Small area estimation of expenditure means and ratios under a unit-level bivariate linear mixed model*

María Dolores Esteban[1], María José Lombardía[2], Esther López-Vizcaíno[3], Domingo Morales[1], Agustín Pérez[1]

[1]Universidad Miguel Hernández de Elche, Spain.
[2]Universidade da Coruña, CITIC, Spain,
[3]Instituto Galego de Estatística, Spain,

July, 2020

Abstract

Under a unit-level bivariate linear mixed model, this paper introduces small area predictors of expenditure means and ratios, and derives approximations and estimators of the corresponding mean squared errors. For the considered model, the REML estimation method is implemented. Several simulation experiments, designed to analyze the behavior of the introduced fitting algorithm, predictors and mean squared error estimators, are carried out. An application to real data from the Spanish household budget survey illustrates the behavior of the proposed statistical methodology. The target is the estimation of means of food and non-food household annual expenditures and of ratios of food household expenditures by Spanish provinces.

Key words: multivariate linear mixed models, nested error regression models, best linear unbiased predictors, ratio estimators, small area estimation, household budget surveys.
AMS subject classification: 62E30, 62J12

1 Introduction

It seems unnecessary to justify the relevance acquired by consumption in the 21st century, at a time when economically developed countries are enrolled in the consumer society, a social model in which a very important part of well-being and quality of life is associated with the acquisition of goods and services. The speed and diversity of the transformations experienced in recent years in the patterns, objects, shapes and places where consumption takes place are the basis for interest in the analysis of consumption from different perspectives. Make a good estimate of consumer spending is important in the economy of a country, since this spending represents, for example, approximately 60% of gross domestic product for Spain. 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, reliability and acceptable punctuality, the main variables and consume indicators in order to implement their strategies.

Among the main consume indicators we can cite the local means of food and non-food annual expenses of households and the ratios of annual food household expenses. The last indicator is

*Supported by the Instituto Galego de Estatística, by the grants MTM2017-82724-R and PGC2018-096840-B-I00 of the MINECO and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

defined as the quotient between the average annual expenditure on food of the households from a given territory and the corresponding average annual expenditure on all items of expenditure. The estimation of ratios in finite populations is usually done by estimating separately the numerator and the denominator with direct estimators. A direct estimator of the total or the average of a target variable in a domain uses only the data of that domain, it is basically unbiased with respect to the distribution of the sample design and its variance decreases when the sample size increases. The ratio estimators inherit part of these properties, so that both their biases and their variances also decrease as the sample size increases. However, domain sample sizes are typically small in the small area estimation (SAE) setup.

SAE deals with the estimation of domain indicators when the sample sizes are small for constructing precise direct estimators. One way to make up for the lack of sample size is to fit a model to the entire sample. Thus, when estimating the population indicators of a domain, data from other domains and the relationships between the different available variables are also taken into account. This is the so called model-based approach to SAE. The monograph of Rao and Molina (2015) gives a general description of SAE methods.

If there are several target variables, multivariate area-level or unit-level mixed models can take into account their correlations. These correlations give an important additional information for the estimation of domain parameters. Fay (1987) and Datta et al. (1991) showed that small area estimators obtained from multivariate models have, in general, better precision than the ones obtained from univariate models for each response variable. These estimator might be, for example, the hierarchical and empirical Bayes predictors introduced by Datta et al. (1991) or the empirical best linear unbiased predictors derived by González-Manteiga et al. (2008b) or Benavent et al. (2016) under multivariate linear mixed models.

There is an extensive literature on the use of statistical models for the estimation of small area socioeconomic indicators. Without being exhaustive, we quote some works that apply procedures based on area-level models. Molina et al. (2007), López-Vizcaíno et al. (2013, 2015) and Esteban et al. (2020) treated the problem of estimating labor force indicators. Morales et al. (2015), Porter et al. (2015), Boubeta et al. (2016, 2017) or Arima et al. (2017) presented applications to the estimation of poverty proportions or gaps. More recently, Marchetti and Secondi (2017) studied the household consumption expenditure at provincial level in Italy by using Fay-Herriot models and Ubaidillah et al. (2019) estimated food and non-food expenditures by small areas under a bivariate Fay-Herriot model.

On the other hand, unit-level models give also high flexibility for modeling micro data. Datta et al. (1999) studied the empirical Bayes prediction of small area mean vectors. Molina (2009) predicted exponentials of mixed effects under a multivariate nested-error regression model with logarithmic transformation. Tzavidis et al. (2008), Chambers et al. (2016) introduced predictors based on M-quantile regression models. Chandra at al. (2012) applied geographically weighted mixed effects models to the Australian agricultural and grazing industries survey. Hobza et al. (2016, 2018) derived predictors of small area poverty proportions based on unit-level logit mixed models, Hobza and Morales (2013) and Morales and Santamaría (2019) estimated domain means of household normalized net annual incomes under random regression coefficient models and temporal linear mixed models, respectively. Ngaruye (2017) derived empirical best linear predictors of domain means under a multivariate linear model for repeated measures data. Ito and Kubosawa (2018) employed a multivariate nested error regression model in the statistical analysis of posted land price data along the Keikyu train line from 1998 to 2001.

The above non complete lists of papers on SAE applications of area-level and unit-level multivariate statistical models show the benefits of taking into account the correlation structure of target variables. In fact, Ubaidillah et al. (2019) considered multivariate FH model and showed the strength of correlation between response variables plays a major role by proving more efficient estimators by use of correlation between response variables than univariate models. However, in the SAE literature, we have not found ratio estimators based on models that take into account the dependency of the involved target variables. Ratio estimators are, in general, constructed from independent estimators of the numerator and denominator. This can be done by fitting to each dependent variable a nested error regression (NER) model, which is the basic unit-level linear mixed model in SAE. However, selecting separate and independent models for each target variable does not take into account their correlation. This fact reduces the predictive capacity of the modeling and does not allow to properly estimate the mean square errors (MSEs) of the ratio predictors.

Although bivariate response variables can be modelled jointly using joint modelling approach through a shared parameter to handle the association between variables of interest, this paper follows a fully multivariate approach. The proposed solution to the problem of estimating ratios is introducing empirical best linear unbiased predictors (EBLUPs) of means and plug-in predictors of ratios, based on a unit-level bivariate linear mixed model. This approach improves the prediction of domain parameters with respect to separate modeling. The article develops predictors and offers approximations to their MSEs. It empirically studies the efficiency of the new proposal versus the usual predictor constructed from univariate and independent models, showing the weaknesses and strengths of both procedures. Finally, the paper illustrates the introduced methodology with an application to data from the 2016 Spanish family budget survey, estimating the means and ratios of food expenditure in Spanish households at the province level.

The rest of the paper is organized as follows. Section 2 describes the survey data and the estimation problem of interest. Section 3 introduces a bivariate nested error regression model and derives the EBLUP of means and the plug-in predictors of ratios. Section 4 approximates the MSEs of the introduced predictors. Section 5 carries out simulation experiments to investigate the behavior of the residual maximum likelihood (REML) fitting algorithm, the predictors of domain means and ratios and the MSE estimators. Section 6 gives an application to real data where the target is the small area estimation of averages and ratios of household annual expenditures in Spanish provinces. Sections 7 summarizes some conclusions. The paper contains two appendices. Appendix A gives a Fisher-scoring algorithm for calculating the REML estimators of the model parameters. Appendix B outlines some mathematical derivations for obtaining an approximation to the MSE of the EBLUP of a domain mean.

2 The data and the problem of interest

The Spanish household budget survey (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 as on various characteristics related to the conditions of household life. We deal with data from the SHBS of 2016. The SHBS collects expenditure and demographic information by personal interview from private dwellings across Spain. The dwellings are selected through a two-stage stratified random sampling in the primary sampling units. The primary sampling units are census sections and the secondary sampling units are dwellings. Our analysis is based on the household level file which contains almost 21.000 households in total. The target domains are the 52 Spanish provinces. The sample sizes of the SHBS are set to calculate precise estimator at the Autonomous Community (NUTS 2) level and does not produce official estimates at the province (NUTS 3) level. In this situation, estimating domain-level consume indicators is a SAE problem.

The response variables are y_{dj1} and y_{dj2}, the food and non-food annual expenses of household

j from domain d. Food includes both food and nonalcoholic beverages and non-food represents the remaining expenditures. The target parameters are the *domain means of food and non-food household annual expenses* and the *domain ratios of food household annual expenses*, i.e.

$$\overline{Y}_{d1}=\frac{1}{N_d}\sum_{j=1}^{N_d}y_{dj1},\quad \overline{Y}_{d2}=\frac{1}{N_d}\sum_{j=1}^{N_d}y_{dj2},\quad R_d=\frac{\overline{Y}_{d1}}{\overline{Y}_{d1}+\overline{Y}_{d2}},\quad d=1,\dots,D.$$

The Hájeck-type direct estimator of the domain mean $\overline{Y}_{dk}$, $k=1,2$ is

$$\hat{\overline{Y}}_{dk}^{dir}=\frac{1}{\hat{N}_d^{dir}}\sum_{j\in s_d}w_{dj}\,y_{djk},\quad \hat{N}_d^{dir}=\sum_{j\in s_d}w_{dj},\quad k=1,2, \tag{2.1}$$

where s_d is the domain sample and the w_{dj}'s are the elevation factors. The design-based covariances of these estimators can be approximated by

$$\widehat{\text{cov}}_\pi(\hat{\overline{Y}}_{d1}^{dir},\hat{\overline{Y}}_{d2}^{dir})=\big(\hat{N}_d^{dir}\big)^{-2}\sum_{j\in s_d}w_{dj}(w_{dj}-1)(y_{dj1}-\hat{\overline{Y}}_{dk_1}^{dir})(y_{dj2}-\hat{\overline{Y}}_{d2}^{dir}). \tag{2.2}$$

The last formulas are obtained from Särndal et al. (1992), pp. 43, 185 and 391, with the simplifications $w_{dj}=1/\pi_{dj}$, $\pi_{dj,dj}=\pi_{dj}$ and $\pi_{di,dj}=\pi_{di}\pi_{dj}$, $i\neq j$, in the second order inclusion probabilities. The direct estimator of the domain ratio R_d is

$$\hat{R}_d^{dir}=\frac{\hat{\overline{Y}}_{d1}^{dir}}{\hat{\overline{Y}}_{d1}^{dir}+\hat{\overline{Y}}_{d2}^{dir}},\quad d=1,\dots,D. \tag{2.3}$$

Section 6 shows that direct estimators (2.1) and (2.3) are not precise at the province level. This is why we look for alternative model-based estimation methods that borrow strength from auxiliary variables and that might produce more precise estimates of the domain target parameters. The available explanatory variables are

- Income. Total net annual household income (in euros).
- Family composition (FC). FC1: Single person or adult couple with at least one of the members being 65 years of age or older, FC2: Other compositions with a single person or a couple without children, FC3: Couple with children under 16 years old or adult with children less than or equal to 16 years old, FC4: Other households.
- Number of consumption units (multiplied by 10). $NCU=10\big\{1+0.5(N_1-1)+0.3N_2\big\}$, where N_1 is the number of people in the household aged 14 or older and N_2 is the number of people in the household under 14 years old.
- Rural. R1: Sparsely populated area, R0: Other areas.

We first analyze the potential predictive power of theses auxiliary variables through an explanatory data analysis. Figure 2.1 plots the observed food and non food expenditures versus the income. We observe that, despite the large variability observed in both plots, the two expenditure variables seem to increase linearly with the income. The estimated Pearson correlation coefficient between food expenditures and income is of 0.36 and between non-food expenditures and income is of 0.65. Further, the corresponding 95% confidence intervals are (0.36,0.37,) and (0.64,0.66), respectively. Therefore, income seems to have a good explanatory power for the target variables.

Figure 2.2 plots the food and non food expenditures for each family composition category. Both response variables have different means and variances across the family composition categories.

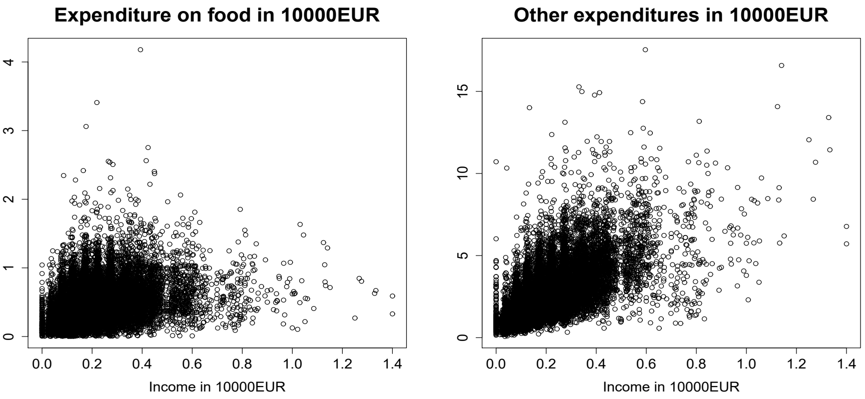

Figure 2.1: Income versus expenditures.

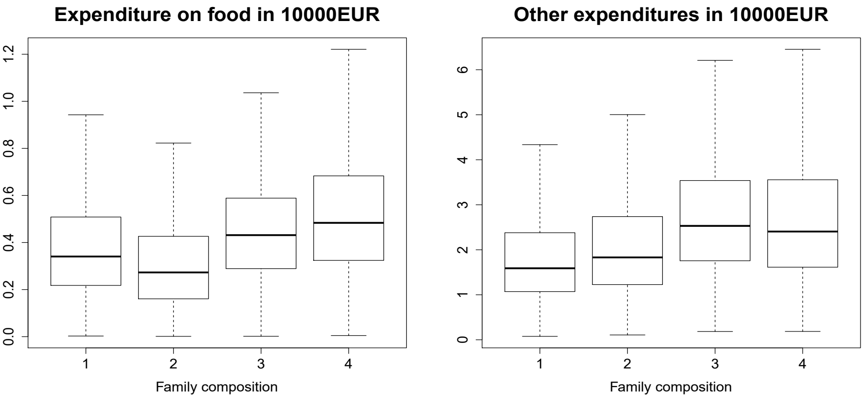

Figure 2.2: Family composition versus expenditures.

Therefore, FC is a candidate to enter as an auxiliary variable in models that explain the behavior of the expenditure variables.

Figure 2.3 plots the food and non food expenditures versus the consumption unities. As the expenditure variables increase with the consumption unities, NCU seems to be a good explanatory variable of the the expenditure variables.

Figure 2.4 plots the food and non food expenditures versus the degree of urbanization. The food expenditure does not show remarkable differences between the categories R0 and R1 of the variable rural. However, the other expenditures seems to be greater in the non rural areas and, therefore, it could be considered as a plausible auxiliary variable.

Therefore, the variables income, FC, NCU and rural could probably be good covariates for modelling the food and non-food expenditures. After fitting separate and independent nested error regression models, the tests of significance for the regression parameters confirm the explanatory power of these auxiliary variables. However, the problem of separate or joint modelling the food and non-food expenditures still remains. To analyze this issue, we calculate the Pearson correlation coefficients, P and P_d, and the corresponding p-values of the expense variables between and within domains. This is to say, for the sets of values

$$\left\{\left(\hat{\overline{Y}}_{d1}^{dir}, \hat{\overline{Y}}_{d2}^{dir}\right) : d = 1, \ldots, D\right\}, \quad \left\{(y_{dj1}, y_{dj2}) : j = 1, \ldots, n_d\right\}, \quad d = 1, \ldots, D.$$

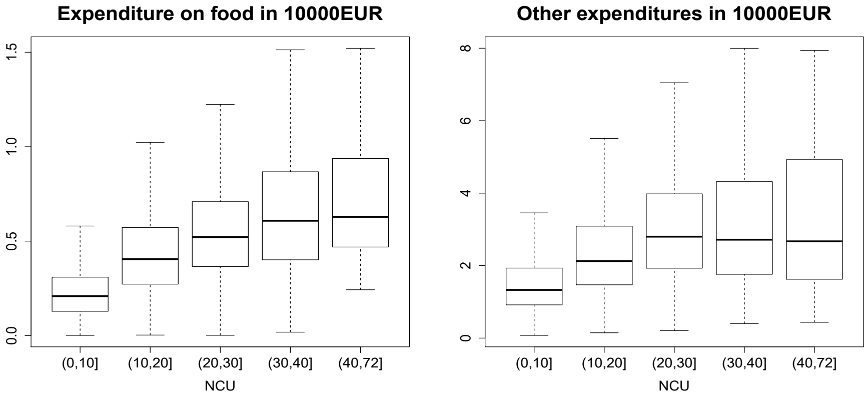

Figure 2.3: Number of consumption units versus expenditures.

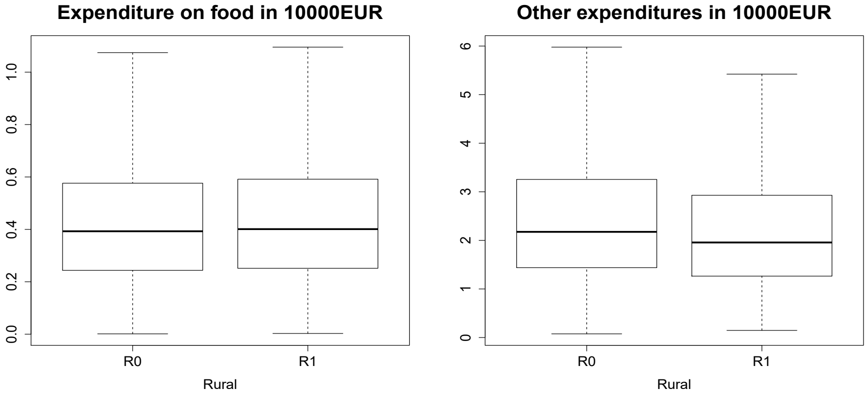

Figure 2.4: Degree of urbanization versus expenditures.

The between-domains correlation coefficient is $P = 0.560$ with p-value 0.12×10^{-4}. The within-domains correlation coefficients $\{P_1, \ldots, P_D\}$ are all positive with quartiles $q_0 = 0.188$, $q_1 = 0.351$, $q_2 = 0.398$, $q_3 = 0.440$, $q_4 = 0.539$ and the corresponding p-values are all lower than 0.05. This fact motivates the need to make a joint modeling of the expenditure variables and, in accordance, the introduction of a bivariate NER model. Section 3 describes the basic properties of the new model and proposes predictors of domain means and ratios.

3 The bivariate nested error regression model

3.1 The population model

Let U be a population of size N partitioned into D domains or areas $U_1, \ldots, U_D$ of sizes $N_1, \ldots, N_D$ 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, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$. For $k = 1, 2$, let $x_{djk} = (x_{djk1}, \ldots, x_{djkp_k})$ be a row vector containing p_k explanatory variables and let $X_{dj} = \operatorname{diag}(x_{dj1}, x_{dj2})_{2\times p}$ 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 bivariate nested error regression (BNER) model is

$$y_{dj} = X_{dj}\beta + u_d + e_{dj}, \quad d = 1, \ldots, D,\, j = 1, \ldots, N_d, \tag{3.1}$$

where the vectors of random effects $u_d = (u_{d1}, u_{d2})'$ and the vectors of random errors $e_{dj} = (e_{dj1}, e_{dj2})'$ are all mutually independent with multivariate normal distributions

$$u_d \sim N_2(0, V_{ud}), \quad e_{dj} \sim N_2(0, V_{edj}), \quad d = 1, \ldots, D,\, j = 1, \ldots, N_d.$$

The 2×2 covariance matrices V_{ud} depend on 3 unknown parameters, $\theta_1 = \sigma_{u1}^2$, $\theta_2 = \sigma_{u2}^2$ and $\theta_3 = \rho_u$, i.e.

$$V_{ud} = \begin{pmatrix} \sigma_{u1}^2 & \rho_u \sigma_{u1}\sigma_{u2} \\ \rho_u \sigma_{u1}\sigma_{u2} & \sigma_{u2}^2 \end{pmatrix}.$$

The 2×2 covariance matrices V_{edj} depend on 3 unknown parameters, $\theta_4 = \sigma_{e1}^2$, $\theta_5 = \sigma_{e2}^2$ and $\theta_6 = \rho_e$, i.e.

$$V_{edj} = \begin{pmatrix} \sigma_{e1}^2 & \rho_e \sigma_{e1}\sigma_{e2} \\ \rho_e \sigma_{e1}\sigma_{e2} & \sigma_{e2}^2 \end{pmatrix}.$$

Let I_m be the $m \times m$ identity matrix. We define the $2N_d \times 1$ vectors y_d and e_d, the $2N_d \times p$ matrix X_d and the $2N_d \times 2$ matrix Z_d, i.e.

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

Model (3.1) can be written in the domain-level form

$$y_d = X_d\beta + Z_d u_d + e_d, \quad d = 1, \ldots, D, \tag{3.2}$$

where $u_d \sim N_2(0, V_{ud})$, $e_d \sim N_{2N_d}(0, V_{ed})$ are independent and $V_{ed} = \underset{1\leq j\leq N_d}{\text{diag}}(V_{edj})$. The vectors y_d are independent with $y_d \sim N_{2N_d}(\mu_d, V_d)$, $\mu_d = X_d\beta$ and $V_d = Z_d V_{ud} Z_d' + V_{ed}$.

We define the $2N \times 1$ vectors y and e, the $2D \times 1$ vector u, the $2N \times p$ matrix X and $2N \times 2D$ matrix Z, i.e.

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

Model (3.1) can be written in the linear mixed model form

$$y = X\beta + Zu + e. \tag{3.3}$$

where $u \sim N_{2D}(0, V_u)$, $e \sim N_{2N}(0, V_{ed})$ are independent, $V_u = \underset{1\leq d\leq D}{\text{diag}}(V_{ud})$ and $V_e = \underset{1\leq d\leq D}{\text{diag}}(V_{ed})$. It holds that $y \sim N_{2N}(\mu, V)$, $\mu = X\beta$ and $V = ZV_uZ' + V_e$.

3.2 The sample model

In practice, inference is carried out based on a sample $s = \cup_{d=1}^D s_d$ of size $n = \sum_{d=1}^D n_d$ drawn from the finite population U. We write $U = s \cup r$ and $U_d = s_d \cup r_d$ to denote the sampled and non-sampled parts of the population. Let y_s and y_{ds} be the sub-vectors of y and y_d corresponding to sample elements and y_r and y_{dr} the sub-vectors of y and y_d corresponding to the out-of-sample elements. Without lack of generality, we can sort the components of vectors y and y_d to write $y = (y_s', y_r')'$ and $y_d = (y_{ds}', y_{dr}')'$. Define also the corresponding decompositions of X, Z, V_e,

V and X_d, Z_d, V_{ed}, V_d by using the subscripts s and r. This paper assumes the prediction approach to inference in finite populations that is described, for example, in Valliant et al. (2000). Therefore, we assume that sample indexes are fixed, so that the sample sub-vector y_s follows the model derived from the population model (3.3). This is to say, the sample BNER model is

$$y_s = X_s\beta + Z_s u + e_s, \tag{3.4}$$

where $u \sim N_{2D}(0, V_u)$, $e_s \sim N_{2n}(0, V_{es})$ are independent, $V_u = \underset{1\leq d\leq D}{\text{diag}}\,(V_{ud})$, $V_{es} = \underset{1\leq d\leq D}{\text{diag}}\,(V_{eds})$ and $V_{eds} = \underset{1\leq j\leq n_d}{\text{diag}}\,(V_{edj})$. It holds that $y_s \sim N_{2n}(\mu_s, V_s)$, $\mu_s = X_s\beta$ and $V_s = Z_s V_u Z_s' + V_{es}$. Similarly, the sample subvectors y_{ds} follow the models derived from (3.2), i.e.

$$y_{ds} = X_{ds}\beta + Z_{ds}u_d + e_{ds}, \quad d = 1, \ldots, D, \tag{3.5}$$

where $u_d \sim N_2(0, V_{ud})$, $e_{ds} \sim N_{2n_d}(0, V_{eds})$ are independent. The vectors y_{ds} are independent with $y_{ds} \sim N_{2n_d}(\mu_{ds}, V_{ds})$, $\mu_{ds} = X_{ds}\beta$ and $V_{ds} = Z_{ds}V_{ud}Z_{ds}' + V_{eds}$.

Under model (3.4), the best linear unbiased estimator (BLUE) of β, and the best linear unbiased predictors (BLUP) of u are

$$\hat{\beta}_B = (X_s' V_s^{-1} X_s)^{-1} X_s' V_s^{-1} y_s, \quad \hat{u}_B = V_u Z_s' V_s^{-1}(y_s - X_s\hat{\beta}_B). \tag{3.6}$$

Instead of doing a direct inversion of the $2n_d \times 2n_d$ matrix $V_{ds} = V_{eds} + Z_{ds}V_{ud}Z_{ds}'$ when calculating the BLUE of β, it is computationally more efficient to apply the formula

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}, \tag{3.7}$$

with $A = V_{eds}$, $B = Z_{ds}$, $C = V_{ud}$ and $D = Z_{ds}'$. As $Z_{ds}'V_{eds}^{-1}Z_{ds} = \sum_{j=1}^{n_d} V_{edj}^{-1} = n_d V_{edj}^{-1}$. We obtain

$$V_{ds}^{-1} = V_{eds}^{-1} - V_{eds}^{-1}Z_{ds}\big(V_{ud}^{-1} + Z_{ds}'V_{eds}^{-1}Z_{ds}\big)^{-1}Z_{ds}'V_{eds}^{-1} = V_{eds}^{-1} - V_{eds}^{-1}Z_{ds}\Big(V_{ud}^{-1} + n_d V_{edj}^{-1}\Big)^{-1}Z_{ds}'V_{eds}^{-1}.$$

where $V_{eds}^{-1} = \underset{1\leq j\leq n_d}{\text{diag}}\,(V_{edj}^{-1})$. The new formula reduces the computational burden, as it only requires running inversion algorithms for 2×2 matrices.

In practice, BLUPs and BLUEs are not calculable as the vector θ of model parameters is unknown. Appendix A gives a Fisher-scoring algorithm for calculating the REML estimator of θ. Let $\hat{\theta}$ be an estimator of θ. By plugging $\hat{\theta}$ in V_u and V_{es}, we get $\hat{V}_u = V_u(\hat{\theta})$, $\hat{V}_{es} = V_{es}(\hat{\theta})$ and $\hat{V}_s = Z_s\hat{V}_u Z_s' + \hat{V}_{es}$. By substituting $\hat{V}_s$ and $\hat{V}_u$ in (3.6), we obtain the empirical BLUE (EBLUE) of β and the empirical BLUP (EBLUP) of u, i.e.

$$\hat{\beta} = (X_s'\hat{V}_s^{-1}X_s)^{-1}X_s'\hat{V}_s^{-1}y_s, \quad \hat{u} = \hat{V}_u Z_s'\hat{V}_s^{-1}(y_s - X_s\hat{\beta}). \tag{3.8}$$

Alternative formulas are

$$\hat{\beta} = \Big(\sum_{d=1}^{D} X_{ds}'\hat{V}_{ds}^{-1}X_{ds}\Big)^{-1}\sum_{d=1}^{D} X_{ds}'\hat{V}_{ds}^{-1}y_{ds}, \quad \hat{u} = \underset{1\leq d\leq D}{\text{col}}(\hat{u}_d), \quad \hat{u}_d = \hat{V}_{ud}Z_{ds}'\hat{V}_{ds}^{-1}\big(y_{ds} - X_{ds}\hat{\beta}\big).$$

3.3 Predictors of domain means and ratios

Under the BNER model (3.3), this section derives the EBLUPs of the 2×1 mean vectors $\overline{Y}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj}$ and introduces the plug-in predictors of the domain ratios $R_d = \overline{Y}_{d1}/(\overline{Y}_{d1} + \overline{Y}_{d2})$, $d = 1, \dots, D$. Assuming that sample indexes are fixed, the non-sampled sub-vectors y_{dr} follow the models derived from (3.2), i.e.

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

where $u_d \sim N_2(0, V_{ud})$, $e_{dr} \sim N_{2(N_d - n_d)}(0, V_{edr})$ are independent and $V_{edr} = \underset{n_d+1 \leq j \leq N_d}{\text{diag}} (V_{edj})$. The vectors y_{dr} are independent with $y_{dr} \sim N_{2(N_d-n_d)}(\mu_{dr}, V_{dr})$, $\mu_{dr} = X_{dr}\beta$, $V_{dr} = Z_{dr}V_{ud}Z'_{dr} + V_{edr}$. Further, 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 conditional mean of y_{dr}, given the sample data y_s, is the $2(N_d - n_d) \times 1$ vector

$$\begin{aligned} E[y_{dr}|y_s] &= E[y_{dr}|y_{ds}] = \mu_{dr} + V_{drs}V_{ds}^{-1}(y_{ds} - \mu_{ds}) = X_{dr}\beta + Z_{dr}V_{ud}Z'_{ds}V_{ds}^{-1}(y_{ds} - X_{ds}\beta) \\ &= X_{dr}\beta + Z_{dr}V_{ud}Z'_{ds}\Big\{V_{eds}^{-1} - V_{eds}^{-1}Z_{ds}\Big(V_{ud}^{-1} + n_dV_{edj}^{-1}\Big)^{-1}Z'_{ds}V_{eds}^{-1}\Big\}(y_{ds} - X_{ds}\beta). \end{aligned}$$

For the following calculations, we note that

$$Z'_{ds}V_{eds}^{-1}(y_{ds} - X_{ds}\beta) = \sum_{j=1}^{n_d} V_{edj}^{-1}(y_{dj} - X_{dj}\beta).$$

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

$$\begin{aligned} E[y_{dj}|y_{ds}] &= X_{dj}\beta + V_{ud}Z'_{ds}\Big\{V_{eds}^{-1} - V_{eds}^{-1}Z_{ds}\big(V_{ud}^{-1} + n_dV_{edj}^{-1}\big)^{-1}Z'_{ds}V_{eds}^{-1}\Big\}(y_{ds} - X_{ds}\beta) \\ &= X_{dj}\beta + V_{ud}\Big\{I_2 - n_dV_{edj}^{-1}\big(V_{ud}^{-1} + n_dV_{edj}^{-1}\big)^{-1}\Big\}\sum_{j=1}^{n_d} V_{edj}^{-1}(y_{dj} - X_{dj}\beta). \end{aligned}$$

We have

$$\hat{y}_{ds}^{eb} = y_{ds}, \quad \hat{y}_{dr}^{eb} = \hat{E}[y_{dr}|y_{ds}] = X_{dr}\hat{\beta} + Z_{dr}\hat{V}_{ud}Z'_{ds}\hat{V}_{ds}^{-1}(y_{ds} - X_{ds}\hat{\beta}),$$

or equivalently, $\hat{y}_{dj}^{eb} = y_{dj}$ if $j \in s_d$ and $\hat{y}_{dj}^{eb} = \hat{E}[y_{dj}|y_{ds}]$ if $j \in r_d$, where

$$\hat{E}[y_{dj}|y_{ds}] = X_{dj}\hat{\beta} + \hat{V}_{ud}\Big\{I_2 - n_d\hat{V}_{edj}^{-1}\big(\hat{V}_{ud}^{-1} + n_d\hat{V}_{edj}^{-1}\big)^{-1}\Big\}\sum_{j=1}^{n_d} \hat{V}_{edj}^{-1}(y_{dj} - X_{dj}\hat{\beta}).$$

The EBLUP of $\overline{Y}_d$ is

$$\begin{aligned} \hat{\overline{Y}}_d^{eb} &= \big(\hat{\overline{Y}}_{d1}^{eb}, \hat{\overline{Y}}_{d2}^{eb}\big)' = \frac{1}{N_d}\sum_{j=1}^{N_d} \hat{y}_{dj}^{eb} = \frac{1}{N_d}\sum_{j=1}^{n_d} y_{dj} + \frac{1}{N_d}\sum_{j=n_d+1}^{N_d} \{X_{dj}\hat{\beta} + \hat{u}_d\} \\ &= f_d\hat{\overline{Y}}_d + \frac{1}{N_d}\sum_{j=1}^{N_d}\{X_{dj}\hat{\beta} + \hat{u}_d\} - f_d\frac{1}{n_d}\sum_{j=1}^{n_d}\{X_{dj}\hat{\beta} + \hat{u}_d\} \\ &= (1 - f_d)\big[\overline{X}_d\hat{\beta} + \hat{u}_d\big] + f_d\big[\hat{\overline{Y}}_d + (\overline{X}_d - \hat{\overline{X}}_d)\hat{\beta}\big]. \end{aligned} \tag{3.9}$$

where $\hat{\overline{Y}}_d = \frac{1}{n_d}\sum_{j=1}^{n_d} y_{dj}$, $\hat{\overline{X}}_d = \frac{1}{n_d}\sum_{j=1}^{n_d} X_{dj}$, $f_d = \frac{n_d}{N_d}$. The plug-in predictor of the ratio $R_d = \overline{Y}_{d1}/(\overline{Y}_{d1} + \overline{Y}_{d2})$ is

$$\hat{R}_d^{in} = \frac{\hat{\overline{Y}}_{d1}^{eb}}{\hat{\overline{Y}}_{d1}^{eb} + \hat{\overline{Y}}_{d2}^{eb}}. \tag{3.10}$$

If $n_d = 0$ and $j \in r_d$, then $r_d = U_d$ and the conditional 2×1 mean vector is $E[y_{dj}|y_s] = X_{dj}\beta$. In this case, the EBLUP of $\overline{Y}_d$ is the synthetic estimator $\overline{Y}_d^{syn} = \overline{X}_d\hat{\beta}$, with $\overline{X}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} x_{dj}$.

4 Estimation of MSEs

Prasad and Rao (1990) gave an approximation to the MSE of the EBLUP of $X_d\beta + Z_d u_d$ under the univariate NER model when the variance component parameters are estimated by using the Henderson's method 3. Datta and Lahiri (2000) extended the results of Prasad and Rao (1990) to the case of the general longitudinal model. They further considered ML and REML estimators of the variance components. For the general linear mixed model, Das et al. (2004) derived the MSE of the EBLUP when the REML or the maximum likelihood fitting methods are employed. Their proof contains the general longitudinal model considered by Datta and Lahiri (2000) as a special case. However, none of the three papers study the approximation to the matrix of mean squared crossed errors of the EBLUP of the mean vector defined in (3.9). Although, the BNER model (3.3) can be written in the form of the general linear mixed model considered by Das et al. (2004), the approximation to the matrix of mean squared crossed errors is not covered by that paper. This is why Appendix B presents the mathematical derivations for approximating and estimating the MSEs of $\hat{\overline{Y}}_d^{eb}$ and $\hat{R}_d^{in}$. The obtained MSE estimators are presented below.

4.1 MSE of the EBLUP of a domain mean

Let us define $T_{ds} = V_{ud} - V_{ud}Z'_{ds}V_{ds}^{-1}Z_{ds}V_{ud}$, $Q_s = (X'_sV_s^{-1}X_s)^{-1}$ and

$$\hat{\overline{X}}_{ds} = \sum_{j=1}^{n_d} V_{edj}^{-1}X_{dj}, \quad \overline{X}_{dr} = \text{diag}\big(\overline{X}_{d1r}, \overline{X}_{d2r}\big), \quad \overline{X}_{dkr} = \frac{1}{N_d - n_d}\sum_{j=n_d+1}^{N_d} x_{djk}, \; k = 1, 2.$$

When predicting $\overline{Y}_d$ with $\hat{\overline{Y}}_d^{eb}$, we use the MSE matrix estimator

$$mse(\hat{\overline{Y}}_d^{eb}) = \begin{pmatrix} mse(\hat{\overline{Y}}_{d1}^{eb}) & mse(\hat{\overline{Y}}_{d1}^{eb}, \hat{\overline{Y}}_{d2}^{eb}) \\ mse(\hat{\overline{Y}}_{d1}^{eb}, \hat{\overline{Y}}_{d2}^{eb}) & mse(\hat{\overline{Y}}_{d2}^{eb}) \end{pmatrix} = g_1(\hat{\theta}) + g_2(\hat{\theta}) + 2g_3(\hat{\theta}) + g_4(\hat{\theta}), \tag{4.1}$$

where

$$\begin{aligned}
g_1(\theta) &= (1 - f_d)^2\Big\{V_{ud} - n_dV_{ud}V_{edj}^{-1}V_{ud} + n_d^2V_{ud}V_{edj}^{-1}T_{ds}V_{edj}^{-1}V_{ud}\Big\}, \\
g_2(\theta) &= (1 - f_d)^2\Big[\overline{X}_{dr} - T_{ds}\hat{\overline{X}}_{ds}\Big]Q_s\Big[\overline{X}_{dr} - T_{ds}\hat{\overline{X}}_{ds}\Big]'. \\
g_3(\theta) &= \Big(\text{tr}\Big\{(\nabla b'_{k_1})V_s(\nabla b'_{k_2})'E\Big[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'\Big]\Big\}\Big)_{k_1,k_2=1,2}, \quad (\nabla b'_k) = \underset{1\le\ell\le6}{\text{col}}\Big(\frac{\partial b'_i}{\partial\theta_\ell}\Big)_{6\times2n}. \\
g_4(\theta) &= \frac{1 - f_d}{N_d}V_{edj}.
\end{aligned}$$

where the 6×6 matrix $E\big[(\hat{\theta}-\theta)(\hat{\theta}-\theta)'\big]$ can be approximated by the output $F^{-1}(\hat{\theta})$ of the REML Fisher scoring algorithm described in Appendix A and the derivatives of b' are

$$\frac{\partial b'}{\partial \theta_\ell}=\begin{pmatrix}\frac{\partial b_1'}{\partial \theta_\ell}\\ \frac{\partial b_2'}{\partial \theta_\ell}\end{pmatrix}=(1-f_d)\big\{V_{ud\ell}Z_{ds}'V_{ds}^{-1}-V_{ud}Z_{ds}'V_{ds}^{-1}Z_{ds}V_{ud\ell}Z_{ds}'V_{ds}^{-1}\big\},\quad \ell=1,2,3,$$

and

$$\frac{\partial b'}{\partial \theta_\ell}=\begin{pmatrix}\frac{\partial b_1'}{\partial \theta_\ell}\\ \frac{\partial b_2'}{\partial \theta_\ell}\end{pmatrix}=-(1-f_d)V_{ud}Z_{ds}'V_{ds}^{-1}\underset{1\leq j\leq n_d}{\text{diag}}\,(V_{edj\ell})V_{ds}^{-1},\quad \ell=4,5,6.$$

The diagonal elements of matrix (4.1), $mse(\hat{\overline{Y}}_{d1}^{eb})$ and $mse(\hat{\overline{Y}}_{d2}^{eb})$, are the estimator of $MSE(\hat{\overline{Y}}_{d1}^{eb})$ and $MSE(\hat{\overline{Y}}_{d2}^{eb})$ respectively.

4.2 MSE of the plug-in predictor of a domain ratio

The plug-in predictor of the ratio $R_d=\overline{Y}_{d1}/(\overline{Y}_{d1}+\overline{Y}_{d2})$ is

$$\hat{R}_d^{in}=\frac{\hat{\overline{Y}}_{d1}^{eb}}{\hat{\overline{Y}}_{d1}^{eb}+\hat{\overline{Y}}_{d2}^{eb}}=f\big(\hat{\overline{Y}}_{d1}^{eb},\hat{\overline{Y}}_{d2}^{eb}\big),\quad f(y_1,y_2)=\frac{y_1}{y_1+y_2}.$$

An approximation to the MSE of $\hat{R}_d^{in}$ can be obtained by Taylor linearization. The first partial derivatives of f are

$$\frac{\partial f}{\partial y_1}=\frac{y_2}{(y_1+y_2)^2},\qquad \frac{\partial f}{\partial y_2}=\frac{-y_1}{(y_1+y_2)^2},$$

The first order Taylor expansion of $f\big(\hat{\overline{Y}}_{d1}^{eb},\hat{\overline{Y}}_{d2}^{eb}\big)$ around $\big(\overline{Y}_{d1},\overline{Y}_{d2}\big)$ is

$$\begin{aligned}\hat{R}_d^{in}&=f\big(\hat{\overline{Y}}_{d1}^{eb},\hat{\overline{Y}}_{d2}^{eb}\big)\approx f\big(\overline{Y}_{d1},\overline{Y}_{d2}\big)+\frac{\partial f\big(\overline{Y}_{d1},\overline{Y}_{d2}\big)}{\partial y_1}\big(\hat{\overline{Y}}_{d1}^{eb}-\overline{Y}_{d1}\big)+\frac{\partial f\big(\overline{Y}_{d1},\overline{Y}_{d2}\big)}{\partial y_2}\big(\hat{\overline{Y}}_{d2}^{eb}-\overline{Y}_{d2}\big)\\ &=R_d+\frac{\overline{Y}_{d2}}{(\overline{Y}_{d1}+\overline{Y}_{d2})^2}\big(\hat{\overline{Y}}_{d1}^{eb}-\overline{Y}_{d1}\big)-\frac{\overline{Y}_{d1}}{(\overline{Y}_{d1}+\overline{Y}_{d2})^2}\big(\hat{\overline{Y}}_{d2}^{eb}-\overline{Y}_{d2}\big).\end{aligned}$$

Therefore, we obtain the approximation

$$\begin{aligned}MSE(\hat{R}_d^{in})&=E\big[(\hat{R}_d^{in}-R_d)^2\big]\approx\frac{\overline{Y}_{d2}^2}{(\overline{Y}_{d1}+\overline{Y}_{d2})^4}E\big[\big(\hat{\overline{Y}}_{d1}^{eb}-\overline{Y}_{d1}\big)^2\big]\\ &+\frac{\overline{Y}_{d1}^2}{(\overline{Y}_{d1}+\overline{Y}_{d2})^4}E\big[\big(\hat{\overline{Y}}_{d2}^{eb}-\overline{Y}_{d2}\big)^2\big]-2\frac{\overline{Y}_{d1}\overline{Y}_{d2}}{(\overline{Y}_{d1}+\overline{Y}_{d2})^4}E\big[\big(\hat{\overline{Y}}_{d1}^{eb}-\overline{Y}_{d1}\big)\big(\hat{\overline{Y}}_{d2}^{eb}-\overline{Y}_{d2}\big)\big]\\ &=\frac{\overline{Y}_{d2}^2}{(\overline{Y}_{d1}+\overline{Y}_{d2})^4}MSE\big(\hat{\overline{Y}}_{d1}^{eb}\big)+\frac{\overline{Y}_{d1}^2}{(\overline{Y}_{d1}+\overline{Y}_{d2})^4}MSE\big(\hat{\overline{Y}}_{d2}^{eb}\big)\\ &-2\frac{\overline{Y}_{d1}\overline{Y}_{d2}}{(\overline{Y}_{d1}+\overline{Y}_{d2})^4}MSE\big(\hat{\overline{Y}}_{d1}^{eb},\hat{\overline{Y}}_{d2}^{eb}\big),\end{aligned}\tag{4.2}$$

where $MSE\big(\hat{\overline{Y}}_{d1}^{eb},\hat{\overline{Y}}_{d2}^{eb}\big)$ are the non-diagonal elements of the matrix.

When predicting R_d with $\hat{R}_d^{in}$, we use the MSE estimator $mse(\hat{R}_d^{in})$ obtained as a plug-in estimator of the approximation (4.2). This is to say, we substitute each MSE by the corresponding component of the matrix $mse(\hat{\overline{Y}}_d^{eb})$ given in (4.1). Similarly, Appendix C gives an estimator of the MSE of the plug-in predictor $\hat{Q}_d^{in}=\hat{\overline{Y}}_{d1}^{eb}/\hat{\overline{Y}}_{d2}^{eb}$ of the quotient $Q_d=\overline{Y}_{d1}/\overline{Y}_{d2}$.

5 Simulations

5.1 Simulation 1

The target of Simulation 1 is to check the behavior of the REML algorithm for fitting the BNER model (3.5). We take $p_1 = p_2 = 2$, $p = 4$, $\beta_1 = (\beta_{11}, \beta_{12})' = (1,1)'$, $\beta_2 = (\beta_{21}, \beta_{22})' = (1,1)'$, For $d = 1, \dots, D$, $j = 1, \dots, n_d$, generate $X_{dj} = \text{diag}(x_{dj1}, x_{dj2})_{2\times 4}$, where $x_{dj1} = (x_{dj11}, x_{dj12})$, $x_{dj2} = (x_{dj21}, x_{dj22})$, $x_{dj11} = x_{dj21} = 1$, $x_{dj12} \sim U(2,4)$ and $x_{dj22} \sim U(2,5)$. We take $\theta_1 = 0.75$, $\theta_2 = 1.00$, $\theta_4 = 0.50$, $\theta_5 = 0.75$ and $\theta_3 = -0.8$, $\theta_6 = 0.8$. For $d = 1, \dots, D$, simulate $u_d \sim N_2(0, V_{ud})$ and $e_{dj} \sim N_2(0, V_{edj})$, where

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

The steps of Simulation 1 are

1. Generate x_{djk}, $d = 1, \dots, D$, $j = 1, \dots, n_d$, $k = 1, 2$.

2. Repeat $I = 10^3$ times ($i = 1, \dots, 10^3$)

   2.1. Generate $u_d^{(i)} \sim N_2(0, V_{ud})$, $e_d^{(i)} \sim N_{2n_d}(0, V_{ed})$, $y_d^{(i)} = X_d\beta + Z_d u_d^{(i)} + e_d^{(i)}$, $d = 1, \dots, D$.

   2.2. For every $\eta \in \{\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \theta_1, \dots, \theta_6\}$, calculate the REML estimator $\hat{\eta}^{(i)} \in \{\hat{\beta}_{11}^{(i)}, \hat{\beta}_{12}^{(i)}, \hat{\beta}_{21}^{(i)}, \hat{\beta}_{22}^{(i)}, \hat{\theta}_1^{(i)}, \dots, \hat{\theta}_6^{(i)}\}$.

3. Output:

$$RMSE(\hat{\eta}) = \left(\frac{1}{I}\sum_{i=1}^{I}(\hat{\eta}^{(i)} - \eta)^2\right)^{1/2}, \quad BIAS(\hat{\eta}) = \frac{1}{I}\sum_{i=1}^{I}(\hat{\eta}^{(i)} - \eta),$$

Tables 5.1.1-5.1.2 present the simulation results. The column labeled by η contains the values of the true model parameters. Simulation 1 shows that the REML Fisher scoring algorithm works properly because BIAS and RMSE decrease as n_d or D increase.

	η	$D = 25$	$D = 50$	$D = 100$	$D = 200$	$D = 25$	$D = 50$	$D = 100$	$D = 200$
β_{11}	1	-0.0164	0.0145	0.0101	0.0036	0.2269	0.1585	0.1129	0.0795
β_{12}	1	-0.0015	0.0006	-0.0002	-0.0002	0.0486	0.0319	0.0221	0.0171
β_{21}	1	0.0330	-0.0042	-0.0106	-0.0006	0.2283	0.1744	0.1204	0.0886
β_{22}	1	0.0003	-0.0017	0.0011	-0.0007	0.0386	0.0290	0.0198	0.0137
θ_1	0.75	-0.0027	0.0161	-0.0123	0.0026	0.1965	0.1720	0.1062	0.0818
θ_2	1	-0.0289	0.0034	-0.0197	-0.0066	0.2935	0.2143	0.1345	0.1094
θ_4	0.5	-0.0008	-0.0019	-0.0009	-0.0013	0.0428	0.0339	0.0233	0.0153
θ_5	0.75	0.0031	0.0018	-0.0004	-0.0009	0.0670	0.0534	0.0341	0.0263
θ_3	-0.8	0.0075	0.0045	-0.0002	0.0030	0.0937	0.0770	0.0498	0.0308
θ_6	0.8	-0.0019	-0.0004	0.0000	0.0000	0.0254	0.0179	0.0124	0.0084

Table 5.1.1. $BIAS(\hat{\eta})$ (left) and $RMSE(\hat{\eta})$ (right) with $n_d = 10$.

	η	$n_d=10$	$n_d=25$	$n_d=50$	$n_d=100$	$n_d=10$	$n_d=25$	$n_d=50$	$n_d=100$
β_{11}	1	-0.0164	0.0051	-0.0266	0.0097	0.2269	0.1882	0.1887	0.1755
β_{12}	1	-0.0015	-0.0007	0.0018	-0.0012	0.0486	0.0294	0.0191	0.0141
β_{21}	1	0.0330	0.0108	0.0303	0.0030	0.2283	0.2351	0.2133	0.1987
β_{22}	1	0.0003	-0.0006	-0.0008	-0.0010	0.0386	0.0249	0.0175	0.0119
θ_1	0.75	-0.0027	0.0047	-0.0048	0.0305	0.1965	0.2267	0.2235	0.2172
θ_2	1	-0.0289	0.0269	0.0058	0.0276	0.2935	0.3193	0.2797	0.2774
θ_4	0.5	-0.0008	-0.0003	-0.0012	-0.0010	0.0428	0.0291	0.0195	0.0136
θ_5	0.75	0.0031	-0.0020	-0.0006	0.0001	0.0670	0.0422	0.0293	0.0214
θ_3	-0.8	0.0075	0.0063	-0.0060	-0.0023	0.0937	0.0899	0.0816	0.0876
θ_6	0.8	-0.0019	-0.0005	-0.0002	0.0002	0.0254	0.0149	0.0108	0.0066

Table 5.1.2. $BIAS(\hat{\eta})$ (left) and $RMSE(\hat{\eta})$ (right) with $D = 25$.

If the domain sample sizes are all equal to 10 and the number of domains increases from 25 to 200, Table 5.1.1 shows that the RMSEs of all estimators decrease. If the number of domains is $D = 25$, a very small value in practice, Table 5.1.2 shows that increasing n_d helps to estimate the parameters of the variance components of the vectors e_d of random errors, but not the corresponding parameters of the vectors u_d of random effects.

5.2 Simulation 2

The target of Simulation 2 is to investigate the behavior of the domain predictors under the BNER model (3.2). For generating the population, we take $N_d = 200$, $d = 1, \dots, D$, so that $N = 200D$. The set of all units (population) and selected units (sample) are

$$U = \{u_{dj} : d = 1, \dots, D, j = 1, \dots, N_d\}, \quad s = \{u_{dj} : d = 1, \dots, D, j = 1, \dots, n_d\} \subset U.$$

For each $u_{dj} \in U$, we generate the auxiliary variables in the same way as in Simulation 1. The steps of Simulation 2 are

1. Generate x_{djk}, $d = 1, \dots, D$, $j = 1, \dots, N_d$, $k = 1, 2$. Construct the population matrices X_d and Z_d of dimensions $2N_d \times p$ and $2N_d \times 2$ respectively.
2. Repeat $I = 10^4$ times ($i = 1, \dots, 10^4$)

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

   2.2. Calculate the domain means and ratios, i.e.
   $$\eta_{dk}^{(i)} = \overline{Y}_{dk}^{(i)} = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{djk}^{(i)}, \quad \eta_{d3}^{(i)} = R_d^{(i)} = \frac{\overline{Y}_{d1}^{(i)}}{\overline{Y}_{d1}^{(i)} + \overline{Y}_{d2}^{(i)}}, \quad d = 1, \dots, D,\ k = 1, 2.$$

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

   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. Calculate the EBLUPs of $\overline{Y}_{dk}^{(i)}$ and the plug-in ratio predictor of $R_d^{(i)}$, i.e.
   $$\hat{\eta}_{dk}^{(i)} = \hat{\overline{Y}}_{dk}^{eb(i)}, \quad \hat{\eta}_{d3}^{(i)} = \hat{R}_d^{in(i)}, \quad d = 1, \dots, D,\ k = 1, 2.$$

3. For $d = 1, \ldots, D$ and $k = 1, 2, 3$, calculate the absolute performance measures

$$RE_{dk} = \Big(\frac{1}{I}\sum_{i=1}^{I}(\hat{\eta}_{dk}^{(i)} - \eta_{dk}^{(i)})^2\Big)^{1/2}, \; B_{dk} = \frac{1}{I}\sum_{i=1}^{I}(\hat{\eta}_{dk}^{(i)} - \eta_{dk}^{(i)}), \; M_{dk} = \frac{1}{I}\sum_{i=1}^{I}\hat{\eta}_{dk}^{(i)},$$

4. For $d = 1, \ldots, D$, $k = 1, 2, 3$, calculate the relative performance measures

$$RRE_{dk} = \frac{RE_{dk}}{M_{dk}}100, \; RB_{dk} = \frac{B_{dk}}{M_{dk}}100, \; RRE_k = \frac{1}{D}\sum_{d=1}^{D} RRE_{dk}, \quad ARB_k = \frac{1}{D}\sum_{d=1}^{D}|RB_{dk}|.$$

Table 5.2.1 presents the simulation results for $\hat{\overline{Y}}_{d1}^{eb}$ $(k = 1)$, $\hat{\overline{Y}}_{d2}^{eb}$ $(k = 2)$ and $\hat{R}_d^{in}$ $(k = 3)$. As expected, the performance measures decrease as the sample size n_d increases. However, if the sample size remains fixed and the number of domains D increases then the reduction of bias and MSE is small. This is due to the fact that the number of domain parameters (means and ratios) also increase when D increases.

Figures 5.2.1 and 5.2.2 presents the boxplots of biases B_{dk} and root-MSEs E_{dk} respectively. The figures shows that the three predictors are basically unbiased and that root-MSEs decrease if sample sizes increase. Further, the variance (and not the bias) gives the main contribution to the root-MSE.

D	k	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
25	1	0.0343	0.0260	0.0161	0.0081	4.6486	3.0845	2.0983	1.2370
	2	0.0394	0.0254	0.0173	0.0110	5.0017	3.3268	2.2643	1.3389
	3	0.0221	0.0120	0.0076	0.0035	2.1821	1.3585	0.8987	0.5243
50	1	0.0331	0.0236	0.0179	0.0073	4.5547	3.0608	2.0938	1.2360
	2	0.0425	0.0265	0.0189	0.0093	4.9271	3.3193	2.2674	1.3411
	3	0.0192	0.0118	0.0071	0.0041	2.1727	1.3563	0.9019	0.5239
100	1	0.0358	0.0253	0.0179	0.0102	4.4957	3.0426	2.0841	1.2337
	2	0.0371	0.0263	0.0182	0.0112	4.8706	3.3010	2.2643	1.3392
	3	0.0186	0.0107	0.0073	0.0040	2.1592	1.3549	0.8997	0.5236
200	1	0.0350	0.0228	0.0156	0.0305	4.4750	3.0357	2.0810	1.2350
	2	0.0388	0.0263	0.0187	0.0316	4.8511	3.2984	2.2634	1.3419
	3	0.0215	0.0125	0.0076	0.0122	2.1552	1.3516	0.8980	0.5240

Table 5.2.1. ARB_k (left) and RRE_k (right), $\rho_u = -0.8$, $\rho_e = 0.8$.

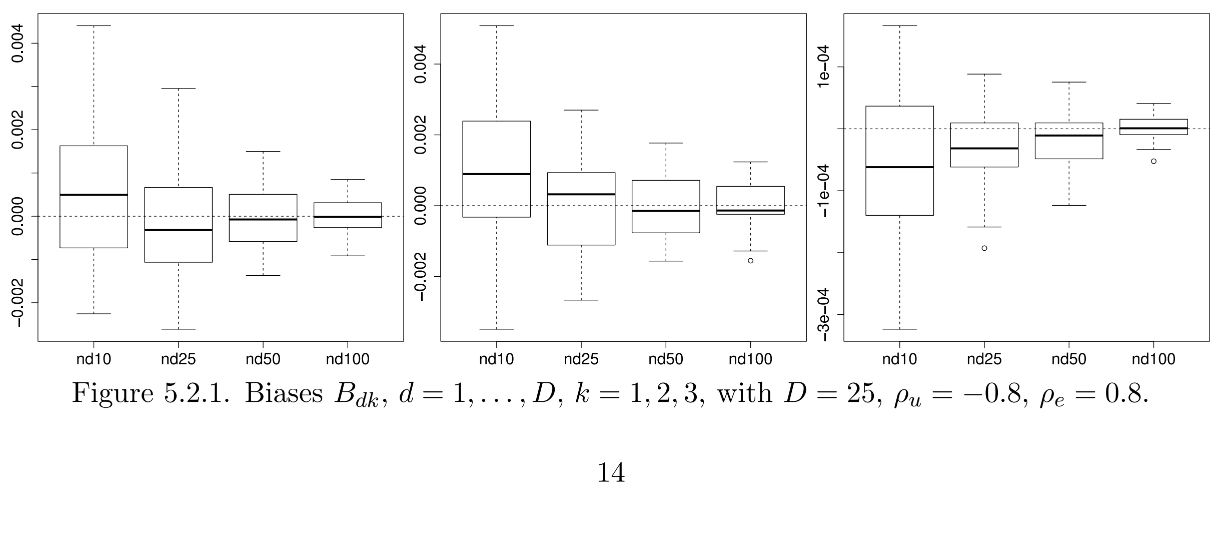

Figure 5.2.1. Biases B_{dk}, $d = 1, \ldots, D$, $k = 1, 2, 3$, with $D = 25$, $\rho_u = -0.8$, $\rho_e = 0.8$.

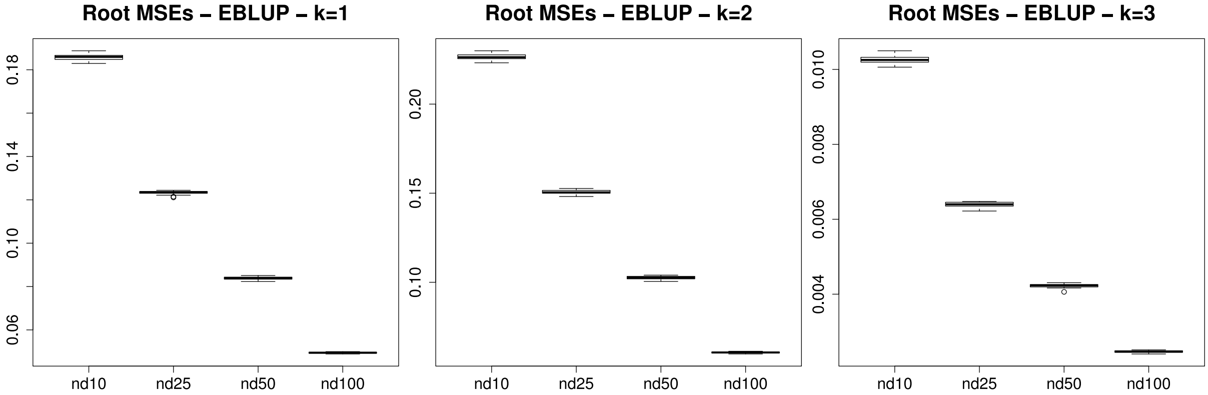

Figure 5.2.2. RE_{dk}, $d = 1, \dots, D$, $k = 1, 2, 3$, with $D = 25$, $\rho_u = -0.8$, $\rho_e = 0.8$.

We run new simulations for comparing the EBLUPs based on the BNER model with the EBLUPs based on the two independent NER models. All simulation settings remain the same, with the exception of $\theta_3 = \rho_u$ and $\theta_6 = \rho_e$. In the case $\rho_u = \rho_e = 0$, the independent NER models generate the data. In all the remaining cases, the BNER model generates the data. We run Fisher-scoring algorithms for calculating the REML estimators of the NER and BNER model parameters. Table 5.2.3 presents the simulation results and the medians of the computational times (c.time in seconds) that Fisher-scoring algorithms take to calculate the REML estimators of the model parameters in all iterations. The "Predictor" column indicates in which model (NER or BNER) the predictor is based. We observe that the BNER model provides better results than the NER model if the correlations ρ_u and ρ_e have a different sign. In the remaining cases, both procedures behave similarly. On the other hand, fitting two separate and independent NER models has much lower computational cost.

				ARB_k			RRE_k		c.time
Predictor	ρ_u	ρ_e	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	seconds
NER	0.0	0.0	0.1363	0.1246	0.1278	2.1921	2.3107	1.7847	0.035
BNER			0.1372	0.1235	0.1273	2.1930	2.3117	1.7857	12.446
NER	-0.8	-0.8	0.1367	0.1441	0.1536	2.1805	2.3161	2.2886	0.035
BNER			0.1364	0.1456	0.1534	2.1839	2.3192	2.2914	12.418
NER	0.8	0.8	0.1203	0.1121	0.0636	2.1817	2.3228	0.8477	0.035
BNER			0.1206	0.1135	0.0620	2.1829	2.3265	0.8473	12.452
NER	0.8	-0.8	0.1329	0.1453	0.1762	2.1829	2.3123	2.4353	0.035
BNER			0.1372	0.1310	0.1623	2.1112	2.2586	2.3844	12.467
NER	-0.8	0.8	0.1246	0.1125	0.0676	2.1835	2.3224	0.9362	0.035
BNER			0.1096	0.1058	0.0627	2.1323	2.2509	0.9011	12.499

Table 5.2.3. Simulation results for $D = 25$ and $n_d = 50$.

5.3 Simulation 3

The target of Simulation 3 is to investigate the behavior of the MSE estimators of the EBLUPs under the BNER model (3.2). The population and sample data is generated in the same way as in Simulation 2. The steps of Simulation 3 are

1. 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 $MSE_{dk} = (RE_{dk})^2$, $d = 1, \dots, D$, $k = 1, 2, 3$, from the output of Simulation 2.
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 $y_{dj}^{(i)} = X_{dj}\beta + u_d^{(i)} + e_{dj}^{(i)}$, $d = 1, \dots, D$ $j = 1, \dots, N_d$.
    - 3.2. Extract the sample (y_{dj}, X_{dj}), $d = 1, \dots, D$, $j = 1, \dots, n_d$.
    - 3.3. Calculate the REML estimators $\hat{\beta}_{11}^{(i)}, \hat{\beta}_{12}^{(i)}, \hat{\beta}_{21}^{(i)}, \hat{\beta}_{22}^{(i)}, \hat{\theta}_1^{(i)}, \dots, \hat{\theta}_6^{(i)}$.
    - 3.4 Calculate $mse_{dk}^{(i)} = mse\big(\hat{\overline{Y}}_{dk}^{eb(i)}\big)$, $k = 1, 2$, and $mse_{d3}^{(i)} = mse\big(\hat{R}_d^{in(i)}\big)$.
4. For $d = 1, \dots, D$, $k = 1, 2, 3$, calculate the absolute performance measures

$$RE_{dk} = \Big(\frac{1}{I}\sum_{i=1}^{I}(mse_{dk}^{(i)} - MSE_{dk})^2\Big)^{1/2}, \quad B_{dk} = \frac{1}{I}\sum_{i=1}^{I}(mse_{dk}^{(i)} - MSE_{dk}),$$

5. For $d = 1, \dots, D$, $k = 1, 2, 3$, calculate the relative performance measures

$$RRE_{dk} = \frac{100 RE_{dk}}{MSE_{dk}},\ RB_{dk} = \frac{100 B_{dk}}{MSE_{dk}},\ RRE_k = \frac{1}{D}\sum_{d=1}^{D} RRE_{dk},\ ARB_k = \frac{1}{D}\sum_{d=1}^{D} |RB_{dk}|.$$

Tables 5.3.1 presents the simulation results for $\hat{\overline{Y}}_{d1}^{eb}$ ($k = 1$), $\hat{\overline{Y}}_{d2}^{eb}$ ($k = 2$) and $\hat{R}_d^{in}$ ($k = 3$). We obtain similar results as in Simulation 2. The performance measures decrease as the sample size n_d increases. If the sample size remains fixed and the number of domains D increases then the reduction of bias and MSE is small.

D	k	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
25	1	330.97	126.43	54.98	18.42	362.03	141.03	61.71	20.55
	2	337.20	128.91	56.14	18.35	369.30	143.86	62.93	20.53
	3	100.90	41.85	19.07	6.17	220.53	118.66	79.71	58.99
50	1	327.05	120.63	52.82	18.04	343.49	127.95	56.04	19.15
	2	333.46	122.57	54.24	18.36	350.65	130.10	57.55	19.51
	3	98.43	40.45	18.37	6.80	214.56	118.19	81.23	61.94
100	1	306.09	117.29	49.98	16.74	312.07	120.15	51.43	17.22
	2	311.92	119.28	50.62	17.16	317.90	122.19	52.08	17.64
	3	89.25	37.01	16.21	5.70	187.73	106.81	74.64	58.51
200	1	319.79	117.42	50.53	16.88	324.21	119.12	51.31	17.17
	2	327.10	119.47	51.24	17.21	331.55	121.22	52.03	17.49
	3	93.18	38.18	17.29	7.26	193.19	108.73	76.48	60.19

Table 5.3.1. ARB_k (left) and RRE_k (right), $\rho_u = -0.8$, $\rho_e = 0.8$.

As Table 5.3.1 contains aggregated information, we give information about domain-level non-relative performance measures (bias and MSE). Figures 5.3.1 and 5.3.2 presents the boxplots of biases B_{dk} and root-MSEs E_{dk} of the MSE estimators respectively. The figures shows that the three predictors have a positive bias that decreases as the sample size increases. Further, the bias (and not the variance) gives the main contribution to the root-MSE of the MSE estimators. The bad news, is that the introduced analytic MSE estimator has not performed well in domains with very small sample sizes (e.g. $n_d \leq 25$). This fact lets an open door to investigate MSE estimators based on resampling procedures; for example, by adapting the bootstrap procedures of González-Manteiga et al. (2008a) to the multivariate case.

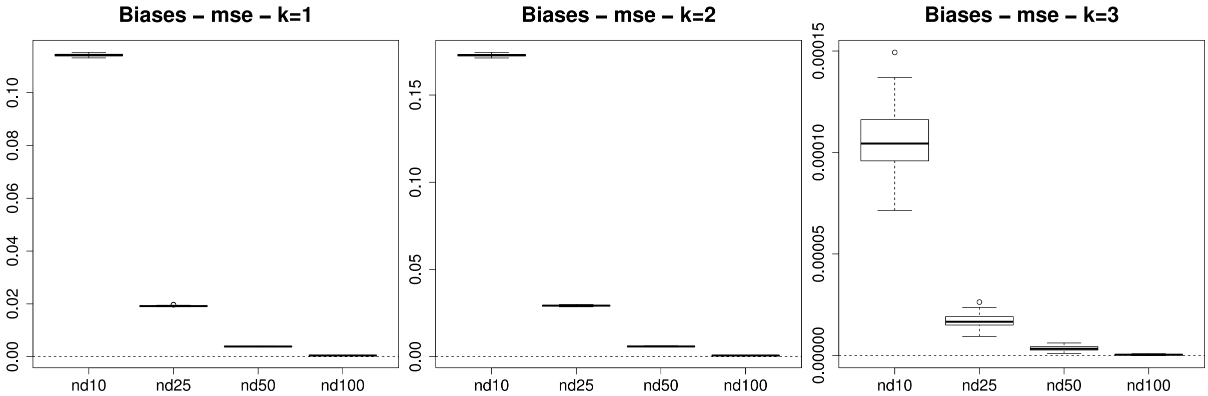

Figure 5.3.1. Biases B_{dk}, $d=1,\ldots,D$, $k=1,2,3$, with $D=25$, $\rho_u=-0.8$, $\rho_e=0.8$.

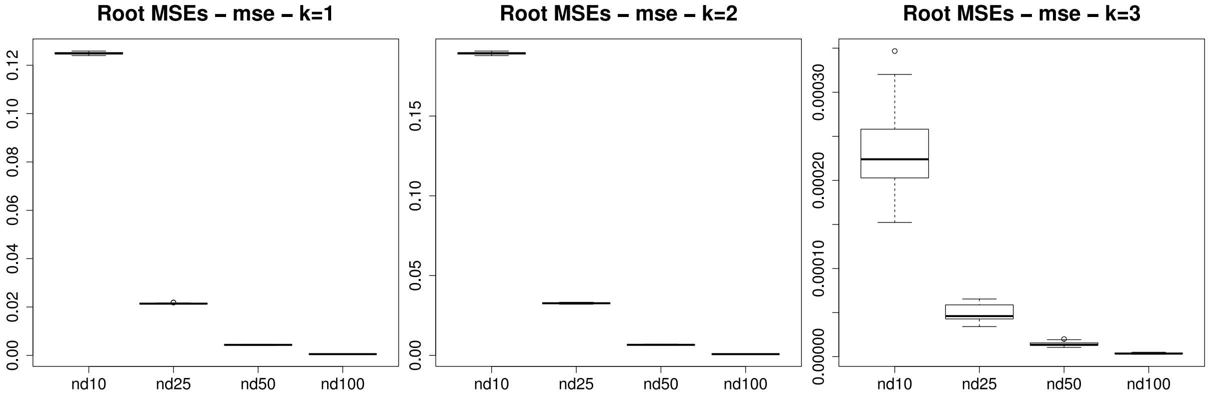

Figure 5.3.2. RE_{dk}, $d=1,\ldots,D$, $k=1,2,3$, with $D=25$, $\rho_u=-0.8$, $\rho_e=0.8$.

6 Application to Spanish household budget survey data

This section applies the developed SAE methodology to data from the SHBS of 2016. The first step is to fit a BNER model to the target vectors (y_{dj1}, y_{dj2}) containing the food and non-food annual expenses of households and the auxiliary variables x_{djk}, $d=1,\ldots,D$, $j=1,\ldots,n_d$, $k=1,2$, described in Section 2. The variables Income and NCU are treated as covariables and the variables FC and Rural as factors with reference categories FC4 and R0 respectively. For each target variable, Table 6.1 presents the estimates of the regression parameters and their standard errors. It also presents the asymptotic p-values for testing the hypothesis $H_0: \beta_{kr}=0$. Table 6.2 presents the estimates of the variance and correlation parameters with their 95% asymptotic confidence intervals. 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 6.1 (left) maps the means of the household annual expenditures in food by Spanish provinces. Figure 6.1 (right) maps the estimated relative root-MSEs (RRMSE) in %. This Figure shows that expenditures on food is rather variable within Autonomous Regions.

Figure 6.2 (left) plots the ratios of household expenditures in food by Spanish provinces (in %). Figure 6.2 (right) plots the corresponding RRMSEs in %. An interesting feature observed here is that within some Autonomous Regions, the percentages of food expenditure could be rather variable. 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.

Expenditure	Variable	Estimate	z-value	St.error	p-value
Food	Intercept	0.02	1.54	0.01	0.12
	Income	0.53	38.17	0.01	0.00
	NCU	0.02	38.91	0.00	0.00
	FC1	0.05	9.29	0.01	0.00
	FC2	-0.02	2.60	0.01	0.01
	FC3	-0.03	7.71	0.00	0.00
Non-food	Intercept	0.31	8.74	0.04	0.00
	Income	6.89	105.94	0.07	0.00
	NCU	0.04	27.01	0.00	0.00
	R1	0.04	2.41	0.02	0.02

Table 6.1: Regression parameters of the fitted BNER model.

	Estimate	Lower.lim	Upper.lim
σ^2_{u1}	0.002	0.001	0.003
σ^2_{u2}	0.025	0.014	0.037
ρ_u	0.552	0.324	0.781
σ^2_{e1}	0.057	0.056	0.058
σ^2_{e2}	1.261	1.237	1.285
ρ_e	0.201	0.188	0.214

Table 6.2: Variance and correlation parameters.

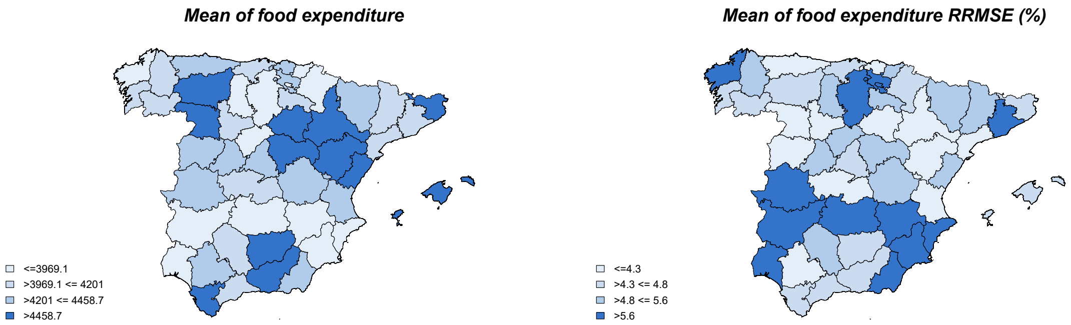

Figure 6.1: Means $\hat{\overline{Y}}^{eb}_{d1}$ (left) and their relative root-MSEs in % (right) of household annual expenditures in food by Spanish provinces.

In contrast, there are other regions, such as Cataluña and Basque Country where the variability of the estimated ratios is smaller.

For the sake of comparability, two separate and independent NER models are fitted with the same auxiliary variables appearing in Table 6.1. The two NER models are incorrect models since we have assumed that the fitted BNER model is the true model. Therefore the predictors obtained using the EBLUP calculation formulas under a NER model are not EBLUPs and are called INDEP.

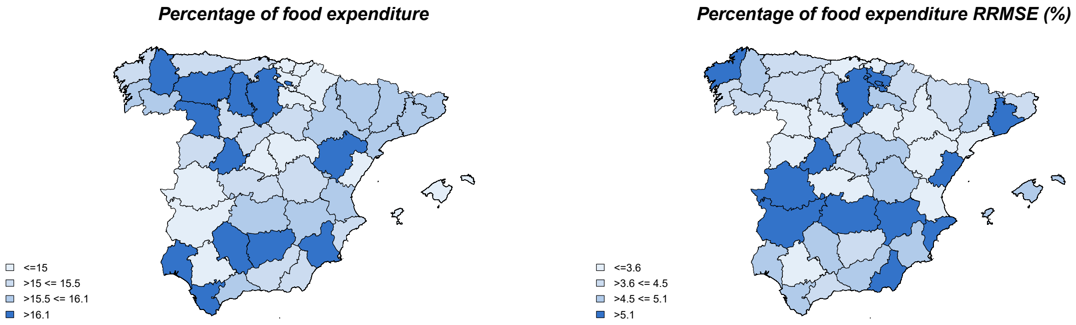
Percentage of food expenditure
<=15
>15 <= 15.5
>15.5 <= 16.1
>16.1
Percentage of food expenditure RRMSE (%)
<=3.6
>3.6 <= 4.5
>4.5 <= 5.1
>5.1

Figure 6.2: Ratios $\hat{R}_d^{in}$ in % (left) and their relative root-MSEs in % (right) of household annual expenditures in food by Spanish provinces.

Figure 6.3 plots the direct, INDEP and EBLUP estimates of $\overline{Y}_1$ (left) and $\overline{Y}_2$ (right). The domains are sorted by sample sizes and the sample size is printed in the axis OX. This figure shows that the three estimators follow the same pattern and come closer as the sample size increases, but the INDEP and EBLUP have a smoother behavior.

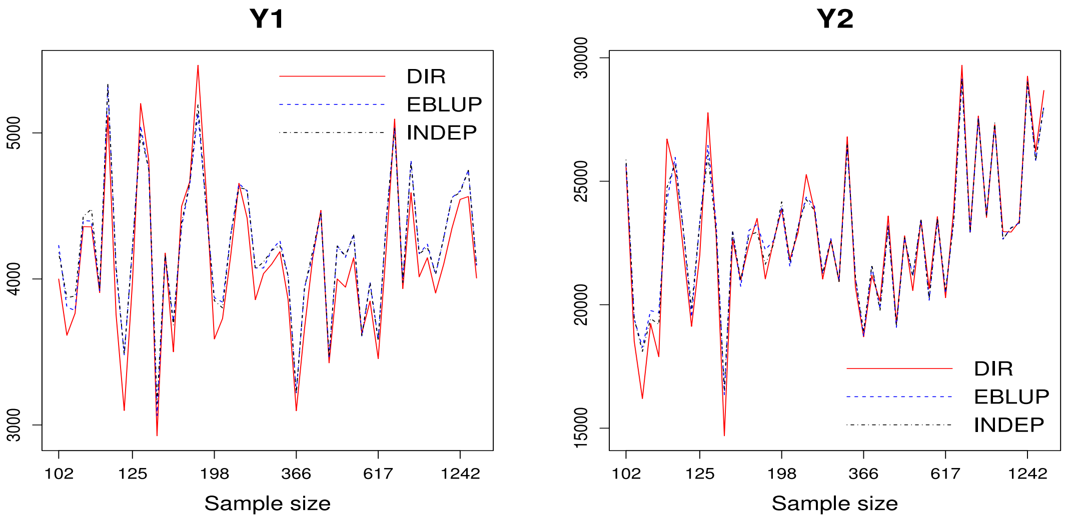

Figure 6.3: Direct and EBLUP estimates

Figure 6.4 plots the estimated RRMSE of the direct estimators and the INDEP and EBLUP predictors of $\overline{Y}_1$ (left) and $\overline{Y}_2$ (right). As before, the domains are sorted by sample size. For the estimation of the MSE of the INDEP predictors, we use the estimator mse_k^{ind}, $k=1,2$ formulas of the functions $g_1 - g_4$ that are described, for example, in Chapter 7 of Rao and Molina (2015). We recall that those formulas are here incorrect, because the assumed true model is the BNER and not the two independent and marginal NER models. We remark that a BLUP is the predictor of linear parameter that minimizes the MSE in the class of unbiased predictors and the EBLUP, with REML estimators, inherit this property asymptotically. Therefore, the MSEs of the EBLUPs should be smaller than the MSEs of the INDEP predictors under the BNER model.

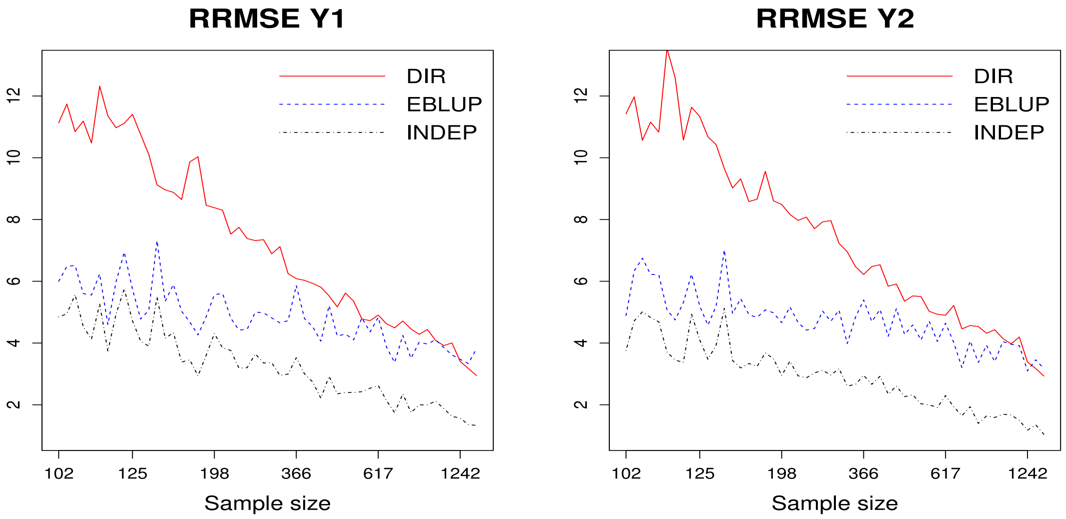

Figure 6.4: RRMSEs of direct and EBLUP estimates

Figure 6.4 shows that the EBLUPs have lower RRMSEs than the direct estimators and that the RRMSEs come closer as the sample size increases. Further, it shows that mse_k^{ind} underestimates the MSE of the INDEP predictors of $\overline{Y}_1$ and $\overline{Y}_2$. This is something quite interesting for practitioners. If we do not take into account the correlation between the two target variable, we can deliver good estimates of domain quantities, but we fail in estimating MSEs.

Figure 6.5 (left) plots the direct and indep and plug-in estimates of ratios of food expenditure in %. Figure 6.5 (right) plots estimated RRMSEs for the direct estimator and the plug-in predictors. The MSEs of the direct estimators of ratios are estimated by plug-in the design-based covariances estimators (2.2) in the formula (4.2). For the INDEP predictor, we cannot calculate MSEs because there is no way to estimate the required covariance term (cf. formula (4.2)). This figure shows that the model-based plug-in estimators have lower RRMSEs than the direct estimators and that the RRMSEs come closer as the sample size increases.

Tables 6.3 and 6.4 present some condensed numerical results. The tables has been constructed in two steps. The domains are sorted by sample size, starting by the domain with the smallest sample size. A selection of 14 domains out of 52 is done from the positions 1, 5, 9,. . . ,52. The name and code of provinces are labeled by Prov and d respectively and the sample sizes by n.

Table 6.3 presents the direct and model-based estimates of mean food and non-food household expenditures and the corresponding ratios of food expenditures by provinces. The estimators are denoted by dir1, eb1, dir2, eb2, Rdir and Rin. The lower and upper limits of the 95% confidence intervals (CIs) for R_d (in %) are in the columns labelled by Rin$^-$ and Rin$^+$ respectively. We calculate the CIs by applying the standard normality formulas to the plug-in estimates Rin and to its RMSEs. This table shows that the model-based estimates follow the pattern of direct estimates and that both estimates are closer when the sample size is large.

Table 6.4 presents the RRMSEs of direct and model-based estimators of $\overline{Y}_{d1}$, $\overline{Y}_{d2}$ and R_d. The RRMSEs are labeled by dir1, eb1, dir2, eb2, Rdir and Rin. By observing the columns of RRMSEs, we conclude that the model-based predictors are preferred to the direct estimators.

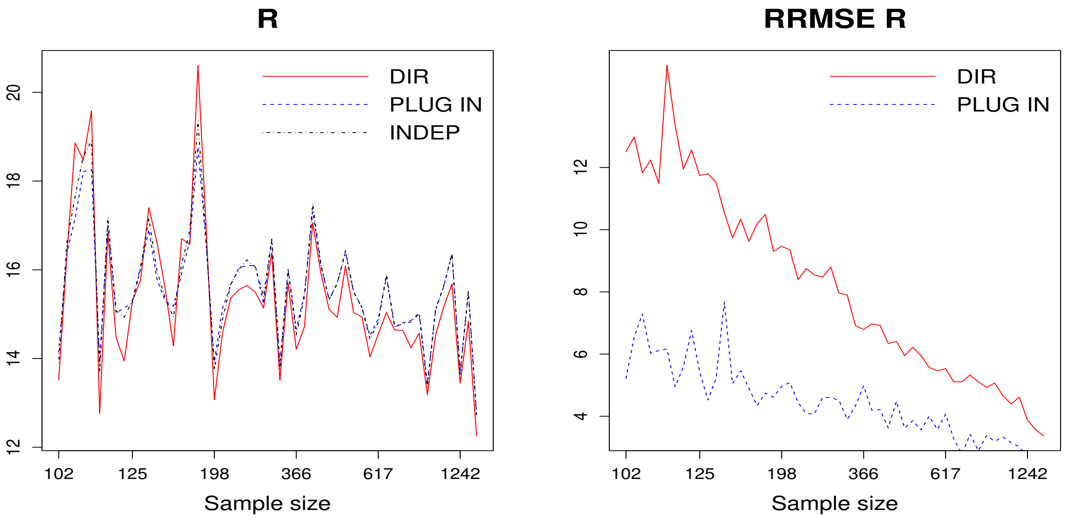

Figure 6.5: Direct and plug-in estimates of ratios (left) and their estimated RRMSEs (left).

Prov	d	n	dir1	eb1	dir2	eb2	Rdir	Rin	Rin$^-$	Rin$^+$
Guadalajara	19	102	3999	4229	25591	25686	13.52	14.14	12.69	15.59
Palencia	34	118	4357	4396	17893	19671	19.58	18.26	16.07	20.46
Cuenca	16	123	3099	3480	19123	19505	13.95	15.14	13.13	17.15
Ourense	32	169	2926	3064	14691	16354	16.61	15.78	13.40	18.16
Burgos	9	187	4666	4651	23492	23255	16.57	16.67	15.25	18.09
Granada	18	198	3729	3841	21833	21545	14.59	15.13	13.62	16.64
Albacete	2	249	3858	4075	21039	21250	15.49	16.09	14.63	17.55
Ciudad Real	13	355	3858	4018	20714	21085	15.70	16.01	14.64	17.37
Pontevedra	36	463	4469	4451	23593	23197	15.93	16.10	14.95	17.25
Coruña, A	15	536	4145	4306	23429	23464	15.03	15.51	14.42	16.59
Zaragoza	50	678	4228	4410	23889	23436	15.04	15.84	14.82	16.86
Cantabria	39	761	4014	4173	23536	23602	14.57	15.02	14.03	16.02
Murcia	30	913	4347	4557	23379	23310	15.68	16.35	15.38	17.33
Madrid	28	1653	4006	4094	28676	28021	12.26	12.75	12.04	13.46

Table 6.3: Estimates of $\overline{Y}_{d1}$, $\overline{Y}_{d2}$ and R_d and CIs for R_d (in %).

7 Conclusions

This paper introduces small area predictors of expenditure means and ratios based on the BNER model (3.1). For a given domain, the EBLUP of a linear domain parameter based on the BNER model borrows strength from the auxiliary data, the data from other domains and the correlation between the target variables. By using this model, applied statistician can obtain estimates of domain parameters that behave in a smooth an stable form across domains and target parameters. This is usually considered as a good property for official statistics. The paper also approximates the matrix of MSEs of the EBLUP and introduces an explicit-formula estimator.

The bivariate unit-level models are the most appropriate models for deriving small area predic-

Prov	d	n	dir1	eb1	dir2	eb2	Rdir	Rin
Guadalajara	19	102	11.13	6.00	11.42	4.89	12.51	5.22
Palencia	34	118	10.48	5.55	10.83	6.21	11.49	6.12
Cuenca	16	123	11.11	6.94	11.64	6.23	12.56	6.76
Ourense	32	169	9.12	7.31	9.65	7.01	10.56	7.67
Burgos	9	187	9.86	4.68	8.66	4.83	10.19	4.34
Granada	18	198	8.30	5.61	8.17	5.17	9.35	5.08
Albacete	2	249	7.32	5.00	7.93	5.03	8.47	4.61
Ciudad Real	13	355	6.25	4.72	6.48	4.83	6.91	4.33
Pontevedra	36	463	5.81	4.06	5.84	4.22	6.34	3.63
Coruña, A	15	536	5.35	4.10	5.50	4.10	5.95	3.56
Zaragoza	50	678	4.62	3.87	5.22	4.02	5.11	3.28
Cantabria	39	761	4.28	4.03	4.32	3.93	4.93	3.38
Murcia	30	913	4.00	3.61	4.19	3.92	4.62	3.02
Madrid	28	1653	2.95	3.82	2.93	3.17	3.37	2.82

Table 6.4: RRMSEs of estimators of $\overline{Y}_{d1}$, $\overline{Y}_{d2}$ and R_d (all in %).

tors of indicators depending on two target variables. The ratio of totals or means is a typical example of this kind of parameters. These parameters could be estimated by fitting univariate models to each of the response variables. The drawback of this approach is that the correlation between the response variables and the correlation between the EBLUPs of the domain means are not taken into account.

If the target domain parameters are totals or means, then the INDEP predictors based on the "incorrect" separate NER models produce loss of efficiency, with respect to the EBLUPs based on the "true" BNER model, mainly when the correlations of random effects and errors have different sign. Otherwise, the loss of efficiency is rather small. The main problem is not the INDEP predictor itself, but the corresponding MSE estimators based on the incorrect models. These estimators behave rather bad. When the target variables are positively correlated, they tend o under-estimate the MSEs, as it happens in the application to real data. This is a severe error.

In the case that the target domain parameter is a ratio, the INDEP ratio estimators could also be used. However, an appropriate estimator of the MSE of the ratio estimator could not be constructed under the independent univariate modeling. This problem can be treated and solved by using plug-in predictor of ratios based on the EBLUPs of the BNER model.

Three simulation experiments are carried out to empirically investigate and to check the behavior of the fitting algorithm, the predictors (EBLUP and plug-in) and the MSE estimators. Simulation 1 investigates the behavior of the REML fitting algorithm and empirically shows the consistency of the REML estimators of model parameters. Simulation 2 studies the gain of efficiency of the EBLUPs and plug-in predictors when using bivariate models instead of univariate ones. The conclusion is that predictors based on the BNER model (3.1) outperform the corresponding ones based on two independent NER models when correlations of random effects and random errors have different sign. Simulation 3 empirically shows that the bias and MSE of the introduced estimator of the MSE matrix decrease as the sample size increases.

The new small area estimation methodology is applied to data from the SHBS of 2016. The target is to estimate means of food and non-food household annual expenditures and ratios of household annual expenditures by Spanish provinces. The estimation procedure takes into account the correlation between the two target variables. The paper also compares the model-

based estimates with the corresponding ones obtained by applying direct Hajéck-type estimators and it shows that introduced estimators have lower MSEs than the direct estimators.
As far as the results obtained for the expenditures in the Spanish provinces, we can say that the provinces with the highest mean of household annual expenditures in food are, mainly, in the north of Spain. Also, we can conclude that the percentage of food expenditure is different within the provinces of some Autonomous Regions and this can help to the regional authorities to implement different plolicies in the provinces.

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, 1191-1209.

Boubeta, M, Lombardía, M.J., Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST*, **25**, 548-569.

Boubeta, M, Lombardía, M.J., Morales, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics and Data Analysis*, **107**, 32-47.

Benavent, R., Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis*, **94**, 372-390.

Chambers, R., Salvati, N. and Tzavidis, N. (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*, **179**, 2, 453-479.

Chandra, H., Salvati, N., Chambers, R., Tzavidis, N. (2012). Small area estimation under spatial nonstationarity *Computational Statistics and Data Analysis*, **56**, 2875-2888.

Das, K. Jiang, J., Rao, J.N.K. (2004). Mean squared error of empirical predictor. *Annals of Statistics*, **32**, 818-840.

Datta, G.S., Fay, R.E., Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. In: Proceedings of Bureau of the Census 1991 Annual Research Conference, U.S. Bureau of the Census, Washington, DC, 63-79.

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., Lahiri P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, **10**, 613-627.

Esteban, M.D., Lombardía, M.J., López-Vizcaíno, E., Morales, D., Pérez, A. (2020). Small area estimation of proportions under area-level compositional mixed models. *TEST*. DOI: 10.1007/s11749-019-00688-w.

Fay, R. E. (1987). Application of multivariate regression of small domain estimation. In: R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh (Eds.), Small Area Statistics, John Wiley, New York, 91-102.

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

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., Santamaría, L. (2008b). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. Computational Statistics and Data Analysis, **52**, 12, 5242-5252.

Hobza, T., Morales, D. (2013). Small area estimation under random regression coefficient models. *Journal of Statistical Computation and Simulation*, **83**, 11, 2160-2177.

Hobza, T. and Morales, D. (2016). Empirical Best Prediction Under Unit-Level Logit Mixed Models. *Journal of official statistics*, **32**, 3, 661-692.

Hobza, T., Morales, D. and Santamaría, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST*, **27**, N. 2., 270-294.

Ito, T., Kubokawa, T. (2018). Empirical best linear unbiased predictors in multivariate nested-error regression models. *Communications in Statistics–Theory and Methods*, DOI: 10.1080/03610926.2019.1662048.

López-Vizcaíno, E., Lombardía, M.J., Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, **13**, 2, 153-178.

López-Vizcaíno, E., Lombardía, M.J. and 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 Association, series A*, **178**, 3, 535-565.

Marchetti, S. and 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.

Molina, I., Saei, A. and Lombardía, M.J. (2007) Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of Royal Statistical Society, Series A*, **170**, 975-1000.

Molina, I. (2009). Uncertainty under a multivariate nested-error regression model with logarithmic transformation. *Journal of Multivariate Analysis*, **100**, 963-980.

Morales, M., Pagliarella, M.C., Salvatore, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. *SORT-Statistics and Operations Research Transactions*, **39**, 1, 19-34.

Morales, D., Santamaría, L. (2019). Small area estimation under unit-level temporal linear mixed models. *Journal of Statistical Computation and Simulation*, **89**, 9, 1592-1620.

Ngaruye, I., Nzabanita, J., von Rosen, D., Singull, M. (2017). Small area estimation under a multivariate linear model for repeated measures data. *Communications in Statistics - Theory and 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. *Australian & New Zealand Journal of Statistics*, **57**, 15-29.

Prasad, N.G.N., Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.

Rao, J.N.K., Molina, I. (2015). *Small area estimation*, Second Edition. John Wiley, Hoboken, New York.

Särndal, C.E., Swensson, B., Wretman, J. (1992). *Model assisted survey sampling.* Springer-Verlag.

Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, **17**, 3, 393-411.

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. New York.