Poverty and Inequality Mapping Based on a Unit-Level Log-Normal Mixture Model

Gardini, Aldo; Fabrizi, Enrico; Trivisano, Carlo

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Gardini, A., Fabrizi, E., & Trivisano, C. (2022). Poverty and Inequality Mapping Based on a Unit-Level Log-Normal Mixture Model. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *185*(4), 2073-2096. https://doi.org/10.1111/rssa.12872

Received: 29 July 2021 | Accepted: 17 May 2022

DOI: 10.1111/rssa.12872

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

Poverty and inequality mapping based on a unit-level log-normal mixture model

Aldo Gardini[1] | **Enrico Fabrizi**[2] | **Carlo Trivisano**[1]

[1]Università di Bologna, Bologna, Italy

[2]Università Cattolica del S. Cuore, Piacenza, Italy

Correspondence
Aldo Gardini, Department of Statistical Sciences, Università di Bologna, Italy.
Email: aldo.gardini2@unibo.it

Abstract

Estimating poverty and inequality parameters for small sub-populations with adequate precision is often beyond the reach of ordinary survey-weighted methods because of small sample sizes. In small area estimation, survey data and auxiliary information are combined, in most cases using a model. In this paper, motivated by the analysis of EU-SILC data for Italy, we target the estimation of a selection of poverty and inequality indicators, that is mean, headcount ratio and quintile share ratio, adopting a Bayesian approach. We consider unit-level models specified on the log transformation of a skewed variable (equivalized income). We show how a finite mixture of log-normals provides a substantial improvement in the quality of fit with respect to a single log-normal model. Unfortunately, working with these distributions leads, for some estimands, to the non-existence of posterior moments whenever priors for the variance components are not carefully chosen, as our theoretical results show. To allow the use of moments in posterior summaries, we recommend generalized inverse Gaussian distributions as priors for variance components, guiding the choice of hyperparameters.

KEYWORDS

EU-SILC, generalized inverse Gaussian, Hierarchical Bayes, nested error model, prior sensitivity

1 | INTRODUCTION

The availability of poverty and inequality indicators estimated for small subsets of large populations enhances the understanding of their distribution across geography and social groups, thereby providing useful information for in-depth analysis and policy interventions. There exist many different measures of poverty and inequality based on widely different approaches (see Atkinson, 1987; Grusky et al., 2006; Sen & Foster, 1997, for general introductions). In this paper, we consider measures of poverty and inequality obtained as functions of a single quantitative *size* variable, such as income or consumption.

In most cases, the size variable is measured at the individual/household level using a sample survey. When sample sizes specific to geographical areas or social groups being targeted are too small for most of the sub-populations (generically labelled as areas), small area estimation methods can be used (see Pratesi, 2016, for a general introduction to the topic). The basic idea of small area estimation is to complement survey data with auxiliary information sources (such as censuses, population registers or other archives) not affected by sampling error. Models are often used in this process (see Molina & Rao, 2015, chapter 4).

In this article, we discuss small area estimation of poverty and inequality indicators targeting the working-age population of Italian administrative provinces classified by gender. Italy is partitioned into 110 provinces, whose administrations play an important role in implementing social policies devised at the national or regional level and coordinating the activities of lower administrative levels (municipalities and health districts). To obtain estimates separated by gender is of interest in the context of Italian society, characterized by a marked economic gender divide. Estimation for administrative units and gender is also considered by Esteban et al. (2012) and Marhuenda et al. (2017) for similar reasons.

We use data from the Italian section of the EU-SILC sample survey complemented by the Italian population Census. The poverty and inequality measures we consider are defined as functions of the equalized income. Specifically, we focus on the headcount at-risk-of-poverty rate as a poverty indicator and on the quintile share ratio as a measure of inequality which are both among those selected at the European Council in Laeken, Belgium, in 2001. Moreover, we also study the mean equivalized income as a general affluence measure. With respect to the small area literature, we consider unit-level model-based methods and a Bayesian approach to estimation. For a general discussion of the properties and limitations of small area unit-level methods, see Tarozzi and Deaton (2009). The Bayesian approach to small area estimation offers several advantages, especially when implemented using posterior sampling methods, including the straightforward estimation of non-linear functionals of the size variable and the associated uncertainty measures.

When working with measures based on a positively skewed variable, as income typically is, several authors propose to implement small area methods based on unit-level linear mixed models (and the nested error model, Battese et al. (1988), in particular) specified on the logarithmic transformation of the size variable (Berg & Chandra, 2014; Elbers et al., 2003; Molina & Martin, 2018; Molina & Rao, 2010). The log-transformation can be extended to accommodate negative or zero values of the size variable with recourse to the so-called log-shifted transformation (Rojas-Perilla et al., 2020). Although other transformations (Dagne, 2001; Rojas-Perilla et al., 2020; Sugasawa & Kubokawa, 2019) or different modelling

strategies have been considered (Manandhar & Nandram, 2019), they are not covered in this paper.

The popularity of the log-transformation is related to the assumption of normality on the log-scale, which offers a dramatic simplification in assessing the estimators' properties and computations. Unfortunately, the log-normality is not tenable in most applications (see Graf et al., 2019, for a discussion), and specifically, it is not in ours (see comments to Figure 2 in Section 5). To overcome this problem, we assume that equivalized income can be modelled using a finite mixture of log-normal (LN) distributions, an option already discussed in the literature (Lubrano & Ndoye, 2016). This choice allows us to increase flexibility while keeping some appealing properties of the LN model. Specifically, it can be shown that a finite mixture of LNs is equivalent to a finite mixture of normal distributions on the log of a variable and that a finite mixture of normals can be used to approximate a very large variety of distributions (Ferguson, 1983). Our specification is also related to previous small area literature and namely to Chakraborty et al. (2019), who propose a mixture of two normals to accommodate outlying residuals in nested error regression models. Our approach extends theirs to several mixture components; moreover, we study the model on the log-transformed scale that entails specific problems in the prior distribution specification.

The nested error model on the log-transformed response variable is considered, under a Bayesian approach to estimation, in Molina et al. (2014). Dagne (2001) and Manandhar and Nandram (2019) note that commonly used priors for the variance components lead to posterior distributions with non-existing moments for several functionals of the un-transformed size variable, including mean, median and all those involving the integration of the right tail of the distribution. To solve this problem, Gardini et al. (2021) propose to use generalized inverse Gaussian (GIG) distributions as priors for the variance components of the LN linear mixed model.

The non-existence of posterior moments can be challenging to diagnose in data analysis, particularly when computing means and standard deviations on samples drawn from posterior distributions without moments: these estimates are only occasionally outlying and the instability predicted by theory is not always apparent, leading to possibly misleading inferences.

In this article, we provide theoretical results extending those of Gardini et al. (2021) to the mixture model and the non-linear functionals we target in the analysis. A discussion on the prior parameters choice, model selection and other implementation issues is presented. We show that our modelling strategy not only guarantees the existence of posterior moments for all the considered indicators, but it improves efficiency for those, such as the poverty headcount ratio, not involved in the posterior moments' non-existence problem.

The paper is organized as follows. In Section 2, we introduce our application, providing details about the sample and the available auxiliary information and introducing the estimands we target: the mean, the headcount ratio and the quintile share ratio, as a measure of inequality. We recognize that many other choices are possible, but these indicators are enough to illustrate the main points. In Section 3, we review the literature on small area estimation of poverty and inequality indicators relying on mixed models specified on the log-scale, while the approach we propose is introduced in Section 4. In Section 5, the results obtained in the application to the Italian data are presented, whereas Section 6 contains some outcomes from both model-based and design-based simulation studies. Eventually, Section 7 contains some concluding remarks.

2 | POVERTY MAPPING USING EU-SILC DATA FOR ITALY

2.1 | The sample

In this study, we illustrate an estimation strategy for poverty and inequality indicators defined on the working-age population of Italian administrative provinces classified by gender. To this aim, we use data from the 2012 Italian section of the EU-SILC survey (income reference year is 2011) complemented by auxiliary information from the 2011 Italian population census. The EU-SILC survey is conducted yearly across many European countries by the relevant National Institutes of Statistics, using harmonized questionnaires and survey methodologies (Atkinson & Marlier, 2010, chapter 2). Although following common guidelines, sampling designs are country specific. In Italy, the EU-SILC is a rotating panel survey with a 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 are the primary sampling units and households are the secondary ones. The primary sampling units are divided into strata according to the administrative region they belong to and their population size, while the secondary sampling units are selected by systematic sampling in each primary sampling unit.

In Italy, there are 110 administrative provinces with largely different populations, ranging from the 4.3 million inhabitants of the Rome province, down to less than 0.1 million (Medio Campidano, Isernia and Ogliastra). Provinces are unplanned domains for the EU-SILC, and so are the domains we target, obtained classifying their population by sex and focusing on the age range 16–65.

Domain-specific sample sizes vary from 9 to 907 with a median of 96 in terms of individuals. The size variable used to define the indicators is the equivalized income, defined as the total disposable household income divided by an equivalence factor, specifically the modified-OECD equivalence factor (see Fusco et al., 2010, for more details). Note that equivalized income is the same for all individuals in the same household: because of this strong within-cluster effect, direct estimates based only on area-specific samples are likely to be less efficient than estimates based on simple random samples of individuals of the same size.

Auxiliary information consists of counts of the population classified by administrative province, age class (three levels in the range we consider), sex and education level (four levels), based on the 2011 national Census.

2.2 | The estimands

There exist several poverty indicators; among those based on a single size variable, Foster et al. (1984) provide a popular, fairly general class with many interesting special cases (headcount ratios, poverty gap, poverty severity). As far as the inequality measures are considered, we note that the European Union Council held in Laeken in 2001 focuses on two: the Gini index and the quintile share ratio that are, since then, ordinarily estimated by Eurostat for large populations using the EU-SILC survey data (Atkinson et al., 2004).

To keep things simple, we restrict our attention to three indicators: the mean equivalized income, the at-risk-of-poverty rate and the quintile share ratio. Technically, the mean equivalent income is neither a poverty nor an inequality indicator, but we consider it as it provides valuable information on the affluence of a given population, and it is the most studied indicator in the small area literature; the headcount at-risk-of-poverty rate is a relative poverty measure based on

equivalized income and a country-specific threshold. The quintile share ratio measures inequality by calculating the ratio of the shares of equivalized income detained by the upper quintile to that of the lower quintile.

Let us now introduce some notation. We target a finite population U of size N, partitioned into D sub-populations $U_1, \ldots, U_D$ whose sizes $N_1, \ldots, N_D$ are such that $N = \sum_{d=1}^{D} N_d$. A random sample s of size n is drawn according to a possibly complex design from U. The domain-specific sub-samples are $s_1, \ldots, s_D$, with sizes $n_1, \ldots, n_D, n_d \geq 0$ and $\sum_{d=1}^{D} n_d = n$. Suppose that the parameters we are interested in can be expressed as a function of a variable y, whose unit-level values are denoted with $y_{di}, i = 1, \ldots, N_d, d = 1, \ldots, D$.

The area-level population mean is defined as:

$$\overline{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} y_{di}; \tag{1}$$

while the at-risk-of-poverty headcount ratio (*HCR*) is given by:

$$HCR_d = N_d^{-1} \sum_{i=1}^{N_d} \mathbf{1}\{y_{di} < PT\}, \tag{2}$$

where PT is a poverty threshold defined as 60% of the national median equivalized income. Finally, the quintile share ratio can be expressed as

$$QSR_d = \frac{\sum_{i=1}^{N_d} y_{di} \mathbf{1}\{y_{di} > Q_{d,0.8}\}}{\sum_{i=1}^{N_d} y_{di} \mathbf{1}\{y_{di} < Q_{d,0.2}\}} = \frac{NUM_d}{DEN_d}, \tag{3}$$

where $Q_{d,0.2}$ and $Q_{d,0.8}$ are the first and fourth quintiles of the d-th area.

3 | REVIEW OF EXISTING METHODS

In this section, we shortly review some small area estimators of population parameters (1)–(3). We first introduce direct estimators, that is those based only on area-specific values of the target variable and survey weights; next, among the many proposals in the literature, we introduce the estimators that will be considered in later comparisons.

3.1 | Direct estimators

As most of the domains considered in small area estimation are non-planned, such as those of our application, we consider the Hajek or ratio estimator of the area level mean:

$$\hat{\overline{Y}}_{d,Dir} = \frac{\sum_{i=1}^{n_d} g_{di} y_{di}}{\sum_{i=1}^{n_d} g_{di}}, \tag{4}$$

where g_{di} are the (officially released) sampling weights associated to unit i in area d. Typically, they are weights calibrated to auxiliary information, correcting also for non-response. In the same line, a direct estimator of HCR_d can be defined as

$$\widehat{HCR}_{d,Dir} = \frac{\sum_{i=1}^{n_d} g_{id} \mathbf{1}\{y_{di} < PT\}}{\sum_{i=1}^{n_d} g_{id}}; \tag{5}$$

while a direct estimator of the quintile share ratio can be defined as follows:

$$\widehat{QSR}_{d,Dir} = \frac{\sum_{i=1}^{n_d} g_{di} y_{di} \mathbf{1}\left\{y_{di} > \hat{Q}_{d,0.8}\right\}}{\sum_{i=1}^{n_d} g_{di} y_{di} \mathbf{1}\left\{y_{di} < \hat{Q}_{d,0.2}\right\}}, \tag{6}$$

where $\hat{Q}_{d,0.2}$, $\hat{Q}_{d,0.8}$ are the 20th and 80th percentiles of equivalized income distribution estimated from the d-th area-specific sample. Note that this estimator uses the whole sample information only to obtain $\hat{Q}_{d,0.2}$ and $\hat{Q}_{d,0.8}$, while y_{di} values are only from the tails of the distributions, thus making $\widehat{QSR}_{d,Dir}$ very imprecise when n_d is small.

Standard errors associated with (4)–(6) can be obtained according to linearization-based methods provided that all relevant sample information is available to the data analyst. If this is not the case, good approximations can be obtained under limited information (see Fabrizi et al., 2020, for a discussion). More details for the quintile share can be found in Langel and Tillé (2011).

3.2 | Empirical Bayes predictors

Empirical Bayes prediction (EBP) is a popular model-based approach for estimating the desired area-level quantities. The basic idea is to carry out a Bayesian analysis of the model conditional on the variance components, and then use a frequentist method for their estimation. The nested error linear regression model (Battese et al., 1988) is often assumed for the response variable y_{di} or a transformation $T(y_{di})$ of it. Among the possible transformations for the response, our focus is on the shifted logarithmic one, that is widely used and induces a LN mixed model on the response. In this case, the response variable used to specify the model is $w_{di} = \log(y_{di} + c)$, where c is a fixed constant that is required to guarantee that the shifted responses $y^*_{di} = y_{di} + c$ are positive for all the observed data:

$$\begin{aligned} &w_{di} = \mathbf{x}_{di}^T \boldsymbol{\beta} + u_d + e_{di}; \\ &u_d|\tau^2 \sim \mathcal{N}(0, \tau^2), \;\; e_{di}|\sigma^2 \sim \mathcal{N}(0, \sigma^2); \;\; d = 1, \ldots, D; \;\; i = 1, \ldots, N_d; \end{aligned} \tag{7}$$

where $\mathbf{x}_{di} \in \mathbb{R}^p$ is the vector of observed covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients and u_d is the area-specific random effect. A general discussion of the theory related to the analysis of this model can be found in Molina and Rao (2010); due to its popularity, the EB method can be readily applied to estimation using some available `R` packages such as `emdi` (Kreutzmann et al., 2019).

Once the model parameters are estimated, a Monte Carlo procedure can be used to evaluate the target predictors: at each iteration, a synthetic population is generated, and the indicator is computed for each small area. The average of such replicates is considered as EBP estimate of the indicator. To quantify the uncertainty of the estimates, the MSE of the EBP needs to be evaluated and bootstrap schemes are usually adopted. For further details on these aspects, see Molina and Rao (2010) and Berg and Chandra (2014). Recently, with reference to the prediction of area-level

means under model (7), Molina and Martin (2018) obtained a second-order asymptotic approximation of the mean crossed predictor errors that lead to good approximations for the MSEs of predicted area means.

3.3 | Hierarchical Bayes predictors

The shifted LN unit-level model (7) is analysed from a Hierarchical Bayes (HB) perspective by Molina et al. (2014). Specifically, they aimed at estimating poverty indicators belonging to the family introduced by Foster et al. (1984): their area-specific posterior distributions are obtained combining the sample data and the posterior predictive distributions of the out-of-sample units. Then, the Bayes estimators under quadratic loss, that is the posterior means, are computed and proposed as HB estimates.

Molina et al. (2014) propose an improper prior setting for the model parameters:

$$p(\boldsymbol{\beta}) \propto 1, \quad p(\sigma^2) \propto \frac{1}{\sigma^2}, \quad p(\rho) \propto \mathbf{1}_{(0,1)}(\rho),$$

where $\rho = \tau^2/(\tau^2 + \sigma^2)$ is the intraclass correlation coefficient. These assumptions induce the uniform shrinkage prior, which usually produces estimators endowed with nice frequentist properties (Natarajan & Kass, 2000).

4 | THE PROPOSED APPROACH: A BAYESIAN FINITE MIXTURE OF LOG-NORMALS

In many real data applications, the log-normality assumption may not be suitable to model the target response. For this reason, we propose to extend model (7) to a mixture of normal distributions in the log scale. The following model is specified for the logarithmic transformation of the shifted response:

$$\begin{aligned} &w_{di} = \mathbf{x}_{di}^T\boldsymbol{\beta} + u_d + e_{di};\ d = 1, \ldots, D;\ i = 1, \ldots, N_d; \\ &u_d|\tau^2 \sim \mathcal{N}(0, \tau^2),\ e_{di}|\sigma_1^2, \ldots, \sigma_K^2 \sim \sum_{k=0}^{K} \pi_k \mathcal{N}(0, \sigma_k^2); \end{aligned} \tag{8}$$

where K is an integer defining the number of mixture components, $\pi_1, \ldots, \pi_K$ are the weights, with $\sum_k \pi_k = 1$, and $\sigma_1^2, \ldots, \sigma_K^2$ are the variances characterizing the K distinct components. In the proposed model formulation, the linear predictor is common for all the components, whereas the variances of the individual error e_{di} are allowed to change, in line with the work by Chakraborty et al. (2019). By doing so, the flexibility of the chosen model is remarkably enhanced, although the relations with the normal model and its good computational features are preserved. The simple LN model (7) is a particular case of model (8) when $K = 1$. Note that model (8), as well as (7) are supposed to hold at the population level. We assume that the sampling design is non-informative, with the implication that the same models can be used to describe the sample data. As for the prior distributions, we start from the regression slopes and the mixture weights for which we specify:

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{b}_0, \mathbf{V}_0), \quad \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\mathbf{1}); \tag{9}$$

where **1** is a K-dimensional vector of ones that allows to set a uniform distribution on the simplex for the vector of weights.

Prior distributions for the variance components, that is τ^2 and $\sigma_1^2, \dots, \sigma_K^2$ will be discussed more in detail later, as their choice can cause the posterior predictive distributions of out-of-sample values to have non-finite moments (Fabrizi & Trivisano, 2016a; Gardini et al., 2021). To guarantee the identifiability of $\sigma_1^2, \dots, \sigma_K^2$, they are assumed to be ordered: $\sigma_1^2 < \cdots < \sigma_K^2$. However, we are not interested in identifying the mixture components to which each individual unit belongs, as the main focus of inference is on predictive distribution that does not depend on components identification.

4.1 | Useful preliminary results

To simplify the notation of the subsequent sections, let us first introduce the vectors $\mathbf{y}_s$, $\mathbf{y}_s^*$ and $\mathbf{w}_s$ containing the responses registered for the sampled units and their transformations defined in Section 3.2, $\mathbf{X}_s$ is the design matrix with the covariates information. The likelihood of the model introduced in Equation (8) becomes more tractable when introducing an n-dimensional vector of latent variables $\mathbf{z}_s$, in which the generic component $z_{di} \in \{1, \dots, K\}$ represents the label of the component to which the i-th unit of the d-th area belongs:

$$z_{di}|\boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi}).$$

In this way, the likelihood of $\mathbf{y}_s$ conditioned to the vector of labels $\mathbf{z}_s$ can be indicated as a multivariate LN whose scalar component is

$$y_{di}^*|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, u_d, z_{di} \sim \mathcal{LN}\left(\mathbf{x}_{di}^T\boldsymbol{\beta} + u_d, \sigma_{z_{di}}^2\right),$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$. This model representation allows retrieving more straightforwardly the full conditionals of the model parameters that are shared by all the mixture components, such as the vector of coefficients $\boldsymbol{\beta}$ and the area-specific random effects u_d. In particular, Proposition 1 contains distributional results useful for the developments presented in the subsequent sections.

Proposition 1. *Under model* (8) *and priors in* (9), *the following distributions can be derived conditionally on the latent variables* $\mathbf{z}_s$.

(a) *Full conditional of the area-specific random effect:*

$$\begin{aligned} &u_d|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \tau^2, \mathbf{y}_s, \mathbf{z}_s \sim \mathcal{N}\left(\overline{u}_d, V_u\right), \\ &\overline{u}_d = V_u\left(\sum_{i=1}^{n_d}\sigma_{z_{di}}^{-2}\left(w_{di} - \mathbf{x}_{di}^T\boldsymbol{\beta}\right)\right), \quad V_u = \left(\sum_{i=1}^{n_d}\sigma_{z_{di}}^{-2} + \tau^{-2}\right)^{-1}; \end{aligned} \tag{10}$$

(b) *Full conditional of the regression coefficients with the random effects integrated out:*

$$\begin{aligned} &\boldsymbol{\beta}|\boldsymbol{\sigma}^2, \tau^2, \mathbf{y}_s, \mathbf{z}_s \sim \mathcal{N}\left(\overline{\boldsymbol{\beta}}, \mathbf{V}_\beta\right), \\ &\overline{\boldsymbol{\beta}} = \mathbf{V}_\beta\left(\mathbf{X}_s^T\text{diag}\left(\left[\sigma_{z_j}^{-2}\right]_{j\in\{1,\dots,n\}}\right)\mathbf{w}_s\right. \end{aligned}$$

$$+ \mathbf{X}_s^T \text{diag}\left(\left[-\frac{\tau^2}{\sigma_{z_j}^4\left(1+\tau^2\left(\sum_{i=1}^{n_{d(j)}}\sigma_{z_{d(j)i}}^{-2}\right)\right)}\right]_{j\in\{1,\dots,n\}}\right)\mathbf{w}_s + \mathbf{V}_0^{-1}\mathbf{b}_0\Bigg),$$

$$\mathbf{V}_\beta = \left(\mathbf{X}_s^T \text{diag}\left(\left[\frac{\sigma_{z_j}^2\left(1+\tau^2\left(\sum_{i=1}^{n_{d(j)}}\sigma_{z_{d(j)i}}^{-2}\right)\right)-\tau^2}{\sigma_{z_j}^4\left(1+\tau^2\left(\sum_{i=1}^{n_{d(j)}}\sigma_{z_{d(j)i}}^{-2}\right)\right)}\right]_{j\in\{1,\dots,n\}}\right)\mathbf{X}_s + \mathbf{V}_0^{-1}\right)^{-1}, \tag{11}$$

where j is an index running the n units, $d(j)$ indicates the area in which the j-th unit is included, and $\text{diag}([\kappa(j)]_{j\in\{1,\dots,n\}})$ *defines a diagonal matrix with the n-dimensional vector* $[\kappa(j)]_{j\in\{1,\dots,n\}}$ *on it.*

Proof. See Supplementary Material.

4.2 | HB estimators

The posterior predictive distribution of the unsampled units plays a crucial role in developing HB predictors for the target indicators introduced in Section 2.2. Indicating with $\tilde{y}_{di}$, $i = n_d + 1, \dots, N_d$ the out-of-sample observations for area d, their posterior predictive distribution is defined as:

$$f\left(\tilde{y}_{di}|\mathbf{y}_s\right) = \int_{\boldsymbol{\Omega}} f\left(\tilde{y}_{di}|\boldsymbol{\omega}\right) f\left(\boldsymbol{\omega}|\mathbf{y}_s\right) \mathrm{d}\boldsymbol{\omega}, \tag{12}$$

where $\boldsymbol{\omega}$ is the vector of the model parameters and $\boldsymbol{\Omega}$ is the parameter space. Under the considered statistical model, it is not possible to retrieve an analytical form of this density, but the distribution can be approximated through numerical methods, exploiting the draws from the posterior distribution of the parameters $f(\boldsymbol{\omega}|\mathbf{y}_s)$.

Let us now indicate with $\tilde{y}_{di}^{(m)}$ the m-th MCMC replicate drawn from the posterior predictive distribution (12). Exploiting the latent variables representation of the mixture, it is possible to generate first replicate of $\tilde{z}_{di}$:

$$\tilde{z}_{di}^{(m)}|\boldsymbol{\pi}^{(m)} \sim \text{Categorical}\left(\boldsymbol{\pi}^{(m)}\right),$$

that represents the label for the mixture component from which a replicate of the shifted response is drawn:

$$\tilde{y}_{di}^{*(m)}|\boldsymbol{\beta}^{(m)}, u_d^{(m)}, \sigma_k^{2(m)}, \tilde{z}_{di}^{(m)} \sim \mathcal{LN}\left(\tilde{\mathbf{x}}_{di}^T\boldsymbol{\beta}^{(m)} + u_d^{(m)}, \sigma_{\tilde{z}_{di}^{(m)}}^{2(m)}\right),$$

and, lastly, $\tilde{y}_{di}^{(m)} = \tilde{y}_{di}^{*(m)} - c$.

The vector of observed responses $\mathbf{y}_s$ and the m-th replicate for the vector of out-of-sample observations $\tilde{\mathbf{y}}_s^{(m)}$ are combined to retrieve a realization from the posterior distribution of the area mean, HCR and QSR indicators:

$$\overline{Y}_d^{(m)}|\mathbf{y}_s, \tilde{\mathbf{y}}_s^{(m)} = N_d^{-1}\left[\sum_{i=1}^{n_d} y_{di} + \sum_{i=n_d+1}^{N_d} \tilde{y}_{di}^{(m)}\right], \tag{13}$$

$$HCR_d^{(m)}|\mathbf{y}_s, \tilde{\mathbf{y}}_s^{(m)} = N_d^{-1}\left[\sum_{i=1}^{n_d}\mathbf{1}\left\{y_{di} < PT\right\} + \sum_{i=n_d+1}^{N_d}\mathbf{1}\left\{\tilde{y}_{di}^{(m)} < PT\right\}\right], \tag{14}$$

$$QSR_d^{(m)}|\mathbf{y}_s, \tilde{\mathbf{y}}_s^{(m)} = \frac{\sum_{i=1}^{n_d} y_{di}\mathbf{1}\left\{y_{di} > \hat{Q}_{d,0.8}^{(m)}\right\} + \sum_{i=n_d+1}^{N_d}\tilde{y}_{di}^{(m)}\mathbf{1}\left\{\tilde{y}_{di}^{(m)} > \hat{Q}_{d,0.8}^{(m)}\right\}}{\sum_{i=1}^{n_d} y_{di}\mathbf{1}\left\{y_{di} < \hat{Q}_{d,0.2}^{(m)}\right\} + \sum_{i=n_d+1}^{N_d}\tilde{y}_{di}^{(m)}\mathbf{1}\left\{\tilde{y}_{di}^{(m)} < \hat{Q}_{d,0.2}^{(m)}\right\}}, \tag{15}$$

where $\hat{Q}_{d,0.2}^{(m)}$ and $\hat{Q}_{d,0.8}^{(m)}$ represent the first and the fourth quintiles computed on the m-th predicted population for area d, that is $[y_{d1}, \dots, y_{dn_d}, \tilde{y}_{d[n_d+1]}^{(m)}, \dots, \tilde{y}_{dN_d}^{(m)}]$.

The Bayes estimator under quadratic loss, that is the posterior mean, is proposed as the HB estimator of the target quantity. Once an MCMC sample of size M from the posterior distributions of the indicators is available, it is possible to approximate the expectation

$$\mathbb{E}\left[\theta_d|\mathbf{y}_s\right] \simeq M^{-1}\sum_{m=1}^{M}\theta_d^{(m)} = \hat{\theta}_{d,HB},$$

with $\hat{\theta}_{d,HB}$ denoting the HB estimate of the generic target parameter for area d: $\theta_d \in \{\overline{Y}_d, HCR_d, QSR_d\}$. Similarly, other summaries of the θ_d posterior can be computed, such as the variance and the quantiles, useful to determine the credible intervals.

4.3 | Existence of posterior moments

We now present some theoretical results establishing a connection between the prior chosen for the variance components and the existence of posterior moments for our target estimands. To this aim, the partial moments need to be introduced, with the following notation:

$$\mathbb{E}_\alpha^\beta\left[V^r\right] = \int_\alpha^\beta v^r p(v)\mathrm{d}v = \mathbb{E}_\alpha^\beta\left[V^r|\alpha < V < \beta\right]\mathbb{P}[\alpha < V < \beta].$$

We can now state some preliminary results concerning the elements of the estimands.

Theorem 1. *Considering model* (8), $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{b}_0, \mathbf{V}_0)$, *and a prior for* σ^2 *with support* $\mathbb{R}^+$, *then*:

(a) $\mathbb{E}[\tilde{y}_{di}^r|\mathbf{y}_s] < +\infty$ *if and only if the prior of* σ_k^2, $\forall k$, *has a density function with a term* $\exp\{-t\sigma_k^2\}$ *and*:

$$t > \frac{r^2}{2}\left[1 + \tilde{\mathbf{x}}_{di}^T\left(\mathbf{X}_s^T\mathbf{X}_s + \mathbf{V}_0^{-1}\right)\tilde{\mathbf{x}}_{di}\right].$$

(b) $\mathbb{E}_q^{+\infty}[\tilde{y}_{di}^r|\mathbf{y}_s] < +\infty$ *if and only if the prior of* σ_k^2, $\forall k$, *has a density function with a term* $\exp\left\{-t\sigma_k^2\right\}$ *and*:

$$t > \frac{r^2}{2}\left[1 + \tilde{\mathbf{x}}_{di}^T\left(\mathbf{X}_s^T\mathbf{X}_s + \mathbf{V}_0^{-1}\right)\tilde{\mathbf{x}}_{di}\right].$$

(c) $\mathbb{E}_0^q[\tilde{y}_{di}^r|\mathbf{y}_s] < +\infty$ *always.*
(d) $\mathbb{E}[\mathbf{1}\{\tilde{y}_{di} < PT\}|\mathbf{y}_s] < +\infty$ *always.*

Proof. See Supplementary Material.

Let us now state the results about the estimands of interest in this research, outlining the relationship between the moments existence and the priors for the variance components.

Theorem 2. *Considering model* (8), $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{b}_0, \mathbf{V}_0)$, *and a prior for* σ^2 *with support* $\mathbb{R}^+$, *then*:

(a) $\mathbb{E}[\overline{Y}_d^r|\mathbf{y}_s] < +\infty$ *if and only if the prior of* σ_k^2, $\forall k$, *has a density function with a term* $\exp\{-t\sigma_k^2\}$ *and*:

$$t > \frac{r^2}{2}\left[1 + \max_i \left\{\tilde{\mathbf{x}}_{di}^T \left(\mathbf{X}_s^T\mathbf{X}_s + \mathbf{V}_0^{-1}\right)\tilde{\mathbf{x}}_{di}\right\}\right].$$

(b) $\mathbb{E}[QSR_d^r|\mathbf{y}_s] < +\infty$ *if and only if*:

- *it is assumed that* $\mathbb{P}[DEN_d > \varepsilon > 0|\mathbf{y}_s] = 1$;
- *the prior of* σ_k^2, $\forall k$, *has a density function with a term* $\exp\{-t\sigma_k^2\}$ *and*:

$$t > \frac{r^2}{2}\left[1 + \max_i \left\{\tilde{\mathbf{x}}_{di}^T \left(\mathbf{X}_s^T\mathbf{X}_s + \mathbf{V}_0^{-1}\right)\tilde{\mathbf{x}}_{di}\right\}\right].$$

(c) $\mathbb{E}[HCR_d|\mathbf{y}_s] < +\infty$ *always*.

Proof. See Supplementary Material.

We note that the assumption $\mathbb{P}[DEN_d > \varepsilon > 0|\mathbf{y}_s] = 1$ in the statement *(c)* of theorem 2 is not restrictive at all in the context of our application. DEN_d is the sum of all incomes in the first quintile of the N_d units in area d. As all incomes are positive with the exception of a small fraction of negative or zero incomes, the expected probability that DEN_d is negative can be assumed zero. This is confirmed in the analysis of our data set and in all simulations of Section 6.

4.4 | Prior specification

To obtain finite posterior moments for the target quantities, a prior distribution that allows fulfilling the conditions derived in Theorem 2 must be considered for the variances $\sigma_1^2, \ldots, \sigma_K^2$. Our choice is the GIG distribution, a flexible three-parameters distribution with positive support already considered in Gardini et al. (2021) for the variance components in the LN mixed model. If $V \sim GIG(\lambda, \delta, \gamma)$, then its probability density function is:

$$p(v) = \left(\frac{\gamma}{\delta}\right)^{\lambda} \frac{1}{2K_\lambda(\delta\gamma)} v^{\lambda-1} \exp\left\{-\frac{1}{2}(\delta^2 v^{-1} + \gamma^2 v)\right\} \mathbf{1}_{\mathbb{R}^+}, \tag{16}$$

where $\lambda \in \mathbb{R}$ is the shape parameter, $\delta \in \mathbb{R}^+$ is the scale parameter, $\gamma \in \mathbb{R}^+$ is the tail parameter and $K_l(x)$ is the modified Bessel function of the second kind with index l evaluated in x.

The tail parameter γ can be fixed in order to satisfy the conditions of Theorem 2 by setting $\gamma > r\sqrt{[1 + \tilde{\mathbf{x}}_{di}^T(\mathbf{X}_s^T\mathbf{X}_s + \mathbf{V}_0^{-1})\tilde{\mathbf{x}}_{di}]}$, $\forall d, i$. In line with the arguments provided in Gardini et al. (2021), we propose the following priors for the variance components:

$$\sigma_k^2 \sim GIG\,(1, 0.01, \gamma_0)\,, \quad k = 1, \ldots, K; \qquad \tau^2 \sim GIG\,(1, 0.01, \gamma_0)\,, \tag{17}$$

with $\gamma_0 = (r+1)\sqrt{[1 + \max_{d,i}\{\tilde{\mathbf{x}}_{di}^T(\mathbf{X}_s^T\mathbf{X}_s + \mathbf{V}_0^{-1})\tilde{\mathbf{x}}_{di}\}]}$, in order to guarantee the existence of the posterior moments of the functionals for any area d. In this way, the priors on the variances are

approximately gamma distributions $Gamma(\text{shape} = 1, \text{rate} = \gamma_0^2/2)$ and the induced prior on the intraclass correlation coefficients $\rho_k = \tau^2/(\tau^2 + \sigma_k^2)$ are approximately uniform distributions in the range (0; 1).

4.5 | Model fitting and comparison

To draw samples from the posterior distribution of the model parameters, we adopt the `Stan` probabilistic programming language (Carpenter et al., 2017), exploiting the `rstan` package to interface it with `R` (R Core Team, 2021). It allows programming the target statistical model, manually specifying the GIG distribution that is not included among the default probability distributions. The code used to fit the proposed models is available as supplementary material.

4.5.1 | Model choice and goodness of fit

It is possible to combine the M-dimensional MCMC samples from the posterior distributions of parameters in order to compute measures aimed at evaluating the fit of mixture models, using them to choose the most appropriate value of K. Among the plethora of methods have been proposed in the literature to study model performances, we implemented the conditional predictive ordinates (CPOs), the leave-one-out cross-validation information criterion (LOOIC) and some tools based on the posterior predictive distribution such as the Bayesian p-values associated to the posterior predictive checks.

The CPO (Gelfand, 1996; Ntzoufras, 2009) is a unit-specific measure that helps to understand if the observed response is expected under the model or represents an outlier. Given an observed response y_{di}, the corresponding indicator is defined as the leave-one-out cross-validation predictive density: $CPO_{di} = f(y_{di}|\mathbf{y}_{s\setminus\{di\}})$, since $\mathbf{y}_{s\setminus\{di\}}$ indicates the vector of all the observed responses with the exception of y_{di}. It can be proved that a Monte Carlo approximation of the CPO is $\widehat{CPO}_{di} = \left(\frac{1}{M}\sum_{m=1}^{M} f(y_{di}|\boldsymbol{\omega}^{(m)})^{-1}\right)^{-1}$. A small CPO value highlights that the observed value for that unit is unexpected under the fitted model. Ntzoufras (2009), as a rule of thumb, defines an observation such that $\widehat{CPO}_{di} < 1/40$ as possible outlier and $\widehat{CPO}_{di} < 1/70$ as extreme value.

The predictive density can also be employed in producing overall measures of the predictive accuracy of the model. The LOOIC, introduced by Vehtari et al. (2017) is defined as $LOOIC = \sum_{d=1}^{D}\sum_{i=1}^{n_d} \log f(y_{di}|\mathbf{y}_{s\setminus di})$. The authors proposed to estimate it using a Pareto smoothed importance sampling (PSIS-LOOIC), avoiding to re-fit the model n times. The model presenting the smaller LOOIC value is preferable. A measure of uncertainty is available, to better understand the relevance of the differences among the models. The `R` package `loo` (Vehtari et al., 2020) provides a function that allows to readily compute this indicator.

To understand if the fitted model can capture the basic features of the analysed data, posterior predictive checks can be carried out (Gabry et al., 2019). They rely on the random generation of M replicated data sets $\tilde{\mathbf{y}}_s^{(m)}$ from the posterior predictive distribution of the fitted model, defined in Section 4.2. Then, it is possible to compare these replicates $\tilde{\mathbf{y}}_s^{(m)}$ to the observed data $\mathbf{y}_s$ in several ways.

5 | APPLICATION TO THE ITALIAN DATA

In this section, we discuss the application of our methodology to the data described in Section 2. As anticipated, the domains we target are the Italian provinces (NUTS-3 regions) stratified by gender; as auxiliary variables we consider age class and education level. Out of sample information comes from the 2011 national Census. The recourse to small area estimation can be easily motivated, considering the poor precision of direct estimators. If we consider headcount ratios direct estimators (5), the coefficients of variation have a median (over the ensemble of the small area) of 0.31. This implies that, for nearly half of the areas, the estimates would be unpublishable according to the criteria adopted, among others, by Statistics Canada (2007). The situation is substantially worse for the quintile share ratio, whose direct estimator is much more unstable, as discussed in Section 2.

A small minority of the observed equivalized income values is negative or zero. To accommodate these observations, we adopt the transformation $y^*_{di} = y_{di} + c$ introduced in Section 3.2. The constant c is fixed to guarantee the transformed values to be positive while minimizing the skewness of the residuals in a simple LN regression (Molina et al., 2014). In principle, c should be treated as an additional model parameter and a prior distribution specified for it. Rojas-Perilla et al. (2020) review frequentist methods for its estimation; as they recognize that accounting for the uncertainty associated to it has a minor impact on the quality of small area estimation (Section 7.3), we prefer to treat it as a constant for simplicity's sake. Our approach is based on the assumption of a parametric model for the equivalized income, comparing the results obtained under the model in Equation (8) for different numbers of mixture components. In all cases, the priors for the variance components are assumed to be GIG distributed to guarantee the existence of the HB estimators of the target estimands (see Section 4.3).

In order to understand if the proposed prior setting is calibrated with respect to the magnitude of the analysed data, we compare the observed data in the log-scale to fake-data generated from prior predictive distributions, as advised in Gabry et al. (2019). We consider the single LN model, that is the model in Equation (8) with $K = 1$, and the prior in (9) for $\boldsymbol{\beta}$. To specify a weakly informative prior on the centred transformed data, in line with the popular `rstanarm` package (Goodrich et al., 2020), we set $\mathbf{V}_0 = \text{diag}([2.5^2 s^2_w, 2.5^2 s^2_w/s^2_{x1}, \ \dots \ , 2.5^2 s^2_w/s^2_{xp}])$ and $\mathbf{b}_0 = [\overline{w}, 0, \ \dots \ , 0]$, where $\overline{w}$ and s^2_w are the mean and the variance of $\mathbf{w}_s$, whereas s^2_{xj} is the variance of the j-th covariate. Eventually, as prior for the variance components, we compare the suggested strategy relying on the GIG distribution (described in Section 4.4) to the popular small parameters inverse-gamma priors. The distributions of fake data generated under the two prior settings are displayed in Figure 1: the range of the data drawn from the model with GIG priors is compliant to those of the observed data, resulting a prior distribution well calibrated for the faced problem.

In Table 1, the model comparison tools based on the posterior predictive densities (see Section 4.5) are reported. We note how the mixtures of LN models show a clear advantage with respect to the simple LN model. When increasing the number of mixture components from two (LNM) to three (LNM-3), improvements are not apparent: a slight one is observed in terms of LOOIC, but well within the range of one standard deviation, while the percentages of observations that can be classified as outliers or extreme in terms of $\widehat{CPO}$ slightly increase. A parsimony criterion suggests the adoption of the LNM model to analyse our data.

The superiority of the fit provided by the LNM model with respect to the more widely used LN model is confirmed by considering posterior predictive checks. Based on the generation of M data set from the posterior predictive distributions, we evaluated both the kernel density (Figure 2, $M = 250$) and calculated posterior p-values at the area level for the prediction of simple summary

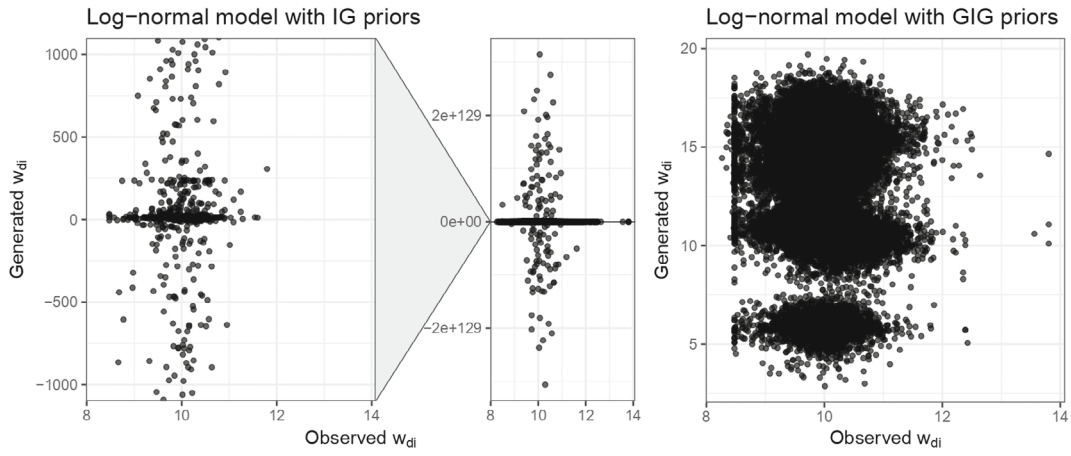

FIGURE 1 Prior predictive checks: observed data in the log scale compared to fake data generated from the prior predictive distribution of the single component log-normal model, assuming either GIG priors (left panel) or IG priors (right panel) for the variance components.

TABLE 1 Leave-one-out information criterion (LOOIC) and % of units with *CPO* below 1/40 and 1/70 for the models being compared: log-normal (LN), two-components log-normal mixture (LNM) and three-components log-normal mixture (LNM-3)

	LOOIC (S.E.)	**% *CPO* < 0.025**	**% *CPO* < 0.014**
LN	46412 (428)	2.97	2.21
LNM	44589 (360)	1.96	1.06
LNM-3	44562 (356)	2.01	1.18

measures (Figure 3, $M = 4000$). In Figure 2, the comparison of the generated-data distributions to the observed sample shows not only that the LNM model is better, but also how the LN model shows a poorer fit in both tails of the distribution. The issues in modelling the left side of the distribution are mainly caused by the irregular behaviour induced by the presence of a peak of values around 0 (before the shifting): this leads to the inflation of the unique variance parameter from which a remarkable difference in the density around the mode follows.

A clearer picture of the problems met by the simple LN model emerges from Table 2. Specifically, from this table, we can note how the posterior mean of the unique individual level standard deviation σ_1 can be obtained approximately as a weighted average between the posterior means of the two corresponding parameters in the LNM model: the added flexibility translates into a better fit. We also note that the posterior distribution of the parameter related to area-level random effects τ are very close.

From Figure 3, we note that, while the LN model predicts adequately area-level means, its performance in predicting the 20th and 80th percentiles and the standard deviation is poorer than that of LNM, as high bars appear in the tails of the histogram. In summary, the flexibility allowed by the mixture models improves the fit on the tails of the distribution; not only on the right one but, notably, also on the left.

The better fit obtained with few additional parameter also implies a better performance in terms of precision of point predictors (i.e. posterior means of HB predictors). Table 3 compares

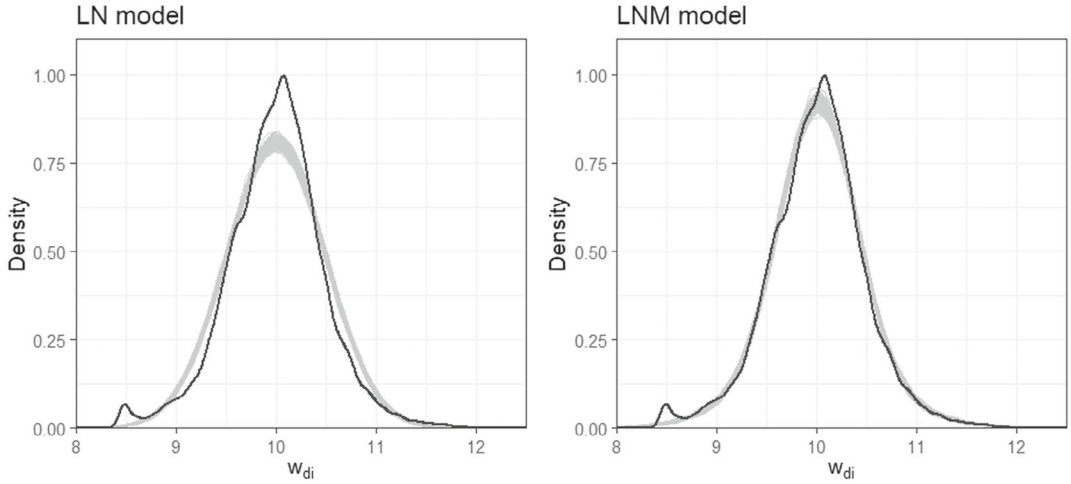

FIGURE 2 Kernel density estimates of $M = 250$ replicated data set from the posterior predictive distributions (grey lines) compared to the kernel density estimation of the actual sample data (black solid line).

TABLE 2 Posterior distribution summaries for variance components for both the log-normal (LN) and two components mixture log-normal (LNM) models

	LN				LNM			
	Mean	**S.D.**	**q05**	**q95**	**Mean**	**S.D.**	**q05**	**q95**
σ_1	0.439	0.002	0.436	0.442	0.310	0.004	0.304	0.317
σ_2	–	–	–	–	0.666	0.011	0.649	0.684
τ	0.157	0.008	0.144	0.172	0.152	0.008	0.140	0.166
π_1	–	–	–	–	0.722	0.014	0.697	0.745

the posterior standard deviations to estimates of the design-based standard errors computed according to the methodology illustrated in Section 2 by calculating the percentage standard deviation reduction (SDR), which is defined as $SDR(\hat{\theta}_{d,Est}) = 100\left(1 - \sqrt{\mathbb{V}[\hat{\theta}_{d,Est}|\mathbf{y}_s]/\mathbb{V}[\hat{\theta}_{d,Dir}]}\right)$, for $\theta \in \{\overline{Y}, HCR\}$ and $Est \in \{HB{-}LN, HB{-}LNM\}$. Not only the average, but also relevant percentiles of their distribution across the ensemble of the areas are reported. We note that model-based estimators, either based on LN or LNM models, are effective in improving the precision of direct estimators. Their performances are comparable as far as the estimation of small area means is concerned, while a clear difference in favour of the estimates based on the LNM emerges for the headcount ratio, in line with the better fit on the left tail of the distribution we just discussed. Figures for the quintile ratio are not reported, as standard errors associated with direct estimators would be very large, leading to a comparison of little use.

The comparison between HB estimates obtained under the LN and LNM models, shown in Figure 4, highlights that the model specification notably impacts the estimation of the target parameters. Mean point estimates are mainly in agreement, with lower posterior standard

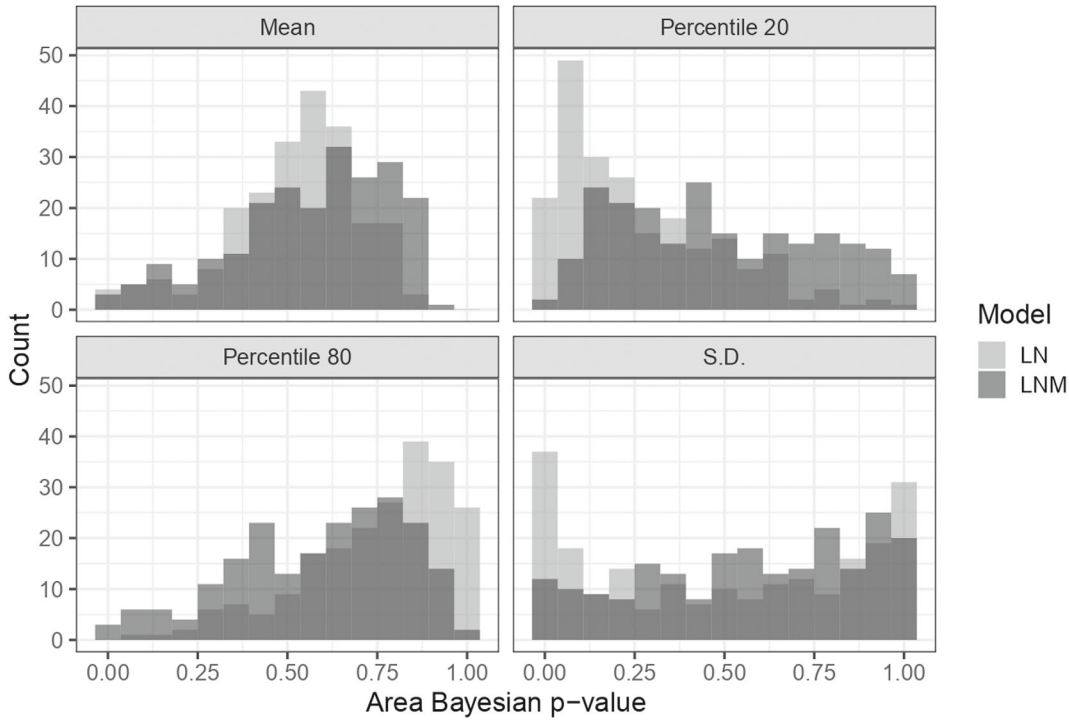

FIGURE 3 Histograms of posterior p-values based on $M = 4000$ replicated data set from the posterior predictive distribution of area-specific summary statistics.

TABLE 3 Summaries of the standard deviation reductions (SDRs): distributions over the ensamble of the small areas

		$SDR(\hat{\bar{Y}}_{d,Est})$ %		$SDR(\widehat{HCR}_{d,Est})$ %	
	n_d	HB-LN	HB-LNM	HB-LN	HB-LNM
Mean	144	23.1	30.1	43.1	51.9
10th-perc.	34	−1.4	7.4	17.5	33.9
25th-perc.	66	12.4	20.9	39.9	45.9
Median	96	24.3	31.1	49.2	53.6
75th-perc.	183	38.2	36.5	41.2	63.8
90th-perc.	286	49.4	53.6	63.6	71.9

deviations for the LNM model when the samples sizes are particularly small (and standard deviations large). The differences become systematic in the case of the other two parameters, for which estimates based on the LN models tend to be regularly higher and very markedly so in the case of the quintile share ratio. For this latter parameter, the posterior standard errors are lower when the LNM is adopted and the differences increase for the areas characterized by smaller samples. More conclusive evidences in favour of the mixture model for the estimation of the quintile share ratio will be provided in the following section, when the estimators associated with the

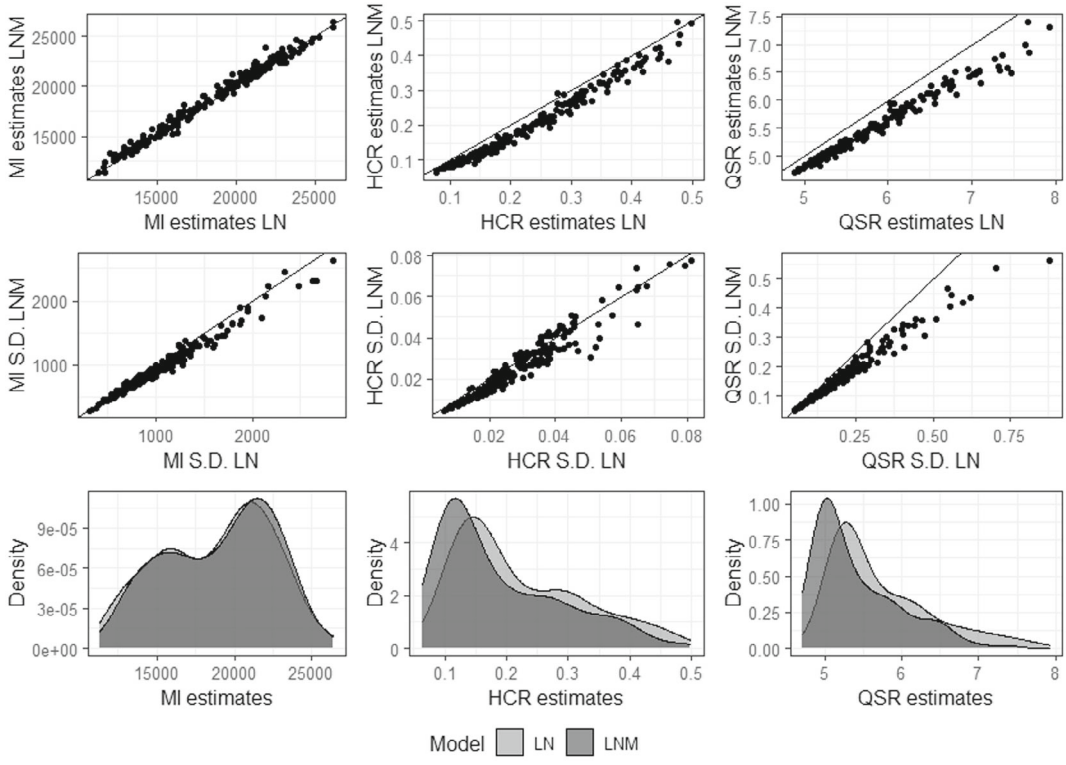

FIGURE 4 Comparison among Hierarchical Bayes estimates from log-normal (LN) and LN mixture (LNM) models. MI indicates the mean income $\overline{Y}_d$.

two models are compared by means of a Monte Carlo simulation exercise. Another argument for the choice $K = 2$ instead of $K = 3$ can be drawn from Figure S1 in the Supplementary material: the model-based estimates and the posterior standard deviations of the target quantities under LNM-3 are considerably close to those obtained under the simpler LNM model (all correlation coefficients >0.995).

Although this paper aims to illustrate a suitable methodology for the problem at hand, we add a few comments about the results from an economic analysis perspective to highlight the relevance of small area estimates. Specifically, in Figure 5, we compare the estimates obtained for the male and female sub-populations for each province. Females appear to be in a worse situation with respect to all the parameters being studied: the means appear to be somewhat larger for males, while the headcount ratios and the quintile share ratios are lower. This result could be expected in view of the gender wage gap that still characterizes Italy's labour market (Mussida & Picchio, 2014). Nonetheless, it is a little surprising as the same equivalized income, on which all measures are based, is shared by all household members, so income inequalities within the households are not reflected in the indicators. The observed gap can be primarily attributed to single-females or single-parent households headed by females (see Treanor, 2018). The fact that relatively poorer sub-populations are characterized by higher inequality has been observed several times for Italy (see Fabrizi & Trivisano, 2016b); the higher inequality experienced by females is in line with this evidence.

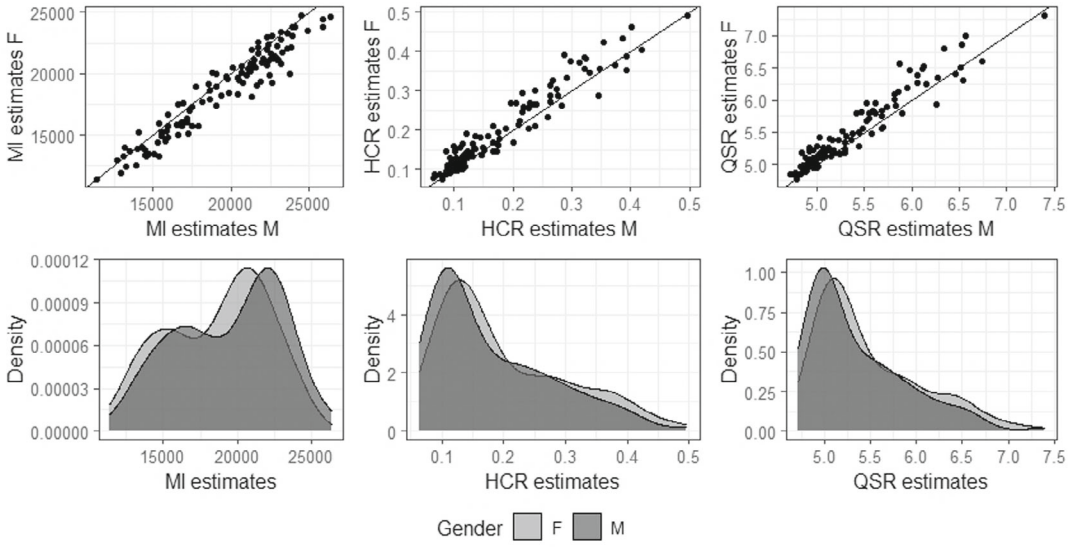

FIGURE 5 Comparison among Hierarchical Bayes estimates for Male and Female under model log-normal mixture. MI indicates the mean income $\overline{Y}_d$.

6 | SIMULATIONS

In this section, we present two simulation exercises: the first, model-based, is aimed at assessing the frequentist properties of the Bayesian predictors under the assumption that the proposed model holds; the second, design-based, explores the behaviour of the predictors under a different data generating process. In both cases, we compare the predictors we propose with the HB and EB predictors based on the simple LN model.

6.1 | Model-based simulation

In this simulation study, we consider the model in Equation (8) with $c = 0$ as a data generating process. We work with only one continuous covariate that is generated (once) from a $\mathcal{N}(1, 1)$ distribution. Slope coefficients and variance components are chosen in order to obtain a population not too different from that we analyse in the application of Section 5, both in terms of distribution and moments of the y variable. Nonetheless, for the individual-level variance parameters, we consider two different scenarios inspired by the work by Chakraborty et al. (2019) in which one is specifically aimed at the investigation of the impact of outliers on the predictors.

For each of the $B = 2000$ Monte Carlo iterations, we generate a population of size $N = 8000$, partitioned in $D = 40$ small areas with $N_d = 200$ units each. A stratified random sample without replacement is drawn from it, fixing an overall sample size $n = 300$ ($n_d = 5$ and $n_d = 10$ for 20 areas each). We set $\beta_0 = 9$, $\beta_1 = 0.22$ and $\tau^2 = 0.05$. The remaining parameters are fixed according to the following scenarios:

(a) $K = 2$, $\pi_1 = 0.9$ and $(\sigma_1^2, \sigma_2^2) = (0.1, 1)$: in this case, the two mixture components are markedly different with a ratio between the two variance components of 10 and a much smaller π_1 than the one we estimated on the data;

(b) $K = 2$, $\pi_1 = 0.7$ and $(\sigma_1^2, \sigma_2^2) = (0.1, 0.45)$; this scenario is more moderate in terms of difference between the component-specific variances and thereby closer to that of our application.
(c) $K = 1$, that is the single LN distribution, and $\sigma_1^2 = 0.25$; this scenario investigates how predictors based on a mixture model behave when actually there is no mixture.

We indicate with θ_d the d-th area-specific target quantities of the simulation study: $\theta_d \in \{\overline{Y}_d, HCR_d, QSR_d\}$, that is area mean, headcount ratio and quantile share ratio. The notation $\theta_d^{(b)}$ denotes their true value generated at iteration $b = 1, \ldots, B$. The estimation procedures we compare: direct estimates $\hat{\theta}_{d,Dir}^{(b)}$ (Equations 1–3), EB estimator under the LN BHF model (7), that is $\hat{\theta}_{d,EB-LN}^{(b)}$, HB estimator under the LN BHF model (7), that is $\hat{\theta}_{d,HB-LN}^{(b)}$ and the HB estimator under model (8) and $K = 2$: $\hat{\theta}_{d,HB-LNM}^{(b)}$. The HB estimates are obtained after the model is fitted using `Stan`, whereas the EB methods are implemented using the function `ebp` from the `emdi` package (Kreutzmann et al., 2019) for `R`.

To evaluate the frequentist properties of the point estimators, the bias and the root mean square error (RMSE) are estimated averaging Monte Carlo iteration-specific results for $\hat{\theta}_{d,Est}^{(b)}$, $Est \in \{Dir, EB{-}LN, HB{-}LN, HB{-}LNM\}$. The first is defined as $B_d = B^{-1}\sum_{b=1}^{B}(\hat{\theta}_{d,Est}^{(b)} - \theta_d^{(b)})$, and the latter $RMSE_d = \sqrt{B^{-1}\sum_{b=1}^{B}(\hat{\theta}_{d,Est}^{(b)} - \theta_d^{(b)})^2}$. We also compared the frequentist properties of the 90% credible interval produced under the two Bayesian procedures using the posterior quantiles: $[L_{d,HB-LN}^{(b)}; U_{d,HB-LN}^{(b)}]$ and $[L_{d,HB-LNM}^{(b)}; U_{d,HB-LNM}^{(b)}]$. To this aim, the frequentist coverage defined as $Cov_d = B^{-1}\sum_{b=1}^{B}\mathbf{1}\{L_{d,Est}^{(b)} < \theta_d^{(b)} < U_{d,Est}^{(b)}\}$ is computed.

In Figure 6, the results concerning the estimators of QSR_d and HCR_d under scenario (a) are reported by means of box-plots. The proposed HB-LNM estimation appears to improve the other model-based procedures from several perspectives: lower RMSEs are obtained for all the target quantities (see also Figure S2 in the supplementary material for $\overline{Y}_d$), together with a considerable bias reduction. This feature allows producing credible intervals that are near to the nominal coverage level (median: 0.88 for HCR_d, 0.85 for QSR_d and 0.89 for $\overline{Y}_d$). We notice that the more substantial improvements are registered for the indicators relying on tail features of the income distribution, whereas for $\overline{Y}_d$ the compared strategies tend to behave more similarly, in line with the findings of Section 5.

Similar results, even if less pronounced, can be detected for scenario (b), whose results are reported in Figure S3 in the supplementary material for sake of brevity. Eventually, no evident differences among the estimation procedures can be detected in scenario (c) (see Figure S4). This finding is in line with Chakraborty et al. (2019) and highlights that specifying a mixture when the data generating process consists of a single LN distribution does not lead to a significant loss of efficiency.

6.2 | Design-based simulation

In this simulation exercise, we consider a data generating process different from the model (8). Specifically, we assume data to be generated from a generalized beta distribution of the 2nd kind (GB2, McDonald, 1984). It represents a flexible distribution particularly indicated to fit income data (Graf et al., 2011), of which we adopt the notation, and recently used also in a SAE framework (Graf et al., 2019). We define the following generating model:

$$Y_{di} \sim GB2\,(a, b_0 \exp\{u_d\}, p, q),\ \ u_d \sim \mathcal{N}\left(0, \tau^2\right);\ \ d = 1, \ldots, D;\ \ i = 1, \ldots, N_d$$

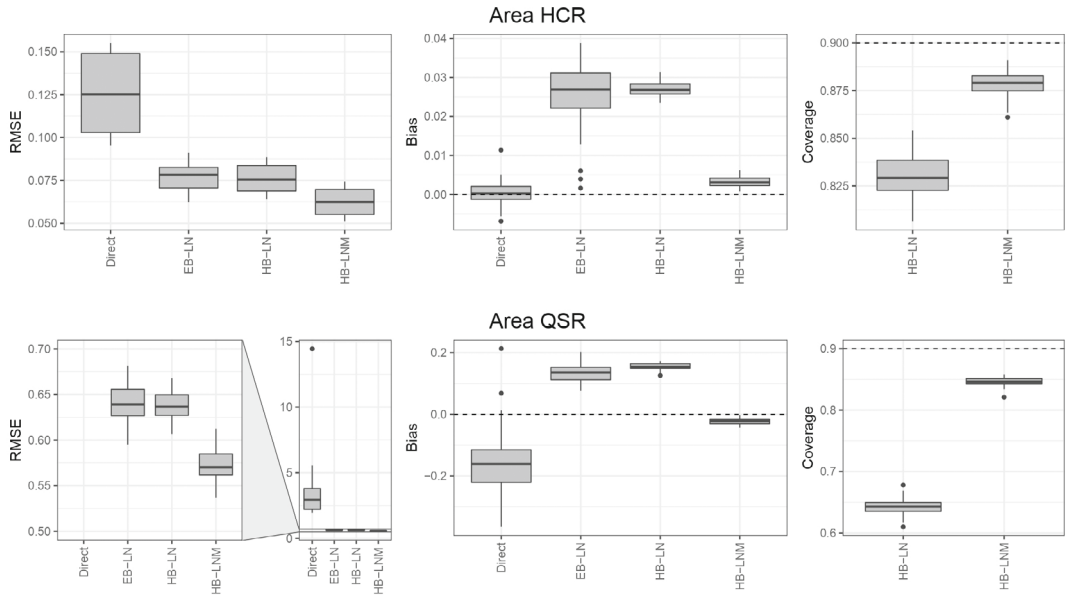

FIGURE 6 Box-plots of the distribution of the computed frequentist properties for QSR_d and HCR_d in the D areas under scenario (a).

to obtain a population of size $N = 3{,}00{,}000$ divided into $D = 30$ areas of size $N_d = 10{,}000$. To produce a population in line with our application, the GB2 parameters are set accordingly to the values fitted on Italian EU-SILC 2006 data by Graf et al. (2011): $(a, b_0, p, q) = (3.4, 17318, 0.7, 1)$, while the variance for the area-specific effects u_d is $\tau^2 = 0.03$. For each one of the $B = 2000$ Monte Carlo iterations, a stratified random sample is drawn ($n_d = 5$ and $n_d = 10$ for 15 areas each), obtaining a sample of size $n = 225$. The compared methods, target quantities and monitored frequentist properties are the same as those described for the model-based simulation study in Section 6.1.

In Figure 7, the distribution of RMSE, bias and frequentist coverage under the different estimation methods are displayed through box-plots. The observed behaviours are similar to those of scenario (a) of the model-based simulation: a gain in the RMSE and a bias reduction can be observed for the HB-LNM proposal in the estimation of HCR_d and QSR_d, whereas similar results are obtained for the area means (see Figure S5). Improvements are registered also for the interval estimates when focusing on HCR_d and QSR_d: those obtained under the HB-LNM approach are considerably nearer to the nominal coverage than the ones under HB-LN.

7 | CONCLUDING REMARKS

In this article, we discussed the production of small area estimates of poverty and inequality indicators that are functions of a positively skewed size variable (i.e. equivalized income). The model we propose, that is a finite scale mixture of LNs, is a relatively simple generalization of the LN model that is often adopted for this purpose, and it involves few additional

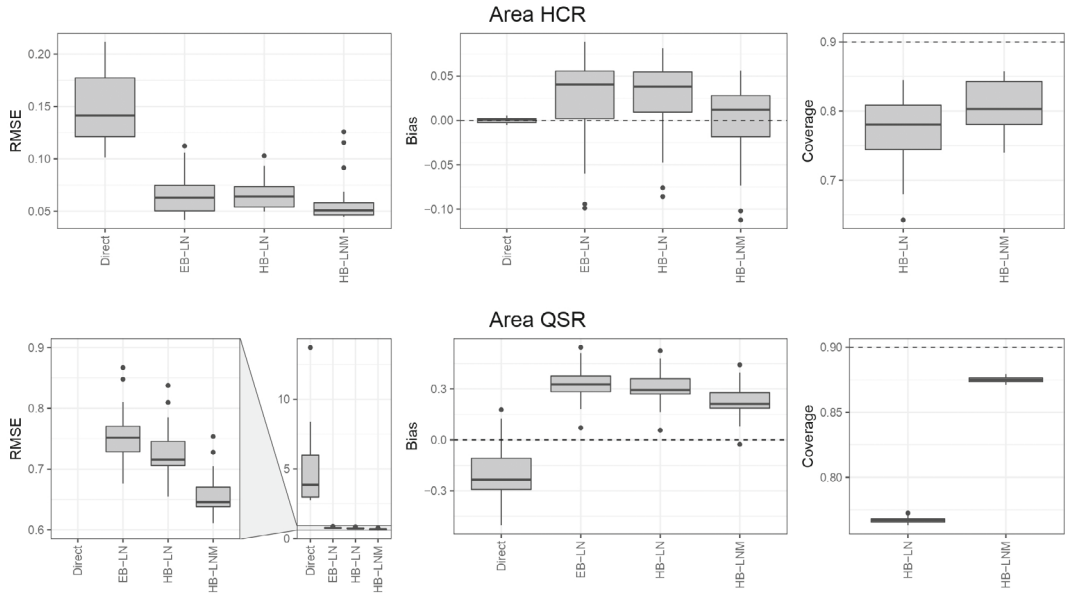

FIGURE 7 Box-plots of the distribution of the computed frequentist properties for QSR_d and HCR_d in the D areas under the design based simulation study.

parameters. It allows the straightforward estimation of non-linear functionals of the size variable measured on the individuals, including inequality measures as the quintile ratio considered in the application.

Despite the few additional parameters, the model is flexible enough to allow for a marked improvement of model fit and thereby more efficient estimation of small area parameters. The Bayesian analysis of this model requires care in specifying the prior to guarantee the existence of posterior moments. We discuss this, extending previous theoretical results available in the literature (Fabrizi & Trivisano, 2016a; Gardini et al., 2021). The considered methodology leads to improvements in the estimation of parameters that involve directly the tails of the income distribution, as the considered quintile share ratio and headcount ratio. In fact, the mixture model allows to better accommodate the irregular behaviour the income data often exhibit in the right tail of the distribution. The improvements in the estimation of the area means are less marked when moving from the single LN to the mixture, as it is a summary measure directly parameterized by the model.

In our application, we took advantage of the fact that all the auxiliary variables we considered were categorical, so posterior predictive distributions for units outside the sample did not require actual access to unit-level information. With continuous covariates this would not be the case (we consider this situation in the Simulations section). With very large populations, this can be represent not only a data availability problem but also a computational challenge. Nonetheless, we note that inference relying on MCMC methods, although computationally demanding, allows, once samples from the posterior distributions are obtained, to get uncertainty measures and posterior probability intervals easily and does not require further intensive computations.

More generally, the use of unit-level models as the basis for small area estimation relies on some delicate assumption about the quality and availability of auxiliary information (see Tarozzi

& Deaton, 2009). However, unit-level models are widely used in practice and constitute a valuable tool in small area estimation, especially when the target indicator has strongly biased direct estimators, as the considered case of quintile share ratio. In these situations, area-level models might suffer this feature leading to biased model estimates too.

To keep model specification and computation as simple as possible, we introduce some simplifications that represent possible limitations of the paper: for instance, we treat the shifting constant c as known, while it is actually estimated from the data. The impact of overlooking the uncertainty it induces on the estimation of small area parameters is likely to be very small (see Rojas-Perilla et al., 2020, in this sense) but we leave its investigation for further research.

ACKNOWLEDGEMENT

Open Access Funding provided by Universita degli Studi di Bologna within the CRUI-CARE Agreement.

REFERENCES

Atkinson, A.B. (1987) On the measurement of poverty. *Econometrica*, 55(4), 749–764.

Atkinson, A.B. & Marlier, E. (Eds.) (2010) Living conditions in Europe and the Europe 2020 agenda. In: *Income and living conditions in Europe*. Luxembourg: Publications Office of the European Union, pp. 21–35.

Atkinson, A.B., Marlier, E. & Nolan, B. (2004) Indicators and targets for social inclusion in the European union. *Journal of Common Market Studies*, 42, 47–75.

Battese, G.E., Harter, R.M. & Fuller, W.A. (1988) An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28–36.

Berg, E. & Chandra, H. (2014) Small area prediction for a unit-level lognormal model. *Computational Statistics & Data Analysis*, 78, 159–175.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M. et al. (2017) Stan: a probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.

Chakraborty, A., Datta, G.S. & Mandal, A. (2019) Robust hierarchical Bayes small area estimation for the nested error linear regression model. *International Statistical Review*, 87, S158–S176.

Dagne, G.A. (2001) Bayesian transformed models for small area estimation. *Test*, 10, 375–391.

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

Esteban, M.D., Morales, D., Pérez, A. & Santamará, L. (2012) Small area estimation of poverty proportions under area-level time models. *Computational Statistics & Data Analysis*, 56, 2840–2855.

Fabrizi, E. & Trivisano, C. (2016a) Bayesian conditional mean estimation in log-normal linear regression models with finite quadratic expected loss. *Scandinavian Journal of Statistics: Theory and Applications*, 43, 1064–1077.

Fabrizi, E. & Trivisano, C. (2016b) Small area estimation of the Gini concentration coefficient. *Computational Statistics & Data Analysis*, 99, 223–234.

Fabrizi, E., Ferrante, M.R. & Trivisano, C. (2020) A functional approach to small area estimation of the relative median poverty gap. *Journal of the Royal Statistical Society: Series A*, 183, 1273–1291.

Ferguson, T.S. (1983) Bayesian density estimation by mixtures of normal distributions. In: Rizevi, M.H., Rustagi, J.J. & Siegmund, D. (Eds.) *Recent advances in statistics*, Amsterdam: Elsevier, pp. 287–302.

Foster, J., Greer, J. & Thorbecke, E. (1984) A class of decomposable poverty measures. *Econometrica*, 52(3), 761–766.

Fusco, A., Guio, A.-C. & Marlier, E. (2010) Characterising the income poor and the materially deprived in european countries. In: Atkinson, A.B. & Marlier, E. (Eds.) *Income and living conditions in Europe*, Luxembourg: Publications Office of the European Union. pp. 133–153.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. (2019) Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A*, 182, 389–402.

Gardini, A., Trivisano, C. & Fabrizi, E. (2021) Bayesian analysis of anova and mixed models on the log-transformed response variable. *Psychometrika*, 86, 619–641.

Gelfand, A.E. (1996) Model determination using sampling-based methods. In: *Markov chain Monte Carlo in practice*, London. pp. 145–161.

Goodrich, B., Gabry, J., Ali, I. & Brilleman, S. (2020) rstanarm: Bayesian applied regression modeling via Stan. Available from: https://mc-stan.org/rstanarm R package version 2.21.1.

Graf, M., Nedyalkova, D., Münnich, R., Seger, J. & Zins, S. (2011) Parametric estimation of income distributions and indicators of poverty and social exclusion. *Report on the Simulation Results, Deliverable 7.1 of the AMELI Project*.

Graf, M., Marín, J.M. & Molina, I. (2019) A generalized mixed model for skewed distributions applied to small area estimation. *Test*, 28, 565–597.

Grusky, D.B., Kanbur, S.R. & Sen, A.K. (2006) *Poverty and inequality*. Redwood City, California: Stanford University Press.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M. & Tzavidis, N. (2019) The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91(7), 1–33.

Langel, M. & Tillé, Y. (2011) Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141, 2976–2985.

Lubrano, M. & Ndoye, A.A.J. (2016) Income inequality decomposition using a finite mixture of log-normal distributions: a Bayesian approach. *Computational Statistics & Data Analysis*, 100, 830–846.

Manandhar, B. & Nandram, B. (2019) Hierarchical Bayesian models for continuous and positively skewed data from small areas. *Communications in Statistics - Theory and Methods*, 50(4), 944–962.

Marhuenda, Y., Molina, I., Morales, D. & Rao, J. (2017) Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A*, 180, 1111–1136.

McDonald, J.B. (1984) Some generalized functions for the size distribution of income. *Econometrica*, 52, 647–665.

Molina, I. & Martin, N. (2018) Empirical best prediction under a nested error model with log transformation. *Annals of Statistics*, 46, 1961–1993.

Molina, I. & Rao, J. (2010) Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, 369–385.

Molina, I., Nandram, B. & Rao, J. (2014) Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. *Annals of Applied Statistics*, 8, 852–885.

Mussida, C. & Picchio, M. (2014) The gender wage gap by education in Italy. *The Journal of Economic Inequality*, 12, 117–147.

Natarajan, R. & Kass, R.E. (2000) Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95, 227–237.

Ntzoufras, I. (2009) *Bayesian modeling using WinBUGS*, vol. 698. Hoboken: John Wiley & Sons.

Pratesi, M. (2016) *Analysis of poverty data by small area estimation*. Hoboken: John Wiley & Sons.

Rao, J. & Molina, I. (2015) *Small area estimation*. Hoboken: John Wiley & Sons.

R Core Team. (2021) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/

Rojas-Perilla, N., Pannier, S., Schmid, T. & Tzavidis, N. (2020) Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A*, 183, 121–148.

Sen, A. & Foster, J.E. (1997) *On economic inequality*. Oxford: Oxford university press.

Statistics Canada. (2007) 2005 Survey of Financial Security - Public Use Microdata File, User Guide. Published by authority of the Minister responsible for Statistics Canada. Available from: http://www.statcan.gc.ca/pub/13f0026m/13f0026m2007001-eng.htm

Sugasawa, S. & Kubokawa, T. (2019) Adaptively transformed mixed-model prediction of general finite-population parameters. *Scandinavian Journal of Statistics: Theory and Applications*, 46, 1025–1046.

Tarozzi, A. & Deaton, A. (2009) Using census and survey data to estimate poverty and inequality for small areas. *The Review of Economics and Statistics*, 91, 773–792.

Treanor, M.C. (2018) Income poverty, material deprivation and lone parenthood. In: Nieuwenhuis, R. & Maldonado, L.C. (Eds.) *The triple bind of single-parent families: resources, employment and policies to improve wellbeing*, 1st edn. Bristol, UK: Policy Press, pp. 81–100.

Vehtari, A., Gelman, A. & Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27, 1413–1432.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T. et al. (2020) loo: efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.4.1.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Gardini, A., Fabrizi, E. & Trivisano, C. (2022) Poverty and inequality mapping based on a unit-level log-normal mixture model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(4), 2073–2096. Available from: https://doi.org/10.1111/rssa.12872