ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 3, *pp.* 1273–1291

A functional approach to small area estimation of the relative median poverty gap

Enrico Fabrizi

Università Cattolica del Sacro Cuore, Piacenza, Italy

and Maria Rosaria Ferrante and Carlo Trivisano

Università di Bologna, Italy

[Received February 2018. Final revision February 2020]

Summary. We consider the estimation of the relative median poverty gap (RMPG) at the level of Italian provinces by using data from the European Union Survey on Income and Living Conditions. The overall sample size does not allow reliable estimation of income-distribution-related parameters at the provincial level; therefore, small area estimation techniques must be used. The specific challenge in estimating the RMPG is that, as it summarizes the income distribution of the poor, samples for estimating it for small subpopulations are even smaller than those available in other parameters. We propose a Bayesian strategy where various parameters summarizing the distribution of income at the provincial level are modelled by means of a multivariate small area model. To estimate the RMPG, we relate these parameters to a distribution describing income, namely the generalized beta distribution of the second kind. Posterior draws from the multivariate model are then used to generate draws for the distribution's area-specific parameters and then of the RMPG defined as their functional.

Keywords: Complex sample surveys; Generalized beta distribution of the second kind; Hierarchical Bayes; Income inequality; Poverty

1. Introduction

The relative median at risk of poverty gap is one of the indicators that have been endorsed by the European Union for the assessment of social cohesion (European Commission, 2004). It is defined as the median distance of the individual poor equivalized income from a threshold defined as the 60% of the national median, relative to this threshold. The relative median at risk of poverty gap is an important complement to the information that is provided by the head count ratio measure of poverty (at risk of poverty rate) as it offers an insight on how deep is the poverty that is experienced by the median poor, regardless of how many live below the poverty line.

At risk of poverty rates (relative median poverty gaps (RMPGs)), as well as many other poverty and income inequality measures are annually calculated by Eurostat for most European Union (EU) member states by using data from the European Union Survey on Income and Living Conditions (EU SILC), conducted under harmonized guidelines (see Atkinson and Marlier (2010) for a general introduction). Estimates of these parameters are published also for large regions or social groups within countries. This paper is about estimating RMPGs in small areas,

Address for correspondence: Enrico Fabrizi, Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy.
E-mail: enrico.fabrizi@unicatt.it

0964–1998/20/1831273

i.e. for a collection of population subsets ('areas') for which the subset-specific sample sizes are not sufficiently large to obtain decent precision from ordinary survey-weighted estimators (that are labelled as *direct estimators* in the small area literature).

We note that the problem of sample sizes that are *not sufficiently large* is more severe for the RMPG than for other summaries of the income distribution as it is a (scaled) quantile of the poor income distribution whose direct estimation is based only on those who are poor: usually a minority of the sample units. For instance, if the prevalence of the poor ranges from 5% to 33% the expected area-specific sample sizes that are available to estimate the sample mean will be from three to 20 times larger than those available for the estimation of the RMPG.

Specifically, we consider the problem of estimating the RMPG for Italian administrative provinces by using data from the Italian section of the EU SILC. In Italy there are 110 provinces corresponding to the Eurostat nomenclature of territorial units for statistics level 3 (Eurostat, 2019). Provincial administrations play an important role in implementing policies that are decided at higher levels (national or regional) and in co-ordinating the activities of lower administrative levels (municipalities and health districts). We consider data from the 2013 wave of the EU SILC and auxiliary information known at the provincial level obtained from various sources, including fiscal archives of the Italian Ministry of Finance and population registers.

Small area estimation is about complementing the insufficient information that is provided by area-specific samples with auxiliary information known from external sources (censuses, administrative archives, ...). The complementing is typically achieved by using models that can be specified at either the area or the unit level (Pfeffermann, 2013).

In this paper we consider area level models (Rao and Molina (2015), chapter 5). These models are less demanding in terms of required information as only direct estimates, associated measures of uncertainty and summaries at the area level of the auxiliary variables are needed. They can represent the only viable strategy for the secondary data analysis that does not have access to the details of the sampling design and relevant unit level information. Moreover, some typical problems that are met when using unit level models, such as possible inconsistencies in definitions and measurement techniques for auxiliary variables between the sample survey and the auxiliary source, are sidestepped. See Tarozzi and Deaton (2009) and Tzavidis *et al.* (2018) for more general discussions of these topics. In our application, we have limited access to some information on the sampling design and dispose only of area level summary statistics for the auxiliary information that we consider in the models.

As it relies on area level models, this research is different from previous literature on small area estimation of the RMPG (Molina and Rao, 2010; Molina *et al.*, 2014) that focuses on unit level modelling.

The inputs of an effective area level model are

(a) a set of area level approximately unbiased estimates endowed with reliable sampling variability measures and
(b) a vector of area level auxiliary information with good predictive power for the parameter in question.

If we denote η_d the RMPG in area d, $\hat{\eta}_d$ its direct estimate and $\mathbf{x}_d$ a vector of area level auxiliary information, a typical area level model is not a viable strategy as direct estimators of the RMPG are biased (as the median is) and very imprecise in small samples (see results in the on-line appendix 1); moreover auxiliary variables with good predictive power are difficult to find for η_d.

Our alternative strategy can be summarized as follows. We consider $\boldsymbol{\theta}_d$: a vector of additional small area parameters for which approximately unbiased direct estimators and predictive auxiliary information are available. As they are not of direct interest, we label $\boldsymbol{\theta}_d$ as *nuisance* small area parameters. We specify a small area model for $\boldsymbol{\theta}_d$. The components of $\boldsymbol{\theta}_d$ can be related functionally to each other via $\boldsymbol{\xi}_d$, a vector of parameters characterizing a distribution that we assume for income in area d, so that $\boldsymbol{\theta}_d = \boldsymbol{\theta}(\boldsymbol{\xi}_d)$. The solution in $\boldsymbol{\xi}_d$ of this system of equations can then be used to estimate $\eta_d = \eta(\boldsymbol{\xi}_d)$ functionally under the distribution that is assumed to describe income.

A few technical comments are in order.

(a) We consider five *nuisance* small area parameters θ_{kd} so that $\boldsymbol{\theta}_d = \{\theta_{kd}\}$, $k = 1, \ldots, 5$; they include three head count ratios based on different thresholds, a concentration index and the mean of the log-income; their choice is aimed at providing a description of the whole income distribution at the area level. More details will be given in Section 2.2.

(b) We specify a multivariate small area model for $\boldsymbol{\theta}_d$. Multivariate models have a long tradition in small area estimation dating back at least to Ghosh *et al.* (1996) and they usually lead to more efficient estimators as they exploit the correlation between parameters.

(c) The parametric distribution that we consider for income is the generalized beta distribution of the second kind (known as the 'GB2' distribution) (McDonald, 1984) that is widely used in the literature. We also consider three distribution that are special cases of GB2 (Dagum, Singh–Maddala (SM) and the beta distribution of the second kind) that depend on three parameters. The recourse to these special cases is motivated by computational sustainability; more details on this point will be given in Sections 4.2 and 5.

(d) The number of *nuisance* parameters is larger than the size of $\boldsymbol{\xi}$ characterizing GB2: this entails a solution of the system $\boldsymbol{\theta}_d = \boldsymbol{\theta}(\boldsymbol{\xi}_d)$ based on the minimization of a loss function that allows more flexible and numerically stable solutions.

The core of this methodology, i.e. the estimation of $\boldsymbol{\xi}$ by solving $\boldsymbol{\theta}_d = \boldsymbol{\theta}(\boldsymbol{\xi}_d)$, was introduced in Graf and Nedyalkova (2014). Here we apply it to a small area estimation problem in the framework of a hierarchical Bayesian model. Specifically, we approximate posterior distributions of $\boldsymbol{\theta}_d$ by means of Markov chain Monte Carlo (MCMC) algorithms. By solving $\boldsymbol{\theta}_d = \boldsymbol{\theta}(\boldsymbol{\xi}_d)$ for each MCMC draw we obtain Markov chains for the parameters characterizing the assumed income distribution at the area level. The $\eta_d = \eta(\boldsymbol{\xi}_d)$ can be exploited to generate a Markov chain converging to the posterior of the target parameter η_d.

Predictors of nuisance parameters are design consistent (see Section 3), i.e. their point predictors converge to area-specific population descriptive quantities regardless of misspecifications of the multivariate model. Asymptotically the estimator of η_d converges to the functional of these population quantities that depends on the assumption of GB2-distributed income in the area. As a consequence, the dependence on the assumption of these distributions remains, but the estimator is robust with respect to misspecifications of the multivariate small area model.

The rest of the paper is organized as follows. Section 2 introduces the data set that we consider in this application and direct estimation of the *small area parameters* that are involved in the study. In Section 3 we introduce the multivariate small area estimation model that provides the basis for the estimation of the RMPG. Section 4 includes a short review of GB2 and its special cases and an illustration of our functional estimation methodology. The estimation of the RMPG at the level of Italian provinces is illustrated in Section 5, with some discussion. As the method is rather complex, we explore the frequentist properties of the proposed estimators

by means of a simulation exercise, based on the same sample data (Section 6). Concluding remarks are provided in Section 7.

2. The data and direct estimation of small area parameters

2.1. The data

We analyse data from the 2013 wave of the EU SILC. The survey is conducted in many countries across the EU by the relevant national institutes of statistics by using harmonized questionnaires and survey methodologies. Although following common guidelines, sampling designs can differ from country to country. In Italy, the EU SILC is a rotating panel survey with 75% overlap of samples in successive years. The fresh part of the sample is drawn according to a stratified two-stage sample design, where municipalities (local authority unit level 2; see Eurostat (2019)) are the primary sampling units, whereas households are the secondary sampling units. The primary sampling units are divided into strata according to their population size and the secondary sampling units are selected by systematic sampling in each primary sampling unit.

We target administrative provinces. The 110 Italian provinces have largely different populations ranging from the 4.3 million inhabitants of Rome, down to less than 0.1 million (Medio Campidano, Isernia and Ogliastra). Provinces are unplanned domains for the EU SILC. For the 2013 wave that we consider in this paper, province-specific sample sizes range from 6 up to 882 in terms of households and from 10 to 2018 in terms of individuals. The median province-specific sample size is 115 households (274 individuals).

2.2. Direct estimation

Consider a population P of size N and a partition of it into D small areas $\{P_1, \ldots, P_d, \ldots, P_D\}$ of size N_d, $\Sigma_{d=1}^{D} N_d = N$. A sample of overall size n is drawn from the population according to a complex design such as the stratified multistage design with a rotating panel component used in the EU SILC.

Area-specific samples sizes are denoted n_d so that $\Sigma_{d=1}^{D} n_d = n$. A survey weight w_{dj} is associated with each unit in the sample ($j = 1, \ldots, n_d$; $d = 1, \ldots, D$), reflecting both inclusion probabilities and non-response corrections. We target a variable y, the equivalized disposable income, defined as the total disposable household income divided by the equivalized household size calculated according to the modified Organisation for Economic Co-operation and Development scale (see Fusco *et al.* (2010)).

Although our ultimate focus is the estimation of the RMPG, we consider several population descriptive quantities at the area level that we label *small area parameters*. To avoid confusion, we denote the RMPG at the area level with η_d and the vector of *nuisance* small area parameters as $\boldsymbol{\theta}_d = \{\theta_{kd}\}$ with $k = 1, \ldots, 5$. Whenever $n_d > 0$ these parameters can be estimated by using area-specific samples using Hàjek type (Hàjek, 1958) or other design-based estimators that we can assume are approximately unbiased. We label these estimators as direct and denote them $\hat{\eta}_d$ and $\hat{\theta}_{kd}$.

The RMPG is defined as $\eta = \{\mathrm{pt}_1 - \mathrm{Me_p}(y)\}/\mathrm{pt}_1$, where $\mathrm{Me_p}(y)$ is the median income of the poor, i.e. $\mathrm{Me_p}(y) = \mathrm{Me}(y | y \leqslant \mathrm{pt}_1)$ and pt_1 is the national poverty threshold, defined in the EU SILC framework as 60% of the national median of equivalized income. A survey weighted estimator of η_d is given by

$$\hat{\eta}_d = \frac{\mathrm{pt}_1 - \widehat{\mathrm{mp}}_d}{\mathrm{pt}_1} \tag{1}$$

where

$$\widehat{\mathrm{mp}}_d = \begin{cases} \frac{1}{2}(y_{(jd)} + y_{(j+1,d)}) & \text{if } \sum_{i=1}^{j} w_{(i)} = 0.5 \sum_{i=1}^{n_{d\mathrm{p}}} w_{(i)}, \\ y_{(j+1,d)} & \text{if } \sum_{i=1}^{j} w_{(i)} < 0.5 \sum_{i=1}^{n_{d\mathrm{p}}} w_{(i)} < \sum_{i=1}^{j+1} w_{(i)}, \end{cases}$$

$n_{d\mathrm{p}} \leqslant n_d$ is the number of poor in the sample specific to domain d and $y_{(i)} \leqslant y_{(i+1)}$, $i = 1, \ldots, n_{d\mathrm{p}}$, is the non-decreasing sequence of poor incomes. $\hat{\eta}_d$ is likely to be more imprecise than $\hat{\theta}_{kd}$ as it is based on the income of only those below pt_1 in the sample: typically a minority. Moreover, in very small samples it can be substantially biased. A small design-based simulation exercise, based on EU SILC data and reported in the on-line appendix 1, explores the size of bias and variance of this estimator in small samples.

The *nuisance* parameters that we consider in this application are

(a) the at risk of poverty rate $\theta_1 = E\{\mathbf{1}(y \leqslant \mathrm{pt}_1)\}$, a poverty count based on the threshold pt_1 and that represents the most popular poverty measure in the EU,
(b) the proportion of people living with an equivalized income below the national median, $\theta_2 = E[\mathbf{1}\{y \leqslant \mathrm{Me}(y)\}]$,
(c) an affluence rate defined as the proportion of individuals for which $y > \mathrm{pt}_3$ where pt_3 is some high threshold, that we fix at twice the national sample median in line with Peichl *et al.* (2010), $\theta_3 = E\{\mathbf{1}(y > \mathrm{pt}_3)\}$ (affluence rates are useful to describe the right-hand tail of the y-distribution at the area level),
(d) the Gini concentration index, which can be defined as $\theta_4 = \Delta\{2E(y)\}^{-1}$ where $\Delta = E\{|y_s - y_t|\}$ with y_s and y_t identically distributed as y, and
(e) the mean of the log-income, i.e. $\theta_5 = E\{\log(y)\}$.

We now present direct estimators for the nuisance parameters θ_{kd}. For $k = 1, 2$ they can be written as

$$\hat{\theta}_{kd} = \frac{\sum_{j=1}^{n_d} w_{dj} \mathbf{1}(y_{dj} < \mathrm{pt}_k)}{\sum_{j=1}^{n_d} w_{dj}}. \tag{2}$$

When $k = 1$, we have the at risk of poverty rate whereas for $k = 2$ we define $\mathrm{pt}_2 = \mathrm{Me}(y)$, i.e. $\mathrm{pt}_1 = 0.6\mathrm{pt}_2$. We note that, when estimated at the whole population level, $\hat{\theta}_{2.} = 0.5$, in specific domains it can be read as a departure of the local median from that of the entire population. The direct estimator of θ_{3d} is defined as

$$\hat{\theta}_{3d} = \frac{\sum_{j=1}^{n_d} w_{dj} \mathbf{1}(y_{dj} > \mathrm{pt}_3)}{\sum_{j=1}^{n_d} w_{dj}}. \tag{3}$$

We note that pt_1, pt_2 and pt_3 rely on the estimated national median of the equivalized income. As this estimate is based on a very large national sample, we shall overlook the uncertainty that is associated with these thresholds and treat them as fixed constants.

The most popular direct estimators of θ_4, for instance the estimator that was considered in Alfons and Templ (2013), are biased in small samples. In line with Fabrizi and Trivisano (2016)

we consider a nearly unbiased direct estimator that accounts also for the fact that individuals in the same household share the same income:

$$\hat{\theta}_{4d} = \frac{1}{2\hat{\bar{Y}}_d} \frac{\sum_{j=1}^{n_d} \sum_{k=1}^{n_d} w_{dj} w_{dk} |y_{dj} - y_{dk}|}{\hat{N}_d^2 - \sum_{h=1}^{m_d} \tilde{w}_{dh}^2}, \tag{4}$$

where $\hat{\bar{Y}}_d = \hat{N}_d^{-1} \Sigma_{j=1}^{n_d} w_{dj} y_{dj}$ and $\hat{N}_d = \Sigma_{j=1}^{n_d} w_{dj}$ is the Horwitz–Thompson estimator of the domain size; moreover, m_d is the number of households sampled in domain d and $\tilde{w}_{dh} = \Sigma_{j=1}^{n_h} w_{dj}$ is the sum of weights associated with the n_h individuals living in household h ($h = 1, \ldots, m_d$).

An approximately unbiased estimator of θ_5 can be defined as

$$\hat{\theta}_{5d} = \frac{\sum_{j=1}^{n_d} w_{dj} \log(y_{dj})}{\sum_{j=1}^{n_d} w_{dj}}. \tag{5}$$

The direct estimators $\hat{\theta}_{kd}$ are nearly unbiased but their variance can be large when n_d is small. In the case of the EU SILC, their variances will be larger than those which we would have obtained with simple random samples of the same number of individuals. In the first place, the same equivalized income is shared by all individuals in the same household (perfect intracluster correlation). Moreover, the design effect of the EU SILC survey for Italy is larger than 1 even considering variables at the household level; although the design is stratified at the first stage, clustering of households within municipalities, unequal selection probabilities and weighting corrections to counter non-response cause losses of efficiency (see Clemenceau and Museux (2007) and Goedemé (2013) for more details).

To estimate the variances of $\hat{\theta}_{kd}$ we consider a two-step approach: first a bootstrap algorithm, described in Fabrizi *et al.* (2011), is used to obtain preliminary variance estimates. These *raw* variances are then used to estimate design effects and other parameters of variance smoothing models that will be described in Section 5. We note that the bootstrap algorithm does not incorporate all details of the EU SILC sample design for Italy, because of limited access to municipality level clustering and longitudinal tracking information; on the basis of previous literature (see Goedemé (2013) and Biewen and Jenskins (2006)) we assume that once essential features of the designs have been accounted for (stratification, clustering at the household level, unequal selection probabilities and weighting), good approximations to actual sampling variances can be obtained. As pointed out in Tzavidis *et al.* (2018), variance smoothing is a delicate step in building an area level model, so special attention will be devoted to the assessment and quality of fit of these smoothing models in Section 5.

3. A multivariate small area model for parameters related to equivalized income distribution

In this section we describe a multivariate model for θ_{kd}, $k = 1, \ldots, 5$. In line with the typical specification of small area models, ours has two levels:

(a) a sampling model that provides a likelihood for the direct estimators and relates them to the underlying population parameters;

(b) a linking model that relates the small area parameters to auxiliary information and to each other by means of exchangeable random effects according to the principle of *borrowing strength*.

The recourse to a multivariate model is motivated by the fact that the five parameters represent different aspects of the area level distribution of the target variable y. The estimates $\hat{\theta}_{kd}$ represent summaries of the same area-specific samples, so it is natural to assume that they are correlated, and to specify a multivariate sampling model. We do this by means of a Gaussian copula function in line with Fabrizi *et al.* (2016). See Souza and Moura (2016) for other applications of copula functions in the small area context. We present the sampling model in two steps: first, we introduce the marginal sampling models; then the copula function is used to account for their dependence structure.

For the rates θ_{kd}, $k = 1, 2, 3$, in line with Fabrizi *et al.* (2016), we specify a zero-inflated beta sampling model to account for the fact that rates range in the (0, 1) interval and that, when m_d is small, the direct estimate can be 0, i.e. $\hat{\theta}_{kd} = 0$ even if it is assumed, as we do, that $\theta_{kd} > 0$:

$$f(\hat{\theta}_{kd}|\theta^*_{kd}, \hat{\phi}_{kd}) = (1-\theta^*_{kd})^{m_d}\mathbf{1}(\hat{\theta}_{kd}=0) + \{1-(1-\theta^*_{kd})^{m_d}\}\text{dbeta}(A_{kd}, B_{kd})\mathbf{1}(\hat{\theta}_{kd}>0) \qquad (6)$$

where $A_{kd} = \theta^*_{kd}(\hat{\phi}_{kd}-1)$ and $B_{kd} = (1-\theta^*_{kd})(\hat{\phi}_{kd}-1)$. See Ospina and Ferrari (2012) and Wieczorek and Hawala (2011) for alternative specifications of zero-inflated beta regression allowing also for $\theta_{kd} = 0$.

The quantities $\hat{\phi}_{kd}$ can be interpreted as an effective sample size in terms of individuals and are estimated by using variance smoothing models. See Section 5 for more details on these models and estimation leading to $\hat{\phi}_{kd}$. The parameter θ^*_{kd} is defined as $\theta^*_{kd} = E(\hat{\theta}_{kd}|\hat{\theta}_{kd}>0, \theta_{kd}, \hat{\phi}_{kd})$ so the parameter that we are actually interested in is given by

$$\theta_{kd} = \theta^*_{kd}\{1-(1-\theta^*_{kd})^{m_d}\} = E(\hat{\theta}_{kd}|\theta^*_{kd}, \hat{\phi}_{kd}).$$

Note that in equation (6) we assume that $P(\hat{\theta}_{kd}=0)$ depends explicitly on the underlying rate θ^*_{kd} and the number m_d of households sampled from domain d.

The sampling model for the Gini concentration coefficient is based on a beta likelihood, with a parameterization that we take from Fabrizi and Trivisano (2016):

$$\hat{\theta}_{4d} \sim \text{beta}\left(\frac{2\hat{\phi}_{4d}}{1+\theta_{4d}} - \theta_{4d}, \frac{2\hat{\phi}_{4d} - \theta_{4d}(1+\theta_{4d})}{1+\theta_{4d}}\frac{1-\theta_{4d}}{\theta_{4d}}\right). \qquad (7)$$

As a consequence $E(\hat{\theta}_{4d}|\hat{\phi}_{4d}) = \theta_{4d}$ and $V(\hat{\theta}_{4d}|\hat{\phi}_{4d}) = \theta^2_{4d}(1-\theta^2_{4d})(2\hat{\phi}^{-1}_{4d})$. See Section 5 for details on the variance model that was used to obtain the quantities $\hat{\phi}_{4d}$, which will be treated as known.

The sampling model for the mean of the log-incomes $\hat{\theta}_{5d}$ is a normal Fay–Herriot model:

$$\hat{\theta}_{5d} \sim N(\theta_{5d}, \hat{\phi}^{-1}_{5d}). \qquad (8)$$

Variances $\hat{\phi}^{-1}_{5d}$ are estimated by using the bootstrap algorithm that was discussed in Fabrizi *et al.* (2016). The assumption of known variances for normal small area models is in line with most literature (see Rao and Molina (2015), chapter 5). It is also consistent with expressions (6) and (7) as we consider a two-parameter distribution where one of the two parameters is assumed known.

The Gaussian copula (Clemen and Reilly, 1999) that was used to model the direct estimators' dependence structure is parameterized in terms of the correlation matrix $\mathbf{R}$ of a Gaussian multivariate distribution. In detail, we assume that

$$f(\hat{\theta}_{1d},\ldots,\hat{\theta}_{kd}) = \frac{g_1(\hat{\theta}_{1d})\times\ldots\times g_k(\hat{\theta}_{kd})}{|\mathbf{R}|^{1/2}} = \exp\left\{-\frac{1}{2}\mathbf{z}_k^{\mathrm{T}}(\mathbf{R}^{-1}-\mathbf{I}_k)\mathbf{z}_k\right\} \tag{9}$$

with $\mathbf{z}_k^{\mathrm{T}} = (\Phi^{-1}\{F_1(\hat{\theta}_{1d})\},\ldots,\Phi^{-1}\{F_5(\hat{\theta}_{kd})\})$; the marginal densities $f_k(\hat{\theta}_{kd})$, $k=1,\ldots,5$, are defined in expressions (6)–(8) and $F_k(\hat{\theta}_{kd})$ are the associated cumulative distribution functions. The matrix $\mathbf{R}$ is to be estimated from the data. For the specific application that we consider in this paper, the estimation procedure will be outlined in Section 5.

The linking models for the three rates and the Gini coefficients are based on a logit link,

$$\mathrm{logit}(\theta_{kd}) = \mathbf{x}_{kd}^{\mathrm{T}}\beta_k + v_{kd} \tag{10}$$

($k=1,\ldots,4$), whereas an identity link is considered for θ_{5d}:

$$\theta_{5d} = \mathbf{x}_{5d}^{\mathrm{T}}\beta_5 + v_{5d}. \tag{11}$$

The vector $\mathbf{x}_{kd}$ contains for each parameter and each area auxiliary information known at the area level. Note that x_{kd} and β_k may vary with k; but the first element of x_{kd} is 1 in all cases.

The multivariate relationship between the population parameters θ_{kd} is incorporated in the distributional assumption for $\mathbf{v}_d = (v_{kd})$, $k=1,\ldots,5$:

$$\mathbf{v}_d \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_v) \tag{12}$$

where MVN denotes the multivariate normal distribution. For Σ_v we specify a prior within the family that was proposed by Huang and Wand (2013) with the purpose of keeping the analytical and computational tractability of the inverse Wishart distribution but improving the non-informativity properties:

$$\begin{aligned}\boldsymbol{\Sigma}_v|a_1,\ldots,a_k &\sim \text{Inv-Wishart}\{\nu+1, 2\nu\,\mathrm{diag}(a_1^{-1},\ldots,a_k^{-1})\},\\ a_k &\sim \text{Inv-Gamma}\left(\frac{1}{2},\frac{1}{A_k}\right), \qquad k=1,\ldots,5.\end{aligned} \tag{13}$$

This prior marginally induces $\sigma_k \sim \text{half-}t(\nu, A_k)$. The choice $\nu=2$ allows for a diffuse prior, close to the popular half-Cauchy($\nu=1$) prior; moreover it induces a marginal uniform prior on the correlations between the random effects. We choose $A_k=1$ after careful consideration of the scale of the parameters' distribution and some sensitivity analysis.

For all parameters the point predictor of the small area mean is obtained summarizing the posterior distribution of θ_{kd} by using quadratic loss, so that $\tilde{\theta}_{kd} = E(\theta_{kd}|\mathbf{d})$, $k=1,\ldots,5$, and where short cut notation $\mathbf{d}$ is used for the data.

It can be shown that, conditionally on $\boldsymbol{\Sigma}_v$, $\tilde{\theta}_{kd}$, $k=1,\ldots,5$, is design consistent provided that $\hat{\theta}_{kd}$ are. For the definition of design consistency we refer to Fuller (2009), page 41. For a proof of this design consistency property see the on-line appendix 2.

4. The proposed estimation strategy for the relative median poverty gap

4.1. The generalized beta distribution of the second kind and its special cases

GB2 (McDonald, 1984) is a four-parameter distribution which is acknowledged as an excellent descriptor of income distributions (Dastrup *et al.*, 2007; Jenkins, 2009; Graf and Nedyalkova, 2011). The GB2 density can be written as

$$f(x;a,b,p,q) = \frac{a}{bB(p,q)}\frac{(x/b)^{ap-1}}{\{1+(x/b)^a\}^{p+q}}\mathbf{1}(x>0) \tag{14}$$

where $a, b, p, q > 0$ and $B(p, q)$ is the beta function. With the exception of b, which is a scale parameter, the other three parameters are all shape parameters: a can be interpreted as an overall shape parameter and p rules the right-hand tail whereas q the left-hand tail. For a general description of the properties of GB2 see Kleiber and Kotz (2003), chapter 6.1, and Graf *et al.* (2011).

In the economy of this study we are interested in the expression of the *small area parameters* η_d and $\boldsymbol{\theta}_d$ that were introduced in Section 2.2 when the equivalized income variable is assumed to be GB2 distributed. We use the notation $\theta_{kd|\mathrm{GB2}}$ and $\eta_{d|\mathrm{GB2}}$ to denote the expression of θ_{kd} under the GB2 assumption:

$$\theta_{1d|\mathrm{GB2}} = F(\mathrm{pt}_1, a_d, b_d, p_d, q_d), \tag{15}$$

$$\theta_{2d|\mathrm{GB2}} = F(\mathrm{pt}_2, a_d, b_d, p_d, q_d), \tag{16}$$

$$\theta_{3d|\mathrm{GB2}} = 1 - F(\mathrm{pt}_3, a_d, b_d, p_d, q_d), \tag{17}$$

$$\theta_{4d|\mathrm{GB2}} = \frac{B(2p_d + 1/a_d, 2q_d - 1/a_d)}{B(p_d + 1/a_d, 2q_d - 1/a_d)} \times \{p_d^{-1} G_1(a_d, p_d, q_d) + (p_d + 1/a_d)^{-1} G_2(a_d, p_d, q_d)\}, \tag{18}$$

$$\theta_{5d|\mathrm{GB2}} = \frac{\psi(p_d) - \psi(q_d)}{a_d} + \log(b_d), \tag{19}$$

$$\eta_{d|\mathrm{GB2}} = 1 - \frac{F^{-1}(\theta_{1d|\mathrm{GB2}}/2, a_d, b_d, p_d, q_d)}{F^{-1}(\theta_{1d|\mathrm{GB2}}, a_d, b_d, p_d, q_d)}. \tag{20}$$

Note that F in equations (15)–(17) is the cumulative distribution function whereas in expression (19) $G_1(\cdot)$ and $G_2(\cdot)$ are generalized hypergeometric series (see McDonald (1984) for a detailed definition) depending on all the distribution parameters except the scale b_d whereas $\psi(\cdot)$ in equation (20) is the digamma function.

GB2 encompasses several special cases. In this research we consider the beta of the second kind (known as 'B2') distribution ($a = 1$), the Dagum distribution ($q = 1$) and the SM distribution ($p = 1$). For these special cases expressions (15)–(20) are simpler and notably so for the Gini coefficient (19) that reduces to

$$\theta_{4d|\mathrm{B2}} = \frac{B(2p_d, 2q_d - 1)}{2pB^2(p_d, q_d)}, \tag{21}$$

$$\theta_{4d|\mathrm{Dagum}} = \frac{\Gamma(p_d)\Gamma(2p_d + 1/a_d)}{\Gamma(2p_d)\Gamma(p_d + 1/a_d)}, \tag{22}$$

$$\theta_{4d|\mathrm{SM}} = 1 - \frac{\Gamma(q_d)\Gamma(2q_d - 1/a_d)}{\Gamma(2q_d)\Gamma(q_d - 1/a_d)} \tag{23}$$

where $\Gamma(\cdot)$ is the gamma function. The considered special cases of GB2 are also those identified by McDonald *et al.* (2013) as those characterized by skewness–kurtosis spaces encompassing the largest portion of the income data set in their cross-country analysis of the Luxembourg income study database. Kakamu (2016), using a simulation study based on data generated from GB2, characterized parameter regions in which the fit of the Dagum distribution is superior

to that of the SM distribution and vice versa. Intuitively, data with a heavy right-hand tail should be better fitted by an SM distribution and those with a more moderate skewness by the Dagum distribution. Kleiber (1996) expected the Dagum distribution to fit better than the SM distribution in most real data sets; actually its skweness–kurtosis space includes that of the SM distribution in the direction of more moderate and even negative skewness. B2 is considered especially for its popularity in the literature (Chotikapanich *et al.*, 2012).

4.2. Indirect estimation of the relative median poverty gap

Let $\boldsymbol{\xi}_d = (a_d, b_d, p_d, q_d)$ denote the parameters of GB2 that we assume to describe the income distribution in area d. As areas are many, this description would imply a very large set of parameters to be estimated; this cannot be done by using area-specific samples, as they are typically small. We use the multivariate model to accomplish this task. Under this GB2 assumption,

$$\boldsymbol{\theta}_d = \boldsymbol{\theta}(\boldsymbol{\xi}_d)$$

according to formulae (15)–(20). Using the multivariate model of Section 3 we can draw from $p(\boldsymbol{\theta}_d|\mathbf{d})$. For each draw $\boldsymbol{\theta}_{rd}$, $r = 1, \ldots, R$, we can solve $\boldsymbol{\theta}_{rd} = \boldsymbol{\theta}(\boldsymbol{\xi}_{rd})$ in $\boldsymbol{\xi}_{rd}$, thus obtaining a draw from $p(\boldsymbol{\xi}_d|\mathbf{d})$. We can then use

$$\eta_d = \eta(\boldsymbol{\xi}_d)$$

defined according to equation (21) to simulate from $p\{\eta_d = \eta(\boldsymbol{\xi}_d)|\mathbf{d}\}$, by drawing $\eta_{rd} = \eta(\boldsymbol{\xi}_{rd})$.

Several technical details about the implementation of this approach now follow. We note that $p(\theta_{kd}|\mathbf{d})$ depends on the way that we modelled the direct estimators $\hat{\theta}_{kd}$ but not on the GB2 that we assume for the income distribution in the areas. If the size of $\boldsymbol{\theta}_d$ and $\boldsymbol{\xi}_d$ were the same, a solution to the system $\boldsymbol{\theta}_d = \boldsymbol{\theta}(\xi_d)$ can be slow or even impossible to find with numeric methods. In line with Graf and Nedyalkova (2014), section 5, we use a vector $\boldsymbol{\theta}_d$ of five elements to solve for the four parameters characterizing GB2 by minimizing a relative quadratic loss function:

$$L(\boldsymbol{\theta}_{rd}, \boldsymbol{\xi}_{rd}) = \sum_{k=1}^{5} \left\{ \frac{\theta_{krd} - \theta_{krd|\mathrm{GB2}}(\boldsymbol{\xi}_{rd})}{\theta_{krd}} \right\}^2. \tag{24}$$

With respect to Graf and Nedyalkova (2014) we select a different set of *nuisance* parameters, namely the θ_{kd}, $k = 1, \ldots, 5$, that were discussed in Section 3. Except for θ_{5d} all parameters have approximately the same scale (as they range between 0 and 1), whereas θ_{5d} is much bigger in scale. For this reason when solving the system we consider the scaled values $\theta^*_{r5d} = \theta_{r5d} - \log(K)$ where K is a suitably chosen constant that makes scales of all parameters more homogeneous. The solution of the system with the original set of parameters $\boldsymbol{\xi}_{rd} = (a_{rd}, b_{rd}, p_{rd}, q_{rd})$ can be obtained from $\boldsymbol{\xi}^*_{rd} = (a_{rd}, b^*_{rd}, p_{rd}, q_{rd})$ by using a property of GB2 as $b_{rd} = Kb^*_{rd}$. In line with Graf *et al.* (2011) and Graf and Nedyalkova (2014) we set the constraints $a_{rd}\, p_{rd} > 1$ and $a_{rd} q_{rd} > 2$ which ensure that the implicitly defined $X_{rd} \sim \mathrm{GB2}(a_{rd}, b_{rd}, p_{rd}, q_{rd})$ are such that $E(X^{-1}_{rd}) < \infty$ and $E(X^2_{rd}) < \infty$.

The minimum is searched for by using numerical methods and namely the popular Levenberg–Marquardt algorithm. Theoretical properties and efficient implementations of this algorithm have been studied in many papers (e.g. Moré (1978)). Kanzow *et al.* (2004) showed global convergence properties of the algorithm when the constraints set is a convex set as in our problem.

Because of the mathematical complexity of expression (19) the solution leading to the indirect estimation of the GB2 parameters can be slow to find, making the whole method impractical.

For this reason we consider three special cases of GB2: B2, Dagum and SM distributions, characterized by three parameters and much simpler formulae for the Gini coefficient (see equations (21), (22) and (23). We keep the same set of five small area parameters and a loss function analogous to equation (25), i.e. $L^{(i)}(\boldsymbol{\theta}_{rd}, \boldsymbol{\xi}_{rd})$, $i = 1, 2, 3$, for the indirect estimation of the three distribution parameters.

For each draw θ_{rkd}, $r = 1, \ldots, R$, we estimate three parallel non-linear systems: one for each of the three special cases of GB2, thus generating separate chains for the three sets of distribution parameters. Although the three systems are solved instead of one, this strategy is computationally much more efficient than that based on GB2. If we denote by $\hat{\boldsymbol{\xi}}_{rd}$ a solution to equation (25) the distribution that minimizes $\Sigma_{r=1}^{R} L^{(i)}(\boldsymbol{\theta}_{rd}, \hat{\boldsymbol{\xi}}_{rd})$ in i is chosen, separately for each area, as the income distribution model. As a consequence, we adapt possibly different models to the data from different areas.

A point predictor for η_d can be obtained summarizing the posterior distribution $p(\eta_d|\mathbf{d})$; if quadratic loss is adopted it will be given by the posterior mean $\tilde{\eta}_d = E(\eta_d|\mathbf{d})$.

The small area estimator that is obtained in this way is not design consistent as it depends on assuming GB2 as a description of income within the areas even in large samples. Nonetheless it is robust with respect to misspecifications of the small area model as $\tilde{\boldsymbol{\theta}}_d$ is design consistent and thus converging to $\boldsymbol{\theta}_d$ regardless of model misspecifications. Asymptotically the posterior distribution $p(\eta_d|\mathbf{d})$ will collapse on the solution of $\eta_d = \eta(\boldsymbol{\xi}_d)$: the dependence on GB2 does remain, but that on the multivariate model does not.

5. An application to Italian European Union Survey on Income and Living Conditions data: estimation of relative median poverty gap in Italian provinces

In this section we illustrate the estimation of the RMPG η_d and the *nuisance* parameters θ_{kd} for the Italian administrative provinces. Input data come from the 2013 EU SILC survey sample for Italy and consist of $(\hat{\theta}_{kd}, \hat{\phi}_{kd}, \mathbf{R})$, $k = 1, \ldots, 5$, $d = 1, \ldots, D$. We obtain an estimate of $\mathbf{R}$ starting from Spearman correlations $\rho_r(\cdot, \cdot)$ among the $\hat{\theta}_{kd}$. Rough estimates of $\rho_r(\hat{\theta}_{kd}, \hat{\theta}_{k'd})$ can be obtained by using the bootstrap algorithm output (see Section 2.2). We denote these estimates as $\mathrm{cor}_{\mathrm{boot}}(\hat{\theta}_{kd}, \hat{\theta}_{k'd})$. As most of the areas are small, to obtain stable estimates, we first assume that correlations $\rho_r(\hat{\theta}_{kd}, \hat{\theta}_{k'd})$ are constant across areas, i.e. $\rho_r(\hat{\theta}_{kd}, \hat{\theta}_{k'd}) = \rho_r(\hat{\theta}_k, \hat{\theta}_{k'})$, and propose averaged estimates

$$\hat{\rho}_r(\hat{\theta}_k, \hat{\theta}_{k'}) = \left(\sum_{d=1}^{D} w_d \right)^{-1} \sum_{d=1}^{D} w_d \, \mathrm{cor}_{\mathrm{boot}}(\hat{\theta}_{kd}, \hat{\theta}_{k'd})$$

with $w_d = n_d$. To obtain even more stable results, we then restrict the average to the set of the largest areas and namely to those with a sample size above the median, thus assuming that $w_d = n_d \mathbf{1}\{n_d > \mathrm{Me}(n_d)\}$. As the matrix $\mathbf{R}$ describes the dependence structure of $\hat{\theta}_{kd}$ on a transformed scale, we finally exploit the invariance of Spearman correlation under non-decreasing monotone transformations and the sine transformation to switch from Spearman to Pearson correlations (see Elfadaly and Garthwaite (2017) for details).

The parameters $\hat{\phi}_{kd}$ are estimated by using variance smoothing models. Specifically, for the rates $\hat{\theta}_{kd}$, $k = 1, 2, 3$, the variances that are estimated by using the bootstrap algorithm $v_{\mathrm{boot}}(\hat{\theta}_{kd})$ are smoothed by using the models

$$\frac{\hat{\theta}_{kd}(1 - \hat{\theta}_{kd})}{v_{\mathrm{boot}}(\hat{\theta}_{kd})} = \nu_k n_d + e_{kd}$$

where, for the residuals e_{kd}, we assume that $E(e_{kd}) = 0$ and $V(e_{kd}) = \varrho_k$. For the Gini concentration coefficient, a different smoothing model is adopted:

$$\frac{\hat{\theta}_{4d}^2(1-\hat{\theta}_{4d}^2)}{v_{\text{boot}}(\hat{\theta}_{4d})} = \nu_4 n_d + e_{4d}.$$

See Fabrizi and Trivisano (2016) for a motivation of this model. The least squares estimators $\hat{\nu}_k$ are then used to compute $\hat{\phi}_{kd} = \nu_k n_d$, $k = 1, \ldots, 4$. For our data the squared correlations describing the fit of these models equal 0.82, 0.95, 0.78 and 0.78 for $k = 1, \ldots, 4$ respectively.

These data are complemented by auxiliary information from administrative archives. A description of auxiliary variables, defined at the provincial level, can be found in the on-line appendix 3. The candidate auxiliary variables are many; some are highly correlated with each other, so selection is needed. Although the model is multivariate, we selected covariates to be used in equations (10) and (11) from the univariate models. Auxiliary variable selection is based on the methodology that was introduced in George and McCulloch (1993). Details on the variable-selection process can be found in appendix 3 as well.

All code used in the estimation exercise was written in R. Posterior distributions for the multivariate model are based on a Metropolis–Hastings type of MCMC algorithms. Specifically we used the software `jags` called through the R package `rjags` (Plummer *et al.*, 2016). For all the parameters single Markov chains of length 50000 were run. To assess the convergence of each chain, beside visual inspection of the chains, we use Heidelberg–Welch diagnostics (Heidelberg and Welch, 1983; Cowles and Carlin, 1996) that reduce to testing the null hypothesis of a stationary path by using the Cramer–von Mises statistic. A conservative burn-in of 10000 is used before calculating these statistics. The Heidelberg–Welch diagnostics are based on a single chain; a multichain approach was not advisable in our problem as a careful setting of the initial value is needed to speed up the convergence. In the overwhelming majority of chains the p-value that is associated with the Heidelberg–Welch diagnostics is above 0.05; for the chains of the parameters θ_{1d}, θ_{2d}, θ_{4d} and θ_{5d} in more than 98% of the cases and for θ_{3d} slightly more than 95% of the cases. In calculating posterior summaries, one every 30th draw was kept. This severe *thinning* of the chains is partly motivated by their relatively poor mixing; this depends on the fact that *nuisance* parameters are strongly correlated, as they are all summaries of the same distributions. Moreover, we want to keep the posterior sample size small as its size defines the number of times that the non-linear system discussed in Section 4.2 needs to be solved. The overall sample from the posterior is of size $R = 3000$.

Each draw from the posterior distribution of θ_{kd}, $k = 1, \ldots, 5$, is used to solve the constrained non-linear system that was discussed in Section 4.2. Specifically we work with the Levenberg–Marquardt non-linear least-squares algorithm as implemented in the `nlsLM` function of the R package `minpack.lm` (Elzhov *et al.*, 2016). Initial values were set by solving the system on the ensemble of the posterior means $E(\theta_{kd}|\mathbf{d})$ with a precision 1.0×10^{-10}, whereas a precision of 1.0×10^{-5} was used to assess convergence of solutions for the systems based on individual draws.

The application ran in about 2 h by using a four-cores 5500u processor (2.44 GHz; 8 Gbytes random-access memory). We tried to run the same application by using GB2 instead of its special cases as the reference distribution: the computing times rose to about 40 h. This motivates our choice of considering a solution based on the three-parameters special cases of GB2.

A special case of GB2 is chosen separately for each area according to the methodology that was illustrated in Section 4.2. The Dagum distribution was chosen in the large majority of areas (95 times), the SM distribution for 14 areas and B2 in only one area. This result is in line with

expectations from the literature (Kleiber, 1996; McDonald *et al.*, 2013) as discussed in Section 4.1. For the analysis of this data set the methodology could then be simplified and only the Dagum distribution considered. Nonetheless this may depend on specific features of our data and it is not necessarily a general result (see Kakamu (2016)).

Markov chains for η_d (RMPG) were generated from those of the parameters of the chosen distributions. The Heidelberg–Welch diagnostics that were computed for the chains η_d result in a p-value that was greater than 0.05 in 96% of the cases. As this percentage is in line with the type I error of the test, we can conclude that the convergence is satisfying also for these chains.

As a further check we applied the functional approach that was used to generate posterior chains for η_d to the *nuisance* parameters θ_{kd} and compared the posterior that was obtained in this way with those directly obtained from the multivariate model that was described in Section 3. We focus our comparisons on posterior means and standard deviations, calculating ratios of the posterior summaries that were obtained according to the two methods. These ratios show some variation across areas. For posterior means we have that for all parameters and all areas the difference is less than 5% with the exception of θ_4 (the Gini concentration coefficient) for which the difference is between 5% and 10% in 20% of the areas; posterior means obtained with the functional were slightly smaller (3% on average). For all parameters, posterior standard deviations are very close on average (less than 2%) with the exception of θ_4 and θ_5 for which the posterior standard deviations based on the functional approach are 5% larger on average. In the large majority of areas the difference is less than 10% and for θ_1, θ_2 and θ_3 less than 5%.

In Table 1 we present how efficient our approach is in reducing the standard errors that are associated with the estimators. We define

$$\text{ser}(\eta_d) = \frac{\text{sd}(\eta_d|\mathbf{d})}{\text{se}(\hat{\eta}_d)} \tag{25}$$

where $\text{se}(\hat{\eta}_d)$ is computed according to the bootstrap algorithm of Fabrizi *et al.* (2011). We calculate also $\text{ser}(\theta_{kd})$ that are defined similarly; $\text{se}(\hat{\theta}_{kd})$ is calculated according to the methodology that was illustrated in Section 2.2. We recognize that this comparison involves two quantities that are logically different as the numerator is a posterior standard deviation and the denominator a standard error with respect to the randomization distribution that is induced by sampling. Nonetheless this type of comparisons is common in the small area literature.

The improvement in precision that is enabled by $\tilde{\eta}_d$ with respect to $\hat{\eta}_d$ is dramatic; on average the posterior standard deviation is slightly more than a quarter of that of the direct estimator.

Table 1. Distribution of the standard error reduction (ser_{kd}) defined in equation (25) across the 110 provinces (areas)†

Parameter	η	$\theta_{1.}$	$\theta_{2.}$	$\theta_{3.}$	$\theta_{4.}$	$\theta_{5.}$
Minimum	0.064	0.102	0.113	0.122	0.078	0.169
1st quartile	0.168	0.380	0.413	0.303	0.303	0.549
Median	0.265	0.482	0.493	0.398	0.362	0.627
Mean	0.284	0.483	0.511	0.414	0.383	0.627
3rd quartile	0.358	0.586	0.601	0.506	0.467	0.745
Maximum	0.711	0.904	0.93	0.885	0.831	0.926

†η is the RMPG, θ_1 the at risk of poverty rate, θ_2 the share of population with income below the median, θ_3 the affluence rate, θ_4 the Gini concentration coefficient and θ_5 the mean of log-income.

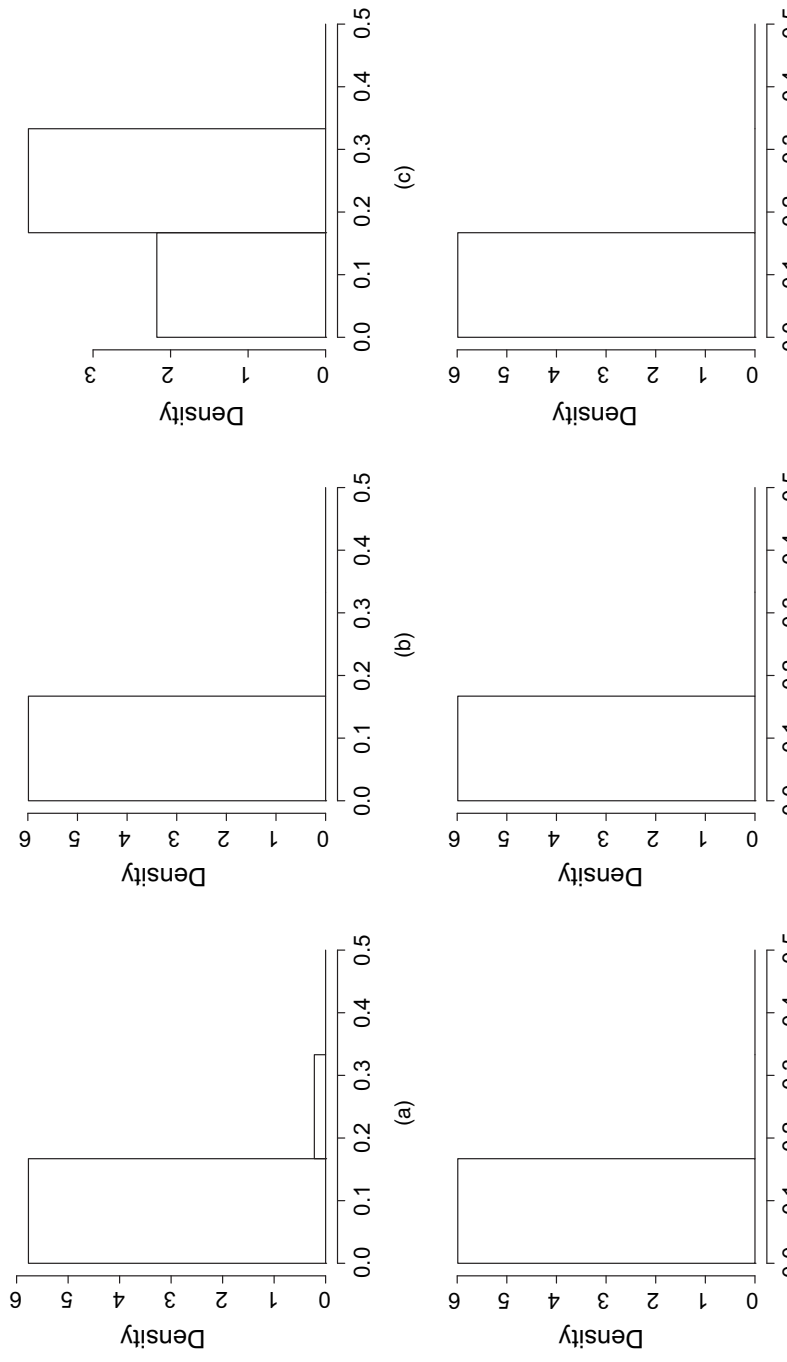

Fig. 1. Histograms of the posterior CVs over the 110 provinces (the breaks in the histograms plot coincide with those suggested by Statistics Canada (2007)): (a) $CV(\theta_1|\mathbf{d})$; (b) $CV(\theta_2|\mathbf{d})$; (c) $CV(\theta_3|\mathbf{d})$; (d) $CV(\theta_4|\mathbf{d})$; (e) $CV(\theta_5|\mathbf{d})$; (f) $CV(\eta|\mathbf{d})$

Only in large areas, and especially so if in the south of the country where the prevalence of poverty is higher, $\mathrm{sd}(\eta_d|\mathbf{d})$ is more than a half of $\mathrm{se}(\hat{\eta}_d)$. The posterior standard deviations $\mathrm{sd}(\theta_{kd}|\mathbf{d})$ are on average half the size of the standard error $\mathrm{se}(\hat{\theta}_{kd})$ of direct estimators; different levels of reduction in different areas can be explained by different area-specific sample sizes.

Statistics Canada (2007) suggested that estimates whose associated coefficient of variation (CV) is less than 16.6% are sufficiently reliable for general use and those with a CV between 16.6% and 33.3% can be published but accompanied by a warning to users whereas those with an even larger CV should be deemed completely unreliable and not published. In Fig. 1 we plot the histograms of $\mathrm{CV}(\eta_{kd}|\mathbf{d})$ and $\mathrm{CV}(\theta_{kd}|\mathbf{d})$, using the thresholds that were suggested by Statistics Canada (2007). We note that, although popular, these criteria can be too exigent for the estimation of small proportions when a high CV can be the effect of a small estimate; in this case, which encompasses our θ_1 and θ_3, alternative criteria in terms of standard errors can be used (see European Commission (2013), page 13). We keep the Statistics Canada criteria as, from Fig. 1, it is apparent that for all parameters the small area estimates that we produce are suitable for publication with few problematic cases for the affluence rate θ_3, attributable to the low point estimates. Notably the posterior CVs are acceptable in all cases for the RMPG.

6. A simulation exercise

The methodology that we have presented for the estimation of the RMPG is complex as it involves a multivariate hierarchical Bayesian model and, for each MCMC draw, the solution of a non-linear system based on a parametric assumption on the distribution of equivalized income in the areas. The good performances in terms of posterior CV that appear in Fig. 1 can be misleading if the point estimates were heavily biased. In this section, we introduce a simulation study to assess the frequentist properties of the RMPG predictor. Specifically we focus on the bias, mean-square error and frequentist coverage of probability intervals based on posterior quantiles. These properties will be evaluated also for the predictors of *nuisance* parameters θ_{kd}.

The simulation exercise is based on the same EU SILC sample as considered in our application. We assume it as a synthetic population, from which we repeatedly draw stratified samples and estimate the small area parameters for areas that are larger than those considered in the application. As the synthetic population is held fixed, the simulation can be labelled as design based.

We target administrative regions as areas of interest: a higher level administrative body with respect to the provinces that were considered in the application; each region includes several provinces; the two exceptions, Valle d'Aosta and Molise, that include only one and two provinces respectively, have been excluded from the synthetic population. Administrative regions are planned domains of the EU SILC survey in Italy. We draw stratified samples from the synthetic population with strata defined by these regions. The size of the 18 administrative regions in the synthetic population ranges, in terms of households, from 386 to 1846 with a median size of 998. Stratified samples, drawn without replacement, are allocated proportionally with a sampling rate of 0.115, chosen so that the median size of region-specific samples in the simulation matches the median of the province-specific samples in the application. With respect to the application, sample sizes are less variable as they range from 44 to 212 (and not from 6 to 882 as in the case of province-specific samples in the application).

For each of the $S = 1000$ samples that were drawn from the synthetic population we replicate the methodology that was illustrated in Section 5; also the details related to MCMC computation and the non-linear system remain the same.

Denote by ${}_P\boldsymbol{\theta}_d$ and ${}_P\eta_d$ the synthetic population target parameters, where ${}_P\boldsymbol{\theta}_d = \{{}_P\theta_{kd}\}$, $k = 1, \ldots, 5$, whereas the Bayes estimators based on quadratic loss are denoted as ${}_s\tilde{\boldsymbol{\theta}}_d = E({}_P\boldsymbol{\theta}_d|\mathbf{d}_s)$ and ${}_s\tilde{\eta}_d = E({}_P\eta_d|\mathbf{d}_s)$ where $\mathbf{d}_s$ denotes the data from the sth replicated sample. If we use the short cut $.\tilde{\theta}_{kd}$ to denote the Bayes estimator for θ_{kd} when averaged over the S replications we can define

$$\mathrm{RRMSE}(\tilde{\theta}_{kd}) = \frac{1}{S}\sum_{s=1}^{S}\frac{\sqrt{\{({}_s\tilde{\theta}_{kd} - {}_P\theta_{kd})^2\}}}{{}_P\theta_{kd}}, \tag{26}$$

$$\mathrm{RBIAS}(\tilde{\theta}_{kd}) = \frac{1}{S}\sum_{s=1}^{S}\frac{{}_s\tilde{\theta}_{kd} - {}_P\theta_{kd}}{{}_P\theta_{kd}}, \tag{27}$$

$$\mathrm{COV}(\tilde{\theta}_{kd}; 1-\alpha) = \frac{1}{S}\sum_{s=1}^{S}\mathbf{1}({}_sq_{\alpha/2} \leqslant {}_s\,\theta_{kd} \leqslant {}_s\,q_{1-\alpha/2}) \tag{28}$$

where ${}_sq_{\alpha/2}$ and ${}_sq_{1-\alpha/2}$ are the α- and $(1-\alpha)$-quantiles of $p({}_P\theta_{kd}|\mathbf{d}_s)$. Specifically we consider $\alpha = 0.05$. Definitions for $\mathrm{RRMSE}(\tilde{\eta}_d)$, $\mathrm{RBIAS}(\tilde{\eta}_d)$ and $\mathrm{COV}(\tilde{\eta}_d, 1-\alpha)$ follow accordingly.

In Table 2 we present results for the indicators (26)–(28): we show the three quartiles (Q_1, Me, Q_3) of the distribution of these three indicators across the 18 regions that were considered in the simulation.

The relative root-mean-squared error RRMSE that is associated with the RMPG has the same magnitude as those of the at risk of poverty rate $\tilde{\theta}_{1d}$ and affluence rate $\tilde{\theta}_{3d}$, which is a good result if we interpret it considering the little information that the direct estimation of the RMPG provides. Smaller RRMSEs can be either attributed to a size effect ($\tilde{\theta}_{2d}$ has a mean-squared error that is similar to that of $\tilde{\theta}_{1d}$ but a larger denominator) or to the more power that auxiliary variables have for some parameters (specifically this is so for the mean of the log-incomes $\tilde{\theta}_{5d}$). The relative bias is, in all cases, when averaged across areas, close to 0, i.e. the shrinkage does not

Table 2. First and third quartiles and median of RRMSE, RBIAS and COV(·, 0.95) with respect to the 18 regions considered in the simulation†

	Quartile	θ_1	θ_2	θ_3	θ_4	θ_5	η
Direct estimators							
RBIAS	Q_1	−0.005	−0.002	−0.006	−0.005	0.000	0.016
	Me	−0.002	0.000	−0.001	−0.003	0.000	0.035
	Q_3	0.003	0.002	0.007	−0.002	0.000	0.123
RRMSE	Q_1	0.205	0.090	0.250	0.072	0.005	0.320
	Me	0.257	0.116	0.329	0.081	0.006	0.426
	Q_3	0.283	0.124	0.486	0.092	0.009	0.466
Bayesian estimators							
RBIAS	Q_1	−0.057	−0.023	−0.046	−0.029	−0.002	−0.054
	Me	0.019	0.003	0.073	−0.004	0.000	0.012
	Q_3	0.101	0.027	0.108	0.028	0.002	0.101
RRMSE	Q_1	0.093	0.043	0.139	0.034	0.002	0.108
	Me	0.115	0.055	0.156	0.041	0.003	0.141
	Q_3	0.160	0.074	0.241	0.066	0.006	0.205
COV(·, 0.95)	Q_1	0.904	0.880	0.933	0.871	0.904	0.911
	Me	0.977	0.983	0.975	0.985	0.955	0.937
	Q_3	0.987	0.985	0.986	0.995	0.979	0.953

†θ_1 is the at risk of poverty rate, θ_2 the share of population with income below the median, θ_3 the affluence rate, θ_4 the Gini concentration coefficient, θ_5 the mean of log-income and η the RMPG.

imply a systematic tendency to overestimate or to underestimate the corresponding population parameters. As far as the RMPG is concerned, the relative biases are, despite their indirect estimation, small in most of the areas. Negative or positive biases on individual areas are due to a shrinkage effect that is more pronounced when the sample size is small.

Interval estimates based on posterior quantiles $(q_{\alpha/2}, q_{1-\alpha/2})$ usually have an approximate $1-\alpha$ frequentist coverage if the bias of the posterior mean is small and the posterior standard deviation is close to the frequentist standard error. Table 2 shows that in some cases the coverage is below the frequentist nominal level; these cases are those characterized by relatively higher bias levels. In some other cases we have a coverage above the nominal (frequentist) level; this is due to a tendency of posterior standard deviations to be slightly larger than the frequentist standard errors (we can estimate from Monte Carlo replications).

To complete the comparison, for η_d, we simulated also an estimator that is associated with a *standard* Fay–Herriot type of model assuming approximate normality of $\hat{\eta}_d$, $\text{var}(\hat{\eta}_d)$, as known and set equal to their actual values resulting from Monte Carlo replications. We selected auxiliary variables from those described in the on-line appendix 3 and namely the variables x_1, the antilogit of x_6 and x_9 that proved to be those providing the best fit. The average RRMSE result was equal to 0.249 and the average RBIAS to 0.059. The average COV(0.95) is very close (slightly above) the nominal level; nonetheless some of the intervals are so wide that the lower bound is negative. This estimator is therefore effective in improving the efficiency of the direct estimator but clearly inferior to $\tilde{\eta}_d$. This finding is in line with our expectation: not only are the $\hat{\eta}_d$ very unreliable but it is difficult to obtain auxiliary variables with a good predictive power.

7. Conclusions

In this research we focused on the estimation of the RMPG, which is a popular measure of the severity of poverty, motivated by the need to estimate it at the small area level by using Italian data from the EU SILC.

We present a small area estimation method based on area level modelling, which requires only survey-based direct estimators and area level summaries from auxiliary sources. Area level modelling is therefore less data demanding with respect to unit level models that, when applied to the estimation of the non-linear functional of the target variable population values, require knowledge of individual level values of the auxiliary variables: a requirement that implies non-trivial data quality and disclosure problems.

The specific nature of the RMPG, for which direct estimators are in most cases completely unreliable, led us to consider a functional estimation method. We built on a method of using summary statistics to estimate parameters of an underlying income distribution due to Graf and Nedyalkova (2014), applied it within the framework of MCMC-sampling-based Bayesian inference and used it in the opposite direction to estimate the RMPG (i.e. using estimated income distribution parameters to obtain an estimate of a population descriptive quantity).

Our methodology implies various choices, some of them driven by computational reasons. Specifically we propose to use three-parameter special cases of GB2 to describe the income distribution in the small area as this choice reduced computational times by a factor of 20. This computational gain was crucial, especially in view of the simulation exercise that we introduced in Section 6, to assess frequentist properties of the Bayesian predictors introduced.

Simulation results confirm that the method that we propose can produce reliable small area estimates of the RMPG. The methodology proposed can be applied to the estimation of other parameters with problems that are similar to those of the RMPG, such as the quintile share ratio. More details on the estimation of this parameter can be found in the on-line appendix 4.

References

Alfons, A. and Templ, M. (2013) Estimation of social exclusion indicators from complex surveys: the R package laeken. *J. Statist. Softwr.*, **54**, 15.

Atkinson, A. B. and Marlier, E. (2010) *Income and Living Conditions in Europe*. Luxembourg: Publication Office of the European Union.

Biewen, M. and Jenkins, S. P. (2006) Variance estimation for generalized entropy and Atkinson inequality indices: the complex survey data case. *Oxf. Bull. Econ. Statist.*, **68**, 371–383.

Chotikapanich, D., Griffiths, W. E., Rao, D. S. P. and Valencia, V. (2012) Global income distributions and inequality, 1993 and 2000: incorporating country-level inequality modeled with Beta distributions. *Rev. Econ. Statist.*, **94**, 52–73.

Clemen, R. C. and Reilly, T. (1999) Correlations and copulas for decision and risk analysis. *Mangmnt Sci.*, **45**, 208–224.

Clemenceau, A. and Museux, J. P. (2007) EU-SILC (community statistics on income and living conditions: general presentation of the instrument). In *Comparative EU statistics on Income and Living Conditions: Issues and Challenges*. Luxembourg: Publication Office of the European Union.

Cowles, M. K. and Carlin, B. P. (1996) Markov Chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Statist. Ass.*, **91**, 883–904.

Datrup, S. R., Hartshorn, R. and McDonald, J. B. (2007) The impact of taxes and transfer payments on the distribution of income: a parametric comparison. *J. Econ. Ineq.*, **5**, 353–359.

Elfadaly, F. G. and Garthwaite, P. H. (2017) Elicting Dirichlet and Gaussian copula prior distributions for multinomial models. *Statist. Comput.*, **27**, 449–467.

Elzhov, T. V., Mullen, K. M., Spiess, A. N. and Bolker, B. (2016) `minpack.lm`: R interface to the Levenberg-Marquardt nonlinear least-squares algorithm found in MINPACK, plus support for bounds. (Available from `https://CRAN.R-project.org/package=minpack.lm`.)

European Commission (2004) A new partnership for cohesion: convergence, competitiveness, cooperation: third report on economic and social cohesion. *Report*. Office for the Official Publications of the European Communities, Luxembourg.

European Commission (2013) *Handbook on Precision Requirements and Variance Estimation for ESS Household Survey*. Luxembourg: Publications Office of the European Union.

Eurostat (2019)*Methodological Manual on Territorial Typologies, 2018 Edition*. Luxembourg: Publications Office of the European Union.

Fabrizi, E., Ferrante, M. R., Pacei, S. and Trivisano, C. (2011) Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Computnl Statist. Data Anal.*, **55**, 1736–1747.

Fabrizi, E., Ferrante, M. R. and Trivisano, C. (2016) Hierarchical Beta regression models for the estimation of poverty and inequality parameters in small areas. In *Analysis of Poverty Data by Small Area Methods* (ed. M. Pratesi), pp. 299–314. Chichester: Wiley.

Fabrizi, E. and Trivisano, C. (2016) Small area estimation of the Gini concentration coefficient. *Computnl Statist. Data Anal.*, **99**, 223–234.

Fuller, W. A. (2009) *Sampling Statistics*. New York: Wiley.

Fusco, A., Guio, A. C. and Marlier, E. (2010) Chracterizing the income poor and the materially deprived in European countries. In *Income and Living Conditions in Europe* (eds A. B. Atkinson and E. Marlier). Luxembourg: Publication Office of the European Union.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

Ghosh, M., Nangia, N. and Kim, D. (1996) Estimation of median income of four-person families: a Bayesian time series approach. *J. Am. Statist. Ass.*, **91**, 1423–1431.

Giorgi, G. M. and Gigliarano, C. (2017) The Gini concentration index: a review of the inference literature. *J. Econ. Surv.*, **31**, 1130–1148.

Goedemé, T. (2013) How much confidence can we have in EU-SILC?: Complex sample designs and the standard error of the Europe 2020 poverty indicators. *Socl Indictrs Res.*, **110**, 89–110.

Graf, M. and Nedyalkova, D. (2011) Parametric estimation of income distributions and derived indicators using the GB2 distribution. In *Report on the Simulation Results, Deliverable 7.1 of the AMELI Project* (ed. B. Hulliger), ch. 7.1.

Graf, M. and Nedyalkova, D. (2014) Modeling of income and indicators of poverty and social exclusion using the Generalized Beta Distribution of the second kind. *Rev. Incm. Wlth*, **60**, 821–832.

Graf, M., Nedyalkova, D., Muennich, R., Seger, J. and Zins, S. (2011) Parametric estimation of income distributions and indicators of poverty and social exclusion, Deliverable 2.1 of the AMELI project.

Hàjek, J. (1958) On the theory of ratio estimates. *Apl. Math.*, **3**, 384–398.

Heidelberger, P. and Welch, P. D. (1983) Simulation run length control in the presence of an initial transient. *Oper. Res.*, **31**, 1109–1144.

Huang, A. and Wand, M. P. (2013) Simple marginally noninformative prior distributions for covariance matrices. *Baysn Anal.*, **8**, 439–452.

Jenkins, S. P. (2009) Distributionally-sensitive inequality indices and the GB2 income distribution. *Rev. Incm. Wlth*, **55**, 392–398.
Kakamu, K. (2016) Simulation studies comparing Dagum and Singh-Maddala income distributions. *Computnl Econ.*, **48**, 593–605.
Kanzow, C., Yamashita, N. and Fukushima, M. (2004) Levenberg-Marquardt methods with strong local convergence properties for solving nonlinear equations with convex constraints. *J. Computnl Appl. Math.*, **172**, 375–397.
Kleiber, C. (1996) Dagum vs. Singh-Maddala income distributions. *Econ. Lett.*, **53**, 265–268.
Kleiber, C. and Kotz, S. (2003) Statistical size distributions in economics and actuarial sciences. New York: Wiley.
McDonald, J. B. (1984) Some generalized functions for the size distribution of income. *Econometrica*, **52**, 647–663.
McDonald, J. B., Sorensen, J. and Turley, P. A. (2013) Skewness and kurtosis properties of income distribution models. *Rev. Incm. Wlth*, **59**, 360–364.
Molina, I., Nandram, B. and Rao, J. N. K. (2014) Small area estimation of general parameters with application to poverty indicators: a hierachical Bayes approach. *Ann. Appl. Statist.*, **8**, 852–885.
Molina, I. and Rao, J. N. K. (2010) Small area estimators of poverty indicators. *Can. J. Statist.*, **38**, 369–385.
Moré, J. J. (1978) The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical Analysis* (ed. G. A. Watson), pp. 105–116. Berlin: Springer.
Ospina, R. and Ferrari, S. L. P. (2012) A general class of zero-or-one inflated beta regression models. *Computnl Statist. Data Anal.*, **56**, 1609–1623.
Peichl, A., Schaefer, T. and Scheicher, C. (2010) Measuring richness and poverty: a micro data application to Europe and Germany. *Rev. Incm. Wlth*, **56**, 597–599.
Pfeffermann, D. (2013) New important developments in small area estimation. *Statist. Sci.*, **28**, 40–68.
Plummer, M., Stukalov, A. and Denwood, M. (2016) rjags. *R Package Version 4-6*. Comprehensive R Archive Network, Vienna.
Rao, J. N. K. and Molina, I. (2015) *Small Area Estimation*. New York: Wiley.
Souza, D. B. and Moura, F. A. S. (2016) Multivariate Beta regression with application in small area estimation. *J. Off. Statist.*, **32**, 747–768.
Statistics Canada (2007) *2005 Survey of Financial Security—Public Use Microdata File, User Guide*. Ottawa: Statistics Canada.
Tarozzi, A. and Deaton, A. (2009) Using census and survey data to estimate poverty and inequality for small areas. *Rev. Econ. Statist.*, **91**, 773–792.
Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T. and Rojas-Perilla, N. (2018) From start to finish: a framework for the production of small area official statistics (with discussion). *J. R. Statist. Soc.* A, **181**, 927–979.
van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes with Applications to Statistics*. New York: Springer.
Wieczorek, J. and Hawala, S. (2011) A Bayesian zero-one inflated Beta model for estimating poverty in U.S. counties. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 2812–2822.
Zellner, A. (1971) Bayesian and non-Bayesian analysis of the log-normal distribution and log-normal regression. *J. Am. Statist. Ass.*, **66**, 327–330.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material'.