

# Multivariate Small Area Estimation of Social Indicators: The Case of Continuous and Binary Variables

Sociological Methodology

2023, Vol. 53(2) 323–343

© The Author(s) 2023



DOI: 10.1177/00811750231169726

<http://sm.sagepub.com>**Angelo Moretti**<sup>1</sup> 

## Abstract

Large-scale sample surveys are not designed to produce reliable estimates for small areas. Here, small area estimation methods can be applied to estimate population parameters of target variables to detailed geographic scales. Small area estimation for noncontinuous variables is a topic of great interest in the social sciences where such variables can be found. Generalized linear mixed models are widely adopted in the literature. Interestingly, the small area estimation literature shows that multivariate small area estimators, where correlations among outcome variables are taken into account, produce more efficient estimates than do the traditional univariate techniques. In this article, the author evaluate a multivariate small area estimator on the basis of a joint mixed model in which a small area proportion and mean of a continuous variable are estimated simultaneously. Using this method, the author “borrows strength” across response variables. The author carried out a design-based simulation study to evaluate the approach where the indicators object of study are the income and a monetary poverty (binary) indicator. The author found that the multivariate approach produces more efficient small area estimates than does the univariate modeling approach. The method can be extended to a large variety of indicators on the basis of social surveys.

## Keywords

empirical plug-in predictor, multilevel, nested-errors, poverty, income

Large-scale sample surveys are usually not designed to produce accurate and precise estimates for small population domains, for example, small geographic areas or small groups in the population (e.g., on the basis of cross-classification between ethnic groups and age bands). However, many social phenomena, such as poverty, well-being, and social exclusion show an important spatial heterogeneity at a small geographic level (Molina and Strzalkowska-Kominiak 2020; Moretti, Shlomo, and Sakshaug 2020; Pratesi 2016). Importantly, policymakers in charge of implementing subnational social policies ask for disaggregated estimates of social indicators. Thus, small area

---

<sup>1</sup>Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands

### Corresponding Author:

Dr Angelo Moretti, Utrecht University, Department of Methodology and Statistics, Sjoerd Groenmangebouw, Padualaan 14, Utrecht, 3584 CH, the Netherlands

Email: [a.moretti@uu.nl](mailto:a.moretti@uu.nl)

estimates of such phenomena should be produced and released to support decision makers (Pratesi 2016).

In recent years, there has been growing attention to the methodological development of small area estimation methods. Statistical models are now prominent in this field. Indeed, model-based estimators “borrow strength” from related areas using linking models, and they provide more efficient estimates than do direct estimators (see Pfeffermann 2013; Rao and Molina 2015).

The small area estimation approach using linear mixed models is now widely used. These models are powerful given that they include random area effects to take into account between-area variation, that is, the unexplained variability between the small areas. A pioneer work in this context is the Battese, Harter, and Fuller model (Battese, Harter, and Fuller 1988), with further work on mean squared error (MSE) estimation proposed by Prasad and Rao (1990). This model is applicable to continuous response variables, and normality is assumed for the individual error and random area effects. In case of failure of the model assumptions, extensions are proposed in the literature, for example, to account for heteroscedasticity and outliers. Readers may refer to Rao and Molina (2015) for an extensive review of methodological developments.

In the presence of binary outcomes, the generalized linear mixed model (GLMM) with logit link function, that is, a logistic mixed model, is widely adopted. Specifically, an empirical plug-in predictor (EPP) under a GLMM is used in small area estimation of proportions in official statistics (Chandra, Chambers, and Salvati 2012; Chandra, Kumar, and Aditya 2018; Molina and Strzalkowska-Kominiak 2020; Rao and Molina 2015; Salvati, Chandra, and Chambers 2012). For example, the U.K. Office for National Statistics and the Australian Bureau of Statistics use this method (Chambers, Salvati, and Tzavidis 2016; Chandra et al. 2018). Binary variables can be found in most surveys. For instance, some poverty and well-being indicators are based on binary variables (Betti and Lemmi 2013). Additionally, some social indicators estimated on Labour Force Surveys are also constructed on these types of variables (see Chambers et al. 2016; Molina and Strzalkowska-Kominiak 2020). Hence, there is a high demand for small area proportions.

Because many social phenomena are naturally multidimensional (Betti and Lemmi 2013), and therefore correlated, this property can be used to further improve small area estimates (Benavent and Morales 2016; Fabrizi, Ferrante, and Pacci 2005; Moretti et al. 2020; Moretti, Shlomo, and Sakshaug 2021; Ubaidillah et al. 2019). In this context, multivariate small area estimation methods can be applied. Moretti et al. (2020) proposed the use of multivariate small area estimation methods to estimate latent well-being indicators, under the use of the multivariate extension of the Battese, Harter, and Fuller model. This model was studied in detail by Datta, Day, and Basawa (1999). Recently, Moretti (2022) extended a GLMM with logit link function to the case of bivariate proportions, showing good results in terms of efficiency compared with its univariate setting.

In this article, I investigate the multivariate small area estimation problem of continuous and binary response variables jointly. In particular, I apply a joint mixed-modeling strategy to the small area estimation of the mean of a continuous variable

and proportion of a binary variable. This is an important problem in sociology and related social sciences. In fact, when building poverty indicators, they can be based on income and measured by a binary indicator taking a value of 0 or 1, denoting whether the personal or household income falls below a poverty threshold (for a review and discussion of poverty indicators, see Pratesi 2016). Therefore, income and poverty indicators are correlated; this can be accounted for in the small area models by assuming a joint modeling strategy.

## NOTATION AND MULTIVARIATE SMALL AREA ESTIMATION PROBLEM

I now present the notation used in the article and the small area estimation problem we are studying. Let us consider a finite target population  $U$  with size  $N$  that is partitioned into  $D$  nonoverlapping (disjoint) small areas,  $U_d$ ,  $d=1, \dots, D$  of size  $N_d$  such that  $\cup_{d=1}^D U_d = U$  and  $\sum_{d=1}^D N_d = N$ . A random sample  $s$  of size  $n$  is selected from  $U$ .  $n_d$  denotes the sample size in small area  $d$ , and  $\sum_{d=1}^D n_d = n$ .

Let  $\mathbf{y}_{di} = (y_{di1}, y_{di2})^T$  denote a vector of the values of  $k=1, 2$  variables of interest  $\mathbf{Y}$  for unit  $i$  in area  $d$ . We assume  $y_{di1}$  is continuous and normally distributed, whereas  $y_{di2}$  takes value 0 or 1 only. We are interested in estimating the vector of means of  $\mathbf{Y}$  denoted by  $\bar{\mathbf{Y}}_d = (\bar{Y}_{d1}, \bar{Y}_{d2})$ .  $\bar{Y}_{d1}$  is the mean of the continuous variable  $Y_1$ , and  $\bar{Y}_{d2}$  is the proportion related to variable  $Y_2$ , which is also a mean.

The generic element related to variable  $k$  is given by equation (1):

$$\bar{Y}_{dk} = N_d^{-1} \sum_{i \in U_d} y_{dik} = N_d^{-1} \left( \sum_{i \in s_d} y_{dik} + \sum_{i \in r_d} y_{dik} \right), \quad (1)$$

where  $s_d$  denotes the sample elements and  $r_d$  the out-of-sample elements in area  $d$ . Looking at equation (1), we see that the population mean can be split into sample ( $s_d$ ) and out-of-sample elements ( $r_d$ ) in area  $d$ .

The design-based direct estimator for the  $k$ th small area mean  $\bar{Y}_{dk}$  is given by

$$\hat{Y}_{dk}^{\text{DIR}} = \frac{\sum_{i \in s_d} w_{di} y_{dik}}{\sum_{i \in s_d} w_{di}}, \quad (2)$$

where  $w_{di}$  denotes the survey weight for unit  $i$  in area  $d$ .

It is well known that estimator (2) becomes unstable when  $n_d$  is small, because it is based on area-specific sample quantities only (see Rao and Molina 2015; Särndal, Swensson, and Wretman 2003). Additionally, the estimator cannot be computed for small areas with  $n_d=0$ . Therefore, model-based small area estimation methods that “borrow strength” across small areas via statistical models are adopted to produce reliable, that is, accurate and precise, small area estimates of the target parameter given by equation (1) (Rao and Molina 2015). For the estimator of the variance of equation (2), see Rao and Molina (2015).

*Small Area Estimation under a Joint Mixed Model for Continuous and Binary Outcome Variables*

To build indirect small area estimators, statistical models with random area-specific effects that account for between-area variability are used in small area estimation (Rao and Molina 2015). In the univariate small area estimation setting, the Battese, Harter, and Fuller model (Battese et al. 1988) can be used to obtain small area estimates of the population mean in case of continuous variables, that is,  $\bar{Y}_{d1}$ . In the case of binary responses, the GLMM with logit link function is often used to estimate small area proportions,  $\bar{Y}_{d2}$  (Chambers et al. 2016; Chandra et al. 2018). Prior literature shows that use of multivariate mixed models in small area estimation provides more efficient estimates than does the univariate case where separate models are estimated on each response variable (Datta et al. 1999; Moretti et al. 2020; Rao and Molina 2015).

Here, we assume that  $\mathbf{y}_{di}$  follows a joint mixed effects model. This is obtained by assuming a multivariate distribution for the random effects of responses  $k=1$  and  $k=2$  (Ivanova, Molenberghs, and Verbeke 2016). The model is presented below, for  $i=1, \dots, N_d$  and  $d=1, \dots, D$ :

$$\begin{cases} y_{di1} = \mathbf{x}_{di1}^T \boldsymbol{\beta}_1 + e_{di1} + u_{d1} \\ \text{logit}[\pi_{di2}] = \log\left[\frac{\pi_{di2}}{1-\pi_{di2}}\right] = \eta_{di2} = \mathbf{x}_{di2}^T \boldsymbol{\beta}_2 + u_{d2} \end{cases} \tag{3}$$

where  $\mathbf{x}_{di1}$  and  $\mathbf{x}_{di2}$  denote the vectors of observed values of  $p$  unit-level auxiliary variables for unit  $i$  in area  $d$  related to variables  $k=1$  and  $k=2$ . These can be the same for both responses, depending on data availability and specific modeling problems (Fuller and Harter 1987; Ivanova et al. 2016; Moretti et al. 2020).

Model 3 is a multilevel model for multivariate mixed response types, explaining two outcomes simultaneously (Goldstein 2011), in this case  $Y_1$  and  $Y_2$ . Goldstein et al. (2009) studied a similar modeling approach. Because we are aiming to predict the small area means of the  $Y_1$  and  $Y_2$  variables, we cannot account for the correlation structure between the two by including one of them as a predictor in a univariate model instead of a joint modeling approach. In fact, in the model-based small area estimation approach, we need auxiliary information of the predictors for all the units in the population to build the EPP. These are not available from census or administrative data for  $Y_1$  and  $Y_2$  (e.g., the income variable is not available as an auxiliary variable from external data sources).

We assume multivariate (bivariate in this case) normality of the random area effects, that is,  $\mathbf{u}_d = (u_{d1}, u_{d2})^T \sim N(0, \Sigma_u)$ , where  $\Sigma_u$  denotes a  $2 \times 2$  positive-definite variance-covariance matrix. The off-diagonal elements are the covariances between  $u_{d1}$  and  $u_{d2}$ . The matrix is given as follows:

$$\Sigma_u = \begin{bmatrix} \sigma_{u1}^2 & \rho_u \cdot \sqrt{\sigma_{u1}^2 \cdot \sigma_{u2}^2} \\ \rho_u \cdot \sqrt{\sigma_{u2}^2 \cdot \sigma_{u1}^2} & \sigma_{u2}^2 \end{bmatrix},$$

where  $\rho_u$  denotes the correlation coefficient. In addition, we assume that  $y_{di2} | u_{d2} \sim \text{Binomial}(1, \pi_{di2})$  with  $\pi_{di2} = E(y_{di2} | u_{d2})$ .

Model 3 can be written for  $i = 1, \dots, n_d$  without loss of generality. Therefore, the model parameters are estimated on a random sample  $s$  drawn from the population  $U$  (Rao and Molina 2015). Here, I estimate model parameters via the maximum likelihood (ML) approach. For estimation details that are beyond the scope of the present article, readers can refer to McCulloch (1994, 1997) and Booth and Hobert (1999) for the theory, and Berridge and Crouchley (2011) for its implementation.

Because of the multivariate nature of the likelihood function, a Gaussian quadrature has been adopted in the literature. ML estimation of the parameters is computationally complex for this model (Berridge and Crouchley 2011; Coull and Agresti 2000). Readers interested in the computations can refer to Appendix A.3 in Berridge and Crouchley (2011). The implementation is fully available in R via the package *sabre*.<sup>1</sup> This approach is also studied and evaluated in Rabe-Hesketh and Skrongdal (2008) and Skrongdal and Rabe-Hesketh (2004).

To provide a small area estimator for  $\bar{Y}_d$ , we consider an EPP. Empirical best predictors in the case of GLMMs are difficult to obtain analytically, and they are not available in a closed form. EPPs are easier to obtain and widely used by statistical agencies (e.g., the U.K. Office for National Statistics and the Australian Bureau of Statistics) (Chambers et al. 2016; Chandra et al. 2018). This problem becomes even more challenging in the case of joint modeling of mixed types responses.

Therefore, we can write the EPP of the small area means under model 3 for area  $d$  as follows:

$$\begin{cases} \hat{Y}_{d1}^{EPP} = N_d^{-1} \left( \sum_{i \in s_d} y_{di1} + \sum_{i \in r_d} \mathbf{x}_{di1}^T \hat{\boldsymbol{\beta}}_1 + \hat{u}_{d1} \right) \\ \hat{Y}_{d2}^{EPP} = N_d^{-1} \left( \sum_{i \in s_d} y_{di2} + \sum_{i \in r_d} \hat{\mu}_{di2} \right) \end{cases}, \tag{4}$$

with  $\hat{\mu}_{di2} = \hat{E}(y_{di2} | u_{d2}) = \exp(\mathbf{x}_{di2}^T \hat{\boldsymbol{\beta}}_2 + \hat{u}_{d2}) [1 + \exp(\mathbf{x}_{di2}^T \hat{\boldsymbol{\beta}}_2 + \hat{u}_{d2})]^{-1}$ , where  $\hat{\boldsymbol{\beta}}_1$ ,  $\hat{\boldsymbol{\beta}}_2$ ,  $\hat{u}_{d1}$ , and  $\hat{u}_{d2}$  denote the estimates of the regression coefficients and predictions of random effects, respectively.

In practical applications, the auxiliary variables are available at the unit level for the sample units only, and area-specific aggregates are available for the population from the census or administrative data. As a consequence, equation (4) cannot be applied; however, a modification of it can be derived, and its performance is studied in the literature (Chandra et al. 2018). This is given as follows:

$$\begin{cases} \hat{Y}_{d1}^{EPP1} = f_d \hat{Y}_{d1} + (1 - f_d) (\bar{\mathbf{X}}_{r,d1}^T \hat{\boldsymbol{\beta}}_1 + \hat{u}_{d1}) \\ \hat{Y}_{d2}^{EPP1} = f_d \hat{Y}_{d2} + (1 - f_d) \exp(\bar{\mathbf{X}}_{r,d2}^T \hat{\boldsymbol{\beta}}_2 + \hat{u}_{d2}) [1 + \exp(\bar{\mathbf{X}}_{r,d2}^T \hat{\boldsymbol{\beta}}_2 + \hat{u}_{d2})]^{-1}, \end{cases} \tag{5}$$

where  $f_d = n_d / N_d$  is the sampling fraction in area  $d$  and  $\hat{Y}_{dk} = \sum_{i \in s_d} y_{dik} / n_d$ .  $\bar{\mathbf{X}}_{r,dk} = (N_d - n_d)^{-1} (N_d \bar{\mathbf{X}}_{dk} - n_d \bar{\mathbf{x}}_{dk})$  denotes the means of the auxiliary variables for the out-of-sample units, with  $\bar{\mathbf{X}}_{dk} = N_d^{-1} \sum_{i \in U_d} \bar{\mathbf{x}}_{dik}$  and  $\bar{\mathbf{x}}_{dk} = n_d^{-1} \sum_{i \in s_d} \mathbf{x}_{dik}$ , denoting the means of the auxiliary variables in the population and sample, respectively, for area  $d$  and  $k = 1, 2$ .

When the sampling fractions  $f_d$  are small (negligible) and  $\bar{\mathbf{X}}_{dk} \approx \bar{\mathbf{X}}_{r,dk}$ , the EPPs given in equation (5) can be written as follows (Chandra et al. 2018; Moretti 2022):

$$\begin{cases} \hat{Y}_{d1}^{EPP1.1} = f_d \hat{Y}_{d1} + (1 - f_d) (\bar{\mathbf{X}}_{d1}^T \hat{\boldsymbol{\beta}}_1 + \hat{u}_{d1}) \\ \hat{Y}_{d2}^{EPP1.1} = f_d \hat{Y}_{d2} + (1 - f_d) \exp(\bar{\mathbf{X}}_{d2}^T \hat{\boldsymbol{\beta}}_2 + \hat{u}_{d2}) [1 + \exp(\bar{\mathbf{X}}_{d2}^T \hat{\boldsymbol{\beta}}_2 + \hat{u}_{d2})]^{-1}. \end{cases} \tag{6}$$

This is reasonable to assume in real data small area applications, given that  $f_d$  may be very small. Hence, in this article, I consider and evaluate EPP1.1 only.

*MSE Estimation via Parametric Bootstrap*

The MSE of equation (6) can be estimated via a parametric bootstrap algorithm. The parametric bootstrap is widely studied and applied in the small area estimation literature, where simulation studies are designed to evaluate the algorithm’s performance (González-Manteiga et al. 2007; Hobza, Morales, and Santamaría 2018; Moretti, Shlomo, and Sakshaug 2020). Here, I extend the algorithm originally developed in González-Manteiga et al. (2007) to the case considered in this article. Note that Moretti et al. (2020) extended and evaluated the algorithm in González-Manteiga et al. (2007) to the bivariate small area estimation of means problem in the case of continuous variables.

The steps of the parametric bootstrap algorithm are as follows:

1. Estimate model given in equation (3) on the random sample  $s$ . The following estimates are obtained:  $\hat{\boldsymbol{\Sigma}}_u$ ,  $\hat{\boldsymbol{\beta}}_1$ ,  $\hat{\boldsymbol{\beta}}_2$ , and  $\hat{\sigma}_e^2$ .
2. Generate the bootstrap area-specific random effects as follows:  $\mathbf{u}_d^{*(b)} \sim N(0, \hat{\boldsymbol{\Sigma}}_u)$ , and individual error term for response  $k=1$  only, that is,  $e_{di1}^{*(b)} \sim N(0, \hat{\sigma}_e^2)$ . The asterisk denotes the bootstrap quantities, and  $b$  denotes the  $b$ th bootstrap replication,  $b = 1, \dots, B$ .
3. Calculate the true means of the bootstrap population for variables  $k = 1, 2$  and area  $d$ :

$$\begin{cases} \bar{Y}_{d1}^{*(b)} = \bar{\mathbf{X}}_{d1} \hat{\boldsymbol{\beta}}_1 + u_{d1}^{*(b)} \\ \bar{Y}_{d2}^{*(b)} = \exp(\bar{\mathbf{X}}_{d2} \hat{\boldsymbol{\beta}}_2 + u_{d2}^{*(b)}) [1 + \exp(\bar{\mathbf{X}}_{d2} \hat{\boldsymbol{\beta}}_2 + u_{d2}^{*(b)})]^{-1}. \end{cases} \tag{7}$$

4. Generate the bootstrap responses  $y_{di1}^{*(b)}$  and  $y_{di2}^{*(b)}$  according to model 3 for  $i \in s_d$  as follows:

$$\begin{cases} y_{di1}^{*(b)} = \bar{\mathbf{x}}_{di1}^T \hat{\boldsymbol{\beta}}_1 + e_{di1}^{*(b)} + u_{d1}^{*(b)} \\ y_{di2}^{*(b)} | u_{d2}^{*(b)} \sim \text{Binomial}(1, \pi_{di2}^{*(b)}) \end{cases}, \tag{8}$$

where  $\pi_{di2}^{*(b)} = \exp(\bar{\mathbf{x}}_{di2}^T \hat{\boldsymbol{\beta}}_2 + u_{d2}^{*(b)}) [1 + \exp(\bar{\mathbf{x}}_{di2}^T \hat{\boldsymbol{\beta}}_2 + u_{d2}^{*(b)})]^{-1}$ .

5. Estimate the joint mixed model given in equation (3) on the responses generated at step 4 and obtain the bootstrap EPP1.1 according to equation (6). These are denoted by  $\hat{Y}_{d1}^{EPP1.1*(b)}$  and  $\hat{Y}_{d2}^{EPP1.1*(b)}$ .
6. Repeat steps 2 to 5  $B$  times.
7. The bootstrap estimators for the MSE of  $\hat{Y}_{d1}^{EPP1.1*(b)}$  and  $\hat{Y}_{d2}^{EPP1.1*(b)}$  are given by

$$\begin{cases} \widehat{\text{MSE}}_{boot}(\hat{Y}_{d1}^{\text{EPP1.1}*(b)}) = B^{-1} \sum_{b=1}^B \left( \hat{Y}_{d1}^{\text{EPP1.1}*(b)} - \bar{Y}_{d1}^{*(b)} \right)^2 \\ \widehat{\text{MSE}}_{boot}(\hat{Y}_{d2}^{\text{EPP1.1}*(b)}) = B^{-1} \sum_{b=1}^B \left( \hat{Y}_{d2}^{\text{EPP1.1}*(b)} - \bar{Y}_{d2}^{*(b)} \right)^2, \end{cases} \quad (9)$$

for  $d = 1, \dots, D$ , with  $B$  denoting the number of replicates, that is,  $B = 500$  (Hobza et al. 2018).

## SIMULATION STUDY

In this section, I present the results of a design-based simulation study. Design-based simulation studies are important because they allow one to evaluate the performance of the estimators in case of repeated samples drawn from a fixed population that does not exactly follow the assumed model (Molina and Strzalkowska-Kominiak 2020).

### Generating the Population

As a fixed target population  $U$ , I use data from Lehtonen and Veijanen (2016). This is available online from Pratesi (2016). These data were derived from AMELIA data (see Burgard et al. 2017; Lehtonen and Veijanen 2016). AMELIA is a synthetic population that allows comparative and reproducible research on the basis of European Union Statistics for Income and Living Conditions variables (Burgard et al. 2017).

The population size in area  $d$  ranges between 12,550 and 42,340, and the areas are  $D = 40$ . As variables of interest, I consider the logarithmic transformation of the income,  $Y_1$ , and a binary variable taking value 1 if the unit  $i$  in area  $d$  is poor and 0 otherwise,  $Y_2$ . The unit  $i$  is poor if the value of the income is below the poverty line, calculated as 60 percent of the median of the income (Chatterjee 2011). The auxiliary variable,  $X_1$ , is the age of unit  $i$ . Because of the nature of the response variables, the random effects are highly correlated with  $\hat{\rho} = -0.80$ .  $\hat{\rho}$  is obtained via a joint mixed model estimated on the fixed population  $U$ .

Figure 1 shows the histogram of the logarithmic transformation of income,  $Y_1$ , and Figure 2 shows the bar chart of the binary variable,  $Y_2$ , denoting whether the unit is poor or not.  $Y_1$  is approximately normally distributed, but slightly skewed and with some outliers; hence, it mimics real data distributions.

### Simulation Steps and Quality Measures

The simulation consists of the following steps, where  $l = 1, \dots, L$ ;  $L = 500$  denotes the repetition:

1. From the target population  $U = \cup_{d=1}^D U_d$  where  $U_d = \{(y_{di1}, y_{di2}, x_{di}), i = 1, \dots, N_d\}$ , select a sample  $s_d^{(l)}$  without replacement of size  $n_d = 3$  for  $d = 1, \dots, D$ . Note that  $U$  is fixed over the simulations because we are following a design-based simulation setting.
2. Estimate the joint mixed model given in equation (3) and the univariate mixed models on each separate response in each sample  $s_d^{(l)}$  and obtain the univariate and multivariate

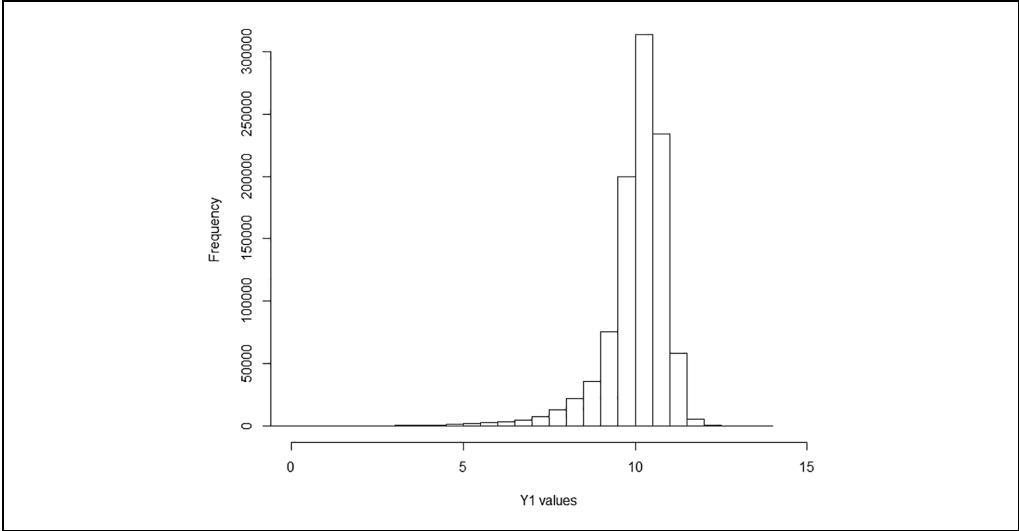


Figure 1. Histogram of log income,  $Y_1$ .

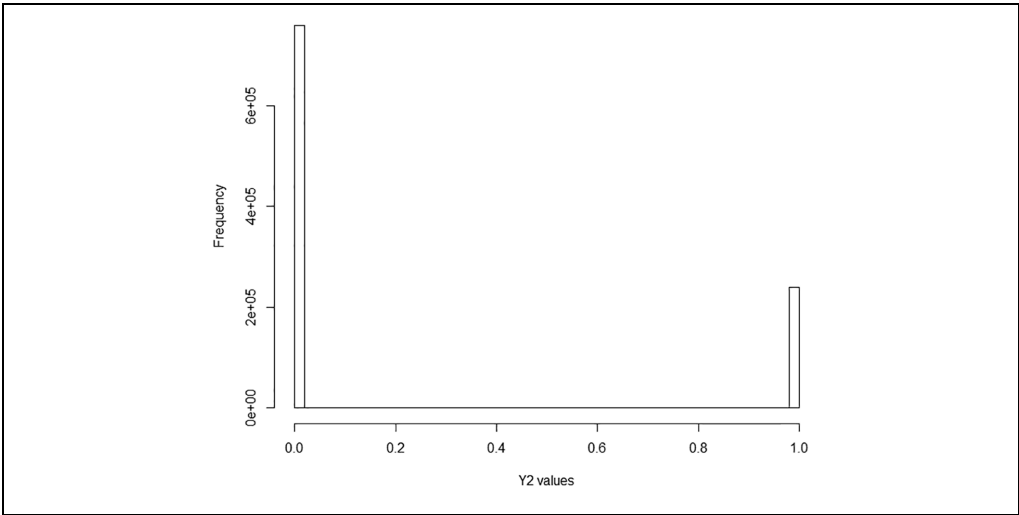


Figure 2. Bar chart of indicator denoting whether the unit's income is below the poverty line,  $Y_2$ .

predictors for both small area means. Readers can refer to the Battese, Harter, and Fuller model (Battese et al. 1988) for the univariate setting and to Chandra et al. (2018) for the small area estimation problem of proportions. The multivariate small area predictors are given by equation (6). The predictors are denoted by  $\hat{y}_{d1}^{MEPP1.1(l)}$ ,  $\hat{y}_{d2}^{MEPP1.1(l)}$  and  $\hat{y}_{d1}^{UEPP1.1(l)}$ ,  $\hat{y}_{d2}^{UEPP1.1(l)}$ , for the multivariate and univariate case, respectively, and  $l$ th repetition. The direct estimates are also calculated via equation (2) and denoted by  $\hat{y}_{d1}^{DIR(l)}$ ,  $\hat{y}_{d2}^{DIR(l)}$ .



3. The following measures of performance are calculated to evaluate the estimators for  $k = 1, 2$  in both the univariate and multivariate case (here,  $\hat{y}_{dk}^{(l)}$  denotes any estimator for the true value in the population, i.e.,  $\bar{y}_{dk}$ , for  $k$ th variable and  $d$ th area):

Absolute relative bias (ARB)

$$\text{ARB}(\hat{y}_{dk}) = \left| \frac{L^{-1} \sum_{l=1}^L (\hat{y}_{dk}^{(l)} - \bar{y}_{dk})}{\bar{y}_{dk}} \right|, \quad (10)$$

Root MSE (RMSE)

$$\text{RMSE}(\hat{y}_{dk}) = \sqrt{L^{-1} \sum_{l=1}^L (\hat{y}_{dk}^{(l)} - \bar{y}_{dk})^2}, \quad (11)$$

Relative root MSE (RRMSE)

$$\text{RRMSE}(\hat{y}_{dk}) = \frac{\text{RMSE}(\hat{y}_{dk})}{\bar{y}_{dk}}, \quad (12)$$

where  $\bar{y}_{dk}$  denotes the true values for  $k = 1, 2$ .

% Relative reduction in terms of RMSE (RelRed%)

$$\text{RelRed}(\hat{y}_{dk})\% = L^{-1} \sum_{l=1}^L \frac{\text{RMSE}(\hat{y}_{dk}^{\text{MEPP1.1}(l)}) - \text{RMSE}(\hat{y}_{dk}^{\text{UEPP1.1}(l)})}{\text{RMSE}(\hat{y}_{dk}^{\text{UEPP1.1}(l)})} \times 100. \quad (13)$$

To present summary statistics, the median across the small areas  $D$  is shown as a robust central tendency measure that avoids the effect of extreme values in some small areas (Chambers, Chandra, and Tzavidis 2011; Giusti et al. 2014). In this case, the same notation as above is used but the index  $d$  is dropped.

## Results

I now present the results of the simulation study with a focus on the quality measures presented above. Table 1 shows the median across the small areas of the ARB and RRMSE of the estimates obtained via the direct, univariate, and multivariate estimators for  $k = 1$  and  $k = 2$  means.

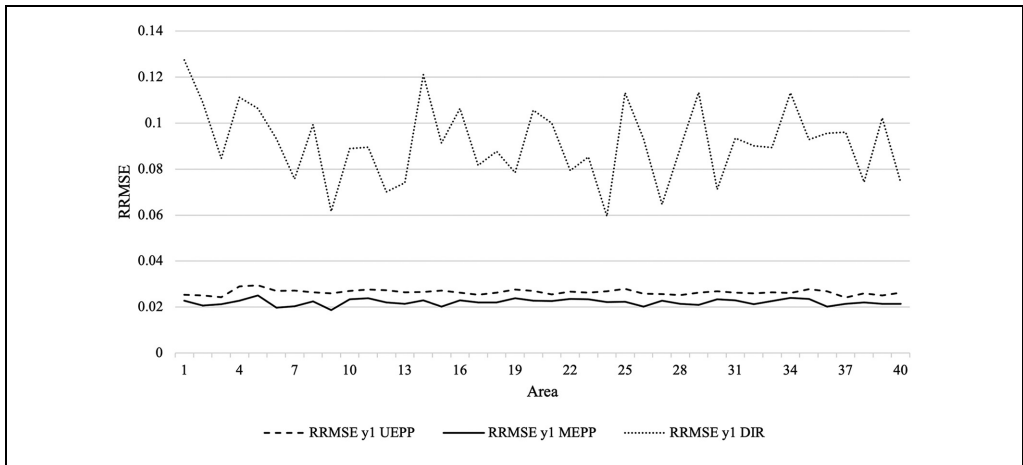
As shown, the ARB is small (negligible) across all the estimators. The multivariate approach produces estimates with a smaller ARB compared with the univariate case. Thus, it does not introduce much bias in the small area estimates. Furthermore, the multivariate small area predictor provides estimates with a smaller RRMSE compared with the univariate and direct estimators. Hence, the multivariate approach returns more efficient estimates compared with the univariate approach. This latter point can be seen if we consider the percentage relative reductions in terms of RMSE. Indeed, these are satisfactory and equal to  $\text{RelRed}(\hat{y}_{d1})\% = -35\%$  and  $\text{RelRed}(\hat{y}_{d2})\% = -10\%$ .

Figures 3 and 4 plot the RRMSE across the small areas where we compare the RRMSE of the multivariate versus univariate approach and their direct estimators

**Table 1.** Median Values across Small Areas of ARB and RRMSE of the Estimates Obtained via the Direct, Univariate, and Multivariate Estimators for  $k = 1$  and  $k = 2$  Means

Estimator	ARB	RRMSE
$\hat{y}_1^{DIR}$	.010	.091
$\hat{y}_2^{DIR}$	.030	.842
$\hat{y}_1^{UEPP}$	.018	.030
$\hat{y}_2^{UEPP}$	.100	.164
$\hat{y}_1^{MEPP}$	.014	.020
$\hat{y}_2^{MEPP}$	.077	.150

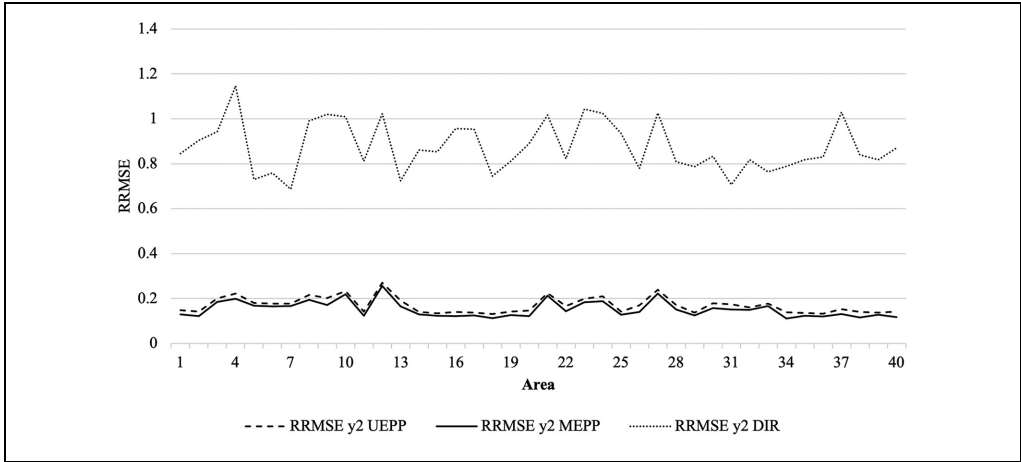
Note: ARB = absolute relative bias; RRMSE = relative root mean squared error.



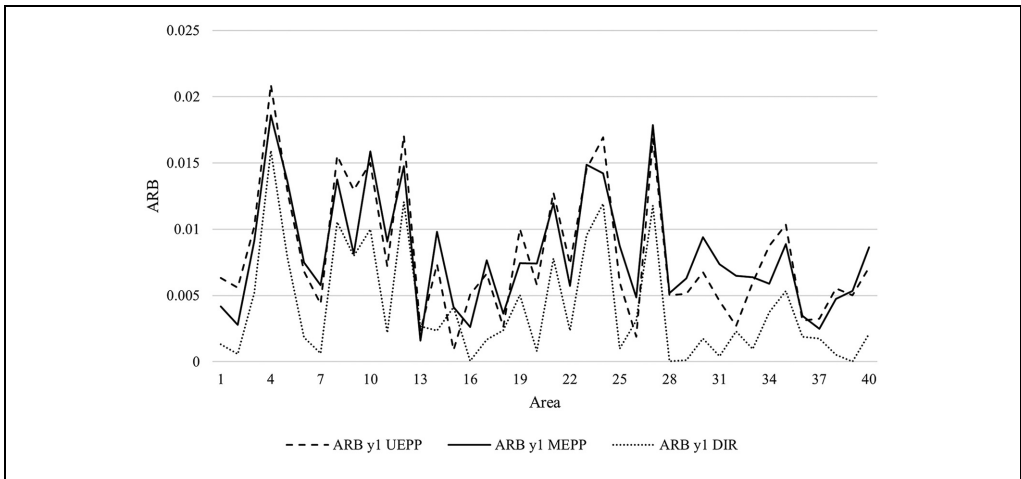
**Figure 3.** Relative root mean squared error (RRMSE) of small area estimates for direct (DIR; dotted line), univariate (UEPP; dashed line), and multivariate (MEPP; solid line) estimators for the mean of log income,  $k = 1$ , ordered by decreasing sampling fraction.

setting for  $k = 1$  and  $k = 2$ , respectively. Figures 5 and 6 show the plots of the ARB across the small areas for the three approaches for  $k = 1$  and  $k = 2$ , respectively. Descriptive statistics related to these figures are shown in Table 1.

As expected, Figures 3 to 6 demonstrate that the direct estimates suffer from a large RRMSE but very small ARB across all the small areas. The multivariate approach provides estimates with a smaller RRMSE than both the univariate and direct approaches, which is in line with the multivariate small area estimation literature. In addition, the ARB of the multivariate estimates is negligible across all the areas, showing it does not introduce a large bias in the estimates. I did not find any relationship between the sampling fraction and performances of the multivariate approach over its univariate setting.



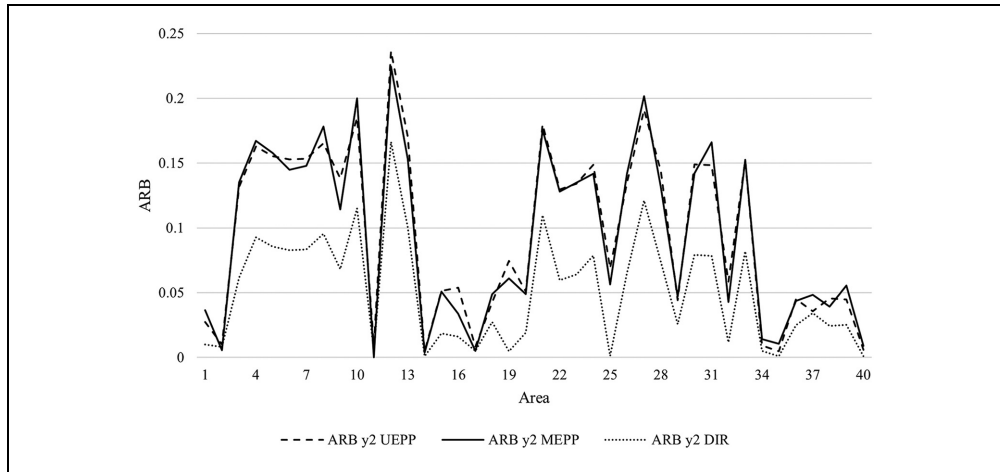
**Figure 4.** Relative root mean squared error (RRMSE) of small area estimates for direct (DIR; dotted line), univariate (UEPP; dashed line), and multivariate (MEPP; continuous line) estimators for the poverty proportion,  $k = 2$ , ordered by decreasing sampling fraction.



**Figure 5.** Absolute relative bias (ARB) of small area estimates for direct (DIR; dotted line), univariate (UEPP; dashed line), and multivariate (MEPP; solid line) estimators for the mean of log income,  $k = 1$ , ordered by decreasing sampling fraction.

Figures A1 and A2 and Table A1 in Appendix A provides additional simulation outputs that are helpful for interpreting the results.

In summary, the results of this simulation study are promising. Although the income variable,  $Y_1$ , does not follow a normal distribution perfectly (see Figure 1), as it occurs with real data, the multivariate approach provides estimates with a smaller RRMSE than the univariate case, and the ARB is negligible across the small areas. Potential applications of this approach to social indicators are discussed in the next section.



**Figure 6.** Absolute relative bias (ARB) of small area estimates for direct (DIR; dotted line), univariate (UEPP; dashed line), and multivariate (MEPP; solid line) estimators for the poverty proportion,  $k = 2$ , ordered by decreasing sampling fraction.

## CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this article, I studied the multivariate small area estimation problem in the case of mixed-type variables. The multivariate small area estimation literature has paid particular attention to the case of continuous variables. However, noncontinuous variables are widely diffused in social surveys. In this work, I considered the unit-level small area estimation approach, assuming the auxiliary variables are known for all the sampled units. I did not investigate the area-level approach in the present article. For the area-level approach, the reader may refer to Fay (1987), and more recent studies by Benavent and Morales (2016, 2020).

I compared a multivariate EPP with its univariate setting, where two mixed models are estimated separately on each response variable in a design-based simulation study. The predictors based on the univariate setting are used by different statistical agencies due to their good properties. I found the multivariate approach provides more reliable small area estimates than does the univariate approach. In particular, the multivariate predictor produces small area estimates with a smaller RRMSE, as well as bias slightly smaller than the univariate approach. Larger gains in efficiency can be seen for the continuous case.

The modeling strategy evaluated in this article is flexible, because joint mixed models can account for variables measured on different scales at the same time. Thus, we can improve the efficiency of the small area estimates by “borrowing strength” across response variables in a model-based, small area estimation approach.

These types of variables are widely present in social surveys, where continuous variables are strongly related to binary variables. Here, I considered income and poverty indicators; however, future work will investigate other modeling scenarios, that is, variables measured on other scales, such as models for count data. Indeed, the use of multivariate mixed (multilevel) models is of particular relevance in sociology, where

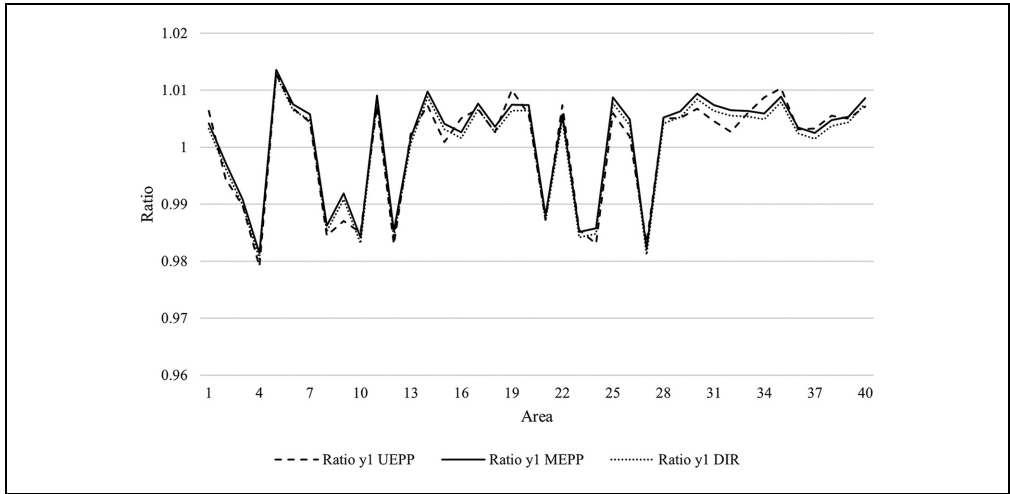
multivariate research questions can be formulated and answered by interpreting model parameter estimates, in particular the association parameters. These association parameters, such as correlations, can be estimated via the use of multivariate models. For example, the literature has shown a strong relationship between happiness and health (see Argyle 1997; Graham 2012). If one considers a continuous latent variable measuring happiness, and a binary variable measuring a health domain, the model object of this study can be applied to investigate relationships in the case of multilevel data. Pierewan and Tampubolon (2015) used a multivariate mixed model to investigate the relationships between happiness and health. The model I proposed in this small area estimation problem can be used to investigate subnational differences of such indicators, and to produce more efficient estimates than in the scenario where two separate multilevel models are estimated.

In the context of income and expenditures analysis, researchers might be interested in studying relationships between deprivation and expenditures; these show correlations (see Moretti and Shlomo 2023) that can be taken into account in the modeling stage. Income is also related to food poverty. Thus, a binary variable indicating whether a household is food deprived or at risk for food deprivation can be built, and this is strongly related to household income (for the link between poverty and food insecurity, see Wight et al. 2014). Therefore, similar to the approach followed in the simulation study, the method proposed here is appropriate to provide geographic understanding around food deprivation and monetary poverty dimensions.

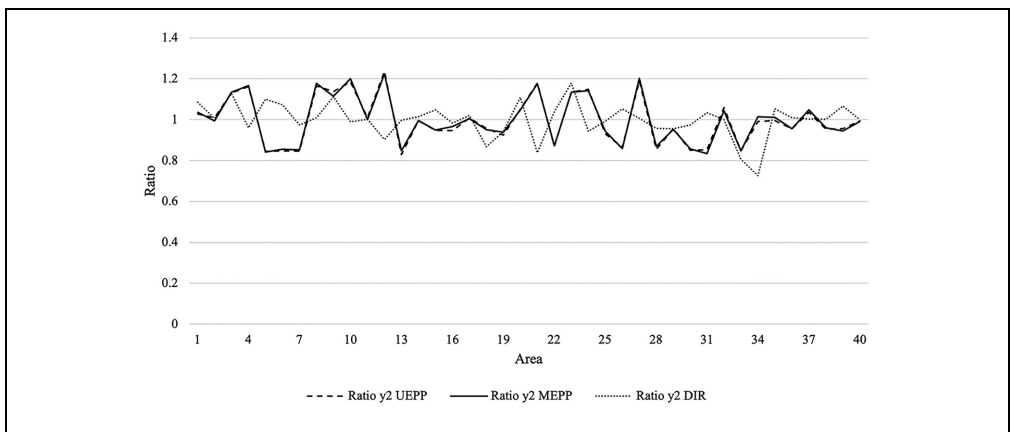
It is important to acknowledge that the mixed models adopted in this article can be extended to a spatial dependence setting (Chandra, Chambers, and Salvati 2019; Chandra et al. 2018). Here, one strategy to incorporate spatial information into the models is to extend the random effects model allowing for spatially correlated random area effects. For instance, this can be achieved by using a simultaneous autoregressive model (Anselin 1988; Cressie 2015). As highlighted by Chandra et al. (2018), in practical applications, it is often reasonable to consider that the effects of neighboring geographic areas are correlated (defined by a contiguity matrix that is included in the modeling stage). Readers interested in simultaneous autoregressive models in small area estimation can refer to the following literature, which presents extended evaluations and applications: Singh, Shukla, and Kundu (2005), Pratesi and Salvati (2008), Pratesi and Salvati (2009), Molina (2009), Marhuenda, Molina, and Morales (2013), and Porter, Wikle, and Holan (2015). In addition, Bayesian approaches to small area estimation considering conditionally autoregressive models can model the spatial dependence (see Besag, York, and Mollié 1991; Leroux, Lei, and Breslow 2000; Mercer et al. 2014). These spatial extensions are widely used for small area estimation problems of health and social indicators in sociology, demography, and epidemiology. Dwyer-Lindgren et al. (2016) used the spatial Besag-York-Mollie model (Besag et al. 1991) to estimate mortality rates. The Centers for Disease Control and Prevention PLACES project adopted a similar strategy to estimate small area indicators of morbidity (<https://www.cdc.gov/places/methodology/index.html>). Regarding poverty indicators, readers may refer to Marhuenda et al. (2013) and Giusti, Masserini, and Pratesi (2017). In this article, I did not include a spatial process into the random area effects. Spatial extensions of the proposed methods are interesting

and useful for users, and they will be the subject of future research given they require particular methodological attention.

**APPENDIX A: EXTRA OUTPUTS OF THE SIMULATION STUDY**



**Figure A1.** Ratios between the small area estimates and true values for the estimates obtained via direct (DIR; dotted line), univariate (UEPP; dashed line), and multivariate (MEPP; solid line) estimators for the mean of log income,  $k = 1$ , ordered by decreasing sampling fraction.



**Figure A2.** Ratios between the small area estimates and true values for the estimates obtained via direct (DIR; dotted line), univariate (UEPP; dashed line), and multivariate (MEPP; solid line) estimators for the poverty proportion,  $k = 2$ , ordered by decreasing sampling fraction.

**Table A1.** Quantiles of the Distributions of the Small Area Estimates (with True Values) Obtained via the Different Approaches for  $k = 1, 2$ 

Estimator	Quantile								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
$\hat{y}_1^{\text{MEPP}}$	9.648	9.661	9.665	9.674	9.677	9.680	9.683	9.688	9.690
$\hat{y}_1^{\text{UEPP}}$	9.657	9.659	9.661	9.662	9.665	9.667	9.670	9.671	9.674
$\hat{y}_1^{\text{DIR}}$	9.587	9.590	9.600	9.649	9.648	9.638	9.659	9.799	9.823
$\hat{y}_1^{\text{True}}$	9.588	9.598	9.606	9.618	9.625	9.631	9.643	9.791	9.818
$\hat{y}_2^{\text{MEPP}}$	.222	.232	.233	.235	.238	.240	.249	.275	.278
$\hat{y}_2^{\text{UEPP}}$	.224	.224	.234	.234	.237	.239	.249	.277	.278
$\hat{y}_2^{\text{DIR}}$	.201	.206	.225	.233	.236	.245	.250	.269	.274
$\hat{y}_2^{\text{True}}$	.200	.207	.227	.235	.238	.247	.250	.272	.277

## APPENDIX B: R PROGRAM

This R program can be used to apply the methods used in this article. First, to estimate multivariate mixed models, such as the joint mixed model introduced in the article, the software *sabre* can be installed from [http://sabre.lancs.ac.uk/installing\\_intro.html](http://sabre.lancs.ac.uk/installing_intro.html) and then installed in R manually.

A possible R program that can be used to replicate the results of the simulation study for the multivariate case is given below. An R program for the univariate case is available in Chandra et al. (2018).

```

population=read.table("universe.txt", header=T)
population$AGE_new <- as.numeric(population$AGE)
povertyThreshold <- 0.6*median(population[["income"]])
population$poor <- as.numeric(ifelse(population$income <povertyThreshold, 1, 0))
population$log_income <- log(population$income)
population$income1 <- population$income
population$income1[population$income1==0] <- 0.0000001
population$log_income1 <- log(population$income1)
population$y1 <- population$log_income1 #log_income1
population$y2 <- population$poor #poor
population$x1 <- population$AGE_new
population$area <- population$DIS

True_y1_bar <- aggregate(population$y1,
  list(population$area), mean)[,2]#true mean y1
True_y2_bar <- aggregate(population$y2,
  list(population$area), mean)[,2]#true mean y2
true <- cbind(True_y1_bar, True_y2_bar)

```

```

x1bar <- aggregate(population$x1,
  list(population$area), mean) [,2] #true mean x1
D <- nrow(true)
Xbar <- cbind(int=rep.int(1,D), x1bar=x1bar)
M=50
K=2 #number of variables
Nd <- data.frame(table(population$area)) [,2]
domsize <- data.frame(table(population$area))
colnames(domsize)[1] <- "area_code"; colnames(domsize)[2] <- "Nd"
N <- sum(Nd)
nd <- rep.int(3, D)
fd <- nd/Nd
n <- sum(nd)
areas <- rep.int(1:D, time=Nd)
S <- 500
EMSE_y1_EPP <- EMSE_y2_EPP <- ERMSE_y1_EPP <- matrix(NA, D, S)
ERMSE_y2_EPP <- Bias_y1_EPP <- Bias_y2_EPP <- matrix(NA, D, S)
EMSE_y1_HT <- EMSE_y2_HT <- ERMSE_y1_HT <- matrix(NA, D, S)
ERMSE_y2_HT <- Bias_y1_HT <- Bias_y2_HT <- matrix(NA, D, S)
y1_dir <- y2_dir <- y1_ind <- y2_ind <- matrix(NA, D, S)
y1_true <- y2_true <- matrix(NA, D, S)
for (s in 1:S) {
  sample = strata(population, stratanames="area",
    size=nd, method=c("srswor"), description=FALSE)
  sample = getdata(population, sample)
  sample = cbind(sample, w=1/sample$Prob)
  x1bar_s <- aggregate(sample$x1, list(sample$area), mean) [,2] #sample mean x1
  xbar_s <- cbind(int=rep.int(1,D), x1bar_s=x1bar_s)
  y1bar_s <- aggregate(sample$y1, list(sample$area), mean) [,2] #sample mean y1
  y2bar_s <- aggregate(sample$y2, list(sample$area), mean) [,2] #sample mean y2
  y_bar_s <- cbind(y1bar_s, y2bar_s)
  xbar_sam <- xbar_s
  y1_HT <- direct(y=y1, dom=area, sweight=w, domsize=domsize,
    data=sample, replace = FALSE)
  y1_HT_est <- y1_HT$Direct
  y2_HT <- direct(y=y2, dom=area, sweight=w, domsize=domsize,
    data=sample, replace = FALSE)
  y2_HT_est <- y2_HT$Direct
  y_HT_est <- cbind(y1_HT_est, y2_HT_est)
  multi_model <- sabre(sample$y1 ~ sample$x1,
    sample$y2 ~ sample$x1,
    case = sample$area,
    first.family="gaussian", second.family="binomial",
    second.link="1", correlated = "yes")
}

```



```

model_results <- multi_model$fit.estimate.print.message
model_results <- unlist(strsplit(model_results, " "))
model_results <- as.data.frame(na.omit(model_results))
model_results <- as.numeric(as.matrix(model_results))
model_results <- as.data.frame(na.omit(model_results))
res <- model_results
beta <- matrix(c(res[1,], res[4,], res[7,], res[10,]), 2, 2)
Sigma_e <- matrix(c(res[13,]^2, 0, 0, 3.29), 2, 2)
cor_u <- res[22,]
Sigma_u <- matrix(c( (res[16,])^2, (res[16,]*res[19,])*cor_u,
                    (res[16,]*res[19,])*cor_u, (res[19,])^2), 2, 2)
y_EPP_0m1 <- matrix(NA, D, M)
y_EPP_0m2 <- matrix(NA, D, M)
for (m in 1 : M) {
  u_m <- mvrnorm(D, numeric(K), Sigma_u)
  u_star1m <- u_m[,1]
  u_star2m <- u_m[,2]
  u_rep_1m <- rep(u_m[,1], time=nd)
  u_rep_2m <- rep(u_m[,2], time=nd)
  y_EPP_0m <- Xbar %*% beta + cbind(u_star1m, u_star2m)
  y_EPP_02m <- exp( Xbar %*% beta + cbind(u_star1m, u_star2m) ) /
    (1 + exp( Xbar %*% beta + cbind(u_star1m, u_star2m) ) )
  y_EPP_0m1[,m] <- y_EPP_0m[,1]
  y_EPP_0m2[,m] <- y_EPP_02m[,2]
}
y_EPP_01 <- rowMeans(y_EPP_0m1)
y_EPP_02 <- rowMeans(y_EPP_0m2)
y_EPP <- cbind(y_EPP_01, y_EPP_02)
y_EPP_final <- matrix(NA, D, K)
for (d in 1 : D) {
  y_EPP_final[d,] <- fd[d] * y_HT_est[d,] + (1-fd[d]) * y_EPP[d,] #y_bar_s
}
y1_EPP_final <- y_EPP_final[,1]
y2_EPP_final <- y_EPP_final[,2]
MSE_y1_EPP <- (y1_EPP_final - True_y1_bar)^2
RMSE_y1_EPP <- sqrt(MSE_y1_EPP)
BIAS_y1_EPP <- (y1_EPP_final - True_y1_bar)
MSE_y2_EPP <- (y2_EPP_final - True_y2_bar)^2
RMSE_y2_EPP <- sqrt(MSE_y2_EPP)
BIAS_y2_EPP <- (y2_EPP_final - True_y2_bar)
EMSE_y1_EPP[,s] <- MSE_y1_EPP
EMSE_y2_EPP[,s] <- MSE_y2_EPP
ERMSE_y1_EPP[,s] <- RMSE_y1_EPP


```

```

ERMSE_y2_EPP[,s] <- RMSE_y2_EPP
Bias_y1_EPP[,s] <- BIAS_y1_EPP
Bias_y2_EPP[,s] <- BIAS_y2_EPP
y1_ind[,s] <- y1_EPP_final
y2_ind[,s] <- y2_EPP_final
MSE_y1_HT <- (y1_HT_est - True_y1_bar)^2
RMSE_y1_HT <- sqrt(MSE_y1_HT)
BIAS_y1_HT <- (y1_HT_est - True_y1_bar)
MSE_y2_HT <- (y2_HT_est - True_y2_bar)^2
RMSE_y2_HT <- sqrt(MSE_y2_HT)
BIAS_y2_HT <- (y2_HT_est - True_y2_bar)
EMSE_y1_HT[,s] <- MSE_y1_HT
EMSE_y2_HT[,s] <- MSE_y2_HT
ERMSE_y1_HT[,s] <- RMSE_y1_HT
ERMSE_y2_HT[,s] <- RMSE_y2_HT
Bias_y1_HT[,s] <- BIAS_y1_HT
Bias_y2_HT[,s] <- BIAS_y2_HT
y1_dir[,s] <- y1_HT_est
y2_dir[,s] <- y2_HT_est
y1_true[,s] <- True_y1_bar
y2_true[,s] <- True_y2_bar
)#end of simulation
MSE_SUMMARY <- data.frame(EMSEy1_EPP=rowMeans(EMSE_y1_EPP),
                           EMSEy2_EPP=rowMeans(EMSE_y2_EPP),
                           ERMSEy1_EPP=rowMeans(ERMSE_y1_EPP),
                           ERMSEy2_EPP=rowMeans(ERMSE_y2_EPP),
                           EMSEy1_HT=rowMeans(EMSE_y1_HT),
                           EMSEy2_HT=rowMeans(EMSE_y2_HT),
                           ERMSEy1_HT=rowMeans(ERMSE_y1_HT),
                           ERMSEy2_HT=rowMeans(ERMSE_y2_HT),
                           ERRMSEy1_EPP=rowMeans(ERMSE_y1_EPP)/rowMeans(y1_true),
                           ERRMSEy2_EPP=rowMeans(ERMSE_y2_EPP)/rowMeans(y2_true),
                           ERRMSEy1_HT=rowMeans(ERMSE_y1_HT)/rowMeans(y1_true),
                           ERRMSEy2_HT=rowMeans(ERMSE_y2_HT)/rowMeans(y2_true))
Relbias <- data.frame(RB_HT_y1 = rowMeans(Bias_y1_HT) / rowMeans(y1_true),
                      RB_HT_y2 = rowMeans(Bias_y2_HT) / rowMeans(y2_true),
                      RB_EPP_y1 = rowMeans(Bias_y1_EPP) / rowMeans(y1_true),
                      RB_EPP_y2 = rowMeans(Bias_y2_EPP) / rowMeans(y2_true))
Estimates <- data.frame(y1_true = rowMeans(y1_true), y2_true = rowMeans(y2_true),
                        y1_EPP = rowMeans(y1_ind), y2_EPP = rowMeans(y2_ind),
                        y1_HT = rowMeans(y1_dir), y2_HT = rowMeans(y2_dir) )

```

**ORCID iD**

Angelo Moretti  <https://orcid.org/0000-0001-6543-9418>

**Note**

1. Installation instructions can be found at [http://sabre.lancs.ac.uk/sabreRuse\\_intro.html](http://sabre.lancs.ac.uk/sabreRuse_intro.html).

**References**

- Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Vol. 4. Berlin, Germany: Springer Science & Business Media.
- Argyle, Michael. 1997. "Is Happiness a Cause of Health?" *Psychology and Health* 12(6):769–81.
- Battese, George E., Rachel M. Harter, and Wayne A. Fuller. 1988. "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association* 83(401):28–36.
- Benavent, Roberto, and Domingo Morales. 2016. "Multivariate Fay–Herriot Models for Small Area Estimation." *Computational Statistics & Data Analysis* 94:372–90.
- Benavent, Roberto, and Domingo Morales. 2020. "Small Area Estimation under a Temporal Bivariate Area-Level Linear Mixed Model with Independent Time Effects." *Statistical Methods and Applications* 30:195–222.
- Berridge, Damon Mark, and Robert Crouchley. 2011. *Multivariate Generalized Linear Mixed Models Using R*. Boca Raton, FL: CRC.
- Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics." *Annals of the Institute of Statistical Mathematics* 43(1):1–20.
- Betti, Gianni, and Achille Lemmi. 2013. *Poverty and Social Exclusion: New Methods of Analysis*. New York: Routledge.
- Booth, James G., and James P. Hobert. 1999. "Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(1):265–85.
- Burgard, Jan Pablo, Jan-Philipp Kolb, Hariolf Merkle, and Ralf Münnich. 2017. "Synthetic Data for Open and Reproducible Methodological Research in Social Sciences and Official Statistics." *AStA Wirtschafts- und Sozialstatistisches Archiv* 11(3–4):233–44.
- Chambers, Ray, Hukum Chandra, and Nikos Tzavidis. 2011. "On Bias-Robust Mean Squared Error Estimation for Pseudo-Linear Small Area Estimators." *Survey Methodology* 37(2):153–70.
- Chambers, Ray, Nicola Salvati, and Nikos Tzavidis. 2016. "Semiparametric Small Area Estimation for Binary Outcomes with Application to Unemployment Estimation for Local Authorities in the UK." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(2):453–79.
- Chandra, Hukum, Ray Chambers, and Nicola Salvati. 2012. "Small Area Estimation of Proportions in Business Surveys." *Journal of Statistical Computation and Simulation* 82(6):783–95.
- Chandra, Hukum, Ray Chambers, and Nicola Salvati. 2019. "Small Area Estimation of Survey Weighted Counts under Aggregated Level Spatial Model." *Survey Methodology* 45(1):31–59.
- Chandra, Hukum, Sushil Kumar, and Kaustav Aditya. 2018. "Small Area Estimation of Proportions with Different Levels of Auxiliary Data." *Biometrical Journal* 60(2):395–415.
- Chatterjee, Deen K. 2011. *Encyclopedia of Global Justice: A–I*. Berlin, Germany: Springer Science & Business Media.
- Coull, Brent A., and Alan Agresti. 2000. "Random Effects Modeling of Multiple Binomial Responses Using the Multivariate Binomial Logit-Normal Distribution." *Biometrics* 56(1):73–80.
- Cressie, Noel. 2015. *Statistics for Spatial Data*. Hoboken, NJ: John Wiley.
- Datta, Gauri Sankar, Bannmo Day, and Ishwar Basawa. 1999. "Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation." *Journal of Statistical Planning and Inference* 75(2):269–79.

- Dwyer-Lindgren, Laura, Amelia Bertozzi-Villa, Rebecca W. Stubbs, Chloe Morozoff, Michael J. Kutz, Chantal Huynh, Ryan M. Barber, et al. 2016. "US County-Level Trends in Mortality Rates for Major Causes of Death, 1980–2014." *JAMA* 316(22):2385–401.
- Fabrizi, Enrico, Maria R. Ferrante, and Silvia Pacei. 2005. "Estimation of Poverty Indicators at Sub-national Level Using Multivariate Small Area Models." *Statistics in Transition* 7(3):587–608.
- Fay, R. E. 1987. "Application of Multivariate Regression to Small Domain Estimation." Pp. 91–102 in *Small Area Statistics*, edited by R. Platek, J.N.K. Rao, C. E. Sarndal, and M. P. Singh. New York: John Wiley.
- Fuller, Wayne A., and R. M. Harter. 1987. "The Multivariate Components of Variance Model for Small Area Estimation." Pp. 103–23 in *Small Area Statistics*, edited by R. Platek, J.N.K. Rao, C. E. Sarndal, and M. P. Singh. New York: John Wiley.
- Giusti, Caterina, Lucio Masserini, and Monica Pratesi. 2017. "Local Comparisons of Small Area Estimates of Poverty: An Application within the Tuscany Region in Italy." *Social Indicators Research* 131(1):235–54.
- Giusti, Caterina, Nikos Tzavidis, Monica Pratesi, and Nicola Salvati. 2014. "Resistance to Outliers of M-Quantile and Robust Random Effects Small Area Models." *Communications in Statistics–Simulation and Computation* 43(3):549–68.
- Goldstein, Harvey. 2011. *Multilevel Statistical Models*. Hoboken, NJ: John Wiley.
- Goldstein, Harvey, James Carpenter, Michael G. Kenward, and Kate A. Levin. 2009. "Multilevel Models with Multivariate Mixed Response Types." *Statistical Modelling* 9(3):173–97.
- González-Manteiga, Wenceslao, María José Lombardía, Isabel Molina, Domingo Morales, and Laureano Santamaría. 2007. "Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters under a Logistic Mixed Model." *Computational Statistics & Data Analysis* 51(5):2720–33.
- Graham, Carol. 2012. *Happiness around the World: The Paradox of Happy Peasants and Miserable Millionaires*. Oxford, UK: Oxford University Press.
- Hobza, Tomáš, Domingo Morales, and Laureano Santamaría. 2018. "Small Area Estimation of Poverty Proportions under Unit-Level Temporal Binomial-Logit Mixed Models." *Test* 27(2):270–94.
- Ivanova, Anna, Geert Molenberghs, and Geert Verbeke. 2016. "Mixed Models Approaches for Joint Modeling of Different Types of Responses." *Journal of Biopharmaceutical Statistics* 26(4):601–18.
- Lehtonen, Risto, and Ari Veijanen. 2016. "Model-Assisted Method for Small Area Estimation of Poverty Indicators." Pp. 109–27 in *Analysis of Poverty Data by Small Area Estimation*, edited by M. Pratesi. Berlin, Germany: Springer.
- Leroux, Brian G., Xingye Lei, and Norman Breslow. 2000. "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence." Pp. 179–91 in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York: Springer.
- Marhuenda, Yolanda, Isabel Molina, and Domingo Morales. 2013. "Small Area Estimation with Spatio-temporal Fay–Herriot Models." *Computational Statistics & Data Analysis* 58:308–25.
- McCulloch, Charles E. 1994. "Maximum Likelihood Variance Components Estimation for Binary Data." *Journal of the American Statistical Association* 89(425):330–35.
- McCulloch, Charles E. 1997. "Maximum Likelihood Algorithms for Generalized Linear Mixed Models." *Journal of the American Statistical Association* 92(437):162–70.
- Mercer, Laina, Jon Wakefield, Cici Chen, and Thomas Lumley. 2014. "A Comparison of Spatial Smoothing Methods for Small Area Estimation with Sampling Weights." *Spatial Statistics* 8:69–85.
- Molina, Isabel. 2009. "Uncertainty under a Multivariate Nested-Error Regression Model with Logarithmic Transformation." *Journal of Multivariate Analysis* 100(5):963–80.
- Molina, Isabel, and Ewa Strzalkowska-Kominiak. 2020. "Estimation of Proportions in Small Areas: Application to the Labour Force using the Swiss Census Structural Survey." *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(1):281–310.
- Moretti, Angelo. 2022. "Estimation of Small Area Proportions under a Bivariate Logistic Mixed Model." *Quality & Quantity*. doi:10.1007/s11135-022-01530-6.
- Moretti, Angelo, and Natalie Shlomo. 2023. "Improving Statistical Matching when Auxiliary Information Is Available." *Journal of Survey Statistics and Methodology*. doi:10.1093/jssam/smac038.

- Moretti, Angelo, Natalie Shlomo, and Joseph W. Sakshaug. 2020. "Parametric Bootstrap Mean Squared Error of a Small Area Multivariate EBLUP." *Communications in Statistics—Simulation and Computation* 49(6):1474–86.
- Moretti, Angelo, Natalie Shlomo, and Joseph W. Sakshaug. 2020. "Multivariate Small Area Estimation of Multidimensional Latent Economic Well-Being Indicators." *International Statistical Review* 88(1): 1–28.
- Moretti, Angelo, Natalie Shlomo, and Joseph W. Sakshaug. 2021. "Small Area Estimation of Latent Economic Well-Being." *Sociological Methods & Research* 50(4):1660–93.
- Pfeffermann, Danny. 2013. "New Important Developments in Small Area Estimation." *Statistical Science* 28(1):40–68.
- Pierewan, Adi Cilik, and Gindo Tampubolon. 2015. "Happiness and Health in Europe: A Multivariate Multilevel Model." *Applied Research in Quality of Life* 10(2):237–52.
- Porter, Aaron T., Christopher K. Wikle, and Scott H. Holan. 2015. "Small Area Estimation via Multivariate Fay–Herriot Models with Latent Spatial Dependence." *Australian & New Zealand Journal of Statistics* 57(1):15–29.
- Prasad, N. G. Narasimha, and Jon N. K. Rao. 1990. "The Estimation of the Mean Squared Error of Small-Area Estimators." *Journal of the American Statistical Association* 85(409):163–71.
- Pratesi, Monica. 2016. *Analysis of Poverty Data by Small Area Estimation*. Hoboken, NJ: John Wiley.
- Pratesi, Monica, and Nicola Salvati. 2008. "Small Area Estimation: The EBLUP Estimator Based on Spatially Correlated Random Area Effects." *Statistical Methods and Applications* 17(1):113–41.
- Pratesi, Monica, and Nicola Salvati. 2009. "Small Area Estimation in the Presence of Correlated Random Area Effects." *Journal of Official Statistics* 25(1):37–53.
- Rabe-Hesketh, Sophia, and Anders Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.
- Rao, J.N.K., and Isabel Molina. 2015. *Small Area Estimation*. Hoboken, NJ: John Wiley.
- Salvati, Nicola, Hukum Chandra, and Ray Chambers. 2012. "Model-Based Direct Estimation of Small-Area Distributions." *Australian & New Zealand Journal of Statistics* 54(1):103–23.
- Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 2003. *Model Assisted Survey Sampling*. Berlin, Germany: Springer Science & Business Media.
- Singh, Bharat Bhushan, Girja Kant Shukla, and Debasis Kundu. 2005. "Spatio-temporal Models in Small Area Estimation." *Survey Methodology* 31(2):183.
- Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.
- Ubaidillah, Azka, Khairil Anwar Notodiputro, Anang Kurnia, and I. Wayan Mangku. 2019. "Multivariate Fay–Herriot Models for Small Area Estimation with Application to Household Consumption per Capita Expenditure in Indonesia." *Journal of Applied Statistics* 46(15):2845–61.
- Wight, Vanessa, Neeraj Kaushal, Jane Waldfogel, and Irv Garfinkel. 2014. "Understanding the Link between Poverty and Food Insecurity among Children: Does the Definition of Poverty Matter?" *Journal of Children and Poverty* 20(1):1–20.

### Author Biography

**Angelo Moretti** is an assistant professor in statistics at Utrecht University in the Department of Methodology and Statistics. He is a survey statistician and an elected member of the International Statistical Institute. He has conducted research in small area estimation under multivariate mixed models, survey calibration, mean squared error estimation based on bootstrap approaches, and data integration methods (statistical matching and probabilistic record linkage). He is also interested in applications related to understanding geographic differences in social exclusion, crime, and public attitudes indicators.