



# Small area estimation of average compositions under multivariate nested error regression models

María Dolores Esteban<sup>1</sup> · María José Lombardía<sup>2</sup> · Esther López-Vizcaíno<sup>3</sup> · Domingo Morales<sup>1</sup> · Agustín Pérez<sup>1</sup>

Received: 13 March 2022 / Accepted: 22 January 2023  
© The Author(s) 2023

## Abstract

This paper investigates the small area estimation of population averages of unit-level compositional data. The new methodology transforms the compositions into vectors of  $R^m$  and assumes that the vectors follow a multivariate nested error regression model. Empirical best predictors of domain indicators are derived from the fitted model, and their mean squared errors are estimated by parametric bootstrap. The empirical analysis of the behavior of the introduced predictors is investigated by means of simulation experiments. An application to real data from the Spanish household budget survey is given. The target is to estimate the average of proportions of annual household expenditures on food, housing and others, by Spanish provinces.

**Keywords** Household budget survey · Small area estimation · Multivariate nested error regression model · Compositional data · Bootstrap · Household expenditures

**Mathematics Subject Classification** 62E30 · 62J12

---

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.

---

✉ María José Lombardía  
maria.jose.lombardia@udc.es

<sup>1</sup> Universidad Miguel Hernández de Elche, Elche, Spain

<sup>2</sup> Universidade da Coruña, CITIC, A Coruña, Spain

<sup>3</sup> Instituto Galego de Estatística, Santiago de Compostela, Spain

## 1 Introduction

Official statistics contain estimates of socioeconomic indicators at different levels of aggregation. In many sampling designs, small sample sizes do not allow accurate direct estimators to be calculated at low levels of aggregation. These territories or population groups are called small areas. Small Area Estimation (SAE) gives a solution to this problem by incorporating auxiliary information to the data analysis and by introducing model-based predictors. The books of Rao and Molina (2015) and Morales et al. (2021) give a general description of SAE.

The Spanish household budget survey (SHBS) provides information about the nature and destination of the consumption household expenses, as well as on various characteristics related to the conditions of household life. Spain is hierarchically partitioned in 17 autonomous communities and 50 provinces, plus 2 autonomous cities. The sampling design and the sample sizes of the SHBS are developed to provide estimates for the 17 autonomous communities level, but not for the provinces. The direct estimates at the province level have a low accuracy and, therefore, estimating SHBS indicators at that level is a SAE problem. This paper has two objectives. The first one is to model the unit-level proportions of annual household expenditures on food, housing and others. The second one is to estimate the average of these proportions, by Spanish provinces.

Under area-level models, we find some more proposals for estimating domain proportions and counts. For example, Esteban et al. (2012), Marhuenda et al. (2013, 2014) and Morales et al. (2015) derived predictors based on linear mixed models and (Chambers et al. 2014; Dreassi et al. 2014; Tzavidis et al. 2015) and (Boubeta et al. 2017, 2016) applied binomial, negative binomial or Poisson regression models. There are also methodologies for estimating proportions and counts in the setup of contingency tables or multinomial regression models. Without being exhaustive, we find the papers of Zhang and Chambers (2004), Berg and Fuller (2014) for contingency tables, and the papers of Ferrante and Trivisano (2010), Souza and Moura (2016), Fabrizi et al. (2016), Saei and Chambers (2003), Molina et al. (2007) and López-Vizcaíno et al. (2013, 2015) for multinomial regression models. However, in the household survey samples, some variables of interest and domain indicators are compositions. This is to say, they are positive quantities summing up to one or to a known integer number. Concerning area-level model for compositional data, Esteban et al. (2020) and Krause et al. (2022) transformed compositions into target vectors of multivariate Fay-Herriot models in order to make model-based predictions, like the ones described by González-Manteiga et al. (2008a), Benavent and Morales (2016), Benavent and Morales (2021) or Arima et al. (2017).

The statistical literature presents some contributions to small area estimation of proportions and counts under unit-level models for binary outcomes. For example, Chambers et al. (2016), Hobza and Morales (2016), Hobza et al. (2018) and Burgard et al. (2021) derived predictors under M-quantile or binomial-logit models for binary outcomes. These approaches are based on univariate models and not in models for compositional data that consider the possibility of jointly estimating the counts or proportions of all the categories of a classification variable. This issue was faced by Scealy and Welsh (2017), which introduced a directional mixed effects model for

compositional data and predicted the proportions of total weekly expenditure on food and housing costs for households in a chosen set of domains. A different approach was employed by Hijazi and Jernigan (2009), Camargo et al. (2012), Tsagris and Stewart (2018), Morais et al. (2018), which modelled compositional data using Dirichlet regression models. This manuscript also deals with unit-level compositional data, but it proposes to fit multivariate linear mixed models to logratio transformations of compositions. Some references on the foundations of compositional data analysis are the books (Aitchison 1986) and (Pawlowsky-Glahn and Buccianti 2011) and the papers (Egozcue et al. 2003) and (Egozcue and Pawlowsky-Glahn 2019), where some basic transformations of compositions are studied.

This paper introduces small area predictors of averages of unit-level vectors of compositions. For this sake, the paper considers three logratio transformations of compositions into vectors of  $R^m$ . They are the additive, centered and isometric logratio transformations. We propose a multivariate nested error regression (MNER) model for analyzing the transformed SHBS compositional data, where the vectors of random effects and the vector of model errors have unstructured covariance matrices with unknown components. The estimates of the MNER model parameters are obtained by using the residual maximum likelihood (REML) estimation method, as it is described in Esteban et al. (2022a). The fitted model is then used to predict averages of proportions of annual household expenditures on food, housing and others, by Spanish provinces. The empirical best and plug-in predictors of small area compositional parameters are derived similarly as in Esteban et al. (2022b).

The estimation of the mean squared error (MSE) of a model-based predictor is an important issue that has no easy solution. Under nonlinear models, the problem is even more difficult. We follow the resampling approach appearing in González-Manteiga et al. (2007, 2008b) to implement a parametric bootstrap procedure.

This paper introduces statistical methodology that is new in four main aspects: (1) the employment of three transformations of unit-level compositional survey data, (2) the use of MNER models with unstructured covariance matrix for modelling the transformed data and capturing the sample correlations, (3) the derivation of domain-level predictors of averages of compositions based on the MNER model fitted to the transformed unit-level data, and (4) the introduction of parametric bootstrap estimators of the MSEs of the new predictors.

The remainder of the paper is organized as follows: Section 2 establishes the probabilistic framework, describes the SAE problem of interest and presents the MNER model. Section 3 derives empirical best predictors (EBP) and plug-in predictors of average compositions and gives a parametric bootstrap method for estimating the MSEs of the EBPs. Section 4 presents three simulation experiments. The target of Simulation 1 is to check the behavior of the REML algorithm for fitting the MNER model. Simulation 2 investigates the performance of the EBPs and plug-in predictors, and Simulation 3 analyzes the parametric bootstrap estimator of the MSEs. Section 5 applies the proposed methodology to data from the SHBS of 2016 in Spain. Section 6 gives some conclusions. The paper contains four appendices in a supplementary material file. Appendix A describes the additive, centered and isometric logratio transformations of compositions. Appendix B gives further simulation results. Appendix

C analyzes the SHBS data with different transformations. Appendix D performs the application to SHBS data without applying logratio transformations of compositions.

## 2 The probabilistic framework

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 us consider the probability vector  $a_{dj}^+ = (a_{dj1}, \dots, a_{djm+1})' \in R^{m+1}$  representing proportions associated with the  $m+1$  categories of a classification variable that is defined on the sample unit  $j$  of domain  $d$ ,  $d = 1, \dots, D$ ,  $j = 1, \dots, N_d$ . For example,  $a_{dj}^+$  may contain the proportions of annual household expenditures in the different expense categories. The components of  $a_{dj}^+$  are nonnegative and fulfill the constraint  $a_1 + \dots + a_{m+1} = 1$ . These vectors  $a_{dj}^+$  are called compositions or  $(m+1)$ -part compositions, and vectors  $a_{dj} = (a_{dj1}, \dots, a_{djm})'$  are called  $m$ -part compositions. Compositional data, consisting of compositions, play an important role in public statistics. Compositions take values in the simplex embedded in  $R^{m+1}$

$$S_e^m = \{(a_1, \dots, a_{m+1})' \in R^{m+1} : a_1 > 0, \dots, a_{m+1} > 0, a_1 + \dots + a_{m+1} = 1\},$$

and  $m$ -part compositions take values in the  $m$ -dimensional simplex defined by

$$S^m = \{(a_1, \dots, a_m)' \in R^m : a_1 > 0, \dots, a_m > 0, a_1 + \dots + a_m < 1\}.$$

This paper deals with the problem of predicting domain average compositions

$$A_{dk} = \frac{1}{N_d} \sum_{j=1}^{N_d} a_{dj k}, \quad d = 1, \dots, D, \quad k = 1, \dots, m+1, \quad (2.1)$$

under a compositional data analysis approach. This is to say, we apply a one-to-one transformation,  $h = (h_1, \dots, h_m)' : S^m \mapsto R^m$ , to  $m$ -part compositions and we assume that the transformed vectors follow a multivariate regression model. Appendix A presents three widely employed transformations. They are the additive, centered and isometric logratio transformations. The components of the transformed vectors  $y_{dj} = h(a_{dj}) = (y_{dj1}, \dots, y_{djm})'$  are continuous variables measured on the sample unit  $j$  of domain  $d$ ,  $d = 1, \dots, D$ ,  $j = 1, \dots, N_d$ .

For  $k = 1, \dots, m$ , let  $x_{dj k} = (x_{dj k1}, \dots, x_{dj k p_k})$  be a row vector containing  $p_k$  explanatory variables and let  $X_{dj} = \text{diag}(x_{dj1}, \dots, x_{dj m})_{m \times p}$  with  $p = p_1 + \dots + p_m$ . Let  $\beta_k$  be a column vector of size  $p_k$  containing regression parameters and let  $\beta = (\beta_1', \dots, \beta_m')'_{p \times 1}$ . We assume that the transformed vectors  $y_{dj}$ 's follow the population MNER model

$$y_{dj} = X_{dj} \beta + u_d + e_{dj}, \quad d = 1, \dots, D, \quad j = 1, \dots, N_d, \quad (2.2)$$

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

$$u_d = (u_{d1}, \dots, u_{dm})' \sim N_m(0, V_{ud}), \quad e_{dj} = (e_{dj1}, \dots, e_{djm})' \sim N_m(0, V_{edj}).$$

The  $m \times m$  covariance matrices  $V_{ud}$  depend on  $q = m(m + 1)/2$  unknown parameters, denoted by

$$\theta_u = (\theta_{u1}, \dots, \theta_{uq})' = (\sigma_{u1}^2, \sigma_{u2}^2, \dots, \sigma_{um}^2, \rho_{u12}, \rho_{u13}, \dots, \rho_{u23}, \rho_{u24}, \dots, \rho_{um-1,m})'.$$

The matrix  $V_{ud}$  is

$$V_{ud} = \begin{pmatrix} \sigma_{u1}^2 & \rho_{u12}\sigma_{u1}\sigma_{u2} & \dots & \rho_{u1m}\sigma_{u1}\sigma_{um} \\ \rho_{u12}\sigma_{u1}\sigma_{u2} & \sigma_{u2}^2 & \dots & \rho_{u2m}\sigma_{u2}\sigma_{um} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{u1m}\sigma_{u1}\sigma_{um} & \rho_{u2m}\sigma_{u2}\sigma_{um} & \dots & \sigma_{um}^2 \end{pmatrix}.$$

The  $m \times m$  covariance matrices  $V_{edj}$  depend on  $q$  unknown parameters, i.e.

$$\theta_e = (\theta_{e1}, \dots, \theta_{eq})' = (\sigma_{e1}^2, \sigma_{e2}^2, \dots, \sigma_{em}^2, \rho_{e12}, \rho_{e13}, \dots, \rho_{e23}, \rho_{e24}, \dots, \rho_{em-1,m})'.$$

The matrix  $V_{edj}$  is

$$V_{edj} = \begin{pmatrix} \sigma_{e1}^2 & \rho_{e12}\sigma_{e1}\sigma_{e2} & \dots & \rho_{e1m}\sigma_{e1}\sigma_{em} \\ \rho_{e12}\sigma_{e1}\sigma_{e2} & \sigma_{e2}^2 & \dots & \rho_{e2m}\sigma_{e2}\sigma_{em} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{e1m}\sigma_{e1}\sigma_{em} & \rho_{e2m}\sigma_{e2}\sigma_{em} & \dots & \sigma_{em}^2 \end{pmatrix}.$$

The  $2q \times 1$  vector of variance component parameters is  $\theta = (\theta'_u, \theta'_e)'$ . The  $(p + 2q) \times 1$  vector of model parameters is  $\psi = (\beta', \theta')'$ . Let  $I_a$  be the  $a \times a$  identity matrix. We define the  $mN_d \times 1$  vectors  $y_d$  and  $e_d$ , the  $mN_d \times p$  matrix  $X_d$  and the  $mN_d \times m$  matrix  $Z_d$  as follows:

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

Model (2.2) can be written in the domain-level form

$$y_d = X_d\beta + Z_d u_d + e_d, \quad d = 1, \dots, D, \tag{2.3}$$

where the vectors  $u_d$  and  $e_d \sim N_{mN_d}(0, V_{ed})$  are independent and  $V_{ed} = \underset{1 \leq j \leq N_d}{\text{diag}} (V_{edj})$ . We define the  $mN \times 1$  vectors  $y$  and  $e$ , the  $mD \times 1$  vector  $u$ , the  $mN \times p$  matrix  $X$  and  $mN \times mD$  matrix  $Z$  as follows:

$$y = \underset{1 \leq d \leq D}{\text{col}} (y_d), \quad e = \underset{1 \leq d \leq D}{\text{col}} (e_d), \quad u = \underset{1 \leq d \leq D}{\text{col}} (u_d), \quad X = \underset{1 \leq d \leq D}{\text{col}} (X_d),$$

$$Z = \text{diag} (Z_d), \quad 1 \leq d \leq D$$

Model (2.2) can be written in the linear mixed model form

$$y = X\beta + Zu + e. \tag{2.4}$$

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

Under the predictive approach to inference in finite populations, statistical procedures are based on a fixed subset (called sample),  $s = \cup_{d=1}^D s_d$ , of the finite population  $U$ . Let  $n_d$  be the size of the domain subset  $s_d \subset U_d$ ,  $d = 1, \dots, D$ , and let  $n = n_1 + \dots + n_D$  be the total sample size. The complementary domain subsets are  $r_d = U_d - s_d$ ,  $d = 1, \dots, D$ . Let  $y_s$  and  $y_{ds}$  be the sub-vectors of  $y$  and  $y_d$  corresponding to sample elements and  $y_r$   $y_{dr}$  the sub-vectors 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  $X_d$ ,  $Z_d$  and  $V_d$ . As we assume that sample indexes are fixed, then the sample sub-vectors  $y_{ds}$  follow the marginal models derived from the population model (2.3), i.e.

$$y_{ds} = X_{ds}\beta + Z_{ds}u_d + e_{ds}, \quad d = 1, \dots, D, \tag{2.5}$$

where  $u_d \sim N_m(0, V_{ud})$ ,  $e_{ds} \sim N_{m n_d}(0, V_{eds})$  are independent and  $V_{eds} = \text{diag} (V_{edj})$ . The vectors  $y_{ds}$  are independent with  $y_{ds} \sim N_{n_d}(\mu_{ds}, V_{ds})$ ,  $\mu_{ds} = X_{ds}\beta$ ,  $V_{ds} = Z_{ds}V_{ud}Z'_{ds} + V_{eds}$ .

When the variance component parameters are known, the best linear unbiased estimator (BLUE) of  $\beta$  and the best linear unbiased predictor (BLUP) of  $u_d$ ,  $d = 1, \dots, D$ , are

$$\hat{\beta}_B = \left( \sum_{d=1}^D X'_{ds} V_{ds}^{-1} X_{ds} \right)^{-1} \sum_{d=1}^D X'_{ds} V_{ds}^{-1} y_{ds}, \quad \hat{u}_{dB} = V_{ud} Z'_{ds} V_{ds}^{-1} (y_{ds} - X_{ds} \hat{\beta}_B).$$

Let  $\hat{\theta}$  be the REML estimator of  $\theta$ , then the empirical BLUE (BLUE) of  $\beta$  and the empirical BLUP (EBLUP) of  $u_d$ ,  $d = 1, \dots, D$ , are

$$\hat{\beta} = \left( \sum_{d=1}^D X'_{ds} \hat{V}_{ds}^{-1} X_{ds} \right)^{-1} \sum_{d=1}^D X'_{ds} \hat{V}_{ds}^{-1} y_{ds}, \quad \hat{u}_d = \hat{V}_{ud} Z'_{ds} \hat{V}_{ds}^{-1} (y_{ds} - X_{ds} \hat{\beta}),$$

where  $\hat{V}_{ds}$  and  $\hat{V}_{ud}$  are obtained by substituting  $\theta$  by  $\hat{\theta}$  in  $V_{ds}$  and  $V_{ud}$ , respectively. We calculate the inverse of  $V_{ds} = V_{eds} + Z_{ds} V_{ud} Z'_{ds} = A + BCD$  by applying the formula

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}.$$

As  $Z'_{ds} V_{eds}^{-1} Z_{ds} = \sum_{j=1}^{n_d} V_{edj}^{-1} = n_d V_{edj}^{-1}$ , we obtain

$$\begin{aligned} V_{ds}^{-1} &= V_{eds}^{-1} - V_{eds}^{-1} Z_{ds} (V_{ud}^{-1} + Z'_{ds} V_{eds}^{-1} Z_{ds})^{-1} Z'_{ds} V_{eds}^{-1} \\ &= V_{eds}^{-1} - V_{eds}^{-1} Z_{ds} (V_{ud}^{-1} + n_d V_{edj}^{-1})^{-1} Z'_{ds} V_{eds}^{-1}. \end{aligned}$$

As the sample indexes are fixed, the out-of-sample sub-vectors  $y_{dr}$  follow the marginal models derived from the population model (2.3), i.e.

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

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

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

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

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

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

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

The conditional covariance matrix is

$$\begin{aligned} V_{dr|s} &= V_{dr} - V_{drs}V_{ds}^{-1}V_{drs} = Z_{dr}V_{ud}Z'_{dr} + V_{edr} - Z_{dr}V_{ud}Z'_{ds}V_{ds}^{-1}Z_{ds}V_{ud}Z'_{dr} \\ &= Z_{dr}V_{ud}Z'_{dr} + V_{edr} - Z_{dr}V_{ud}Z'_{ds} \left\{ V_{eds}^{-1} - V_{eds}^{-1}Z_{ds}(V_{ud}^{-1} + n_d V_{edj}^{-1})^{-1}Z'_{ds}V_{eds}^{-1} \right\} Z_{ds}V_{ud}Z'_{dr} \\ &= Z_{dr}V_{ud}Z'_{dr} + V_{edr} - n_d Z_{dr}V_{ud}V_{edj}^{-1}V_{ud}Z'_{dr} + n_d^2 Z_{dr}V_{ud}V_{edj}^{-1}(V_{ud}^{-1} + n_d V_{edj}^{-1})^{-1}V_{edj}^{-1}V_{ud}Z'_{dr}. \end{aligned}$$

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  $n_d \neq 0$  and  $j \in r_d$ ,  $j > n_d$ , the conditional  $m \times 1$  mean vector is

$$\mu_{dj|s} = X_{dj}\beta + V_{ud}Z'_{ds} \left\{ V_{eds}^{-1} - V_{eds}^{-1}Z_{ds}(V_{ud}^{-1} + n_d V_{edj}^{-1})^{-1}Z'_{ds}V_{eds}^{-1} \right\} (y_{ds} - X_{ds}\beta)$$

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

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

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

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

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

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

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

### 3 Predictors of average compositions

This section deals with the problem of predicting the domain average compositions  $A_{dk}, d = 1, \dots, D, k = 1, \dots, m + 1$ , defined in (2.1). As explained in Sect. 2 and Appendix A, we first transform the  $m$ -part compositions  $a_{dj} = (a_{dj1}, \dots, a_{djm})'$  into vectors of  $R^m$ . This is done by applying a one-to-one function  $h = (h_1, \dots, h_m)' : \mathcal{S}^m \mapsto R^m$ . The transformed vectors  $y_{dj} = h(a_{dj})$  have components  $y_{dj1} = h_1(a_{dj}), \dots, y_{djm} = h_m(a_{dj})$ . Let  $h^{-1} = (h_1^{-1}, \dots, h_m^{-1})' : R^m \mapsto \mathcal{S}^m$  be the inverse function of  $h$ , so that  $a_{dj1} = h_1^{-1}(y_{dj}), \dots, a_{djm} = h_m^{-1}(y_{dj})$ .

For estimating  $A_{dk}, k = 1, \dots, m + 1$ , we assume that  $y_{dj} = (y_{dj1}, \dots, y_{djm})'$  follows a multivariate nested error regression (MNER) model. For  $d = 1, \dots, D$ , the target parameters are additive, i.e

$$A_{dk} = \frac{1}{N_d} \sum_{j=1}^{N_d} h_k^{-1}(y_{dj}), \quad k = 1, \dots, m; \quad A_{dm+1} = 1 - A_{d1} - \dots - A_{dm}.$$

The EBP of  $A_{dk}$  is

$$\begin{aligned} \hat{A}_{dk}^{eb} &= \frac{1}{N_d} \left\{ \sum_{j \in s_d} h_k^{-1}(y_{dj}) + \sum_{j \in r_d} E_{y_r} [h_k^{-1}(y_{dj}) | y_s; \psi] \right\}, \quad k = 1, \dots, m; \quad \hat{A}_{dm+1}^{eb} \\ &= 1 - \hat{A}_{d1}^{eb} - \dots - \hat{A}_{dm}^{eb}. \end{aligned}$$

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

- (a) Estimate the unknown parameter  $\psi = (\beta', \theta')'$  using sample data  $(y_s, X_s)$ .



(b) Replacing  $\psi = (\beta', \theta')'$  by the estimate  $\hat{\psi} = (\hat{\beta}', \hat{\theta}')'$  obtained in (a), draw  $L$  copies of each out-of-sample variable  $y_{dj}$  as

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

where

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

and

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

(c) The Monte Carlo approximation of the expected value is

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

The Monte Carlo approximation of the EBP of  $A_{dk}$  is

$$\hat{A}_{dk}^{eb} \approx \frac{1}{L} \sum_{\ell=1}^L A_{dk}^{(\ell)}, \quad A_{dk}^{(\ell)} = \frac{1}{N_d} \left( \sum_{j \in s_d} h_k^{-1}(y_{dj}) + \sum_{j \in r_d} h_k^{-1}(y_{dj}^{(\ell)}) \right), \quad k = 1, \dots, m.$$

The plug-in estimator of  $A_{dk}$  is

$$\hat{A}_{dk}^{in} = \frac{1}{N_d} \sum_{j \in U_d} h_k^{-1}(\hat{y}_{dj}^{eb}) = \frac{1}{N_d} \left\{ \sum_{j \in s_d} h_k^{-1}(y_{dj}) + \sum_{j \in r_d} h_k^{-1}(\hat{\mu}_{dj|s}) \right\}, \quad k = 1, \dots, m,$$

and  $\hat{A}_{dm+1}^{in} = 1 - \hat{A}_{d1}^{in} - \dots - \hat{A}_{dm}^{in}$ .

**Remark 3.1** 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  $A_{dk}^{(\ell)}$  as

$$A_{dk}^{(\ell)} = \frac{1}{N_d} \left[ \sum_{j=1}^{n_d} h_k^{-1}(y_{dj}) + \sum_{t=1}^T \sum_{j=1}^{N_{dt}-n_{dt}} h_k^{-1}(y_{dj}^{(\ell)}) \right],$$

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

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

and  $\hat{V}_{d|s}$  was defined in Step (b) of the above Monte Carlo procedure.

**Remark 3.2** If some auxiliary variables are continuous, we can use the Hájek-type approximation to  $A_{dk}^{(\ell)}$ , i.e.

$$A_{dk}^{(\ell)} \approx \frac{1}{N_d} \sum_{j \in s_d} w_{dj} h_k^{-1}(y_{dj}^{(\ell)}).$$

where  $w_{dj}$  is the sample weight of unit  $j$  of domain  $d$ . A GREG-type approximation to  $A_{dk}^{(\ell)}$  is

$$A_{dk}^{(\ell)} \approx \frac{1}{N_d} \left( \sum_{j \in s_d} \{h_k^{-1}(y_{dj}) - h_k^{-1}(y_{dj}^{(\ell)})\} + \sum_{j \in s_d} \tilde{w}_{dj} h_k^{-1}(y_{dj}^{(\ell)}) \right),$$

where  $\tilde{w}_{dj} = w_{dj} N_d / \hat{N}_d$ ,  $\hat{N}_d = \sum_{j \in s_d} w_{dj}$ .

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. (2008b). The steps for implementing this method are

1. Fit the model (2.5) to sample data  $(y_s, X_s)$  and calculate an estimator  $\hat{\psi} = (\hat{\beta}', \hat{\theta}')'$  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  $e_{dj}^* \sim N(0, \hat{V}_{edj})$ , where  $\hat{V}_{ud} = V_{ud}(\hat{\theta})$  and  $\hat{V}_{edj} = V_{edj}(\hat{\theta})$ .
3. Construct the bootstrap superpopulation model  $\xi^*$  using  $\{u_d^*\}$ ,  $\{e_{dj}^*\}$ ,  $\{X_{dj}\}$  and  $\hat{\beta}$ , i.e

$$\xi^* : y_{dj}^* = X_{dj}\hat{\beta} + u_d^* + e_{dj}^*, \quad d = 1, \dots, D, \quad j = 1, \dots, N_d. \tag{3.1}$$

4. Under the bootstrap superpopulation model (3.1), 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

$$A_{dk}^{*(b)} = \frac{1}{N_d} \sum_{j=1}^{N_d} h_k(y_{dj}^{*(b)}), \quad k = 1, \dots, m, \quad b = 1, \dots, B.$$

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,  $\hat{A}_{dk}^{eb*(b)}$ ,  $k = 1, \dots, m$ , as described in Sect. 3, using the bootstrap sample vector  $y_s^*$  and the known values  $X_{dj}$ .
6. A Monte Carlo approximation to the theoretical bootstrap estimator

$$MSE_*(\hat{A}_{dk}^{eb*}) = E_{\xi^*}[(\hat{A}_{dk}^{eb*} - A_{dk}^*)(\hat{A}_{dk}^{eb*} - A_{dk}^*)'], \quad k = 1, \dots, m,$$

is

$$mse_*(\hat{A}_{dk}^{eb*}) = \frac{1}{B} \sum_{b=1}^B (\hat{A}_{dk}^{eb*(b)} - A_{dk}^{*(b)})(\hat{A}_{dk}^{eb*(b)} - A_{dk}^{*(b)})', \quad k = 1, \dots, m. \tag{3.2}$$

The estimator (3.2) is used to estimate  $MSE(\hat{A}_{dk}^{eb})$ ,  $k = 1, \dots, m$ .

### 4 Simulations

The simulation experiments empirically investigate the asymptotic behavior of: (1) the REML estimators of model parameters in Sect. 4.1 and Appendix B.1, (2) the EBP and plug-in predictors of domain average compositions in Sect. 4.2 and Appendix B.2, and (3) the parametric bootstrap MSE estimators in Sect. 4.3 and Appendix B.3.

To meet these three objectives, we consider a basic scenario in which we run simulations for different sample sizes. Take  $m = 2$ ,  $p_1 = p_2 = 2$ ,  $p = 4$ ,  $\beta_1 = (\beta_{11}, \beta_{12})' = (10, 10)'$ ,  $\beta_2 = (\beta_{21}, \beta_{22})' = (10, 10)'$ . 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  $x_{dj1} = (x_{dj11}, x_{dj12})$ ,  $x_{dj2} = (x_{dj21}, x_{dj22})$  and

$$x_{dj11} = x_{dj21} = 1, \quad x_{dj12} \sim \text{Bin}(1, 1/2), \quad x_{dj22} \sim \text{Bin}(1, 1/2),$$

For  $d = 1, \dots, D$ , simulate  $u_d \sim N_2(0, V_{ud})$  and  $e_{dj} \sim N_2(0, V_{edj})$ , where

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

where  $\theta_1 = 0.75$ ,  $\theta_2 = 0.75$ ,  $\theta_4 = 0.5$ ,  $\theta_5 = 0.5$  and  $\theta_3 = -0.4$ ,  $\theta_6 = 0.4$ . Simulation 1 generates only 4 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

$$X_{01} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad X_{02} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad X_{03} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad X_{04} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

**Table 1**  $RB(\hat{\eta})$  (left) and  $RRE(\hat{\eta})$  (right) with  $n_d = 10$ 

$\eta$		$D = 25$	$D = 50$	$D = 100$	$D = 25$	$D = 50$	$D = 100$
$\beta_{11}$	-1	-2.02	1.47	0.82	17.44	12.38	8.87
$\beta_{12}$	1	0.18	0.32	0.26	8.23	5.82	4.20
$\beta_{21}$	-1	2.12	0.27	-0.24	17.00	13.62	9.18
$\beta_{22}$	1	0.26	-0.61	0.15	8.42	6.71	4.32
$\theta_1$	0.75	0.89	1.62	-1.24	20.66	17.20	10.62
$\theta_2$	0.75	-1.39	0.78	-1.31	22.72	15.83	10.45
$\theta_4$	0.5	-0.12	-0.19	-0.09	4.26	3.39	2.33
$\theta_5$	0.5	0.36	0.21	-0.03	4.59	3.31	2.17
$\theta_3$	-0.4	1.80	1.09	0.34	19.56	13.75	8.76
$\theta_6$	0.4	-0.11	0.17	0.07	5.68	4.21	2.81

**Table 2**  $RB(\hat{\eta})$  (left) and  $RRE(\hat{\eta})$  (right) with  $D = 50$ 

$\eta$		$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
$\beta_{11}$	-1	1.47	0.12	0.19	0.04	12.38	12.85	11.86	13.39
$\beta_{12}$	1	0.32	0.53	-0.12	-0.12	5.82	3.83	2.73	1.86
$\beta_{21}$	-1	0.27	-0.43	1.84	0.80	13.62	12.27	13.02	13.05
$\beta_{22}$	1	-0.61	-0.11	0.08	0.01	6.71	3.71	2.52	1.81
$\theta_1$	0.75	1.62	0.66	-0.19	-0.07	17.20	15.14	14.88	14.95
$\theta_2$	0.75	0.78	0.55	-0.54	1.01	15.83	16.54	14.28	13.82
$\theta_4$	0.5	-0.19	-0.09	-0.08	-0.13	3.39	1.99	1.39	0.93
$\theta_5$	0.5	0.21	-0.05	0.02	-0.01	4.21	2.12	1.41	0.99
$\theta_3$	-0.4	1.09	0.13	0.62	1.44	13.75	11.65	12.39	11.61
$\theta_6$	0.4	0.17	0.03	0.12	0.06	4.21	2.37	1.59	1.05

#### 4.1 Simulation 1 for REML estimators

The target of Simulation 1 is to check the behavior of the REML algorithm for fitting the MNER model (2.5). This simulation runs  $I = 200$  iterations. Appendix B.1 gives the steps of Simulation 1 and the definitions of the absolute and relative performance measures. For every REML estimator  $\hat{\eta} \in \{\hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{21}, \hat{\beta}_{22}, \hat{\theta}_1, \dots, \hat{\theta}_6\}$ , Tables 1 and 2 present the relative bias  $RB(\hat{\eta})$  and the relative root-mean-squared error  $RRE(\hat{\eta})$  in %. Appendix B.1 gives the corresponding absolute performance measures. Simulation 1 shows that the REML Fisher-scoring algorithm works properly because  $RB(\hat{\eta})$  and  $RRE(\hat{\eta})$  decrease as  $n_d$  or  $D$  increase.

#### 4.2 Simulation 2 for EBPs

Simulation 2 investigates the EBP and plug-in predictors,  $\hat{A}_{dk}^{eb}$  and  $\hat{A}_{dk}^{in}$ , respectively,  $k = 1, 2, 3$ . It takes  $I = 200$  iterations and generates  $L = 200$  random vectors for

**Table 3**  $RAB_k$  (left) and  $RRE_k$  (right) for clr with  $D = 50$

$\hat{A}_{dk}$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
$\hat{A}_{d1}^{eb}$	1.1680	0.6938	0.5031	0.3548	19.4109	12.1619	8.2437	4.9351
$\hat{A}_{d1}^{in}$	5.0306	4.6955	4.0220	2.4395	23.6714	16.6757	13.0039	8.2716
$\hat{A}_{d2}^{eb}$	1.1390	0.7501	0.4865	0.2508	19.2729	12.3950	8.1939	4.8637
$\hat{A}_{d2}^{in}$	5.5594	4.3595	4.0676	2.4828	23.3680	17.0513	13.0109	8.1814
$\hat{A}_{d3}^{eb}$	0.6798	0.4789	0.3080	0.1774	11.8003	7.5792	5.0246	2.9736
$\hat{A}_{d3}^{in}$	4.0433	3.4835	3.0912	1.8774	14.8722	10.9202	8.4756	5.3553

**Table 4**  $RAB_k$  (left) and  $RRE_k$  (right) for alr with  $D = 50$

$\hat{A}_{dk}$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
$\hat{A}_{d1}^{eb}$	0.6672	0.3614	0.2709	0.1694	10.9320	6.7357	4.4970	2.6389
$\hat{A}_{d1}^{in}$	0.6226	0.6225	0.4622	0.2987	11.4034	7.2562	5.0881	3.0810
$\hat{A}_{d2}^{eb}$	0.6735	0.3903	0.2963	0.1369	11.0080	6.8462	4.4879	2.6031
$\hat{A}_{d2}^{in}$	0.9609	0.5758	0.5536	0.3087	11.4535	7.4184	5.1074	3.0540
$\hat{A}_{d3}^{eb}$	0.5419	0.3860	0.2492	0.1373	9.1757	5.8458	3.8838	2.2651
$\hat{A}_{d3}^{in}$	0.9445	0.7637	0.7149	0.4164	9.5444	6.2760	4.3598	2.6019

**Table 5**  $RAB_k$  (left) and  $RRE_k$  (right) for ilr with  $D = 50$

$\hat{A}_{dk}$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$	$n_d = 10$	$n_d = 25$	$n_d = 50$	$n_d = 100$
$\hat{A}_{d1}^{eb}$	1.0411	0.7943	0.5048	0.2696	18.5707	11.9505	7.8420	4.6657
$\hat{A}_{d1}^{in}$	10.6431	9.2562	7.9828	5.1322	21.9423	15.6805	11.8124	7.3526
$\hat{A}_{d2}^{eb}$	0.5424	0.3580	0.2541	0.0992	9.9159	6.1886	4.0830	2.3988
$\hat{A}_{d2}^{in}$	3.4303	2.8141	2.4114	1.5490	11.4624	7.9513	5.8924	3.6645
$\hat{A}_{d3}^{eb}$	0.6043	0.3826	0.2850	0.1311	10.1898	6.3440	4.1874	2.4775
$\hat{A}_{d3}^{in}$	2.2682	2.1507	1.8451	1.1882	11.4280	7.7876	5.7255	3.5466

the Monte Carlo approximations of integrals. The population sizes are  $N_d = 200$  and  $D = 50$ . Let  $h$  be the clr, alr or ilr transformation. Appendix B.2 gives the steps of Simulation 2 and the definitions of the absolute and relative performance measures. Tables 3, 4 and 5 present the relative absolute bias  $RAB_k$  and the relative root-mean-squared error  $RRE_k$  in %,  $k = 1, 2, 3$ , for the clr, alr and ilr transformations, respectively. Appendix B.2 gives the corresponding absolute performance measures.

The performances measures decrease as the sample sizes,  $n_d$ 's, increase and the EBP gets better results (RAB and RRE) than the plug-in predictor. Note that for each transformation, the data generation, and therefore the true underlying model, is different. For this reason, the results in Tables 3, 4 and 5 are not comparable. It is curious to observe that if the data are generated by the MNER model derived from

**Table 6**  $RAB_k$  (left) and  $RRE_k$  (right) for clr with  $D = 50$  and  $n_d = 10$ 

$B$	50	100	200	300	400	50	100	200	300	400
$\hat{A}_{d1}^{eb}$	12.09	12.26	12.09	12.13	12.38	32.42	25.28	22.51	19.94	20.40
$\hat{A}_{d1}^{in}$	32.15	32.15	32.12	31.82	29.64	38.68	35.77	34.68	33.61	31.63
$\hat{A}_{d2}^{eb}$	13.76	14.82	13.34	13.65	13.30	31.57	27.90	23.25	21.00	19.60
$\hat{A}_{d2}^{in}$	31.24	27.15	30.45	27.89	33.45	37.13	31.86	33.16	29.99	34.84
$\hat{A}_{d3}^{eb}$	9.22	9.81	8.80	9.31	8.53	23.13	18.66	15.46	14.02	12.89
$\hat{A}_{d3}^{in}$	35.98	34.99	36.36	34.61	35.81	38.42	36.37	37.14	35.14	36.23

**Table 7**  $RAB_k$  (left) and  $RRE_k$  (right) for alr with  $D = 50$  and  $n_d = 10$ 

$B$	50	100	200	300	400	50	100	200	300	400
$\hat{A}_{d1}^{eb}$	6.73	6.55	5.43	5.69	5.12	26.64	20.09	15.58	14.25	13.33
$\hat{A}_{d1}^{in}$	9.91	10.51	7.44	10.75	7.98	26.03	20.82	15.65	16.44	13.98
$\hat{A}_{d2}^{eb}$	7.53	6.15	5.98	5.55	6.25	24.88	19.53	15.70	12.76	13.52
$\hat{A}_{d2}^{in}$	11.19	9.75	10.00	9.19	11.97	25.13	20.02	17.07	14.27	16.49
$\hat{A}_{d3}^{eb}$	5.69	4.97	4.34	4.05	4.41	21.67	16.55	13.13	11.00	9.74
$\hat{A}_{d3}^{in}$	10.16	10.11	9.51	8.95	9.94	22.22	18.06	15.20	13.32	13.02

the alr transformation and its corresponding EBP is used, the results are slightly better than in the clr and ilr cases.

### 4.3 Simulation 3 for MSEs

Simulation 3 investigates the MSE estimators of predictors  $\hat{A}_{dk}^{eb}$  and  $\hat{A}_{dk}^{in}$ ,  $k = 1, 2, 3$ . One of the goals is to give a recommendation on the number of bootstrap replicates  $B$  to implement. The simulation takes  $I = 200$  iterations and generates  $L = 200$  random vectors for the Monte Carlo approximations of integrals. The population sizes are  $N_d = 200$  and  $D = 50$ . Let  $h$  be the clr, alr or ilr transformation. Appendix B.3 gives the steps of Simulation 3 and the definitions of the absolute and relative performance measures.

Tables 6, 7 and 8 present the relative absolute bias  $RAB_k$  and the relative root-mean-squared error  $RRE_k$  in %,  $k = 1, 2, 3$ , for the clr, alr and ilr transformations, respectively. The number of bootstrap replicates is  $B = 50, 100, 200, 300, 400$ . Appendix B.3 gives the corresponding absolute performance measures. As in Simulation 2, we remark that the results in Tables 6, 7 and 8 are not comparable because the data generation is different. Nevertheless, we observe that if the data are generated by the MNER model derived from the alr transformation and its corresponding EBP is used, Simulation 3 gives slightly better results than in the clr or ilr cases. That is, the functional form of the transformation plays a non-negligible role. In any case, the

**Table 8**  $RAB_k$  (left) and  $RRE_k$  (right) for ilr with  $D = 50$  and  $n_d = 10$

$B$	50	100	200	300	400	50	100	200	300	400
$\hat{A}_{d1}^{eb}$	10.91	11.61	11.52	10.62	10.18	28.34	24.10	21.27	17.66	16.34
$\hat{A}_{d1}^{in}$	27.67	23.95	26.22	25.36	28.87	33.62	28.68	29.28	27.27	30.23
$\hat{A}_{d2}^{eb}$	9.13	9.07	8.82	8.77	8.85	26.55	20.47	16.61	15.63	14.60
$\hat{A}_{d2}^{in}$	24.63	27.58	24.17	26.05	23.23	31.44	30.98	26.57	27.88	24.95
$\hat{A}_{d3}^{eb}$	11.18	10.18	10.47	9.57	9.34	25.69	20.16	17.91	14.55	14.10
$\hat{A}_{d3}^{in}$	19.77	18.85	16.86	20.88	21.03	27.76	24.08	21.01	22.85	22.84

selection of the transformation in an application to real data must be made based on the diagnosis of the corresponding MNER model that we select.

Figures 1 and 2 show the boxplots of  $RRE_{dk}$  and  $RAB_{dk}$  for the predictors  $\hat{A}_{dk}^{eb}$ ,  $k = 1, 2, 3$ , with the clr transformation. From the obtained performance measures, we recommend to implement the bootstrap algorithm with at least  $B = 300$  iterations. Appendix B.3 give the same recommendation for the alr and ilr transformations.

### 5 The Spanish Household Budget Survey (SHBS)

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 as on various characteristics related to the conditions of household life. In the Spanish economy, it is important to have good estimates of consume 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 consume indicators in order to implement their strategies. Among the main consume indicators are the proportions of food and housing annual expenses of households. This section presents an application of the new statistical methodology to the estimation of domain parameters defined as average of proportions of annual household expenditures. We take 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  $a_{dj1}$ ,  $a_{dj2}$  and  $a_{dj3}$  be the proportions of annual expenditures on food, housing and other for household  $j$  of domain  $d$ . Housing includes expenditure on current housing costs, water, electricity, gas and other fuels. Food includes both food and nonalcoholic beverages and other represent the remaining expenditures. The vectors  $a_{dj} = (a_{dj1}, a_{dj2})' \in R^2$  are 2-part compositions that can be transformed into vectors  $y_{dj} = h(a_{dj})$  of  $R^2$  by one of the transformations  $h$  described in Appendix A. Let  $x_{djk}$ ,  $d = 1, \dots, D$ ,  $j = 1, \dots, n_d$ ,  $k = 1, 2$ , be the  $4 \times 1$  vector whose components are the binary auxiliary variables that indicate the composition of the household to which household  $j$  belongs in domain  $d$ . As auxiliary variables, we thus consider the household composition HC with categories

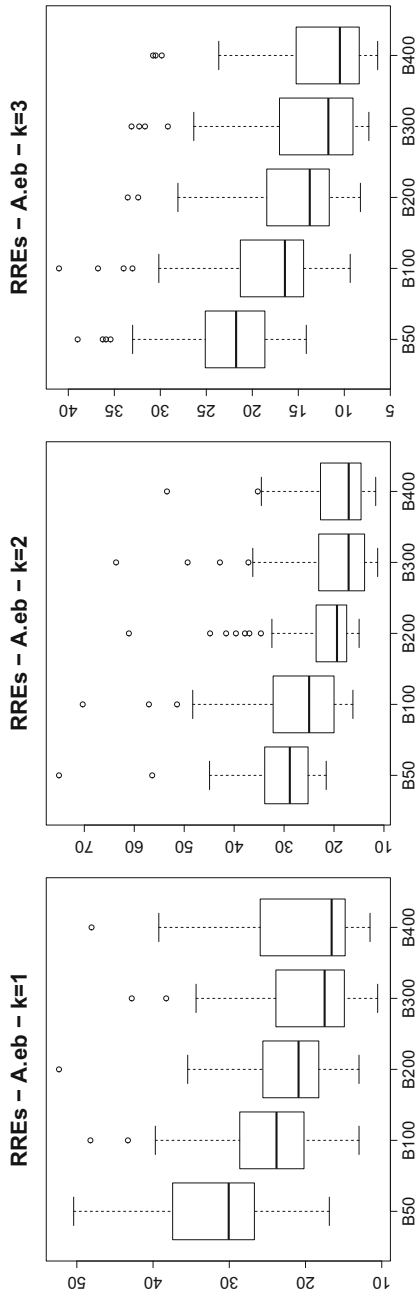


Fig. 1  $RRE_{dik}$  (in %) of MSE estimators for  $\hat{A}_{dik}^{eb}$ ,  $k = 1, 2, 3$ , for  $clr$



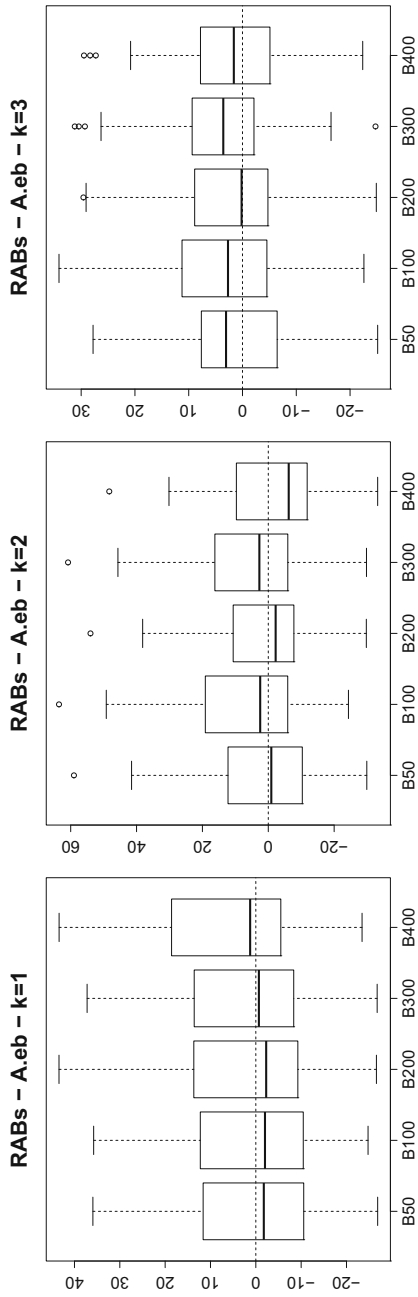


Fig. 2  $RAB_{dk}$  (in %) of MSE estimators for  $\hat{A}_{dk}^{eb}$ ,  $k = 1, 2, 3$ , for  $clr$

**Table 9** Regression parameters

$y$	$x$ -variable	Estimate	$z$ -value	St.Error	$p$ -value
$y_1$	Intercept	-0.66	55.78	0.012	0.00
	HC1	0.46	54.78	0.001	0.00
	HC2	0.01	0.70	0.008	0.48
	HC3	-0.16	18.82	0.008	0.00
$y_2$	Intercept	-0.41	37.85	0.011	0.00
	HC1	-0.24	31.57	0.008	0.00
	HC2	-0.30	38.88	0.008	0.00
	HC3	-0.03	4.45	0.008	0.00

**Table 10** Variance and correlation parameters

Parameter	Estimate	L.CI	U.CI
$\sigma_{u1}^2$	0.006	0.003	0.008
$\sigma_{u2}^2$	0.005	0.003	0.007
$\rho_u$	-0.695	-0.865	-0.524
$\sigma_{e1}^2$	0.210	0.206	0.214
$\sigma_{e2}^2$	0.174	0.171	0.177
$\rho_e$	-0.347	-0.359	-0.336

**HC1:** Single person or adult couple with at least one members with age over 65,

**HC2:** Other compositions with a single person or a couple without children,

**HC3:** Couple with children under 16 years old or adult with children under 16 years old,

**HC4:** Other households.

The variable HC is treated as a factor with reference category HC4.

For calculating the EBPs of the domain parameters of interest, we need the *true* population sizes,  $N_{dt}$ , of the crossings of provinces with the categories of the variable HC. 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 they are considered the most accurate demographic figures in Spain. On the other hand, the SHBS sampling weights are calibrated to the population sizes of the autonomous community (NUTS 2) crossed with sex and age groups, which are not the domains of interest.

This section presents an statistical analysis by applying the centered logratio transformation. This choice is due to the good fit of the MNER model to the transformed data. For the sake of completeness, Appendix C presents the corresponding data analysis for the alr and ilr transformations. Table 9 presents the estimates of the regression parameters, the  $z$ -values, the standard errors and the asymptotic  $p$ -values. The factor HC is significative for  $y_1$  and  $y_2$ . Table 10 presents the asymptotic 95% confidence intervals (L.CI, U.CI) for the variance component parameters. None of them contains the zero.

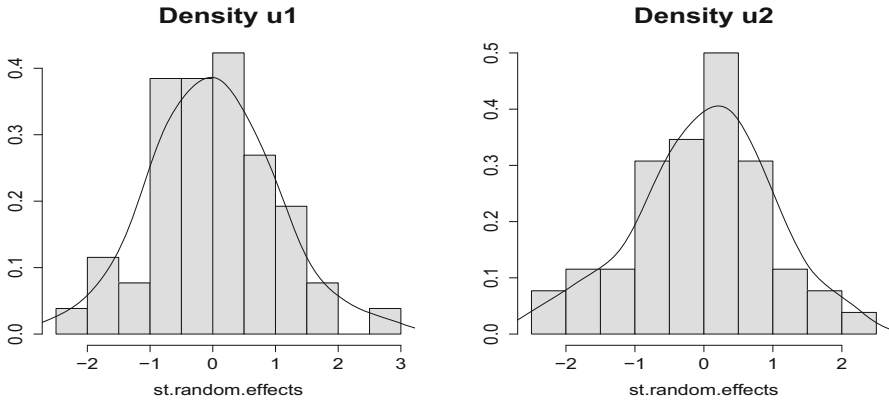


Fig. 3 Histograms of standardized random effects

For calculating the asymptotic  $p$ -values and confidence intervals of Tables 9 and 10, we take the asymptotic distributions of the REML estimators  $\hat{\theta}$  and  $\hat{\beta}$ , i.e.

$$\hat{\theta} \sim N_6(\theta, F_s^{-1}(\theta)), \quad \hat{\beta} \sim N_p(\beta, (X'_s V_s^{-1} X_s)^{-1}),$$

where  $F_s$  is the REML Fisher information matrix. For  $\hat{\beta}_i = \beta_0$ , the asymptotic  $p$ -value for testing the hypothesis  $H_0 : \beta_i = 0$  is

$$p\text{-value} = 2P_{H_0}(\hat{\beta}_i > |\beta_0|) = 2P(N(0, 1) > |\beta_0|/\sqrt{q_{ii}}).$$

where  $(X'V^{-1}(\hat{\theta})X)^{-1} = (q_{ij})_{i,j=1,\dots,p}$  and  $\beta_i$  denotes the  $i$ -th component of the vector  $\beta$ . The asymptotic  $(1 - \alpha)$ -level confidence intervals for the components  $\theta_\ell$  of  $\theta$  are

$$\hat{\theta}_\ell \pm z_{\alpha/2} v_{\ell\ell}^{1/2}, \quad \ell = 1, \dots, 6,$$

where  $F^{-1}(\hat{\theta}) = (v_{ab})_{a,b=1,\dots,6}$  and  $z_\alpha$  is the  $\alpha$ -quantile of the  $N(0, 1)$  distribution.

Figure 3 plots the histograms of the  $D = 52$  standardized EBPs of the random effects of the fitted MNER model for food (left) and housing (right) expenditures. 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 nonparametric estimate of the density functions, the definitive conclusions can not be drawn.

Figure 4 gives the histograms of standardized residuals for components  $y_1$  and  $y_2$ . It also prints the corresponding probability density function estimates. We do not appreciate a large deviation from the normal distribution.

Figure 5 presents the dispersion plots of standardized residuals versus predicted values (in  $10^4$  euros). Most standardized residuals fall within the interval  $(-3, 3)$ , so we consider that outliers do not play a relevant role in the performance of the

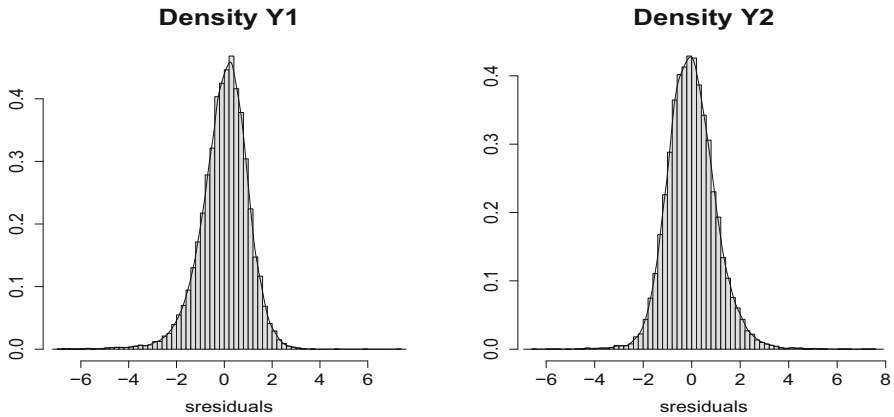


Fig. 4 Histograms of standardized residuals

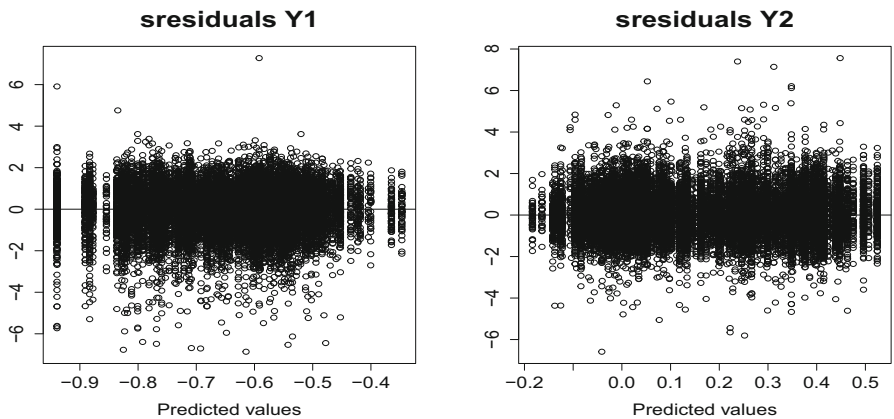


Fig. 5 Standardized residuals versus predicted values (in  $10^4$  euros)

EBPs. Appendix C of the supplementary material gives the corresponding plots for the additive and isometric logratio transformations. The corresponding plots are similar to the ones shown in Figs. 4 and 5 for the centered logratio transformation. However, Fig. 5 presents more uniform clouds of points in both components than the corresponding figures for the two other transformations. From this graphical diagnosis, we finally prefer doing the data analysis with the centered logratio transformation. However, since the choice of the clr transformation can be debatable, Appendix C presents the full analysis of the data under the two other transformations.

Figure 6 plots the plug-in and the EBP predictions of  $a_{d1}$  and  $a_{d2}$ . The domains are sorted by sample sizes and the sample size is printed in the axis OX. This figure shows that both estimators follow a similar pattern. This information is completed by Fig 7, which shows the relative root-MSEs (RRMSE).

Figure 8 (left) maps the proportions of the household annual expenditures in food by Spanish provinces. Figure 8 (right) maps the estimated RRMSE in %. These figures

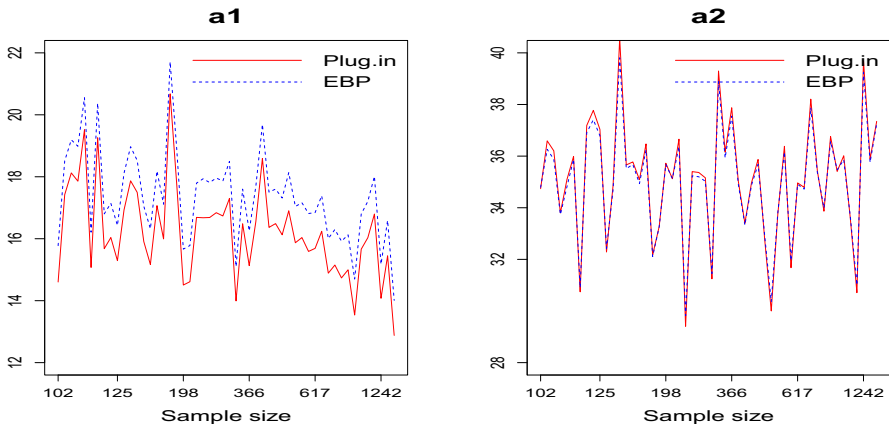


Fig. 6 Plug-in and EBP predictions of  $a_{d1}$  and  $a_{d2}$  in %

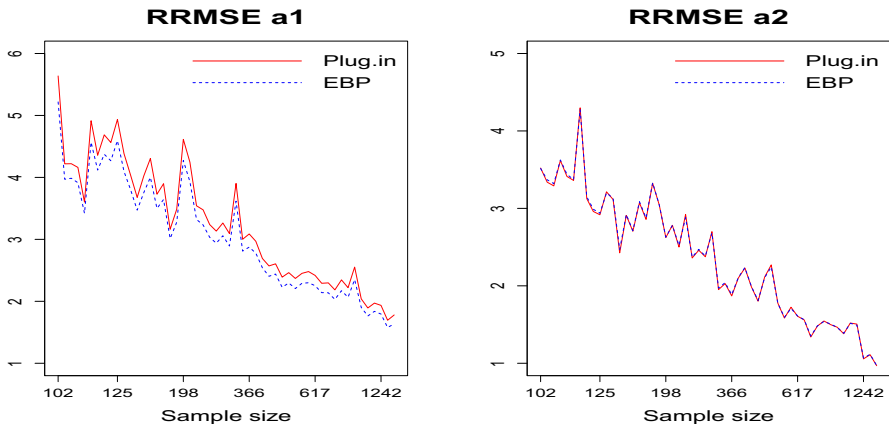
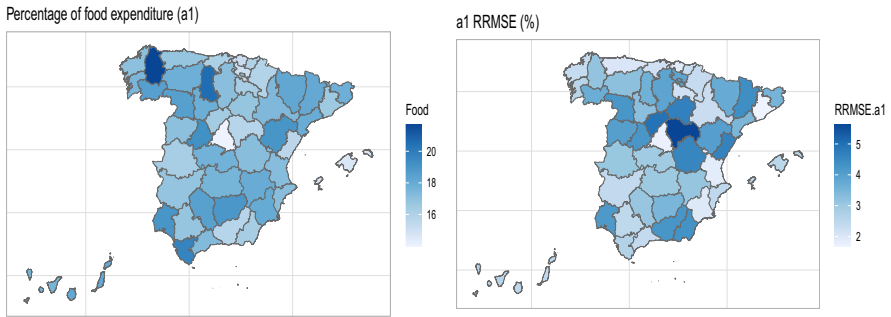


Fig. 7 RRMSE of plug-in and EBP predictions of  $a_{d1}$  and  $a_{d2}$  in %

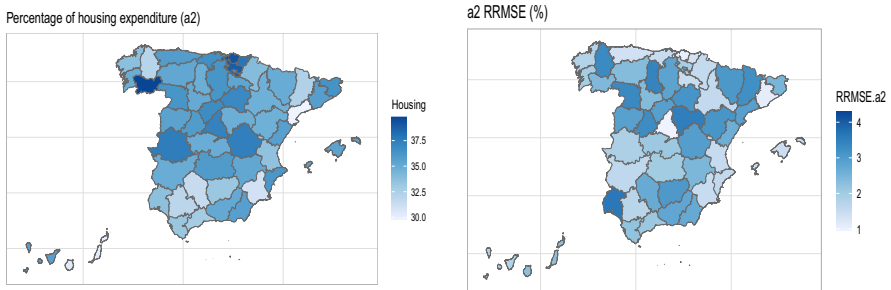
show that expenditures on food are rather variable between provinces. This happens mostly in the autonomous regions of Andalucía, Aragón or Castilla León, where there are many provinces and some of them are more deprived than others. In contrast, there are other regions, such as Basque Country where the variability of the estimated ratios is smaller. This information could be of great use to local governments in developing economic plans aimed at households and improving the quality of life.

Figure 9 (left) maps the proportions of the household annual expenditures on housing by Spanish provinces. Figure 9 (right) maps the estimated RRMSE in %. As is the case with food expenditure, these figures show that expenditures on housing is rather variable between provinces. This map shows clear differences between the north-central regions, where the proportion of spending is higher, and the southern regions, where household expenditures are lower.

Tables 11 and 12 present some condensed numerical results. The tables are constructed in two steps: First, the domains are sorted by sample size, starting by the



**Fig. 8** EBP predictions of  $a_{d1}$  by Spanish provinces in %



**Fig. 9** EBP predictions of  $a_{d2}$  by Spanish provinces in %

domain with the smallest sample size. Finally, a selection of 14 domains out of 52 is done from the positions 1, 5, 9,  $\dots$ , 52. The name and code of provinces are labeled by *province* and  $d$ , respectively, and the sample sizes by  $n_d$ . Table 11 presents the model-based predictions of food and housing expenditures by provinces and Table 12 displays the corresponding estimates of RRMSEs. The plug-in predictors are denoted by  $in1$  and  $in2$  and the EBPs by  $ebp1$  and  $ebp2$ .

## 6 Conclusions

Compositional data play an important role in public statistics. The proposed methodology is applied to estimate the proportions of annual household expenditures on food, housing and others from the 2016 SHBS at the province level. This paper introduces small area predictors of averages of unit-level vectors of compositions. For this purpose, the manuscript considers the centered logratio transformations of compositions into vectors of  $R^m$ . For the sake of completeness, Appendix C of the supplementary material presents the corresponding statistical analysis under the additive and isometric logratio transformations. A MNER model is proposed for analyzing the transformed compositional data, where the vectors of random effects and the vector of model errors have unstructured covariance matrices with unknown components. As usual in linear mixed models, the parameter estimates of the MNER model are obtained using the

**Table 11** Predictions of  $a_{d1}$  and  $a_{d2}$  in %

	$d$	$n_d$	in1	ebp1	in2	ebp2
Guadalajara	19	102	14.6	15.8	34.7	34.8
Palencia	34	118	19.5	20.6	35.1	34.8
Cuenca	16	123	16.0	17.1	37.8	37.4
Ourense	32	169	17.5	18.5	40.6	39.8
Burgos	9	187	16.0	17.1	36.5	36.2
Granada	18	198	14.6	15.8	35.1	35.1
Albacete	2	249	16.8	18.0	35.4	35.2
Ciudad Real	13	355	16.5	17.6	36.2	35.9
Pontevedra	36	463	16.4	17.5	35.0	34.9
A Coruña	15	536	15.9	17.0	33.8	33.8
Zaragoza	50	678	16.2	17.4	34.8	34.7
Cantabria	39	761	15.0	16.1	36.8	36.6
Murcia	30	913	16.8	18.0	30.7	31.0
Madrid	28	1653	12.9	14.0	37.3	37.3

**Table 12** RRMSE estimates for  $a_{d1}$  and  $a_{d2}$  in %

	$d$	$n_d$	in1	ebp1	in2	ebp2
Guadalajara	19	102	5.64	5.22	3.52	3.52
Palencia	34	118	3.60	3.42	3.41	3.44
Cuenca	16	123	4.56	4.27	2.96	2.99
Ourense	32	169	3.68	3.47	2.43	2.47
Burgos	9	187	3.90	3.64	2.86	2.88
Granada	18	198	4.24	3.93	2.78	2.78
Albacete	2	249	3.13	2.94	2.46	2.47
Ciudad Real	13	355	3.00	2.81	2.03	2.04
Pontevedra	36	463	2.57	2.40	1.98	1.99
A Coruña	15	536	2.37	2.21	1.78	1.77
Zaragoza	50	678	2.29	2.14	1.56	1.56
Cantabria	39	761	2.22	2.06	1.50	1.51
Murcia	30	913	1.97	1.84	1.51	1.49
Madrid	28	1653	1.78	1.64	0.97	0.97

REML method. The selection of the centered logratio transformation was motivated by the interpretability and diagnosis of the selected MNER model. In this sense, we followed the recommendations of Greenacre (2019). This is to say, we have tried to provide a simple solution to a practical problem of compositional data.

Of the two proposed predictors, EBP and plug-in, EBP presents a slightly better performance, as can be seen in the simulation study. For the calculation of the MSE, we recommend a parametric bootstrap, following the ideas of González-Manteiga et al. (2008a) and for a number of repetitions greater than  $B = 300$ .

As a result of the statistical analysis for Spanish provinces, we conclude that food expenditure in Spain accounts for 14.6% of total household expenditure and presents great variability within autonomous communities. This happens mostly in the Autonomous Regions of Andalucía, Aragón or Castilla León, where there are many provinces and some of them are more deprived than others. In contrast, there are other regions, such as Basque Country where the variability of the estimated proportions is smaller. On the other hand, spending on housing in Spain accounts for 31% of total household spending and there are important differences between the north-central provinces (with higher incomes) and those in the south.

In this case, we applied the introduced methodology to the SHBS, but it is useful in other topics of the official statistics, like the classification of the population by the educational level and according to economic activity. In both situations, it is necessary to take into account the simplex constraints.

We finally remind that there are other regression models for compositions, such as directional mixed effects models or Dirichlet regression mixed models. These models are likely to be adapted to the SAE context described in Sect. 2, including fitting algorithms, predictors of domain quantities, MSE estimators, and so on. They can be competitive options with respect to fitting a multivariate normal mixed model to logratio transformations of compositions. We believe that these tasks are interesting subjects for future research.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11749-023-00847-0>.

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, New York
- Arima S, Bell WR, Datta GS, Franco C, Liseo B (2017) Multivariate Fay-Herriot Bayesian estimation of small area means under functional measurement error. *J R Stat Soc Ser A* 180(4):1191–1209
- Benavent R, Morales D (2016) Multivariate Fay-Herriot models for small area estimation. *Comput Stat Data Anal* 94:372–390
- Benavent R, Morales D (2021) Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. *Stat Methods Appl* 30(1):195–222
- Berg EJ, Fuller WA (2014) Small area prediction of proportions with applications to the Canadian Labour Force Survey. *J Survey Stat Methodol* 2:227–256
- Boubeta M, Lombardía MJ, Morales D (2016) Empirical best prediction under area-level Poisson mixed models. *TEST* 25:548–569
- Boubeta M, Lombardía MJ, Morales D (2017) Poisson mixed models for studying the poverty in small areas. *Comput Stat Data Anal* 107:32–47



- Burgard JP, Krause J, Münnich R, Morales D (2021) l2-Penalized temporal logit-mixed models for the estimation of regional obesity prevalence over time. *Stat Methods Med Res* 30(7):1744–1768
- Camargo AP, Stern JM, Lauretto MS (2012) Estimation and model selection in Dirichlet regression. *AIP Conf Proc* 1443:206–213. <https://doi.org/10.1063/1.3703637>
- Chambers R, Dreassi E, Salvati N (2014) Disease mapping via negative binomial regression M-quantiles. *Stat Med* 33:4805–4824
- Chambers R, Salvati N, Tzavidis N (2016) Semiparametric small area estimation for binary outcomes with application to unemployment estimation for Local Authorities in the UK. *J R Stat Soc Ser A* 179:453–479
- Dreassi E, Ranalli MG, Salvati N (2014) Semiparametric M-quantile regression for count data. *Stat Methods Med Res* 23:591–610
- Egozcue JJ, Pawlowsky-Glahn V (2019) Compositional data: the sample space and its structure. *TEST* 28(3):599–638
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35(3):279–300
- Esteban MD, Morales D, Pérez A, Santamaría L (2012) Small area estimation of poverty proportions under area-level time models. *Comput Stat Data Anal* 56:2840–2855
- Esteban MD, Lombardía MJ, 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 MD, Lombardía MJ, López-Vizcaíno E, Morales D, Pérez A (2022a) Small area estimation of expenditure means and ratios under a unit-level bivariate linear mixed model. *J Appl Stat* 49(1):143–168
- Esteban MD, Lombardía MJ, López-Vizcaíno E, Morales D, Pérez A (2022b) Empirical best prediction of small area bivariate parameters. *Scand J Stat* 49:1699–1727
- Fabrizi E, Ferrante MR, Trivisano C (2016) Hierarchical Beta regression models for the estimation of poverty and inequality parameters in small areas. In: Pratesi Monica (ed) *Analysis of poverty data by small area methods*. Wiley, New York
- Ferrante MR, Trivisano C (2010) Small area estimation of the number of firms' recruits by using multivariate models for count data. *Surv Methodol* 36(2):171–180
- González-Manteiga W, Lombardía MJ, 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. *Comput Stat Data Anal* 51:2720–33
- González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2008a) Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Comput Stat Data Anal* 52:5242–5252
- González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2008b) Bootstrap mean squared error of small-area EBLUP. *J Stat Comput Simul* 78:443–462
- Greenacre M (2019) Comments on: Compositional data: the sample space and its structure. *TEST* 28:644–652
- Hobza T, Morales D (2016) Empirical best prediction under unit-level logit mixed models. *J Off Stat* 32(3):661–669
- Hobza T, Santamaría L, Morales D (2018) Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST* 27(2):270–294
- Hijazi RH, Jernigan RW (2009) Modeling compositional data using Dirichlet regression models. *J Appl Probab Stat* 4(1):77–91
- Krause J, Burgard JP, Morales D (2022) Robust prediction of domain compositions from uncertain data using isometric logratio transformations in a penalized multivariate Fay-Herriot model. *Stat Neerl* 76(1):65–96
- López-Vizcaíno E, Lombardía MJ, Morales D (2013) Multinomial-based small area estimation of labour force indicators. *Stat Model* 13(2):153–178
- López-Vizcaíno E, Lombardía MJ, Morales D (2015) Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *J R Stat Soc Ser A* 178(3):535–565
- Marhuenda Y, Molina I, Morales D (2013) Small area estimation with spatio-temporal Fay-Herriot models. *Comput Stat Data Anal* 58:308–325
- Marhuenda Y, Morales D, Pardo MC (2014) Information criteria for Fay-Herriot model selection. *Comput Stat Data Anal* 70:268–280

- Molina I, Saei A, Lombardía MJ (2007) Small area estimates of labour force participation under multinomial logit mixed model. *J R Stat Soc Ser A* 170:975–1000
- Morais J, Thomas-Agnan C, Simioni M (2018) Using compositional and Dirichlet models for market share regression. *J Appl Stat* 45(9):1670–1689
- Morales D, Pagliarella MC, Salvatore R (2015) Small area estimation of poverty indicators under partitioned area-level time models. *SORT Stat Oper Res Trans* 39(1):19–34
- Morales D, Esteban MD, Pérez A, Hobza T (2021) *A course on small area estimation and mixed models*. Springer, Berlin
- Pawlowsky-Glahn V, Bucciante A (eds) (2011) *Compositional data analysis*. Wiley, Chichester
- Rao JNK, Molina I (2015) *Small area estimation*, 2nd edn. Wiley, Hoboken
- Saei A, Chambers R (2003) Small area estimation under linear and generalized linear mixed models with time and area effects. S3RI Methodology Working Paper M03/15, Southampton Statistical Sciences Research Institute
- Scealy JL, Welsh AH (2017) A directional mixed effects model for compositional expenditure data. *J Am Stat Assoc* 112(517):24–36
- Souza DB, Moura FAS (2016) Multivariate Beta regression with applications in small area estimation. *J Off Stat* 32:747–768
- Tsagris M, Stewart C (2018) A Dirichlet regression model for compositional data with zeros. *Lobachevskii J Math* 39(3):398–412
- Tzavidis N, Ranalli MG, Salvati N, Dreassi E, Chambers R (2015) Robust small area prediction for counts. *Stat Methods Med Res* 24(3):373–395
- Zhang L, Chambers R (2004) Small area estimates for cross-classifications. *J Roy Stat Soc B* 66(2):479–496

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.