small area mean, and denoting by U_j the population of units in domain j, the WB method leads to $E\left(y_{ij}^*\right) = N_j^{-1}\sum_{\mathbf{k}\in U_\mathbf{j}}\mathbf{x_{kj}}\widehat{\beta}^\mathbf{T}$, which is a regression synthetic estimator. It may be reasonable to assume that in many cases a regression synthetic estimator will be less efficient than competing indirect estimators.

3.2.2 The EBP method

The EBP method was proposed by MOLINA and RAO (2010). Like the WB approach, the EBP approach also relies on the use of a nested error regression model on the logarithmically transformed welfare variable. Let us start the description of the method by decomposing the population small area-specific poverty indicator as follows,

$$z_j\left(\alpha{,}t\right) = N_j^{-1}\left[\sum_{i\in s_j} z_i\left(\alpha{,}t\right) + \sum_{k\in r_j} z_k\left(\alpha{,}t\right)\right]. \tag{3.3}$$

The first component in equation (3.3) is observed in the sample whereas the second component is unknown and should be estimated by using a small area model, which is estimated with the sample data. Similarly to the WB method, the EBP method starts by estimating the nested error regression model to obtain estimates of the fixed effects, $\widehat{\beta}$, of the variance components, $\widehat{\sigma}_u^2\ \widehat{\sigma}_\varepsilon^2$, and of the area random effects $\widehat{u}_j$.

The EBP method then simulates out of sample data from the conditional distribution of the out of sample data given the sample data. This is done by using the following bootstrap population model for generating L synthetic Censuses:

$$y_{ij}^* = \mathbf{x}_{ij}\widehat{\boldsymbol{\beta}}^T + \widehat{u}_j + u_j^* + \varepsilon_{ij}^*, \quad u_j^* \sim N\left(0, \widehat{\sigma}_u^2(1-\gamma_j)\right), \quad \varepsilon_{ij}^* \sim N\left(0, \widehat{\sigma}_\varepsilon^2\right) \quad , \tag{3.4}$$

$$\gamma_j = \frac{\widehat{\sigma}_u^2}{(\widehat{\sigma}_u^2 + \widehat{\sigma}_\varepsilon^2/n_j)}.$$

The exact steps of the Monte-Carlo simulation are as follows. Start by estimating equation (3.1) using the sample data; draw L out of sample vectors of y_{ij}^* using equation (3.4); combine the sample y_{ij} with the out of sample y_{ij}^* values; compute the EBP estimate for the

l^{th} synthetic Census, $\widehat{z}_j^{EBP\ (l)}(\alpha,t)$; average the results over L Monte Carlo simulations. Focusing again on the simplest case of estimating a small area mean and since $E\left(u_j^*\right) = 0$, $E\left(\varepsilon_{ij}^*\right) = 0$, the EBP approach leads to $E\left(y_{ij}^*\right) = N_j^{-1}\left[\sum_{i\in s_j} y_i + \sum_{k\in r_j} \mathbf{x}_{kj}\widehat{\boldsymbol{\beta}}^T + \widehat{u}_j\right]$, which is expected to be more efficient than the regression synthetic estimates obtained with the WB approach.

MSE estimation for the EBP estimates relies on a parametric bootstrap scheme (see also González-Manteiga et al., 2008). In particular, using the bootstrap population model equation (3.2) B bootstrap populations are generated and the target population parameters are computed for each bootstrap population. From each bootstrap population a sample is selected and the EBP approach is implemented using the sample and bootstrap population data. MSE estimates of the EBP estimates are computed over the B bootstrap replications.

3.3 Outlier robust methodologies

Recent small area estimation literature has been concerned with methods that are outlier robust. Relying on parametric assumptions as is the case with the EBP can have an impact on the quality of the small area estimates. In this section we review some methodologies that are recently proposed or are currently under development.

An alternative approach to small area estimation is based on the use of a quantile/M-quantile regression model (cf. Chambers and Tzavidis, 2006). Chambers and Tzavidis (2006) extended the use of M-quantile regression models to small area estimation. Following their development, these authors characterize the conditional variability across the population of interest by the M-quantile coefficients of the population units. For unit i with values y_i and $\mathbf{x}_i$, this coefficient is the value θ_i such that $MQ_y(\theta_i|\mathbf{x_i};\psi) = y_i$. The M-quantile coefficients are determined at the population level. Consequently, if a hierarchical structure does explain part of the variability in the population data, then we expect units within clusters (domains) defined by this hierarchy to have similar M-quantile coefficients. An area specific semi-parametric (empirical) random effect, θ_j, can be computed by the expected value of the M-quantile coefficients in area j.

Similarly to the EBP approach, we start by decomposing the population small area-specific poverty indicator as follows:

$$z_j\left(\alpha,t\right) = N_j^{-1}\left[\sum_{i\in s_j} z_i\left(\alpha,t\right) + \sum_{k\in r_j} z_k\left(\alpha,t\right)\right]. \tag{3.5}$$

The first component in equation (3.5) is observed in the sample whereas the second component is unknown and predicted values can be obtained by using a small area model. The EBP approach (cf. Molina and Rao, 2010) makes Gaussian, assumptions for the nested error regression model error terms. If it is known that the error distribution is normal, the EBP will offer the optimal approach for estimating poverty indicators for small areas. What if, however, the true error distribution is unknown?

A non-parametric approach to estimating equation (3.5) is offered by using a smearing-type estimator that can be motivated by the work of Duan (1983). More specifically:

$$z_j\left(\alpha,t\right) = N_j^{-1}\left[\sum_{i\in s_j} z_i\left(\alpha,t\right) + \sum_{k\in r_j} E\left(z_k\left(\alpha,t\right)\right)\right]. \tag{3.6}$$

We are now interested in finding an estimator for $E\left[z_k\left(\alpha,t\right)\right]$. For simplicity, let us focus on the simplest case, i.e. that of estimating the HCR, $z_k\left(0,t\right)$. In this case, $z_k\left(0,t\right) = I\left(y_k \leq t\right)$. The y_k values are unknown and hence we can use the M-quantile small area model for predicting these values. It follows that:

$$E\left[z_k\left(0,t\right)\right] = \int I\left(\mathbf{x}_i^T \boldsymbol{\beta}_\psi\left(\theta_j\right) + \varepsilon \leq t\right) dF\left(\varepsilon\right). \tag{3.7}$$

Since we make no assumptions about the error distribution, $F\left(\varepsilon\right)$, we can estimate $F\left(\varepsilon\right)$ by the empirical distribution of the residuals:

$$\hat{F}\left(\varepsilon\right) = n^{-1}\sum_{i=1}^{n} I\left(\hat{\varepsilon}_i \leq \varepsilon\right).$$

It follows that:

$$\widehat{E}\left[z_k\left(0,t\right)\right] = \int I\left(\mathbf{x}_k^T \widehat{\boldsymbol{\beta}}_\psi\left(\widehat{\theta}_{\mathbf{j}}\right) + \widehat{\varepsilon}_i \leq t\right) d\widehat{F}\left(\varepsilon\right) = n_j^{-1}\sum_{k\in r_j}\sum_{i\in s_j} I\left(\mathbf{x}_k^T \widehat{\boldsymbol{\beta}}_\psi\left(\widehat{\theta}_j\right) + \widehat{\varepsilon}_i \leq t\right), \tag{3.8}$$

where $\widehat{\varepsilon}_i$ are the estimated residuals from the M-quantile fit. An estimator of $z_j\left(0,t\right)$ is then obtained by substituting equation (3.8) into equation (3.5) leading to:

$$\widehat{z}_j\left(0{,}t\right) = N_j^{-1}\left[\sum_{i\in s_j} z_i\left(0{,}t\right) + \widehat{E}\left[z_k\left(0{,}t\right)\right]\right]. \tag{3.9}$$

The same approach can be followed for estimating $z_j\left(1{,}t\right)$ or for any other of the FGT poverty measures. Since robust estimation for the M-quantile model is automatic, conventionally the M-quantile model is estimated using the raw values of the welfare survey variable. However, the decision as to whether to use the logarithmically transformed or the untransformed values depends on what the model diagnostics suggest.

Mean Squared Error estimation for equation (3.9) is discussed in detail in Marchetti et al. (2012a) and is based on a non-parametric bootstrap scheme. Here we recall the main steps of this bootstrap scheme. Starting from sample s, selected from a finite population U without replacement, we fit the M-quantile small area model and obtain estimates of $\widehat{\theta}_j$ and $\widehat{\boldsymbol{\beta}}_\psi\left(\widehat{\theta}_j\right)$ which are used to compute the model residuals. We then generate B bootstrap populations, U^{*b}. From each bootstrap population we select L bootstrap samples using simple random sampling within the small areas and without replacement such that $n_j^* = n_j$. Bootstrap populations are generated by sampling from the empirical distribution of the residuals or a smoothed version of this distribution conditionally or unconditionally on the small areas. Bootstrap estimators of the bias and variance of the estimated target small area parameter, $\widehat{\tau}_j$, are defined respectively by:

$$\widehat{B}\left(\widehat{\tau}_j\right) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}\left(\widehat{\tau}_j^{*bl} - \tau_j^{*b}\right)$$

$$\widehat{V}\left(\widehat{\tau}_j\right) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}\left(\widehat{\tau}_j^{*bl} - \bar{\widehat{\tau}}_j^{*bl}\right)^2,$$

where τ_j^{*b} is the small area parameter of the b^{th} bootstrap population, $\widehat{\tau}_j^{*bl}$ is the small area parameter estimated by using the l^{th} sample from the b^{th} bootstrap population and $\bar{\widehat{\tau}}_j^{*bl} = L^{-1}\sum_{l=1}^{L}\widehat{\tau}_j^{*bl}$. The bootstrap MSE estimator of the estimated small area target parameter is then defined as:

$$\widehat{M}\left(\widehat{\tau}_j\right) = \widehat{V}\left(\widehat{\tau}_j\right) + \widehat{B}\left(\widehat{\tau}_j\right)^2 \tag{3.10}$$

More recently there has been work on other approaches that attempt to impose less strict parametric assumptions. To start with Gershunskaya and Lahiri (2011) and Elbers and Van der Weide (2014) proposed the use of EBP by using normal mixtures. A more recent alternative is the use of EBP under an alternative parametric distribution in particular, the generalised beta distribution of the second kind (cf. Molina et al., 2015). A further recent alternative is to use a random effects model for the quantiles of the empirical distribution function of income. Modelling a grid of quantiles of the empirical distribution allows us to obtain a good approximation of the target distribution from which synthetic income values can be micro-simulated (cf. Weidenhammer et al., 2015). Evaluating the performance of the outlier robust methodologies is work in progress.

3.4 An application of the EBP methodology: Poverty mapping for NUTS2 areas in Spain

In Spain the European Survey on Income and Living Conditions (EU-SILC) is carried out yearly by INE with the aim of producing estimates of poverty and living conditions both at national and at regional levels. Regions are planned domains for which EU-SILC estimates are published, while the provinces (LAU-1 level) are unplanned domains. Provinces are further partitioned into municipalities (LAU-2 level). In this section we describe the application of small area methodologies for estimating small area poverty and inequality indicators (HCR, PG, Gini) for NUTS2 regions in Spain. For the purposes of this application we used survey data from the 2004 and 2011 EU-SILC surveys. Although the real target would have been to obtain estimates of those indicators at NUTS3 level, the cooperation between Southampton and CED only served for transferring knowledge on how to implement the small area methodologies. In the last section of this report we discuss further the current data limitations that prevented us from considering more disaggregated domain estimation although in some cases the data are available.

The small area methods presented in the previous sections require auxiliary information available for all units (households) in the population. In this application we draw this information from the publicly available IPUMS data in 2001 and 2011. IPUMS auxiliary variables in the case of Spain, also available in the EU-SILC 2004 (closest survey to 2001) and 2011 and survey dataset, included access to phone, access to a car, number of rooms, number of household members, tenure, education and employment. These explanatory variables are included in the small area model that is used to model equivalised household income.

Before proceeding with the presentation of the results, two comments are in order at this stage. The target geography in this application is NUTS 2 level. Although the main target geography is NUTS 3 and the NUTS 3 identifiers are available in the IPUMS dataset, NUTS 3 identifiers are not available in the EU-SILC survey. Hence, this prevents the application of the EBP methodology at NUTS 3 level. Should NSIs be interested in NUTS 3 estimation, they identifiers should be made available to researchers possibly in a safe setting. The second comment relates to the use of IPUMS data. In theory the EBP method should be implemented with the latest Census micro-data available. However, for the purposes of this application, the use of the IPUMS public dataset can offer some useful insights and a benchmark against which estimates that use more detailed data can be compared to.

Table 3.1: EBP estimates of the HCR, PG and of the Gini in NUTS2 areas of Spain in 2004 and 2011.

	2004			**2011**		
NUTS2	**hcr**	**gini**	**pgap**	**hcr**	**gini**	**pgap**
11	0.29	0.32	0.11	0.27	0.38	0.15
12	0.19	0.30	0.06	0.22	0.35	0.11
13	0.19	0.30	0.06	0.29	0.39	0.16
21	0.14	0.29	0.04	0.17	0.33	0.08
22	0.14	0.29	0.04	0.14	0.32	0.07
23	0.23	0.31	0.08	0.26	0.37	0.14
24	0.16	0.29	0.05	0.22	0.35	0.12
30	0.14	0.29	0.04	0.20	0.34	0.10
41	0.27	0.32	0.10	0.27	0.38	0.15
42	0.32	0.33	0.13	0.32	0.40	0.19
43	0.40	0.35	0.17	0.29	0.39	0.17
51	0.15	0.29	0.05	0.19	0.34	0.10
52	0.23	0.31	0.08	0.29	0.39	0.16
53	0.19	0.30	0.07	0.22	0.35	0.11
61	0.32	0.33	0.13	0.32	0.40	0.19
62	0.29	0.32	0.11	0.31	0.40	0.18
63	0.32	0.34	0.13	0.24	0.36	0.13
64	0.30	0.33	0.12	0.30	0.39	0.17
70	0.27	0.32	0.10	0.30	0.39	0.17

In 2011 both the IPUMS and the EU-SILC contained 19 NUTS2 areas. In addition, the 2001 IPUMS also included 19 NUTS2 areas but the 2004 EU-SILC only included 18 NUTS2 as Melilla was missing.

The poverty line used for computing estimates of HCR and PG corresponds to 60% of the

Spanish median equivalised household income. This estimate is derived by using the 2004 and 2011 EU-SILC income values which are available for the entire sample in Spain with weights equal to the cross-sectional EU-SILC household weights. Model-based estimates are presented in Table 3.1 and in Figure 3.1, 3.2 and 3.3.

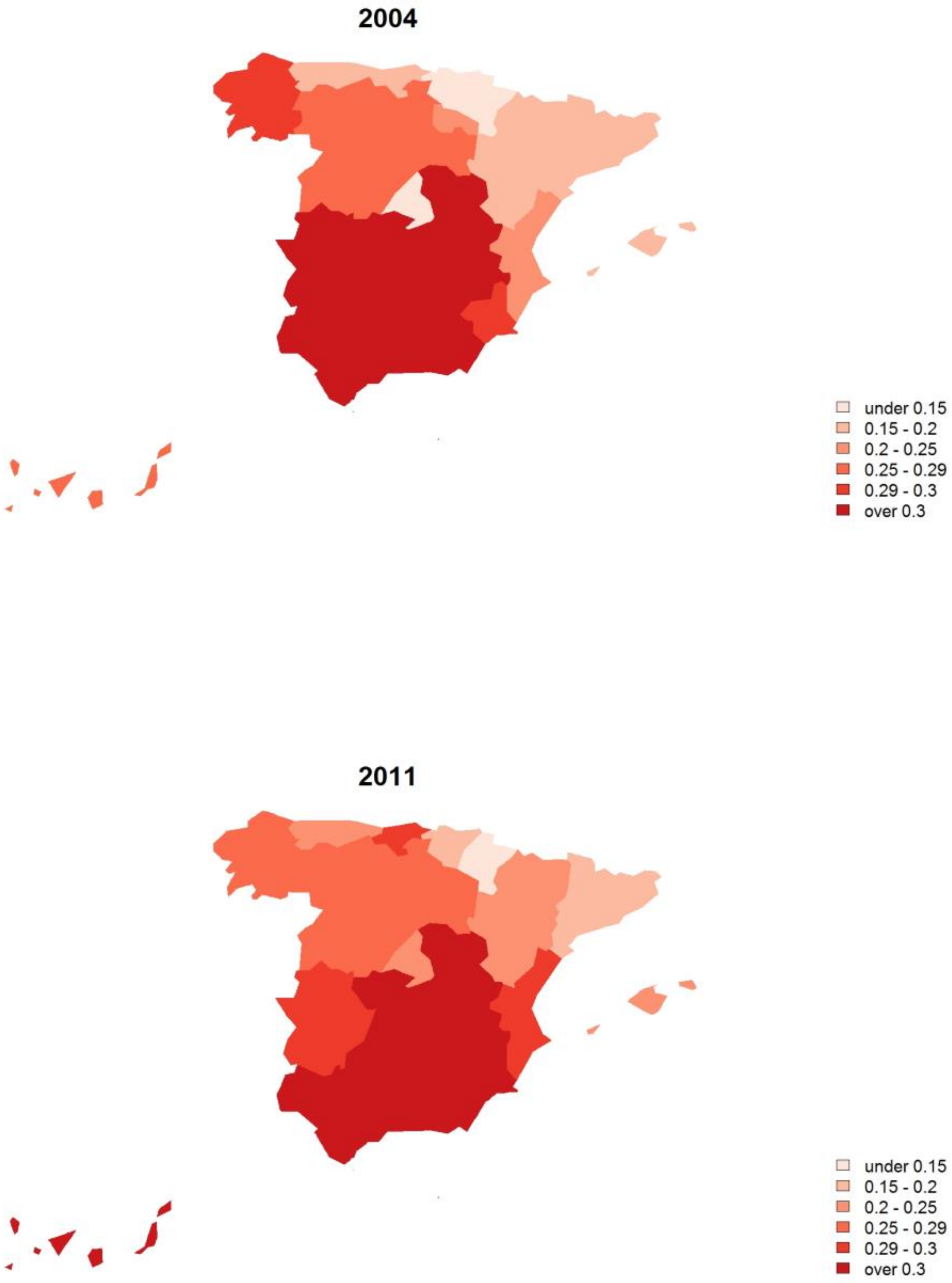
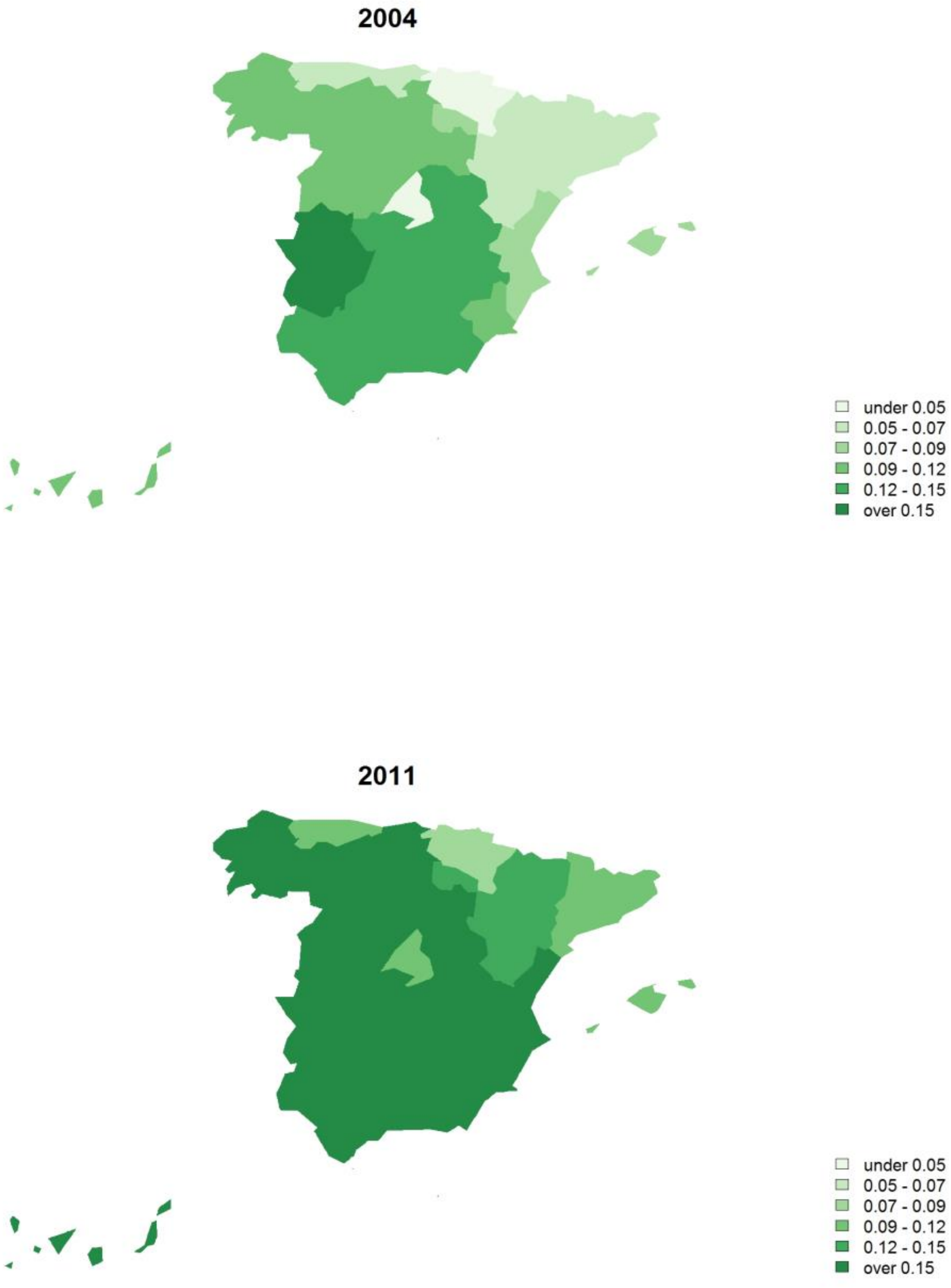


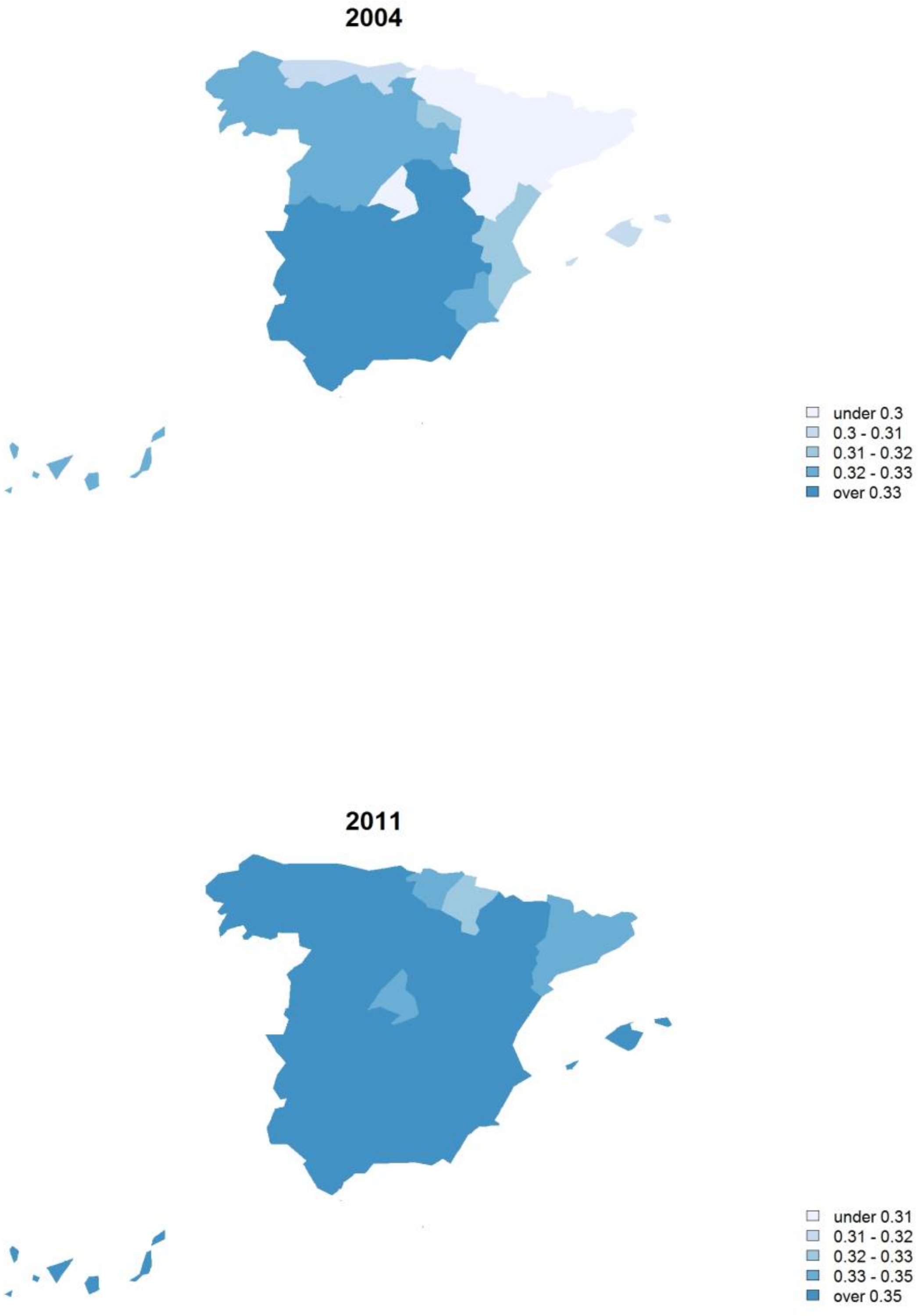
Figure 3.1: EBP estimates of HCR for NUTS2 in Spain in 2004 and 2011



2004

2011

Figure 3.2: EBP estimates of PG for NUTS2 in Spain in 2004 and 2011



2004

2011

Figure 3.3: EBP estimates of Gini for NUTS2 in Spain in 2004 and 2011

3.5 Concluding remarks and challenges

This report reviews some recently proposed model-based methods for small area poverty estimation with particular emphasis on Empirical Best Prediction and on the estimation of linear and non-linear indicators. We have further alternative robust methods that attempt to reduce the dependency on parametric assumptions about the model error terms. One area of current research activity is the use of appropriate transformations that will allow the implementation of the EBP approach assuming Normality of the model error terms. This has the advantage of a model that can be readily estimated by using standard software. However, selecting an appropriate transformation requires the use of diagnostic analysis and careful model selection before implementing the final small area model.

The final remark concerns the availability of data and the target geography. As we mentioned in this report the CED team could not get access to the NUTS3 identifiers in the EU-SILC survey. However, in the past we have worked with the EU-SILC data at more refined than NUTS2 geographies by special arrangements with National Statistical Institutes. Hence, NSI could provide access to data with suitable geographical detail within a secure environment. A second data challenge is the requirement for access to Census microdata. IPUMS provides an excellent publicly available dataset that can be used for evaluation and training purposes. However if NSIs are interested in producing official statistics that are accredited as National Statistics, then use of Census microdata may be more appropriate. As this is a highly confidential dataset, implementation of small area methodologies has to be done within a safe setting.

4 Accuracy measures for changes over time

Yves G. Berger

University of Southampton, UK

The work presented in this deliverable is submitted to the International Statistical Review journal (Berger and Escobar, 2015).

Measuring change over time is a central problem for many users of social, economic and demographic data. The primary interest of many users is often in changes or trends from one time period to another. Smith et al. (2003) recognised that assessing change is one of the most important challenge in survey statistics. A common problem is to compare two cross-sectional estimates for the same study variable taken on two different waves or occasions. These cross-sectional estimates often include imputed values to compensate for item non-response (e.g. Lohr, 2009, ch. 8). The estimation of the variance of an estimator of change is useful to judge whether the observed change is statistically significant.

We propose to use a multivariate linear regression approach (Berger and Priam, 2016) to estimate these covariances. The proposed estimator is not a model-based estimator, as it is valid even if the underlying model does not fit the data. We show how this approach can be used to accommodate the effect of imputation. The regression approach gives design-consistent estimation of the variance of change when the sampling fraction is small. We illustrate the proposed approach using random hot-deck imputation, although the proposed estimator can be implemented with any other imputation techniques.

4.1 Rotating surveys

The estimation of variance of change would be relatively straightforward if cross-sectional estimates were based on the same sample. Furthermore, because of rotations that is used in repeated surveys, cross-sectional estimates are not independent. Let s_1 and s_2 denote respectively the first and the second wave samples. The samples s_1 and s_2 are usually not completely overlapping sets of units, because repeated surveys use rotation designs which consist in selecting new units ($k \in s_2 \setminus s_1$) to replace old units ($k \in s_1 \setminus s_2$) that have been in the survey for a specified number of waves. Without lost of generality, we assume that s_1 and s_2 have the same sample size n. Let n_{12} denote the sample size of the common sample, $s_{12} = s_1 \cap s_2$. The units sampled on s_{12} represent usually a large fraction of s_1;

that is, n_{12}/n is usually large. The Figure 4.1 gives a visual representation of the samples considered.

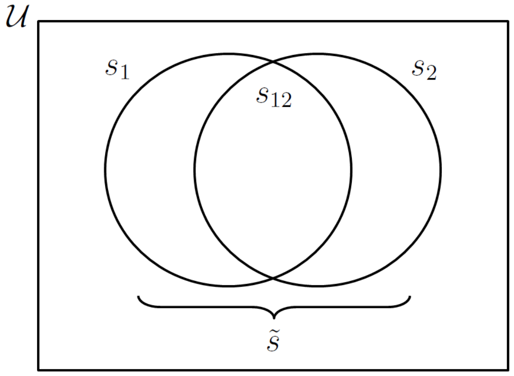


Figure 4.1: The overall sample $\tilde{s} = s_1 \cup s_2$.

Let $y_{\ell;k}$ denote the value of the variable of interest y_ℓ for the wave $(\ell = 1, 2)$. Suppose, we wish to estimate the absolute change

$$\Delta \;=\; \tau_2 \;-\; \tau_1, \tag{4.1}$$

between two population totals τ_1 and τ_2 from waves 1 and 2, where $\tau_\ell = \sum_{k \in \mathcal{U}} y_{\ell;k}$. Here, $\mathcal{U}$ denotes the population of size N, assumed to be the same at both waves. It is possible to extend the approach we proposed for other measures of changes, such as relative change or change between means.

4.2 Non-response

Our main objective is to address the problem of variance estimation under non-response rather than the non-response issue. Little has been done on variance estimation of change. However, there are many design-based variance estimators of cross-sectional estimates (e.g. Wolter, 2007). The use of models to address non-response is also popular. A model-assisted approach can be found in Deville and Särndal (1994), Fay (1994), Steel and Fay (1995), Särndal and Lundström (2005). A Bayesian treatment of imputation can be found for example in Rubin (1987). See Brick and Montaquila (2009) for a wide discussion on non-response. A discussion on which inference-approach to use for non-response in surveys can be found in Haziza (2009). These approaches deal with cross-sectional estimators, and cannot be directly implemented with estimators of changes.

We propose to use a design-based approach combined with random hot-deck imputation. A recent review on cross-sectional hot-deck imputation can be found in ANDRIDGE and LITTLE (2010). The random hot-deck imputation has the advantage of guaranteeing unbiased estimation of population distributions (RAO and SHAO, 1992). The approach proposed is also valid under deterministic regression imputation.

Suppose that the change Δ in (4.1) is estimated by

$$\widehat{\Delta}^I = \widehat{\tau}_2^I - \widehat{\tau}_1^I, \tag{4.2}$$

where

$$\widehat{\tau}_\ell^I = \sum_{k \in \tilde{s}} \frac{y_{\ell;k}^I}{\pi_{\ell;k}} \qquad (\ell = 1, 2) \tag{4.3}$$

is the cross-sectional imputed NARAIN (1951), HORVITZ and THOMPSON (1952) estimators at wave ℓ. Here $y_{\ell;k}^I$ the values of the imputed variable.

4.3 Variance of the hot-deck imputed estimator of change

We propose to estimate the variance of $\widehat{\Delta}^I$ defined by equation (4.2) using a reverse approach for non-response (FAY, 1991, SHAO and STEEL, 1999). Let $\mathcal{U}_\ell^r$ be the population of respondents at wave ℓ, where $\mathcal{U}_\ell^r \subset \mathcal{U}$. In other words, at both waves, the population is randomly split into a population of respondents and a population of non-respondents according to an unknown response mechanism. Note that the response mechanisms can be such that the set of respondents of the wave 2 depends on the set of respondents at wave 1.

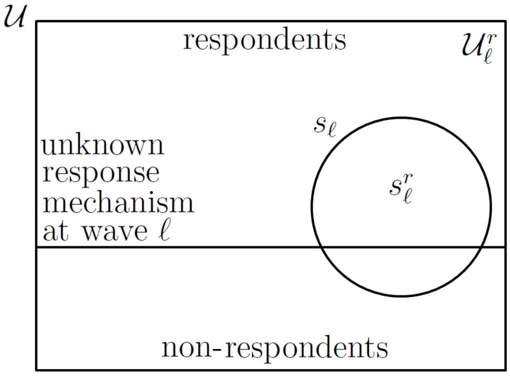


Figure 4.2: Non-response at wave $\ell = 1, 2$.

Let $E_r\{\cdot\}$ and $V_r\{\cdot\}$ denote respectively the expectation and variance operators with respect to the response mechanism. Rotation samples s_1 and s_2 are selected from the population $\mathcal{U}$ according to a rotation sampling design. The samples of respondents are given by $s_\ell^r = \mathcal{U}_\ell^r \cap s_\ell, (\ell = 1, 2)$. Let $E_d\{\cdot\}$ and $V_d\{\cdot\}$ denote the expectation and the variance operators with respect to the sampling design. Let $E_I\{\cdot\}$ and $V_I\{\cdot\}$ denote the expectation and the variance operators with respect to the random imputation.

We proposed to estimate the variance of the imputed estimator (4.2) by

$$\widehat{V}\{\widehat{\Delta}^I\} \;=\; \widehat{V}_d\{E_I\{\widehat{\Delta}^I|S,R\}|R\} \;+\; \widehat{V}_I\{\widehat{\Delta}^I|S,R\}, \tag{4.4}$$

where

$$\begin{aligned} \widehat{V}_d\{E_I\{\widehat{\Delta}^I|S,R\}|R\} &= \boldsymbol{\nabla}(\widehat{\boldsymbol{\tau}})^\top\, \widehat{\boldsymbol{V_d}}(\widehat{\boldsymbol{\tau}}|R)\, \boldsymbol{\nabla}(\widehat{\boldsymbol{\tau}}), \qquad (4.5) \\ \widehat{V}_I\{\widehat{\Delta}^I|S,R\} &= \sum_{h=1}^{H}\sum_{\ell=1}^{2} V_I\{y^*_{\ell;k}|S,R\} \sum_{k\in\tilde{s}} \frac{z^{(h)}_{\ell;k}}{\pi^2_{\ell;k}}(1-a_{\ell;k})\cdot \end{aligned}$$

Berger and Escobar (2015) proposed a multivariate (or general) linear regression model to estimate the covariance matrix $\boldsymbol{V_d}(\widehat{\boldsymbol{\tau}}|R)$ involved in the estimator (4.5). Berger and Escobar (2015) showed that the proposed estimator (4.4) is an approximately unbiased estimator of the variance $V(\widehat{\Delta}^I)$.

The advantages of the proposed variance estimator (4.4) are that it is approximately unbiased under the unknown response mechanisms and that it does not involve the estimation of the response probabilities. Moreover, note that the estimator (4.4) can be generalised for other types of imputation, as long as $E_I\{\widehat{\Delta}^I|S,R\}$ is a function of Narain-Horvitz-Thompson estimators of totals.

4.4 Simulation study

4.4.1 Labor Force Population

We use the *Labor Force Population* dataset from Valliant et al. (2000, Appendix B.5) available at the John Wiley worldwide website. The dataset is duplicated 50 times to obtain a large population suitable for different levels of rotation and small sampling fractions in the sampling design. We consider two variables: the weekly wages and the hours worked per week (HW). The units with the value 99 for the weekly wage and 999 for the hours worked per week were removed from the population frame. These units were not treated as missing. We obtain a population frame of size $N = 23\,550$. The target variables $y_{1;k}$ and $y_{2;k}$ are given by,

$$\begin{aligned} y_{1;k} &= \textit{Weekly wages}, \\ y_{2;k} &= y_{1;k} + \sqrt{y_{1;k}} + \psi_k, \end{aligned}$$

where ψ_k denotes randomly generated values according to a Normal distribution $N(0, 5^2)$. The true absolute change between the two wave totals is given by $\Delta = 377\,960.66$. We estimate Δ by the hot-deck imputed point estimator $\widehat{\Delta}^I$ defined by equation (4.2). The first wave sample s_1 is selected using the Rao (1965) and Sampford (1967) unequal probability sampling design. We consider two scenarios for the inclusion probabilities: the $\pi_{1;k}$ are constant ($\pi_{1;k} = n/N$), and the $\pi_{1;k}$ are proportional to the variable *hours worked per week* which has values all larger than 5. We consider that we have a single stratum.

For the second wave sample s_2 we select a simple random sample of n_{12} units taken from s_1, where $g = n_{12}/n = \{0.40, 0.60, 0.80, 0.95\}$, and a sample of $n - n_{12}$ units from $\mathcal{U} \setminus s_1$ selected with probabilities proportional to $\pi_{2;k} = \pi_{1;k}/(1 - \pi_{1;k})$. We have that $\pi_{2;k} \fallingdotseq \pi_{1;k}$ (Berger and Priam, 2016).

Let $a_{1;k} = 1$ if $u_{1;k} \leq q_1$ and $a_{1;k} = 0$ otherwise, where q_1 is a fixed quantity which specify the response rate at wave 1, and $u_{1;k}$ are independent uniform random variables $U(0, 1)$. Let $a_{2;k} = 1$ if $u_{2;k} \leq (0.95)\, a_{1;k} + (0.65)\,(1 - a_{1;k})$ and $a_{2;k} = 0$ otherwise, where $u_{2;k}$ are independent uniform random variables $U(0, 1)$. Note that $a_{1;k}$ and $a_{2;k}$ are dependent because a respondent at wave 1 is more likely to be also a respondent on wave 2. The items non-response are imputed using random hot-deck. We consider that we have a single imputation class. A new set of respondents $(a_{1;k}, a_{2;k})$ is generated randomly before each selection of s_1 and s_2.

For each simulation, 10 000 samples are selected to compute: the empirical relative bias $\mathrm{RB} = \mathrm{Bias}(\widehat{\mathrm{var}}(\widehat{\Delta}^I))/\mathrm{var}(\widehat{\Delta}^I)$ where $\mathrm{Bias}(\widehat{\mathrm{var}}(\widehat{\Delta}^I)) = \mathrm{E}(\widehat{\mathrm{var}}(\widehat{\Delta}^I)) - \mathrm{var}(\widehat{\Delta}^I)$, the empirical relative root mean squared error $\mathrm{RRMSE} = (\mathrm{MSE}(\widehat{\mathrm{var}}(\widehat{\Delta}^I)))^{1/2}/\mathrm{var}(\widehat{\Delta}^I)$, and the coverage of the 95% confidence interval $\widehat{\Delta}^I \pm 1.96\, \widehat{\mathrm{var}}(\widehat{\Delta}^I)^{1/2}$. The term $\mathrm{var}(\widehat{\Delta}^I)$ denotes the empirical variance computed from the 10 000 observed values of $\widehat{\Delta}^I$. Computations were performed in R (R Core Team, 2015) using some routines from the R packages ‘sampling’ (Tillé and Matei, 2013) and ‘samplingVarEst’ (Escobar and Barrios, 2014). We com-

Table 4.1: RB, RRMSE and Coverage of 95% confidence interval of the variance estimators. Hot-deck imputed point estimator $\widehat{\Delta}^I$. $\pi_{1;k} = n/N$.

$q_{1;k}$	$q_{2;k}$	g	f	RB		RRMSE		Coverage	
				Prop.	Naïve	Prop.	Naïve	Prop.	Naïve
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
0.70	0.86	40	0.5	-2.8	-33.8	15.5	35.3	95.0	88.7
			1.0	-0.7	-32.3	11.2	33.2	94.8	89.3
			1.5	-0.4	-32.1	9.1	32.7	94.7	89.5
			2.0	-2.7	-33.7	8.2	34.1	94.6	88.8
		60	0.5	-1.8	-31.3	17.6	33.7	94.7	89.1
			1.0	-1.2	-30.9	12.5	32.2	94.8	89.7
			1.5	-1.1	-30.8	10.2	31.7	94.7	89.2
			2.0	0.0	-30.1	8.7	30.8	94.8	89.9
		80	0.5	-1.8	-28.8	20.1	32.7	94.7	89.8
			1.0	-0.4	-27.5	14.4	29.8	95.0	90.4
			1.5	-0.4	-27.5	11.6	29.0	95.0	90.5
			2.0	-2.2	-29.0	10.0	30.0	94.6	90.0
		95	0.5	-1.8	-25.3	22.8	31.9	94.8	90.8
			1.0	-1.9	-25.5	16.0	29.0	94.5	90.8
			1.5	-0.9	-24.8	13.1	27.2	94.8	90.7
			2.0	-1.6	-25.3	11.2	27.0	94.7	90.9
0.90	0.92	40	0.5	-0.7	-15.9	14.5	20.2	94.8	92.9
			1.0	0.2	-15.2	10.2	17.6	95.3	93.2
			1.5	-2.1	-17.2	8.5	18.5	94.7	92.6
			2.0	-0.7	-15.9	7.2	17.1	95.1	92.8
		60	0.5	0.4	-14.4	17.2	21.0	94.9	93.2
			1.0	0.2	-14.6	12.1	18.2	94.9	92.9
			1.5	0.6	-14.2	9.9	16.7	95.1	93.1
			2.0	0.0	-14.8	8.5	16.7	94.8	92.7
		80	0.5	-2.2	-15.3	21.4	25.5	94.7	93.0
			1.0	-2.0	-15.0	15.0	20.8	95.0	93.1
			1.5	-0.2	-13.7	12.3	18.2	94.8	92.9
			2.0	-1.0	-14.4	10.7	17.7	94.6	92.9
		95	0.5	-2.9	-13.4	27.7	33.0	94.5	92.9
			1.0	-2.4	-13.2	19.4	24.8	94.5	93.2
			1.5	-1.1	-12.0	15.9	21.3	95.1	93.6
			2.0	-0.9	-12.0	13.8	19.3	94.9	93.5

pare the proposed estimator $\widehat{V}(\widehat{\Delta}^I)$ from (4.4) versus a naïve approach which consists in treating the imputed values as real values. Note that there is no other competitor for the proposed approach, as design-based variance estimators for imputed change estimators is non existent in the literature.

Tables 4.1 and 4.2 give the RB, the RRMSE and the coverage for different values of the overlapping fraction g between waves. In Table 4.1, $\pi_{1;k} = n/N$ and in Table 4.2 the $\pi_{1;k}$ are proportional to the variable *hours worked per week*.

Table 4.2: RB, RRMSE and Coverage of 95% confidence interval of the variance estimators. Hot-deck imputed point estimator $\widehat{\Delta}^I$. $\pi_{1;k} \propto HW$.

$q_{1;k}$	$q_{2;k}$	g	f	RB		RRMSE		Coverage	
				Prop.	Naïve	Prop.	Naïve	Prop.	Naïve
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
0.70	0.86	40	0.5	-1.6	-29.1	32.3	52.7	94.3	88.7
			1.0	-2.5	-29.9	23.3	42.8	93.5	87.3
			1.5	-3.4	-30.5	19.0	39.7	93.0	86.8
			2.0	-1.5	-29.4	16.4	36.7	92.4	86.5
		60	0.5	-2.0	-27.7	36.2	57.1	94.3	89.3
			1.0	-1.2	-27.5	26.1	44.0	94.2	88.9
			1.5	-0.8	-27.1	21.4	39.1	94.1	88.9
			2.0	-0.7	-27.5	18.2	36.4	93.5	87.8
		80	0.5	-0.1	-25.6	40.7	59.2	94.8	90.4
			1.0	0.0	-25.2	29.3	45.5	94.9	89.7
			1.5	-0.4	-25.1	23.6	40.4	94.8	90.2
			2.0	-0.6	-25.8	20.4	37.1	94.5	89.7
		95	0.5	-1.5	-24.3	43.9	63.6	94.5	90.8
			1.0	0.4	-22.9	31.7	48.5	95.2	91.3
			1.5	0.5	-23.4	26.0	41.4	94.9	91.2
			2.0	-0.8	-24.3	22.3	38.1	94.9	90.6
0.90	0.92	40	0.5	-0.5	-15.5	34.3	51.2	94.1	91.7
			1.0	-1.5	-15.6	23.5	37.7	93.1	90.4
			1.5	-0.5	-14.8	19.9	33.1	92.9	90.2
			2.0	-1.7	-16.0	16.9	29.7	91.8	88.9
		60	0.5	-0.1	-14.2	41.2	61.1	94.3	92.3
			1.0	-2.4	-15.3	29.2	48.1	93.8	91.6
			1.5	0.2	-13.4	23.6	38.1	93.9	91.3
			2.0	-1.0	-14.4	20.5	34.3	93.0	90.6
		80	0.5	-0.3	-12.8	51.9	78.6	94.5	93.1
			1.0	-0.3	-11.7	36.3	59.9	94.7	93.1
			1.5	-1.3	-12.8	29.3	49.3	94.2	92.0
			2.0	-0.4	-12.4	25.6	42.1	94.2	92.1
		95	0.5	-0.8	-11.5	64.3	99.0	94.7	94.4
			1.0	-1.9	-11.5	44.0	71.4	94.3	93.5
			1.5	-1.5	-11.9	35.5	58.8	94.6	93.2
			2.0	-0.8	-11.3	30.7	49.7	94.6	93.4

The proposed approach gives negligible RB. As expected, the naïve approach tends to severely underestimate the variance; in particular, when the fraction of non respondents is large; that is, when $q_{1;k}$ is small. Furthermore, by comparing Table 4.1 and 4.2, we observe smaller RB with unequal inclusion probabilities.

The proposed approach has smaller RRMSE than the naïve approach. However, with unequal probabilities we observe larger RRMSE. The coverage of the proposed approach is closer to 95%. The coverage of the naïve approach is lower because of the under-estimation of the variance.

4.4.2 Missing not at random response and multiple imputation-classes

Four variables, y_1, y_2, x_1, x_2 and w_1, are generated from a multivariate normal distribution with means 20, 10, 20, 10 and 20. All the variables have the same variance equals to 5. The correlation between y_1 and y_2 is either $\rho(y_1, y_2) = 0.7$ or $\rho(y_1, y_2) = 0.9$. The other correlations are $\rho(y_\ell, x_{\ell'}) = \rho(y_\ell, w_1) = 0.7$ and $\rho(x_\ell, x_{\ell'}) = \rho(x_\ell, w_1) = 0.5$ $(\ell \neq \ell')$. The wave 1 variables are y_1, x_1 and w_1. The wave 2 variables are y_2 and x_2. We generate $N = 20\,000$ values for each variables.

The values $y_{1;k}$ and $y_{2;k}$ are the values of the variable y_1 and y_2. The parameter of interest is the absolute change between means: $\Delta_\mu = \Delta/N$. The imputed estimator is $\widehat{\Delta}_\mu^I = \widehat{\Delta}^I/N$.

The sample s_1 is a randomised systematic sample with first-order inclusion probabilities $\pi_{1;k}$ proportional to $w_{1;k}$, where $w_{1;k}$ denotes the k-th value of w_1. The sample s_2 is a simple random sample of n_{12} units selected from s_1 combined with a randomised systematic sample of $n_2 - n_{12}$ units selected without replacement from $U \setminus s_1$ with probabilities proportional to $\pi_{1;k}/(1 - \pi_{1;k})$. We have that $\pi_{2;k} \doteq \pi_{1;k}$ (Berger and Priam, 2016). The sample sizes are $n_1 = n_2 = 500$ and $n_{12} = 375$. We consider that we have a single stratum. 10 000 samples s_1 and s_2 are selected. The Hansen and Hurwitz (1943) variance estimator is used for cross-sectional variance estimation.

We consider hot-deck imputation with multiple imputation-classes. The number of imputation classes, denoted by C, is the same on wave 1 and 2. We consider three types of imputation classes.

(i) "*Population imputation classes*": The imputation classes of wave ℓ are C quantile classes based on the variable x_ℓ. The bounds of the classes are the $(100c/C)\%$ quantiles $(c = 1, \ldots, C)$ of the population values $\{x_{\ell;k} : k \in U\}$, where $x_{\ell;k}$ denotes the k-th value of x_ℓ.

(ii) "*Sample imputation classes*": The imputation classes of wave ℓ are are C quantile classes based on the sample values of the variable x_ℓ. The bounds of the classes are the $(100c/C)\%$ quantiles $(c = 1, \ldots, C)$ of the sample values $\{x_{\ell;k} : k \in s_\ell\}$.

(iii) "*Across-waves imputation classes*": For the classes of wave 1, we use C quantile classes based on the sample values of the variable x_1, as in (ii). The wave 2 imputation classes are C quantile classes based on the sample values $\{\widetilde{y}_{1;k} : k \in s_2\}$, where

$$\widetilde{y}_{1;k} = \begin{cases} y^I_{1;k} & \text{for } k \in s_{12}, \\ \widehat{\beta}_0 + \widehat{\beta}_1 x_{2;k} & \text{for } k \in s_2 \setminus s_1. \end{cases} \tag{4.6}$$

Here, $x_{2;k}$ is the value of the variable x_2 for unit k. The quantity $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the ordinary least square coefficients of the regression $y^I_{1;k} = \beta_0 + \beta_1 x_{2;k}$, with $k \in s_{12}$.

For the classes (i), the class indicators do not depend on the samples selected. For the classes (ii) and (iii), the class indicators depend on the samples. Note that the classes of wave 1 are different from the classes of wave 2, unless we have a single class.

We consider a "*missing not at random response mechanism*". The first and second wave response probabilities $q_{1;k}$ and $q_{2;k}$ are given by $q_{\ell;k} = \exp(\eta_{\ell;k})\{1 + \exp(\eta_{\ell;k})\}^{-1}$, where $\eta_{1;k} = 4 - 0.15\, y_{1;k}$ and $\eta_{2;k} = 3 - 0.2\, y_{2;k}$. The resulting response probabilities lies within the range [0.25, 0.95]. We have $a_{\ell;k} = 1$ if $u_{\ell;k} \leq q_{\ell;k}$ and $a_{\ell;k} = 0$ otherwise, where $u_{\ell;k}$ are independent uniform random variables $U(0,1)$. The resulting response mechanism is missing not at random because large $q_{1;k}$ and $q_{2;k}$ are associated with small values of $y_{1;k}$ and $y_{2;k}$. The overall response rate are 73% and 72% for the first and second wave. The correlation between $q_{1;k}$ and $q_{2;k}$ is approximately 0.7. The response probabilities are not constant within the imputation classes. Missing values are generated randomly before each selection of s_1 and s_2.

The simulation results are given in Table 4.3. Large number of classes reduces the bias of the point estimator. With a single imputation class ($C = 1$), the variance estimator has the smallest bias and is more stable (small root mean square error, RMSE), but with low coverages (92.9% and 91.8%). The low coverages is explained by the largest bias of the point estimator. Note that the point estimator is more precise with $C \geqslant 1$ and $\rho(y_1, y_2) = 0.9$, in term of bias and variance. However, there is only negligible differences between the variance for $C \geqslant 5$. We only notice a decrease in the variance, as C increases, for population level imputation classes with $\rho(y_1, y_2) = 0.9$. For $C \geqslant 5$, we observe a slight positive bias for the variance estimator and an increase in the RMSE. For population level classes, the RMSE increases with C. The coverage observed are slightly larger than 95% for $C \geqslant 5$. We do not observe significant differences between the imputation classes (i), (ii) and (iii).

Table 4.3: Overall expectation, variance, root-mean squared error (RMSE) and coverage of 95% confidence interval based on the estimator proposed. Missing not at random response mechanisms. $\rho(y_1, y_2)$ denotes the correlation between the variables of interest. $N = 20\,000$, $n_1 = n_2 = 500$ and $n_{12} = 375$. $\Delta_\mu = \Delta/N$ and $\widehat{\Delta}^I_\mu = \widehat{\Delta}^I/N$.

$\rho(y_1, y_2)$	Imputation	C	Δ_μ	$E(\widehat{\Delta}^I_\mu)$	$V(\widehat{\Delta}^I_\mu)$	$E\{\widehat{V}(\widehat{\Delta}^I_\mu)\}$	RMSE	Coverage (%)
0.7	(i) Population level	1	-10.03	-10.09	0.022	0.021	0.0016	92.9
		5	-10.03	-10.06	0.017	0.019	0.0029	95.9
		10	-10.03	-10.06	0.016	0.019	0.0031	96.0
		20	-10.03	-10.06	0.016	0.019	0.0036	96.3
	(ii) Sample level	5	-10.03	-10.06	0.016	0.019	0.0031	96.2
		10	-10.03	-10.06	0.015	0.019	0.0037	96.7
		20	-10.03	-10.06	0.015	0.019	0.0039	96.8
	(iii) Across waves	5	-10.03	-10.09	0.017	0.019	0.0029	94.5
		10	-10.03	-10.08	0.016	0.019	0.0033	95.3
		20	-10.03	-10.08	0.016	0.019	0.0032	95.2
0.9	(i) Population level	1	-9.99	-10.06	0.019	0.019	0.0015	91.8
		5	-9.99	-10.03	0.014	0.017	0.0030	95.7
		10	-9.99	-10.03	0.014	0.016	0.0031	96.0
		20	-9.99	-10.03	0.013	0.016	0.0031	96.2
	(ii) Sample level	5	-9.99	-10.03	0.013	0.017	0.0034	96.3
		10	-9.99	-10.03	0.013	0.016	0.0035	96.2
		20	-9.99	-10.03	0.013	0.016	0.0033	96.3
	(iii) Across waves	5	-9.99	-10.01	0.013	0.016	0.0032	96.8
		10	-9.99	-10.00	0.013	0.016	0.0037	97.3
		20	-9.99	-10.00	0.013	0.016	0.0035	97.2

The missing not at random response mechanism tends to under-represent the large values of the variables of interest and therefore the observed correlation between $y_{1;k}$ and $y_{2;k}$ is lower than $\rho(y_1, y_2)$. As a result, the correlation between $\widehat{\tau}^I_2$ and $\widehat{\tau}^I_1$ is slightly underestimated. This explains the slight positive bias for the variance estimator (Berger, 2004, p. 462). However, this bias is negligible because the coverages of the confidence intervals are of an acceptable order. This bias is only observed for $C \neq 1$. For $C = 1$, the larger variance compensates the bias.

4.5 Discussion

The proposed variance estimator is applicable for unequal rotating stratified sampling designs when random hot-deck imputation is used at both waves and the sampling fractions are negligible. The proposed variance estimator may be extended in various ways.

Point estimators, such as calibration estimators (Huang and Fuller, 1978, Deville and Särndal, 1992) which employ auxiliary population information may often be expressible as functions of totals. The proposed variance estimator (4.5) can be modified to accommodate this situation.

The main advantages of the proposed variance estimator are that it is approximately unbiased under the response mechanisms and that it does not require the estimation of the response probabilities.

The proposed approach is not limited to hot-deck imputation, as it can be extended to other method of imputation, as long as the expectation of the imputed estimator of change under random imputation can be expressed as a function of totals.

The variance estimator is based on the assumption that the imputation class are fixed. However, this assumption does not hold when the imputation classes are based on sampled data. This is also the case when the imputation at wave 2 is based on classes constructed from sample variables observed at wave 1. In Section 4.4.2, we suggest using the the wave 1 variable to impute at wave 2, by using imputation classes based on the variable of interest of wave 1 (see (iii) "*Across-waves imputation classes*"). Our simulation study showed that sample based imputation classes have a negligible effect on the variance estimates, even with across-waves imputation classes. Adjusting the variance estimator to accommodate this situation is beyond the scope of this paper. This is a topic which would need further investigation.

5 Estimating income poverty and inequality from income classes

Simon Lenau **Ralf Münnich**

Economic and Social Statistics Department

Trier University

5.1 Introduction

Fighting poverty and social exclusion is one of the main goals of the EU. Following the Council of the European communities (1985, p. 24), 'the poor' are defined as

> "persons, families and groups of persons whose resources (material, cultural and social) are so limited as to exclude them from the minimum acceptable way of life in the Member States in which they live."

Clearly, poverty and social exclusion do not only mean lack of (current) financial ressources (cf. Atkinson et al., 2005, p. 79). Nevertheless, income poverty and inequality constitutes a central element thereof.

Indicators used to measure poverty and social exclusion as well as changes over time (cf. AMELI, 2011, Social Protection Committee, 2001, SAMPLE, 2009) are commonly estimated from surveys. Thus, these estimates are subject to errors, of which we focus on sampling errors here. The European statistics code of practice (Eurostat, 2012) requests those errors to be estimated and documented as well.

Focussing on the three core indicators proposed by the Indicators Subgroup of the Social Protection Committee (European Commission, 2009b), we choose the following selection from the indicators on poverty and social exclusion:

- The **At-risk-of-poverty rate** (ARPR),
  which solely depends on the lower to medium range of the income distribution.
- The **Gini-Coefficient** (Gini),
  reflecting income inequality of the entire distribution.
- The **Quintile-Share-Ratio** (QSR),
  based mainly on extreme points of the distribution.

These indicators and their estimates are briefly presented in the following. Since their sampling errors can be approximated by linearization, the corresponding linearized values are also treated (cf. Deville, 1999, Osier, 2009). As this works only for the case of a continuous income variable, bootstrapping procedures are discussed briefly.

5.2 Framework for Point- and Variance-Estimation

Focussing on estimation on national level, we restrict ourselves to design-based approaches, not taking into account model-based (small-area) estimation (cf. Burgard et al., 2015, and the references therein). To consider the designs of EU-SILC appropriately, which are summarized in (Graf et al., 2011b, pp. 31f), multi-stage sampling has to be treated. This is possibly combined with inclusion probabilites proportional to the size of households or municipalities. The following methodology is used to account for such complex sampling designs in point and variance estimation. Lacking better terminology, we consider stratification being a sampling stage, even though all strata are selected, thus contradicting a random selection (cf. Särndal et al., 1992, p. 125).

Let i_p be the sampling unit i on sampling stage p in sample $\mathcal{S}$. Its first order sampling inclusion probability $\pi_{p,i}$ is

$$\pi_{p,i} \quad = \quad P(i_p \in \mathcal{S}) \quad . \tag{5.1}$$

For unit k at stage p, we get the estimator for the total of the variable $\mathbf{y}$

$$\hat{\tau}_{p,k}(\mathbf{y}) \quad = \quad \sum_{i \in q_{k_p}} \frac{\hat{\tau}_{(p+1),i}(\mathbf{y})}{\pi_{(p+1),i}} \quad , \tag{5.2}$$

where q_{k_p} is the set of all sampling units on stage $(p+1)$ included in k_p. At the lowest stage, the total estimator is the value $\mathbf{y}$ of element k_p. Corresponding to equation (5.1) , be $\pi_{p,ij}$ the second-order inclusion probability, that is, the inclusion probability of i_p and j_p being in the sample simultanuously:

$$\pi_{p,ij} \quad = \quad P(i_p \in \mathcal{S} \cap j_p \in \mathcal{S}) \quad . \tag{5.3}$$

The variance of the total estimator $\hat{\tau}_{p,k}(\mathbf{y})$ can be estimated by:

$$\begin{aligned}\widehat{V}\left(\hat{\tau}_{p,k}(\mathbf{y})\right) \;=\;& \sum_{i \in q_{k_p}} \sum_{j \in q_{k_p}} \left(1 - \frac{\pi_{(p+1),i} \cdot \pi_{(p+1),j}}{\pi_{(p+1),ij}}\right) \cdot \frac{\hat{\tau}_{(p+1),i}(\mathbf{y})}{\pi_{(p+1),i}} \cdot \frac{\hat{\tau}_{(p+1),j}(\mathbf{y})}{\pi_{(p+1),j}} \\ & + \sum_{i \in q_{k_p}} \frac{\widehat{V}\left(\hat{\tau}_{(p+1),i}(\mathbf{y})\right)}{\pi_{(p+1),i}} \end{aligned} \qquad (5.4)$$

(cf. Cochran, 1963, Demnati and Rao, 2004, Deville and Särndal, 1992, Eurostat, 2013, Horvitz and Thompson, 1952, Münnich and Zins, 2011, Wolter, 2007).

To estimate ARPR, Gini and QSR, we start with the distribution function $F\left(x\right) = p(\mathbf{y} \leq x)$ of a variable $\mathbf{y}$. Its design-based estimate following from equation (5.2) is

$$\widehat{F}\left(x\right) \;=\; \frac{\sum\limits_{i=1}^{n} \mathbb{I}\left(y_i \leq x\right) \cdot \pi_i^{-1}}{\sum\limits_{i=1}^{n} \pi_i^{-1}} \qquad , \qquad . \qquad (5.5)$$

Quantiles can be estimated by its inverse

$$\widehat{F}^{-1}\left(p\right) \;=\; \inf_q \left\{q \;:\; \widehat{F}\left(q\right) \geq p\right\} \quad , \qquad 0 \leq p \leq 1 \quad . \qquad (5.6)$$

A design-based estimate of the Lorenz curve is then given by

$$\widehat{L}\left(p\right) \;=\; \frac{\sum\limits_{i=1}^{n} \mathbb{I}\left(y_i \leq \widehat{F}^{-1}\left(p\right)\right) \cdot y_i \cdot \pi_i^{-1}}{\sum\limits_{i=1}^{n} y_i \cdot \pi_i^{-1}} \qquad . \qquad (5.7)$$

The **at-risk-of-poverty rate** is the share of persons with an equivalized disposable income below a certain at-risk-of-poverty threshold (ARPT). For the EU, this threshold is commonly set to 60% of the median income in the reference population, other possible definitions are discussed in Atkinson et al. (2005). Using equations (5.5) and (5.6), the at-risk-of-poverty is estimated as

$$\widehat{\mathrm{ARPR}} \;=\; \widehat{F}\left(0.6 \cdot \widehat{F}^{-1}\left(0.5\right)\right) \quad . \qquad (5.8)$$

Linearized values, which can be used for variance estimation via equation (5.4) are

$$\begin{aligned} z_{k,\widehat{\mathrm{ARPR}}} &= \frac{1}{N}\left(\mathbb{I}(y_k \leq 0.6 \cdot \widehat{F}^{-1}(0.5)) - \mathrm{ARPR}\right) \\ &\quad - \frac{0.6 \cdot \widehat{F}'\left(0.6 \cdot \widehat{F}^{-1}(0.5)\right)}{N \cdot \widehat{F}'\left(\widehat{F}^{-1}(0.5)\right)} \cdot \left(\mathbb{I}\left(y_k \leq \widehat{F}^{-1}(0.5)\right) - \frac{1}{2}\right) \end{aligned} \tag{5.9}$$

for every observation k, where $\mathbb{I}$ is a indicator function (cf. Deville, 1999, Osier, 2009).

The **Gini** is commonly used to express inequality and concentration of wealth regarding the whole income distribution. It is defined through the area between the angle bisector and the Lorenz curve (eqn. 5.7) and a standardization to the range $[0;1]$. Estimation is done by

$$\widehat{\mathrm{GINI}} = 1 - 2 \cdot \int_0^1 \widehat{L}(p)\, dp \tag{5.10}$$

(cf. Morgan, 1962). The corresponding linearization can be done via

$$z_{k,\widehat{\mathrm{GINI}}} = \frac{2}{N\mu}\left(\left(F(y_k) - \frac{\widehat{\mathrm{GINI}}+1}{2}\right)y_k + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(y_i \geq y_k)y_i\right) - \frac{\mu}{2}\left(\widehat{\mathrm{GINI}}+1\right)\right) \tag{5.11}$$

(cf. Kovacevic and Binder, 1997, p. 50).

The **Quintile-Share-Ratio** represents the ratio between the mean income of the richest and the poorest 20%. Thus it is primarily based on the tails of the distribution. The mean is used instead of the total to avoid instabilities due to varying estimated case numbers corresponding to the quantiles. Using equation (5.7) , the estimate is

$$\widehat{\mathrm{QSR}} = \frac{1 - \widehat{L}(0.8)}{\widehat{L}(0.2)} \tag{5.12}$$

(cf. Eurostat, 2003, Social Protection Committee, 2001) with linearized values

$$z_{k,\widehat{\mathrm{QSR}}} = \frac{y_k - \left(\left(y_k - \widehat{F}^{-1}(0.8)\right) \cdot \mathbb{I}\left(y_k \leq \widehat{F}^{-1}(0.8)\right) + 0.8 \cdot \widehat{F}^{-1}(0.8)\right)}{\widehat{L}(0.2)} - \frac{\widehat{\mathrm{QSR}} \cdot \left(\left(y_k - \widehat{F}^{-1}(0.2)\right) \cdot \mathbb{I}\left(y_k \leq \widehat{F}^{-1}(0.2)\right) + 0.2 \cdot \widehat{F}^{-1}(0.2)\right)}{\widehat{L}(0.2)} \tag{5.13}$$

(cf. Hulliger and Münnich, 2006, p. 3155).

Those variance approximations by linearization are constructed assuming continuous incomes. Up to date, income classification can not be taken into account. In this case, merely *naive linearizations* can be used. This means equations (5.9), (5.11) and (5.13) are used, even though their applicability for estimating from income classes seems questionable.

In overcoming this problem, resampling methods seem more suitable (cf. Münnich, 2008). As the variance of point estimates can not in general be given in a closed form (cf. Bruch et al., 2011, p. 2), those approaches approximate it by Monte Carlo methods and are applicable even for estimators where no linearization is available. From this class of variance estimation methods, the Monte Carlo bootstrap (cf. Efron, 1979, Shao and Tu, 1995a) and the rescaling bootstrap (cf. Rao and Wu, 1988) are used here.

From a sample with T stages of size n_t on stage t

$$\mathcal{S} = \left(i_{1,1}, \ldots, i_{n_1,1}, \ldots, i_{1,t}, \ldots, i_{n_t,t}, \ldots, i_{1,T}, \ldots, i_{n_T,T}\right) \quad , \tag{5.14}$$

an estimate $\hat{\theta}$ is a function of this sample

$$\hat{\theta} = T(\mathcal{S}) \quad . \tag{5.15}$$

In the Monte Carlo bootstrap, B independent subsamples are drawn according to the original design, but *with replacement*

$$\mathcal{S}_b^{MC} = \left(i^M_{1,1}, \ldots, i^M_{n_1,1}, \ldots, i^M_{1,t}, \ldots, i^M_{n_t,t}, \ldots, i^M_{1,T}, \ldots, i^M_{n_T,T}\right), \; b = 1, \ldots, B. \tag{5.16}$$

Problems when applying this to SILC arise from this approach are the i.i.d-assumption and accounting for design weights. The rescaling bootstrap is a solution for these. Here, subsamples of size m_t at stage t are drawn without replacement:

$$\mathcal{S}_b^{Rsc} = \left(i^R_{1,1},\dots,i^R_{m_1,1},\dots,i^R_{1,t},\dots,i^R_{m_t,t},\dots,i^R_{1,T},\dots,i^R_{m_T,T}\right),\ b = 1,\dots,B. \tag{5.17}$$

As a sampling fraction, $m_{i_p}/n_{i_p} = 0.5$ in sampling unit i_p (unit i on sampling stage p) is common. Let g_s be the sampling unit on stage s containing i_p and $\delta_{i_p,b} := \sum\limits_{k=1}^{m_p} \mathbb{I}\left(i_p = i^R_{k,p}\right)$ be the number of times that i_p is selected in bootstrap iteration b. Based on this, the original design weights $w_{i_p} = \pi^{-1}_{i,p}$ of unit i_p in this iteration are rescaled by

$$w^*_{i_p,b} = \begin{cases} \left(1 + \lambda_{g_{1,b}} \cdot \left(\frac{n_{g_1}}{m_{g_1}}\delta_{g_{1,b}} - 1\right)\right) \cdot w_{i_p} & ,p = 1 \\ \left(1 + \lambda_{g_{1,b}} \cdot \left(\frac{n_{g_1}}{m_{g_1}}\delta_{g_{1,b}} - 1\right)\right. & \\ \quad \left. + \left(\sum\limits_{s=2}^{p} \lambda_{g_{s,b}} \left(\prod\limits_{t=1}^{s-1} \sqrt{\frac{n_{g_t}}{m_{g_t}}} \delta_{g_{t,b}}\right) \cdot \left(\delta_{g_{s,b}} \frac{n_{g_s}}{m_{g_s}} - 1\right)\right)\right) \cdot \left(\prod\limits_{s=1}^{p-1} \frac{w_{g_s}}{w^*_{g_{s,b}}}\right) \cdot w_{i_p} & ,p > 1, \end{cases} \tag{5.18}$$

where

$$\lambda_{g_s} = \begin{cases} \sqrt{m_{g_s} \cdot \frac{1 - n_{g_s}/N_{g_s}}{n_{g_s} - m_{g_s}}} & ,\ s = 1 \\ \sqrt{m_{g_s} \cdot \frac{1 - n_{g_s}/N_{g_s}}{n_{g_s} - m_{g_s}} \cdot \left(\prod\limits_{t=1}^{s-1} \frac{n_{g_t}}{N_{g_t}}\right)} & ,\ s > 1 \ . \end{cases} \tag{5.19}$$

In contrast, the Monte Carlo bootstrap can be interpreted as rescaling weights by to the number of times i_p is drawn

$$w^*_{i_p,b} = w_{i_p} \cdot \sum_{k=1}^{n_p} \mathbb{I}\left(i_p = i^M_{k,p}\right) \quad . \tag{5.20}$$

Using equation (5.15) with the respective B sets of rescaled weights produces the B bootstrap statistics in each case, $\left(\left(\hat{\theta}^{MC}_1,\dots,\hat{\theta}^{MC}_B\right) \text{ and } \left(\hat{\theta}^{Rsc}_1,\dots,\hat{\theta}^{Rsc}_B\right)\right)$. The variance of the point-estimator can be approximated by the Monte Carlo variance of the bootstrap statistics:

$$\widehat{V}_{BT}\left(\hat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b\right)\right)^2 \quad . \tag{5.21}$$

By means of equations (5.5) and (5.6), confidence intervals (CIs) can als be computed directly from the quantiles of the bootstrap statistics (cf. Preston, 2009, Rao and Wu, 1988, Rao et al., 1992, Shao and Tu, 1995b)). For the designs used in the simulation, the MC-bootstrap has to be nested within the strata or selected PSUs, respectiveley.

5.3 Problems estimating poverty and social exclusion from income classes

Not all indicators on poverty and social exclusion in Europe are based on the EU-SILC. The prerequisite of a continuous measurement of income is not necessarily met by those other surveys. German federal statistical offices and charities, for example, make regular use of the German Microcensus to estimate those indicators, particularly for (regional) ARPRs. The reasons for this are presumably the case number and duty of disclosure, as then again this survey has the serious disadvantage of an income variable which is classified into 24 groups (cf. Deutscher Paritätischer Wohlfahrtsverband, 2009, 2013).

We assume a general classification into L income classes with boundaries $(t_1, \ldots, t_{L+1})$. Thus, for the j-th class, only the sample units belonging to and the boundaries $(t_j,\, t_{j+1})$ of the class are known. Hence, the distribution functional values of income variable y can only be estimated at the class boundaries ($\widehat{F}\,(t_j)$ and $\widehat{F}\,(t_{j+1})$), using survey weights if necessary. What remains unknown is the distribution within the income classes. Figures 5.1 and 5.2 should give a first impression on how severe this loss of information might be. Depending on the choice of income classes in terms of the number of classes and class boundaries, we obtain different levels of inaccuracy due to classification.

Selecting equidistant boundaries results in equal classwidth along the entire distribution. In contrast, with ascending distances, the classwidth increases the higher the income. As higher income values are less frequent than lower ones, this results in less varying class frequencies compared to the equidistant case. The highest income class is unbounded from above to cover all possible income values. The scheme of 24 classes of ascending width coincides with the one used in the German Microcensus.

While the red line indicates the true values resulting from the continuous income, the green area illustrates possible scenarios which would lead to the *the same information on income classes*. As already mentioned, precise distribution function values can be calculated directly at least at the class boundaries. However, this is not the case for the Lorenz curve, since the average income of every income class would be needed. Thus, even

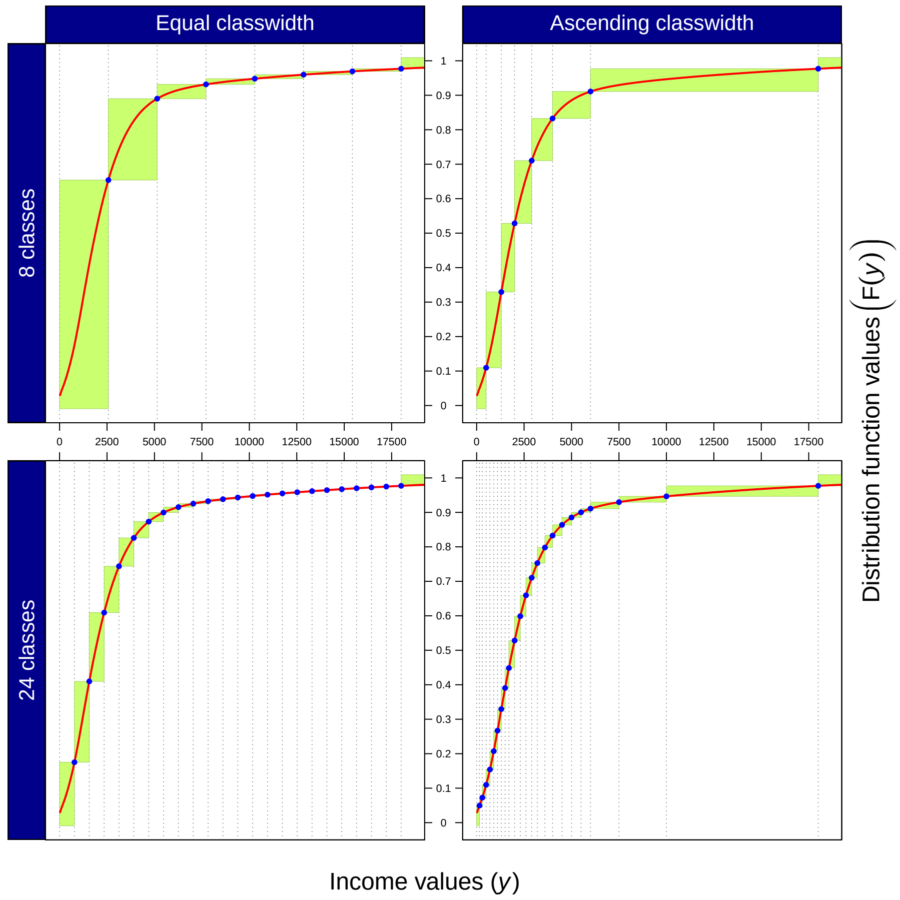


Figure 5.1: Information loss due to classification (distribution)

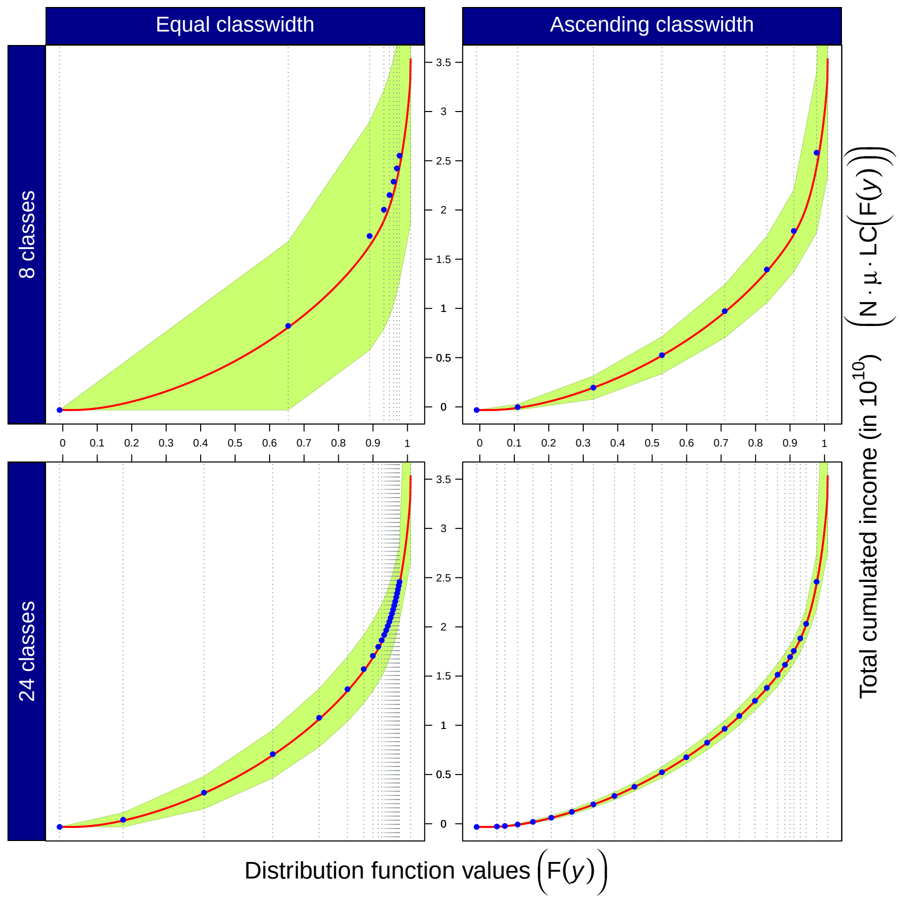


Figure 5.2: Information loss due to classification (Lorenz curve)

🟥 True Values

🟩 Possible range, given the income classes

⋯ Class boundaries

● Values at class boundaries

(Approximation by using class center as expected value for each class in case of the Lorenz curve) at the boundaries, the values can only be approximated. Using class centers as expectation for each income class results in the displayed points. Please note that non-standardized Lorenz curves are presented, as standardization would result in different scalings of area, points and true curve.

5.4 Possible solutions for estimations from income classes

Due to classification, information about the distribution within the classes is lost. Hence, we can not use equations (5.5) to (5.7) directly. In fact, additional assumptions or external informations are needed to perform point- and variance estimation for poverty indicators (cf. Stauder and Hüning, 2004, Strengmann-Kuhn, 1999).

German statistical offices thus use linear interpolation of the distribution function.

Since this approach is mainly used for estimating the ARPR, further options are proposed and compared hereto. These include non-parametric as well as parametric approximation. The three approaches are presented in Sections 5.4.1 to 5.4.3.

5.4.1 Linear interpolation of the distribution function

As a first approach used by German statistical offices, the known values of the distribution function are interpolated linearly. For the distribution function, a straight line is assumed within the income class j between $\widehat{F}(t_j)$ and $\widehat{F}(t_{j+1})$:

$$\widehat{F}(y_u) = \widehat{F}(t_j) + \frac{y_u - t_j}{t_{j+1} - t_j} \cdot \left(\widehat{F}(t_{j+1}) - \widehat{F}(t_j)\right) \quad , \qquad t_j \leq y_u < t_{j+1} \quad . \tag{5.22}$$

This corresponds to an assumed uniform distribution within the classes (cf. Information und Technik Nordrhein-Westfalen, 2009). As linear interpolation is a special case of spline-interpolation, the estimation methodology coincices with the one presented in Section 5.4.2.

Primary aim of this approach was estimating the ARPR, which anyways is only affected by information loss in two income classes – containing the median and the poverty threshold – and based on a rather stable area of the income distribution. The question whether this approach is transferable to other indicators on poverty and social exclusion, is thus one aim of this study.

5.4.2 Non-parametric modelling: Splines

The second goal is to compare different approaches in dealing with income classes. One alternative to this linear interpolation is spline modeling, which is used here similar to empirical likelihood (see Chapter 4 and Berger and Escobar, 2015). The main idea here is to estimate a non-parametric regression function

$$\widehat{y}_j \; = \widehat{m}\left(x_j\right) \quad , \qquad j = 1, \ldots, L \tag{5.23}$$

between independent variable $\mathbf{x}$ and dependent variable $\mathbf{y}$.

B-splines are used here as base function, since they provide greater numerical stability, compared to other base functions. For this base function, so called 'knots' are used. Knots are points breaking down the range of possible values of $\mathbf{x}$ into sub-intervals. Using the class-boundaries $(t_1, \ldots, t_{L+1})$ as knots, B-splines for our purpose can recursively be defined as

$$B_i^0\left(x_j\right) \; = \begin{cases} 1 & , \; x_j \in [t_i, t_{i+1}) \\ 0 & , \; \text{else} \end{cases} \tag{5.24}$$

and

$$B_i^k\left(x_j\right) \; = \frac{x_j - t_i}{t_{i+k} - t_i} \cdot B_i^{k-1}\left(x_j\right) + \frac{t_{i+k+1} - x_j}{t_{i+k+1} - t_{i+1}} \cdot B_{i+1}^{k-1}\left(x_j\right) \qquad , \tag{5.25}$$

where anything divided by zero is set to zero: $\frac{z}{0} := 0 \; \forall \; z$. The degree of the splines is denoted by k, whereas depending on k additional outer knots have to be chosen. We do so by replicating the outermost knots. Spline functions and their derivatives estimated by the following methods are in general continuous up to order $k-1$. In the simulation, we use cubic B-splines ($k = 3$).

The estimated regression function $\widehat{m}$ is a linear combination of B-splines:

$$\widehat{m}\left(x_j\right) \; = \sum_{i=1-k}^{L} \gamma_i \cdot B_i^k\left(x_j\right) \tag{5.26}$$

Integrals and derivatives of B-splines are again B-splines of higher and lower degree, respectively

$$\frac{d}{d\,x}\,B_i^k(x) = \left(\frac{k}{t_{i+k}-t_i}\right)B_i^{k-1}(x) - \left(\frac{k}{t_{i+k+1}-t_{i+1}}\right)B_{i+1}^{k-1}(x) \tag{5.27}$$

$$\int\limits_{-\infty}^{x} B_i^k(s)\,ds = \left(\frac{t_{i+k+1}-t_i}{k+1}\right)\sum_{j=1}^{\infty} B_j^{k+1}(x) \qquad . \tag{5.28}$$

Hence, (anti)derivatives of the regression function can be represented as

$$\frac{d\,\widehat{m}(x)}{d\,x} = \sum_{i=1-k}^{L} k\cdot\frac{\gamma_i-\gamma_{i-1}}{t_{i+k}-t_i}\cdot B_i^{k-1}(x) \tag{5.29}$$

$$\int\limits_{-\infty}^{x} \widehat{m}(s)\,ds = \sum_{i=1-k}^{L}\frac{1}{k+1}\left(\sum_{j=-\infty}^{i}\gamma_j\,(t_{j+k+1}-t_j)\right)\cdot B_i^{k+1}(x) \qquad . \tag{5.30}$$

To estimate $\widehat{m}$, we start with model matrix

$$\mathbf{B} = \begin{bmatrix} B_{1-k}^k(x_1) & \dots & B_L^k(x_1) \\ \vdots & \ddots & \vdots \\ B_{1-k}^k(x_{L+1}) & \dots & B_L^k(x_{L+1}) \end{bmatrix} \qquad . \tag{5.31}$$

In case of interpolating splines, we get

$$\begin{array}{ccccccc} \begin{bmatrix} \widehat{m}(x_1,\boldsymbol{\gamma}) \\ \vdots \\ \widehat{m}(x_{L+1},\boldsymbol{\gamma}) \end{bmatrix} & = & \begin{bmatrix} B_{1-k}^k(x_1) & \dots & B_L^k(x_1) \\ \vdots & \ddots & \vdots \\ B_{1-k}^k(x_{L+1}) & \dots & B_L^k(x_{L+1}) \end{bmatrix} & \begin{bmatrix} \gamma_{1-k} \\ \vdots \\ \gamma_L \end{bmatrix} & \overset{!}{=} & \begin{bmatrix} y_1 \\ \vdots \\ y_{L+1} \end{bmatrix} \\ \widehat{\mathbf{m}} & = & \mathbf{B} & \boldsymbol{\gamma} & \overset{!}{=} & \mathbf{y} \end{array} \tag{5.32}$$

Since $\mathbf{B}$ is in general not a square matrix, $k-1$ additional conditions are needed to get a unique solution. The assumption of natural splines is used, which demands the second derivatives of $\widehat{\mathbf{m}}$ to be 0 at the outermost used values of $\mathbf{x}$:

$$\widehat{m}''(x_1,\boldsymbol{\gamma}) = \widehat{m}''(x_{L+1},\boldsymbol{\gamma}) = 0 \qquad . \tag{5.33}$$

This forces $\widehat{\mathbf{m}}$ to be linear outside the interval of used knots. Let $\mathbf{B}^*$ be the extension of the matrix **B** including this additional conditions, then equation (5.32) can be solved by

$$\gamma = (\mathbf{B}^{*})^{-\mathbf{1}}\mathbf{y} \qquad . \tag{5.34}$$

As mentioned, linear interpolation (Section 5.4.1) equals spline interpolation of degree $k = 1$.

Smoothing splines does not require the regression function to interpolate all used points, but rather approximates them. Again, we start from regression equation

$$\begin{array}{ccccc} \begin{bmatrix} \widehat{m}(x_1, \boldsymbol{\beta}) \\ \vdots \\ \widehat{m}(x_{L+1}, \boldsymbol{\beta}) \end{bmatrix} & = & \begin{bmatrix} B^k_{1-k}(x_1) & \dots & B^k_L(x_1) \\ \vdots & \ddots & \vdots \\ B^k_{1-k}(x_{L+1}) & \dots & B^k_L(x_{L+1}) \end{bmatrix} & \begin{bmatrix} \beta_{1-k} \\ \vdots \\ \beta_L \end{bmatrix} \\ \widehat{\mathbf{m}} & = & \mathbf{B} & \boldsymbol{\beta} \end{array} \tag{5.35}$$

Now, the coefficients are given by

$$\boldsymbol{\beta} = \left(\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Sigma}\right)^{-1}\mathbf{B}^T\mathbf{y} \qquad , \tag{5.36}$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \int B^{k''}_1(x)\,B^{k''}_1(x)\,dx & \dots & \int B^{k''}_1(x)\,B^{k''}_{L+1}(x)\,dx \\ \vdots & \ddots & \vdots \\ \int B^{k''}_{L+1}(x)\,B^{k''}_1(x)\,dx & \dots & \int B^{k''}_{L+1}(x)\,B^{k''}_{L+1}(x)\,dx \end{bmatrix} \quad . \tag{5.37}$$

This is equivalent to the often used expression

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\underbrace{\sum_{j=1}^{L+1} \left(\widehat{m}(x_j, \boldsymbol{\beta}) - y_j\right)^2}_{\text{OLS-part}} + \underbrace{\lambda \int \widehat{m}''(x, \boldsymbol{\beta})^2\,dx}_{\text{smoothing-part}} \right) \qquad . \tag{5.38}$$

The smoothing reduces differences in the slopes between differend knots, resulting in a regression function which oscillates less than a pure OLS estimate (cf. Cheney and Kincaid, 2012, De Boor, 1978, Eilers and Marx, 1996, Reinsch, 1967, Ruppert et al., 2003, Wahba, 1990). The smoothing parameter λ must be greater than 0 and determines the degree of smoothing. It can either be set or optimized by generalized cross-validation (GCV) (cf. Craven and Wahba, 1978) using

$$\underset{\lambda}{\operatorname{argmin}} \left(\frac{n^{-1} \cdot \left|\left| \Big(\mathbf{I} - \mathbf{A}(\lambda) \Big) \mathbf{y} \right|\right|^2}{\Big(n^{-1} \cdot tr \big(\mathbf{I} - \mathbf{A}(\lambda) \big) \Big)^2} \right) \qquad , \tag{5.39}$$

with identity matrix $\mathbf{I}$ and hat-matrix $\mathbf{A}(\lambda) = \mathbf{B} \left(\mathbf{B}^T \mathbf{B} + \lambda \boldsymbol{\Sigma} \right)^{-1} \mathbf{B}^T$ depending on λ.

Due to their flexibility, splines can be used to model density, distribution- and quantile function as well as the Lorenz curve. Therefore, we again get several estimation approaches. As estimation of quantiles and values of the distribution from classified income is possible *without further assumptions*, those may be subject to interpolation as well as smoothing. In contrast, functional values of the density or Lorenz curve require additional assumptions. For the density, the cumulated probabilites have to be split across the classes. As a first approach, the class centers

$$\widetilde{y}_j = \frac{t_j + t_{j+1}}{2} \qquad , j = 1, \ldots, L \tag{5.40}$$

are assigned an assumed density as if the distribution within the classes was uniform:

$$\widetilde{f}\left(\widetilde{y}_j\right) = \frac{\widehat{F}\left(t_{j+1}\right) - \widehat{F}\left(t_j\right)}{t_{j+1} - t_j} \qquad , j = 1, \ldots, L. \tag{5.41}$$

Calculating values of the Lorenz curve lacks knowledge about the average income of every class. Thus, class centers ($\widetilde{y}_j$ in equation (5.40)) are presumed as expectation for each class.

As those additional assumptions raise uncertainty, for density and Lorenz curve approximate spline smoothing seems more reasonable than interpolation. In summary, distribution and quantile function are estimated by interpolation *and* smoothing, whereas density and Lorenz curve are only estimated by smoothing. For this purpose, the corresponding function arguments are used as independent and the resulting functional values as dependent variable in equations (5.24) to (5.38). Due to the comparably small number of income classes, every x_j-value is used as a knot t_j.

From this estimated functions, the indicators of interest can be computed as shown in

Section 5.2. For this purpose, as the inverse of B-splines (and their (anti)derivatives) are *not* in general B-splines, numerical approximation must be applied where needed.

5.4.3 Parametric Income Distributions

As further alternatives, parametric income distributions seem promising. There is a large variety of potential distributions which have already been evaluated with regard to fitting to income classes (cf. Bandourian et al., 2002, Dagum, 1977, McDonald, 1984). Consistent results indicate the use of a generalized beta distribution of the second kind as well as two of its special cases, the Singh-Maddala- and Dagum-distribution.

The GB2-Distribution is given by

$$F_{GB2}\left(y,a,b,p,q\right) = \frac{\left(\frac{\left(\frac{y}{b}\right)^{a}}{1+\left(\frac{y}{b}\right)^{a}}\right)^{p}}{pB(p,q)} \cdot {}_{2}F_{1}\left[\begin{matrix} p & , 1-q\,; \dfrac{\left(\frac{y}{b}\right)^{a}}{1+\left(\frac{y}{b}\right)^{a}} \\ p+1 & ; \end{matrix}\right] \quad , \quad y>0 \quad , \tag{5.42}$$

with

$$B\left(a\ ,\ b\right) = \frac{\Gamma\left(a\right)\Gamma\left(b\right)}{\Gamma\left(a+b\right)} \quad , \tag{5.43}$$

$$\Gamma\left(x\right) = \int\limits_{0}^{\infty} t^{x-1}e^{-t}dt \quad , \quad \Re(x)>0 \qquad \text{and} \tag{5.44}$$

$${}_{p}F_{q}\left[\begin{matrix} a_1\ ,\dots,\ a_p\ ;\ x \\ b_1\ ,\dots,\ b_q\ ; \end{matrix}\right] = \sum_{i=0}^{\infty}\frac{\left(a_1\right)_i\dots\left(a_p\right)_i}{\left(b_1\right)_i\dots\left(b_q\right)_i}\cdot\frac{x^i}{i!} \quad , \tag{5.45}$$

where $(a)_i$ is the Pochhammer symbol

$$\left(a\right)_i = \prod_{j=0}^{i-1}\left(a+j\right) = \frac{\Gamma\left(a+i\right)}{\Gamma\left(a\right)} \quad , \qquad i\in\mathbb{N} \qquad . \tag{5.46}$$

As a larger numbers of free parameters (4 for the GB2) increases flexibility but also instability, the Singh-Maddala (SM, $p = 1$) and Dagum-distribution (DA, $q = 1$) are also taken into account (cf. Abramowitz and Stegun, 1964, Bandourian et al., 2002, Dagum, 1977, McDonald, 1984, Singh and Maddala, 1976, Seger, 2015).

There are different approaches to estimate the vector of free parameters $\mathbf{\Theta}$ from a sample with income classes. We can only estimate distribution $\left(\widehat{F}\left(t_j\right)\right)$ and quantile $\left(\widehat{F}^{-1}\left(\widehat{F}\left(t_j\right)\right)=t_j\right)$ values for the class boundaries $\left(t_1,\ldots,t_{L+1}\right)$ directly. Denoting the parametric counterparts, depending on parameter vector $\mathbf{\Theta}$, by $\widehat{F}_{\mathbf{\Theta}}\left(t_j\right)$ and $\widehat{F}_{\mathbf{\Theta}}^{-1}\left(\widehat{F}\left(t_j\right)\right)$, fitting can be done to (cumulative) classwidths or -frequencies as well as maximum-likelihood:

Fitting-Method **F**: Fitting to distribution values

$$\mathbf{\Theta}_F = \underset{\mathbf{\Theta}}{\operatorname{argmin}} \left(\sum_{j=1}^{L+1} \left(\frac{\widehat{F}\left(t_j\right)-\widehat{F}_{\mathbf{\Theta}}\left(t_j\right)}{\widehat{F}\left(t_j\right)} \right)^2 \right) \tag{5.47}$$

Fitting-Method **Q**: Fitting to quantiles

$$\mathbf{\Theta}_Q = \underset{\mathbf{\Theta}}{\operatorname{argmin}} \left(\sum_{j=1}^{L+1} \left(\frac{t_j-\widehat{F}_{\mathbf{\Theta}}^{-1}\left(\widehat{F}\left(t_j\right)\right)}{t_j} \right)^2 \right) \tag{5.48}$$

Fitting-Method **ΔF**: Fitting to frequencies

$$\mathbf{\Theta}_{\Delta F} = \underset{\mathbf{\Theta}}{\operatorname{argmin}} \left(\sum_{j=1}^{L} \left(\left(\Delta\widehat{F}\left(t_j\right)\right) - \left(\Delta\widehat{F}_{\mathbf{\Theta}}\left(t_j\right)\right) \right)^2 \right) \tag{5.49}$$

Fitting-Method **ΔQ**: Fitting to classwidths

$$\mathbf{\Theta}_{\Delta Q} = \underset{\mathbf{\Theta}}{\operatorname{argmin}} \left(\sum_{j=1}^{L} \left(\left(\Delta t_j\right) - \left(\Delta\widehat{F}_{\mathbf{\Theta}}^{-1}\left(\widehat{F}\left(t_j\right)\right)\right) \right)^2 \right) \quad , \tag{5.50}$$

where Δ is the difference operator (e.g. $\Delta\widehat{F}\left(t_j\right)=\widehat{F}\left(t_{j+1}\right)-\widehat{F}\left(t_j\right)$).

Fitting-Method **ML**: Maximum-Likelihood-Estimation

$$\mathbf{\Theta}_{ML} = \underset{\mathbf{\Theta}}{\operatorname{argmax}} \left(\widehat{N}! \prod_{j=1}^{L} \frac{\left(\Delta\widehat{F}_{\mathbf{\Theta}}\left(t_j\right)\right)^{\widehat{N}_j}}{\widehat{N}_j!} \right) \tag{5.51}$$

with estimated (sub-)population sizes $\widehat{N}=\sum\pi_i^{-1}$ and $\widehat{N}_j = \widehat{N}\cdot\Delta\widehat{F}\left(t_j\right)$.

As these equations can not generally be solved analytically, numerical procedures are applied (cf. Bandourian et al., 2002, Dagum, 1977, McDonald, 1984).

From these solutions, the indicators of interest can be computed by

$$\mathrm{ARPR}_{GB2} = F_{GB2}\left(0.6 \cdot b \cdot \left(\frac{z_\alpha}{1-z_\alpha}\right)^{\frac{1}{a}}, a, b, p, q\right) \tag{5.52}$$

where z_α is the α-quantile of the beta(p, q)-distribution,

$$\begin{aligned}\mathrm{GINI}_{GB2} = \frac{B\left(2q-\frac{1}{a}\ ,\ 2p+\frac{1}{a}\right)}{B\left(p\ ,\ q\right)B\left(p+\frac{1}{a}\ ,\ q-\frac{1}{a}\right)} & \left\{\left(\frac{1}{p}\right)\ {}_3F_2\left[\begin{matrix}1 & , p+q & , 2p+\frac{1}{a}\ ; 1\\ p+1\ , 2\left(p+q\right) & & ;\end{matrix}\right]\right.\\ & \left. -\left(\frac{1}{p+\frac{1}{a}}\right)\ {}_3F_2\left[\begin{matrix}1 & , p+q & , 2p+\frac{1}{a}\ ; 1\\ p+\frac{1}{a}+1\ , 2\left(p+q\right) & & ;\end{matrix}\right]\right\}\end{aligned} \tag{5.53}$$

and

$$\mathrm{QSR}_{GB2} = \frac{1-F_{GB2}\left(y_{0.8}, a, b, p+\frac{1}{a}, q-\frac{1}{a}\right)}{F_{GB2}\left(y_{0.2}, a, b, p+\frac{1}{a}, q-\frac{1}{a}\right)} \tag{5.54}$$

(cf. Graf et al., 2011a, McDonald, 1984).

5.5 Monte-Carlo-Simulation

5.5.1 Setup of the simulation study

The methodology presented in Section 5.4 enables estimation of indicators of poverty and social exclusion from surveys whith income classes, taking into account general sampling designs. Evaluation and comparison of these methods is done by means of a Monte Carlo (MC) simulation study. The population is the synthetic AMELIA dataset, described in Section 1. The four classification options presented in Section 5.3 are used, as well as a reference case where continuous income values are available. The sampling designs used are listed in Table 5.1 and correspond to those used in the AMELI project (cf. Hulliger

Table 5.1: Sampling designs

		Stage 1				Stage 2		
ID	**PSU**	**Strata**	$\pi_{1,i}$	**fr$_1$**	**Allo-cation**	**SSU**	$\pi_{2,i}$	**fr$_2$**
1.2	HID	–	srs	0.16%	–	–	–	–
1.4a	HID	NUTS2	srs	0.16%	prop.	–	–	–
2.7	CIT	NUTS2 × DOU	srs	16.0%	prop.	HID	srs	1%

srs: simple random sampling without replacement

fr$_1$,**fr$_2$**: sampling fractions **prop.**: proportional $\pi_{1,i}$, $\pi_{2,i}$: sample inclusion probabilites

HID: Household identifier **CIT**: municipality identifier **DOU**: degree of urbanization

et al., 2011).

According to these designs, in $R = 10\ 000$ MC iterations samples covering 0.16% of the housholds are drawn. For each sample, estimation according on the presented methodology is performed. Table 5.2 gives an overview of the used methods and serves as legend for the following figures.

Table 5.2: Overview of used methods

Color / Symbol			**Abbreviation**	**Description**	
	■		**Ct.**	Estimation from continuous income variable	
	●		**Lin. Int.**	Linear interpolation of the distribution function	
GB2	**DA**	**SM**		Fitting method of parametric distributions	
■	■	■	F	. . . to distribution values	
●	●	●	Δ F	. . . to quantiles	
▲	▲	▲	Q	. . . to class frequencies	
■	■	■	Δ Q	. . . to classwidths	
●	●	●	ML	. . . by maximum-likelihood	
			Splines		
			Int.	Spline-interpolation	
	■		F	. . . of the distribution function	
	●		Q	. . . of the quantile function	
			Smoothing	Spline-smoothing	
	▲		Density	. . . of the density, where	λ_1 fulfills eqn. 5.39
	■			and	$\lambda_2 = 6.610706 \cdot 10^{-5}$
	●		F	. . . of the distribution, where	λ_1 fulfills eqn. 5.39
	▲			and	$\lambda_2 = 6.978335 \cdot 10^{-1}$
	■		Q	. . . of the quantiles, where	λ_1 fulfills eqn. 5.39
	●			and	$\lambda_2 = 6.610706 \cdot 10^{-5}$
	▲		LC	. . . of the Lorenz curve, where	λ_1 fulfills eqn. 5.39
	■			and	$\lambda_2 = 4.511080 \cdot 10^{-1}$

5.5.2 Results of the simulation study

To compare the presented methods under different classification schemes in the most direct way, we focus on design 1.2, which is simple random sampling of households. The other sampling schemes are presented in 5.6. The resulting point estimates of the at-risk-of-poverty rate are plotted in Figure 5.3. Their confidence intervall lengths and coverage rates are displayed in Figure 5.4. Hardly any method under consideration yields reasonable point estimates when using 8 classes of equal width. Bias and variance are almost generally by far to high. The least biased point estimates are obtained by using the Dagum distribution fitted by the ΔQ-Method. It is likewise the only method allowing for reasonable CIs in terms of length and coverage by bootstrapping. However, its variance and thus MSE are still comparativeley high. The lowest MSEs correspond to spline-smoothing of the density using the fixed λ_2 or the linear interpolation of the distribution function. When looking at 25 equidistant class boundaries, we get considerably better results than before for quite a number of methods. Some of them perform even better than the reference case (estimation based on continuous income). The parametric Singh-Maddala distribution fitted by ML, the Dagum distribution using the ΔF- and the GB2 distribution fitted by F-,ΔF- or ML-Method as well as the spline-interpolation of the distribution function all have MSEs smaller than this reference case. Also, the spline interpolation and smoothing of the quantile function using fixed λ seem able to compete. The GB2 ML-method even achieves zero bias, thus being even better than estimation from continuous income. Its inference using the MC- or rescaling bootstrap works pretty well. This also applies for all mentioned spline-estimates as well as the SM ML estimates. Turning to 8 classes of ascending width, results in rather smaller variances compared to equal classwidth, but does not necessarily reduce biases. The density smoothing with fixed λ performs best in terms of bias and MSE. It is closely followed by spline interpolation of the distribution function and the ΔQ-fitting of the Singh-Maddala distribution. However, inference works better for the latter, parametric approach. The classification scheme used by the German Microcensus tends to reduce bias, but not variance compared to the 8 classes before. Also, it seems clearly in favor of the nonparametric approaches. In terms of bias and MSE, the linear interpolation as well as spline interpolation and smoothing of the distribution function (using optimized λ) perform well. The same hold for the smoothing of density, Lorenz curve and quantile function using fixed λ, as well as the spline interpolation of the latter. Inference using resampling as well as naive linearization works quite well for all

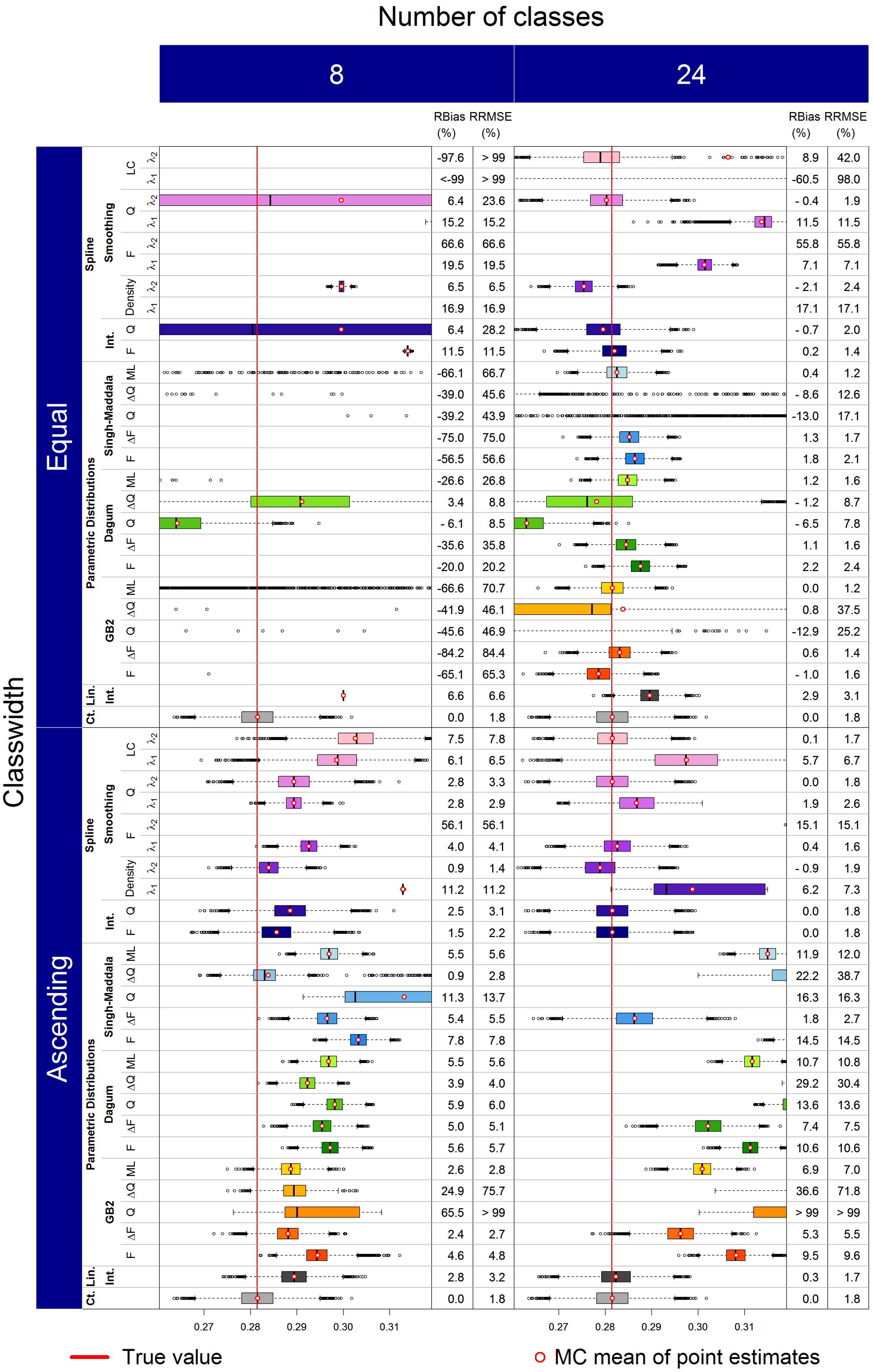


Figure 5.3: ARPR: Point estimates for design 1.2

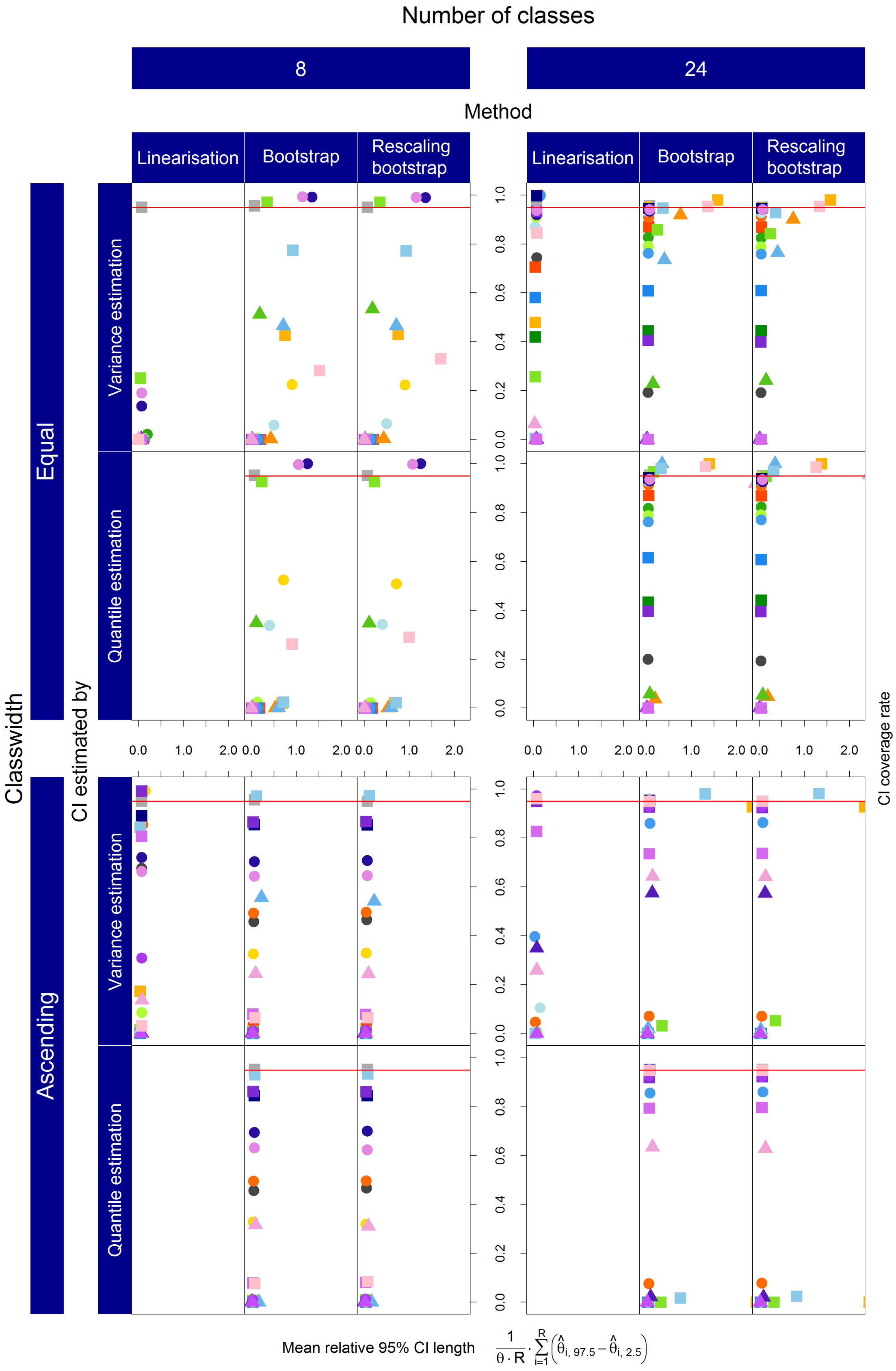


Figure 5.4: ARPR: Confidence interval length and coverage

those methods.

Figures 5.5 and 5.6 illustrate the according GINI point estimates and confidence intervals. Looking at 8 equally spaced classes, bias and variance are almost generally by far to high again. All parametric F-fittings perform comparabely well in bias and MSE. However, their CIs cover less than the nominal rate and thus yield no satisfactory inference. Compared to this, the 24 equally spaced classes tend to decrease the variances, but not biases of the approaches under consideration. The best options in this case, regarding bias and MSE, are the F-fitting of the SM and the ML-fit of the GB2. But again, the inference works not too good for both of them, staying behind the nominal CI coverage rate. Turning to 8 classes of ascending width, again, variances but not biases are reduced compared to equal classwidth. The ML estimate of the GB2 performs well in bias and MSE, while its Q-fit, the ΔQ-method of the Singh-Maddala distribution and the spline-smoothing of the distribution using fixed λ only do so for bias or MSE, respectiveley. Inference via resampling works well for all of the above methods with low bias, namely the parametric ones. In case of 24 ascending classes, neither biases nor variances tend to be smaller than in the case of 8. There is no method which is best in bias and MSE, the MSE is best for ML-estimation of the GB2, while bias is smallest for the ΔF-fit of the SM. However, the CIs for both do not meet the nominal level, even though estimating quantiles by bootstrapping performs not too bad for the latter one.

The corresponding point estimations for the quintile share ratio and their confidence intervals are depicted in figures 5.7 and 5.8. In case of 8 classes of equal width, again most methods suffer from high bias or variance. The lowest biases are achieved by spline-interpolation of the quantile function and the GB2 fit by Q. In terms of MSE, the spline-interpolation of the distribution works best. Regarding the CIs, the latter performs worst due to its bias, which is not represented in variance estimation. On the other hand, the estimates with lower bias obtain better inference but, due to their immense variances, at expense of very long CIs (considerably more than 200% mean relativ CI length, which is why they are not displayed in the plot). In case of 24 equally spaced classes, the variances tend again to be smaller than with 8 classes. Spline-smoothing of the quantile function using λ set by GCV has the lowest bias and MSE. Its inference using resampling works quite well, especially with the CIs based on variance estimation. 8 classes of unequal width, tend to decrease variance and bias compared to those with equal width. Still, results are not too promising. The GB2 distribution performs best, using ΔQ-fit results in the lowest

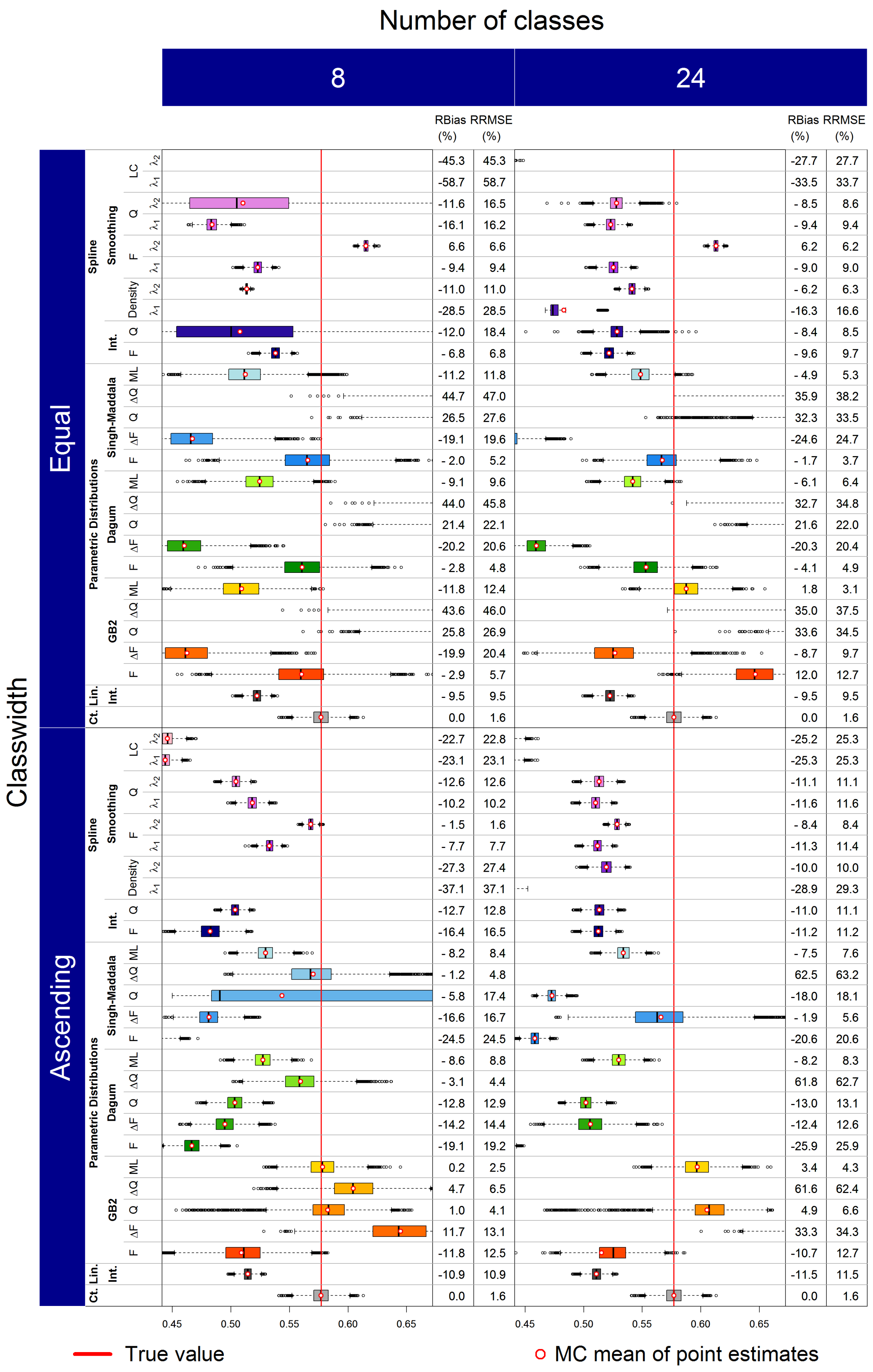


Figure 5.5: GINI: Point estimates for design 1.2

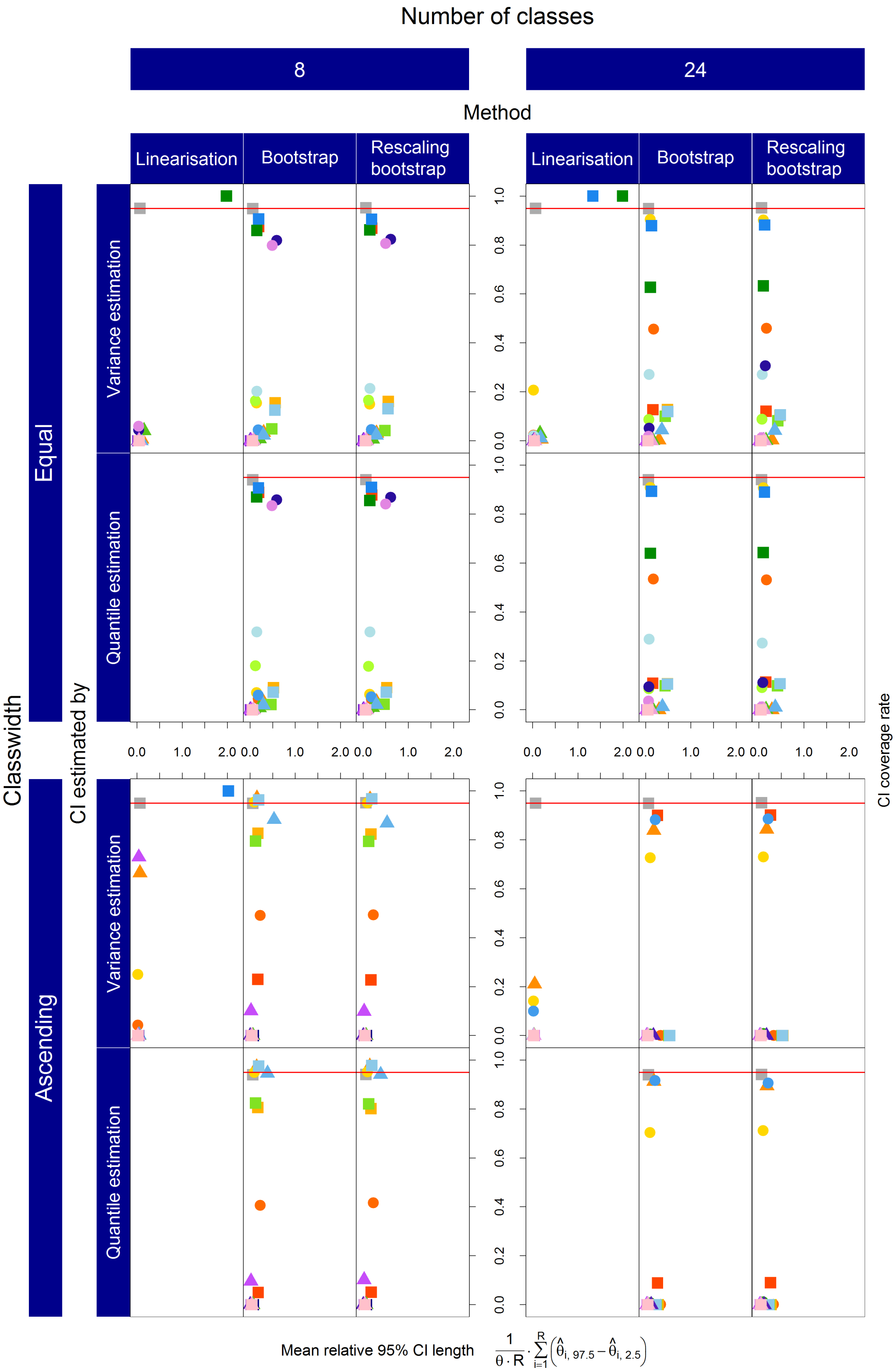


Figure 5.6: GINI: Confidence interval length and coverage

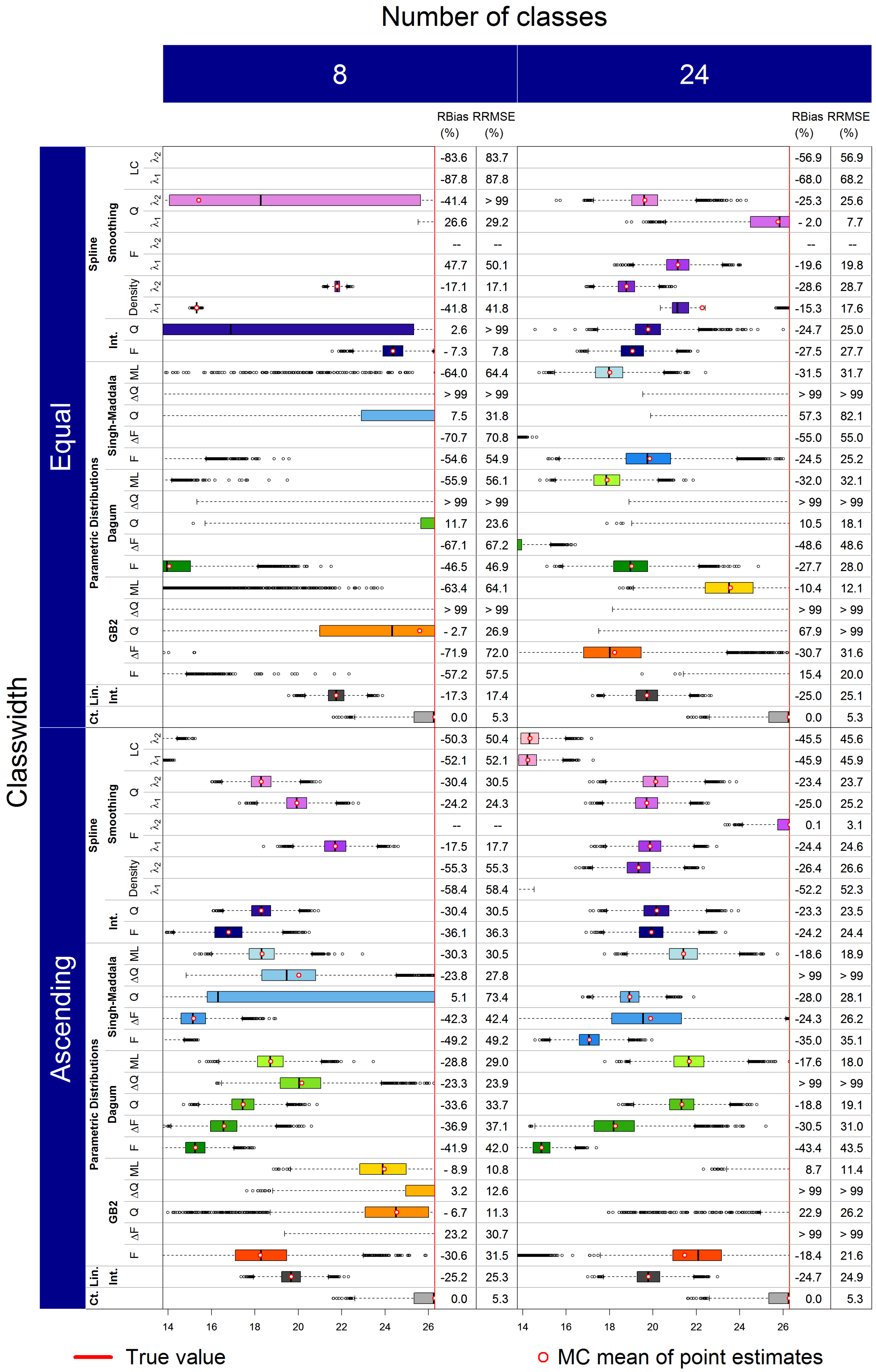


Figure 5.7: QSR: Point estimates for design 1.2

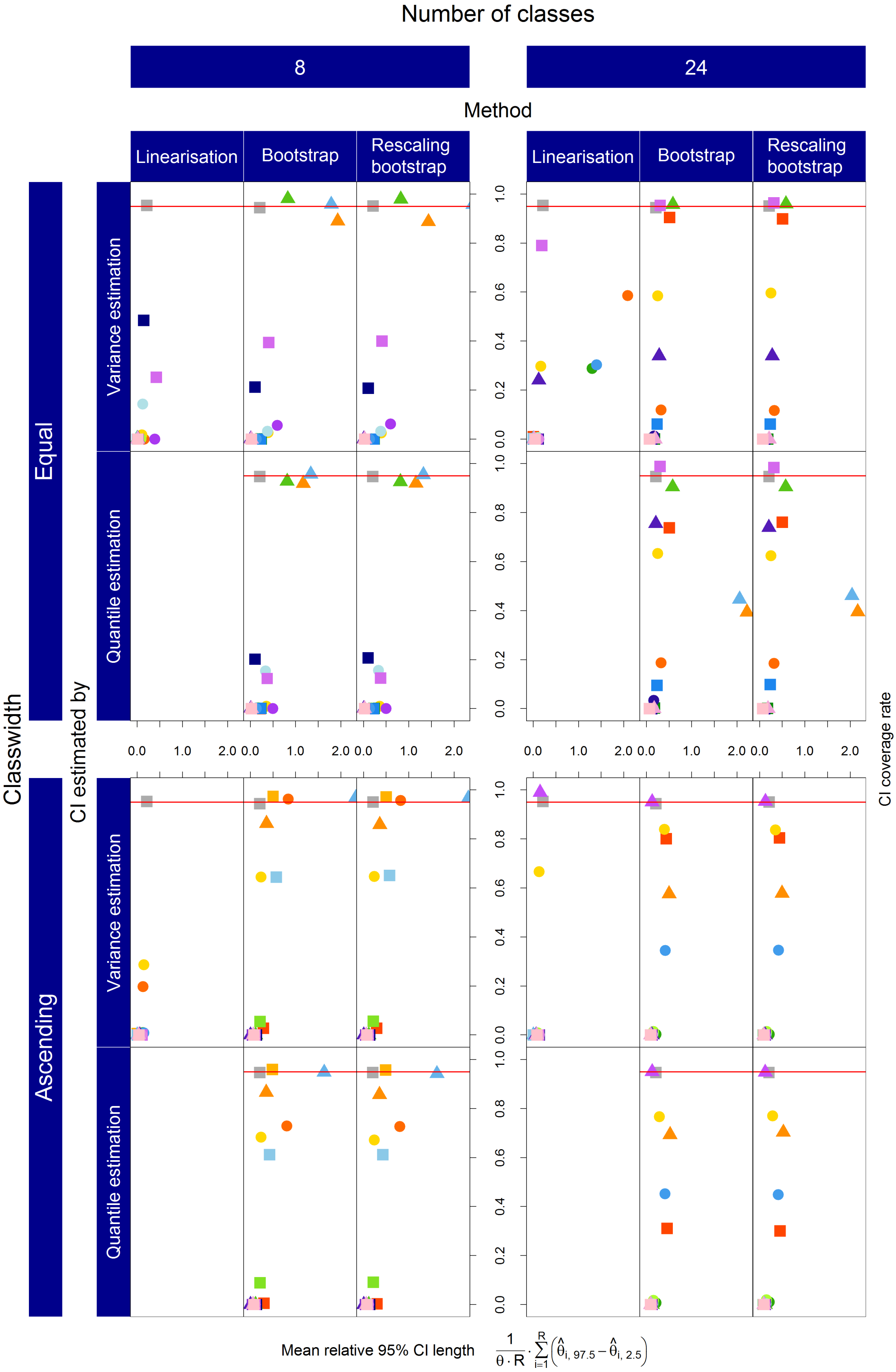
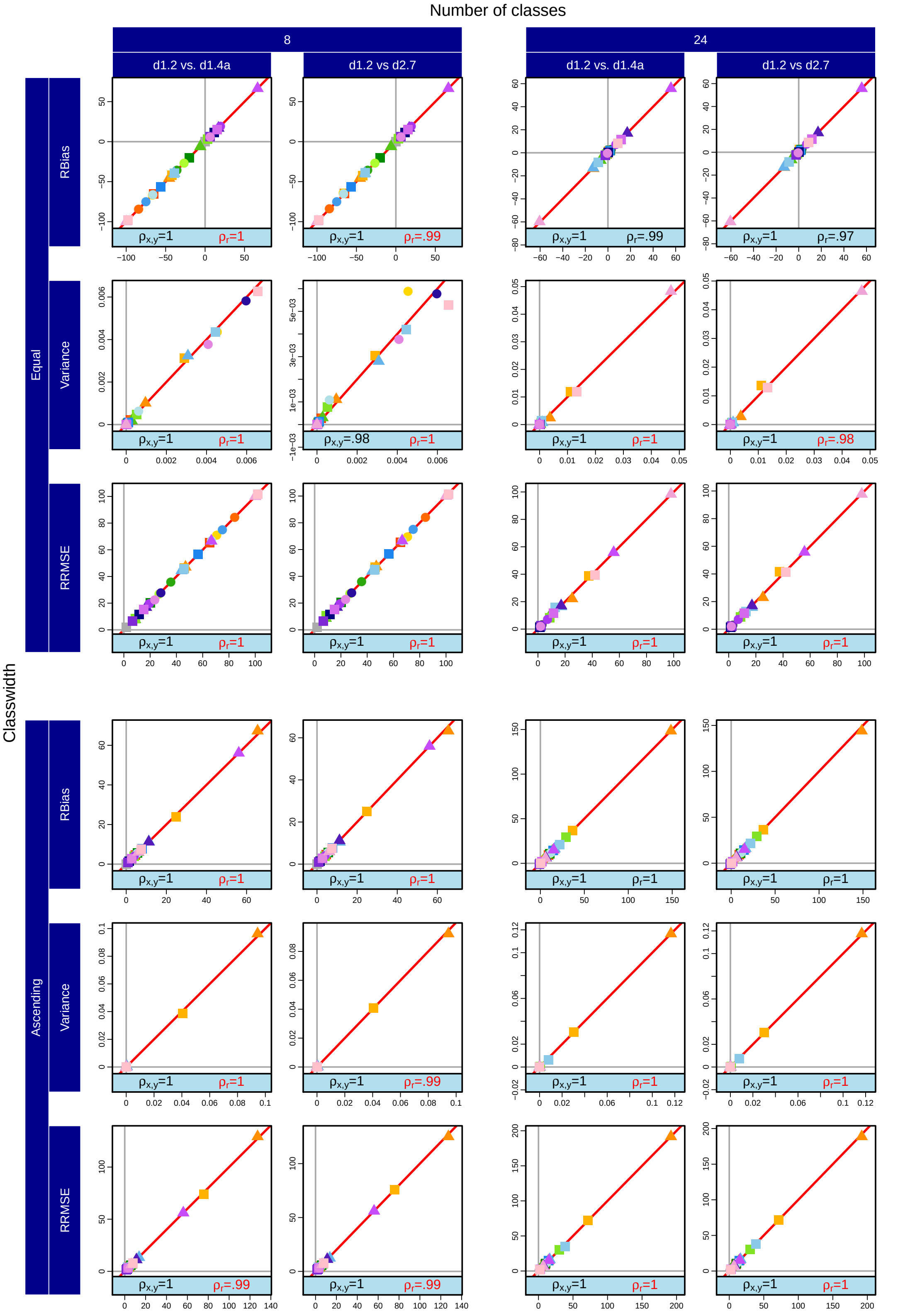


Figure 5.8: QSR: Confidence interval length and coverage

bias, while fitting methods Q and ML provide lower MSEs. Again, out of these, inference only works for the approach with low bias, since the CIs do not account for bias. Turning to the classification scheme used in German Microcensus, neither bias nor variance seem to clearly decrease when compared to 8 classes. The approach minimizing bias and MSE is the smoothing of the distribution using fixed λ. It performs better than estimation from continuous income and allows for almost perfect inference for all resampling CIs. Only the CIs based on naive linearization tend to be too long.

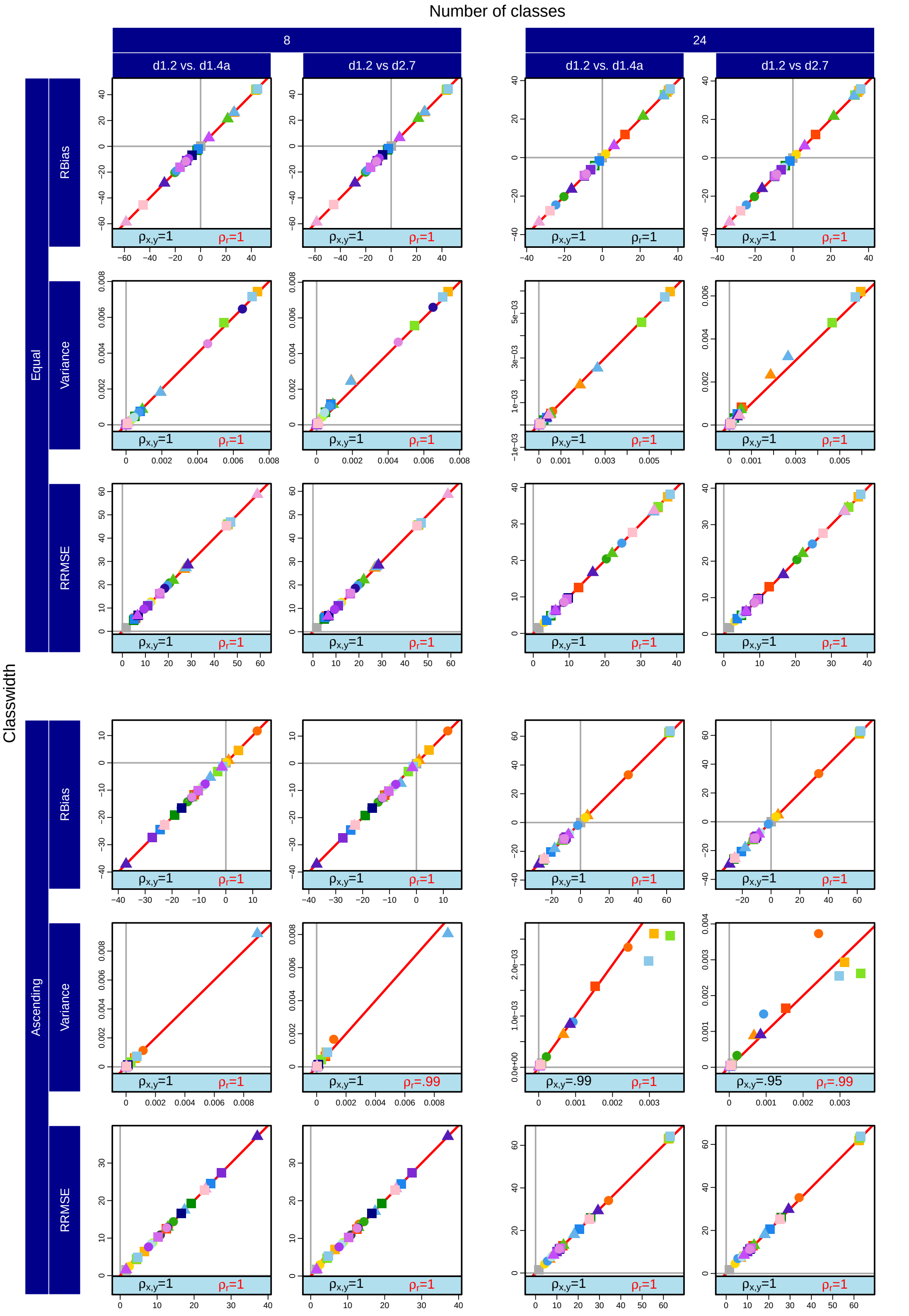
5.6 Comparison of sampling designs



— x = y $\rho_{x,y}$: Correlation ρ_r: Rank correlation   (red, if best method is the same

— 0   (±1%) for both designs.)

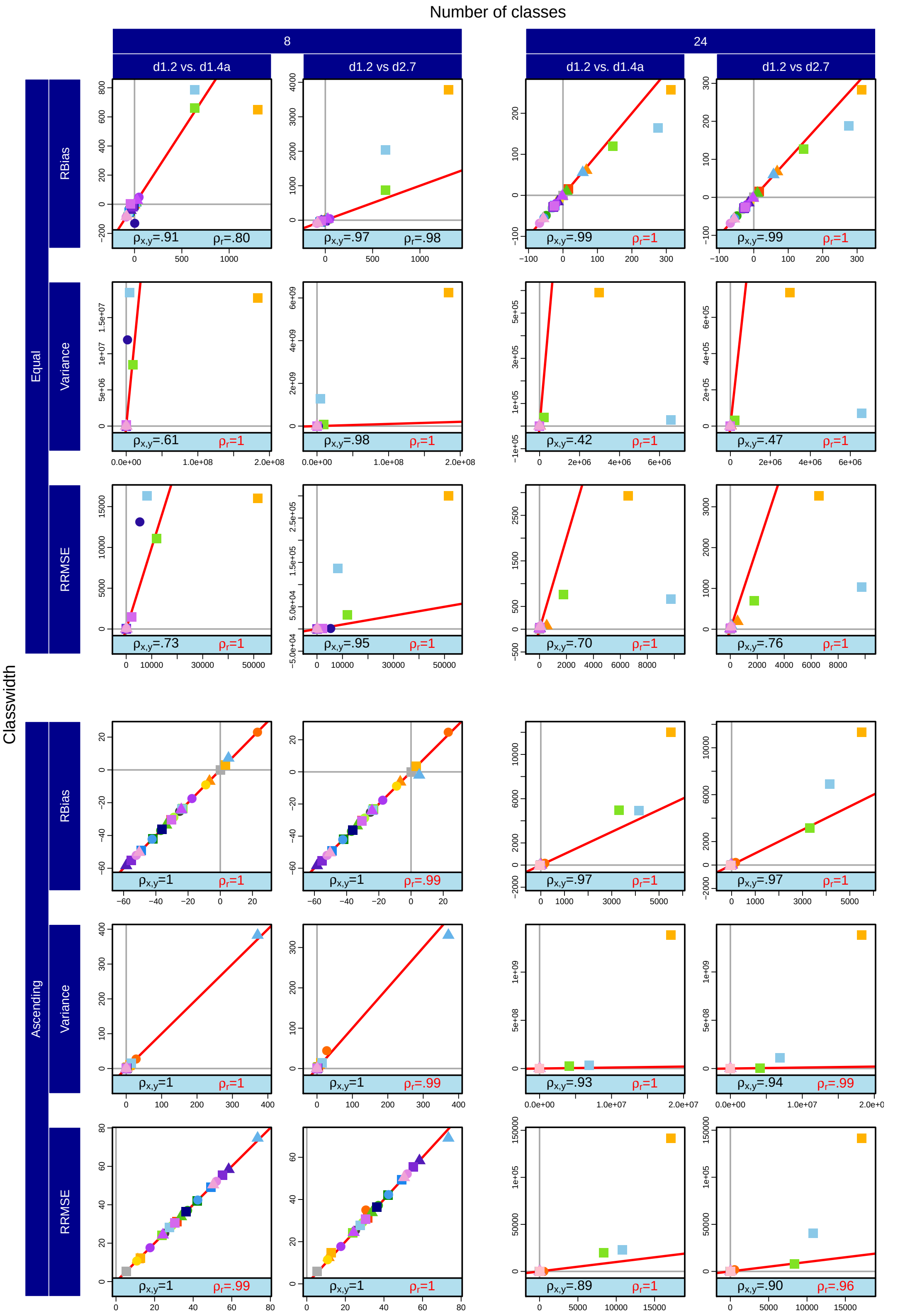
Figure 5.9: ARPR: Comparison of point estimates between designs



x = y $\rho_{x,y}$: Correlation ρ_r: Rank correlation   (red, if best method is the same

0   (±1%) for both designs.)

Figure 5.10: GINI: Comparison of point estimates between designs



— x = y $\rho_{x,y}$: Correlation ρ_r: Rank correlation (red, if best method is the same

— 0 (±1%) for both designs.)

Figure 5.11: QSR: Comparison of point estimates between designs

5.7 Conclusion

Estimation of indicators on poverty and social exclusion is commonly done based on the EU-SILC. But, mainly because of larger sample sizes and better suitability for regional estimation, other data sources are used in some countries. Those surves do not necessarily have an continuous income variable. To determine the most suitable way to estimate those indicators on such data, one has to consider the classification structure.

In general, none of the methods under consideration allow estimating poverty indicators from few equidistant classes. In this case, bias and variance are simply too high, since all distributional information except at the class boundaries is lost. Thus, a higher number of classes or ascending classwidths are necessary to assess poverty in a suitable precision. Regarding the point estimates, parametric income distributions perform better for few or equidistant classes. The nonparametric approaches under consideration seem to be a superior alternative where more classes of ascending width are used. However, the GINI yields an exception to this, presumabely because the parametric income distributions ensure theoretical properties of the income distribution, whereas splines (as used here) fail to do so.

When it comes to inference, the spline-based estimates seem in general more appropriate, unless their bias is too high. As expected, naive linearization is not the method of choice when it comes to income classes. Resampling performs considerably better.

Future research in this topic to obtain methods of linearization accounting for income classes as well as nonparametric modelling of income distributions ensuring their theoretical properties seems promising. The latter might also include studies on how to choose the 'best' smoothing parameter λ. Selecting different smoothings for certain areas of the distribution is also worth thinking.

6 Missing Data and Imputation

Alexandru Cernat **Adrian Byrne** **Natalie Shlomo**

Cathie Marsh Institute for Social Research & Social Statistics, School of Social Sciences
University of Manchester

6.1 Missing data mechanisms

Missing data is a pervasive issue in the social sciences. Although there is significant statistical literature on how to deal with missing data, applied research seems to still be lagging behind. Many researchers are still conducting complete case analysis, a procedure known to be biased even when data is missing completely at random. Below we present a small overview of the state-of-the-art. The main source for the summary is the book by Enders (2010). See also Schafer and Graham (2002) and Little and Rubin (2002) for overviews.

Rubin (1976) first introduced the classification of missing data mechanism. He postulated that missing values could appear in three ways:

Missing Completely At Random (MCAR) is represented mathematically as:

$$p(R|\phi) \tag{6.1}$$

Where R is the missing data indicator and ϕ is a (set of) parameter(s) that describe(s) the relationship between R and the data. Here, the probability of responding doesn't depend on the data and is thus "completely" random.

Missing At Random (MAR) mechanism implies that the likelihood of having missing data depends on other observed and auxiliary variables (Y_{obs}, X):

$$p(R|Y_{obs}, X, \phi) \tag{6.2}$$

Finally, when the data are **Missing Not At Random** (MNAR) then the availability of the data depends both on observed and unobserved variables:

$$p(R|Y_{obs}, Y_{mis}, \phi) \tag{6.3}$$

Recently Thoemmes and Mohan (2015) proposed a graphical way to present the missing data problem and helped link it to the more general literature on Directed Acyclic Graphs (cf. Pearl, 2003). Figure 1 shows the three different mechanisms introduced previously. Here Y* represents the observed data that includes missingness. This is caused by the true variable of interest Y which is unobserved (thus the dashed rectangle) and a propensity to provide missing values on Y, represented by R_Y. The figure also presents an explanatory/auxiliary variable X which has an effect on Y. In the MCAR case the propensity to have missing data is random and thus is not influenced by other factors, such as X. In the MAR case, on the other hand, we see that propensity to answer, R_Y, is caused by the explanatory/auxiliary variable X. In order to have unbiased estimates of the relationship between X and Y the variable X needs to be included in statistical methods that can handle MAR (such as multiple imputation or likelihood methods). Lastly, there are two versions of the MNAR mechanism. In the first one the likelihood to respond or not depends on your level of Y. In the second case there are other unobserved causes that influence both Y and R_Y (L_1) that are not measured.

Section 2 of this overview presents compensating for missing data in a cross-sectional survey setting and Section 3 in a longitudinal survey setting. Section 4 presents some of the literature containing simulations and comparison studies.

6.2 Compensating for Missing Data in Cross-sectional Survey Data

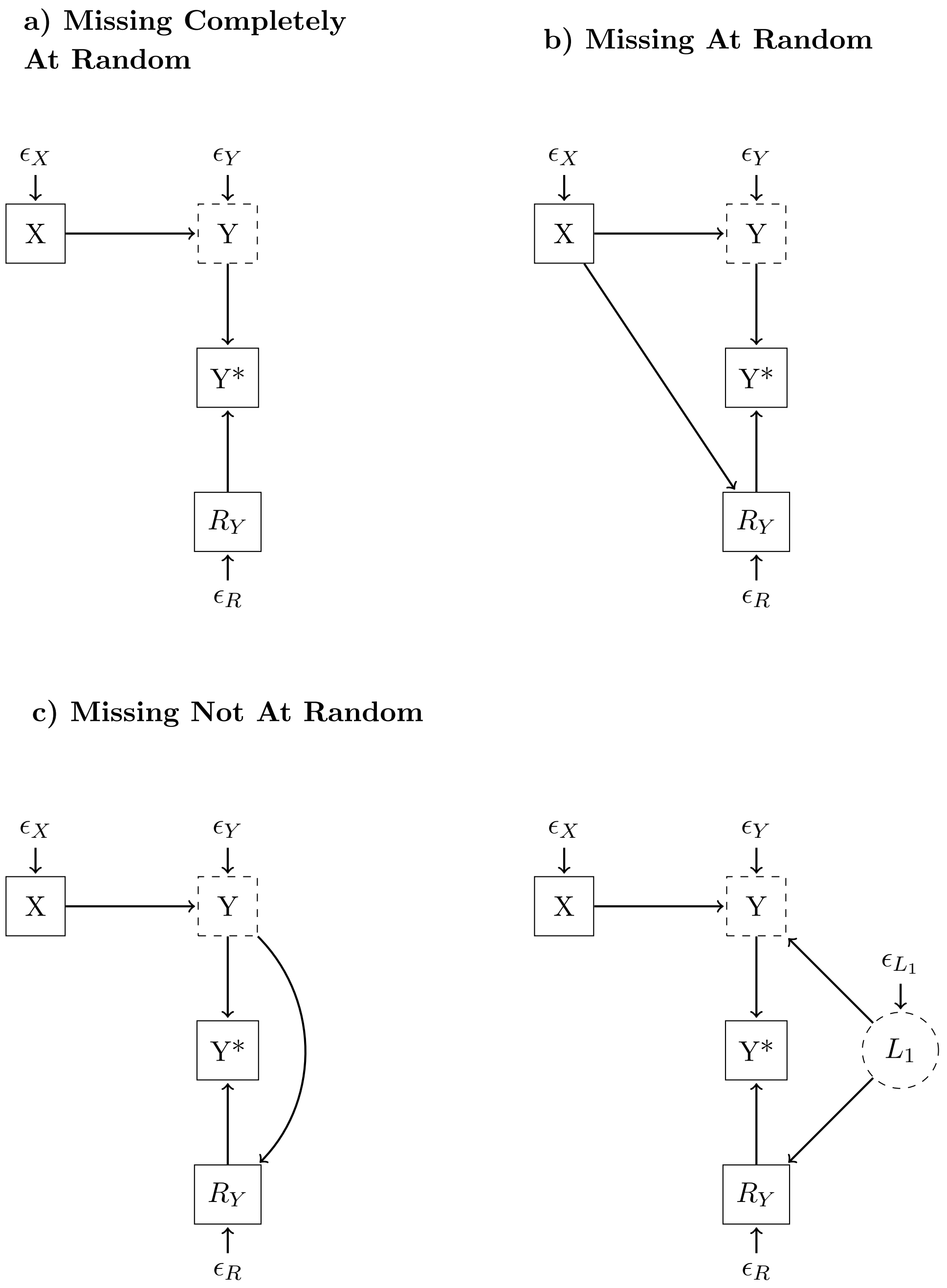
6.2.1 Single Imputation Methods for Item Missing Data

Listwise deletion/complete case analysis implies using only the cases that have available information on all the variables, thus eliminating any case that has any missing from the analysis. It has the advantage of being extremely easy to use (the default in most computer software) and leading to the same sample in all the analyses (unlike pairwise deletion described next). Two big disadvantages with this approach: it leads to biased estimates when MCAR does not hold and loses much power (due to the deletion of any case with missing). An important advantage of this approach is that ”it can produce unbiased estimates of regression slopes under any missing data mechanism, provided that missingness is a function of a predictor variable and not the outcome variable” (Enders, 2010, p.40).

Pairwise deletion/available-case analysis eliminates cases on an analysis basis. This leads to more power than complete case analysis. The main issues relate to the assumption of MCAR and the use of different sub-samples for different analyses (which can lead to nonpositive definite matrices (impossible correlations) and problems with computing standard errors).

Mean imputation works by replacing missing values with the observed mean of the

Figure 6.1: Visual representation of the three missing mechanisms



a) Missing Completely At Random

b) Missing At Random

c) Missing Not At Random

variable. This method gives biased results even under MCAR. It also leads to reduced variance and covariance. "...simulations studies suggest that mean imputation is possibly the worst missing data handling method available" (Enders, 2010, p. 43).

Single regression imputation/conditional mean imputation replaces the missing values with predicted scores from a regression equation. The first step is to model the incomplete variables from the complete ones. Then the predictions based on the regression models are used to replace the missing values. Disadvantages include overestimating correlations even under MCAR and underestimating variability for the imputed data which leads to attenuated variances and covariances. To overcome some of these drawbacks, random errors can be generated from the regression model typically under the normal distribution with zero mean and variance equal to the mean squared error of the regression model. These can then be added to the predictions which replace the missing values.

Hot-deck imputation imputes the missing cases with values from similar responders (see Durrant (2005) and Kenward and Carpenter (2007) for overviews). To do this it "...replaces each missing value with a random draw from a donor set based on a subsample of respondents that scored similarly on a set of auxiliary variables" (Enders, 2010, p. 49). Alternatively, nearest-neighbour hot-deck imputation selects a single donor that is 'closest' to the case with the missing value. This approach preserves the univariate distributions of the data. Nevertheless it may bias estimates of correlations and underestimate standard errors. When multiple variables are missing in a give case, it is the practice that all values are replaced using a single donor to mitigate the impact on correlations.

Predictive Mean Matching is similar to the nearest-neighbour hot-deck imputation approach where the distance measure is based on the predictive mean. In this case, a regression model is used to calculate the distance of the most appropriate donor from which the value for the missing data is taken.

There are variations and combinations of the methods described in this section which may inform best practice.

6.2.2 Weighting for Unit Missing Data

Weighting is commonly used to account for sample design but is also used to compensate for unit non-response in surveys. The inverse probability weighting is typically created by modelling the probability of response to a survey using a logit or probit model. Variables that are measured both for participants and non-participants are used to predict this response propensity. The inverse of the response propensity is used as a non-response weight and compensates for differences between the respondents and nonrespondents. Nonresponse weights can easily be combined with other types of weights, such as those correcting for unequal selection probability and benchmarking to known population totals, and are easy to implement. This approach assumes missing at random.

A specific issue of weighting is that some values of the inverse response probabilities can be very high. This may be due to a misspecified model. In these cases either arbitrary truncation of high values can be used or alternative statistical methods to estimate them, such as a tobit regression. In addition, weights can be 'smoothed' by using the inverse response propensities to formulate weighting classes where each respondent within the weighting class receives the same correction factor for non-response based on the average response propensity within the class.

The inverse probability weighting approach has some characteristics that make it different to imputation methods. One important difference is that the weighting is based on a model that explains the probability of being a complete case while imputation methods model the distribution of the missing data given the observed data (cf. Seaman and White, 2013). Additionally, weighting can traditionally use only fully observed variables and imputation can be more efficient.

In a simulation study Alanya et al. (2015) compared the use of multiple imputation (described in the next section) and propensity-score weighting in order to correct for unit level non-response. They have found that one does not outperform the other overall. In this context they highlight that weighting has a number of practical advantages, is easier to carry out and more user friendly.

6.2.3 Multiple Imputation

Multiple imputation is a collection of approaches that have three common phases in dealing with missing data: imputation, analysis and pooling. The imputation step is also comprised of two steps: an imputation step (I-step) and a posterior step (P-step). In the imputation phase a stochastic regression is used to predict the variable with missing data. This creates the conditional distribution (also named the posterior predictive distribution) from which random draws are taken in the posterior phase. Formally the imputation phase can be written as:

$$Y_t^* \sim p(Y_{miss}|Y_{obs}, \theta_{t-1}^*) \tag{6.4}$$

Where Y_t^* is the imputed value at each I-step t, Y_{miss} is the proportion of missing data, Y_{obs}

is the observed part of the data and θ^*_{t-1} is the mean vector and the covariance matrix from the previous P-step. The P-step "randomly draws a new mean vector and a new covariance matrix from their respective distributions" (Enders, 2010, p. 193). Formally:

$$\theta^*_t \sim p(\theta | Y_{obs}, Y^*_{t-1}) \tag{6.5}$$

Where θ^*_t are the simulated parameters from the P-step and Y^*_{t-1} contains the imputed values from the preceding I-step. Repeating the two steps a number of times creates multiple copies of the data each with unique values for the missing information.

The second stage of using multiple imputation is the analysis one. Here the preferred analysis of the researcher is applied separately for each of the different m datasets. This procedure is usually automated in most statistical software and is no different than the standard application.

Lastly, the results from the analyses undertaken in each of the imputed datasets are pooled. This is done with the **multiple imputation point estimate** as the mean of the different m estimates:

$$\bar{\theta} = \frac{1}{m}\sum_{t=1}^{m}\hat{\theta}_t \tag{6.6}$$

Where $\bar{\theta}$ is the pooled parameter estimate based on the parameter from each dataset t, $\hat{\theta}_t$. Standard errors are computed in a similar way. Here we need to take into account two types of variance: **within-imputation variance** and **between-imputation variance**. The first one can be written as:

$$V_W = \frac{1}{m}\sum_{t=1}^{m} SE^2_t \tag{6.7}$$

and can be described as a simple average standard error. The **between-imputation variance** on the other hand uses the squared difference between the predicted value of the parameter in each dataset t and the average one:

$$V_W = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}_t - \bar{\theta})^2 \tag{6.8}$$

Both of these contribute to the total sampling variance which is a sum of the two variances and a correction for the use of a finite number of imputations:

$$V_T = V_W + V_B + \frac{V_B}{m} \tag{6.9}$$

Goldstein et al. (2014) highlight that there are two approaches to multiple imputation: the one that uses the joint posterior distribution of all variables when sampling for missing values and the chained equation approach that uses the conditional distribution for each variable in turn. A distinct advantage of the former is the implementation of multilevel data structures and interactions in the imputation process (cf. Goldstein et al., 2009). The procedure can also be used to deal with the endogeneity issue.

The approach described up till now is based on a parametric model. One can also carry out a non-parametric method, such as the hot-deck imputation method, and repeat it multiple times. This is called **fractional hot deck imputation** (cf. Kim and Fuller, 2004). This approach has a number of advantages. Firstly, it preserves the distributional properties of the data as observed variables are used to replace the missing values. This may be important for certain distributions and for categorical data (cf. Durrant, 2005). More generally the approach makes no distributional assumptions, as opposed to other multiple imputation approaches. On the other hand Goldstein et al. (2014) highlight that non-parametric imputation approaches such as hot-deck or related donor methods are less efficient and can have issues with small donor pools while the parametric versions work better in smaller samples and are easier to implement. Rubin (2004) presents multiple imputation for nonresponse in surveys and censuses.

6.2.4 Full-Information Maximum Likelihood Imputation

Full Information Maximum Likelihood (FIML) or **direct maximum likelihood** is a means of estimating models by including all the data available. Simulations have shown that it is unbiased under MCAR and MAR, it has high power, as it uses all the information available, and is easy to use.

Intuitively this works by calculating a log-likelihood computation for each missing pattern. When some variables are missing the likelihood estimation will ignore the corresponding coefficients. The final model log-likelihood will be a combination of the different sub-models. Standard errors are computed using the observed information matrix which gives unbiased estimates under MAR (unlike when using the expected information matrix which is unbiased only under MCAR).

An important distinction from multiple imputation is that FIML integrates the missing data and the substantive model into one. Very often the missing mechanism will be different from the substantive model of interest, as such auxiliary variables must be introduced in the model that do not impact the model of interest but help explain the missing data mechanism. One way to do this is to use the saturated correlates model (Enders, 2010, p. 134). This is implemented by correlating the auxiliary variables with the explanatory variables, other auxiliary variables and the residual terms of the outcome variables. One limitation with this approach is that it can lead to estimation and convergence problems. As such, it is recommended to choose only the best auxiliary variables (high correlations with the other variables and low amount of missing data).

FIML assumes multivariate normality. If this does not hold it can impact standard errors. Enders (2010) highlights two main approaches to dealing with this problem: using the sandwich/robust estimator or bootstrapping for the variance estimation.

6.2.5 Models for Missing Not at Random

FIML and multiple imputation are some of the best and popular methods to dealing with missing data. Nevertheless they make the assumption of MAR. In order to bypass this limitation a number of alternative models were put forward. The most prominent ones are the selection model and the pattern mixture model.

6.2.5.1 Selection models

The **selection model** was put forward by Heckman (1979). This approach combines two analyses: the substantive one of interest and a model for response probabilities. Formally this can be written as:

$$p(Y, R) = p(R|Y)p(Y) \tag{6.10}$$

The first part of the equation, $p(R|Y)$, represents the model that predicts missing data and is also called the conditional distribution of missingness. The second part, $p(Y)$ is the substantive model of interest, or the marginal distribution of the data (cf. Enders, 2010). Intuitively this procedure works by first modelling the probability of having missing data and then using the residuals in the substantive model of interest. It has been found that the model works well as long as assumptions are met.

Unfortunately, in a large part of realistic applications the selection models might not hold which can be problematic as the model assumptions are mostly untestable. The most important such assumption is that the regression model of missing data has been correctly specified. Additionally, problems arise when the variables used in the substantive and the missing models are correlated. Finally, the correct estimation is dependent on the bivariate normality assumption of the residuals. It has been found that when assumptions do not hold the selection model can be more biased than the MAR counterparts (cf. Enders, 2010).

6.2.5.2 Pattern-mixture Models

Another approach to dealing with MNAR data is to use **pattern mixture models** (cf. Little, 1993) to factorize the joint distribution depending on the different missing data patterns:

$$p(Y, R) = p(Y|R)p(R) \tag{6.11}$$

Here we model the conditional distribution of the data given a value of the response R, $p(Y|R)$, over the marginal distribution of missingness, $p(R)$. Here we have the reversal of the selection model as the data is dependent on the different missing patterns (cf. Enders, 2010). Molenberghs et al. (1998) show that missing data mechanisms can be applied to pattern-mixture models.

One way to see this approach is as a way of integrating the patterns of missing data in the analysis. This is done by creating subgroups based on the different patterns. The issue with this approach is that patterns with missing values have one or more inestimable parameters. This is solved by grouping different patterns of missing in different ways. The **complete case missing variable restriction** assumes that the inestimable parameters are the same as those in the complete case. Another approach is the **marginal parameter estimate** which averages out the pattern-specific estimates. Alternatively, **neighbouring case missing variable restriction** receives information based on similar groups with incomplete cases (cf. Enders, 2010, Demirtas and Schafer, 2003).

Alternatively, joint modeling has been used to deal with this issue. This implies estimating concurrently both the substantive model and the patterns of missing data (cf. Tsiatis and Davidian, 2004, Wu et al., 2011).

Like the selection model, the pattern-mixture approach is dependent on untestable assumptions. Furthermore, specifying the wrong values for the inestimable ones can lead to bias even under MAR. On the other hand it is argued that it is a valuable tool as it makes it's assumptions explicit and the possibility of employing different approaches for the inestimable parameters can be implemented easily in sensitivity analyses (cf. Enders, 2010).

6.2.5.3 Multiple Imputation Under MNAR

Work has been carried out also to extend the multiple imputation to the MNAR case. Carpenter et al. (2007) proposed using a reweighting approach of the multiple imputation procedure in order to test for different MNAR scenarios. The approach implies using the variables of interest y in a regression model together with a selection measure δ. The coefficient can be given different values depending on the researcher's expectation regarding the missing mechanism not included in the MAR. This proves to be a flexible and easy way to implement sensitivity analyses. The approach is also "robust to the possible mis-specification of the dependence of the missingness mechanism on the observed data" (Carpenter, et al., 2007, p. 273) and can be seen as a semi-parametric alternative to the local sensitivity analyses. One limitation is that the method works well with a large number of imputations, they recommend more than 50 (in the paper they used 1000). Using fewer can lead to extreme weight values. Additionally, the approach might underestimate standard errors as confidence in the value of δ is overstated.

6.3 Compensating for Missing Data in Longitudinal Survey Data

Attrition in longitudinal studies is seen as a serious problem because the loss of individuals over time results in a sample size significantly smaller than the initial sample size after a few occasions. In this instance, using only complete-case data will result in a loss of efficiency. Furthermore, missingness may not occur at random so that the remaining sample may be biased with respect to the variables being analysed. In longitudinal studies, at any given occasion the characteristics of subsequent losses will be known and these can be compared with those who are followed up. If biases are detected then suitable weights can be introduced to compensate for this or a general model-based approach can be employed to deal with attrition. One common approach is to carry forward values from previous waves to replace missing cases. This assumes that scores do not change and can lead to exaggerated group differences and distorted parameter estimates even under MCAR.

6.3.1 Multilevel Multiple Imputation for Missing Data

Since the introduction of multiple imputation (MI), it has become increasingly established as the leading practical approach to modelling partially observed data, under the assumption that the data are MAR. Carpenter et al. (2011) suggest that if the partially observed data are multilevel/hierarchical, this structure should be reflected not only in

the model of interest, but also in the imputation model. In particular, the imputation model should reflect the differences between level 1 variables and level 2 variables (which are constant across level 1 units). Multilevel data arise when observations are clustered in some way. In longitudinal or repeated measures data, subjects are followed up over time and measured repeatedly over time. Each subject therefore contributes one or more observations to the dataset with these repeated observations clustered within individuals. Such clustering generally induces dependence in the observed data, e.g. repeated observations from the same subject are usually correlated with each other. This dependence should be accounted for in the analysis. If missingness only occurs in the response variable(s) in the model of interest, and we can assume that such data are MAR given the predictors and other observed data, then direct likelihood methods (such as random effects models) are usually preferred to MI methods as they are generally computationally faster. Moreover, provided the modelling assumptions hold, likelihood based methods make optimal use of the available data for estimation and inference. However, direct likelihood methods are less well suited in situations where missingness occurs in the predictors in the model of interest. In such instances, MI methods are often more attractive. Furthermore, MI can be used to handle missing data in multilevel/longitudinal datasets. This is very useful particularly when the model of interest is a multilevel or hierarchical model, with the cluster modelled as a random effect as the most appropriate imputation model is correspondingly a multilevel random effects model. This can be achieved by the multivariate/joint modelling approach to imputation. In this approach, a multilevel model is specified for the partially observed variables given the fully observed variables. This multilevel model is fitted to the observed data, following which multiple imputations of the missing data are generated, typically using Markov Chain Monte Carlo (MCMC) methods.
In summary, MI will produce more valid results when:

- Predictors of missingness are included in the imputation model to increase plausibility of MAR;
- The distributional assumptions made by the imputation model are reasonable;
- The imputation model respects the structure of the subsequent MOI;
- The imputed values appear plausible in light of the external contextual knowledge and the observed data.

Ordinarily, any assumptions made about the missing data mechanism cannot be empirically verified definitively using the observed data. Therefore, it is usually wise to consider performing sensitivity analyses to assess the robustness of the results generated by different (but plausible) missingness mechanism assumptions. It is also possible to conduct sensitivity analyses comparing different multiple imputation methods. Simple tabular and graphical analyses will often be sufficient but one could also perform a repeated measures ANCOVA analysis on the response variable(s) of interest with the imputation method treated as a factor by subsetting the data beforehand. Alternatively, one could do pairwise 'global' comparisons between imputation methods (cf. Chambers, 2001). Specifically addressing longitudinal data with non-monotonic missingness, modern methods/software packages can now tackle this problem in wide and long data formats.

6.3.2 Multiple imputation using the fully conditional specification algorithm (MICE)

A popular MI approach is fully conditional specification (FCS), which specifies separate univariate imputation models for each variable with missing data conditional on all other variables (cf. Van Buuren et al., 1999). Therefore, we can choose a model appropriate to the variable type (that is, continuous, count, ordered categorical, unordered categorical). Welch et al. (2014) report that this method is easier computationally than directly specifying a multivariate distribution for a mixture of continuous and categorical variables with missing data, as required in parametric MIs original form.

In longitudinal studies where individuals' characteristics are measured at fixed times, we can treat measurements at each 'time' as distinct variables and impute using FCS multiple imputation chained equations (MICE). An imputation model for a response variable at a particular time point includes the predictor variables at the same time point and the response variable measurements at all other time points as explanatory variables. However, as the number of variables and time points increase, the imputation model has many explanatory variables, potentially causing numerical problems because of overfitting. To overcome this, Nevalainen et al. (2009) recently proposed a modification of the FCS approach to MI, the two-fold FCS algorithm. Missing values at a given time point are imputed from a model that only uses information from that time point and immediately adjacent time points. The rationale is that measurements at time points before or after the time point with imputed measurements are more unlikely to provide substantial additional information than measurements at immediately adjacent time points. This simplifies the imputation models thereby decreasing the computational intensity and reduces problems due to overfitting and collinearity. However, this simplification may induce bias in parameter estimates if the measurements excluded from imputation models have independent effects.

Each univariate imputation step in the standard FCS is as follows: the postulated imputation model is fit to individuals with the variable observed, conditioning on the observed and current imputations of the imputation models explanatory variables; then a draw of the imputation models parameters is taken from their posterior distribution (assuming standard non-informative priors); and lastly, the missing values are imputed using these newly drawn parameter values.

When measurements are intended to be taken from subjects according to a particular schedule, measurements would be available at the same time points for all subjects. In such situations, one possible approach is to impute the data using single-level MI software, treating the repeated observations as distinct variables as outlined by the FCS method above. This can be achieved by putting the data into wide form, where a separate variable/column is used for each measurement time point. FCS MICE can be executed in both Stata and R. The potential drawback of this approach is that structural assumptions, such as random intercepts and/or slopes that one might wish to make, cannot be incorporated.

6.3.3 REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types

Carpenter et al. (2011) developed the REALCOM-IMPUTE software which can perform multilevel multiple imputation, and can handle ordinal and unordered categorical data appropriately. The software may be used either as a standalone package, or in conjunction with the multilevel software MLwiN or Stata. The full conditional specification approach does not explicitly model the joint distribution but forms univariate models for each incomplete variable in turn conditional on all the others. There is no guarantee in general that these correspond to a proper joint model. However, using a multivariate latent normal model allows for the proper handling of a two-level structure whereby level 2 variables are constant over the observations at level 1. Having a multilevel imputation model avoids biasing the parameter estimates in the multilevel model of interest thereby reducing the likelihood of producing potentially invalid estimates of precision. A multilevel imputation model is also appropriate if the data are unbalanced. The REALCOM-IMPUTE software fits multivariate response models to 2-level data, allowing for both level 1 and level 2 variables, and through this allows proper imputation of missing data. Continuous data are modelled using the multivariate normal distribution. The default is to have all the variables as responses, although fully observed variables can be included as covariates; in this way interactions with fully observed variables may be handled. For each level 1 response a mean and level 2 random intercept is fitted, together with a level 1 residual. For level 2 variables, only a mean and level 2 residual is fitted. Level 1 and level 2 residuals are assumed independent, with mean zero, with separate covariance matrices. If all variables are normal these are unstructured; otherwise these have appropriate structures for the latent normal model for discrete data (cf. Goldstein et al., 2009).

The authors note that the imputation model is equivalent to a conditional model for each variable in which it is linearly regressed on all the other variables. Once specified, the REALCOM-IMPUTE software fits the model using Markov Chain Monte Carlo. The authors use a Gibbs sampling approach, updating each set of parameters in turn, conditional on the others. Where possible, the method samples direct from the appropriate conditional distribution. Otherwise the Metropolis steps or rejection sampling method is used.

6.3.4 Further Comments on Imputing Longitudinal Data

When dealing with a non-trivial multilevel setting, multiple imputation (MI) puts the onus on the researcher to devise an appropriate imputation model. The process is therefore thought-intensive as well as computer-intensive. In terms of computing intensity, the FCS algorithm is quicker in completing its MI process. The speed of the Matlab code is a limitation for the REALCOM-IMPUTE software. However, more recent multilevel software has been developed by the Centre for Multilevel Modelling at the University of Bristol which can more quickly complete multilevel MI. This software is called Stat-JR and it uses a C++ compiler (MinGW) rather than a Matlab realtime installer to compute the multiply imputated datasets. A possibly preferable alternative approach to a single large-scale MI is to generate separate imputed datasets for each outcome of interest, with imputation models specified to ensure they are compatible with a given outcome and model of interest. Consideration must also be given to impute compatibly using standard imputation models when substantive models contain interactions and/or nonlinear covariate effects (cf. Bartlett et al., 2015). Goldstein et al. (2009) state that if we make no distinction between the types of non-response (e.g. refusal, non-contact etc.), then we assume that the relationship (as expressed in the parameters of the imputation/prediction model) is the same for different types. If this is not plausible given the data then we can allow for this in the imputation model by including auxiliary variables associated with different types of non-response. Where we cannot assume MAR and attrition is informative (i.e. missing not at random (MNAR)) it may be possible to adopt a fully model based procedure. In this case we need to write down an explicit (selection) model for the probability of attrition and link this to the model of interest that we wish to fit. The essence of selection models is that the residual terms for the substantive model and the model for response are correlated. For such a bivariate model, often a normal probit model, to be identified and hence for a valid correction term to be able to be included in the substantive model, the model for predicting response must include variables that are unrelated to the outcome of interest. Such variables, often referred to as instruments, can be difficult to find but it is plausible to suppose that the variables that are linked to the data collection process, could be a valuable source of valid instruments. Consequently, HAWKES and PLEWIS (2006) found that the probability of a response is predicted by variables like the proportion of questions that were answered at a previous sweep and the number of addresses that were attempted at the current sweep could be used to identify a selection model.

Some additional multilevel methods briefly mentioned here:

- **Mixed-Effects Pattern-Mixture Models for Incomplete Longitudinal Data**: Mixed-effects pattern-mixture models maybe more useful by adding a missing-data pattern as a between-subjects factor which assesses the degree to which 'missingness' influences (available) outcomes as well as the degree to which 'missingness' interacts with model terms. This method does not invent data but maximizes information obtained from available data. KENWARD et al. (2003) present pattern mixture modelling for incomplete longitudinal data.
- **Multilevel selection models**: Models for handling sample selection or informative missingness have been developed for both cross-sectional and longitudinal or panel data. For cross-sectional data as discussed above, HECKMAN (1979) suggested a joint model for the response and sample selection processes where the disturbances of the processes are correlated. For longitudinal data, (cf. DIGGLE and KENWARD, 1994) developed a model in which the continuous response (observed or unobserved), and possibly the lagged response, is a predictor of attrition or dropout. They combine growth curve models with another model for wave specific response probabilities. Another approach is the random coefficient selection model which uses directly the intercept and slope latent variables to predict the probability of response. FOLLMANN and WU (1995) proposed to use the **shared parameter models** to model

jointly event time data and serial data. This is done by assuming that the joint distribution of the repeated measurements and event times are conditionally independent given a shared random effect. Vonesh et al. (2006) extend this model by using generalized linear and non-linear mixed-effects models. The Heckman model can be estimated using the heckman command in Stata and the Diggle-Kenward model is available in the Oswald package running in S-PLUS. Both models can also be estimated using gllamm with the advantage that the following three generalizations are possible. First, the models can be extended to multilevel settings where there may be unobserved heterogeneity between the clusters at the different levels in both the substantive and selection processes and where selection may operate at several levels. Second, the Heckman model can be modified for non-normal response processes. Third, both the Heckman and Diggle-Kenward models can be extended to situations where the substantive response is a latent variable measured by a number of indicators.

6.4 Simulation and Comparison Studies

Faris et al. (2002) compared three imputation methods (norm, MICE and Transcan) with the use of administrative records to predict mortality in a clinical trials. They have found that imputations with MICE is slightly better but overall the estimates were similar. Additionally, they have found that adding administrative data to the MICE imputation did not improve the prediction model. Nevertheless there are examples of situations in which linking data with administrative information can improve models for missing data significantly (cf. Cornish et al., 2015).

In a small simulation study, Kristman et al. (2005) showed how estimates can be biased under different attrition situations (10%, 25%, 45% & MCAR, MAR and MNAR). They find that under MAR and MCAR imputation and weighting give unbiased results but this is not true under MNAR.

Bartlett et al. (2014) discuss complete case analysis (CCA), multiple imputation (MI), and inverse probability weighting (IPW) in the context of MNAR data and show that in some settings CCA may be more valid under MNAR where missingness in a covariate depends on the value of that covariate, but is conditionally independent of the outcome variable. The other methods assume MAR. They propose an augmented CCA approach which makes the same conditional independence assumption for missingness as CCA, but which improves efficiency through specification of an additional model for the probability of missingness, given the fully observed variables.

Holman and Glas (2005) propose a procedure based on item response theory modelling using the partial credit and generalized partial credit models which aims to reduce the bias caused by ignoring the missing-data mechanism.

In another study, Engels and Diehr (2003) compare different procedures of implementing multiple imputation in dealing with attrition in a longitudinal study. They find that using individual longitudinal data (both before and after the missing point) are superior. They also found that all methods are biased towards making people appear healthier than they are due to the non-ignorable nature of the missing.

7 A multidimensional approach to measure children's living conditions

Caterina Giusti

Department of Economics and Management

University of Pisa

7.1 Methodology

This study aims to gain insights into children's living conditions and to generate knowledge relevant to broaden the discussion on the crucial topic on children's well-being, adding some interesting points to the current literature, by using EU-SILC data. The idea of the study was born upon the TNA visit of Dr Antoanneta Potsi at the Department of Economics and Management of the University of Pisa. Part of this study has already been published in the paper by Potsi et al. (2016), while some other part will be objective of future publications, such as a book chapter in the book "Human development in times of crisis" to be published soon by Palgrave MacMillan. From a methodological point of view we present an approach based on the fuzzy methodology introduced by Cheli and Lemmi (1995) and then updated by Betti et al. (2006). This methodology, developed for the study of poverty on a multidimensional perspective, is able to preserve the richness of the data available from the EU-SILC survey. To define the dimension of children's well-being we decided to engage with the Capability Approach (CA) as an alternative normative framework for the evaluation of human development, well-being and freedom by thinking in terms of human functionings and capabilities. Functionings are features of the state of an existence of a person (cf. Hawthorn, 1987) while capabilities represent what people are able to do or to be (cf. Sen, 1999). According to Dean (2009), capabilities represent the essential fulcrum between material resources (commodities) and human achievements. The use of fuzzy methodologies under the capability approach has already been explored by some authors in a theoretical and applied perspective (cf. Chiappero-Martinetti, 2006, Addabbo and Facchinetti, 2013), but none of these works presented a framework for the study of children living conditions.

The CA focuses on measuring the well-being of adults whose freedom to choose a life they have reason to value is central to the notion of capabilities. Biggeri et al. (2006) argue that children are subjects of capabilities and that the capability approach can be very useful as a framework of thought and as a normative tool, in analysing children's well-being, poverty and deprivation and in individuating social policies for children human development. It is argued that deficiencies in important capabilities during childhood not only reduce the well-being of those suffering from the deficiencies, but may also have larger societal implications (cf. Klasen, 2001, Biggeri, 2007). Nussbaum (2006) developed an open-ended and extendable list of domains or basic central capabilities for human flourishing as a minimum account of social justice that has proven to be a valuable framework for the operationalization of the approach and inspired this study.

The methodological approach used in this work (henceforth Integrated Fuzzy and Relative – IFR) was born on the assumption that poverty is a multidimensional phenomenon and a vague predicate that manifests itself in different shades and degrees (fuzzy concept) rather than an attribute that is simply present or absent for individuals in the population, as the traditional poverty approach assumes. The fuzzy set vision of poverty is particularly adequate for studying children's living conditions and deprivation mainly for two reasons. Firstly, it includes a non fixed value of poverty risk and deprivation, through the introduction of a membership functions (m.f.), i.e. a quantitative specification of individuals/households degrees of poverty and deprivation depending on the other individuals or households included in the analysis. A membership function's value of 0 is always associated with the lowest risk of poverty and deprivation, whereas a value of 1 is associated with the highest risk. Secondly, the multidimensional framework of the IFR approach proposed by Betti et al. (2006) works up on several non-monetary indicators, assumed to be the manifest representation of a restricted number of underlying domains of deprivation, besides a monetary indicator based on the equivalent disposable income. The multidimensional analysis of poverty seems to be one reasonably grounded way to combine the CA and secondary quantitative data, because it includes monetary and non-monetary dimensions going beyond the traditional approach based only on the economic or financial situation.

As concerns the monetary dimension, the Fuzzy Monetary Indicator (FM) is computed for each individual i by using the following equation:

$$FM_i = (1 - F_i)^{\alpha-1}(1 - L(F_i)) = \left(\frac{\sum_\gamma w_\gamma | y_\gamma > y_i}{\sum_\gamma w_\gamma | y_\gamma > y_1}\right)^{\alpha-1} \left(\frac{\sum_\gamma w_\gamma y_\gamma | y_\gamma > y_i}{\sum_\gamma w_\gamma y_\gamma | y_\gamma > y_1}\right) \tag{7.1}$$

where y_γ is the equivalised income, F_i is the income distribution function, w_γ is the sample weight of individual of rank γ ($\gamma = 1, \ldots, n$) in the ascending income distribution, L_i represents the value of the Lorenz curve of income for individual i. Therefore, the

FM indicator takes into account both the share of individuals less poor and the share of the total equivalised income received by all individuals less poor than individual i. The parameter α is chosen so that the mean of the membership function FM is equal to the Head Count of Ratio (HCR), the indicator measuring the number of poor people, i.e. the number of people with income below the poverty line:

$$E(FM) = HCR. \tag{7.2}$$

In this manner the scale parameter α allows the comparison of the fuzzy monetary measure with one of the more traditional poverty measures, the HCR. In this study it is assumed the HCR to be equal to the EU standard definition of people at-risk of poverty calculated on the sub-population of household with children. The computation of the deprivation indicators of the IFR approach involves a long process. In details, the IFR method follows seven main steps: 1) identification of the relevant survey items; 2) transformation of the items into the [0,1] interval, 3) exploratory and/or confirmatory factor analysis to define the latent dimensions; 4) computation of the weights within each dimension; 5) calculation of the scores for each dimension; 6) calculation of an overall score and the α parameter as for the FM measure; 7) calculation of the deprivation indices. The identification of the EU-SILC 2009 items referring to children's deprivation relevant under the CA covered the first step of the above process. Then, in step 2, the items chosen in step 1 were transformed into the $[0, 1]$ interval by using the following equation:

$$s_{j,i} = 1 - \frac{1 - F(c_{j,i})}{1 - F(1)} = \frac{F(c_{j,i}) - F(1)}{1 - F(1)} \tag{7.3}$$

for $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, n$. Here $c_{(j,i)}$ is the value of the category of the $j - th$ item for the $i - th$ individual and $F(c_{(j,i)})$ is the value of the $j - th$ item cumulation function for the $i - th$ individual. When the $j - th$ item is a dichotomous variable we have $s_{(j,i)} = 0$ for deprivation and 1 otherwise when the item is instead polychotomous we assign to each unit a value corresponding to the percentage of units that are better off than that unit instead of the real value of the category. In step 3, an exploratory factor analysis was performed in order to confirm the latent structure of the domains previously defined. Then the weights within each domain were computed, covering step 4. Using the scores and the weights previously computed, the next step (step 5) was the computation of a single, aggregate score for each domain h and individual i as the weighted mean taken over the j items:

$$s_{hi} = \sum w_{hj} \cdot s_{hj,j} / w_{hj}. \tag{7.4}$$

Here w_{hj} is the weight of the $j-th$ deprivation variable in the $h-th$ dimension. In step 6 the overall score for the $i-th$ individual is calculated as the unweighted mean of the individual-specific scores:

$$s_i = \frac{\sum_{h=1}^{m} s_{hi}}{m}. \tag{7.5}$$

Following the same process used for the FM indicator, a Fuzzy Supplementary (henceforth, FS) measure can then be computed as:

$$FS_i = (1 - F_{(S),i})^{\alpha-1}(1 - L_{(S),i}) \tag{7.6}$$

where the parameter α is determined again equal to the value that allows the overall FS measure rate to be numerically identical to the HCR. Finally, in step 7 we separately computed the membership function FS_{hi} for each dimension of the deprivation by:

$$FS_{hi} = (1 - F_{(S),hi})^{\alpha-1}(1 - L_{(S),hi}) = \left(\frac{\sum_{\gamma=i+1}^{n} w_{h\gamma} | s_{h\gamma} > s_{hi}}{\sum_{\gamma=2}^{n} w_{h\gamma} | s_{h\gamma} > s_{h1}}\right)^{\alpha-1} \cdot \left(\frac{\sum_{\gamma=i+1}^{n} w_{h\gamma} s_{h\gamma} | s_{h\gamma} > s_{hi}}{\sum_{\gamma=2}^{n} w_{h\gamma} s_{h\gamma} | s_{h\gamma} > s_{h1}}\right) \tag{7.7}$$

where $F_{(S),hi}$ is the distribution function of s evaluated for individual i dimension h; $1 - F_{(S),hi}$ is for the $i-th$ individual the proportion of individuals who are less deprived, in the $h-th$ dimension, than the individual concerned; $w_{h\gamma}$ is the sample weight of the $i-th$ individual of rank γ in the ascending score distribution in the $h-th$ dimension; $L_{(S),hi}$ is the value of the Lorenz curve of s for individual i in dimension h; $1 - L_{(S),hi}$ is the share of the total lack of deprivation score assigned to all individuals less deprived than the person concerned.

7.2 Application to Italian EU-SILC data

The data of this study are derived from the European Survey on Income and Living Conditions (EU-SILC). We use in particular the cross-sectional EU-SILC 2009 data for Italy. In 2009 the questionnaire of the survey was enlarged by adding a specific module on material deprivation to the standard core survey (cf. EUROPEAN COMMISSION, 2009a). This list of target secondary variables relating to material deprivation allows a better understanding of childhood and capability deprivation in Italy. EU-SILC 2009 data were selected as adequate to reveal aspects of the multidimensional deprivation of children, though the available data can only be considered as potential proxies for capabilities. However, the use of EU-SILC 2009 data has many advantages: an example is the European comparative study of Neubourg de et al. (2012) that used EU-SILC 2009 data to define 14 domains of deprivation for children living in 32 European countries, including financial items. In addition, other empirical studies, even though based on different databases, defined capabilities using similar observed indicators - see, for instance, Maccagnan (2011) or Addabbo et al. (2014).

The data selected focused on children aged 0-14. Although significant works in the field consider children as those aged 0 to 17 years inclusive, in line with the United Nations Convention on the Rights of Children (1989) the age boundaries 0-14 were chosen for this study. This age group may be subject to higher vulnerability and intergenerational dependence. Moreover, until the age of 14 Italian children attend the same compulsory schools – elementary plus secondary lower school – while in the following years they can differentiate their educational path, e.g. choosing between schools more addressed to University studies or to the labour market. Thus, the age of 14 corresponds to decisive breaking point in Italian children's life. The final data set used in the empirical analysis covers 19,128 individuals from 5,030 households with at least one child aged less than 15 (see Table 7.1).

Table 7.1: Households with children: number and percentage of households, number and percentage of individuals (with respect to the overall population).

# households	**% of total households**	**# individuals**	**% of total individuals**
5030	23	19128	37

Table 7.2: Head Count Ratio of households with children, by some households' characteristics.

Macro regions			**Single parent**		**Educational level**		
North	Centre	South	Yes	No	Low	Medium	High
11%	16%	41%	31%	23%	36%	17%	7%

The traditional view of poverty (HCR) depicts, as expected, a country where the risk of poverty among children is much higher than among the whole population (24% vs 18%). Although this national trend characterises several Member States, the estimated six percentage points gap in Italy is higher than the European average, equal to approximately three percentage points (European Commission, 2010). Moreover, the risk of poverty among children varies considerably across Italian macro-regions (see Table 7.2). The disparity between Northern and Southern regions is conspicuous, as the HCR ranges from 11% to 41%. Central regions play frequently a mediator effect between two very different economies: the north where an attitude towards entrepreneurship has found a fertile ground, while in the south it did not happened. The gap between single parent households (with HCR equal to 31%) and other household compositions (HCR 23%) is also evident. Moreover, the differences in the HCR observed classifying the households according to three alternative educational levels attained by the reference person in the household were investigated (see again Table 7.2). Education is a basic capability that affects the development and expansion of other capabilities. Household educational level can be used as proxy of the household social class. The study showed that the risk of poverty is very high for children living in households characterised by low educational level (HCR 36%). In contrast, high household educational levels guard their children from poverty (HCR 7%). Using the above mentioned data, each capability was assumed to be a latent variable that is measured by multiple indicators. Such indicators are manifestation of the latent factor, thus a variation in the capability determines a variation in all functioning measures. A selection of the items, from the large set of EU-SILC variables, substantively meaningful and useful for the construction of fuzzy monetary (FM) and supplementary indicators (FS) was made. This was a crucial step, since the choice of the capabilities to include in the evaluation was not straightforward. Dimensions were selected based on the restrictions of the existing data. An attempt was made to include a comprehensive list of basic capabilities considering all the available items for children. The items selected were classified into seven components that represent a respective latent capability (see Table 7.3): the ability to play (PLAY), to be well nourished and clothed (NUTRITION & CLOTHING), to have an adequate financial budget at household level (FINANCIAL), to have a social life (AFFILIATION & SOCIAL LIFE), to live in an adequate housing (SHELTER) and in a good environment (SAFETY) and to be bodily healthy (BODILY HEALTH). They are not referred just to children but they consider a broader complex system in which children spent their life.

The list of deprivation items is arranged into seven domains of children's functionings. The

Table 7.3: Capabilities, indicators/functions and Cronbach's alpha values.

CAPABILITIES	**INDICATORS/FUNCTIONINGS**	**CRONBACH'S ALPHA VALUES**
PLAY	Outdoor leisure equipment	0.78
	Indoor games	
	Book at home suitable for their age	
NUTRITION & CLOTHING	Fresh fruit & vegetable once a day	0.75
	Three meals a day	
	One meal with meat, chicken, or fish (or veg. equivalent) at least once a day	
	Some new (not second-handed) clothes	
	Two pairs of properly fitting shoes	
FINANCIAL	Inability to cope with unexpected expenses	0.65
	Arrears on mortgage or rent payments	
	Arrears on utility bills	
	Arrears on hire purchase instalments	
	Ability to keep the home adequately warm	
AFFILIATION & SOCIAL LIFE	Celebrations on special occasions	0.75
	Internet connection	
	Regular leisure activity	
	Invite friends round to play & eat from time to time	
	Participate in school trips and school events that cost money	
	Child holiday away from home at least 1 week for year	
	Outdoor space in the neighbourhood where children can play safely	
SHELTER	The accommodation is too dark	0.62
	The dwelling has an insufficient number of rooms compared to the number of persons	
	The dwelling has leaking roof/damp walls/floors/ foundations or rot in the window frames	
	Suitable place to study or do homework	
SAFETY	Crime, Violence, vandalism	0.64
	Pollution	
	Noise	
	Damaged public amenities in the neighbourhood	
BODILY HEALTH	Unmet need for consulting a dentist	0.66
	Unmet need for consulting a GP or specialist excluding dentists and ophthalmologists	

internal reliability (i.e. the degree to which the items in the scale are representative of each latent construct) has been estimated by the Cronbach's alpha index. The reliability of each scale-items ranges from 0.52 to 0.78 (Table 7.3, third column). Since not all of the alpha coefficients exceeded the 0.70 cut-off recommended by Nunnally and Bernstein (1994), Exploratory Factor Analysis (EFA) was used to give a framework of the domains regardless the theoretical assumptions. The exploratory factor analysis mostly supports our null hypothesis. Indeed, seven domains were identified reflecting the arrangement presented in Table 7.3 unless two items that loaded respectively in the third and the first dimension as regards EFA. We proceed to rearrange these two items in the assumed domain in order to create more meaningful groups and according to the experience acquired in this framework.

We used Confirmatory Factor Analysis (CFA) for determining whether the hypothesized factor structure provided a good fit to the data. In other words, we tested if the relationship between the observed variables (functionings) and their underlying latent constructs (capabilities) exists (see Table 7.4). CFA outcomes provide information on each indicator's significance. When running CFA, many different fit statistics can be used to help determine whether the model provides adequate fit for the data. To assess the fit of the latent structural model, Chi-square computation was omitted since its value is considered to be highly dependent on sample size. Instead, the Adjusted for Degrees of Freedom (AGFI), the Root Square Mean Error of Approximations (RMSEA) and the Non-Normed Fixed Index (NFI) were calculated. The AGFI accounted for 0.91, where achieving a value close to 1 indicates a good fit. The RMSEA, expressing the unexplained or residual variance of the factor structure, accounted for 0.05: values of the statistic ranging between 0.05 and 0.08 indicate reasonable errors of approximation in the population. NNFI, equal to 0.86, met the criteria for acceptable fit (0.80 or greater).

Table 7.4: Robustness analysis of the hypothesized structure: CFA.

Index	**Value**
Adjusted for Degrees of Freedom (AGFI)	0.91
Root Square Mean Error of Approximations (RMSEA)	0.05
Bentler&Bonett's (1980) Non-Normed Fixed Index NNFI	0.86
Number of dimensions	7

Table 7.5 illustrates the fuzzy and multidimensional poverty measures overall and according to some households' characteristics. Besides the already mentioned higher poverty for households with children, children are also deprived with respect to non-monetary domains with different degree of severity. At national level (see the first column of Table 7.5) the BODILY HEALTH domain has the lowest value (0.043): this means that deprivation in this domain compared to the other dimensions is negligible. It should be noted that "bodily health" focuses on the access to health care and in particular on the need for a specialist doctor, which implies a previous identification of the need, which is not measured here. For PLAY and NUTRITION & CLOTHING domains similar results (both 0.068) are obtained. Italian children basic needs are satisfied in terms of food and clothing, and they are not deprived from medical and specialist care. They have also a good equipment of leisure activities and games. Surely, these are not direct measures of how – and if - children use their equipment, but these results show that they have the possibility to do it. In contrast, the SAFETY and AFFILIATION & SOCIAL LIFE domains obtained the highest values, 0.194 and 0.170, respectively. It means that these dimensions have a larger impact respect to others included in the analysis on children deprivation. Italian children appear more vulnerable with respect to life outside the family. It could be represented as duality: internal to own family and external to it. Aspects which are more related to familial or internal dimensions are more amenable to alternative non-monetary resourcing, such as food and health, while such solutions are less available for aspects of social life and the quality of environment. The negative evaluation of quality of environment signals the lack of safety threatened by crime, violence, vandalism, pollution, noise or by damaged public amenities in the neighbourhood. Lastly for SHELTER and FINANCIAL domains, median results were obtained, that is the third and fourth highest values respectively.

It is interesting to observe the ranking of the non-monetary domains per three macro regions. The analysis confirms that a north/south dualism in Italy has a primary component in the financial and economic status also for households with children. The FM domain plays a crucial role for determining children's deprivation in central and southern regions, while in northern regions SHELTER and SAFETY conditions seem to be more important than monetary condition. The lowest value is observed on the BODILY HEALTH domain in the three macro-regions, while the lowest level of children's deprivation in this domain appears to be in central regions. Among the macro regions, the value of the capability of PLAY is more than double in the South in comparison with the other two macro-regions: 0.104 versus 0.047 in the North and 0.046 in the Centre. The same gap is observed also for other dimensions, confirming a better evaluation for "inside" dimensions and a worst one for environment quality and community life. These findings suggest a sort of "dual duality" in Italian children's quality of life: near to the traditional north/south, the new internal/external life comes from the fuzzy approach. As the first dualism has been studied for a long time, addressing the new one could be the way also to reduce the distance between the north and south of Italy on children's quality of life.

We also investigated the influence of the household-types (single and non-single parent) to the economic situation of the family and of children's capabilities (see Table 7.5). Single parent households are almost exclusively composed by woman and children (83.7%) and they are more vulnerable than the other type of households: the absolute poverty increased from 5% to 8% in $2011-2012$ and from 5.8% to 9.1% in the case of single parent households. The analysis confirms the vulnerability of single parent household: the values obtained for the FM and for the FS indicators are all higher for households with a single parent, with the only exception of the NUTRITION & CLOTHING domain, where the results are equal. These last domains are part of care work, which remains a task mainly burdening women both on single and in two-parent families. Women and children mainly compose single parent households and this gender composition may imply a low disadvantage on good meals and adequate clothing in such households. The BODILY HEALTH dimension is the one where the highest gap between the two groups of households is observed: in this case the value of households with a single parent is the double with respect to the value obtained for all the other households. As said before, the access to a specialist involves economic expenditure and the single parent are financial weaker. This could be the reason of the huge gap between the two household types. Literature (cf. Bradshaw et al., 2012, Krueger and Lindahl, 2001) and empirical evidence (cf. Tarki Social Research Institute, 2010) show that the level of education of the household members influences the child's survival and development chances as well as their risk of poverty. The consideration of three separated groups of households - those where the head of the household has low, medium or high educational level[1] - led to interesting results (see again Table 7.5). In the first two groups, the analysis shows that safety domain is the most relevant, implying a central role for children's capabilities enhancement. Further, the lowest gap between low and high-educated families for this dimension is observed. Yet, the analysis conducted by

[1] Families are classified based on the major income earner (the person with the highest income in the household). The household educational level has been defined through the International Standard Classification of Education (ISCED). Individuals whose attained educational level is lower than the ISCED level 3 are classified as low-educated while individuals whose ISCED level is greater than 3 are classified as high-educated.

Table 7.5: Fuzzy and multidimensional poverty measures (only households with children).

		Households characteristics (only households with children)								
		Macro regions			**Single parent**		**Educational level**			
Description	**Italy**	North	Centre	South	Yes	No	Low	Medium	High
FUZZY MONETARY	0.236	0.141	0.186	0.367	0.30	0.23	0.34	0.18	0.09
PLAY	0.068	0.047	0.046	0.104	0.10	0.07	0.11	0.04	0.01
NUTRITION & CLOTHING	0.068	0.051	0.044	0.099	0.07	0.07	0.11	0.04	0.02
FINANCIAL	0.160	0.120	0.154	0.210	0.22	0.16	0.24	0.11	0.08
AFFILIATION & SOCIAL PARTICIPATION	0.170	0.111	0.126	0.261	0.21	0.17	0.25	0.12	0.08
SHELTHER	0.169	0.143	0.172	0.198	0.19	0.17	0.22	0.14	0.12
SAFETY	0.194	0.170	0.180	0.230	0.22	0.19	0.19	0.19	0.21
BODILY HEALTH	0.043	0.039	0.026	0.055	0.08	0.04	0.07	0.02	0.02

educational status enhances the distance for the PLAY dimension, which is eleven times higher for households with low education level with respect to households with high one. For NUTRITION & CLOTHING, BODILY HEALTH, AFFILIATION & SOCIAL LIFE and FINANCIAL the gap is very pronounced – the values of the low educated households are at least three times higher in comparison with high educated ones. It is worth noting that by considering the last two disaggregation criteria, the ranking of domains remain the same, with minor changes. This confirms the hypothesis of internal/external factors concurring to determine children's well-being following the CA approach.

Another contextual variable we investigated is the social class of the household[2] (see Table 7.6). The values reported in Table 7.6 are the ratios of the estimates obtained for social class 2 (blue collar class) and 3 (working class) with respect to the estimates for social class 1 (white collar), chosen as the baseline. Actually social class 1 represents the best benchmark insofar as it includes very high-skilled workers. Therefore, as expected, almost all the ratios are major than one, indicating a higher deprivation for social classes 2 and 3 with respect to social class 1. The main exception regards the SAFETY dimension, with values of the ratios equal to 0,85 and 0,95 for class 2 and class 3 respectively. Even if the reasons behind this result cannot be easily understood without further information, we can suppose that some of the items included in such dimension, as for example pollution and noise, are certainly more important issues in big cities where it is also higher the concentration of households belonging to social class 1. Therefore, this result can hide a geographic rather than a social class effect. Also for the SHELTER dimension the gap is not so pronounced, especially the one between social class 2 and social class 1. At the opposite side, we can observe a pronounced gap for the FS measures PLAY and NUTRITION & CLOTHING. To sum up, these results suggest the existing social class differentials in children's well-being, which may affect the future capabilities and opportunities of children and can have consequences on the total welfare.

[2] The household's social class has been based on the definition used by the ESEC, the European Socio-economic Classification (cf. Whelan and Maitre, 2008). We use the current job for those employed at the moment of the EU-SILC 2009 interview, while we use the information related to the last job for those that are not currently employed but were employed in the past. The ESEC criteria define nine classes. For simplicity we have aggregated them in three classes: class 1- (white collar) comprising employers, higher grade professional, administrative and managerial occupations, higher grade white-collar workers and lower supervisory and lower technician occupations; class 2 (blue collar class) - comprising small employer and self-employed occupations (not included in the first class), class 3 - (working class) comprising lower services, sales and clerical occupations and lower technical occupations, routine occupations. Households in which the head of the household never worked are not included in the analysis, since it is a "residual" class and it has been classified with missing values.

We finally present the results characterising the households depending on the age[3] and the gender of the children (last three columns of Table 7.6). The results referring to the mean age of children are computed as ratios comparing the first and the third class of mean age values over the second class, chosen as the baseline. We can notice that the deprivation of households with children aged less than 3 years and more than 9 years is always minor with respect to that of households with children aged 3 to 9 years. This may depend on the lower needs or lower relevance of younger children in terms of the functionings considered in the present study (see Table 7.3).

Lastly, as concerns the gender of the children, to control for the effect of the number of children in the household, we decided to consider only households with one female child and those with one male child. For these households the ratio is computed considering the category one male as the baseline. The results suggest that the deprivation for households with one female is higher under all the dimensions.

In future developments the methodology used in the present study will be applied to other European Countries using again EU-SILC 2009 data. This will allow the comparison of children well-being and capability deprivation on a geographical basis. Moreover, since the ad-hoc module on children has been repeated in the EU-SILC 2014 wave, the application to these data will allow the comparison of the results on a temporal perspective, investigating the effect of the financial crisis on children's living conditions.

[3] The households where classified into three classes according to their children's mean age: less the 3 years (class 1), between 3 and 9 years (class 2), more than 9 years (class 3)

Table 7.6: Fuzzy supplementary deprivation rates according to some socio-demographic characteristics.

Variable	**Social Class**		**Children's mean age**		**Children's gender**
	Class 2/ Class 1	Class 3/ Class 1	< 3 years/ Bet. 3 and 9	> 9 years/ Bet. 3 and 9	One female / One male
PLAY	1.32	5.31	0.36	0.72	1.19
NUTRITION & CLOTHING	1.59	4.60	0.46	0.73	1.25
FINANCIAL	1.65	2.63	0.95	0.90	1.11
AFFILIATION & SOCIAL PARTICIPATION	1.29	2.64	0.74	0.83	1.12
SHELTHER	1.07	1.63	0.85	0.86	1.05
SAFETY	0.85	0.95	0.95	0.93	1.10
BODILY HEALTH	1.26	2.68	0.36	0.86	1.11

8 Using Big Data sources for small area estimation of poverty and living condition estimates

Caterina Giusti **Stefano Marchetti** **Monica Pratesi** **Nicola Salvati**

Department of Economics and Management
University of Pisa

8.1 Introduction

The timely, accurate monitoring of social indicators, such as poverty or inequality, on a fine-grained spatial and temporal scale is a crucial tool for understanding social phenomena and policymaking, but poses a great challenge to official statistics. The idea of this work is to propose an interdisciplinary approach which is able to combine the body of statistical research in small area estimation with the body of research on the huge amounts of digital information about human activities produced by a wide range of high-throughput tools and technologies - often referred to as 'Big Data'. The main part of the work has been already published in the Journal of Official Statistics as Marchetti et al. (2015).

Generally statistical data are collected by means of sample surveys or censuses. Administrative data and registers can also be exploited to produce statistical data. Censuses are complex and expensive to carry out, so sample surveys represent a common way of collecting data. In order to draw inferences on the target population, surveys should be representative of the whole population. However, to measure social complexity with a focus on the identification and quantification of social exclusion and deprivation, there is a demand for local-level estimates of the most relevant poverty and well-being indicators. Generally the local level (local administrative area, zone of local governance) constitutes a so-called unplanned domain of estimation in sample surveys. Oversampling to increase the sample size in the domains of interest could be a feasible solution for assessing poverty and deprivation at a local level, say at Local Administrative Units levels 1 and 2 (LAU 1 and LAU 2 levels in the Nomenclature of Territorial Units for Statistics used by Eurostat), as is often required by policymakers. However, the high cost in terms of time and financial resources makes this approach impractical for obtaining accurate estimates. Big Data can represent an alternative source of data for the same areas, usually reaching a very high level of geographical detail.

We identify three possible approaches to the use of Big Data in the small area estimation framework.

The first opportunity is to use Big Data sources to create local indicators and compare them to those obtained with small area estimation methods. The idea is to reconcile data from the two independent sources - Big Data and sample surveys - to use available local measures extrapolated from Big Data to compare and benchmark measures on related aspects of the phenomenon under study (e.g. poverty and social exclusion) obtained from survey data and vice versa. Measures from Big Data sources are usually obtained very quickly; however, they can be affected by a serious self-selection bias. Conversely, small area estimates are methodologically sound, but they require timely survey and population data that can be difficult to obtain. Comparing the two alternative sets of measures referring to the same areas can provide useful insights on the potential of Big Data to benchmark small area estimates. If there is accordance between Big Data and survey data in a given small domain/area with respect to the recorded level of deprivation and poverty, then analysts and policy makers may rely on a strong evidence. Otherwise, if there is a discrepancy between the results obtained from the two sources of data, then there is a need for further investigation of those domains/areas.

The second possibility is to use Big Data sources to generate new covariates for small area models. However, the extension of the covariates to include variables such as social media search loads or remote-sensing images (e.g. in crop-yield surveys, and also in social surveys) or tracking of human mobility opens up difficulties and challenges. Due to technical problems and legal restrictions, it is unfeasible at this stage to have unit-level data that can be linked with administrative archives, census or survey data. To overcome this problem we can use the so-called area-level models, such as the Fay-Herriot model (cf. Fay and Herriot, 1979). However, attention should be paid to the fact that under the Fay-Herriot model it is assumed that the auxiliary variables are measured without error, that is, that they are available for all the areas and they come from census or archives covering the entire population of interest. When auxiliary variables come from surveys, they suffer from sampling errors and may also suffer from nonsampling errors, and thus we consider them as measured with error. Generally, auxiliary variables coming from Big Data are not measured on all (or on a big proportion) of the units of the target population, nor are they collected using a random sample. For these reasons we consider that Big Data are subject to measurement error.

The last opportunity is to use survey data to check and remove the self-selection bias of the values of the indicators obtained using Big Data. The idea is that Big Data could be used directly to measure poverty and social exclusion, appropriately taking into account the self-selection problem. We envision that survey data could be used to check and remove this bias, provided that unit-level information from Big Data sources will be available.

8.2 Big data on individuals' mobility

We used a large dataset of private vehicles in Central Italy, tracked with a GPS device. The dataset is comprised of information on approximately ten million different car journeys made by 150000 vehicles tracked during May 2011. Focusing on Tuscany, the dataset refers to 37326 vehicles, which correspond to the 1.5 percent of the total vehicles registered in Tuscany in 2011. The GPS traces were collected by OCTO Telematics S.p.a., a company that provides a data collection service for insurance companies. The GPS device is automatically turned on when the car is started, and the global trajectory of a vehicle is formed by the sequence of GPS points that the device transmits every 30 seconds to the server. When the vehicle stops no points are logged or sent. We exploited these stops to split the global trajectory into several sub-trajectories, which corresponded to the single journeys undertaken by a vehicle. Vehicle traces were then mapped on the road network and their position during the stops was associated with the census sectors, provided by the Italian National Institute of Statistics (ISTAT). In this way, each car journey was described by a tuple composed of the timestamp and a pair of coordinates corresponding to the origin and destination of the journey. We then considered several alternative measures to characterized individuals' mobility.

We define the mobility for a given vehicle ν by:

$$M_\nu = -\sum_{l_1=1}^{L}\sum_{l_2=1}^{L} p_\nu(l_1,l_2) log(p_\nu(l_1,l_2)), \tag{8.1}$$

a measure of entropy where (l_1,l_2) represents a pair of locations, $p_\nu(l_1,l_2)$ is the probability of observing a movement of vehicle ν between the locations l_1 and l_2, and L is the total number of locations. The probability $p_\nu(l_1,l_2)$ is given by the ratio between the number of trips of ν between l_1 and l_2 and the total number of trips of ν. When l_1 is equal to l_2, $p_\nu(l_1,l_2)$ is set to 0. Then, we define the mobility of an area d, $d = 1,\dots,D$ as:

$$M_d = \frac{1}{V_d}\sum_{\nu\in d} M_\nu \tag{8.2}$$

where V_d is the number of vehicles resident in area d. A vehicle is considered resident in the area where it most frequently stops during the night. The mobility value tends to zero when the vehicle ν visits few distinct locations, showing low mobility diversity. On the other hand, when the mobility measure equation (8.1) increases, it means that the vehicle ν makes journeys with several locations as destinations. We calculate the standard deviation of the mobility M_d for each area. For a given area d we measure the standard deviation of the mobility by:

$$s_{M_d} = \left[\frac{\sum_{\nu \in d} (M_\nu - M_d)^2}{V_d - 1} \right]^{1/2} \tag{8.3}$$

where M_ν and M_d are defined by equation (8.1) and equation (8.2).

Another measure of mobility we computed is based on the radius of gyration (RG), which for each vehicle measures how spread out its visited locations are from its centre of mass. The centre of mass $\boldsymbol{l}_{cm,\nu}$ of a vehicle ν is defined as a two-dimensional vector representing the weighted mean point of the locations visited by that vehicle. We can measure the mass associated with a location with its visitation frequency, obtaining the following definition:

$$\boldsymbol{l}_{cm,\nu} = \frac{1}{\sum_{i \in L} \delta_{i,\nu}} \sum_{i \in L} \delta_{i,\nu} \boldsymbol{l}_i \tag{8.4}$$

where L is the set of all the visited locations, $\boldsymbol{l}_i$ is a two-dimensional vector describing the geographic coordinates of location i and $\delta_{i,\nu}$ is its visitation frequency by vehicle ν. Then, the RG of a vehicle ν is defined as:

$$Rg_\nu = \left[\frac{1}{\sum_{i \in L} \delta_{i,\nu}} \sum_{i \in L} \delta_{i,\nu} (\boldsymbol{l}_i - \boldsymbol{l}_{cm,\nu})^T (\boldsymbol{l}_i - \boldsymbol{l}_{cm,\nu}) \right]^{1/2} . \tag{8.5}$$

We can then define the radius of gyration in area d as:

$$RG_d = \frac{1}{V_d} \sum_{\nu \in d} RG_\nu . \tag{8.6}$$

The radius of gyration provides a measure of the volume of mobility, indicating the typical distance travelled by a vehicle and provides an estimation of its tendency to move. Figure 8.1 (reported from Pappalardo et al. (2013) and referring to the Big Data mentioned above) shows that vehicles having small radius of gyration (red points) tend to concentrate their center of mass in the main urban centers of central Italy: Carrara, Pisa, Livorno, Pistoia, Empoli, Siena, Grosseto, Arezzo and the pool of towns composing the conurbation of Florence (Firenze, Prato, Sesto Fiorentino, Scandicci). Conversely, vehicles characterized by high radius of gyration (green points) are distributed in the contryside and on the coast.

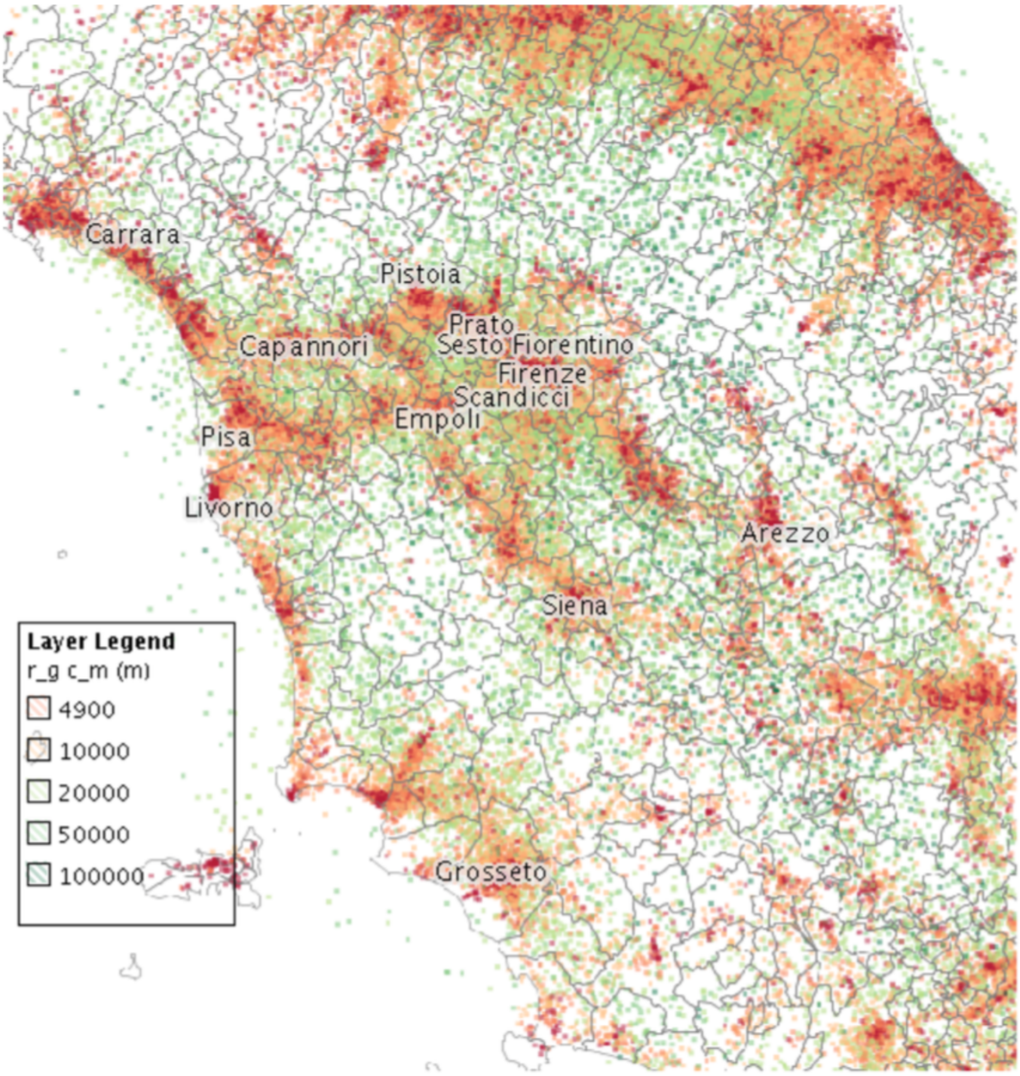


Figure 8.1: Spatial distribution of car users' center of mass on the map of central Italy. Source: Pappalardo et al. (2013).

8.3 Application: comparing Big Data with small area estimates

Figure 8.2 shows the map of the s_{M_d} and of the Head Count Ratio estimates for the ten provinces of the Tuscany region, Italy. The HCR estimates were obtained by applying the M-quantile estimators proposed by Tzavidis et al. (2010b) and Marchetti et al. (2012b) to data from EU-SILC 2008 and the Population Census 2001. M-quantile models (cf. Chambers and Tzavidis, 2006) relax the parametric assumptions of random effects models traditionally used for small area estimation (cf. Rao, 2003), which can represent an advantage in many real data applications (cf. Giusti et al., 2012, Fabrizi et al., 2014). The household-level covariates included in the model for the mean of the household equivalised income - common to the EU-SILC survey and to the population census - are the house-ownership status, the age of the head of the household, the employment status of the head of the household, the gender of the head of the household, the years of education of the head of the household and the household size. For the two maps we can see that higher levels of HCR correspond to lower levels of heterogeneity of the mobility M_d.

To better represent this relationship, Figure 8.3 shows the scatterplot of the HCR values plotted against the s_{M_d} values. Their linear correlation coefficient, used as a mere descriptive index, is equal to -0.74. Again, this result suggests that higher levels of heterogeneity of mobility (M_d), expressed by the standard deviation s_{M_d}, are in the provinces where there are lower levels of poverty. In other words, the diversification of mobility within an area with respect to its mean value can be a proxy of the level of poverty. Conversely, the mean level of the mobility it is not able to discriminate across areas because the values are all very similar and not correlated with the HCRs.

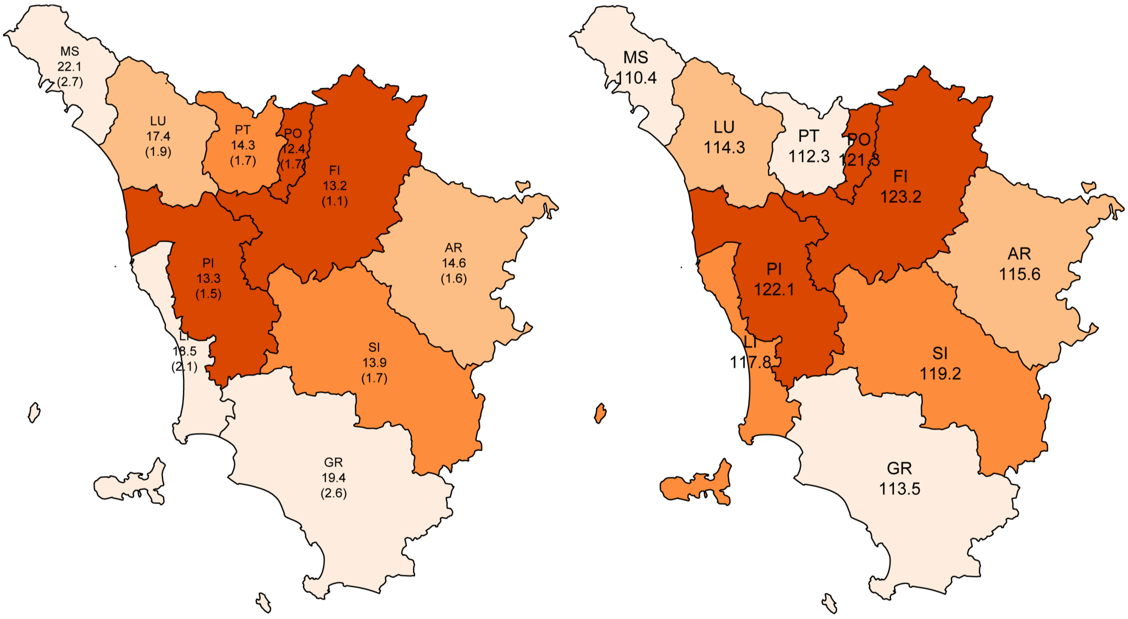


Figure 8.2: Map of the Head Count Ratio (left) and its standard error (in parenthesis) and of the standard deviation of the mobility (right) for the the provinces of the Tuscany region, Italy.

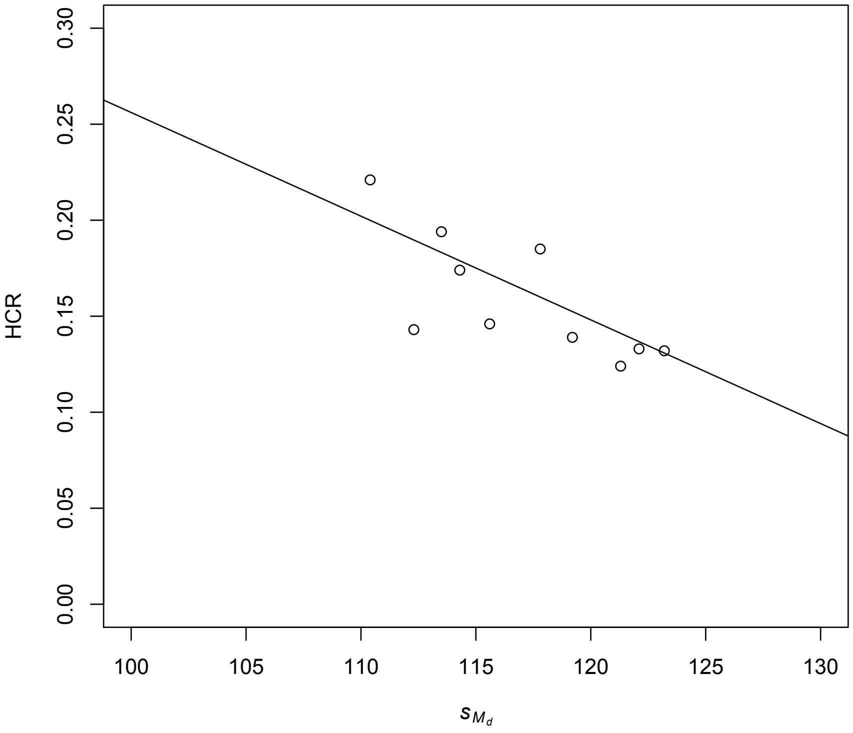


Figure 8.3: Scatterplot of the standard deviation of the mobility vs. estimates of the HCR at province level in Tuscany

8.4 Application: using Big Data as covariates in small area models

As a second approach to the joint use of Big Data and small area estimation models, we carried out an application where an extension of the Fay-Herriot model - an area-level small area model - is used to produce estimates of the HCR and of the mean household equivalised income for the LLSs of the Tuscany region using area-level data from the EU-SILC survey 2011 and, as covariate information, data from the EU-SILC survey itself and from Big Data on mobility. To use these data as covariate information, we propose to use the modified version of the Fay-Herriot model proposed by Ybarra and Lohr (2008) to allow for measurement error in the auxiliary variables.

Ybarra and Lohr (2008) assume that for a direct estimator y_d of the target variable Y_d in area d, under the sampling design, $E[y_d] = Y_d$, and that the auxiliary data source provides an estimator $\hat{X}_d$ of a p-vector X_d of population characteristics, where the estimator $\hat{X}_d$ has mean squared error $MSE(\hat{X}_d) = C_d$ under the sample design. They show that when the auxiliary variables are measured with error, the traditional Fay-Herriot estimator can be worse than the direct estimator in terms of precision and in addition the estimated mean squared error gives a misleading notion of precision. Suppose that X_d is the true value of the auxiliary variable in small area d available for small area estimation. Since X_d may be measured with error, we substitute an estimator $\hat{X}_d$ for X_d and use the following model:

$$y_d = \hat{X}_d^T \beta + r_d(\hat{X}, X_d) + e_d \tag{8.7}$$

where $r_d(\hat{X}, X_d) = u_d + (X_d - \hat{X}_d)^T \beta$, with $u_d \sim N(0, \sigma_u^2)$ and the random error $e_d \sim N(0, \psi_d^2)$, with known ψ_d. Here, u_d is independent from both e_d and $\hat{X}_d$, and random variables in different small areas are independent. They also assume that $\hat{X}_d$ and y_d are independent for each area, as when X_d and Y_d are estimated using different data sources. In our application this is the case for Big Data auxiliary variables, while for auxiliary variables from the EU-SILC survey this hypothesis is violated. However, this problem can be solved changing the model according to Ybarra (2003). The resulting EBLUP (Empirical Best Linear Unbiased Predictor) is:

$$\hat{Y}_{dME} = \hat{\gamma}_d y_d + (1 - \hat{\gamma}_d) \hat{X}_d^T \hat{\beta} \tag{8.8}$$

where $\hat{\gamma}_d = (\hat{\sigma}_u^2 + \hat{\beta}^T C_d \hat{\beta})/(\hat{\sigma}_u^2 + \hat{\beta}^T C_d \hat{\beta} + \psi_d^2)$ and the regression vector β and the variance component σ_u^2 are estimated according to an iterative procedure for the modified least squares as in Cheng and Van Ness (1999). Ybarra and Lohr (2008) prove the consistency of equation (8.8) and propose an analytic and a jackknife estimator of $MSE(\hat{Y}_{dME})$. In our application we are interested in producing mean estimates of the household equivalised income ■ equivalised according to the OECD (Organisation for Economic Co-operation and Development) modified scale (cf. Hagenaars et al., 1994) ■ and estimates of the HCR for the 57 LLSs in Tuscany. Note that 24 out of the 57 LLSs are "out-of-sample areas" with a zero sample size in the EU-SILC 2011. To compute the direct estimates of the mean household incomes and of the HCRs we used data from the 2011 census wave of EU-SILC, available at LLS level. Data from the same survey was also considered as covariate information.

As covariate information available for all 57 LLSs we also used Big Data on individuals' mobility, under the hypothesis that mobility data could be predictive of well-being measures. We used two different models to estimate the small area income means and HCRs. To estimate the mean incomes (Y_d) using estimator $\hat{Y}_{dME}$, let $\hat{X}_d$ be the vector of the auxiliary variables of area d: it contains a constant term, the direct estimate of the proportion of male as the head of the household, the direct estimate of the mean of the squared metres of the house (both from the EU-SILC 2011 survey) and, finally, the values of the RG_d (from Big Data sources). Let C_d be the corresponding variance-covariance matrix of the auxiliary variables, with the covariances set to zero. Let y_d be the direct estimate in area d of the mean of the household equivalised income and ψ_d its standard deviation. In the model for the HCR the auxiliary variables vector $\hat{X}_d$ contains a constant term, the direct estimate of the mean of the age of the head of household (from EU-SILC 2011) and the mobility index M_d (from Big Data on mobility). Here y_d is the direct estimate of the HCR in area d.

An important problem in small area estimation is the synthetic prediction for out-of-sample areas: that is, areas where there are no sampled units, even if there are population units with the characteristics of interest in those areas. The conventional approach for estimating a small area characteristic, say the mean, is the synthetic estimation (cf. Rao, 2003): $\hat{Y}_{d,OUT} = X^T_{d,OUT}\hat{\beta}$, where $X_{d,OUT}$ is the auxiliary information for the out-of-sample area d and $\hat{\beta}$ is the vector of estimated coefficients under a small area model. In the application presented here the problem is serious, since there are 24 out-of-sample areas (42 percent of the total number of small areas). Moreover, the predictor $\hat{Y}_{d,OUT} = \hat{X}^T_{d,OUT}\hat{\beta}$ according to equation (8.8) cannot be applied because the EU-SILC auxiliary variables selected in our models are not available for the out-of-sample areas. In contrast, Big

Data auxiliary variables are available for all the areas. One possible synthetic predictor is $\hat{Y}_{d,OUT} = \hat{\hat{X}}^T\hat{\beta} + \hat{X}_{d,BD}\hat{\beta}_{BD}$, where $\hat{\hat{X}}$ is the matrix of the direct estimators of the EU-SILC auxiliary variables at a regional level, $\hat{X}_{d,BD}$ is the value of the Big Data auxiliary information for area d and finally $\hat{\beta}$ and $\hat{\beta}_{BD}$ are the estimated regression coefficients (see Giusti et al., 2012 for an example). Accordingly, using Big Data it is possible to obtain area-specific synthetic estimates for the out-of-sample areas, taking into account the variability between areas that cannot be specified by only basing predictions on the values of $\hat{\hat{X}}$. This represents one of the major advantages in the use of Big Data sources in small area estimation.

Finally, to estimate the mean squared error of equation (8.8) for both sampled and out-of-sample areas we use a parametric bootstrap approach, since the jackknife approach described in Ybarra and Lohr (2008) was too unstable with our data, often producing negative estimates of the mean squared error. In the parametric bootstrap we first estimated β and σ_u^2, then we parametrically generated the errors $u_d^* \sim N(0, \hat{\sigma}_u^2)$ and $e_d^* \sim N(0, \psi_d^2)$. Using these random errors and the matrix of auxiliary variables, we generated the bootstrap true values $Y_d^* = \hat{X}_d^T\hat{\beta} + u_d^*$ and the bootstrap direct estimates $y_d^* = Y_d^* + e_d^*$. In the next step, we generated a bootstrap matrix of auxiliary variables with errors $\hat{X}_d^* = \hat{X}_d + \epsilon_d$, where $\epsilon_d \sim N_p(0, C_d)$ with N_p a multivariate normal of dimension p. Using equation (8.8) with $\hat{X}_d^*$ and y_d^* we obtained a bootstrap estimate $\hat{Y}_{dME}^*$ of Y_d^*. Repeating this process B times to obtain B values of $\hat{Y}_{dME}^{*b}$ and Y_d^{*b}, $b = 1, \ldots, B$, the bootstrap mean squared error estimator of $\hat{Y}_{dME}$ was

$$mse(\hat{Y}_{dME}) = B^{-1}\sum_{b=1}^{B}(\hat{Y}_d^{*b} - Y^{*b})^2. \tag{8.9}$$

In the application of this article we have used $B = 500$.

We checked the performance of this bootstrap mean squared error estimator with a small simulation following the setting used in Ybarra and Lohr (2008) in their simulation study. The bootstrap scheme seemed to work properly, showing an expected slight underestimation of the real (i.e. Monte Carlo) mean squared error.

As an alternative to the bootstrap, for the out-of-sample areas we predicted the ψ_d values using a linear model based on the same variables used in the estimation process (cf. Wolter, 2007). This method is feasible given that data coming from Big Data sources are available for all the small areas.

Results for the means of both the equivalised income and for the HCR are mapped in

Figure 2. These estimates referring to the LLSs show intraregional differences that would be lost if the scope of the analysis was limited to the regional level.

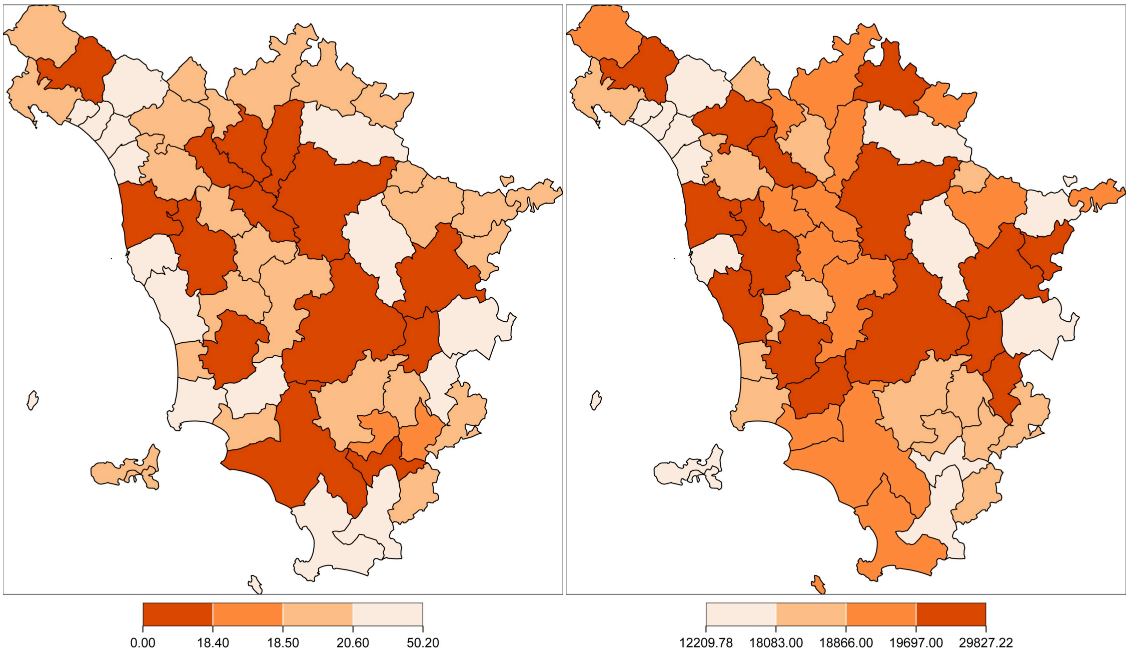


Figure 8.4: Estimates of the mean equivalised income in Euros (right) and of the HCR (left) for the Local Labour Systems of Tuscany region. Small area estimates based on EU-SILC 2011 and Mobility Data 2011. Out-of-sample areas are estimated using a synthetic estimator

What is even more important is that for both the target parameters we achieved a remarkable gain in terms of precision with respect to the direct estimates. Even if this gain is marginally overestimated because the bootstrap mean squared error of $\hat{Y}_{dME}$ underestimates the real mean squared error, the gain in precision is evident. Figure 8.5 shows a comparison of the bootstrap mean squared error estimates of $\hat{Y}_{dME}$ and the mean squared error estimates of the direct estimates y_d for both the mean equivalised income (right) and the HCR (left). Since the mean squared errors for the direct estimators are only available for the sampled areas, we only report these ratios for the 33 sampled areas. A gain in precision is observed for all the areas. In most of the areas the gain in precision is about 5% − 20% for the mean and 10% − 40% for the HCR. In some areas the gain is more than 50%. However, the mean squared error estimator of the small area estimator $\hat{Y}_{dME}$ should be treated with caution due to its observed underestimation. Nonetheless, these first promising results encourage further research on this topic.

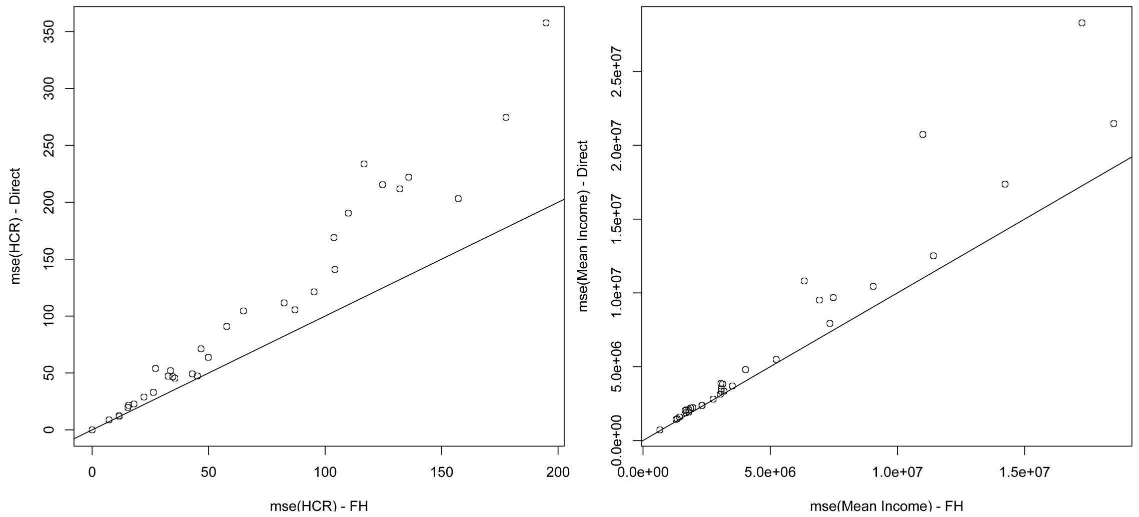


Figure 8.5: The plot on the right shows the mean squared error estimates of the small areas, obtained using equation (8.9) with B=500 bootstrap replications, vs. the direct estimates of the mean of the equivalised income; the plot on the left shows the same for the HCR estimates. Results are reported for the 33 Local Labour Systems (LLSs) sampled in the EU-SILC 2011 survey for Tuscany.

9 PCA and Complex Survey Designs

Duncan Smith **Natalie Shlomo**

Cathie Marsh Institute for Social Research & Social Statistics, School of Social Sciences

University of Manchester

9.1 Introduction

Principal components analysis (PCA) is an important statistical technique that allows analysts to explore high dimensional data. PCA finds an optimal way of combining variables into a smaller number of subsets through an orthogonal transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables called principle components. When carrying out PCA on sample data, a covariance (correlation) matrix and a vector of means are estimated from the data, and these are used to generate the PCA. Standard estimation methods for these components are based on the sample data being representative of the population (generally, a simple random sample). Complex survey designs do not generate simple random samples from a population. The population might be partitioned into strata and the primary sampling unit might contain a cluster of elements with further sub-sampling of elements within the cluster. In particular, elements might not be selected with equal probability, and sampling may be with or without replacement. The research question addressed here is how PCA can be performed with data arising from complex survey designs? We do not consider specific solutions for specific types of design. Rather, we consider whether there exists a general, pragmatic approach that can be used in a wide variety of settings.

9.2 Principal Component Analysis

PCA transforms a set of observations on p (possibly correlated) variables into a set of observations on $k, k < p$ linearly uncorrelated variables. The transformation can be thought of as a translation to mean centre the data, followed by a rotation of the axes (corresponding to the original variables). Mutual orthogonality of the axes is maintained. The rotation is such that the first principal component (axis) explains as much of the total variance in the data as possible, the second principal component explains as much of the remaining variance in the data as possible (subject to the orthogonality constraint), and so on. PCA is performed on the p by p covariance or correlation matrix for the data. The aim is to be able to summarize the data and reduce the dimensionality, losing as little information as possible. More details are available in text books. Many statistical packages will perform

PCA. The output is a set of eigenvectors which can be used to transform data to the new coordinate system. These can be used to generate the corresponding eigenvalues (loadings) which are the variances in the data explained by the corresponding principle components. Factor Analysis is similar to PCA with respect to reducing the dimensionality of the data and typically the data is first centred and standardized and hence the eigenvectors are calculated based on the correlation matrix.

9.3 Covariance Matrices

The entry in the covariance matrix with index (i, j) is the (estimated) covariance of the i th and j th variables. For i = j the covariance is simply the variance.

$$Cov(x, y) = E(xy) - E(x)E(y) \tag{9.1}$$

or equivalently,

$$Cov(x, y) = E(x - E(x)) \times E(y - E(y)). \tag{9.2}$$

The former can be more convenient mathematically, whilst the latter can be more convenient for calculations. The covariance can be estimated from a simple random sample of data via

$$\widehat{Cov}(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} \tag{9.3}$$

or via the above equation multiplied by a factor $n/(an-1)$ to correct for the bias introduced by the use of estimated means.

9.4 PCA with data arising from complex survey designs

The question is how best to perform PCA with data that are not simple random samples. For any complex survey design we can generate inclusion probabilities for the sampled data. These can also reflect response rates, as well as the design itself. An intuitive approach is to simply include the inclusion probabilities in the estimation of the covariance (correlation) matrix and means. Estimation of population totals using inclusion probabilitie π_i is generally performed using Horvitz-Thompson estimation

$$\widehat{T} = \sum_{i=1}^{n} \pi_i^{-1} x_i \tag{9.4}$$

The estimate for the population mean is thus.

$$\widehat{E}(x) = N^{-1} \sum_{i=1}^{n} \pi_i^{-1} x_i \tag{9.5}$$

If the population total, N, is unknown it can be estimated

$$\widehat{N} = \sum_{i=1}^{n} \pi_i^{-1} \tag{9.6}$$

The estimated covariance matrix is

$$\widehat{Cov}(x,y) = \frac{\sum_{i=1}^{n} \pi_i^{-1} x_i y_i}{\sum_{i=1}^{n} \pi_i^{-1}} - \frac{\sum_{i=1}^{n} \pi_i^{-1} x_i}{\sum_{i=1}^{n} \pi_i^{-1}} \frac{\sum_{i=1}^{n} \pi_i^{-1} y_i}{\sum_{i=1}^{n} \pi_i^{-1}} \tag{9.7}$$

For simple random sampling, the weights are all equal to n/N and the estimator becomes equation (9.3). In fact, this is true for any design that generates equal inclusion probabilities regardless of their value.

If we have a population that consists of strata with distinct covariances and means, then it can be shown that the population covariance is given by

$$Cov(x,y) = \left(\sum_{h=1}^{H} p_j \left| Cov_h(x,y) + E_h(x)E_h(y) \right| \right) - E(x)E(y) \tag{9.8}$$

where the p_j are the marginal probabilities for the strata indexed $h = 1, ..., H$ and

$$E(x) = \sum_{h=1}^{H} p_j E_h(x), E(y) = \sum_{h=1}^{H} p_j E_h(y) \tag{9.9}$$

For disproportionate stratification we can simply pool all the data and the estimator of the population covariance would be equivalent to equation (9.7) where we take into account the design via the inclusion probabilities.

Kriegel et al. (2008) proposes an essentially identical weighting scheme to increase the robustness of PCA in the presence of outliers. The issues there are how to identify outliers so that they can be down-weighted, and deciding upon the appropriate degree of down-weighting. Here we can derive appropriate weights from the sampling design, although it is clear that they might be adjusted to accommodate non-response and the issues considered by Kriegel et al.. Nathan and Holt (1980) consider exactly the approach presented here - 'An obvious way to utilize the sample design information is to base the estimators on weighted sample means, variances and covariances, where the weights are the inverses of the sample inclusion probabilities'. Skinner et al. (1986) also considered the inclusion of the design weights in the estimation of the covariance matrices and showed that estimators are biased for non self-weighted sample designs.

9.5 Intracluster Correlation

The intracluster correlation coefficient (ICC) describes how strongly observations within clusters resemble each other. Here we use the ANOVA framework definition,

$$ICC = \frac{SST - SSW}{SST} \tag{9.10}$$

where SST is the sum of squared deviations from the global means and SSW is the sum of squared deviations from the relevant cluster means. If the cluster means were equal, then the ICC would equal 0. If there was no variance within cluster (as in the degenerate case of a single observation in each cluster), then the ICC would be 1. In complex sample designs the records within a given cluster (primary sampling unit) will often tend to be more similar than records selected across clusters. We examine the impact of the intracluster correlation in cluster sample designs in the estimation of population covariance/correlation matrices for PCA in the simulation studies described in the next section.

9.6 Simulation Study

Clustering may impact on the estimation of the population covariance matrices but may have less influence on correlation matrices which standardize the covariance matrices so they are measurement invariant. In the simulation study, we propose to implement a bootstrap bias correction to adjust for potential bias arising from the clustering.
The simulation study has the following set up:

- Generate a population of 3 strata, each strata having 100 clusters of size 20, $N = 6000$.

- Generate a vector composed of 3 variables: $\mathbf{X} = (x, y, z)$ as follows:

  1. Generate Σ_ϵ where for strata 1: $\sigma^2_\epsilon = 1$, for strata 2: $\sigma^2_\epsilon = 4$ and for strata 3: $\sigma^2_\epsilon = 9$ for all $\mathbf{X}$ and the covariance matrix is defined with a correlation of 0.5 for each pair of variables.

  2. From Σ_ϵ generate ϵ_{hij} from a $MVN(\mathbf{0}, \Sigma_\epsilon)$ for $h = 1, \ldots, 3$ strata, $i = 1, \ldots, 100$ clusters and $j = 1, \ldots, 20$ units.

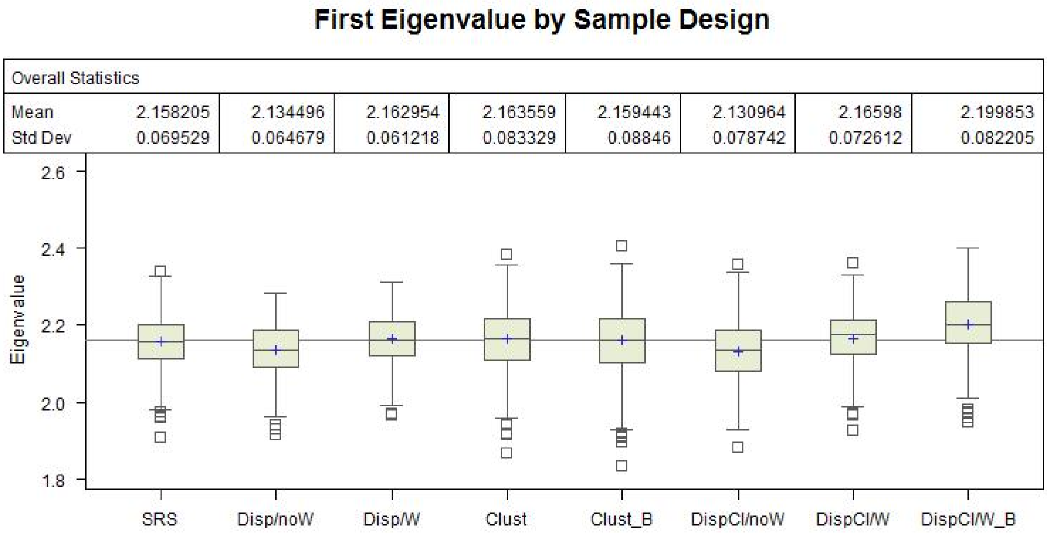
  3. We consider intracluster correlation coefficients (ICC) of varying magnitudes: 0.2, 0.5 and 0.8 and generate μ_{hi} from a $MVN(\mathbf{0}, \Sigma_\mu)$ where $\sigma^2_\mu = ICC/(1 -$

$ICC)\sigma_\epsilon^2$.

4. Generate $\mathbf{X}_{hij} = \beta\phi_{hij} + \mu_{hi} + \epsilon_{hij}$, $\beta = 1$ and ϕ_{hij} is $MVN(\mathbf{0}, \mathbf{1})$.

- Draw 500 samples under varying designs: simple random sample of 1:20, simple random samples drawn in each strata with differential weights: strata 1 has a sample fraction of 1 : 60, strata 2 has a sample fraction of 1 : 30 and strata 3 has a sample fraction of 1 : 10; 15 clusters drawn with equal probability and all units in the selected cluster are kept in the sample; disproportionate stratification of clusters selected using the same sample fractions as above and all units in the selected cluster are kept in the sample.

- For each of the cluster designs, we carry out a bootstrap bias correction based on 100 samples drawn with replacement from each of the original 500 samples.

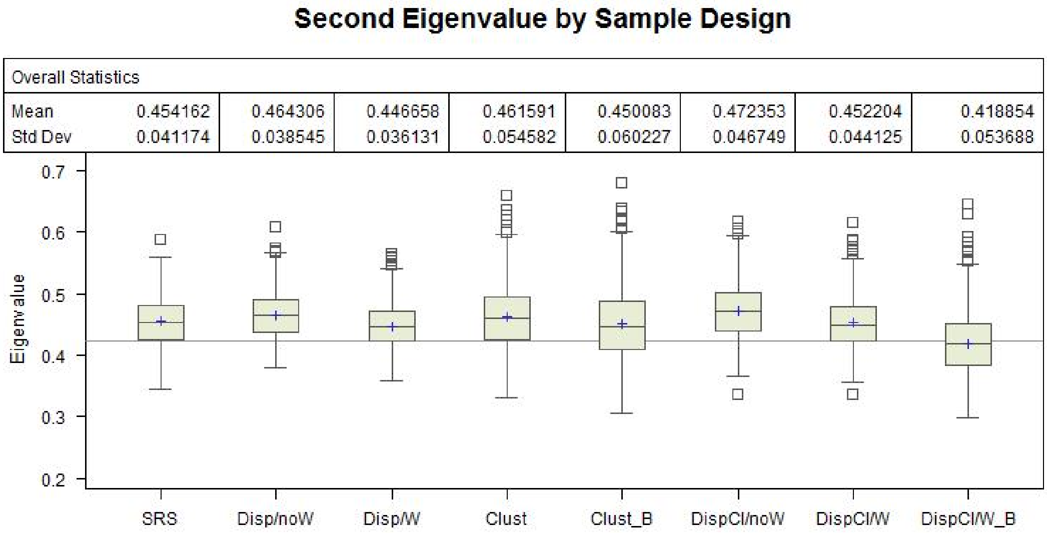


First Eigenvalue by Sample Design

Overall Statistics								
Mean	2.158205	2.134496	2.162954	2.163559	2.159443	2.130964	2.16598	2.199853
Std Dev	0.069529	0.064679	0.061218	0.083329	0.08846	0.078742	0.072612	0.082205

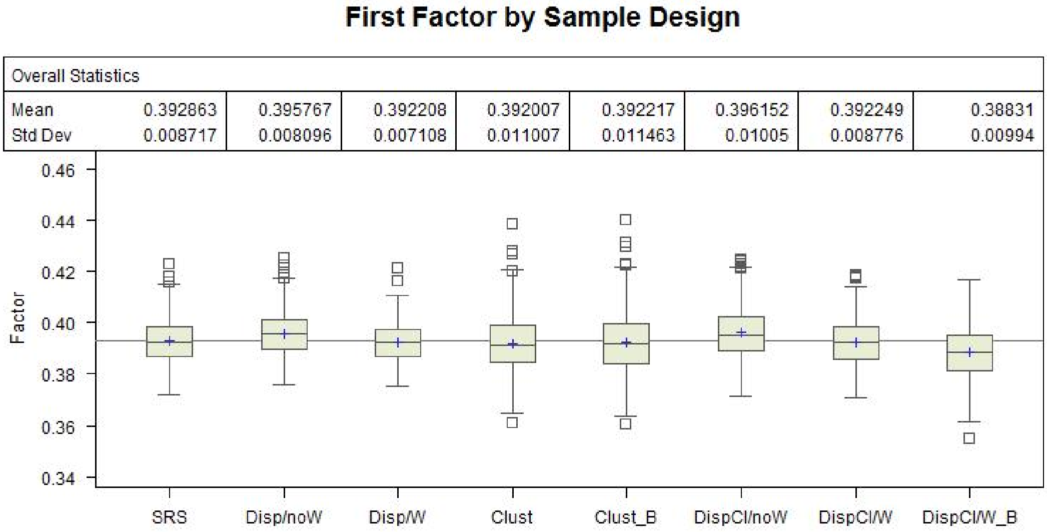
Figure 9.1: First Eigenvalue ICC=0.2

Figures 9.1 to 9.4 contain the first and second eigenvalues and the first and second factors when the ICC is equal to 0.2, Figures 9.5 to 9.8 contain the first and second eigenvalues and the first and second factors when the ICC is equal to 0.5 and finally Figures 9.9 to 9.12 contain the first and second eigenvalues and the first and second factors when the ICC is equal to 0.8. The horizontal line in each of the figures represents the true value in the population. Each of the designs in the figures have the following notation: SRS for the simple random sampling; Disp for the disproportionate stratification of units where noW



Overall Statistics								
Mean	0.454162	0.464306	0.446658	0.461591	0.450083	0.472353	0.452204	0.418854
Std Dev	0.041174	0.038545	0.036131	0.054582	0.060227	0.046749	0.044125	0.053688

Figure 9.2: Second Eigenvalue ICC=0.2



Overall Statistics								
Mean	0.392863	0.395767	0.392208	0.392007	0.392217	0.396152	0.392249	0.38831
Std Dev	0.008717	0.008096	0.007108	0.011007	0.011463	0.01005	0.008776	0.00994

Figure 9.3: First Factor ICC=0.2

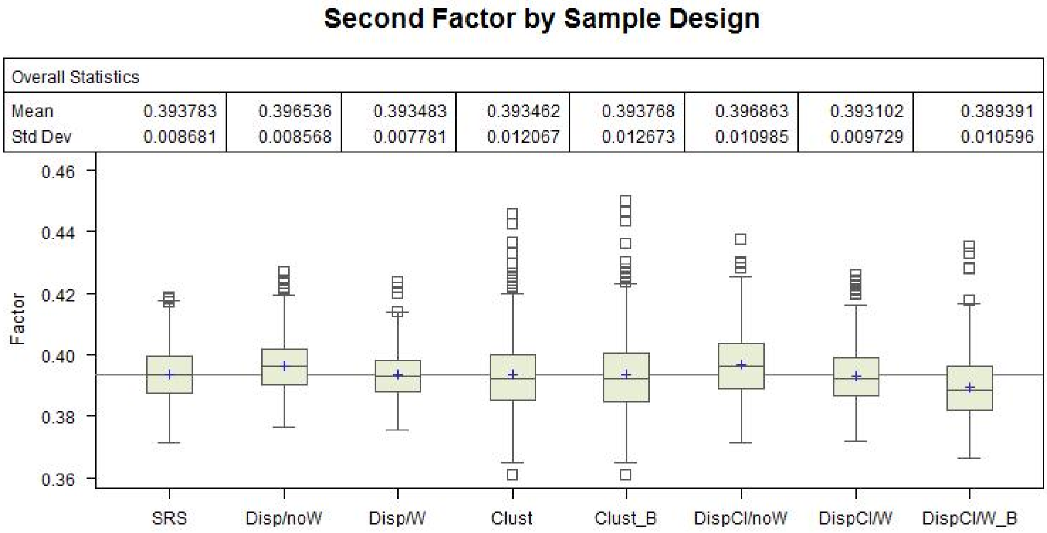


Figure 9.4: Second Factor ICC=0.2

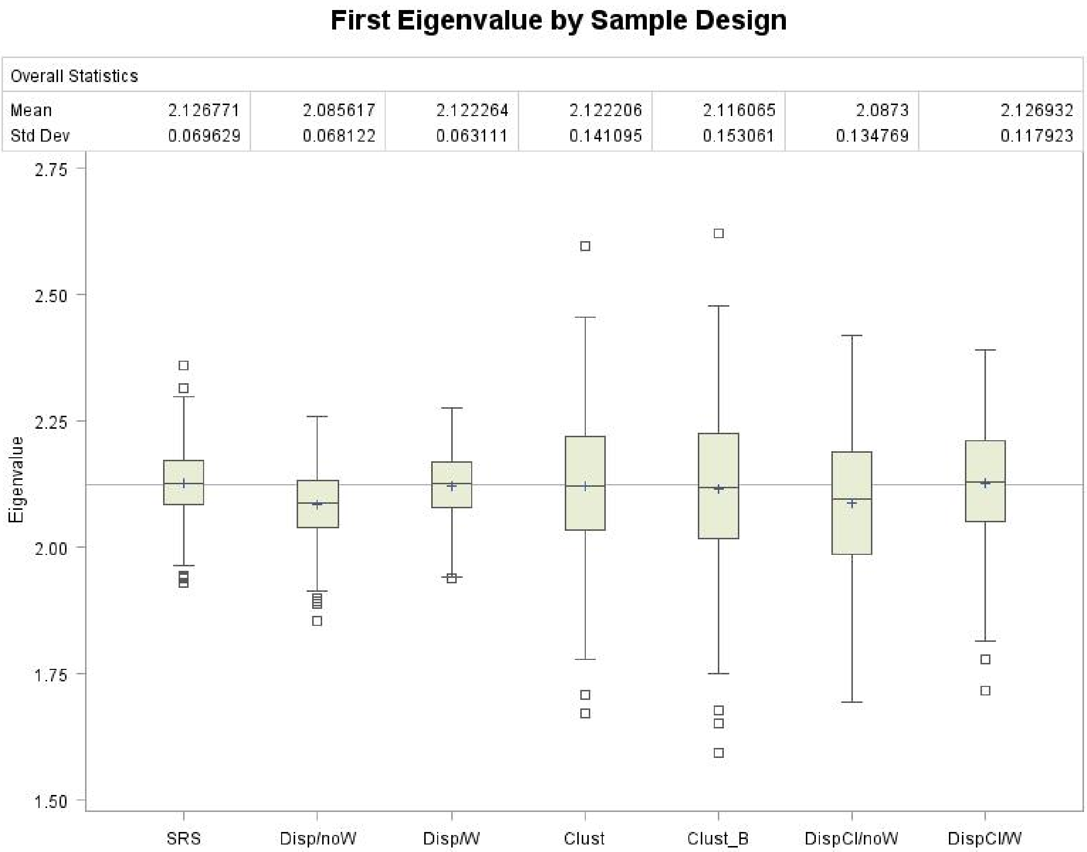
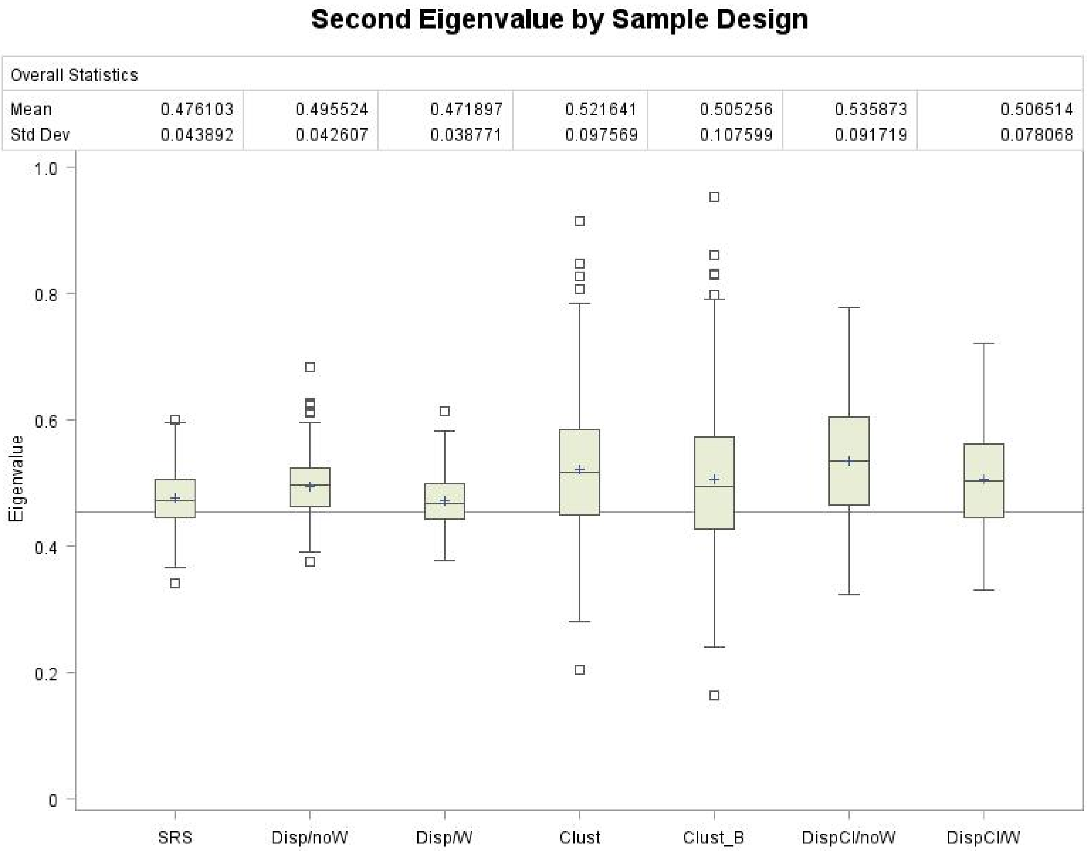


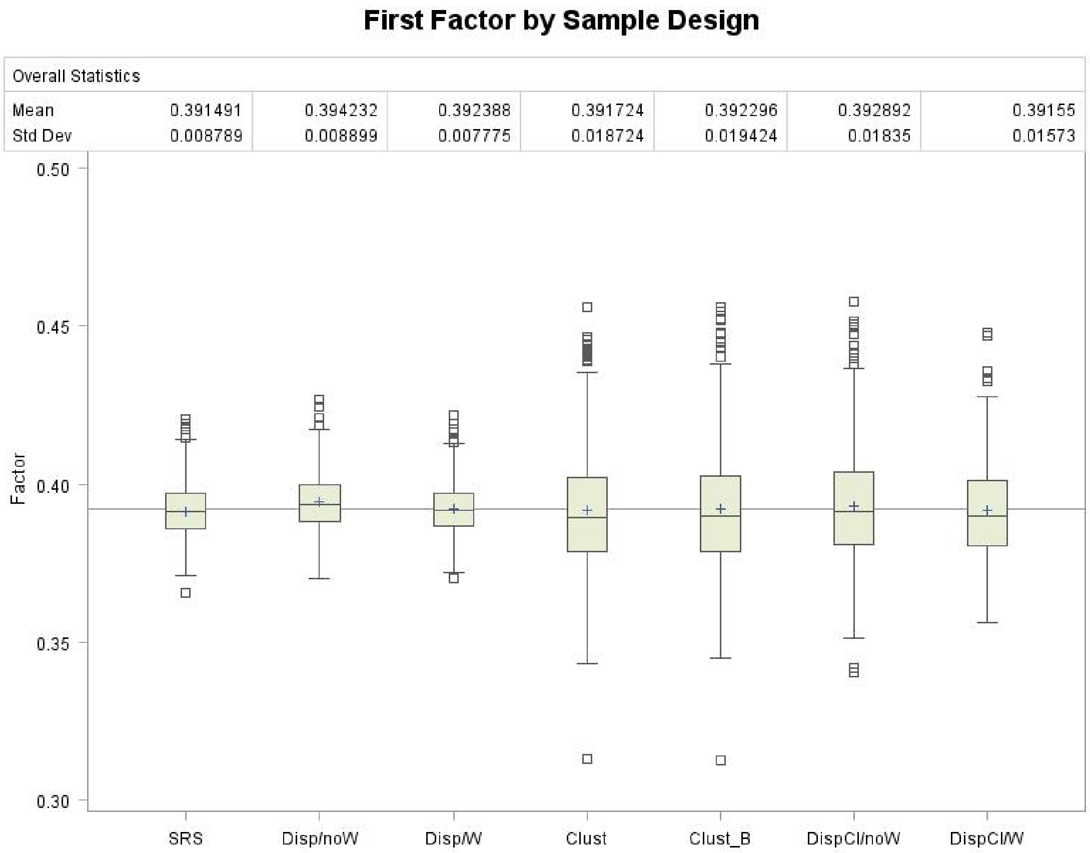
Figure 9.5: First Eigenvalue ICC=0.5



Overall Statistics

Mean	0.476103	0.495524	0.471897	0.521641	0.505256	0.535873	0.506514
Std Dev	0.043892	0.042607	0.038771	0.097569	0.107599	0.091719	0.078068

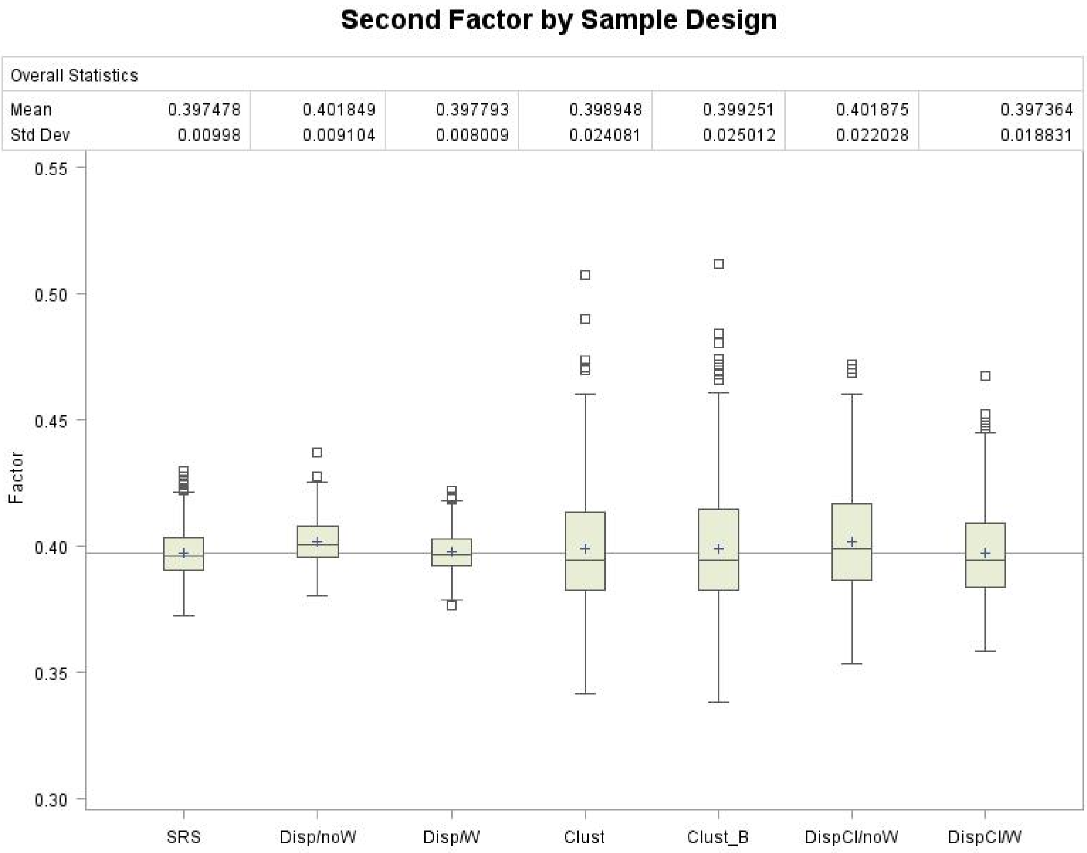
Figure 9.6: Second Eigenvalue ICC=0.5



First Factor by Sample Design

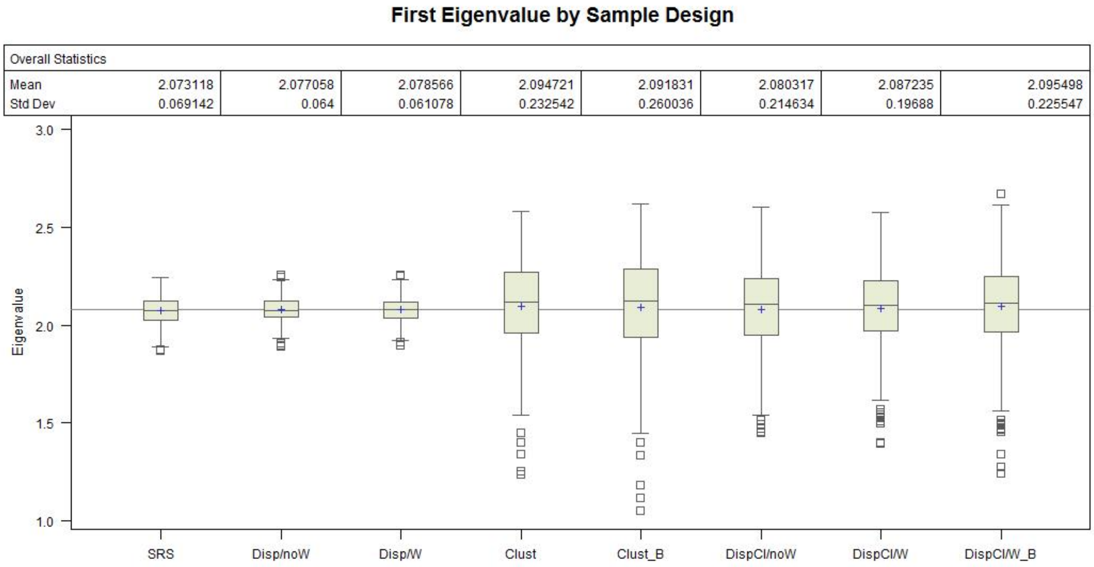
Overall Statistics							
Mean	0.391491	0.394232	0.392388	0.391724	0.392296	0.392892	0.39155
Std Dev	0.008789	0.008899	0.007775	0.018724	0.019424	0.01835	0.01573

Figure 9.7: First Factor ICC=0.5



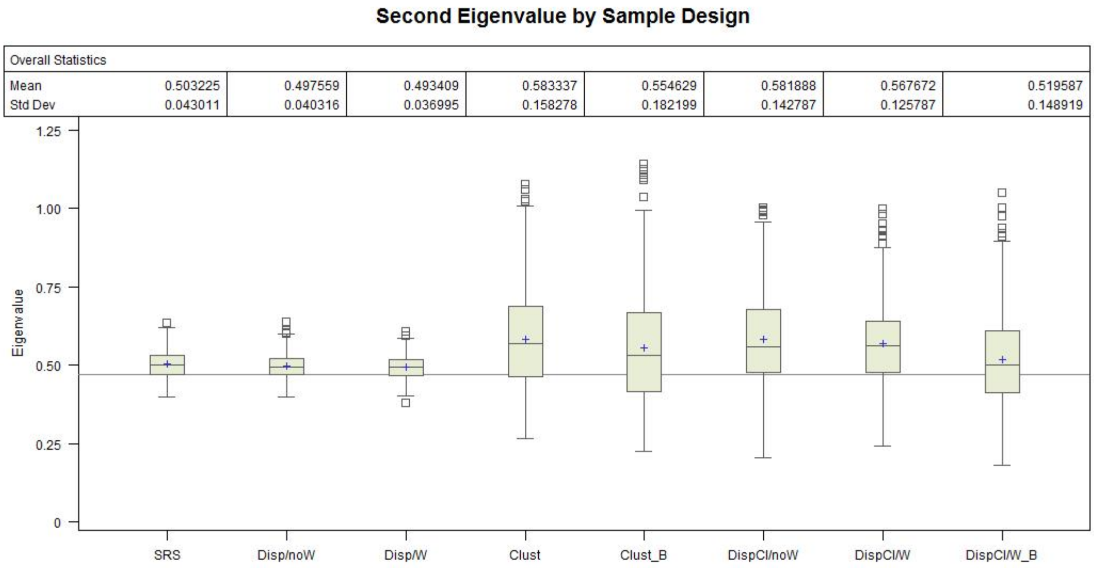
Overall Statistics							
Mean Std Dev	0.397478 0.00998	0.401849 0.009104	0.397793 0.008009	0.398948 0.024081	0.399251 0.025012	0.401875 0.022028	0.397364 0.018831

Figure 9.8: Second Factor ICC=0.5



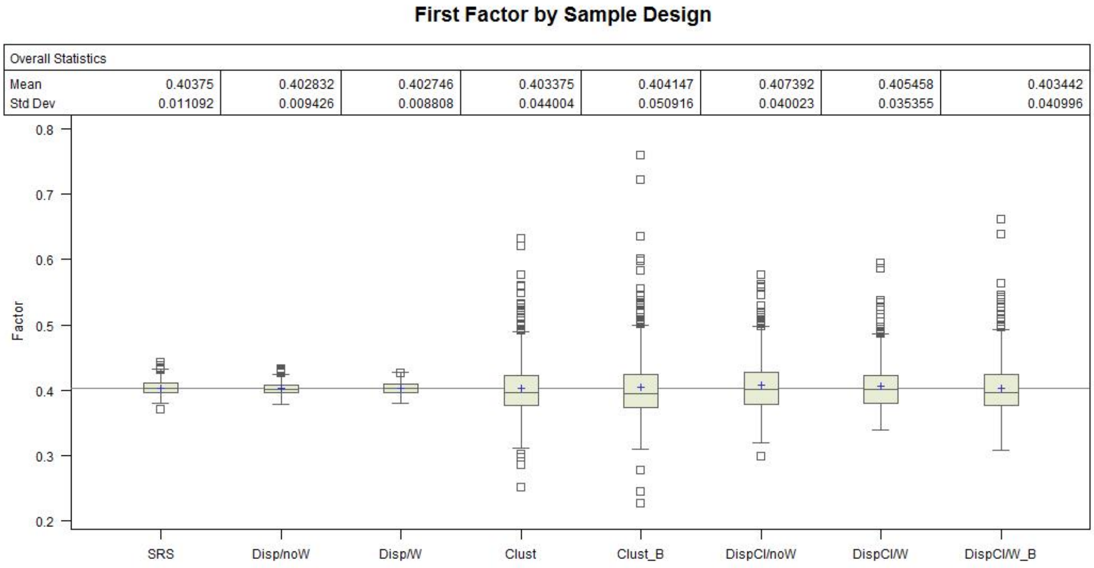
Overall Statistics								
Mean Std Dev	2.073118 0.069142	2.077058 0.064	2.078566 0.061078	2.094721 0.232542	2.091831 0.260036	2.080317 0.214634	2.087235 0.19688	2.095498 0.225547

Figure 9.9: First Eigenvalue ICC=0.8



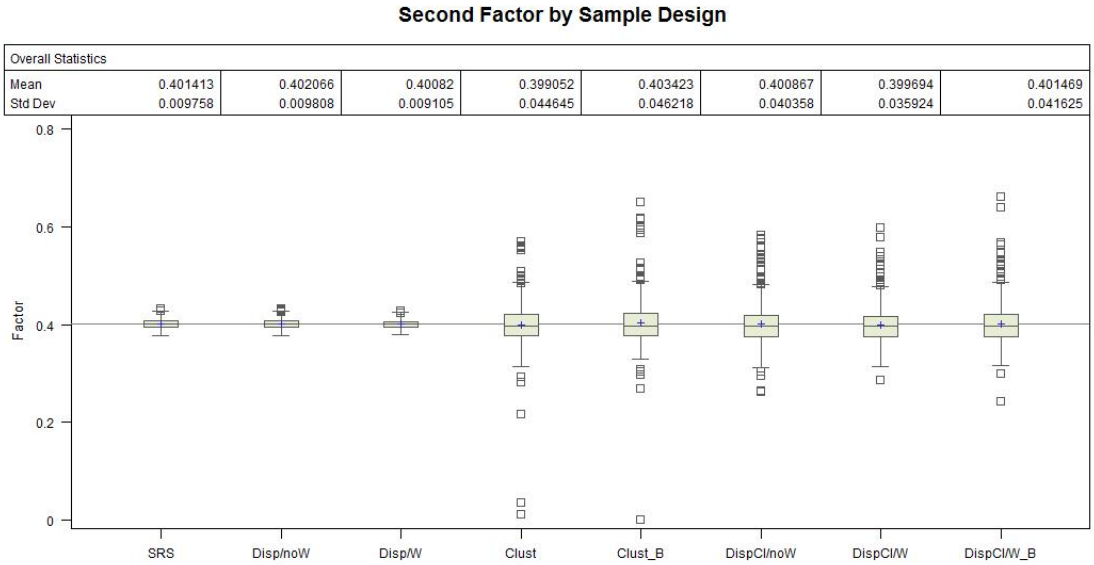
	SRS	Disp/noW	Disp/W	Clust	Clust_B	DispCl/noW	DispCl/W	DispCl/W_B
Mean	0.503225	0.497559	0.493409	0.583337	0.554629	0.581888	0.567672	0.519587
Std Dev	0.043011	0.040316	0.036995	0.158278	0.182199	0.142787	0.125787	0.148919

Figure 9.10: Second Eigenvalue ICC=0.8



	SRS	Disp/noW	Disp/W	Clust	Clust_B	DispCl/noW	DispCl/W	DispCl/W_B
Mean	0.40375	0.402832	0.402746	0.403375	0.404147	0.407392	0.405458	0.403442
Std Dev	0.011092	0.009426	0.008808	0.044004	0.050916	0.040023	0.035355	0.040996

Figure 9.11: First Factor ICC=0.8



Overall Statistics								
Mean	0.401413	0.402066	0.40082	0.399052	0.403423	0.400867	0.399694	0.401469
Std Dev	0.009758	0.009808	0.009105	0.044645	0.046218	0.040358	0.035924	0.041625

Figure 9.12: Second Factor ICC=0.8

denotes the unweighted covariance matrix in the PCA and W denotes the weighted covariance matrix in the PCA; Clust for the equal probability cluster design where B denotes the bootstrapped bias corrected estimate; DispCl for the disproportionate cluster design where noW denotes the unweighted covariance matrix in the PCA, W denotes the weighted covariance matrix in the PCA and where specified, W_B denotes the bootstrapped bias corrected estimate.

9.7 Discussion

Given the high correlation between the variables in **X**, the first eigenvalue describes most of the variance but we examine the first two eigenvalues as well as the first two factors from the PCA. Under ICC=0.2, the first eigenvalue and first factor under simple random sampling gives unbiased estimates. The simple random sampling with varying sample fractions in each strata provides biased estimates if the sample weights are not taken into account in the estimation of the population covariance matrix for the PCA. If the sample weights are used, we obtain unbiased estimates. Given the low ICC of 0.2, the equal probability cluster design provides similar results to the bootstrap bias corrected estimates showing a slight improvement for the first factor. Under the cluster design where the clusters are sampled within strata under varying sample fractions, the weighted covariance matrix provides unbiased results but it is clear that the bootstrap bias corrected estimates over-compensated for the first eigenvalue and first factor. Looking at the second eigenvalue there is an upward bias even under simple random sampling with the bootstrap bias corrected estimate for the cluster designs showing an improvement. However, the second factor is unbiased under simple random sampling and when taking into account the weighted covariance matrices under the differential sample designs. We also see that the bootstrap bias corrected estimates are over-compensating for the bias.

Under ICC=0.5, we note that the bootstrap bias corrected estimates for the disproportionate cluster sampling within strata was not carried out since there was little evidence of a clustering effect at this ICC as well. We see similar patterns as described for the case of ICC=0.2.

For ICC=0.8 representing an extreme value of an ICC, the bootstrap bias corrected estimate for the first eigenvalue and first factor is having little effect on correcting the bias which in any case remains small. However, it does seem to increase the variation. Again we see similar patterns of a bias in the second eigenvalue with the bootstrap bias corrected estimator improving the bias, but for both factors there is little evidence of bias arising from the cluster designs.

From the simulation study, there does not seem to be any indication that the clustering, even for the high intracluster correlation, is having much impact on the estimation of the PCA. It is clear that sampling weights are necessary to correct for bias arising from the survey design.

Bibliography

Abramowitz, M. and **Stegun, I. A.** (**1964**): Handbook of mathematical functions: with formulas, graphs, and mathematical tables. Washington, D.C.: Courier Corporation.

Addabbo, T., **Di Tommaso, M.** and **Maccagnan, A.** (**2014**): *Gender differences in italian children capabilities.* In: Feminist Economics, 20, pp. 90–121.

Addabbo, T. and **Facchinetti, G.** (**2013**): *The fuzzy logic and the capability approach.* Technical report 106, CAP Paper.

Alanya, A., **Wolf, C.** and **Sotto, C.** (**2015**): *Comparing multiple imputation and propensity-score weighting in unit-nonresponse adjustments a simulation study.* In: Public Opinion Quarterly, p. nfv029.

Alfons, A., **Filzmoser, P.**, **Hulliger, B.**, **Kolb, J.-P.**, **Kraft, S.**, **Münnich, R.** and **Templ, M.** (**2011**a): *Synthetic Data Generation of SILC Data.* Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI.
URL `http://ameli.surveystatistics.net`

Alfons, A., **Kraft, S.**, **Templ, M.** and **Filzmoser, P.** (**2011**b): *Simulation of close-to-reality population data for household surveys with application to EU-SILC.* In: Statistical Methods & Applications, 20 (3), pp. 383–407.

AMELI (**2011**): *Advanced Methodology for European Laeken Indicators.* http://ameli.surveystatistics.net, retrieved at 01.08.2015.

Andridge, R. R. and **Little, R. J. A.** (**2010**): *A review of Hot Deck imputation for survey non-response.* In: Int. Statist. Rev., 78, pp. 40–64.

Atkinson, T., **Cantillon, B.**, **Marlier, E.** and **Nolan, B.** (**2005**): Social indicators. Oxford (u.a.): Oxford Univ. Press.

Bandourian, R., **McDonald, J.** and **Turley, R. S.** (editors) (**2002**): A comparison of parametric models of income distribution across countries and over time, Luxembourg: Luxembourg income study working paper.

Bartlett, J. W., **Carpenter, J. R.**, **Tilling, K.** and **Vansteelandt, S.** (**2014**): *Improving upon the efficiency of complete case analysis when covariates are MNAR.* In: Biostatistics, 15 (4), pp. 719–730.

Bartlett, J. W., **Seaman, S. R.**, **White, I. R.**, **Carpenter, J. R.**, **Initiative, A. D. N.** et al. (**2015**): *Multiple imputation of covariates by fully conditional specification: accommodating the substantive model.* In: Statistical methods in medical research, 24 (4), pp. 462–487.

Berger, Y. G. (**2004**): *Variance estimation for measures of change in probability sampling.* In: Canadian Journal of Statistics, 32 (4), pp. 451–467.

Berger, Y. G. and **Escobar, E. L.** (**2015**): *Variance estimation of Hot-Deck imputed estimators of change for repeated rotating surveys.* Southampton Statistical Sciences Research Institute.

Berger, Y. G. and **Priam, R.** (**2016**): *A Simple Variance Estimator of Change for Rotating Repeated Surveys: an Application to the European Union Statistics on Income and Living Conditions Household Surveys.* In: Journal of the Royal Statistical Society, Series A, 179 (1), pp. 251–272.

Betti, G., **Cheli, B.**, **Lemmi, A.** and **Verma, V.** (**2006**): *Multidimensional and longitudinal poverty: an integrated fuzzy approach.* In: **Lemmi, A.** and **Betti, G.** (editors) Fuzzy Set Approach to Multidimensional Poverty Measurement, pp. 111–137, Springer.

Biggeri, M. (**2007**): *Children's valued capabilities.* In: **Walker, M.** and **Unterhalter, E.** (editors) Amartya Sen's Capability Approach and Social Justice in Education., Palgrave McMillan.

Biggeri, M., **Libanora, R.**, **Mariani, S.** and **Menchini, L.** (**2006**): *Children conceptualizing their capabilities: results of a survey conducted during the first children's world congress on child labour.* In: Journal of Human Development and Capabilities, 7 (1), pp. 59–83.

Bradshaw, J., **Chzhen, Y.**, **de Neubourg, C.**, **Main, G.**, **Martorano, B.** and **Menchini, L.** (**2012**): *Relative income poverty among children in rich countries.* Technical report 1, Innocenti Working Paper.

Breckling, J. and **Chambers, R.** (**1988**): *M-Quantiles.* In: Biometrika, 75 (4), pp. 761–771.

Brick, J. M. and **Montaquila, J. M.** (**2009**): *Nonresponse and weighting.* In: **Pfeffermann, D.** and **Rao, C. R.** (editors) Sample Surveys: Design, Methods and Applications, *Handbook of Statistics*, vol. 29A, pp. 163–185, Amsterdam: Elsevier.

Bruch, C., **Münnich, R.** and **Zins, S.** (**2011**): *Variance Estimation for Complex Surveys.* Technical report 3.1, FP7-SSH-2007-217322 AMELI.

Burgard, J. P., **Münnich, R.** and **Zimmermann, T.** (**2015**): *Impact of Sampling Designs in Small Area Estimation with Applications to Poverty Measurement.* In: **Pratesi, M.** (editor) Analysis of Poverty Data by Small Area Estimation, John Wiley & Sons.

Carpenter, J. R., **Goldstein, H.**, **Kenward, M. G.** et al. (**2011**): *REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types.* In: Journal of Statistical Software, 45 (5), pp. 1–14.

Carpenter, J. R., **Kenward, M. G.** and **White, I. R.** (**2007**): *Sensitivity analysis after multiple imputation under missing at random: a weighting approach.* In: Statistical methods in medical research, 16 (3), pp. 259–275.

Chambers, R. (**2001**): *Evaluation Criteria for Statistical Editing and Imputation, National Statistics Methodological Series No. 28.* In: University of Southampton.

Chambers, R. and **Tzavidis, N.** (**2006**): *M-quantile models for small area estimation.* In: Biometrika, 93 (2), pp. 255–268.

Cheli, B. and **Lemmi, A.** (**1995**): *A totally fuzzy and relative approach to the multidimensional analysis of poverty.* In: Economic Notes, 24 (1), pp. 115–134.

Cheney, W. and **Kincaid, D.** (**2012**): Numerical Mathematics and Computing. Boston: International Thomson Publishing, 7 ed., ISBN 0534351840.

Cheng, C. and **Van Ness, J.** (**1999**): Statistical Regression with Measurement Error. London: Arnold.

Chiappero-Martinetti, E. (**2006**): *Capability approach and fuzzy set theory: description, aggregation and inference.* In: **Lemmi, A.** and **Betti, G.** (editors) Fuzzy Set Approach to Multidimensional Poverty Measurement., Springer.

Cochran, W. G. (**1963**): Sampling techniques. New York (u.a.): Wiley, 2 ed.

Commission of the European Communities (**2003**): *Draft joint inclusion report.* Statistical annex. COM (2003) 773 final.

Cornish, R. P., **Tilling, K.**, **Boyd, A.**, **Davies, A.** and **Macleod, J.** (**2015**): *Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years.* In: International journal of epidemiology, p. dyv035.

Council of the European communities (**1985**): *Council Decision of 19 December 1984 on specific Community action to combat poverty (85/8/EEC).* In: Official Journal of the European Communities, pp. 24–25.

Craven, P. and **Wahba, G.** (**1978**): *Smoothing noisy data with spline functions.* In: Numerische Mathematik, 31 (4), pp. 377–403.

Dagum, C. (**1977**): *A New Model of Personal Income Distribution: Specification and Estimation.* In: Economie appliquée, 30 (3), pp. 413–437.

De Boor, C. (**1978**): A practical guide to splines. Appl. Math. Sci., New York, NY: Springer.

Dean, H. (**2009**): *Critiquing capabilities: the distractions of a beguiling concept.* In: Critical Social Policy, 29 (1), pp. 261–278.

Demirtas, H. and **Schafer, J. L.** (**2003**): *On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out.* In: Statistics in medicine, 22 (16), pp. 2553–2575.

Demnati, A. and **Rao, J.** (**2004**): *Linearization variance estimators for survey data.* In: Survey Methodology, 30 (1), pp. 17–26.

Deutscher Paritätischer Wohlfahrtsverband (**2009**): *Unter unseren Verhältnissen… Der erste Armutsatlas für Regionen in Deutschland.* Berlin.

Deutscher Paritätischer Wohlfahrtsverband (**2013**): *Zwischen Wohlstand und Verarmung: Deutschland vor der Zerreißprobe. Bericht zur regionalen Armutsentwicklung in Deutschland.* Berlin.

Deville, J. C. (**1999**): *Variance estimation for complex statistics and estimators: Linearization and residual techniques.* In: Survey methodology, 25 (2), pp. 193–204.

Deville, J. C. and **Särndal, C. E.** (**1992**): *Calibration Estimators in Survey Sampling.* In: Journal of the American Statistical Association, 87 (418), pp. 376–382.

Deville, J. C. and **Särndal, C. E.** (**1994**): *Variance estimation for the regression imputed Horvitz-Thompson estimator.* In: J. Off. Statist., 10, pp. 381–394.

Diggle, P. and **Kenward, M. G.** (**1994**): *Informative drop-out in longitudinal data analysis.* In: Applied statistics, pp. 49–93.

Duan, N. (**1983**): *Smearing estimate: A nonparametric retransformation method.* In: Journal of the American Statistical Association, 78 (383), pp. 605–610.

Durrant, G. B. (**2005**): *Imputation methods for handling item-nonresponse in the social sciences: a methodological review.* In: ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute. NCRM Methods Review Papers NCRM/002.

Efron, B. (**1979**): *Bootstrap Methods: Another Look at the Jackknife.* In: The Annals of Statistics, 7 (1), pp. 1–26, doi:10.1214/aos/1176344552.
URL `http://dx.doi.org/10.1214/aos/1176344552`

Eilers, P. H. and **Marx, B. D.** (**1996**): *Flexible smoothing with B-splines and penalties.* In: Statistical science, pp. 89–102.

Elbers, C., **Lanjouw, J. O.** and **Lanjouw, P.** (**2003**): *Micro–level estimation of poverty and inequality.* In: Econometrica, 71 (1), pp. 355–364.

Elbers, C. and **Van der Weide, R.** (**2014**): *Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality.* World Bank Policy Research Working Paper.

Enders, C. K. (**2010**): Applied missing data analysis. Guilford Press.

Engels, J. M. and **Diehr, P.** (**2003**): *Imputation of missing longitudinal data: a comparison of methods.* In: Journal of clinical epidemiology, 56 (10), pp. 968–976.

Escobar, E. L. and **Barrios, E.** (**2014**): samplingVarEst: Sampling Variance Estimation. R package version 0.9-9.
URL http://cran.r-s$_b$project.org/web/packages/samplingVarEst

European Commission (**2009**a): *2009 EU-SILC MODULE ON MATERIAL DEPRIVATION. Assessment of the implementation.*

European Commission (**2009**b): *Portfolio of indicators for the monitoring of the European strategy for social protection and social inclusion – 2009 update.* Technical report.

Eurostat (**2003**): *'Laeken' Indicators - detailed calculation methodology.* http://www.cso.ie/en/media/csoie/eusilc/documents/Laeken,Indicators,-,calculation,algorithm.pdf.

Eurostat (**2012**): *European statistics code of practice. For the national and community statistical authorities.* http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF, retrieved at 17.03.2014.

Eurostat (**2013**): *Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys - 2013 edition.* `http://ec.europa.eu/eurostat/documents/3859598/5927001/` `KS-sbRA-sb13-sb029-sbEN.PDF/a3155d11-sb4bf0-sb48d2-sb943d-sb2b1e9d096442`, retrieved at 12.01.2016.

Fabrizi, E., **Giusti, C.**, **Salvati, N.** and **Tzavidis, N.** (**2014**): *Mapping Average Equivalized Income Using Robust Small Area Methods.* In: Papers in Regional Science, 93, pp. 685–701.

Faris, P. D., **Ghali, W. A.**, **Brant, R.**, **Norris, C. M.**, **Galbraith, P. D.**, **Knudtson, M. L.**, **Investigators, A.** et al. (**2002**): *Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses.* In: Journal of clinical epidemiology, 55 (2), pp. 184–191.

Fay, B. E. (**1991**): *A design-based perspective on missing data variance.* In: Proceeding of the 1191 Annual Research Conference. U.S. Bureau of the Census, pp. 429–440.

Fay, B. E. (**1994**): *Analyzing imputed survey datasets with model-assisted estimators.* In: Proc. Survey Methods Sec., pp. 900–905, Am. Statist. Assoc.

Fay, R. and **Herriot, R.** (**1979**): *Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.* In: Journal of the American Statistical Association, 74, pp. 269–277.

Follmann, D. and **Wu, M.** (**1995**): *An approximate generalized linear model with random effects for informative missing data.* In: Biometrics, pp. 151–168.

Foster, J., **Greer, J.** and **Thorbecke, E.** (**1984**): *A class of decomposable poverty measures.* In: Econometrica, 52, pp. 761–766.

Gershunskaya, J. and **Lahiri, P.** (**2011**): *Robust Small Area Estimation Using a Mixture Model.* Presentation at the 58th orld Statistics Congress of the ISI, Dublin.

Giusti, C., **Marchetti, S.**, **Pratesi, M.** and **Salvati, N.** (**2012**): *Robust Small Area Estimation and Oversampling in the Estimation of Poverty Indicators.* In: Survey Research Methods, 6, pp. 155–163.

Goldstein, H., **Carpenter, J.**, **Kenward, M. G.** and **Levin, K. A.** (**2009**): *Multilevel models with multivariate mixed response types.* In: Statistical Modelling, 9 (3), pp. 173–197.

Goldstein, H., **Carpenter, J. R.** and **Browne, W. J.** (**2014**): *Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms.* In: Journal of the Royal Statistical Society: Series A (Statistics in Society), 177 (2), pp. 553–564.

González-Manteiga, W., **Lombardía, M.**, **Molina, I.**, **Morales, D.** and **Santamaría, L.** (**2008**): *Bootstrap mean squared error of a small-area EBLUP.* In: Journal of Statistical Computation and Simulation, 78 (5), pp. 443–462.

Graf, M., **Nedyalkova, D.**, **Münnich, R.**, **Seger, J.** and **Zins, S.** (**2011**a): *Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion.* Technical report 2.1, FP7-SSH-2007-217322 AMELI.

Graf, M., **Wenger, A.** and **Nedyalkova, D.** (**2011**b): *Quality of EU-SILC data.* Research Project Report WP5 – D5.1, FP7-SSH-2007-217322 AMELI.
URL `http://ameli.surveystatistics.net`

Hagenaars, A., **K. de Vos, K.** and **Zaidi, M.** (**1994**): *Poverty Statistics in the Late 1980s: Research Based on Micro-data.* Technical report, Eurostat.

Hansen, M. H. and **Hurwitz, W. N.** (**1943**): *On the Theory of Sampling from Finite Populations.* In: The Annals of Mathematical Statistics, 14 (4), pp. pp. 333–362.

Hawkes, D. and **Plewis, I.** (**2006**): *Modelling non-response in the national child development study.* In: Journal of the Royal Statistical Society: Series A (Statistics in Society), 169 (3), pp. 479–491.

Hawthorn, G. (**1987**): The Standard of Living. Cambridge University Press.

Haziza, D. (**2009**): *Imputation and inference in the presence of missing data.* In: **Pfeffermann, D.** and **Rao, C. R.** (editors) Sample Surveys: Design, Methods & Applications, *Handbook of Statistics*, vol. 29A, pp. 215–246, Amsterdam: Elsevier.

Heckman, J. (**1979**): *Sample Selection Bias as a Specification Error.* In: Econometrica, 47 (1), pp. 153–161.

Holman, R. and **Glas, C. A.** (**2005**): *Modelling non-ignorable missing-data mechanisms with item response theory models.* In: British Journal of Mathematical and Statistical Psychology, 58 (1), pp. 1–17.

Horvitz, D. G. and **Thompson, D. J.** (**1952**): *A generalization of sampling without replacement from a finite universe.* In: J. Am. Statist. Assoc., 47, pp. 663–685.

Horvitz, D. G. and **Thompson, D. J.** (**1952**): *A generalization of sampling without replacement from a finite universe.* In: Journal of the American Statistical Association, 47 (260), pp. 663–685.

Huang, E. T. and **Fuller, W. A.** (**1978**): *Nonnegative regression estimation for survey data.* In: Proc. Soc. Statist. Sec., pp. 300–305, American Statistical Association.

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**): *Report on the Simulation Results.* Research Project Report WP7 – D7.1, FP7-SSH-2007-217322 AMELI. URL `http://ameli.surveystatistics.net`

Hulliger, B. and **Münnich, R.** (**2006**): *Variance estimation for complex surveys in the presence of outliers.* In: Proceedings Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 3153–3156.

Hurlin, C., **Pérignon, C.** and **Stodden, V.** (**2014**): *RunMyCode.org: A research-reproducibility tool for computational sciences.* In: **Stodden, V.**, **Leisch, F.** and **Peng, R. D.** (editors) Implementing reproducible research, pp. 367–381, CRC Press.

Information und Technik Nordrhein-Westfalen (**2009**): *Berechnung von Armutsgefährdungsquoten auf Basis des Mikrozensus.*

http://www.amtliche-sozialberichterstattung.de/pdf/Berechnung von Armutsgefaehrdungsquoten_090518.pdf, retrieved at 26.08.2015.

Kenward, M. G. and **Carpenter, J.** (**2007**): *Multiple imputation: current perspectives.* In: Statistical methods in medical research, 16 (3), pp. 199–218.

Kenward, M. G., **Molenberghs, G.** and **Thijs, H.** (**2003**): *Pattern-mixture models with proper time dependence.* In: Biometrika, 90 (1), pp. 53–71.

Kim, J. K. and **Fuller, W.** (**2004**): *Fractional hot deck imputation.* In: Biometrika, 91 (3), pp. 559–578.

Klasen, S. (**2001**): *Social exclusion, children and education: implications of a rights-based approach.* In: European Societies, 3 (4), pp. 413–445.

Kolb, J.-P. (**2013**): Methoden zur Erzeugung synthetischer Simulationsgesamtheiten. Ph.D. thesis, Universität Trier, Universitätsring 15, 54296 Trier.

Kovacevic, M. S. and **Binder, D. A.** (**1997**): *Variance estimation for measures of income inequality and polarization-the estimating equations approach.* In: Journal of Official Statistics, 13 (1), p. 41.

Kriegel, H.-P., **Kröger, P.**, **Schubert, E.** and **Zimek, A.** (**2008**): *A general framework for increasing the robustness of PCA-based correlation clustering algorithms.* In: Scientific and Statistical Database Management, pp. 418–435, Springer.

Kristman, V. L., **Manno, M.** and **Côté, P.** (**2005**): *Methods to account for attrition in longitudinal data: do they work? A simulation study.* In: European journal of epidemiology, 20 (8), pp. 657–662.

Krueger, A. and **Lindahl, M.** (**2001**): *Education for growth: why and for whom?* In: Journal of Economic Literature, 34 (4), pp. 1101–1136.

Little, R. J. (**1993**): *Pattern-mixture models for multivariate incomplete data.* In: Journal of the American Statistical Association, 88 (421), pp. 125–134.

Little, R. J. and **Rubin, D. B.** (**2002**): Statistical analysis with missing data. John Wiley & Sons, 2 ed.

Lohr, S. L. (**2009**): Sampling: Design and Analysis. Boston: Brooks/Cole.

Maccagnan, A. (**2011**): *Measuring the interaction between parents and children in Italian families: a structural equation approach.* Working Paper 645, Department of Political Economics, University of Modena and Reggio Emilia.

Marchetti, S., **Giusti, C.**, **Pratesi, M.**, **Salvati, N.**, **Giannotti, F.**, **Pedreschi, D.**, **Rinzivillo, S.**, **Pappalardo, L.** and **Gabrielli, L.** (**2015**): *Small Area Model-Based Estimators Using Big Data Sources.* In: Journal of Official Statistics, 31 (2), pp. 263–281.

Marchetti, S., **Tzavidis, N.** and **Pratesi, M.** (**2012**a): *Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators.* In: Computational Statistics & Data Analysis, 56 (10), pp. 2889–2902.

Marchetti, S., **Tzavidis, N.** and **Pratesi, M.** (**2012**b): *Non-Parametric Bootstrap Mean Squared Error Estimation for M-Quantile Estimators of Small Area Averages, Quantiles and Poverty Indicators.* In: Computational Statistics and Data Analysis, 56, pp. 2889–2902.

McDonald, J. (**1984**): *Some generalized functions for the size distribution of income.* In: Econometrica: Journal of the Econometric Society, pp. 647–663.

Molenberghs, G., **Michiels, B.**, **Kenward, M. G.** and **Diggle, P. J.** (**1998**): *Monotone missing data and pattern-mixture models.* In: Statistica Neerlandica, 52 (2), pp. 153–161.

Molina, I., **Graf, M.** and **Marin, J.** (**2015**): *Poverty mapping at local level with suitable modelling of income.* Presentation at the 60th World Statistics Congress of the ISI, Rio de Janeiro.

Molina, I. and **Rao, J.** (**2010**): *Small area estimation of poverty indicators.* In: Canadian Journal of Statistics, 38 (3), pp. 369–385.

Morgan, J. (**1962**): *The anatomy of income distribution.* In: The Review of Economics and Statistics, 44 (3), pp. 270–283.

Münnich, R. (**2008**): *Varianzschätzung in komplexen Erhebungen.* In: Austrian Journal of Statistics, 37 (3 & 4), pp. 319–334.

Münnich, R. and **Zins, S.** (**2011**): *Variance Estimation for Indicators of Poverty and Social Exclusion.* Technical report 3.2, FP7-SSH-2007-217322 AMELI.

Narain, R. (**1951**): *On Sampling without Replacement with Varying Probabilities.* In: J. Ind. Soc. Agri. Statist., 3 (3), pp. 169–174.

Nathan, G. and **Holt, D.** (**1980**): *The effect of survey design on regression analysis.* In: Journal of the Royal Statistical Society. Series B (Methodological), pp. 377–386.

Neubourg de, C., **Bradshaw, J.**, **Chzhen, Y.**, **Main, G.**, **Martorano, B.** and **Menchini, L.** (**2012**): *Child deprivation, multidimensional poverty and monetary poverty in Europe.* Technical report 02, Innocenti Working Paper.

Nevalainen, J., **Kenward, M. G.** and **Virtanen, S. M.** (**2009**): *Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification.* In: Statistics in medicine, 28 (29), pp. 3657–3669.

Nunnally, J. and **Bernstein, I.** (**1994**): Psychometric Theory. McGraw-Hill.

Nussbaum, M. (**2006**): *Education and democratic citizenship: capabilities and quality education.* In: Journal of Human Development, 7 (3), pp. 385–395.

Osier, G. (**2009**): *Variance estimation for complex indicators of poverty and inequality using linearization techniques.* In: Survey Research Methods, vol. 3, pp. 167–195.

Pappalardo, L., **Rinzivillo, S.**, **Qu, Z.**, **Pedreschi, D.** and **Giannotti, F.** (**2013**): *Understanding the Patterns of Car Travel.* In: The European Physical Journal – Special Topics, 215, pp. 61–73.

Pearl, J. (**2003**): Causality: models, reasoning and inference, vol. 19. Cambridge Univ Press, 675–685 pp.

Peng, R. (**2015**): *The reproducibility crisis in science: A statistical counterattack.* In: Significance, 12 (3), pp. 30–32.

Potsi, A., **D'Agostino, A.**, **Giusti, C.** and **Porciani, L.** (**2016**): *Childhood and capability deprivation in Italy: a multidimensional and fuzzy set approach.* In: Quality & Quantity.

Preston, J. (**2009**): *Rescaled bootstrap for stratified multistage sampling.* In: Survey Methodology, 35 (2), pp. 227–234.

R Core Team (**2015**): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL http://www.R-s$_b$project.org/

Rao, J. and **Wu, C.** (**1988**): *Resampling inference with complex survey data.* In: Journal of the american statistical association, 83 (401), pp. 231–241.

Rao, J., **Wu, C.** and **Yue, K.** (**1992**): *Some recent work on resampling methods for complex surveys.* In: Survey methodology, 18 (2), pp. 209–217.

Rao, J. N. K. (**1965**): *On two simple schemes of unequal probability sampling without replacement.* In: Journal of the Indian Statistical Association, 3, pp. 173–180.

Rao, J. N. K. (**2003**): Small Area Estimation. London: Wiley.

Rao, J. N. K. and **Shao, A. J.** (**1992**): *Jackknife variance estimation with survey data under hotdeck imputation.* In: Biometrika, 79, pp. 811–822.

Reinsch, C. H. (**1967**): *Smoothing by spline functions.* In: Numerische Mathematik, 10 (3), pp. 177–183.

Rubin, D. B. (**1976**): *Inference and missing data.* In: Biometrika, 63 (3), pp. 581–592.

Rubin, D. B. (**1987**): Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D. B. (**2004**): Multiple imputation for nonresponse in surveys, vol. 81. John Wiley & Sons.

Ruppert, D., **Wand, M. P.** and **Carroll, R. J.** (**2003**): Semiparametric regression. Cambridge: Cambridge university press, 12 ed.

Sampford, M. R. (**1967**): *On Sampling Without Replacement with Unequal Probabilities of Selection.* In: Biometrika, 54 (3/4), pp. 499–513.

SAMPLE (**2009**): http://www.sample-project.eu, retrieved at 19.10.2014.

Särndal, C. E. and **Lundström** (**2005**): Estimation in Surveys with Nonresponse. Chichester: Wiley.

Särndal, C.-E., **Swensson, B.** and **Wretman, J. H.** (**1992**): Model assisted survey sampling. New York (u.a.): Springer.

Schafer, J. L. and **Graham, J. W.** (**2002**): *Missing data: our view of the state of the art.* In: Psychological methods, 7 (2), p. 147.

Seaman, S. R. and **White, I. R.** (**2013**): *Review of inverse probability weighting for dealing with missing data.* In: Statistical methods in medical research, 22 (3), pp. 278–295.

Seger, J. G. (**2015**): Measuring welfare and quality of life on regional level: Methodology for the estimation of composite indicators. Ph.D. thesis, Universität Trier.

Sen, A. K. (**1999**): *Breaking the poverty cycle - investing in early childhood.* Keynote Address, Inter-American Development Bank, Sustainable Development Department, Social Department Division.
URL `http://www.unicef.org/lac/spbarbados/Implementation/ECD/BreakingPovertyCycle_ECD_1999.pdf`

Shao, J. and **Steel, P.** (**1999**): *Variance Estimation for Survey Data with composite Imputation and Nonnegligible Sampling Fractions.* In: Journal of the American Statistical Association, 94, pp. 254–265.

Shao, J. and **Tu, D.** (**1995**a): The Jackknife and Bootstrap. Springer, New York.

Shao, J. and **Tu, D.** (**1995**b): The jackknife and bootstrap. New York: Springer.

Singh, S. K. and **Maddala, G. S.** (**1976**): *A function for size distribution of incomes.* In: Econometrica: Journal of the Econometric Society, 44, pp. 963–970.

Skinner, C., **Holmes, D.** and **Smith, T.** (**1986**): *The effect of sample design on principal component analysis.* In: Journal of the American Statistical Association, 81 (395), pp. 789–798.

Smith, P., **Pont, M.** and **Jones, T.** (**2003**): *Developments in business Surv. Methodol. in the Office for National Statistics, 1994–2000.* In: J. R. Statist. Soc. D (The Statistician), 52, pp. 257–295.

Social Protection Committee (**2001**): *Report on Indicators in the field of poverty and social exclusion.* Technical report, European Commission.

Stauder, J. and **Hüning, W.** (**2004**): *Die Messung von Äquivalenzeinkommen und Armutsquoten auf der Basis des Mikrozensus.* In: Statistische Analysen und Studien NRW, 13, pp. 9–31.

Steel, P. and **Fay, B. E.** (**1995**): *Variance estimation for finite populations with imputed data.* In: Proceedings of the Survey Research Methods Section, pp. 374–379, American Statistical Association.

Stodden, V. (**2015**): *Reproducing Statistical Results.* In: Annual Review of Statistics and Its Application, 2, pp. 1–19, doi:10.1146/annurev-s$_{\mathrm{b}}$statistics-s$_{\mathrm{b}}$010814-s$_{\mathrm{b}}$020127.

Stodden, V., **Miguez, S.** and **Seiler, J.** (**2015**): *Researchcompendia.org: Cyberinfrastructure for reproducibility and collaboration in computational science.* In: Computing in Science & Engineering, 17 (1), pp. 12–19.

Strengmann-Kuhn, W. (**1999**): *Armutsanalysen mit dem Mikrozensus.* In: ZUMA-Nachrichten Spezial. Bd 6, vol. 6, pp. 376–402, GESIS.

Tarki Social Research Institute (**2010**): *Child poverty and child well-being in the European Union. Report for the European Commission.*

Thoemmes, F. and **Mohan, K.** (**2015**): *Graphical representation of missing data problems.* In: Structural Equation Modeling: A Multidisciplinary Journal, 22 (4), pp. 631–642.

Tillé, Y. and **Matei, A.** (**2013**): sampling: Survey Sampling. R package version 2.6. URL `http://cran.r-`s$_{\mathrm{b}}$`project.org/web/packages/sampling`

Tsiatis, A. A. and **Davidian, M.** (**2004**): *Joint modeling of longitudinal and time-to-event data: an overview.* In: Statistica Sinica, pp. 809–834.

Tzavidis, N., **Marchetti, S.** and **Chambers, R.** (**2010**a): *Robust estimation of small area means and quantiles.* In: Australian and New Zeland Journal of Statistics, 52 (2), pp. 167–186.

Tzavidis, N., **Marchetti, S.** and **Chambers, R.** (**2010**b): *Robust Prediction of Small Area Means and Distributions.* In: Australian and New Zealand Journal of Statistics, 52, pp. 167–186.

Valliant, R., **Dorfman, A. H.** and **Royall, R. M.** (**2000**): Finite Population Sampling and Inference: A Prediction Approach. New York: Wiley.

Van Buuren, S., **Boshuizen, H. C.**, **Knook, D. L.** et al. (**1999**): *Multiple imputation of missing blood pressure covariates in survival analysis.* In: Statistics in medicine, 18 (6), pp. 681–694.

Vonesh, E. F., **Greene, T.** and **Schluchter, M. D.** (**2006**): *Shared parameter models for the joint analysis of longitudinal data and event times.* In: Statistics in medicine, 25 (1), pp. 143–163.

Wahba, G. (**1990**): Spline models for observational data, vol. 59. Siam.

Weidenhammer, B., **Tzavidis, N.**, **Schmid, T.** and **Salvati, N.** (**2015**): *Microsimulation via quantiles for domain prediction.* Presentation at the 60th World Statistics Congress of the ISI, Rio de Janeiro.

Welch, C., **Bartlett, J.** and **Petersen, I.** (**2014**): *Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data.* In: The Stata journal, 14 (2), p. 418.

Whelan, C. and **Maitre, B.** (**2008**): *The Life Cycle Perspective on Social Inclusion in Ireland: An Analysis of EU-SILC.* Research Series 3, The economic and social research institute Dublin.

Wolter, K. (**2007**): Introduction to Variance Estimation. New York: Springer, 2nd ed.

Wolter, K. M. (**2007**): Introduction to Variance Estimation. New York: Springer, 2nd ed.

Wu, L., **Liu, W.**, **Yi, G. Y.** and **Huang, Y.** (**2011**): *Analysis of longitudinal and survival data: joint modeling, inference methods, and issues.* In: Journal of Probability and Statistics, 2012.

Ybarra, L. (**2003**): *Small Area Estimation Using Data from Multiple Surveys*, unpublished PhD thesis, Arizona State University.

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

InGRID
Inclusive Growth Research Infrastructure Diffusion

Referring to the EU2020-ambition of Inclusive Growth, the general objectives of InGRID – Inclusive Growth Research Infrastructure Diffusion – are to integrate and to innovate existing, but distributed European social sciences research infrastructures on 'Poverty and Living Conditions' and 'Working Conditions and Vulnerability' by providing transnational data access, organising mutual knowledge exchange activities and improving methods and tools for comparative research. This integration will provide the related European scientific community with new and better opportunities to fulfil its key role in the development of evidence-based European policies for Inclusive Growth. In this regard specific attention is paid to a better measurement of related state policies, to high-performance statistical quality management, and to dissemination/outreach activities with the broader stakeholder community-of-interest, including European politics, civil society and statistical system.

InGRID is supported by the European Union's Seventh Programme for Research, Technological Development and Demonstration under Grant Agreement No 312691.

More detailed information is available on the website: www.inclusivegrowth.be

Co-ordinator

Guy Van Gyes
Monique Ramioul

KU LEUVEN HIVA
RESEARCH INSTITUTE FOR WORK AND SOCIETY

Partners

TÁRKI Social Research Institute Inc. (HU)
Amsterdam Institute for Advanced labour Studies, Universiteit van Amsterdam (NL)
The Swedish Institute for Social Research, Stockholms Universitet (SE)
Fachbereich IV, Wirtschafts- und Sozialstatistik, Universität Trier (DE)
Centre d'Etudis Demogràfics, Campus de la Universitat Autònoma de Barcelona (ES)
Centre d'Etudes de Population, de Pauvreté et de Politiques Socio-Economiques (LU)
Centre for Social Policy, Universiteit Antwerpen (BE)
Institute for Social & Economic Research, University of Essex (UK)
Bremen International Graduate School of Social Sciences, Universität Bremen (DE)
Department of Dynamics of Organisations of Work, Centre d'Etudes de l'Emploi (FR)
The Centre for European Policy Studies (BE)
Dipartimento di Economica e Menagement, Università di Pisa (IT)
Social Statistics Division, University of Southampton (UK)
Luxembourg Income Study, asbl (LU)
WageIndicator Foundation (NL)
School of Social Sciences, The University of Manchester (UK)

InGRID

Inclusive Growth Research
Infrastructure Diffusion
Contract No 312691

For further information about
the InGRID project, please contact
inclusive.growth@kuleuven.be
www.inclusivegrowth.be
p/a HIVA – Research Institute
for Work and Society
Parkstraat 47 box 5300
3000 Leuven
Belgium

The views expressed during the execution of the InGRID project, in whatever form and or by whatever medium, are the sole responsibility of the authors. The European Union is not liable for any use that may be made of the information contained therein.