





Empirical best prediction of small area bivariate parameters

María Dolores Esteban¹  | María José Lombardía²  |
Esther López-Vizcaíno³  | Domingo Morales¹  | Agustín Pérez¹ 

¹Research Center IUCIO, Universidad Miguel Hernández de Elche, Elche, Spain

²Research Center CITIC, Universidade da Coruña, Coruña, Spain

³Instituto Galego de Estatística, Coruña, Spain

Correspondence

Domingo Morales, Instituto Universitario Centro de Investigación Operativa, Universidad Miguel Hernández de Elche, 03202 Elche, Spain.
Email: d.morales@umh.es

Funding information

Generalitat Valenciana, Grant/Award Number: Prometeo/2021/063; Ministerio de Ciencia e Innovación, Grant/Award Numbers: PGC2018-096840-B-I00, PID2020-113578RB-I00; Xunta de Galicia, Grant/Award Numbers: ED431G 2019/01, ED431C 2020/14, COV20/00604

Abstract

This paper introduces empirical best predictors of small area bivariate parameters, like ratios of sums or sums of ratios, by assuming that the target unit-level vector follows a bivariate nested error regression model. The corresponding means squared errors are estimated by parametric bootstrap. Several simulation experiments empirically study the behavior of the introduced statistical methodology. An application to real data from the Spanish household budget survey gives estimators of ratios of food household expenditures by provinces.

KEYWORDS

best linear unbiased predictors, household budget surveys, multivariate linear mixed models, nested error regression models, ratio estimators, small area estimation

1 | INTRODUCTION

Complex indicators based on more than one variable play an important role in public statistics. For a finite population, partitioned in domains or small areas, examples of such indicators are the ratios of domain means or the domain means of ratios. In the first case, we may have the quotient between the mean annual expenditure on food of the households from a given territory and the corresponding mean annual expenditure on all items of expenditure. In the second case, we have the domain mean of the proportions of annual household expenditures used for food.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Scandinavian Journal of Statistics* published by John Wiley & Sons Ltd on behalf of The Board of the Foundation of the Scandinavian Journal of Statistics.

One way to estimate a ratio of domain means is to estimate the numerator and denominator separately and independently and substitute in its expression. This approach leads to the use of plug-in estimators, which have the problem of being biased even though their components are unbiasedly estimated. There are two additional inconveniences. The first one is not considering the correlation between the variables that intervene in the definition of the population parameters of the ratio type. The second one is that the asymptotic property of unbiasedness cannot be assumed for estimators of domain indicators if sample sizes are small.

For estimating domain means of households ratios of food expenditure, as well as for other nonlinear bivariate parameters, the statistical literature presents few model-based contributions. For covering this gap, Erciulescu et al. (2018) gave an interesting proposal. They discussed some methods of applying benchmarking constraints to a triplet (numerator, denominator, ratio), at multiple stages of aggregation, where the denominator and the ratio are modeled and the numerator is derived. This manuscript follows a different approach by introducing predictors of ratio-type domain indicators based on unit-level bivariate models.

Small area estimation (SAE) gives statistical methodology to estimate parameters of population subsets, called domains or small areas. The word “small” refers to sample size and not to population size. To overcome the problem of having a small sample size in a domain, SAE complements the data of the target variable with data of auxiliary variables, information taken from other domains and correlation structures. All this can be done by fitting models to the available data for the entire population and building estimators based on the selected model. This is the unit-level model-based approach. Alternatively, models can be used for aggregated data and then inferential procedures are based on area-level models.

This paper defines *domain parameter* as a function of the values taken by one or more objective variables in all units of the population. The mathematical expression (formula) of a domain parameter is therefore relevant. If the target variables are continuous, then it is possible to estimate linear domain parameters with empirical best linear unbiased predictors (EBLUP) based on linear mixed models (LMM). However, domain parameters are often nonlinear or defined by non-continuous target variables. In those cases, it is quite common to estimate domain parameters with empirical best (or Bayes) predictors (EBP) based on LMMs or on generalized LMMs (GLMM). This paper deals with the estimation of domain parameters that are nonlinear functions of two continuous variables and puts special emphasis in the estimation of ratios of domain means and domain means of ratios.

As the domain parameters of interest depend on several target variables, the use of multivariate models is recommended. Since the first works of Fay (1987), Datta et al. (1991, 1999), the statistical literature contains some applications of these models to the SAE setup. Concerning area-level multivariate models, Molina et al. (2007), López-Vizcaíno et al. (2013, 2015) and Esteban et al. (2020) derived predictors for totals of employed and unemployed people and for unemployment rates based on multinomial-logit or compositional mixed models. Morales et al. (2015), Porter et al. (2015), Benavent and Morales (2016, 2021) or Arima et al. (2017) studied the problem of estimating poverty indicators, including the nonlinear poverty gap. Marchetti and Secondi (2017) and Ubaidillah et al. (2019) estimated household consumption expenditures by applying Fay-Herriot models. Erciulescu and Opsomer (2019) predicted employee compensation components by using a hierarchical Bayes bivariate Fay-Herriot models.

Concerning unit-level multivariate models, Fuller and Harter (1987) introduced the multivariate nested error regression (NER) model and Datta et al. (1998) applied this model to hierarchical Bayes prediction of small area mean vectors. For analyzing unit-level multivariate data in SAE, Ngaruye et al. (2017), Ito and Kubokawa (2021) and Esteban et al. (2022) gave EBLUPs of domain

means and totals to treat problems of repeated measures, posted land prices, and expenditure data, respectively. Furthermore, Erciulescu et al. (2019) employed a bivariate hierarchical Bayesian unit-level model for estimating cropland cash rental rates at the county level.

On the other hand, the EBPs are widely employed when the domain parameters of interest are nonlinear. Since the first works of Jiang and Lahiri (2001) and Jiang (2003), where EBPs of functions of fixed effects and small-area-specific random effects were developed under GLMMs, some authors have extended their procedures and applied EBPs in the SAE context. For example, Boubeta et al. (2016, 2017) and Hobza and Morales (2016); Hobza et al. (2018) derived EBPs of small area poverty proportions based on area-level Poisson mixed models and unit-level logit mixed models, respectively. Erciulescu and Fuller (2016) introduced predictors under alternative specifications of generalized mixed models. Erciulescu and Fuller (2018) constructed bootstrap prediction intervals for small area means from unit-level nonlinear models. Marino et al. (2019) and Hobza et al. (2020) proposed EBPs under semi-parametric and parametric unit-level GLMMs. Chandra et al. (2017, 2018) and Chandra and Salvati (2018) introduced SAE predictors of spatially correlated count data. Torabi (2019) proposed a class of spatial GLMMs to obtain small area predictors of esophageal cancer prevalence. This uncomplete list of contribution shows the high impact that the EBP approach has in SAE. We refer to Rao and Molina (2015) or to Morales et al. (2021) for a more complete list of references.

Based on the NER model, the seminal paper of Molina and Rao (2010) introduced the basic theory for calculating EBPs of domain nonlinear parameters, depending on one continuous target variable, under unit-level LMMs. Molina et al. (2014) and Guadarrama et al. (2016) assumed that a transformation of the study variable follows a NER model. Their results were later extended to the two-fold NER model by Marhuenda et al. (2017), to log-normal models by Molina, and Martín (2018), to data-driven transformations by Rojas-Perilla et al. (2020) and to unit-level mixed models with skewed distributions by Graf et al. (2019) and Diallo and Rao (2018). However, the statistical literature has not yet treated the problem of constructing EBPs, based on bivariate NER models, for estimating domain nonlinear parameters defined by two continuous variables. The new best predictors are unbiased under the distribution of the selected model. This is the main contribution of this paper.

By following González-Manteiga et al. (2007, 2008), this article introduces a parametric bootstrap procedure for estimating the mean squared errors (MSE) of the EBPs. As the optimality properties of the best predictors might not hold for EBPs when the number of domains and the domain sample sizes are small, an empirical research is carried out. In addition to the mathematical developments, Monte Carlo simulations empirically investigate the properties of the EBPs and the corresponding MSE estimators. Finally, the new statistical methodology is applied to data from the 2016 Spanish Household Budget Survey (SHBS). The target is to estimate means of ratios and ratios of means of food expenditures in Spanish households at the province level. Both indicators are employed in macro- and micro-economic studies, respectively.

The paper derives statistical methodology for SAE under the unit-level model-based approach. This is to say, it assumes the prediction theory for finite population inference. See for example, Valliant et al. (2000) for a description of this theory. Therefore, this paper does not take into account the sampling design distribution and related issues in the derivation of the predictors and in the study of their properties. The proposed statistical methodology is based on a bivariate NER model, where the error term is normally distributed. Like many other SAE methods, our proposal follows the parametric statistical inference approach. That has advantages and disadvantages. The advantage is that we introduce a predictor with optimality properties under the assumption that

the assumed hypotheses are fulfilled. The drawback is that these hypotheses are not always fulfilled in practice. Therefore, it is necessary to make a diagnosis of the model before accepting it as a working tool to calculate predictors of small area parameters.

The rest of the paper is organized as follows. Section 2 introduces the bivariate NER model. Section 3 derives the EBPs of additive and nonadditive domain parameters, including predictors of domain ratios. Section 4 describes a parametric bootstrap procedure to estimate MSEs of the introduced predictors. Section 5 carries out simulation experiments to investigate the behavior of the predictors of ratio-type domain parameters and the MSE estimators. Section 6 gives an illustrative application to data from the SHBS of 2016, where the target is the SAE of means of ratios and ratios of means of household annual food expenditures by provinces. Section 7 summarizes some conclusions. We give Data S1 with several appendices. Appendix A gives alternative mathematical derivations for the best predictors of random effects. Appendices B and C contain complementary simulation results. Appendices D, E, and F present further insights on the application to real data.

2 | THE MODEL

Let U be a population of size N partitioned into D domains or areas U_1, \dots, U_D of sizes N_1, \dots, N_D respectively. Let $N = \sum_{j=1}^D N_d$ be the global population size. Let $y_{dj} = (y_{dj1}, y_{dj2})'$ be a vector of continuous variables measured on the sample unit j of domain d , $d = 1, \dots, D$, $j = 1, \dots, N_d$. For $k = 1, 2$, let $x_{dj k} = (x_{dj k1}, \dots, x_{dj k p_k})$ be a row vector containing p_k explanatory variables and let $X_{dj} = \text{diag}(x_{dj1}, x_{dj2})_{2 \times p}$ with $p = p_1 + p_2$. Let β_k be a column vector of size p_k containing regression parameters and let $\beta = (\beta'_1, \beta'_2)'_{p \times 1}$. The population homoscedastic bivariate NER (BNER) model assumes that

$$y_{dj} = X_{dj}\beta + u_d + e_{dj}, \quad d = 1, \dots, D, \quad j = 1, \dots, N_d. \quad (1)$$

where the vectors of random effects $\{u_d\}$ and random errors $\{e_{dj}\}$ are independent with multivariate distributions

$$u_d \sim N_2(0, V_{ud}), \quad e_{dj} \sim N_2(0, V_{edj}),$$

and variance-covariance matrices that do not vary with units or domains, that is,

$$V_{ud} = \begin{pmatrix} \sigma_{u1}^2 & \rho_{u12}\sigma_{u1}\sigma_{u2} \\ \rho_{u12}\sigma_{u1}\sigma_{u2} & \sigma_{u2}^2 \end{pmatrix}, \quad V_{edj} = \begin{pmatrix} \sigma_{e1}^2 & \rho_{e12}\sigma_{e1}\sigma_{e2} \\ \rho_{e12}\sigma_{e1}\sigma_{e2} & \sigma_{e2}^2 \end{pmatrix},$$

with parameters $\theta_1 = \sigma_{u1}^2$, $\theta_2 = \sigma_{u2}^2$, $\theta_3 = \rho_{u12}$, $\theta_4 = \sigma_{e1}^2$, $\theta_5 = \sigma_{e2}^2$ and $\theta_6 = \rho_{e12}$. Let I_m be the $m \times m$ identity matrix. We define the $2N_d \times 1$ vectors and the $2N_d \times p$ and $2N_d \times 2$ matrices

$$y_d = \underset{1 \leq j \leq N_d}{\text{col}}(y_{dj}), \quad e_d = \underset{1 \leq j \leq N_d}{\text{col}}(e_{dj}), \quad X_d = \underset{1 \leq j \leq N_d}{\text{col}}(X_{dj}), \quad Z_d = \underset{1 \leq j \leq N_d}{\text{col}}(I_2).$$

Model (1) can be written in the domain-level form

$$y_d = X_d\beta + Z_d u_d + e_d, \quad d = 1, \dots, D, \quad (2)$$

where $u_d \sim N_2(0, V_{ud})$, $e_d \sim N_{2N_d}(0, V_{ed})$ are independent and $V_{ed} = \text{diag}(V_{edj})_{1 \leq j \leq N_d}$. We define the $2N \times 1$ and $2D \times 1$ vectors and the $2N \times p$ and $2N \times 2D$ matrices

$$y = \text{col}_{1 \leq d \leq D}(y_d), \quad e = \text{col}_{1 \leq d \leq D}(e_d), \quad u = \text{col}_{1 \leq d \leq D}(u_d), \quad X = \text{col}_{1 \leq d \leq D}(X_d), \quad Z = \text{diag}_{1 \leq d \leq D}(Z_d).$$

Model (1) can be written in the linear mixed model form

$$y = X\beta + Zu + e. \tag{3}$$

where $u \sim N_{2D}(0, V_u)$, $e \sim N_{2N}(0, V_e)$ are independent, $V_u = \text{diag}(V_{ud})_{1 \leq d \leq D}$ and $V_e = \text{diag}(V_{ed})_{1 \leq d \leq D}$.

As this paper assumes the prediction theory, where the only source of randomness comes from the distribution of vector y , derived from model (3), the inference is carried out based on a fixed subset (called sample), s , of the finite population U . Let $s = \cup_{d=1}^D s_d$ and $r = \cup_{d=1}^D r_d$, with $s \cap r = \emptyset$ and $s \cup r = U$, denote the subsets containing the ‘‘sample’’ and ‘‘out-of-sample’’ units. Let y_s and y_{ds} be the subvectors of y and y_d corresponding to sample elements and y_r and y_{dr} the subvectors of y and y_d corresponding to the out-of-sample elements. Without lack of generality, we can write $y_d = (y'_{ds}, y'_{dr})'$. Define also the corresponding decompositions of X_d, Z_d, V_{ed} and V_d .

As we assume that sample indexes are fixed, the sample subvectors y_{ds} follow the marginal models derived from the population model (2), that is,

$$y_{ds} = X_{ds}\beta + Z_{ds}u_d + e_{ds}, \quad d = 1, \dots, D,$$

where $u_d \sim N_2(0, V_{ud})$, $e_{ds} \sim N_{2n_d}(0, V_{ed,s})$ are independent and $V_{ed,s} = \text{diag}(V_{edj})_{1 \leq j \leq n_d}$. For $d = 1, \dots, D$, the vectors y_{ds} are independent with $y_{ds} \sim N_{n_d}(\mu_{ds}, V_{ds})$, $\mu_{ds} = X_{ds}\beta$, $V_{ds} = Z_{ds}V_{ud}Z'_{ds} + V_{ed,s}$. When the variance component parameters are known, the best linear unbiased estimator (BLUE) of β and the best linear unbiased predictor (BLUP) of u_d , $d = 1, \dots, D$, are

$$\hat{\beta}_B = (X'_s V_s^{-1} X_s)^{-1} X'_s V_s^{-1} y_s, \quad \hat{u}_{Bd} = V_{ud} Z'_{ds} V_{ds}^{-1} (y_{ds} - X_{ds} \hat{\beta}_B). \tag{4}$$

This paper estimates the model parameters by using the residual maximum likelihood (REML) method. See e.g. McCulloch et al. (2008) for a description of this method and Esteban et al. (2020) for the derivation of the updating equation of the Fisher-scoring algorithm that calculates the REML estimators of the BNER model. By substituting parameters by REML estimators in (4), the empirical BLUE and BLUP are obtained.

The out-of-sample subvectors y_{dr} follow the marginal models derived from the population model (2), that is,

$$y_{dr} = X_{dr}\beta + Z_{dr}u_d + e_{dr}, \quad d = 1, \dots, D,$$

where $u_d \sim N_2(0, V_{ud})$, $e_{dr} \sim N_{2(N_d - n_d)}(0, V_{ed,r})$ are independent and $V_{ed,r} = \text{diag}(V_{edj})_{n_d+1 \leq j \leq N_d}$. The vectors y_{dr} are independent with $y_{dr} \sim N_{N_d - n_d}(\mu_{dr}, V_{dr})$, $\mu_{dr} = X_{dr}\beta$, $V_{dr} = Z_{dr}V_{ud}Z'_{dr} + V_{ed,r}$. The covariance matrix between y_{dr} and y_{ds} is

$$V_{drs} = \text{cov}(y_{dr}, y_{ds}) = \text{cov}(X_{dr}\beta + Z_{dr}u_d + e_{dr}, X_{ds}\beta + Z_{ds}u_d + e_{ds}) = Z_{dr} \text{var}(u_d) Z'_{ds} = Z_{dr} V_{ud} Z'_{ds}.$$

The distribution of y_{dr} , given the sample data y_s , is

$$y_{dr}|y_s \sim y_{dr}|y_{ds} \sim N(\mu_{dr|s}, V_{dr|s}). \quad (5)$$

The conditional $(N_d - n_d) \times 1$ mean vector is

$$\begin{aligned} \mu_{dr|s} &= \mu_{dr} + V_{drs}V_{ds}^{-1}(y_{ds} - \mu_{ds}) = X_{dr}\beta + Z_{dr}V_{ud}Z'_{ds}V_{ds}^{-1}(y_{ds} - X_{ds}\beta) \\ &= X_{dr}\beta + Z_{dr}V_{ud}Z'_{ds} \left\{ V_{ed,s}^{-1} - V_{ed,s}^{-1}Z_{ds}(V_{ud}^{-1} + n_dV_{edj}^{-1})^{-1}Z'_{ds}V_{ed,s}^{-1} \right\} (y_{ds} - X_{ds}\beta). \end{aligned}$$

The conditional covariance matrix is

$$\begin{aligned} V_{dr|s} &= V_{dr} - V_{drs}V_{ds}^{-1}V_{dsr} = Z_{dr}V_{ud}Z'_{dr} + V_{ed,r} - Z_{dr}V_{ud}Z'_{ds}V_{ds}^{-1}Z_{ds}V_{ud}Z'_{dr} \\ &= Z_{dr}V_{ud}Z'_{dr} + V_{ed,r} - n_dZ_{dr}V_{ud}V_{edj}^{-1}V_{ud}Z'_{dr} + n_d^2Z_{dr}V_{ud}V_{edj}^{-1}(V_{ud}^{-1} + n_dV_{edj}^{-1})^{-1}V_{edj}^{-1}V_{ud}Z'_{dr}. \end{aligned}$$

If $n_d \neq 0$ and $j \in r_d = U_d - s_d, j > n_d$, the conditional 2×1 mean vector is

$$\mu_{dj|s} = X_{dj}\beta + V_{ud} \left\{ I_2 - n_dV_{edj}^{-1}(V_{ud}^{-1} + n_dV_{edj}^{-1})^{-1} \right\} \sum_{j=1}^{n_d} V_{edj}^{-1}(y_{dj} - X_{dj}\beta).$$

If $n_d = 0$ and $j \in r_d$, the conditional 2×1 mean vector is

$$\mu_{dj|s} = X_{dj}\beta.$$

If $n_d \neq 0$ and $j \in r_d, j > n_d$, the conditional 2×2 covariance matrix is

$$V_{dj|s} = V_{d|s} = V_{ud} + V_{edj} - n_dV_{ud}V_{edj}^{-1}V_{ud} + n_d^2V_{ud}V_{edj}^{-1}(V_{ud}^{-1} + n_dV_{edj}^{-1})^{-1}V_{edj}^{-1}V_{ud}.$$

If $n_d = 0$ and $j \in r_d$, the conditional 2×2 covariance matrix is

$$V_{dj|s} = V_{d|s} = V_{ud} + V_{edj}.$$

Appendix A of Data S1 gives an alternative derivation of the conditional distribution (5). More concretely, it shows the out-of-sample element $y_{dj} = (y_{dj1}, y_{dj2})'$, $j \in r_d$, conditioned to the sampled vector y_{ds} , has the representation

$$y_{dj} = X'_{dj}\beta + \tilde{u}_d + e_{dj}, \quad j \in r_d, \quad d = 1, \dots, D,$$

where $e_{dj} \sim N_2(0, V_{edj})$, $\tilde{u}_d \sim N_2(\tilde{\mu}_d, \tilde{V}_{uu})$ are all independent, with

$$\tilde{\mu}_d = V_{ud}(V_{ud} + n_d^{-1}V_{edj})^{-1}(\bar{y}_d - \bar{X}_d\beta), \quad \tilde{V}_{uu} = n_d^{-1}V_{ud}(V_{ud} + n_d^{-1}V_{edj})^{-1}V_{edj},$$

$$\bar{y}_d = n_d^{-1} \sum_{j=1}^{n_d} y_{dj}, \quad \bar{X}_d = n_d^{-1} \sum_{j=1}^{n_d} X_{dj}.$$

3 | EBPS OF DOMAIN PARAMETERS

3.1 | EBPs of additive parameters

Let $z_{dj} = (z_{dj1}, z_{dj2})'$ be a vector of continuous positive variables measured on the sample unit j of domain d , $d = 1, \dots, D$, $j = 1, \dots, n_d$. This section consider additive domain 2×1 or 1×1 parameters that can be written in the form

$$\delta_d = \frac{1}{N_d} \sum_{j=1}^{N_d} h(z_{dj}), \quad d = 1, \dots, D, \quad (6)$$

where h is a known measurable function $R^2 \mapsto R^t$, $t = 1, 2$. Examples of real-valued function $h : R^2 \mapsto R$ are $h(z_{dj}) = z_{dj1}$, $h(z_{dj}) = z_{dj2}$, and $h(z_{dj}) = z_{dj1} / (z_{dj1} + z_{dj2})$. The corresponding domain parameters (means of marginal variables or of unit-level ratios) are

$$\bar{Z}_{d1} = \frac{1}{N_d} \sum_{j=1}^{N_d} z_{dj1}, \quad \bar{Z}_{d2} = \frac{1}{N_d} \sum_{j=1}^{N_d} z_{dj2}, \quad A_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \frac{z_{dj1}}{z_{dj1} + z_{dj2}}, \quad d = 1, \dots, D. \quad (7)$$

In applications to real data z_{dj1} and z_{dj2} might not follow normal distributions, as may happens with expenditure variables that are typically asymmetric. This is why we assume that there exist a one-to-one transformation $g : R^2 \mapsto R^2$ such that $y_{dj} = g(z_{dj})$ follows the BNER model (1). We further assume that g is separable, i.e.

$$y_{dj} = g(z_{dj}) = (g_1(z_{dj1}), g_2(z_{dj2}))', \quad z_{dj} = g^{-1}(y_{dj}) = (g_1^{-1}(y_{dj1}), g_2^{-1}(y_{dj2}))',$$

where $g_1 : (0, \infty) \mapsto R$ and $g_2 : (0, \infty) \mapsto R$ are one-to-one functions. For $d = 1, \dots, D$, we write (6) and (7) as functions of y_{dj1} and y_{dj2} , that is,

$$\delta_d = \frac{1}{N_d} \sum_{j=1}^{N_d} h(g^{-1}(y_{dj})), \quad \bar{Z}_{d1} = \frac{1}{N_d} \sum_{j=1}^{N_d} g_1^{-1}(y_{dj1}),$$

$$\bar{Z}_{d2} = \frac{1}{N_d} \sum_{j=1}^{N_d} g_2^{-1}(y_{dj2}), \quad A_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \frac{g_1^{-1}(y_{dj1})}{g_1^{-1}(y_{dj1}) + g_2^{-1}(y_{dj2})}.$$

The best predictor (BP) of δ_d is

$$\hat{\delta}_d^B = E_{y_r} \left[\frac{1}{N_d} \sum_{j=1}^{N_d} h(g^{-1}(y_{dj})) \middle| y_s \right] = \frac{1}{N_d} \left\{ \sum_{j \in s_d} h(g^{-1}(y_{dj})) + \sum_{j \in r_d} E_{y_r} \left[h(g^{-1}(y_{dj})) \middle| y_s \right] \right\}.$$

The conditional distribution (5) depends on the vector $\psi = (\beta', \theta')'$ of unknown model parameters, which must be estimated, that is,

$$E_{y_r} \left[h(g^{-1}(y_{dj})) \middle| y_s \right] = E_{y_r} \left[h(g^{-1}(y_{dj})) \middle| y_s; \psi \right].$$

Let $\hat{\psi} = (\hat{\beta}', \hat{\theta}')'$ be an estimator based on sample data y_s . The EBP of δ_d is

$$\delta_d^{eb} = \frac{1}{N_d} \left\{ \sum_{j \in s_d} h(g^{-1}(y_{dj})) + \sum_{j \in r_d} E_{y_r} \left[h(g^{-1}(y_{dj})) \mid y_s; \hat{\psi} \right] \right\}.$$

For a general function h , the expected value above might be not tractable analytically. When this occurs, the following Monte Carlo procedure can be applied.

- Estimate the unknown parameter $\psi = (\beta', \theta')'$ using sample data (y_s, X_s) .
- Replacing $\psi = (\beta', \theta')'$ by the estimate $\hat{\psi} = (\hat{\beta}', \hat{\theta}')'$ obtained in (a), draw L copies of each nonsample variable y_{dj} as

$$y_{dj}^{(\ell)} \sim N_2(\hat{\mu}_{dj|s}, \hat{V}_{dj|s}), \quad j \in r_d, \quad d = 1, \dots, D, \quad \ell = 1, \dots, L.$$

where

$$\hat{\mu}_{dj|s} = \begin{cases} X_{dj}\hat{\beta} + \hat{V}_{ud}Z'_{ds} \left\{ \hat{V}_{eds}^{-1} - \hat{V}_{eds}^{-1}Z_{ds} \left(\hat{V}_{ud}^{-1} + n_d \hat{V}_{edj}^{-1} \right)^{-1} Z'_{ds} \hat{V}_{eds}^{-1} \right\} (y_{ds} - X_{ds}\hat{\beta}) & \text{if } n_d \neq 0, \\ X_{dj}\hat{\beta} & \text{if } n_d = 0, \end{cases}$$

and

$$\hat{V}_{dj|s} = \begin{cases} \hat{V}_{ud} + \hat{V}_{edj} - n_d \hat{V}_{ud} \hat{V}_{edj}^{-1} \hat{V}_{ud} + n_d^2 \hat{V}_{ud} \hat{V}_{edj}^{-1} \left(\hat{V}_{ud}^{-1} + n_d \hat{V}_{edj}^{-1} \right)^{-1} \hat{V}_{edj}^{-1} \hat{V}_{ud} & \text{if } n_d \neq 0, \\ \hat{V}_{ud} + \hat{V}_{edj} & \text{if } n_d = 0. \end{cases} \quad (8)$$

- The Monte Carlo approximation of the expected value is

$$E_{y_r} \left[h(g^{-1}(y_{dj})) \mid y_s; \hat{\psi} \right] \approx \frac{1}{L} \sum_{\ell=1}^L h \left(g^{-1}(y_{dj}^{(\ell)}) \right), \quad j \in r_d, \quad d = 1, \dots, D.$$

The Monte Carlo approximation of the EBP of δ_d is

$$\hat{\delta}_d^{eb} \approx \frac{1}{L} \sum_{\ell=1}^L \delta_d^{(\ell)}, \quad \delta_d^{(\ell)} = \frac{1}{N_d} \left(\sum_{j \in s_d} h(g^{-1}(y_{dj})) + \sum_{j \in r_d} h \left(g^{-1}(y_{dj}^{(\ell)}) \right) \right). \quad (9)$$

Remark 1. In many practical cases the values of the auxiliary variables are not available for all the population units. If in addition some of the variables are continuous, the EBP method is not applicable. An important particular case, where this method is applicable, is when the number of values of the vector of auxiliary variables is finite. More concretely, suppose that the covariates are categorical such that $X_{dj} \in \{X_{01}, \dots, X_{0T}\}$, then we can calculate $\delta_d^{(\ell)}$ as

$$\delta_d^{(\ell)} = \frac{1}{N_d} \left[\sum_{j=1}^{n_d} h(g^{-1}(y_{dj})) + \sum_{t=1}^T \sum_{j=1}^{N_{dt}-n_{dt}} h \left(g^{-1}(y_{dj}^{(\ell)}) \right) \right], \quad (10)$$

where $N_{dt} = \#\{j \in U_d : X_{dj} = X_{0t}\}$ is available from external data sources (aggregated auxiliary information), $n_{dt} = \#\{j \in s_d : X_{dj} = X_{0t}\}$, $y_{dj}^{(\ell)} \sim N_2(\hat{\mu}_{dt|s}, \hat{V}_{dt|s})$, $d = 1, \dots, D$, $j = 1, \dots, N_{dt} - n_{dt}$, $t =$

$1, \dots, T, \ell = 1, \dots, L$, where

$$\hat{\mu}_{dtls} = \begin{cases} X_{0t}\hat{\beta} + \hat{V}_{ud}Z'_{ds} \left\{ \hat{V}_{eds}^{-1} - \hat{V}_{eds}^{-1}Z_{ds} \left(\hat{V}_{ud}^{-1} + n_d \hat{V}_{edj}^{-1} \right)^{-1} Z'_{ds} \hat{V}_{eds}^{-1} \right\} (y_{ds} - X_{ds}\hat{\beta}) & \text{if } n_d \neq 0, \\ X_{0t}\hat{\beta} & \text{if } n_d = 0, \end{cases} \quad (11)$$

and \hat{V}_{dls} was defined in (8).

3.2 | EBPs of nonadditive parameters

Let $z_{dj} = (z_{dj1}, z_{dj2})'$ be a vector of continuous positive variables measured on the sample unit j of domain d , $d = 1, \dots, D, j = 1, \dots, n_d$. Define $z_d = \text{col}_{1 \leq j \leq n_d} (z_{dj})$. This section consider domain 2×1 or 1×1 parameters that can be written in the form

$$\delta_d = h(z_d), \quad (12)$$

where h is a known measurable function $R^{2N_d} \mapsto R^t, t = 1, 2$. A domain parameter is the ratio

$$R_d = \frac{\bar{Z}_{d1}}{\bar{Z}_{d1} + \bar{Z}_{d2}}, \quad \text{where } h(z_d) = \frac{\sum_{j=1}^{N_d} z_{dj1}}{\sum_{j=1}^{N_d} (z_{dj1} + z_{dj2})}. \quad (13)$$

As in Section 3.1, we assume that there exist a one-to-one transformation $g : R^2 \mapsto R^2$ such that $y_{dj} = g(z_{dj})$ follows the BNER model (1). We further assume that g is separable, that is

$$y_{dj} = g(z_{dj}) = (g_1(z_{dj1}), g_2(z_{dj2}))', \quad z_{dj} = g^{-1}(y_{dj}) = (g_1^{-1}(y_{dj1}), g_2^{-1}(y_{dj2}))',$$

where $g_1 : (0, \infty) \mapsto R$ and $g_2 : (0, \infty) \mapsto R$ are one-to-one functions. For ease of notation, we define the $2n_d \times 1$ vectors

$$z_d = g^{-1}(y_d) = \text{col}_{1 \leq j \leq n_d} (g^{-1}(y_{dj})), \quad d = 1, \dots, D.$$

We can write (12) and (13) as functions of y_{dj1} and y_{dj2} , i.e.

$$\delta_d = h(g^{-1}(y_d)), \quad R_d = \frac{\sum_{j=1}^{N_d} g_1^{-1}(y_{dj1})}{\sum_{j=1}^{N_d} (g_1^{-1}(y_{dj1}) + g_2^{-1}(y_{dj2}))}.$$

The BP of δ_d is

$$\delta_d^B = E_{y_r} [h(g^{-1}(y_d)) | y_s].$$

The conditional distribution (5) depends on the vector $\psi = (\beta', \theta')'$ of unknown model parameters, which must be estimated, that is,

$$E_{y_r} [h(g^{-1}(y_d)) | y_s] = E_{y_r} [h(g^{-1}(y_d)) | y_s; \psi].$$

Let $\hat{\psi} = (\hat{\beta}', \hat{\theta}')'$ be an estimator based on sample data y_s . The EBP of δ_d is

$$\hat{\delta}_d^{eb} = E_{y_r} [h(g^{-1}(y_d)) | y_s; \hat{\psi}].$$

For a general function h , the expected value above might be not tractable analytically. When this occurs, we can apply a following Monte Carlo procedure with the same steps (a) and (b) of Section 3.1 and with the following new steps

(c) Construct the vectors

$$y_{dr}^{(\ell)} = \text{col}_{j \in r_d} (y_{dj}^{(\ell)}), \quad y_{ds}^{(\ell)} = \text{col}_{j \in s_d} (y_{dj}^{(\ell)}), \quad y_d^{(\ell)} = (y_{ds}^{(\ell)}, y_{dr}^{(\ell)})'.$$

(d) The Monte Carlo approximation of the EBP of δ_d is

$$\hat{\delta}_d^{eb} \approx \frac{1}{L} \sum_{\ell=1}^L h(g^{-1}(y_d^{(\ell)})), \quad d = 1, \dots, D.$$

Remark 2. Under the categorical covariate setup of Remark 1, we can write the elements of y as $y_{dij} = (y_{dij1}, y_{dij2})'$, where d, i and j denote domain, category and individual, respectively. We approximate \hat{R}_d^{eb} as

$$\hat{R}_d^{eb} = \frac{1}{L} \sum_{\ell=1}^L \frac{\sum_{j=1}^{n_d} g_1^{-1}(y_{dj1}) + \sum_{t=1}^T \sum_{j=1}^{N_{dt}-n_{dt}} g_1^{-1}(y_{dtj1}^{(\ell)})}{\sum_{j=1}^{n_d} (g_1^{-1}(y_{dj1}) + g_2^{-1}(y_{dj2})) + \sum_{t=1}^T \sum_{j=1}^{N_{dt}-n_{dt}} (g_1^{-1}(y_{dtj1}^{(\ell)}) + g_2^{-1}(y_{dtj2}^{(\ell)}))}, \quad (14)$$

where $N_{dt} = \#\{j \in U_d : X_{dj} = X_{0t}\}$ is available from external data sources (aggregated auxiliary information), $n_{dt} = \#\{j \in s_d : X_{dj} = X_{0t}\}$, $y_{dij}^{(\ell)} = (y_{dij1}^{(\ell)}, y_{dij2}^{(\ell)})' \sim N_2(\hat{\mu}_{dt|s}, \hat{V}_{dt|s})$, $d = 1, \dots, D$, $j = 1, \dots, N_{dt} - n_{dt}$, $t = 1, \dots, T$, $\ell = 1, \dots, L$, where $\hat{\mu}_{dt|s}$ and $\hat{V}_{dt|s}$ were defined in (11) and (8) respectively.

4 | PARAMETRIC BOOTSTRAP MSE ESTIMATOR

Analytical approximations to the MSE are difficult to derive in the case of complex parameters. We therefore propose a parametric bootstrap MSE estimator by following the bootstrap method for finite populations of González-Manteiga et al. (2007, 2008). We present the case of additive domain parameters. The modifications to deal with nonadditive parameters are straightforward. The steps for implementing this method are

1. Fit the model (1) to sample data (y_s, X_s) and calculate an estimator $\hat{\phi} = (\hat{\beta}', \hat{\theta}')'$ of $\psi = (\beta', \theta')'$.
2. For $d = 1, \dots, D, j = 1, \dots, N_d$, generate independently $u_d^* \sim N(0, \hat{V}_{ud})$ and $e_{dj}^* \sim N(0, \hat{V}_{edj})$.
3. Construct the bootstrap superpopulation model ξ^* using $\{u_d^*\}$, $\{e_{dj}^*\}$, $\{X_{dj}\}$ and $\hat{\phi}$, that is,

$$\xi^* : y_{dj}^* = X_{dj} \hat{\phi} + u_d^* + e_{dj}^*, \quad d = 1, \dots, D, j = 1, \dots, N_d. \quad (15)$$

4. Under the bootstrap superpopulation model (15), generate a large number B of i.i.d. bootstrap populations $\{y_{dj}^{*(b)} : d = 1, \dots, D, j = 1, \dots, N_d\}$ and calculate the bootstrap population parameters

$$\delta_d^{*(b)} = \frac{1}{N_d} \sum_{j=1}^{N_d} h\left(g^{-1}(y_{dj}^{*(b)})\right), \quad b = 1, \dots, B.$$

5. From each bootstrap population b generated in Step 4, take the sample with the same indices $s \subset U$ as the initial sample, and calculate the bootstrap EBPs, $\hat{\delta}_d^{eb*(b)}$, as described in Section 3.1 using the bootstrap sample data y_s^* and the known values X_{dj} .
6. A Monte Carlo approximation to the theoretical bootstrap estimator

$$\text{MSE}_* \left(\hat{\delta}_d^{eb*} \right) = E_{\xi^*} \left[\left(\hat{\delta}_d^{eb*} - \delta_d^* \right) \left(\hat{\delta}_d^{eb*} - \delta_d^* \right)' \right]$$

is

$$\text{mse}_* \left(\hat{\delta}_d^{eb*} \right) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\delta}_d^{eb*(b)} - \delta_d^{*(b)} \right) \left(\hat{\delta}_d^{eb*(b)} - \delta_d^{*(b)} \right)'. \quad (16)$$

The estimator (16) is used to estimate $\text{MSE}(\hat{\delta}_d^{eb})$.

Hall and Maiti (2006a, 2006b) and Erculescu and Fuller (2016, 2018) derived parametric double-bootstrap algorithms for estimating the MSE of predictors of “model-based” small area parameters. Their domain parameters of interest are functions of elements of the assumed population model. Although these approaches are asymptotically more efficient, the corresponding methodologies are not directly applicable to domains parameters of the form (6) or (12), which are functions of z_d . The aforementioned methods could be adapted to the EBPs of additive and non-additive parameters, but they may still have the drawback of computational cost in many of the applications to real data, where the population sizes of the domains, N_d , are very large. This is the case of the SHBS, treated in Section 6. Nevertheless, the fast double bootstrap approach of Erculescu and Fuller (2016) is not computationally intensive, as it requires one sample at the second level. The adaptation of their approach to bivariate NER models is thus an interesting alternative.

5 | SIMULATIONS

This section presents simulation experiments for investigating the EBPs and the MSE estimators. We carry out the simulations with a correlation structure similar to that of the application to real data. This is to say, with positive correlation parameters ρ_{u12} and ρ_{e12} . We generate artificial population data as follows. Take $p_1 = p_2 = 2$, $p = 4$, $\beta_1 = (\beta_{11}, \beta_{12})' = (10, 10)'$, $\beta_2 = (\beta_{21}, \beta_{22})' = (10, 10)'$. For $k = 1, 2, d = 1, \dots, D, j = 1, \dots, n_d$, generate $X_{dj} = \text{diag}(x_{dj1}, x_{dj2})_{2 \times 4}$, where $x_{dj1} = (x_{dj11}, x_{dj12})$, $x_{dj2} = (x_{dj21}, x_{dj22})$, $x_{dj11} = x_{dj21} = 1$, $x_{dj12} \sim \text{Bin}(1, 1/2)$, $x_{dj22} \sim \text{Bin}(1, 1/2)$. For $d = 1, \dots, D, j = 1, \dots, N_d$, simulate $u_d \sim N_2(0, V_{ud})$ and $e_{dj} \sim N_2(0, V_{edj})$, where

$$V_{ud} = \begin{pmatrix} \theta_1 & \theta_3 \sqrt{\theta_1} \sqrt{\theta_2} \\ \theta_3 \sqrt{\theta_1} \sqrt{\theta_2} & \theta_2 \end{pmatrix}, \quad V_{edj} = \begin{pmatrix} \theta_4 & \theta_6 \sqrt{\theta_4} \sqrt{\theta_5} \\ \theta_6 \sqrt{\theta_4} \sqrt{\theta_5} & \theta_5 \end{pmatrix},$$

with $\theta_1 = 0.50$, $\theta_2 = 0.75$, $\theta_4 = 0.75$, $\theta_5 = 1.00$ and $\theta_3 = 0.6$, $\theta_6 = 0.4$. The simulations generate only four different matrices X_{dj} . They are

$$X_{dj} = \left(\begin{array}{cc|cc} x_{dj11} & x_{dj12} & 0 & 0 \\ 0 & 0 & x_{dj21} & x_{dj22} \end{array} \right) \in \{X_{01}, X_{02}, X_{03}, X_{04}\},$$

where

$$X_{01} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array} \right), X_{02} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array} \right),$$

$$X_{03} = \left(\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array} \right), X_{04} = \left(\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array} \right).$$

The simulations apply the same transformations as in the application to real data, that is, $g_1(z_{dj1}) = \log z_{dj1}$ and $g_2(z_{dj2}) = \log z_{dj2}$, so that $z_{dj1} = \exp\{y_{dj1}\}$ and $z_{dj2} = \exp\{y_{dj2}\}$, $d = 1, \dots, D$, $j = 1, \dots, N_d$. We apply the function `rmvnorm` of the R package `mvtnorm` for generating multivariate normal vectors.

5.1 | Simulation 1

The target of Simulation 1 is to investigate the behavior of the EBPs, \hat{A}_d^{eb} and \hat{R}_d^{eb} , based on the BNER model. For this sake, we carry out a simulation experiment with $I = 200$ Monte Carlo iterations. We further take, $L = 200$ and $N_d = 200$, $d = 1, \dots, D$. The steps of Simulation 1 are

1. Generate x_{dj1} , $d = 1, \dots, D$, $j = 1, \dots, N_d$, $k = 1, 2$. Construct the population matrices X_d and Z_d of dimensions $2N_d \times p$ and $2N_d \times 2$ respectively. For $d = 1, \dots, D$, $t = 1, \dots, T$, $T = 4$, calculate

$$N_{dt} = \# \{j \in U_d : X_j = X_{0t}\} = \# \{j \in I : j \leq N_d, X_j = X_{0t}\},$$

$$n_{dt} = \# \{j \in S_d : X_j = X_{0t}\} = \# \{j \in I : j \leq n_d, X_j = X_{0t}\}.$$

2. Repeat $I = 200$ times ($i = 1, \dots, 200$)

- 2.1. Generate the populations random vectors $u_d^{(i)} \sim N_2(0, V_{ud})$, $e_d^{(i)} \sim N_{2N_d}(0, V_{ed})$, $y_d^{(i)} = X_d \beta + Z_d u_d^{(i)} + e_d^{(i)}$, where $V_{ed} = \text{diag}(V_{edj})$, $d = 1, \dots, D$. Calculate $z_{dj1}^{(i)} = \exp\{y_{dj1}^{(i)}\}$, $z_{dj2}^{(i)} = \exp\{y_{dj2}^{(i)}\}$, $d = 1, \dots, D$, $j = 1, \dots, N_d$.

- 2.2. Calculate the domain ratio parameters, that is,

$$A_d^{(i)} = \frac{1}{N_d} \sum_{j=1}^{N_d} \frac{z_{dj1}^{(i)}}{z_{dj1}^{(i)} + z_{dj2}^{(i)}}, \quad R_d^{(i)} = \frac{\sum_{j=1}^{N_d} z_{dj1}^{(i)}}{\sum_{j=1}^{N_d} z_{dj1}^{(i)} + \sum_{j=1}^{N_d} z_{dj2}^{(i)}}, \quad d = 1, \dots, D.$$

- 2.3. Extract the sample $(y_{dj}^{(i)}, X_{dj})$, $d = 1, \dots, D$, $j = 1, \dots, n_d$, with $n_d \in \{3, 5, 10, 25, 50, 100\}$.

- 2.4. Calculate the REML estimators $\hat{\beta}_{11}^{(i)}, \hat{\beta}_{12}^{(i)}, \hat{\beta}_{21}^{(i)}, \hat{\beta}_{22}^{(i)}, \hat{\theta}_1^{(i)}, \dots, \hat{\theta}_6^{(i)}$.

2.5. For $d = 1, \dots, D, t = 1, \dots, T$, calculate

$$\hat{\mu}_{dt|s}^{(i)} = X_{0t}\hat{\beta}^{(i)} + \hat{V}_{ud}^{(i)}Z'_{ds} \left\{ \hat{V}_{eds}^{(i-1)} - \hat{V}_{eds}^{(i-1)}Z_{ds} \left(\hat{V}_{ud}^{(i-1)} + n_d \hat{V}_{edj}^{(i-1)} \right)^{-1} Z'_{ds} \hat{V}_{eds}^{(i-1)} \right\} \left(y_{ds} - X_{ds}\hat{\beta}^{(i)} \right),$$

$$\hat{V}_{d|s}^{(i)} = \hat{V}_{ud}^{(i)} + \hat{V}_{edj}^{(i)} - n_d \hat{V}_{ud}^{(i)} \hat{V}_{edj}^{(i-1)} \hat{V}_{ud}^{(i)} + n_d^2 \hat{V}_{ud}^{(i)} \hat{V}_{edj}^{(i-1)} \left(\hat{V}_{ud}^{(i-1)} + n_d \hat{V}_{edj}^{(i-1)} \right)^{-1} \hat{V}_{edj}^{(i-1)} \hat{V}_{ud}^{(i)}.$$

2.6. For $d = 1, \dots, D, j = 1, \dots, N_{dt} - n_{dt}, t = 1, \dots, T, \ell = 1, \dots, L$, generate

$$y_{dij}^{(i\ell)} = \left(y_{dij1}^{(i\ell)}, y_{dij2}^{(i\ell)} \right)' \sim N_2 \left(\hat{\mu}_{dt|s}^{(i)}, \hat{V}_{d|s}^{(i)} \right),$$

and calculate $z_{dij1}^{(i\ell)} = \exp \left\{ y_{dij1}^{(i\ell)} \right\}, z_{dij2}^{(i\ell)} = \exp \left\{ y_{dij2}^{(i\ell)} \right\}$.

2.7. For $d = 1, \dots, D$, calculate the EBPs $\hat{A}_d^{eb(i)}$ and $\hat{R}_d^{eb(i)}$, that is

$$\hat{A}_d^{eb(i)} = \frac{1}{LN_d} \sum_{\ell=1}^L \left[\sum_{j=1}^{n_d} \frac{z_{dij1}^{(i)}}{z_{dij1}^{(i)} + z_{dij2}^{(i)}} + \sum_{t=1}^T \sum_{j=1}^{N_{dt}-n_{dt}} \frac{z_{dij1}^{(i\ell)}}{z_{dij1}^{(i\ell)} + z_{dij2}^{(i\ell)}} \right], \tag{17}$$

$$\hat{R}_d^{eb(i)} = \frac{1}{L} \sum_{\ell=1}^L \frac{\sum_{j=1}^{n_d} z_{dij1}^{(i)} + \sum_{t=1}^T \sum_{j=1}^{N_{dt}-n_{dt}} z_{dij1}^{(i\ell)}}{\sum_{j=1}^{n_d} \left(z_{dij1}^{(i)} + z_{dij2}^{(i)} \right) + \sum_{t=1}^T \sum_{j=1}^{N_{dt}-n_{dt}} \left(z_{dij1}^{(i\ell)} + z_{dij2}^{(i\ell)} \right)}. \tag{18}$$

3. For $\hat{\eta}_d^{(i)} \in \left\{ \hat{A}_d^{eb(i)}, \hat{R}_d^{eb(i)} \right\}, \eta_d^{(i)} \in \left\{ A_d^{(i)}, R_d^{(i)} \right\}, d = 1, \dots, D$, calculate

$$B_d(\eta) = \frac{1}{I} \sum_{i=1}^I \left(\hat{\eta}_d^{eb(i)} - \eta_d^{(i)} \right), \quad RE_d(\eta) = \left(\frac{1}{I} \sum_{i=1}^I \left(\hat{\eta}_d^{eb(i)} - \eta_d^{(i)} \right)^2 \right)^{1/2}, \quad \eta_d = \frac{1}{I} \sum_{i=1}^I \eta_d^{(i)},$$

$$RB_d(\eta) = \frac{B_d(\eta)}{\eta_d} 100, \quad RRE_d(\eta) = \frac{RE_d(\eta)}{\eta_d} 100, \quad AB(\eta) = \frac{1}{D} \sum_{d=1}^D |B_d(\eta)|,$$

$$RE(\eta) = \frac{1}{D} \sum_{d=1}^D RE_d(\eta), \quad RAB = \frac{1}{D} \sum_{d=1}^D |RB_d(\eta)|, \quad RRE = \frac{1}{D} \sum_{d=1}^D RRE_d(\eta).$$

Simulation 1 takes sample sizes depending (variable) and not depending (constant) on d . The considered constant sample sizes are $n_d = 3, 5, 10, 25, 50, 100, d = 1, \dots, D$. In the case of variable sample sizes, the n_d 's are drawn at random from the set $\{2, 3, \dots, 20\}$. Simulation 1 makes this selection before starting the loops, so that the samples sizes are the same in all the iterations. This case is called $n_d = 10_*$. Tables 1, 2, 3, and 4 present the average absolute and relative performances measures for $D = 25, 50, 100, 200$. We observe that the EBPs are basically unbiased and that the MSEs decrease as the sample sizes n_d increase. These results indicate that the optimality properties of the BPs are inherited by the EBPs. The performance measures remain stable as the number of domains D increases, with small fluctuations due to the number of Monte Carlo iterations ($I = 200$). This is somehow expected because the rate between the number of observations and the number of domains quantities, A_d and $R_d, d = 1, \dots, D$, to predict remains constant as the number of domains D increases.

For studying the effect of deviations from normality, we change the multivariate normal distributions of $u_d^{(i)}$ and $e_d^{(i)}$ in step 2.1 by multivariate skew-normal and Student's t distributions,

TABLE 1 AB(η) with $N_d = 200$

D	η	$n_d = 3$	$n_d = 5$	$n_d = 10_*$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
25	\hat{A}^{eb}	0.0030	0.0021	0.0015	0.0015	0.0012	0.0008	0.0004
	\hat{R}^{eb}	0.0083	0.0066	0.0045	0.0046	0.0031	0.0020	0.0015
50	\hat{A}^{eb}	0.0029	0.0023	0.0016	0.0019	0.0011	0.0009	0.0005
	\hat{R}^{eb}	0.0145	0.0072	0.0049	0.0047	0.0029	0.0023	0.0015
100	\hat{A}^{eb}	0.0022	0.0019	0.0018	0.0017	0.0011	0.0008	0.0005
	\hat{R}^{eb}	0.0088	0.0065	0.0060	0.0052	0.0031	0.0023	0.0014
200	\hat{A}^{eb}	0.0024	0.0023	0.0017	0.0016	0.0010	0.0008	0.0005
	\hat{R}^{eb}	0.0086	0.0060	0.0057	0.0049	0.0028	0.0022	0.0015

TABLE 2 RE(η) with $N_d = 200$

D	η	$n_d = 3$	$n_d = 5$	$n_d = 10_*$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
25	\hat{A}^{eb}	0.0475	0.0390	0.0311	0.0292	0.0193	0.0134	0.0087
	\hat{R}^{eb}	0.1215	0.1004	0.0793	0.0748	0.0520	0.0381	0.0273
50	\hat{A}^{eb}	0.0461	0.0389	0.0315	0.0295	0.0196	0.0136	0.0087
	\hat{R}^{eb}	0.1146	0.0957	0.0785	0.0736	0.0511	0.0384	0.0274
100	\hat{A}^{eb}	0.0455	0.0380	0.0310	0.0293	0.0193	0.0136	0.0086
	\hat{R}^{eb}	0.1098	0.0928	0.0771	0.0728	0.0503	0.0386	0.0269
200	\hat{A}^{eb}	0.0449	0.0377	0.0305	0.0291	0.0192	0.0135	0.0087
	\hat{R}^{eb}	0.1087	0.0916	0.0751	0.0718	0.0509	0.0383	0.0269

TABLE 3 RAB(η) in %, with $N_d = 200$

D	η	$n_d = 3$	$n_d = 5$	$n_d = 10_*$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
25	\hat{A}^{eb}	0.5930	0.4312	0.2942	0.2903	0.2515	0.1715	0.0854
	\hat{R}^{eb}	1.7489	1.4044	0.9409	0.9618	0.6658	0.4195	0.3308
50	\hat{A}^{eb}	0.5925	0.4515	0.3200	0.3821	0.2172	0.1825	0.0979
	\hat{R}^{eb}	3.0777	1.5197	1.0194	0.9856	0.6090	0.4865	0.3277
100	\hat{A}^{eb}	0.4510	0.3891	0.3601	0.3357	0.2263	0.1542	0.1012
	\hat{R}^{eb}	1.8718	1.3810	1.2520	1.0983	0.6520	0.4901	0.3021
200	\hat{A}^{eb}	0.4810	0.4675	0.3458	0.3198	0.1980	0.1643	0.0962
	\hat{R}^{eb}	1.8249	1.2685	1.2065	1.0427	0.5874	0.4722	0.3249

TABLE 4 RRE(η) in %, with $N_d = 200$

D	η	$n_d = 3$	$n_d = 5$	$n_d = 10_*$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
25	\hat{A}^{eb}	9.6226	7.8797	6.2522	5.8851	3.8985	2.7157	1.7634
	\hat{R}^{eb}	25.9849	21.3923	16.8070	15.8288	11.0848	8.1627	5.8238
50	\hat{A}^{eb}	9.3201	7.8455	6.3447	5.9508	3.9476	2.7463	1.7532
	\hat{R}^{eb}	24.4787	20.3999	16.6696	15.6100	10.9063	8.1760	5.8287
100	\hat{A}^{eb}	9.1647	7.6393	6.2438	5.9145	3.8943	2.7373	1.7372
	\hat{R}^{eb}	23.3933	19.6736	16.3651	15.5579	10.7073	8.2119	5.7284
200	\hat{A}^{eb}	9.0275	7.5733	6.1196	5.8423	3.8611	2.7210	1.7465
	\hat{R}^{eb}	23.1056	19.4270	15.9273	15.2669	10.8071	8.1225	5.7151

TABLE 5 Empirical skewness and kurtosis, with $I = 2 \cdot 10^6$ replicates

Measure	v_i	Skew-normal, $\alpha_1 = 0$				Student's t		
		$\alpha_2 = 0$	$\alpha_2 = 2$	$\alpha_2 = 5$	$\alpha_2 = 10$	$df = 15$	$df = 10$	$df = 5$
Skewness	v_{i1}	0.0000	0.0497	0.0575	0.0591	0.0004	-0.0025	0.0192
	v_{i2}	0.0000	0.4445	0.8272	0.9337	0.0002	-0.0051	-0.0245
Kurtosis	v_{i1}	0.0000	-0.0313	-0.0259	-0.0249	0.5519	1.0226	6.2916
	v_{i2}	0.0000	0.2472	0.5656	0.6778	0.5433	0.9982	6.6255

keeping the same mean and variance component parameters. The simulations of skew-normal and Student's t random vectors are carried out with the functions `rMSN` and `rmvt` of the R packages `sn` and `mvtnorm`, respectively. We generate bivariate skew-normal random vectors with skewness parameters $(\alpha_1, \alpha_2) = (0, 0), (0, 2), (0, 5), (0, 10)$ and bivariate Student's t vectors with degrees of freedom $\nu = 15, 10, 5$. We recall that $(\alpha_1, \alpha_2) = (0, 0)$ yields to the multivariate normal distribution, which is now simulated from function `rMSN` of R package `sn` and not from `rmvnorm` of `mvtnorm`, as before. Table 5 presents the empirical skewness and kurtosis calculated from simulated random vectors $v_i = (v_{i1}, v_{i2}), i = 1, \dots, 2 \cdot 10^6$. For the sake of comparison, we include the case of normality $(\alpha_1, \alpha_2) = (0, 0)$ and its theoretical measures.

Table 6 presents the relative performance measures for $D = 50, n_d = 10, N_d = 200$. We observe that relative biases and root-MSEs have a moderate increment as the skewness parameter increases. However, these measures are more sensible to the decrease of the degrees of freedom of the multivariate Student's t distribution. For the nonlinear parameter R_d , the increase in RRE is more noticeable; that is, the increase of kurtosis makes EBP less efficient and therefore we should be more cautious when applying it.

5.2 | Simulation 2

For investigating the behavior of the bootstrap-based MSE estimators of predictors \hat{A}_d^{eb} and \hat{R}_d^{eb} , we take $I = 200, L = 200, N_d = 200, D = 50$ and $n_d = 10, d = 1, \dots, D$. A second objective of simulation

TABLE 6 RAB(η) and RRE(η) with $D = 50$, $n_d = 10$, $N_d = 200$

	η	Skew-normal, $\alpha_1 = 0$				Student's t		
		$\alpha_2 = 0$	$\alpha_2 = 2$	$\alpha_2 = 5$	$\alpha_2 = 10$	$df = 15$	$df = 10$	$df = 5$
RAB	\hat{A}^{eb}	0.3035	0.3055	0.3278	0.3419	0.3018	0.3222	0.3636
	\hat{R}^{eb}	1.0152	1.1382	1.5557	1.6195	0.9955	1.0379	2.1487
RRE	\hat{A}^{eb}	5.8948	5.8620	5.7729	5.7572	6.2152	6.3045	6.9143
	\hat{R}^{eb}	15.4235	15.4207	15.1357	15.0809	18.3089	20.6769	37.5347

2 is to give recommendations on how many bootstrap replicates B should be applied so that the MSE estimator is acceptably accurate. The steps of Simulation 2 are

1. For $D = 50$, generate x_{djk} , $d = 1, \dots, D$, $j = 1, \dots, N_d$, $k = 1, 2$. Construct the population matrices X_{dj} of dimensions $2 \times p$.
2. Take $MSE_d(\eta) = RE_d(\eta)^2$, $\eta \in \{A_d, R_d\}$, $d = 1, \dots, D$, from the output of Simulation 1.
3. Repeat $I = 200$ times ($i = 1, \dots, 200$)

3.1. Generate the populations random vectors $u_d^{(i)} \sim N_2(0, V_{ud})$, $e_{dj}^{(i)} \sim N_2(0, V_{edj})$ and $y_{dj}^{(i)} = X_{dj}\beta + u_d^{(i)} + e_{dj}^{(i)}$, $d = 1, \dots, D$, $j = 1, \dots, N_d$ ($N_d = 200$). Calculate $z_{dj1}^{(i)} = \exp\{y_{dj1}^{(i)}\}$, $z_{dj2}^{(i)} = \exp\{y_{dj2}^{(i)}\}$, $d = 1, \dots, D$, $j = 1, \dots, N_d$.

3.2. Extract the sample (y_{dj}, X_{dj}) , $d = 1, \dots, D$, $j = 1, \dots, n_d$ ($n_d = 10$).

3.3. Calculate the REML estimators $\hat{\beta}_{11}^{(i)}, \hat{\beta}_{12}^{(i)}, \hat{\beta}_{21}^{(i)}, \hat{\beta}_{22}^{(i)}, \hat{\theta}_1^{(i)}, \dots, \hat{\theta}_6^{(i)}$ and the EBPs $\hat{A}_d^{eb(i)}, \hat{R}_d^{eb(i)}$, $d = 1, \dots, D$.

3.4. Repeat B times, $B = \{50, 100, 200, 400\}$ ($b = 1, \dots, B$).

(a) Generate the bootstrap population vectors $u_d^{*(ib)} \sim N_2(0, \hat{V}_{ud})$, $e_{dj}^{*(ib)} \sim N_2(0, \hat{V}_{edj})$,

$$y_{dj}^{*(ib)} = X_{dj}\hat{\beta}^{(i)} + u_d^{*(ib)} + e_{dj}^{*(ib)}, \quad d = 1, \dots, D, \quad j = 1, \dots, N_d,$$

where $\hat{V}_{ud} = V_{ud}(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ and $\hat{V}_{edj} = V_{edj}(\hat{\theta}_4, \hat{\theta}_5, \hat{\theta}_6)$. Calculate $z_{dj1}^{*(ib)} = \exp\{y_{dj1}^{*(ib)}\}$, $z_{dj2}^{*(ib)} = \exp\{y_{dj2}^{*(ib)}\}$, $d = 1, \dots, D$, $j = 1, \dots, N_d$.

(b) Calculate the bootstrap domain ratio parameters, that is,

$$A_d^{*(ib)} = \frac{1}{N_d} \sum_{j=1}^{N_d} \frac{z_{dj1}^{*(ib)}}{z_{dj1}^{*(ib)} + z_{dj2}^{*(ib)}}, \quad R_d^{*(ib)} = \frac{\sum_{j=1}^{N_d} z_{dj1}^{*(ib)}}{\sum_{j=1}^{N_d} z_{dj1}^{*(ib)} + \sum_{j=1}^{N_d} z_{dj2}^{*(ib)}}, \quad d = 1, \dots, D.$$

(c) Extract the bootstrap sample $(y_{dj}^{*(ib)}, X_{dj})$, $d = 1, \dots, D$, $j = 1, \dots, n_d$.

(d) Calculate the bootstrap REML estimators $\hat{\beta}_{11}^{*(ib)}, \hat{\beta}_{12}^{*(ib)}, \hat{\beta}_{21}^{*(ib)}, \hat{\beta}_{22}^{*(ib)}, \hat{\theta}_1^{*(ib)}, \dots, \hat{\theta}_6^{*(ib)}$.

(e) Calculate the EBPs $\hat{A}_d^{eb*(ib)}, \hat{R}_d^{eb*(ib)}$ and $\hat{R}_d^{in*(ib)}$, $d = 1, \dots, D$, as in (17) and (18), with $L = 200$.

3.5. For $\hat{\eta}_d^{*(ib)} \in \{A_d^{eb*(ib)}, \hat{R}_d^{eb*(ib)}\}, \eta_d^{*(ib)} \in \{A_d^{*(ib)}, R_d^{*(ib)}\}, d = 1, \dots, D$, calculate

$$mse_d^{*(i)} = \frac{1}{B} \sum_{b=1}^B \left(\hat{\eta}_d^{eb*(ib)} - \eta_d^{*(ib)} \right)^2.$$

3.6. For $\eta_d^{(i)} \in \{A_d^{(i)}, R_d^{(i)}\}, \hat{\eta}_d^{eb(i)} \in \{\hat{A}_d^{eb(i)}, \hat{R}_d^{eb(i)}\}, d = 1, \dots, D$, calculate the coverage

$$C_{\eta_d}^{*(i)} = I \left(\eta_d^{(i)} \in \left(\hat{\eta}^{eb(i)} - z_{0.975} mse_d^{*(i)1/2}, \hat{\eta}^{eb(i)} + z_{0.975} mse_d^{*(i)1/2} \right) \right).$$

4. For $d = 1, \dots, D, \hat{\eta} \in \{\hat{A}_d^{eb}, \hat{R}_d^{eb}\}$, calculate

$$B_d(\hat{\eta}) = \frac{1}{I} \sum_{i=1}^I \left(mse_d^{*(i)} - MSE_d(\hat{\eta}) \right), \quad RE_d(\hat{\eta}) = \left(\frac{1}{I} \sum_{i=1}^I \left(mse_d^{*(i)} - MSE_d(\hat{\eta}) \right)^2 \right)^{1/2},$$

$$RB_d(\hat{\eta}) = \frac{B_d(\hat{\eta})}{MSE_d(\hat{\eta})} 100, \quad RRE_d(\hat{\eta}) = \frac{RE_d(\hat{\eta})}{MSE_d(\hat{\eta})} 100, \quad AB(\hat{\eta}) = \frac{1}{D} \sum_{d=1}^D |B_d(\hat{\eta})|,$$

$$RE(\hat{\eta}) = \frac{1}{D} \sum_{d=1}^D RE_d(\hat{\eta}), \quad RAB(\hat{\eta}) = \frac{1}{D} \sum_{d=1}^D |RB_d(\hat{\eta})|, \quad RRE(\hat{\eta}) = \frac{1}{D} \sum_{d=1}^D RRE_d(\hat{\eta}).$$

5. For $d = 1, \dots, D, \eta_d \in \{A_d, R_d\}$, calculate the coverage rates

$$C_{\eta_d} = \frac{1}{I} \sum_{i=1}^I C_{\eta_d}^{*(i)}, \quad C_{\eta} = \frac{1}{D} \sum_{d=1}^D C_{\eta_d}.$$

Tables 7 and 8 present the average absolute and relative performance measures, respectively, for $D = 50, B = 50, 100, 200, 300, 400$, and $n_d = 10, N_d = 200, d = 1, \dots, D$. We note that the absolute biases of \hat{A}_d^{eb} are smaller than the absolute biases of \hat{R}_d^{eb} , while their corresponding relative absolute biases are similar. This is due to the division by the corresponding Monte Carlo approximation to the true MSE. We observe that the MSEs of the MSE estimators decrease when the number of bootstrap resamples B increases. However the biases of the MSE estimators remain stable when B increases. It is remarkable that $RRE(\hat{\eta})$ is below 15% if $B = 200$ and it is around 12% if $B = 400$.

Table 9 presents the coverage rates C_{η} for $D = 50, B = 50, 100, 200, 300, 400$ and $n_d = 10, N_d = 200, d = 1, \dots, D$. They remain stable and close to the nominal value 0.95 even in the case $B = 50$. Therefore, the confidence interval based on the asymptotic normal distribution works “well” if the data is simulated from the model. To obtain more accurate results, it should be necessary to increase the number of Monte Carlo iterations. We have not carried out Simulation 2 with more iterations because of the high computational time.

TABLE 7 $10^3 AB(\hat{\eta})$ (left) and $10^3 RE(\hat{\eta})$ (right) with $D = 50, n_d = 10, N_d = 200$

B	50	100	200	300	400	50	100	200	300	400
\hat{A}^{eb}	0.0732	0.0744	0.0736	0.0744	0.0731	0.1878	0.1503	0.1251	0.1165	0.1087
\hat{R}^{eb}	0.3903	0.4346	0.4186	0.4370	0.4270	1.1833	0.9584	0.7672	0.7235	0.6757

TABLE 8 RAB($\hat{\eta}$) (left) and RRE($\hat{\eta}$) (right), in %, with $D = 50$, $n_d = 10$, $N_d = 200$

B	50	100	200	300	400	50	100	200	300	400
\hat{A}^{eb}	8.146	8.277	8.178	8.265	8.123	21.417	17.100	14.177	13.181	12.268
\hat{R}^{eb}	7.218	7.995	7.734	8.061	7.923	21.923	17.740	14.246	13.440	12.565

TABLE 9 C_η , with $D = 50$, $n_d = 10$, $N_d = 200$

B	50	100	200	300	400	Nominal
C_A	0.940	0.945	0.948	0.944	0.938	0.950
C_R	0.943	0.944	0.947	0.944	0.940	0.950

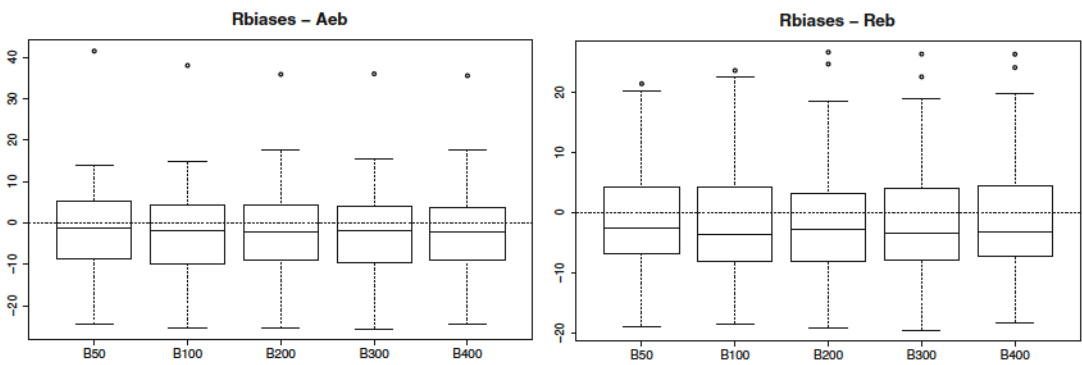
FIGURE 1 Relative biases of the mean squared error estimators for \hat{A}_d^{eb} (left) and \hat{R}_d^{eb} (right)

Figure 1 contains the boxplots of the empirical relative biases (Rbiases), in %, of the parametric bootstrap estimators of the MSEs of the predictors \hat{A}_d^{eb} (left) and \hat{R}_d^{eb} (right). Figure 2 presents the corresponding boxplots for the relative empirical relative root-MSEs (RRMSEs). More concretely, Figures 1 and 2 plot the quantities $RB_d(\hat{\eta})$ and $RRE_d(\hat{\eta})$, $d = 1, \dots, D$, for $\hat{\eta} \in \{\hat{A}_d^{eb}, \hat{R}_d^{eb}\}$, respectively. The first figure shows that the bootstrap MSE estimators are rather unbiased, with a small tendency to under estimation. The second figure suggests running around $B = 400$ iterations in the bootstrap resampling procedure for obtaining good approximations to the MSEs of the EBPs.

For studying the effect of deviations from normality, we repeat Simulation 2 by considering the same nonnormal scenarios as in Simulation 1. Table 10 presents the relative performance measures for $D = 50$, $n_d = 10$, $N_d = 200$. Table 11 gives the coverage rates. We observe that relative biases and root-MSEs have a moderate increment as the skewness parameter increases. However, these measures are more sensible to the decrease of the degrees of freedom of the multivariate Student's t distribution. In the case of R_d , the coverage rates move away from the nominal value 95% for high values of the kurtosis.

6 | ILLUSTRATIVE APPLICATION TO SHBS DATA

The SHBS is annually carried out by the “Instituto Nacional de Estadística” (INE), with the objective of obtaining information on the nature and destination of the consumption expenses, as well

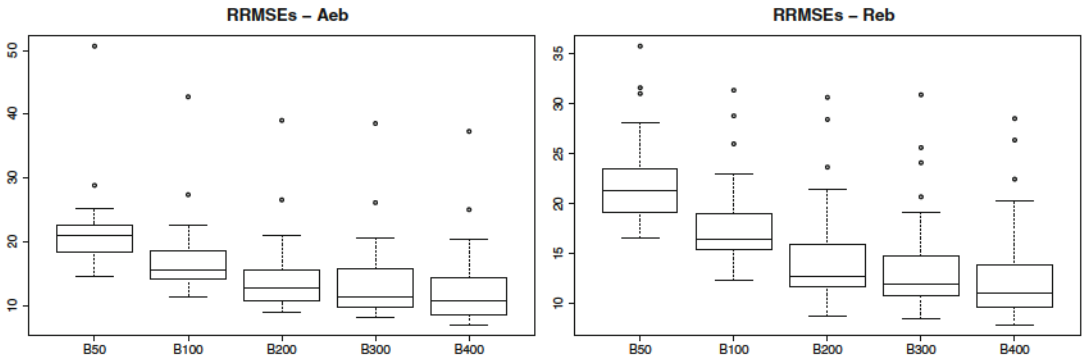


FIGURE 2 Relative root-mean squared errors (MSEs) of the MSE estimators for \hat{A}_d^{eb} (left) and \hat{R}_d^{eb} (right)

TABLE 10 RAB(η) and RRE(η) with $D = 50, n_d = 10, N_d = 200$

		Skew-normal, $\alpha_1 = 0$				Student's t		
	η	$\alpha_2 = 0$	$\alpha_2 = 2$	$\alpha_2 = 5$	$\alpha_2 = 10$	$df = 15$	$df = 10$	$df = 5$
RAB	\hat{A}^{eb}	8.6813	7.9008	7.9927	8.0333	7.5722	7.4171	9.8922
	\hat{R}^{eb}	9.1284	8.5401	8.3253	8.2471	13.7195	24.9397	67.1476
RRE	\hat{A}^{eb}	14.7324	14.2653	14.3236	14.3783	13.2178	13.0384	15.4502
	\hat{R}^{eb}	15.5767	15.3870	15.4869	15.5575	17.4954	26.8344	67.2886

TABLE 11 C_η , with $D = 50, n_d = 10, N_d = 200$

	$\alpha_2 = 0$	$\alpha_2 = 2$	$\alpha_2 = 5$	$\alpha_2 = 10$	$df = 15$	$df = 10$	$df = 5$
C_A	0.9464	0.9384	0.9416	0.9448	0.9448	0.9496	0.9424
C_R	0.9504	0.9344	0.9264	0.9288	0.9312	0.9144	0.7936

as on various characteristics related to the conditions of household life. In the Spanish economy it is important to have good estimates of consumer spending, since this spending represents, approximately, 60% of gross domestic product. However, global political measures are not often satisfactory for regional authorities, which can also develop their own economic strategies. They need some tools to determine, with precision and reliability, the main variables and consumer indicators in order to implement their strategies. Among the main consumer indicators are the local means of food and nonfood annual expenses of households and the ratios of annual food household expenses. For example, regional authorities are interested in providing aid for vulnerable families in basic necessities, such as food at province level. For this, it is necessary to know the mean expenditure of families on this type of goods.

This section presents an application of the new statistical methodology to the estimation of domain parameters defined as additive functions of two types of expense variables. We deal with data from the SHBS of 2016. The domains are the 50 Spanish provinces plus the autonomous cities Ceuta and Melilla, so that $D = 52$. Let z_{dj1} and z_{dj2} be the food and nonfood annually expenses of household j of domain d . The domain mean of food and nonfood household annually expenses

are

$$\bar{Z}_{d1} = \frac{1}{N_d} \sum_{j=1}^{N_d} z_{dj1}, \quad \bar{Z}_{d2} = \frac{1}{N_d} \sum_{j=1}^{N_d} z_{dj2}, \quad d = 1, \dots, D,$$

which are additive parameters with $h(z_{dj}) = \frac{1}{N_d}(z_{dj1}, z_{dj2})$. For each domain d , the ratio of the mean annually household expenses on food to the mean annually household expenses is

$$R_d = \frac{\bar{Z}_{d1}}{\bar{Z}_{d1} + \bar{Z}_{d2}}, \quad d = 1, \dots, D,$$

which are ratios of additive parameters. For each domain d , the mean of the ratios of food household annually expenses to the corresponding total household annually expenses is

$$A_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \frac{z_{dj1}}{z_{dj1} + z_{dj2}}, \quad d = 1, \dots, D,$$

which are additive parameters with $h(z_{dj}) = z_{dj1}/(z_{dj1} + z_{dj2})$. As we do not have a Spanish census file dated around 2016, we estimate the domain parameters \bar{Z}_{d1} , \bar{Z}_{d2} , A_d and R_d by using the EBPs $\hat{\delta}_d^{eb}$ defined in (10) and (14), respectively. For this sake, we fit a BNER model to the target variables z_{dj1} and z_{dj2} with categorical covariates that are related to consumption.

Scealy and Welsh (2017) showed that categorical auxiliary variables that influence the household consumption are the household composition and the area of usual residence. We only use the household typology because the area of usual residence was not significant. As auxiliary variable, we thus consider the household composition FC with categories

- FC1: Single person or adult couple with at least one members with age over 65,
- FC2: Other compositions with a single person or a couple without children,
- FC3: Couple with children under 16 years old or adult with children under 16 years old,
- FC4: Other households.

The variable FC is treated as a factor with reference category FC4.

For calculating the EBPs of the domain parameters of interest, by means of the formulas (14) and (10), we need the *true* population sizes, N_{dt} , of the crossings of provinces with the categories of the variable FC. We calculate these sizes by using the sampling weights of the Spanish Labor Force Survey (SLFS). The SLFS sampling weights are calibrated to the population sizes of the provinces crossed with sex and age groups. These demographic quantities come from the INE population projection system and are considered the most accurate demographic figures in Spain. The SHBS sampling weights are calibrated to the population sizes of the autonomous community (NUTS 2) crossed with sex and age groups. These weights make the direct estimators of socioeconomic indicators at the autonomous community level basically unbiased. However, they introduce a nonnegligible bias in the direct estimators of such indicators at the province level. For a detailed study of the influence of these weights and the construction of alternative estimators, see Appendix E in Data S1.

TABLE 12 Regression parameters and p -values

y	x -variable	Estimation	z -value	SE	p -Value
y_1	Intercept	-0.80	45.29	0.02	0.00
	FC1	-0.36	28.94	0.01	0.00
	FC2	-0.61	49.23	0.01	0.00
	FC3	-0.14	10.79	0.02	0.00
y_2	Intercept	0.83	42.00	0.02	0.00
	FC1	-0.39	37.24	0.01	0.00
	FC2	-0.29	27.93	0.01	0.00
	FC3	0.01	1.14	0.01	0.25

TABLE 13 Estimation and confidence intervals of variance and correlation parameters

Parameter	Estimation	L.CI	U.CI	L.CI.boot	U.CI.boot
σ_{u1}^2	0.013	0.007	0.018	0.007	0.019
σ_{u2}^2	0.018	0.010	0.025	0.010	0.026
ρ_u	0.614	0.421	0.807	0.367	0.774
σ_{e1}^2	0.451	0.442	0.459	0.448	0.467
σ_{e2}^2	0.318	0.312	0.324	0.315	0.327
ρ_e	0.377	0.366	0.389	0.364	0.387

We first fit a BNER model to the expenditure variables z_{dj1} and z_{dj2} . As the shape of the histogram estimators of the probability density functions of the model marginal residuals are slightly skewed, we apply the log transformation. Therefore, we fit a BNER model to $y_{dj1} = \log z_{dj1}$ and $y_{dj2} = \log z_{dj2}$, with z_1 and z_2 expressed in 10^4 euros and with the same auxiliary variables. For each target variable, y_1 and y_2 , Table 12 presents the estimates of the regression parameters and their standard errors. It also presents the asymptotic p -values for testing the hypotheses $H_0 : \beta_{kr} = 0$, $k = 1, 2$, $r = 1, 2, 3, 4$.

Table 13 presents the estimates of the variance and correlation parameters with their 95% confidence intervals, on the left side the asymptotic normality intervals and on the right side the bootstrap percentile intervals (Shao & Tu, 1995). This table shows that all the estimated parameters are significantly greater than zero. We remark that correlations ρ_u and ρ_e are significantly greater than zero, so that the independent univariate modeling of y_1 and y_2 is not appropriate.

Figure 3 plots the histograms of the $D = 52$ standardized EBPs of the random effects of the fitted BNER model for food (left) and nonfood (right) expenditures. The standardization of \hat{u}_{d1} and \hat{u}_{d2} is carried out by subtracting their mean value and dividing by their SD. It also prints the corresponding probability density function estimates. The shapes of the densities are quite symmetrical, which indicates that the distributions of the random effects are not very far from the normal distributions. Since D is too small to obtain a good non-parametric estimate of the density functions, the definitive conclusions can not be drawn.

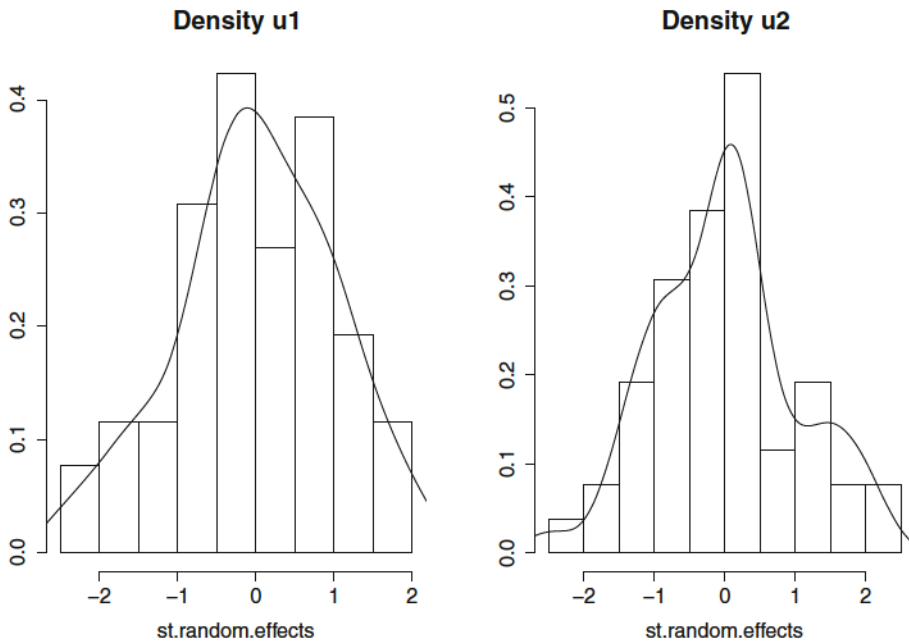


FIGURE 3 Histograms of standardized random effects

Figure 4 plots the histograms of the $n = 22,010$ standardized residuals (sresiduals) of the fitted BNER model for the first (left) and second (right) response variables. The standardization of \hat{e}_{dj1} and \hat{e}_{dj2} is carried out by subtracting their mean value and dividing by their SD. It also prints the corresponding probability density function estimates. The curves of the estimated densities have longer left tails and slightly skewed shape. Nevertheless, we could admit that the distributions of standardized residuals is not too far from normality.

Figure 5 plots the standardized residuals versus the predicted values of the fitted BNER model, which correspond to the logarithms of food (Y1) and nonfood (Y2) expenditures. In both cases, the main cloud of residuals is situated symmetrically around zero without any recognizable pattern.

Appendix D of Data S1 includes the results derived from using the transformations cube-root and fifth-root. Density figures D.2 and D.5 of standardized residuals are quite symmetrical and show that distributions are not too far from normality. Dispersion graphs D.3 and D.6 of standardized residuals versus predicted values show that the clouds of points are situated symmetrically around zero without any recognizable patterns. Table D.5 presents the skewness and kurtosis of the standardized residuals for the three transformations. It shows skewness close to zero in all cases and smaller kurtosis in case of the log transformation. Table D.6 presents the skewness and kurtosis of the random effects and leads to similar conclusions. For the marginal random effects, the Jarque–Bera normality test does not reject normality in any of the cases. In what follows, we present the EBPs based on the BNER model with the log transformation. Data S1 gives additional information about EBPs in two appendixes. Appendix E discusses the influence of the SLFS sampling weights on the estimation of the population sizes appearing in the formula of the EBPs. For this sake, it constructs alternative EBPs with population sizes estimated from SHBS sampling weights.

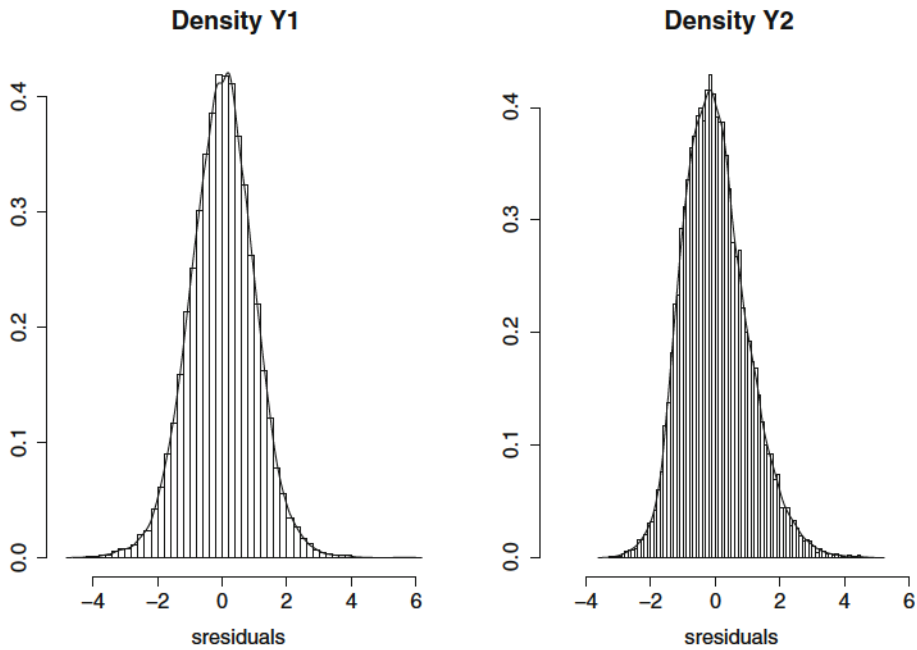


FIGURE 4 Histograms of standardized residuals

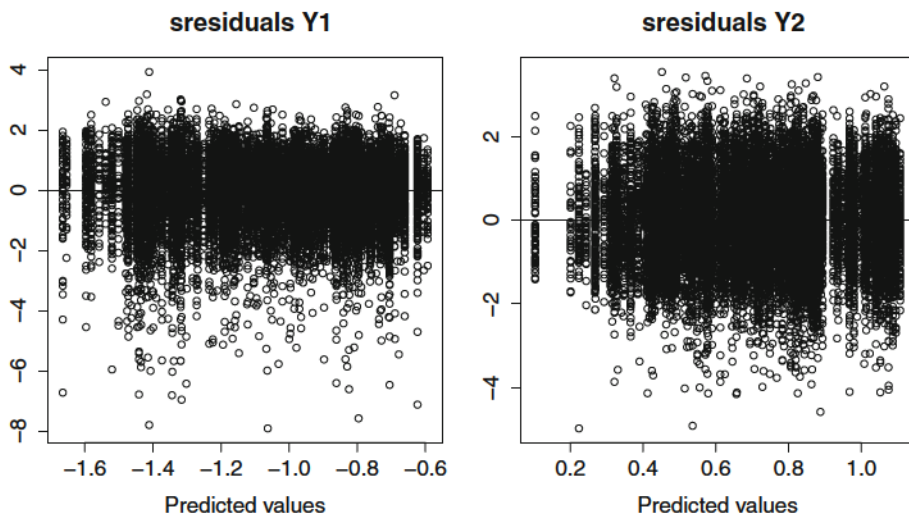


FIGURE 5 Standardized residuals versus predicted values (in 10^4 euros)

Direct estimators of small area parameters are not very precise because the sample size is small in the domains. However, they are approximately unbiased estimators under the distribution of the sample design. For this reason, the comparison of model-based predictors with direct estimators is of interest to researchers. In particular, the predictors are expected to follow the pattern of the direct estimates, but in a smoothed way. Appendix F presents some figures

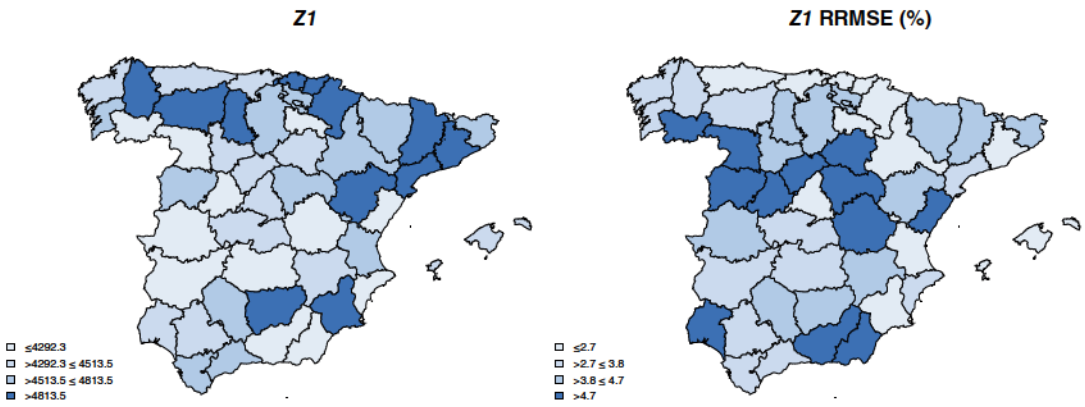


FIGURE 6 \hat{Z}_{d1}^{eb} (left) and their relative root-mean square errors (MSEs) in % (right) of household annual expenditures in food by Spanish provinces

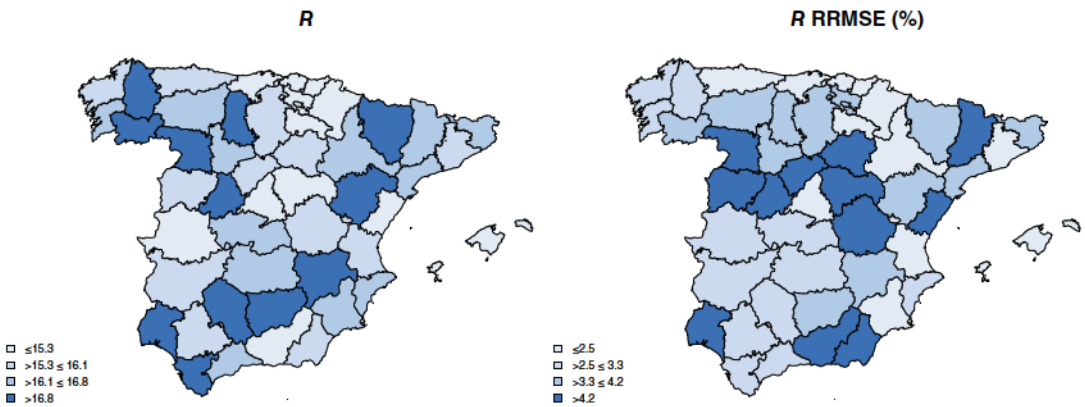


FIGURE 7 \hat{R}_d^{in} in % (right) and their relative root-mean square errors (MSEs) in % (right) of household annual expenditures in food by Spanish provinces

containing plots of direct and EBP estimates and plots of the corresponding estimates of the relative root-MSEs (RRMSE). It also gives tables with condensed numerical results.

Figure 6 (left) maps the means of the household annual expenditures in food by Spanish provinces. Figure 6 (right) maps the estimated RRMSE in %. These figures show that expenditures on food is rather variable between provinces.

Figure 7 (left) and Figure 8 (left) plot the ratios of means and the mean of ratios of household expenditures in food by Spanish provinces in %. Figure 7 (right) and Figure 8 (right) plot the corresponding RRMSEs in %. An interesting feature observed here is that within some autonomous regions, the province percentages of food expenditure, R_d , could be rather variable. The same happens for the province means of household percentages of food expenditure A_d . This happens mostly in the Autonomous Regions of Andalucía, Aragón, Castilla León or in Galicia, where there are many provinces and some of them are more deprived than others. In contrast, there are other regions, such as Cataluña and Basque Country where the variability of the estimated ratios is smaller.

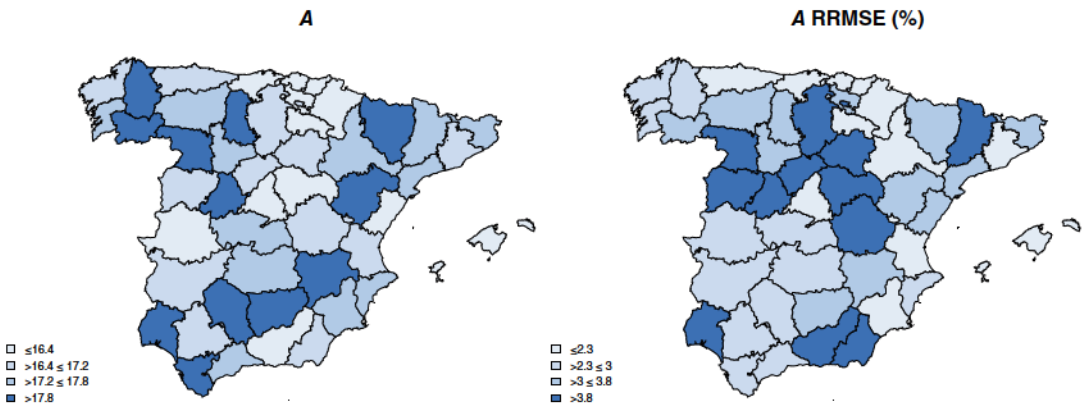


FIGURE 8 Ratios \hat{A}_d in % (right) and their relative root-mean square errors (MSEs) in % (right) of household annual expenditures in food by Spanish provinces

7 | CONCLUSIONS

This paper introduces small area predictors of expenditure means and ratios based on the BNER model (1). Best predictors minimize the MSE, within the class of unbiased predictors, under the model distribution. This optimality property approximately holds if we substitute true model parameters by consistent estimators, as the REML estimators are. The paper proposes estimating the MSEs of the EBPs by parametric bootstrap. As far as we know, this is the first time that EBPs for nonlinear bivariate parameters are introduced.

Two simulation experiments are carried out to empirically investigate and to check the behavior of the EBPs and the MSE estimators when the number of domains or the domain sample sizes are small. This is to say, in scenarios where the asymptotic properties might not hold. Simulation 1 investigates the biases and the MSEs of the EBPs. This simulation shows that the EBPs are basically unbiased, even in cases with rather small sample sizes, and that the MSEs decrease as the sample sizes n_d increase. Simulation 2 gives the recommendation of doing $B = 400$ iterations when applying the introduced parametric bootstrap procedures for estimating the MSEs.

For studying the effect of deviations from normality, Simulations 1 and 2 also changes the multivariate normal distributions of random effects and errors by multivariate skew-normal and Student's t distributions, keeping the same mean and variance component parameters. These simulations illustrate how efficiency measures worsen with increasing skewness or kurtosis of multivariate distributions. Data S1 contains additional simulations. Appendix B carries out new variants of Simulations 1 and 2 for studying the effect of correlations between the components of the random effects and errors in the behavior of the EBPs and MSE estimators. Appendix C moves apart from the unit-level model-based theory, where the sample is a deterministic subset of the population. In the new simulations the samples and the corresponding sizes are random and depends on the target vector y . The main findings are that the EBPs are not much affected because of the implemented informative sampling scenarios, but the parametric bootstrap estimators of their MSEs are drastically affected.

The introduced EBP methodology is applied to data from the SHBS of 2016. The target is to estimate province means of food and nonfood household annual expenditures, ratios of province

means of household annual expenditures and province means of ratios of household annual expenditures. The estimation procedure takes into account the correlation between the two target variables. The paper also compares the model-based estimates with direct estimates and it shows that introduced EBPs have lower MSEs.

The new methodology is not universal. It is limited to the assumed hypotheses. A key point is the need of having an auxiliary census files if the fitted model contains continuous auxiliary variables. This is a drawback that limits the applicability of the methodology, since there are few countries that maintain updated population censuses. However, it does not reduce the applicability to zero. The EBP approach, introduced by Molina and Rao (2010), can be applied using recent population censuses. For example, the World Bank traditionally used the Elbers et al. (2003) census-based methodology to map poverty in developing countries. On the other hand, the method is applicable to business surveys, where it is common to have a census of companies. The restriction of having a census is circumvented in the case that categorical explanatory variables are used. This is the example of the application to SHBS data, since Spain does not maintain an updated population census. It is true that in this case the predictive power of the model is reduced, but thus the whole model introduces valuable information that allows the construction of more efficient predictors than direct estimators.

We finally recall that we have carried out a research under the unit-level model-based approach. The corresponding extensions to the model-assisted or informative sampling approach will allow incorporating sampling design features into the statistical methodology, including point estimation and bootstrap MSE estimation.


ACKNOWLEDGMENTS


Supported by the Instituto Galego de Estatística, by the grants PGC2018-096840-B-I00 and PID2020-113578RB-I00 of the Spanish Ministerio de Economía y Competitividad, by the grant Prometeo/2021/063 of the Generalitat Valenciana, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C 2020/14), and by GAIN (Galician Innovation Agency) and the Regional Ministry of Economy, Employment and Industry grant COV20/00604 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01, all of them through the ERDF.

ORCID

Maria Dolores Esteban  <https://orcid.org/0000-0001-8330-9991>

Maria José Lombardía  <https://orcid.org/0000-0001-9452-9818>

Esther López-Vizcaíno  <https://orcid.org/0000-0002-7406-5849>

Domingo Morales  <https://orcid.org/0000-0002-9794-5654>

Agustín Pérez  <https://orcid.org/0000-0003-4994-3176>

REFERENCES

- Arima, S., Bell, W. R., Datta, G. S., Franco, C., & Liseo, B. (2017). Multivariate Fay–Herriot Bayesian estimation of small area means under functional measurement error. *Journal of the Royal Statistical Society: Series A*, 180(4), 1191–1109.
- Benavent, R., & Morales, D. (2016). Multivariate Fay–Herriot models for small area estimation. *Computational Statistics & Data Analysis*, 94, 372–390.
- Benavent, R., & Morales, D. (2021). Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. *JISS*, 30(1), 195–222.

- Boubeta, M., Lombardia, M. J., & Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST*, 25, 548–569.
- Boubeta, M., Lombardia, M. J., & Morales, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics & Data Analysis*, 107, 32–47.
- Chandra, H., & Salvati, N. (2018). Small area estimation for count data under a spatial dependent aggregated level random effects model. *Communication in Statistics- Theory and Methods*, 47(5), 1234–1255.
- Chandra, H., Salvati, N., & Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. *SpatStat*, 20, 30–56.
- Chandra, H., Salvati, N., & Chambers, R. (2018). Small area estimation under a spatially non-linear model. *Computational Statistics & Data Analysis*, 126, 19–38.
- Datta, G. S., Day, B., & Basawa, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, 75, 269–279.
- Datta, G. S., Day, B., & Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhya A*, 60(3), 344–362.
- Datta, G.S., Fay, R. E., & Ghosh, M. (1991). *Hierarchical and empirical Bayes multivariate analysis in small area estimation*. Proceedings of Bureau of the Census 1991 Annual Research Conference (pp. 63–79). Washington, DC: U.S. Bureau of the Census.
- Diallo, M. S., & Rao, J. N. K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, 45, 1092–1116.
- Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355–364.
- Erculescu, A.L., Berg, E., Cecere, W., & Ghosh, M. (2019). A bivariate hierarchical bayesian model for estimating cropland cash rental rates at the county level. survey methodology. *Statistics Canada*, Catalogue No. 12-001-X, Vol. 45, No. 2. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00001-eng.htm>
- Erculescu, A. L., Cruze, N., & Nandram, B. (2018). benchmarking a triplet of official estimates. *Environmental and Ecological Statistics*, 25(4), 523–547.
- Erculescu, A. L., & Fuller, W. A. (2016). Small area prediction under alternative model specifications. *Statistics in Transition New Series*, 17(1), 9–24.
- Erculescu, A. L., & Fuller, W. A. (2018). Bootstrap prediction intervals for small area means from unit-level nonlinear models. *Journal of Survey Statistics and Methodology*, 7(3), 309–333.
- Erculescu, A. L., & Opsomer, J. D. (2019). *A model-based approach to predict employee compensation components*. Joint Statistical Meetings Proceedings. Government Statistics Section (pp. 1601–1623). Alexandria, VA: American Statistical Association. <https://ww2.amstat.org/MembersOnly/proceedings/2019/data/assets/pdf/1199560.eps>
- Esteban, M. D., Lombardia, M. J., López-Vizcaíno, E., Morales, D., & Pérez, A. (2020). Small area estimation of proportions under area-level compositional mixed models. *TEST*, 29(3), 793–818.
- Esteban, M. D., Lombardia, M. J., López-Vizcaíno, E., Morales, D., & Pérez, A. (2022). Small area estimation of expenditure means and ratios under a unit-level bivariate linear mixed model. *Journal of Applied Statistics*, 31(1), 204–234.
- Fay, R. E. (1987). *Application of multivariate regression of small domain estimation*. In R. Platek, J. N. K. Rao, C. E. Särndal, & M. P. Singh (Eds.), *Small area statistics* (pp. 91–102). Wiley.
- Fuller, W. A., & Harter, R. (1987). *The multivariate components of variance model for small area estimation*. In R. Platek, J. N. K. Rao, C. E. Sarndall, & M. P. Singh (Eds.), *Small area statistics* (pp. 103–123). John Wiley & Sons, Inc.
- González-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D., & Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics & Data Analysis*, 51, 2720–2733.
- González-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 443–462.
- Graf, M., Marín, J. M., & Molina, I. (2019). A generalized mixed model for skewed distributions applied to small area estimation. *TEST*, 28, 565–597.
- Guadarrama, M., Molina, I., & Rao, J. N. K. (2016). A comparison of small area estimation methods for poverty mapping. *Stat Trans New Series*, 1, 41–66.

- Hall, P., & Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Annals of Statistics*, 34(4), 1733–1750.
- Hall, P., & Maiti, T. (2006b). On parametric bootstrap methods for small-area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 221–238.
- Hobza, T., & Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *The Journal of Official Statistics*, 32(3), 661–692.
- Hobza, T., Morales, D., & Marhuenda, Y. (2020). Small area estimation of additive parameters under unit-level generalized linear mixed models. *SORT*, 44(1), 3–38.
- Hobza, T., Morales, D., & Santamaría, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST*, 27(2), 270–294.
- Ito, T., & Kubokawa, T. (2021). Empirical best linear unbiased predictors in multivariate nested-error regression models. *Communications in Statistics - Theory and Methods*, 50, 2224–2249.
- Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111, 117–127.
- Jiang, J., & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53, 217–243.
- López-Vizcaino, E., Lombardía, M. J., & Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Model*, 13(2), 153–178.
- López-Vizcaino, E., Lombardía, M. J., & Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society. Series A*, 178(3), 535–565.
- Marchetti, S., & Secondi, L. (2017). Estimates of household consumption expenditure at provincial level in Italy by using small area estimation methods: “Real” comparisons using purchasing power parities. *Social Indicators Research*, 131, 215–234.
- Marhuenda, Y., Molina, I., Morales, D., & Rao, J. N. K. (2017). Poverty mapping in small areas under a two-fold nested error regression model. *Journal of the Royal Statistical Society. Series A*, 180(4), 1111–1136.
- Marino, M. F., Ranalli, M. G., Salvati, N., & Alfo, M. (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, 13(2), 1166–1197.
- McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models* (2nd ed.). John Wiley.
- Molina, I., & Martín, N. (2018). Empirical best prediction under a nested error model with log transformation. *Annals of Statistics*, 46(5), 1961–1993.
- Molina, I., Nandram, B., & Rao, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, 8, 852–885.
- Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, 369–385.
- Molina, I., Saei, A., & Lombardía, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society. Series A*, 170, 975–900.
- Morales, D., Esteban, M. D., Pérez, A., & Hobza, T. (2021). *A course on small area estimation and mixed models*. Springer.
- Morales, D., Pagliarella, M. C., & Salvatore, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. *SORT*, 39(1), 19–34.
- Ngaruye, I., Nzabanita, J., von Rosen, D., & Singull, M. (2017). Small area estimation under a multivariate linear model for repeated measures data. *Comm Statist Theory Methods*, 46(21), 10835–10850.
- Porter, A. T., Wikle, C. K., & Holan, S. H. (2015). Small area estimation via multivariate fay-Herriot models with latent spatial dependence. *The Australian & New Zealand Journal of Statistics*, 57, 15–29.
- Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. John Wiley.
- 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(1), 121–148.
- Scealy, J. L., & Welsh, A. H. (2017). A directional mixed effects model for compositional expenditure data. *Journal of the American Statistical Association*, 112(517), 24–36.
- Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. Springer.

- Torabi, M. (2019). Spatial generalized linear mixed models in small area estimation. *Canadian Journal of Statistics*, 47(3), 426–437.
- Ubaidillah, A., Notodiputro, K. A., Kurnia, A., & Wayan, I. (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), 2845–2861.
- Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference. A prediction approach*. John Wiley.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., & Pérez, A. (2022). Empirical best prediction of small area bivariate parameters. *Scandinavian Journal of Statistics*, 49(4), 1699–1727. <https://doi.org/10.1111/sjos.12618>