Journal of the Royal Statistical Society Series A: Statistics in Society, 2026, 00, 1–31
https://doi.org/10.1093/jrsssa/qnag065
Published: 12 May 2026
Original article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society
OXFORD

Small area estimation of proportions and rates under area-level Dirichlet mixed models

Esteban Cabello [1], María Dolores Esteban [1], Tomáš Hobza [2], Domingo Morales [1], Agustín Pérez [3]

[1]Center of Operations Research, Miguel Hernández University of Elche, Elche (Alicante), Spain
[2]Department of Mathematics, Czech Technical University in Prague, Prague, Czech Republic
[3]Department of Economic and Financial Studies, Miguel Hernández University of Elche, Elche (Alicante), Spain
Address for correspondence: Esteban Cabello, Center of Operations Research, Miguel Hernández University of Elche, Edificio Torretamarit—Avda. de la Universidad s/n, Elche (Alicante) 03202, Spain. Email: ecabello@umh.es

Abstract

This paper introduces an area-level Dirichlet mixed model for predicting compositional indicators of small areas. Direct estimators of the domain category proportions of a classification variable are the target variables of the new model. Once the model has been selected and fitted to the data, predictors of proportions, totals and rates of small areas are obtained and their mean square errors are estimated by parametric bootstrap. Several simulation experiments, designed to analyse the behaviour of the fitting algorithm, the small area predictors and the bootstrap procedure, are carried out. An application to real data from the Spanish Labour Force Survey, in the last quarter of 2022, is given. The target is the estimation of proportions of employed, unemployed and inactive people and unemployment rates by province, sex and age group.

Keywords Labour force survey, small area estimation, Dirichlet mixed models, area-level models, compositional data, bootstrap
JEL codes: 62E30, 62J12

1 Introduction

The National Statistical Institutes (NSI) estimate proportions and rates in subsets (domains) of a population using survey data. Sampling designs are designed to obtain accurate estimators in their planned domains. However, the demand for statistics is increasing, and NSIs must also provide estimates for unplanned domains (small areas), where sample sizes are too small to obtain accurate direct estimates. Small area estimation (SAE) provides solutions to these problems using statistical models to incorporate additional information into the inferential process. The monographs by Rao and Molina (2015), Pratesi (2016), and Morales et al. (2021) are introductory textbooks on SAE.

We are interested in area-level models for direct estimators of domain proportions, totals and rates. With this goal in mind, we have reviewed some statistical literature relevant to the estimation of proportions and totals of small areas. Count prediction in cross-classification contingency tables was studied by Zhang and Chambers (2004) and Berg and Fuller (2014). Area-level mixed models were

Received: February 21, 2025. **Revised:** February 9, 2026. **Accepted:** April 23, 2026

used by Esteban et al. (2012), Marhuenda et al. (2013, 2014), and Morales et al. (2015) to predict proportions. Robust and semiparametric estimation of counts were addressed by Tzavidis et al. (2015), Chambers et al. (2014), Dreassi et al. (2014), Chambers et al. (2016), and Chandra et al. (2017). Mixed models for binomial or Poissson variables were used by Erciulescu and Fuller (2013), Hobza and Morales (2016), Hobza et al. (2018), Boubeta et al. (2016, 2017), Ranalli et al. (2018), Marino et al. (2019), Berg (2023), and Diz-Rosales et al. (2024) to predict proportions counts and totals, including poverty indicators. Generalized linear mixed models (GLMM) were applied by Burgard et al. (2021), Krause et al. (2022a, 2022b), and Morales et al. (2022) to estimate regional prevalences. Based on nested error regression models, Molina and Rao (2010), Marhuenda et al. (2017), and Guadarrama et al. (2021) introduced empirical best predictors (EBPs) of head count ratios. Concerning the Bayesian methodology for small area estimation of proportions, we quote the contributions of Farrell (2000), Larsen (2003), S. Chen and Lahiri (2012), Liu and Lahiri (2017), X. Chen and Nandram (2022), and De Nicolò et al. (2024). These approaches apply univariate models that do not consider the possibility of jointly estimating the counts or proportions of two or more categories.

In Labour Force statistics, some indicators of interest are functions of the totals, the proportions or the percentages of the categories of a classification variable. That is, it is common to find indicators based on compositions. Compositional data are proportions or percentages of disjoint categories with a constant sum, usually 1 or 100. This simple constant sum constraint conditions the type of multivariate regression model that can be used for these data or for functions of them. The statistical literature presents solutions to these problems based on multinomial or Dirichlet regression models, or through multivariate normal models applied to log-ratio transformations of the compositions. Next, we cite some contributions on the prediction of labour indicators based on multinomial, multivariate normal or multivariate beta models. Molina et al. (2007), Saei and Taylor (2012), López-Vizcaíno et al. (2013, 2015), and Bugallo et al. (2024) derived predictors of labour indicators under multinomial mixed models. Souza and Moura (2016) and Fabrizi et al. (2016) applied multivariate beta regression models to SAE. Ferrante and Trivisano (2010) used multivariate models on count data to estimate totals of firms' recruits. Sun et al. (2024) proposed a multivariate mixed-effects model for mixed-type response variables subject to item nonresponse. Wang et al. (2018) assumes that proper transformations of the direct estimators of small area proportions follows a beta mixed model. Esteban et al. (2020), Krause et al. (2022a, 2022b), Esteban et al. (2023), and Cabello et al. (2024, 2025) proposed to fit multivariate linear mixed models to log-ratio transformations of compositions.

Dirichlet and Dirichlet-multinomial regression models are good alternatives for the analysis of compositional data. Without wishing to be exhaustive, we cite the following contributions. Peña and Yohai (2006) introduces a random coefficient Dirichlet regression model for quality indicators. Hijazi and Jernigan (2009) models the composition of 39 sediments taken from an Arctic lake by using a fixed-effect Dirichlet regression model. Tsagris and Stewart (2018) and Tang and Chen (2019) deal with Dirichlet regression model for compositional data with zeros. Under Dirichlet-multinomial models, Maples (2019) and Slud et al. (2024) gives small area estimates of totals and proportions. Martínez-Minaya et al. (2023) introduces a Laplace approximation to Bayesian inference in Dirichlet regression models. However, none of the aforementioned papers addresses the problem of obtaining predictors of small area compositional indicators based on Dirichlet mixed models.

This paper introduces an area-level Dirichlet mixed model (ADMM) to obtain predictors of proportions of employed, unemployed and inactive people, and of unemployment rates, by province, sex and age group in Spain. Given that the likelihood of the model is a multiple integral, a Laplace approximation-maximization algorithm is proposed to calculate the maximum likelihood estimators of the model parameters. Based on the fitted model, empirical best predictors and plug-in predictors of category proportions and rates are proposed. The mean squared errors (MSE) are estimated by parametric bootstrap, following the methodology of González-Manteiga et al. (2008, 2010). In addition an empirical research is carried out by means of simulation experiments and an application to Spanish Labour Force Survey (SLFS) data is given. This paper presents a statistical methodology that is new in three main aspects: (1) The introduction and implementation of a new model for small area estimation of labour market indicators, (2) The use of the introduced ADMM to derive predictors of domain proportions and counts, (3) The parametric bootstrap estimation of MSEs.

The remainder of the paper is organized as follows. Section 2 gives an introduction to the Labour Force data and to the SAE problem of interest. Section 3 describes the new ADMM and the ML-Laplace fitting algorithm. Section 4 develops the proposed predictors and the corresponding MSE estimation procedures. Section 5 presents three simulation experiments. The target of Simulation 1 is to check the behaviour of the ML-Laplace fitting algorithm. Simulation 2 investigates the performance of the predictors of the category proportions and rates. Simulation 3 analyses the parametric bootstrap estimator of the MSEs. Section 6 applies the proposed methodology to data from the SLFS of the last quarter of 2022 in Spanish provinces. Section 7 gives some conclusions. The paper contains three online supplementary material, appendices. Online supplementary material, Appendix A presents the derivatives needed to implement the ML-Laplace algorithm and provides an alternative fitting algorithm to the new model. Online supplementary material, Appendix B enumerates the steps of the three simulation experiments and shows additional tables and figures. This appendix contains two additional simulation experiments. The first is based on a sampling design distribution. The second is based on the model fitted to real data. Online supplementary material, Appendix C compares the predictors based on the ADMM with the predictors based on the multivariate Fay-Herriot and the area-level multinomial mixed model. Online supplementary material, Appendix D provides additional maps and tables to complete the application to the SLFS data.

2 Data and labour indicators

This paper presents and applies a new SAE methodology, based on an ADMM, to estimate labour indicators by Spanish province, sex and age group. To do so, we use data from the SLFS, which is published quarterly by the Spanish Statistical Office (Instituto Nacional de Estadística—INE). Specifically, the SLFS collects data about the labour market, including employment status, types of work, and other labour-related factors, as well as on the population outside the labour market.

The SLFS sample design is two-stage with stratification of first-stage units. The first-stage units are the census tracts. For their selection, they are grouped into strata according to the province and type of municipality (based on demographic importance) to which they belong. Within each stratum, the tracts are selected with probability proportional to the number of primary residences, according to data from the latest census or administrative register. The second-stage units are the primary family dwellings and fixed accommodations. Within each tract selected in the first stage, a fixed number of dwellings is selected using systematic sampling with a random start. Within the second-stage units, no subsampling is performed, and information is collected from all persons who normally reside there. Regarding the unemployment measurement, only persons aged 16 or older are of interest.

For the last quarter of 2022 (SLFS2022.4), the anonymized unit-level SLFS data file can be downloaded free of charge from the INE website and includes nearly 53,000 dwellings, equivalent to 128,000 people. However, our population of interest, U, consists of the working-age population aged 16 to 64 with permanent residence in Spain. For this reason, individuals under 16 or over 64 years of age are excluded, as they do not belong to the working-age population established in Spain, being either not above the legally minimum Spanish working age or already retired. Consequently, the size of the survey data file is reduced to $n = 103{,}000$ working-age respondents, approximately. Respondents were classified in domains defined by the variables province of residence, sex (1: male, 2: female) and age group. The variable age group has two categories, *age1* and *age2*, which are defined by grouping individuals according to their age within the ranges [16, 34] and [35, 64], respectively. The first age group (16–34) represents the first stage of working life, typically associated with the transition from education to employment, greater job instability, and greater vulnerability to labour market exclusion. The second group (35–64) represents older individuals of working age, who are generally associated with greater job stability, work experience, and consolidated career paths.

As regards territorial division, Spain has 52 provinces, including the autonomous cities of Ceuta and Melilla. There are $D = 52 \cdot 2 \cdot 2 = 208$ domains, $U_d \subset U$, defined by the crosses of province, sex and age group, respectively. The total number of people in the domain U_d is N_d, which only contains individuals

Table 1. Deciles of domain sample sizes and sampling fractions (in %)

	p_0	$p_{0.1}$	$p_{0.2}$	$p_{0.3}$	$p_{0.4}$	$p_{0.5}$	$p_{0.6}$	$p_{0.7}$	$p_{0.8}$	$p_{0.9}$	p_1
n_d	54	97	136	189	242	280	364	434	555	794	1,969
f_d	0.081	0.175	0.209	0.245	0.290	0.361	0.420	0.486	0.547	0.715	1.338

aged between 16 and 64. The sizes N_d are taken from the population projections published by the INE, and are the official sizes of the domains.

For n working-age respondents, it is expected an average of $n/D = 495$ respondents per domain. Regarding the sizes n_d of the samples $s_d \subset U_d$, we define the estimated sampling fractions

$$f_d = 100\frac{n_d}{\hat{N}_d^{dir}}, \quad \hat{N}_d^{dir} = \sum_{j\in s_d} w_{dj}, \quad d = 1, \ldots, D, \tag{2.1}$$

where $\hat{N}_d^{dir}$ is the estimated domain size and w_{dj} is the elevation factor of the jth individual of s_d. Table 1 presents the deciles of n_d and f_d, $d = 1, \ldots, D$. We observe that 10% (20 domains), or 30% (62 domains), of the sample sizes are lower than 97, or 189, and the average sample size 495 is between $p_{0.7} = 434$ and $p_{0.8} = 555$, which says that the sample size distribution is positively skewed. Furthermore, sampling fractions allow us to know the percentage of individuals from domains that actually belong to the sample. As they are all lower than 1.338 (in %), the representativeness of the samples is low. Since sample sizes are small in many domains, mapping labour force indicators using direct estimators is not sufficiently accurate. This motivates the incorporation of model-based predictors using SAE methods.

The domain labour indicators of interest are the proportions of employed ($k = 1$), unemployed ($k = 2$) and inactive ($k = 3$) people and the corresponding unemployment rates, i.e.

$$\overline{Y}_{dk} = \frac{1}{N_d}\sum_{j\in U_d} y_{djk}, \quad R_d = \frac{\overline{Y}_{d2}}{\overline{Y}_{d1} + \overline{Y}_{d2}}, \quad k = 1, 2, 3, \ d = 1, \ldots, D, \tag{2.2}$$

where $y_{djk} = 1$ if the individual j of the domain d belongs to category k and $y_{djk} = 0$ if the individual does not belong to category k.

The quantities (2.2) can be estimated using direct estimators, that is, relying only on the data from sample units in the domain of interest. Two direct estimators widely used in the literature are the Horvitz and Thompson (1952) estimator and the Hájek (1971) estimator. Although it is difficult to establish general conditions under which one of them is better than the other, Särndal et al. (1992) present some evidence in favour of the Hájek estimator. This has motivated its choice on our part, which is why, from now on, the Hájek estimator is called the direct estimator. The direct estimators of $\overline{Y}_{dk}$ and R_d are

$$\hat{\overline{Y}}_{dk}^{dir} = \frac{\sum_{j\in s_d} w_{dj} y_{djk}}{\sum_{j\in s_d} w_{dj}}, \quad \hat{R}_d^{dir} = \frac{\hat{\overline{Y}}_{d2}^{dir}}{\hat{\overline{Y}}_{d1}^{dir} + \hat{\overline{Y}}_{d2}^{dir}}, \quad k = 1, 2, 3, \ d = 1, \ldots, D. \tag{2.3}$$

The design-based covariance $\mathrm{cov}_\pi(\hat{\overline{Y}}_{dk_1}^{dir}, \hat{\overline{Y}}_{dk_2}^{dir})$, $k_1, k_2 = 1, 2, 3$, can be estimated by

$$\widehat{\mathrm{cov}}_\pi(\hat{\overline{Y}}_{dk_1}^{dir}, \hat{\overline{Y}}_{dk_2}^{dir}) = \frac{1}{\left(\hat{N}_d^{dir}\right)^2}\sum_{j\in s_d} w_{dj}(w_{dj} - 1)(y_{djk_1} - \hat{\overline{Y}}_{dk_1}^{dir})(y_{djk_2} - \hat{\overline{Y}}_{dk_2}^{dir}), \tag{2.4}$$

where the case $k_1 = k_2 = k$ denotes estimated variance, i.e. $\widehat{\mathrm{var}}_\pi(\hat{\overline{Y}}_{dk}^{dir}) = \widehat{\mathrm{cov}}_\pi(\hat{\overline{Y}}_{dk}^{dir}, \hat{\overline{Y}}_{dk}^{dir})$. The last formulas are obtained from Särndal et al. (1992), pp. 43, 185 and 391, with the simplifications $w_{dk} = 1/\pi_{dk}$,

Table 2. Deciles of the estimated design-based CVs for each labour force category, in %

	p_0	$p_{0.1}$	$p_{0.2}$	$p_{0.3}$	$p_{0.4}$	$p_{0.5}$	$p_{0.6}$	$p_{0.7}$	$p_{0.8}$	$p_{0.9}$	p_1
$k=1$	1	2	3	3	4	5	7	8	9	11	22
$k=2$	7	12	14	17	18	20	22	26	30	36	60
$k=3$	5	7	8	9	9	10	11	13	14	16	26

$\pi_{dk,dk} = \pi_{dk}$ and $\pi_{di,dj} = \pi_{di}\pi_{dj}$, $i \neq j$, in the second order inclusion probabilities. Morales et al. (2021), Chapter 2, presents Formula (2.4) as well as the R code for its calculation. On the other hand, Herrador et al. (2008) empirically investigate its behaviour in the SLFS2003 and verify its good performance with respect to two design-based resampling techniques: naive two-stage bootstrap and delete-one-cluster jackknife.

The design-based coefficients of variation (CVs), in %, of the direct estimators of each labour force category proportion are estimated by

$$\widehat{\mathrm{CV}}_{\pi}(\hat{\bar{Y}}_{dk}^{dir}) = 100\frac{\left(\widehat{\mathrm{var}}_{\pi}(\hat{\bar{Y}}_{dk}^{dir})\right)^{1/2}}{\hat{\bar{Y}}_{dk}^{dir}}, \quad k = 1, 2, 3.$$

Table 2 reports the estimated design-based CVs for each labour force category, rounded to zero decimal places. Notably, 50% of the domains present CVs greater than 20% for the unemployed category, which highlights the high relative uncertainty associated with direct estimates of unemployment proportions at the domain level. These results reinforce the need to use model-based predictors that provide estimates with a lower degree of uncertainty.

Since $y_{d1} = \hat{\bar{Y}}_{d1}^{dir}$, $y_{d2} = \hat{\bar{Y}}_{d2}^{dir}$ and $y_{d3} = \hat{\bar{Y}}_{d3}^{dir}$ add up to one, we may assume that $y_d = (y_{d1}, y_{d2})'$ follows a Dirichlet distribution with mean parameters $\mu_d = (\mu_{d1}, \mu_{d2})'$, $\mu_{d1} > 0$, $\mu_{d2} > 0$, $\mu_{d1} + \mu_{d2} < 1$ and concentration parameter $\phi_d > 0$. Section 3 introduces this distribution, in which the expectations, variances and covariances are

$$E[y_{dk}] = \mu_{dk} > 0, \quad \mathrm{var}(y_{dk}) = \frac{\mu_{dk}(1-\mu_{dk})}{(\phi_d+1)} > 0, \quad \mathrm{cov}(y_{d1}, y_{d2}) = \frac{-\mu_{d1}\mu_{d2}}{(\phi_d+1)} < 0, \quad k = 1, 2. \tag{2.5}$$

From (2.5) we have

$$\phi_d = \frac{\mu_{dk}(1-\mu_{dk})}{\mathrm{var}(y_{dk})} - 1, \quad k = 1, 2,$$

and in SAE problems we can estimate ϕ_d with

$$\hat{\phi}_d = \frac{1}{2}\sum_{k=1}^{2}\frac{y_{dk}(1-y_{dk})}{\widehat{\mathrm{var}}_{\pi}(y_{dk})} - 1. \tag{2.6}$$

Let us note that this value will be treated as known in the proposed model. It should be noted that, under the Dirichlet distribution, the precision parameter ϕ_d is unique and common to all categories. However, the estimates $\hat{\phi}_{dk}$ derived from each category k generally differ. To obtain a single robust estimate for the model, we calculate $\hat{\phi}_d$ as the average of these component-wise estimates. Additionally, if the direct estimates of the design-based variances are very small, we recommend applying the generalized variance function method (see Morales et al., 2021, p.427).

Figure 1 shows the boxplots of the design-based covariances of pairs $\{(y_{dk_1}, y_{dk_2}) : d = 1, \ldots, D\}$ for the categories $k_1, k_2 = 1, 2, 3$. We observe that the within-domain covariances are negative. Therefore,

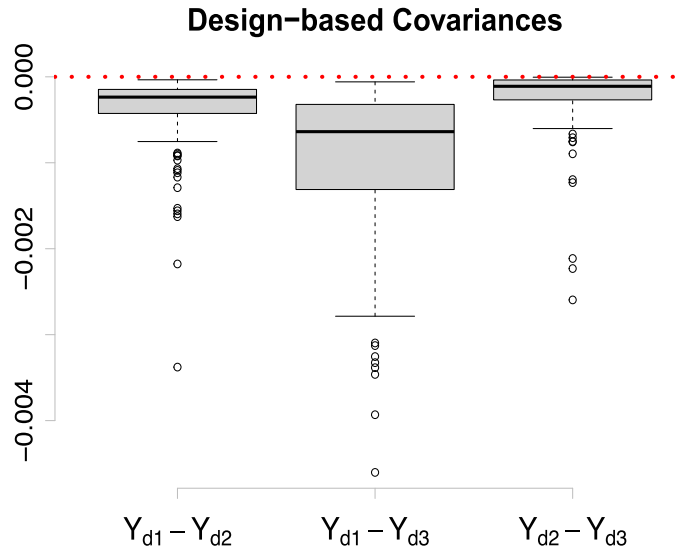

Figure 1. Boxplots of the design-based covariances of the target variables.

the data aligns with the Dirichlet model, which assumes negative correlation within-domain among the target variables due to the unit simplex constraint.

Drawing on auxiliary data to improve the direct estimators of the proportion of employed, unemployed and inactive people, and to obtain model-based predictors of unemployment rates by domains, Section 3 introduces an area-level Dirichlet mixed model.

3 Area-level Dirichlet mixed-effects model

This section introduces an area-level Dirichlet mixed model for a prefixed number q of categories, where the target vectors takes values in the simplex embedded in R^q

$$\mathcal{S}_e^{q-1} = \left\{(y_1, \ldots, y_q) \in R^q : y_1 > 0, \ \ldots, y_q > 0, \ y_1 + \cdots + y_q = 1\right\}.$$

As a sample space $\mathcal{S}_e^{q-1}$ is equivalent to the $(q-1)$-dimensional simplex defined by

$$\mathcal{S}^{q-1} = \left\{(y_1, \ldots, y_{q-1}) \in R^{q-1} : y_1 > 0, \ \ldots, y_{q-1} > 0, \ y_1 + \cdots + y_{q-1} < 1\right\}.$$

The vectors $y \in \mathcal{S}^{q-1}$ are called $(q-1)$-part compositions. For the definition of the new model, we use the parameterization $\mathcal{D}(\phi, \mu)$ of the Dirichlet distribution, proposed by Ferrari and Cribari-Neto (2004) for the beta distribution, where $\mu = (\mu_1, \ldots, \mu_{q-1}) \in \mathcal{S}^{q-1}$ is the vector of means and $\phi > 0$ is the concentration parameter. The probability density function (p.d.f.) of this distribution is

$$f(y|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu_1)\cdots\Gamma(\phi\mu_q)} y_1^{\phi\mu_1 - 1} \cdots y_{q-1}^{\phi\mu_{q-1} - 1} \left(1 - y_1 - \cdots - y_{q-1}\right)^{\phi\mu_q - 1}, \quad y \in \mathcal{S}^{q-1},$$

and $\mu_q = 1 - \mu_1 - \cdots - \mu_{q-1}$. In the population structure of a SAE problem, we have area-level random vectors $y_d = (y_{d1}, \ldots, y_{dq-1})' \in \mathcal{S}^{q-1}$ with components y_{dk} measured in domain d and category (group) k, $d = 1, \ldots, D$, $k = 1, \ldots, q-1$. To explain the behaviour of y_{dk}, $k = 1, \ldots, q-1$, we consider known row vectors $x_{dk} = (x_{dk1}, \ldots, x_{dkp_k})$ containing p_k explanatory variables and unknown column vectors $\beta_k = (\beta_{k1}, \ldots, \beta_{kp_k})'$ containing regression parameters, with $p = p_1 + \cdots + p_{q-1}$. Let $\sigma=$

$(\sigma_1, \ldots, \sigma_{q-1})'$ be a column vector of standard deviation parameters, where $\sigma_k > 0$, $k = 1, \ldots, q-1$, and let $u_{dk} \sim N(0,1)$, $d = 1, \ldots, D$, $k = 1, \ldots, q-1$, be independent random effects. The vectors of domain random effects are $u_d = (u_{d1}, \ldots, u_{dq-1})' \sim N(0, I_{q-1})$, $d = 1, \ldots, D$, where I_m is the $m \times m$ is the identity matrix. In matrix notation, we write $u = \underset{1 \le d \le D}{\text{col}}(u_d) \sim N(0, I_{D(q-1)})$. The p.d.f. of u is

$$f(u) = \prod_{d=1}^{D} f(u_d) = \prod_{d=1}^{D} \prod_{k=1}^{q-1} f(u_{dk}) = (2\pi)^{-D(q-1)/2} \exp\left\{-\frac{1}{2} u'u\right\}.$$

The area-level Dirichlet mixed model, with independent domain-category random effects, assumes that the distribution of the target vector y_d, conditioned to the random vector u_d, is Dirichlet. More concretely, it assumes that

$$y_d|_{u_d} \sim \mathcal{D}(\phi_d, \mu_d), \quad d = 1, \ldots, D, \tag{3.1}$$

where $\phi_d > 0$ is known, $\mu_d = (\mu_{d1}, \ldots, \mu_{dq-1})' \in \mathcal{S}^{q-1}$ and $\mu_{dq} = 1 - \sum_{k=1}^{q-1} \mu_{dk}$, $d = 1, \ldots, D$, are unknown, and that the logit link function holds, i.e.

$$\eta_{dk} = \log \frac{\mu_{dk}}{\mu_{dq}} = x_{dk}\beta_k + \sigma_k u_{dk}, \quad d = 1, \ldots, D, \;\; k = 1, \ldots, q-1.$$

For $d = 1, \ldots, D$, the components of the mean vectors are

$$\mu_{dk} = \frac{\exp\{x_{dk}\beta_k + \sigma_k u_{dk}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\beta_\ell + \sigma_\ell u_{d\ell}\}}, \quad k = 1, \ldots, q-1; \quad \mu_{dq} = \frac{1}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\beta_\ell + \sigma_\ell u_{d\ell}\}}.$$

The vector of model parameters is $\theta = (\beta', \sigma')'$, where $\beta = (\beta_1', \ldots, \beta_{q-1}')'$. The target vector is $y = \underset{1 \le d \le D}{\text{col}}(y_d)$. For $d = 1, \ldots, D$, the conditioned probability of y_d, given u, is

$$f_\theta(y_d|u) = f_\theta(y_d|u_d) = \frac{\Gamma(\phi_d)}{\prod_{k=1}^{q} \Gamma(\phi_d \mu_{dk})} \prod_{k=1}^{q} y_{dk}^{\phi_d \mu_{dk} - 1}, \quad y_d \in \mathcal{S}^{q-1} \quad \text{and} \quad y_{dq} = 1 - \sum_{k=1}^{q=1} y_{dk}.$$

Finally, we assume that $y_1, \ldots, y_d$ are independent, conditioned to u. The model likelihood is

$$f_\theta(y) = \int_{R^{D(q-1)}} f_\theta(y|u) f(u)\, du, \quad f_\theta(y|u) = \prod_{d=1}^{D} f(y_d|u_d). \tag{3.2}$$

The ML estimator $\hat{\theta}$ maximizes the log-likelihood function $\ell(\theta; y) = \log f_\theta(y)$. To perform the maximization, Section 3.1 applies the Laplace approximation of the integral in (3.2). In addition to the Laplace approximation approach, we propose an alternative fitting algorithm that approximates the multivariate integral in the model log-likelihood (arising from integrating out the domain-level random effects) via Gaussian quadrature. The maximization of the Gaussian quadrature algorithm and the R functions used are described in online supplementary material, Appendix A.1

3.1 ML-Laplace approximation algorithm

Let $h : R^m \mapsto R$ be a twice continuously differentiable function with a global maximum at the column vector x_0. This is to say, let us assume that the vector of first partial derivatives is $\dot{h}(x_0) = \frac{\partial h}{\partial x}|_{x=x_0} = 0$ and the matrix of second partial derivatives, $\ddot{h}(x_0) = \frac{\partial^2 h}{\partial x^2}|_{x=x_0}$, is negative definite. A Taylor series expansion

of $h(x)$ around x_0 yields to

$$\begin{aligned} h(x) &= h(x_0) + \dot{h}'(x_0)(x - x_0) + \frac{1}{2}(x - x_0)'\ddot{h}(x_0)(x - x_0) + o\left(\|x - x_0\|^2\right) \\ &\approx h(x_0) + \frac{1}{2}(x - x_0)'\ddot{h}(x_0)(x - x_0). \end{aligned}$$

As the p.d.f of $x \sim N_m(\mu, \Sigma)$ is

$$f(x) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu)\right\},$$

the multivariate Laplace approximation is

$$\begin{aligned} \int_{R^m} e^{h(x)}\, dx &\approx \int_{R^m} e^{h(x_0)} \exp\left\{-\frac{1}{2}(x - x_0)'\left(-\ddot{h}(x_0)\right)(x - x_0)\right\} dx \\ &= (2\pi)^{m/2}\left|-\ddot{h}(x_0)\right|^{-1/2} e^{h(x_0)}. \end{aligned} \tag{3.3}$$

Let us now approximate the likelihood (3.2). As the target vectors $y_1, \ldots, y_D$ are unconditionally independent and $u_d \sim N(0, I_{q-1})$, the marginal distribution of y_d is

$$\begin{aligned} f_\theta(y_d) &= \int_{R^{q-1}} f_\theta(y_d|u_d) f(u_d)\, du_d = \int_{\mathbb{R}^{q-1}} \frac{\Gamma(\phi_d)}{\prod_{j=1}^q \Gamma(\phi_d \mu_{dj})} \prod_{j=1}^q y_{dj}^{\phi_d \mu_{dj} - 1} f(u_d)\, du_d \\ &= \frac{\Gamma(\phi_d)}{(2\pi)^{(q-1)/2}} \int_{\mathbb{R}^{q-1}} \exp\left\{h_d(u_d)\right\} du_d, \end{aligned} \tag{3.4}$$

where

$$h_d(u_d; y_d, \theta) = h_d(u_d) = \sum_{j=1}^q \left\{(\phi_d \mu_{dj} - 1)\log y_{dj} - \log\Gamma(\phi_d \mu_{dj})\right\} - \frac{1}{2}\sum_{j=1}^{q-1} u_{dj}^2. \tag{3.5}$$

To apply the Laplace approximation to the integral in (3.4), we have to maximize $h_d(u_d; y_d, \theta)$ in u_d, given y_d and θ. For simplicity, we use the notation $h_d(u_d)$. The maximization can be done by applying an R-function of optimization. Alternatively, it is possible to implement a Newton-Raphson algorithm after calculating the first and second partial derivatives of h_d with respect to u_d, $d = 1, \ldots, D$, given y_d and θ. Online supplementary material, Appendix A.2 gives the first and second derivatives of the components of μ_d with respect to the components of u_d. The first derivatives of h_d, with respect to u_{dk}, are

$$\frac{\partial h_d(u_d)}{\partial u_{dk}} = \phi_d \sum_{j=1}^q \frac{\partial \mu_{dj}}{\partial u_{dk}} \log y_{dj} - \phi_d \sum_{j=1}^q \psi(\phi_d \mu_{dj}) \frac{\partial \mu_{dj}}{\partial u_{dk}} - u_{dk}, \quad k = 1, \ldots, q-1,$$

where $\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z}\log\Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$ is the digamma function. The second derivatives of h_d, with respect to u_{dk} and u_{dk} , are

$$\frac{\partial^2 h_d(u_d)}{\partial u_{dk}^2} = \phi_d \sum_{j=1}^q \frac{\partial^2 \mu_{dj}}{\partial u_{dk}^2} \log y_{dj} - \phi_d^2 \sum_{j=1}^q \psi_1(\phi_d \mu_{dj}) \left(\frac{\partial \mu_{dj}}{\partial u_{dk}}\right)^2 - \phi_d \sum_{j=1}^q \psi(\phi_d \mu_{dj}) \frac{\partial^2 \mu_{dj}}{\partial u_{dk}^2} - 1,$$

where $\psi_1(z) = \frac{\mathrm{d}}{\mathrm{d}z}\psi(z)$ is the trigamma function. The second derivatives of h_d, with respect to u_{dk_1} and

u_{dk_2}, $k_1, k_2 = 1, \ldots, q-1$, $k_1 \neq k_2$, are

$$\frac{\partial^2 h_d(u_d)}{\partial u_{dk_1}\partial u_{dk_2}} = \phi_d \sum_{j=1}^{q} \frac{\partial^2 \mu_{dj}}{\partial u_{dk_1}\partial u_{dk_2}} \log y_{dj} - \phi_d^2 \sum_{j=1}^{q} \psi_1(\phi_d\mu_{dj}) \frac{\partial \mu_{dj}}{\partial u_{dk_1}} \frac{\partial \mu_{dj}}{\partial u_{dk_2}} - \phi_d \sum_{j=1}^{q} \psi(\phi_d\mu_{dj}) \frac{\partial^2 \mu_{dj}}{\partial u_{dk_1}\partial u_{dk_2}}.$$

Let us define the score vector

$$S_d(u_d, \theta) = \left(\frac{\partial h_d(u_d)}{\partial u_{d1}}, \ldots, \frac{\partial h_d(u_d)}{\partial u_{dq-1}} \right)'$$

and the Jacobian matrix

$$H_d(u_d, \theta) = \begin{pmatrix} \frac{\partial^2 h_d(u_d)}{\partial u_{d1}^2} & \cdots & \frac{\partial^2 h_d(u_d)}{\partial u_{d1}\partial u_{dq-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 h_d(u_d)}{\partial u_{dq-1}\partial u_{d1}} & \cdots & \frac{\partial^2 h_d(u_d)}{\partial u_{dq-1}^2} \end{pmatrix}.$$

For $\theta = (\beta', \sigma')'$ fixed, the function $h_d(u_d)$, defined in (3.5), can be maximized by using the Newton-Raphson algorithm. The updating equation is

$$u_d^{(r+1)} = u_d^{(r)} - H_d^{-1}(u_d^{(r)}, \theta) S_d(u_d^{(r)}, \theta), \quad d = 1, \ldots, D. \tag{3.6}$$

Let us denote by u_{0d} the argument of maxima of $h_d(u_d)$. Thus, it holds $\dot{h}(u_{0d}) = 0$ and $\ddot{h}(u_{0d})$ is negative definite. The model loglikelihood is $\ell = \ell(\theta; y) = \sum_{d=1}^{D} \log f_\theta(y_d) = \sum_{d=1}^{D} \ell_d$. By applying (3.3) in $u_d = u_{0d}$ to (3.4), we get the Laplace approximation ℓ_{0d} of the term ℓ_d, where

$$\begin{aligned} \ell_{0d} &= \ell_d(\theta; y, u_{0d}) = \log\Gamma(\phi_d) + h_d(u_{0d}) - \frac{1}{2}\log|-H_d(u_{0d}, \theta)| \\ &= \log\Gamma(\phi_d) + \sum_{j=1}^{q} \left\{ (\phi_d\mu_{0dj} - 1)\log y_{dj} - \log\Gamma(\phi_d\mu_{0dj}) \right\} - \frac{1}{2}\sum_{j=1}^{q-1} u_{0dj}^2 - \frac{1}{2}\log|-H_d(u_{0d}, \theta)|, \end{aligned} \tag{3.7}$$

and

$$\mu_{0dj} = \frac{\exp\{x_{dj}\beta_j + \sigma_j u_{0dj}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\beta_\ell + u_{0d\ell}\}}, \quad j = 1, \ldots, q-1; \quad \mu_{0dq} = \frac{1}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\beta_\ell + u_{0d\ell}\}}.$$

The following step is to maximize $\ell_0(\theta) \triangleq \sum_{d=1}^{D} \ell_d(\theta; y, u_{0d})$, $\theta \in \Theta$. Let us define $M = \dim(\Theta) = p + q - 1$. Let $\dot{\ell}_0$ and $\ddot{\ell}_0$ denote the $M \times 1$ vector and the $M \times M$ matrix of first and second order partial derivatives of $\ell_0(\theta)$, respectively. The Newton-Raphson updating equation is

$$\theta^{(r+1)} = \theta^{(r)} - \ddot{\ell}_0^{-1}(\theta^{(r)}) \dot{\ell}_0(\theta^{(r)}). \tag{3.8}$$

Alternatively, we can apply the derivative-free optimization algorithm NB, included in the $\mathbb{R}$ function nloptr, and calculate

$$\theta^{(r+1)} = \text{argmax}_{\theta\in\Theta} \ell_0(\theta), \quad \Theta = \{(\beta', \sigma')' \in \mathbb{R}^{p+q-1} : \sigma_1 > 0, \ldots, \sigma_{q-1} > 0\}. \tag{3.9}$$

The final ML-Laplace approximation algorithm combines the two described maximization algorithms

and has the following steps:

1. Set the initial values $r = 0$, $\theta^{(0)}$, $\theta^{(-1)} = \theta^{(0)} + 1_{p+q-1}$, $u_d^{(0)} = 0_{q-1}$, $u_d^{(-1)} = 1_{q-1}$, $d = 1, \ldots, D$.
2. Until $\|\theta^{(r)} - \theta^{(r-1)}\|_2 < \varepsilon_1$, $\|u_d^{(r)} - u_d^{(r-1)}\|_2 < \varepsilon_2$, $d = 1, \ldots, D$, do
   (a) Apply algorithm (3.6) with seeds $u_d^{(r)}$, $d = 1, \ldots, D$, convergence tolerance ε_2 and $\theta = \theta^{(r)}$ fixed. Output: $u_d^{(r+1)}$, $d = 1, \ldots, D$.
   (b) Apply algorithm (3.8) (or maximization (3.9)) with seed $\theta^{(r)}$, convergence tolerance ε_1 and $u_d = u_d^{(r+1)}$ fixed, $d = 1, \ldots, D$. Output: $\theta^{(r+1)}$.
   (c) $r \leftarrow r + 1$.
3. Output: $\hat{\theta} = \theta^{(r)}$, $\hat{u}_d = u_d^{(r)}$, $d = 1, \ldots, D$.

Note that the ML-Laplace approximation algorithm provides estimates of the model parameters and modal predictors of the random effects at convergence.

3.2 Confidence intervals

Given any model parameter η, an asymptotic 100(1-α)% confidence interval (CI) is

$$\mathrm{CI}_{1-\alpha}(\eta) = (\hat{\eta} - z_{1-\alpha/2}\hat{\sigma}^*_{\hat{\eta}}, \hat{\eta} + z_{1-\alpha/2}\hat{\sigma}^*_{\hat{\eta}}), \tag{3.10}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $\hat{\sigma}^*_{\hat{\eta}}$ is the bootstrap estimator (cf. Section 4.3) of the standard deviation of $\hat{\eta}$. Alternatively, we propose the basic bootstrap confidence interval

$$\mathrm{CI}^*_{1-\alpha}(\eta) = \left(\hat{\eta} - \frac{x^*_{1-\alpha/2}}{\sqrt{B}}, \hat{\eta} - \frac{x^*_{\alpha/2}}{\sqrt{B}}\right) \tag{3.11}$$

where B is the number of bootstrap replicates and $x^*_{\alpha/2}, x^*_{1-\alpha/2}$ are the quantiles $\alpha/2$ and $1-\alpha/2$ of the bootstrap distribution of the statistic $R^* = \sqrt{B}(\hat{\eta}^* - \hat{\eta})$.

4 Predictors

Under the assumption that y_d, $d = 1, \ldots, D$, follows the area-level Dirichlet mixed model defined around (3.1), this section derives predictors of area-level proportions and indicators. This section assumes that $\phi_d > 0$ is known, $d = 1, \ldots, D$. In practice, ϕ_d is substituted by $\hat{\phi}_d$ given in (2.6). To highlight the dependence of μ_{dk} on u_d and θ, we write $\mu_{dk}(\theta, u_d)$.

4.1 Predictors of proportions

This section derives a predictor $\hat{\mu}_{dk}$ of $\mu_{dk} = \mu_{dk}(\theta, u_d)$ with minimum MSE in the class of unbiased predictors for known θ. The predictor that satisfies this property is the best predictor (BP), i.e. $\hat{\mu}^{bp}_{dk}(\theta) = E_\theta[\mu_{dk}|y]$. It holds that $E_\theta[\mu_{dk}|y] = E_\theta[\mu_{dk}|y_d]$ and

$$E_\theta[\mu_{dk}|y_d] = \frac{\int_{R^{q-1}} \dfrac{\exp\{x_{dk}\beta_k + \sigma_k u_{dk}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\beta_\ell + \sigma_\ell u_{d\ell}\}} f(y_d|u_d) f(u_d)\, du_d}{\int_{R^{q-1}} f(y_d|u_d) f(u_d)\, du_d} = \frac{A_{dk}(y_d, \theta)}{D_d(y_d, \theta)},$$

where $A_{dk} = A_{dk}(y_d, \theta)$ and $D_d = D_d(y_d, \theta)$ are

$$\begin{aligned} A_{dk} &= \int_{R^{q-1}} \frac{\exp\{x_{dk}\beta_k + \sigma_k u_{dk}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\beta_\ell + \sigma_\ell u_{d\ell}\}} \\ &\quad \cdot \exp\left\{\sum_{j=1}^{q} \left\{(\phi_d \mu_{dj}(\theta, u_d) - 1)\log y_{dj} - \log\Gamma\big(\phi_d \mu_{dj}(\theta, u_d)\big)\right\}\right\} f(u_d)\, du_d, \\ D_d &= \int_{R^{q-1}} \exp\left\{\sum_{j=1}^{q} \left\{(\phi_d \mu_{dj}(\theta, u_d) - 1)\log y_{dj} - \log\Gamma\big(\phi_d \mu_{dj}(\theta, u_d)\big)\right\}\right\} f(u_d)\, du_d. \end{aligned} \tag{4.1}$$

The empirical best predictor (EBP) of μ_{dk} is obtained by replacing θ with its estimate $\hat{\theta}$, i.e. $\hat{\mu}_{dk}^{ebp} = \hat{\mu}_{dk}^{bp}(\hat{\theta})$, and can be approximated using the Monte Carlo method as follows.

1. Estimate $\hat{\theta} = (\hat{\beta}', \hat{\sigma}')'$.
2. For $s = 1, \ldots, S$, generate $u_d^{(s)}$ i.i.d. $N(0, I_{q-1})$ and set $u_d^{(S+s)} = -u_d^{(s)}$.
3. Calculate $\hat{\mu}_{dk}^{ebp} = \hat{A}_{dk}/\hat{D}_d$, $d = 1, \ldots, D$, where

$$\begin{aligned} \hat{A}_{dk} &= \frac{1}{2S}\sum_{s=1}^{2S} \frac{\exp\{x_{dk}\hat{\beta}_k + \hat{\sigma}_k u_{dk}^{(s)}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\hat{\beta}_\ell + \hat{\sigma}_\ell u_{d\ell}^{(s)}\}} \\ &\quad \cdot \exp\left\{\sum_{j=1}^{q} \left\{(\phi_d \mu_{dj}(\hat{\theta}, u_d^{(s)}) - 1)\log y_{dj} - \log\Gamma\big(\phi_d \mu_{dj}(\hat{\theta}, u_d^{(s)})\big)\right\}\right\}, \\ \hat{D}_d &= \frac{1}{2S}\sum_{s=1}^{2S} \exp\left\{\sum_{j=1}^{q} \left\{(\phi_d \mu_{dj}(\hat{\theta}, u_d^{(s)}) - 1)\log y_{dj} - \log\Gamma\big(\phi_d \mu_{dj}(\hat{\theta}, u_d^{(s)})\big)\right\}\right\}. \end{aligned} \tag{4.2}$$

If $n_d = 0$, the BP of μ_{dk} is $\hat{\mu}_{dk}(\theta) = E_\theta[\mu_{dk}]$, which is also called synthetic BP (sBP). The synthetic EBP (sEBP) can be approximated as the EBP when $n_d > 0$, but by running Step 3:

3. Calculate

$$\hat{\mu}_{dk}^{sebp} = \frac{1}{2S}\sum_{s=1}^{2S} \frac{\exp\{x_{dk}\hat{\beta}_k + \hat{\sigma}_k u_{dk}^{(s)}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\hat{\beta}_\ell + \hat{\sigma}_\ell u_{d\ell}^{(s)}\}}, \quad d = 1, \ldots, D.$$

The main property of the derived BP is their minimal MSE in the class of unbiased predictors. Unfortunately, this does not hold for the EBPs, which are obtained by replacing the true parameters with their estimated values and are therefore not unbiased. Assuming that the model parameter estimates are consistent, the EBPs are asymptotically unbiased as $D \to \infty$, but the number of domains may not be large enough in SAE problems.

Even when θ is known, we must approximate the integrals using Monte Carlo or other numerical methods. These approximations have a high computational cost and introduce another source of error that can increase the variance of the final predictor. Therefore, it makes sense to also consider less computationally demanding plug-in predictors. The plug-in predictor of μ_{dk} is

$$\hat{\mu}_{dk}^{in} = \frac{\exp\{x_{dk}\hat{\beta}_k + \hat{\sigma}_k \hat{u}_{dk}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\hat{\beta}_\ell + \hat{\sigma}_\ell \hat{u}_{d\ell}\}}, \tag{4.3}$$

where $\hat{u}_{dk}$ is a predictor (EBP or likelihood mode) of u_{dk}, $d = 1, \ldots, D$, $k = 1, \ldots, q-1$. The EBP of u_{dk} is $\hat{u}_{dk}^{ebp} = E_{\hat{\theta}}[u_{dk}|y_d]$ and can be approximated by $\hat{u}_{dk}^{ebp} = \hat{A}_{u,dk}/\hat{D}_d$, where

$$\hat{A}_{u,dk} = \frac{1}{2S}\sum_{s=1}^{2S} u_{dk}^{(s)} \exp\left\{\sum_{j=1}^{q}\left\{(\phi_d\mu_{dj}(\theta, u_d^{(s)}) - 1)\log y_{dj} - \log\Gamma(\phi_d\mu_{dj}(\theta, u_d^{(s)}))\right\}\right\}$$

and $\hat{D}_d$ is given in (4.2). The likelihood modes of u_{dk} can be taken from the output of the ML-Laplace approximation algorithm given on page 3.2.

4.2 Predictors of unemployment rates

This section assumes that $q = 3$ and μ_{d1}, μ_{d2} and μ_{d3} are the proportions of employed, unemployed and inactive people in the domain d, $d = 1, \ldots, D$. The unemployment rates are

$$R_d = R_d(\theta, u_d) = \frac{\mu_{d2}(\theta, u_d)}{\mu_{d1}(\theta, u_d) + \mu_{d2}(\theta, u_d)}, \quad d = 1, \ldots, D.$$

The best predictor (BP) of R_d is $\hat{R}_d^{bp}(\theta) = E_\theta[R_d|y]$. In this case, it holds that $E_\theta[R_d|y] = E_\theta[R_d|y_d]$ and

$$E_\theta[R_d|y_d] = \frac{\int_{R^{q-1}} R_d(\theta, u_d) f(y_d|u_d) f(u_d)\, du_d}{\int_{R^{q-1}} f(y_d|u_d) f(u_d)\, du_d} = \frac{A_{Rd}(y_d, \theta)}{D_d(y_d, \theta)},$$

where $A_{Rd} = A_{Rd}(y_d, \theta)$ and $D_d = D_d(y_d, \theta)$ are

$$A_{Rd} = \int_{R^{q-1}} R_d(\theta, u_d) \exp\left\{\sum_{j=1}^{q}\left\{(\phi_d\mu_{dj}(\theta, u_d) - 1)\log y_{dj} - \log\Gamma(\phi_d\mu_{dj}(\theta, u_d))\right\}\right\} f(u_d)\, du_d,$$

$$D_d = \int_{R^{q-1}} \exp\left\{\sum_{j=1}^{q}\left\{(\phi_d\mu_{dj}(\theta, u_d) - 1)\log y_{dj} - \log\Gamma(\phi_d\mu_{dj}(\theta, u_d))\right\}\right\} f(u_d)\, du_d.$$

The EBP of R_d is $\hat{R}_d^{ebp} = \hat{R}_d^{bp}(\hat{\theta})$ and can be approximated as follows.

1. Estimate $\hat{\theta} = (\hat{\beta}', \hat{\sigma}')'$.
2. For $s = 1, \ldots, S$, generate $u_d^{(s)}$ i.i.d. $N(0, I_{q-1})$ and set $u_d^{(S+s)} = -u_d^{(s)}$.
3. Calculate $\hat{R}_d^{ebp} = \hat{A}_{Rd}/\hat{D}_d$, where

$$\hat{A}_{Rd} = \frac{1}{2S}\sum_{s=1}^{2S} R_d(\hat{\theta}, u_d^{(s)}) \exp\left\{\sum_{j=1}^{q}\left\{(\phi_d\mu_{dj}(\hat{\theta}, u_d^{(s)}) - 1)\log y_{dj} - \log\Gamma(\phi_d\mu_{dj}(\hat{\theta}, u_d^{(s)}))\right\}\right\},$$

$$\hat{D}_d = \frac{1}{2S}\sum_{s=1}^{2S} \exp\left\{\sum_{j=1}^{q}\left\{(\phi_d\mu_{dj}(\hat{\theta}, u_d^{(s)}) - 1)\log y_{dj} - \log\Gamma(\phi_d\mu_{dj}(\hat{\theta}, u_d^{(s)}))\right\}\right\}.$$

If $n_d = 0$, the sBP of μ_{dk} is $\hat{\mu}_{dk}(\theta) = E_\theta[\mu_{dk}]$. The sEBP of R_d can be approximated as the EBP when $n_d > 0$, but by running Step 3:

3. Calculate

$$\hat{R}_d^{sebp} = \frac{1}{2S}\sum_{s=1}^{2S} \frac{\exp\{x_{d2}\hat{\beta}_2 + \hat{\sigma}_2 u_{d2}^{(s)}\}}{\exp\{x_{d1}\hat{\beta}_1 + \hat{\sigma}_1 u_{d1}^{(s)}\} + \exp\{x_{d2}\hat{\beta}_2 + \hat{\sigma}_2 u_{d2}^{(s)}\}}, \quad d = 1, \ldots, D.$$

The plug-in and the plug-in-EBP predictors of R_d are

$$\hat{R}_d^{in} = \frac{\hat{\mu}_{d2}^{in}}{\hat{\mu}_{d1}^{in} + \hat{\mu}_{d2}^{in}}, \quad \hat{R}_d^{in.ebp} = \frac{\hat{\mu}_{d2}^{ebp}}{\hat{\mu}_{d1}^{ebp} + \hat{\mu}_{d2}^{ebp}}, \quad d = 1, \ldots, D. \tag{4.4}$$

4.3 Bootstrap estimation of the MSE of a predictor of μ_{dk}

The following procedure, based on González-Manteiga et al. (2008, 2010), calculates a parametric bootstrap estimator of $MSE(\hat{\mu}_{dk})$, where $\hat{\mu}_{dk}$ is a predictor (EBP or plug-in) of μ_{dk}.

1. Fit the model to the sample and calculate $\hat{\theta} = (\hat{\beta}', \hat{\sigma}')'$.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) Bootstrap sample: Generate a set of random effects $\{u_d^{*(b)} : d = 1, \ldots, D\}$ i.i.d. $N(0, I_{q-1})$. For $d = 1, \ldots, D$, generate the elements of the bootstrap sample

$$y_d^{*(b)} \sim \mathcal{D}(\phi_d, \mu_d^{*(b)}), \quad \mu_d^{*(b)} = (\mu_{d1}^{*(b)}, \ldots, \mu_{d(q-1)}^{*(b)}) \in \mathcal{S}^{q-1},$$

   where

$$\mu_{dk}^{*(b)} = \frac{\exp\{x_{dk}\hat{\beta}_k + \hat{\sigma}_k u_{dk}^{*(b)}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\hat{\beta}_\ell + \hat{\sigma}_\ell u_{d\ell}^{*(b)}\}}, \quad k = 1, \ldots, q-1.$$

   (b) Bootstrap model: Fit an area-level Dirichlet mixed model to the bootstrap sample

$$(y_d^{*(b)}, x_d), \quad x_d = (x_{d1}, \ldots, x_{d(q-1)}), \quad d = 1, \ldots, D.$$

   Calculate the parameter estimator $\hat{\theta}^{*(b)} = (\hat{\beta}^{*(b)\prime}, \hat{\sigma}^{*(b)\prime})'$ and the random effects predictors $\hat{u}_{dk}^{*(b)}$, $d = 1, \ldots, D$, $k = 1, \ldots, q-1$. Calculate the predictor $\hat{\mu}_{dk}^{*(b)}$ of the bootstrap population quantity $\mu_{dk}^{*(b)}$. For example, the plug-in predictor of $\mu_{dk}^{*(b)}$ is

$$\hat{\mu}_{dk}^{in*(b)} = \frac{\exp\{x_{dk}\hat{\beta}_k^{*(b)} + \hat{\sigma}_k^{*(b)}\hat{u}_{dk}^{*(b)}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{x_{d\ell}\hat{\beta}_\ell^{*(b)} + \hat{\sigma}_\ell^{*(b)}\hat{u}_{d\ell}^{*(b)}\}}.$$

3. Output: $mse^*(\hat{\mu}_{dk}) = \frac{1}{B}\sum_{b=1}^{B} (\hat{\mu}_{dk}^{*(b)} - \mu_{dk}^{*(b)})^2$.

5 Simulations

This section presents some simulation experiments for area-level Dirichlet mixed models with $q = 3$ categories. The target of Simulation 1 is to check and compare the behaviour of the Laplace approximation algorithm and the Gaussian quadrature. The ML-Laplace approximation results are provided in the main manuscript, whereas the Gaussian quadrature results are given in online supplementary material, Appendix B.1 Simulation 2 is designed to analyse the biases and the root mean squared errors of the predictors of proportions and rates. Simulation 3 investigates the parametric bootstrap estimators of the MSEs and gives recommendations on the number B of bootstrap samples to be

generated. In addition, online supplementary material, Appendices B.4 and B.5 present two online supplementary material, supplementary simulations. In the first, we conduct a design-based simulation by drawing samples from a synthetic census that mimics the target population in our real-data application. In the second, we simulate the data from the model fitted to real data, but varying the parameter values to analyse their effect on the accuracy of the predictors.

To generate the explanatory variables, we take $\mu_{x1} = \mu_{x2} = 1$, $\sigma_{x11} = 1$ and $\sigma_{x22} = 3/2$. For $k = 1, 2$, $d = 1, \ldots, D$, generate $x_{d1} = (x_{d11}, x_{d12})$, $x_{d2} = (x_{d21}, x_{d22})$, $x_{d11} = x_{d21} = 1$,

$$x_{d12} = \mu_{x1} + \sigma_{x11}^{1/2} U_{d1}, \quad x_{d22} = \mu_{x2} + \sigma_{x22}^{1/2} U_{d2}, \quad U_{dk} \overset{\text{ind}}{\sim} U(0, 1). \tag{5.1}$$

The regression parameters are $\beta_1 = (\beta_{11}, \beta_{12})' = (-1/2, 1)'$, $\beta_2 = (\beta_{21}, \beta_{22})' = (1, -1/3)'$. The variance parameters are $\sigma_1 = 1/3$, $\sigma_2 = 1/2$. The concentration parameters are $\phi_d = \exp\{2 + 1/d\}$, $d = 1, \ldots, D$. For $d = 1, \ldots, D$, we simulate the vector of random effects $u_d \sim N_2(0, I_2)$, we generate the components of the mean vectors

$$\mu_{dk} = \frac{\exp\{x_{dk}\beta_k + \sigma_k u_{dk}\}}{1 + \sum_{\ell=1}^{2} \exp\{x_{d\ell}\beta_\ell + \sigma_\ell u_{d\ell}\}}, \quad k = 1, 2; \quad \mu_{d3} = \frac{1}{1 + \sum_{\ell=1}^{2} \exp\{x_{d\ell}\beta_\ell + \sigma_\ell u_{d\ell}\}},$$

and we simulate the vectors of target variables $y_d|_{u_d} \sim \mathcal{D}(\phi_d, \mu_d)$, $\mu_d = (\mu_{d1}, \mu_{d2})$.

Let η denote a generic scalar target, which may be either a model parameter component or a model-based predictor μ_{dk}, R_d, $d = 1, \ldots, D$, $k = 1, 2, 3$. Let $\{\hat{\eta}^{(i)}\}_{i=1}^{I}$ be the corresponding estimate/prediction at ith Monte Carlo replicate. The absolute and relative performance measures are

$$RMSE(\hat{\eta}) = \left(\frac{1}{I}\sum_{i=1}^{I}(\hat{\eta}^{(i)} - \eta)^2\right)^{1/2}, \quad BIAS(\hat{\eta}) = \frac{1}{I}\sum_{i=1}^{I}(\hat{\eta}^{(i)} - \eta),$$
$$RRMSE(\hat{\eta}) = 100\frac{RMSE(\hat{\eta})}{|\eta|}, \qquad RBIAS(\hat{\eta}) = 100\frac{BIAS(\hat{\eta})}{|\eta|}.$$

5.1 Simulation 1

Simulation 1 investigates the performance of the Laplace approximation and the Gaussian quadrature algorithms for fitting the area-level Dirichlet mixed model (3.1). To this end, Simulation 1 empirically calculates the root-MSE $RMSE(\hat{\eta})$, the bias $BIAS(\hat{\eta})$, the relative root-MSE $RRMSE(\hat{\eta})$ and the relative bias $RBIAS(\hat{\eta})$ of the ML estimators $\hat{\eta} \in \{\hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{21}, \hat{\beta}_{22}, \hat{\sigma}_1, \hat{\sigma}_2\}$. Online supplementary material, Appendix B.1 describes the Monte Carlo simulation steps. The Simulation 1 is carried out with $I = 1{,}000$ iterations.

Tables 3–4 present the empirical absolute and relative performance measures $BIAS(\hat{\eta})$, $RBIAS(\hat{\eta})$, $RRMSE(\hat{\eta})$ and $RBIAS(\hat{\eta})$. The column labelled by η contains the values of the true model parameters. Online supplementary material, Tables B.1–B.4 present the absolute and relative performance measures under the Gaussian quadrature algorithm.

Simulation 1 shows that the ML-Laplace fitting algorithm works properly because the relative biases are always below 10% in absolute value and the relative root-MSEs decrease as D increases.

Regarding the regression parameters (β_{11}, β_{12}, β_{21} and β_{22}), the bias fluctuates around zero with negligible values, reflecting Monte Carlo variability rather than systematic bias. In contrast, the variance components (σ_1 and σ_2) show a consistent decreasing trend in bias as D increases.

The negative relative biases of σ_1 and σ_2 from Table 3 are somehow expected: the ML variance-component estimators are known to be negatively biased, because ML does not adjust for the estimation of fixed effects. REML typically reduces the variance bias, however, this is an available option for LMMs. The ADMM is a GLMM and, unfortunately, the authors are not aware of any solutions to the variance-component bias for GLMMs in the literature.

Table 3. $BIAS(\hat{\eta})$ (left) and $RBIAS(\hat{\eta})$ (right). The column η shows the model parameters

		D					D				
	η	50	75	100	150	200	50	75	100	150	200
β_{11}	−1/2	−0.014	−0.008	−0.048	−0.027	−0.016	−2.79	−1.74	−9.66	−5.38	−3.19
β_{12}	1	0.013	0.008	0.030	0.020	0.013	1.30	0.85	3.00	2.07	1.34
β_{21}	1	−0.000	0.010	−0.007	−0.011	−0.001	−0.00	1.01	−0.68	−1.15	−0.17
β_{22}	−1/3	−0.008	−0.016	−0.008	−0.003	−0.009	−2.38	−5.03	−2.64	−1.11	−2.86
σ_1	1/3	−0.036	−0.030	−0.026	−0.027	−0.022	−10.85	−9.11	−7.92	−8.28	−6.61
σ_2	1/2	−0.048	−0.041	−0.039	−0.035	−0.030	−9.63	−8.27	−7.83	−7.01	−6.10

Table 4. $RMSE(\hat{\eta})$ (left) and $RRMSE(\hat{\eta})$ (right). The column η shows the model parameters

		D					D				
	η	50	75	100	150	200	50	75	100	150	200
β_{11}	−1/2	0.480	0.371	0.321	0.255	0.224	96.10	74.28	64.16	51.02	44.83
β_{12}	1	0.304	0.238	0.206	0.162	0.142	30.46	23.85	20.61	16.21	14.19
β_{21}	1	0.396	0.313	0.268	0.228	0.199	39.56	31.27	26.85	22.79	19.97
β_{22}	−1/3	0.235	0.183	0.157	0.135	0.121	70.63	55.08	47.03	40.54	36.50
σ_1	1/3	0.075	0.063	0.056	0.049	0.043	22.61	19.04	17.01	14.86	12.85
σ_2	1/2	0.116	0.094	0.084	0.073	0.061	23.17	18.78	16.80	14.56	12.31

Table 5. Average per-run computational times (in seconds) of the Laplace approximation and the Gaussian quadrature algorithms

	D				
Fitting algorithm	**50**	**75**	**100**	**150**	**200**
Laplace	1.670	2.268	2.633	3.830	5.254
Gauss-Hermite	42.636	60.819	99.039	119.719	131.843

For the Gauss-Hermite quadrature fit, we used a 25×25 product grid (625 points). Table 5 presents the average per-run computational times (in seconds) of the ML-Laplace approximation and the Gaussian quadrature algorithms for $D = 50, 75, 100, 150, 200$. The Laplace approximation algorithm demonstrates substantially lower average computational time per run compared with Gaussian quadrature. The Simulation was run on a MacBook Pro (Apple M3, 8-core CPU) with 16 GB RAM, using the latest stable release of `R` available at the time of analysis (4.4.1 version).

On the other hand, comparing the results for $RRMSE(\hat{\eta})$, the Laplace approximation markedly outperforms Gaussian quadrature algorithm. This outcome is expected since Laplace is a second-order accurate method that adapts to the local mode and curvature of the integrand, yielding stable and reliable estimates. By contrast, the Gaussian quadrature algorithm approximates the bivariate integral by a weighted sum in a grid of points and weigths, therefore it is a free-derivative algorithm. In our setting, even with a 25×25 grid, the quadrature does not achieve good precision in the integration of the log-likelihood, inflating error and thus the $RRMSE(\hat{\eta})$ Achieving comparable accuracy would typically require substantially denser and/or adaptive grids (e.g. sparse-grid rules, or higher points per dimension), but these options increase computational cost rapidly. Finally, one crucial point of the Gauss-Hermite method is that it allows us to calculate the ML estimators but does not provide

Table 6. *RAB* (left) and *RRE* (right) of model-based predictors

	D					*D*				
	50	**75**	**100**	**150**	**200**	**50**	**75**	**100**	**150**	**200**
$\hat{\mu}_{d1}^{ebp}$	0.510	0.578	0.508	0.575	0.591	17.49	17.50	17.33	17.41	17.33
$\hat{\mu}_{d1}^{in}$	0.728	0.922	0.858	0.939	1.005	17.88	17.79	17.55	17.57	17.45
$\hat{\mu}_{d2}^{ebp}$	1.134	1.243	1.143	1.117	1.129	29.82	29.77	29.66	29.60	29.41
$\hat{\mu}_{d2}^{in}$	1.785	1.940	1.924	1.954	2.009	30.41	30.20	30.00	29.84	29.58
$\hat{\mu}_{d3}^{ebp}$	0.813	0.655	0.731	0.680	0.595	23.70	23.11	22.76	22.51	22.41
$\hat{\mu}_{d3}^{in}$	0.901	0.693	0.798	0.753	0.684	23.95	23.29	22.88	22.58	22.47
$\hat{R}_d^{ebp}$	0.984	1.111	1.019	1.012	1.059	28.44	28.39	28.33	28.21	28.05
$\hat{R}_d^{in.ebp}$	0.955	1.068	0.965	0.959	0.999	28.45	28.39	28.33	28.21	28.04
$\hat{R}_d^{in}$	1.487	1.713	1.681	1.712	1.788	29.08	28.85	28.69	28.47	28.24

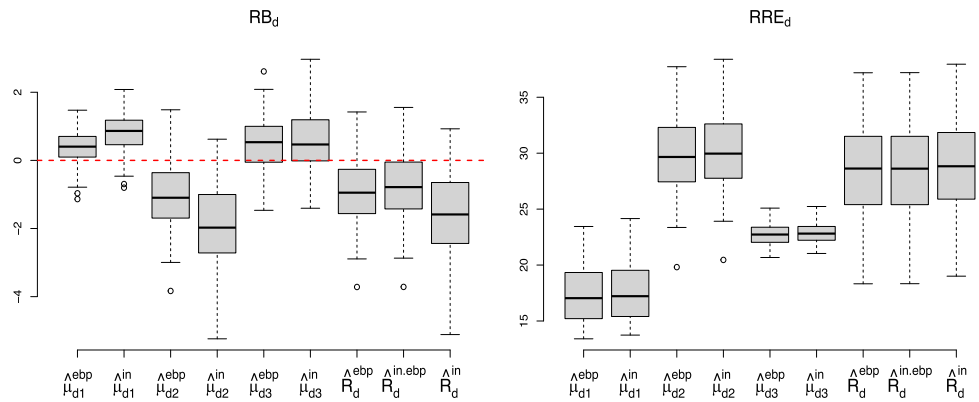

Figure 2. Boxplots of RB_d (left) and RRE_d (right) for $D = 100$.

predictions of the random effects, which is clearly unfavourable in SAE. Thus, for all these reasons we recommend the Laplace approximation as the most suitable fitting algorithm for the ADMM.

5.2 Simulation 2

The target of Simulation 2 is to calculate biases and root-MSEs of predictors of proportions of employed, unemployed and inactive people, and of unemployment rates. Simulation 2 empirically calculates the average across domains of root-MSEs *RE* and of absolute biases *AB* of EBPs and plug-in predictors. It also calculates the empirical relative root-MSEs *RRE* and absolute biases *RAB*. Online supplementary material, Appendix B.2 describes the Monte Carlo simulation steps. The number of iterations of Simulation 2 is $I = 1{,}000$. Online supplementary material, Table B.5 of Appendix B.2 presents the absolute performance measures *RE* and *AB*.

Table 6 presents the empirical relative performance measures *RRE* and *RAB*. Figure 2 presents boxplots of relative biases (left) and relative root-MSEs (right) of EBPs and plug-in predictors, RB_d, RRE_d, $d = 1, \ldots, D$. Online supplementary material, Appendix B.2 contains online supplementary material, Figures B.2 and B.3, which plot the RRE_d of R_d^{in} against the actual rate R_d and show that the former decreases as the latter increases in the interval [0.25, 0.5]. Online supplementary material, Appendix B.2.2 provides a comparative study of the performance of the plug-in predictors under the ADMM, BFH-alr and MMM models (described in online supplementary material, Appendix C).

Table 7. *RAB* (left) and *RRE* (right) of the bootstrap MSE estimators for $D = 100$

	B					***B***				
	50	**100**	**200**	**300**	**400**	**50**	**100**	**200**	**300**	**400**
$\hat{\mu}_{d1}^{in}$	10.175	9.235	7.585	10.540	8.758	25.34	21.43	19.21	18.77	17.97
$\hat{\mu}_{d2}^{in}$	10.167	9.469	8.342	10.026	10.407	27.08	23.32	20.07	20.25	20.66
$\hat{\mu}_{d3}^{in}$	10.974	10.430	9.570	12.257	9.358	31.12	26.91	24.93	23.89	23.29
$\hat{R}_{d}^{in}$	9.667	9.025	7.932	9.786	9.286	25.25	21.32	18.40	18.21	18.10

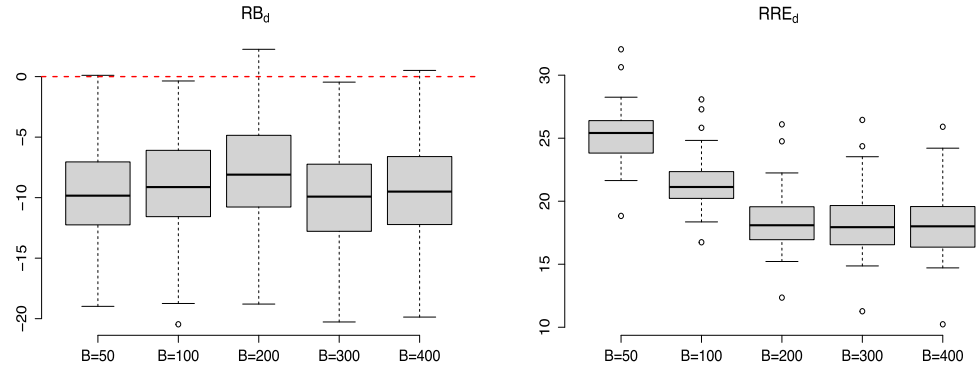

Figure 3. Boxplots of RB_d and RRE_d for $\hat{R}_d^{in}$ and $D = 100$.

Simulation 2 shows that *RAB* of $\hat{\mu}_{dk}$, $k = 1, 2, 3$, remains stable as D increases for both EBPs and plug-in predictors. On the other hand, *RRE* of $\hat{\mu}_{dk}$, $k = 1, 2, 3$, decrease very slowly, which is reasonable since as the number of domains increases, so does the number of predictors to calculate. By definition, EBPs outperform plug-in predictors in terms of bias. Regarding the relative root-MSEs, there are no differences between the proposed predictors. Analogous results are obtained for $\hat{R}^{ebp}$, $\hat{R}^{ebp.in}$ and $\hat{R}^{in}$, respectively. In view of this result and the cost of the computating EBPs, we recommend the use of the plug-in predictor.

5.3 Simulation 3

The target of Simulation 3 is to empirically calculate the biases and root-MSEs of the parametric bootstrap estimators of the MSE of $\hat{\mu}_{dk}^{in}$, $k = 1, 2, 3$, and $\hat{R}_d^{in}$. Simulation 3 empirically calculates the average across domains of root-MSEs *RE* and of absolute biases *AB* of the MSE estimators. It also calculates the empirical relative root-MSEs *RRE* and absolute biases *RAB*. Online supplementary material, Appendix B.3 defines the performance measures and describes the Monte Carlo simulation and the bootstrap steps. Simulation 3 is run with $I = 200$ Monte Carlo iterations and $D = 100$ domains.

Table 7 presents the empirical relative performance measures *RRE* and *RAB*. This table shows a decrease of *RRE* as B increases, with some stabilization around $B = 400$. Online supplementary material, Table B.8 of Appendix B.3 presents the absolute performance measures *RE* and *AB*. The mean computational time of each bootstrap iteration $b = 1, \ldots, B$ closely matches that of Simulation 1, at approximately 2.6 s per run. The total computational times are 7.2 h, 14.4 h, 28.8 h, 43.3 h and 57.7 h for $B = 50, 100, 200, 300, 400$, respectively. The hardware and software used are the same as in Simulation 1.

Figure 3 presents boxplots of relative biases (left) and relative root-MSEs (right) for the plug-in predictors of the unemployment rates, RB_d, RRE_d, $d = 1, \ldots, D$. We observe that the relative biases are negative with values around −10%, so we recommend some caution when using parametric bootstrap MSE estimators. Nonetheless, a relative bias of this order is not large enough to raise major concerns

Table 8. Empirical coverage probabilities of the 95% prediction intervals for μ_{dk}, $k = 1, 2, 3$, R_d and $D = 100$

	C_k					C_k^*				
	$B = 50$	$B = 100$	$B = 200$	$B = 300$	$B = 400$	$B = 50$	$B = 100$	$B = 200$	$B = 300$	$B = 400$
μ_{d1}	0.952	0.952	0.948	0.949	0.951	0.931	0.929	0.933	0.931	0.933
μ_{d2}	0.949	0.950	0.946	0.946	0.949	0.928	0.928	0.928	0.925	0.928
μ_{d3}	0.949	0.952	0.950	0.952	0.952	0.925	0.928	0.931	0.935	0.926
R_d	0.951	0.949	0.948	0.948	0.949	0.931	0.930	0.931	0.930	0.932

or overturn the conclusions of the analysis. The relative root-MSEs decrease with increasing B, stabilizing at $B = 400$, so we recommend applying that number of bootstrap iterations. If we averaged the relative biases without taking absolute values, we would obtain values practically identical to those in Table 7 but with a negative sign.

In the SAE literature we can find papers that present simulations showing that parametric bootstrap estimators can lead to underestimations of MSEs. To mitigate this issue, Hall and Maiti (2006) and Erciulescu and Fuller (2013, 2016) proposed double bootstrap approaches that correct the leading-order bias of bootstrap MSE estimators. However, these solutions can be methodologically complex and computationally demanding, and thus this objective would lie beyond the scope of the present work.

Table 8 presents the empirical coverage probabilities of the 95% prediction intervals for the domain category proportions μ_{dk}, $k = 1, 2, 3$ and rates R_d. We consider two prediction intervals. The first, denoted as C_{dk}, uses as standard error the empirical Monte Carlo MSE from Simulation 2, i.e. $MSE_{dk} = RE_d^2(\hat{\mu}_k^{in})$, $k = 1, 2, 3$ and $MSE_{d0} = RE_d^2(R_d)$. The second, C_{dk}^*, replaces this standard error with the parametric bootstrap estimators $mse_{dk}^* = mse^*(\hat{\mu}_{dk})$, $k = 1, 2, 3$ and $mse_{d0}^* = mse^*(\hat{R}_d)$. Table 8 shows the means coverage probabilities across domains, i.e.

$$C_k = \frac{1}{D}\sum_{d=1}^{D} C_{dk}, \quad C_k^* = \frac{1}{D}\sum_{d=1}^{D} C_{dk}^*.$$

The C_k means remain very close to 0.95, whereas the bootstrap-based C_k^* means show a mild undercoverage (0.925-0.933). This behaviour is consistent with the negative bias observed for the bootstrap MSE estimators. However, the gap between C_k and C_k^* is fairly stable over B.

6 Application to SLFS data

This section illustrates the use of the new SAE methodology with an application of the area-level Dirichlet mixed model to SLFS data from the last quarter of 2022. The target variables are the direct estimators of the proportions of employed, unemployed and inactive people, defined in (2.3) and calculated with SLFS2022.04 data.

The models based on multinomial or Dirichlet distributions are not invariant with respect to the choice of the third (last) category. The proportions of the two first categories are those that the models explain. Therefore, it is reasonable to choose the least interesting category as the last category, since its proportion prediction is obtained by subtracting the predictions of the other categories from one, and errors can accumulate. In the case of the SLFS data, we are interested in estimating unemployment rates. Therefore, we chose the inactive category as the last category.

Due to the difficulty in obtaining auxiliary variables aggregated at the province-sex-age level from freely accessible statistical sources, we use the 19 SLFS microdata files SLFS2018.01-SLFS2022.03 (downloadable from the INE website) to build the working dataset. The auxiliary variables are direct estimators of class proportions of unit-level classification variables, calculated by applying the

Table 9. Quantiles of standard deviations of the direct estimators of the auxiliary and target variables (right) and quotients $Q_{dk,r}$ defined in (6.1) (right)

	q_0	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	q_1	q_0	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	q_1
*edu*3	0.0023	0.0049	0.0068	0.0082	0.0149	9.28	18.04	21.77	26.02	59.15
*work*1	0.0023	0.0048	0.0067	0.0084	0.0147	12.81	18.89	21.28	25.25	40.48
*edu*1	0.0011	0.0024	0.0031	0.0041	0.0092	7.221	24.41	35.12	57.81	162.83
*st*3	0.0016	0.0030	0.0037	0.0048	0.0118	5.13	14.62	23.36	42.15	179.83
y_{d1}	0.0097	0.0224	0.0295	0.0425	0.0768					
y_{d2}	0.0063	0.0139	0.0183	0.0266	0.0741					

formula (2.3) on the augmented SLFS file. As a result of this process, the design-based variances of the direct estimators of the class proportions (auxiliary variables) are much smaller than the variances of the direct estimators of the labour force indicators (target variables). Thus, the explanatory variables of the dataset are estimators, subject to measurement error, of the administrative census variables to which the authors unfortunately do not have access. Extending our research to include the measurement error methods proposed by Ybarra and Lohr (2008) and Burgard et al. (2020, 2022) is beyond the scope of our current research objectives. The domain-level auxiliary variables are the proportion of people in the categories of the classification variables:

Citizenship: *nac*1 (Spanish citizenship) and *nac*2 (not Spanish citizenship).
Education level: *edu*0, (less than primary), *edu*1 (primary), *edu*2 (basic secondary education) and *edu*3 (bachelor or/and university).
Professional status: employer (with or without employees) or independent worker (*st*1), cooperative or family business (*st*2), public sector salaried employee (*st*3), private sector salaried employee (*st*4), and others (*st*5).
Working hours: unemployed or inactive (*work*0), full-time contract (*work*1) and part-time contract (*work*2)

After a selection process of auxiliary variables based on p-values, a model with all significant auxiliary variables is reached. The selected area-level Dirichlet mixed model has four significant variables, which explain the employment indicators in terms of higher levels of education ($x_{d12} = edu3$) and full-time workers ($x_{d13} = work1$) for the employed, and lower levels of education ($x_{d22} = edu1$) and public salaried employees ($x_{d23} = st3$) for the unemployed. To justify that the variability of the selected explanatory variables is substantially smaller than that of the target variables, for $k = 1, 2$ and $r = 2, 3$, Table 9 (left) reports the quantiles of the set $d = 1, \ldots, D$ of estimates of the design-based standard deviations of the direct estimators. For $k = 1, 2$ and $r = 2, 3$, Table 9 (right) presents the quantiles of the quotients

$$Q_{dkr} = \left(\frac{\widehat{\text{var}}_\pi(y_{dk})}{\widehat{\text{var}}_\pi(x_{dkr})} \right)^{1/2}, \quad d = 1, \ldots, D. \tag{6.1}$$

We observe that the quantiles of the direct estimators of the proportions of employed and unemployed people are markedly larger than those of the aggregated auxiliary variables, indicating that uncertainty is concentrated in the response variables.

Table 10 presents the estimates $\hat{\beta}_{k,r}$ of the regression parameters (top) and the estimates $\hat{\sigma}_k$ of the standard deviation parameters (bottom), $k = 1, 2$, $r = 1, 2, 3$. The estimates are calculated by the ML-Laplace approximation algorithm. This table presents the asymptotic and the basic bootstrap 95% confidence intervals (L.B, U.B) and (L.B*, U.B*) described in (3.10) and (3.11) and the logical statements of whether (T) or not (F) the zero belongs to the confidence interval. The standard errors $\hat{\sigma}^*_{\hat{\eta}}$ of the asymptotic CI (L.B, U.B) and the quantiles $x^*_{\alpha/2}$, $x^*_{1-\alpha/2}$ of formulas (3.10) and (3.11), are calculated

Table 10. Regression parameters $\hat{\beta}_{k,r}$ (top) and standard deviation parameters $\hat{\sigma}_k$ (bottom), $k = 1, 2$, $r = 1, 2, 3$, of the selected ADMM

		$\hat{\theta}$	**L.B**	**U.B**	**L.B***	**U.B***	$0 \in$ **CI**
$k = 1$	*Intercept*	−1.2564	−1.3343	−1.1788	−1.3289	−1.1748	F
	*edu*3	0.2331	0.1243	0.3420	0.1234	0.3398	F
	*work*1	3.8313	3.7713	3.8914	3.7588	3.8751	F
$k = 2$	*Intercept*	−1.4905	−1.5406	−1.4404	−1.5331	−1.4309	F
	*edu*1	5.2992	4.7658	5.8326	4.6423	5.7834	F
	*st*3	2.0998	1.8045	2.3952	1.7805	2.3682	F
$k = 1$	σ_1	0.2108	0.1962	0.2256	0.2004	0.2285	F
$k = 2$	σ_2	0.2567	0.2285	0.2849	0.2354	0.2924	F

by the parametric bootstrap procedure described in Section 4.3 with $B = 500$ iterations. None of the CIs for the regression parameters or the variance components include zero, indicating statistical significance for all the model parameters, while their narrowness reflects high accuracy in the estimation.

Concerning the sign of the regression coefficients, and assuming that the remaining ones are fixed, we observe that a higher educational level (*edu*3) and a greater proportion of full-time working hours (*work*1) are positively associated with employment, whereas lower educational attainment (*edu*1) and public sector employment (*st*3) are negatively associated with employment. The latter association may reflect counter-cyclical 'refuge' behaviour: during periods of labour-market uncertainty, an increasing number of workers apply for public-service positions. This can raise the observed public-sector share while unemployment remains high due to slow recruitment and selection processes.

6.1 Model validation

To evaluate the quality of the model, we conduct a residuals analysis. For $d = 1, \ldots, D$, $k = 1, 2, 3$, the plug-in raw residuals and standardized residuals (sresiduals) of the model (3.1) are

$$\varepsilon_{dk} = y_{dk} - E[y_{dk}|\hat{\mu}^{in}_{dk}, \hat{\phi}_d], \quad r_{dk} = \frac{y_{dk} - E[y_{dk}|\hat{\mu}^{in}_{d}, \hat{\phi}_d]}{\sqrt{\mathrm{Var}[y_{dk}|\hat{\mu}^{in}_{d}, \hat{\phi}_d]}} = \frac{y_{dk} - \hat{\mu}^{in}_{dk}}{\sqrt{\hat{\mu}^{in}_{dk}(1 - \hat{\mu}^{in}_{dk})/(\hat{\phi}_d + 1)}}. \tag{6.2}$$

The distributions of the sresiduals r_{dk} are not normal. However, by construction, they are standardized direct estimators. Since the direct estimators y_{dk} are weighted sums, the Central Limit Theorem suggests that their distribution should be close to normal for sample sizes $n_d > 30$, 50 (cf. Table 1). If, in addition, the selected area-level Dirichlet mixed model has a good fit to the aggregate data, we can expect the sresiduals to have a symmetric distribution around zero and a good diagnostic of normality. In the problem at hand, sample sizes vary between 59 and 1969, so it makes sense to analyse the symmetry and normality of the sresiduals.

Figure 4 presents the boxplots of the marginal sresiduals by labour categories (left) and by labour categories (E = employed, U = unemployed, I = inactive) crossed by sex (M = men, W = women) and age group (1 = 16–34, 2 = 35–64) (right). Within the labour categories, the sresiduals are centered around zero and lie inside the interval (−2, 2), and no remarkable deviations from normality are observed. However, when analysing the residuals across occupational category, sex, and age group, the model-based predictions of proportions tend to be lower than the direct ones for the employed and unemployed mature working-age groups (E_{W2}, U_{M2}) and for inactive young women (I_{W1}). On the other hand, the model-based predictions of proportions tend to be greater than the direct ones for young men (E_{M1}, U_{M1}, I_{M1}), unemployed young women (U_{W1}), and inactive older women (I_{W2}).

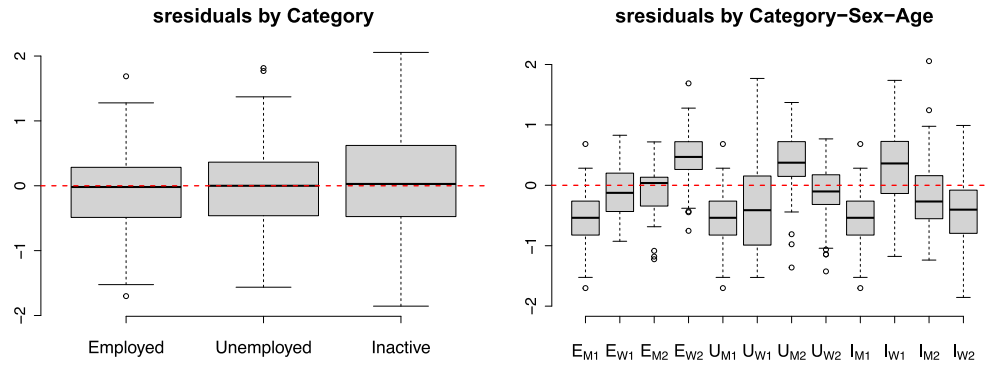

Figure 4. Boxplots of the sresiduals by labour category (left) and by labour category, sex and age group (right).

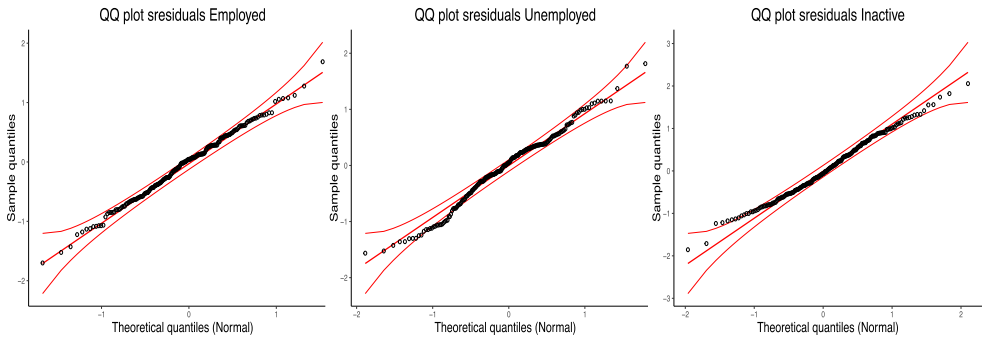

Figure 5. QQ plots of the sresiduals by labour category.

Table 11. Shapiro-Wilk and Kolmogorov-Smirnov normality tests by labour category

Test	S-Wilk		K-Smirnov	
Category	**Statistic**	***p*-value**	**Statistic**	***p*-value**
Employed	0.996	0.955	0.047	0.316
Unemployed	0.986	0.044	0.054	0.130
Inactive	0.991	0.305	0.054	0.141

Figure 5 presents the QQ plots of the sresiduals by labour categories with 95% confidence bands, and shows that most of the sresiduals are located within the bands.

Table 11 gives the Shapiro-Wilk (S-Wilk) and Kolmogorov-Smirnov (K-Smirnov) normality tests by labour categories. It contains the corresponding test-statistic values (Statistic) and the *p*-values. The combined evidence from the QQ plots and the normality tests is feasible with the hypothesis that the distribution of the residuals is close to normal.

Alternatively, online supplementary material, Appendices C.1 and C.2 present the fitting and validation of the bivariate Fay-Herriot model with additive log-ratio transformation (BFH-alr) and the area-level multinomial mixed model (MMM). The residual diagnostics and the significance of the auxiliary variables indicate that the ADMM provides the best fit to the SFLS2022.04 data.

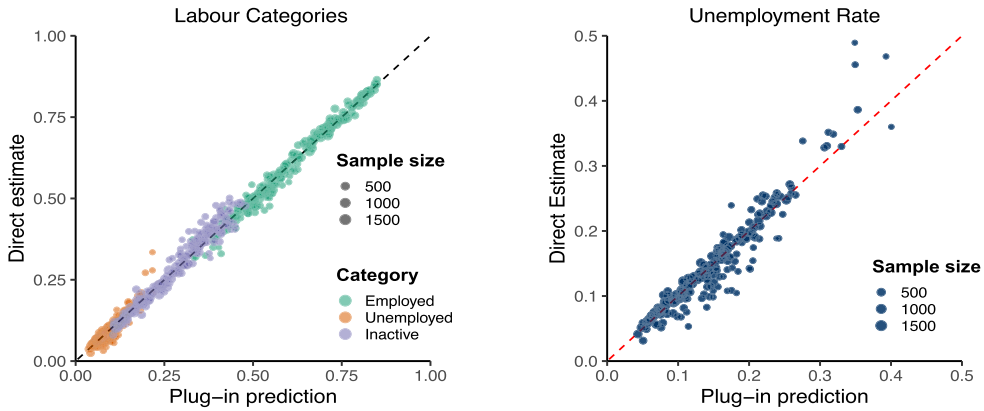

Figure 6. Plug-in predictions versus direct estimates by labour category (left) and of the unemployment rates (right).

6.2 Predictions and error measurements

This section presents the domain-level plug-in predictions for the labour-force indicators (employed, unemployed, inactive) and the unemployment rates, together with their uncertainty measures. Reliability is assessed via design-based coefficients of variation (CVs) and model-based mean squared errors (MSEs) estimated by parametric bootstrap. Plug-in predictions are compared against the direct estimators. Finally, the spatial distribution of predicted unemployment rates and CVs is displayed to aid interpretation across provinces and sex-age groups.

Figure 6 plots the plug-in predictions versus direct estimates by labour category (left) and of the unemployment rates (right) of young and mature working-age men and women. The similarities between the behaviour of the direct estimates and the plug-in predictions are evident. They align closely with the line $y = x$, indicating no appreciable systematic bias and confirming that the plug-in predictors closely reproduces the behaviour of the direct estimators. Moreover, the point cloud suggests moderate shrinkage and reduced dispersion under the model. However, if there are differences in unemployment rates between men and women, these cannot be easily identified from this figure. For this reason, to capture gender-specific dynamics in labour market outcomes at the domain level, Figure 9 provides a spatial representation of the unemployment rate predictions through domain-level maps.

Online supplementary material, Figures D.1–D.8 of Appendix D present the plug-in predictions and direct estimates separated by labour category, age group and sex, sorted by sample size. Here it can be better appreciated the differences between predictions and direct estimates, although the behaviour is similar between them.

Figures 7 and 8 plot the estimated model-based MSEs of the plug-in predictors and the estimated design-based variances of the direct estimators of proportions of employed, unemployed and inactive people and unemployment rates for young working-age men, with domains sorted by sample size. The bootstrap algorithm for calculating the estimators of the model-based MSEs runs with $B = 500$ iterations. We observe that the blue line of the plug-in predictor is below the red line of the direct estimator and that they tend to be equal as the sample size increases. Moreover, significant differences are observed across all labour categories and unemployment rates. Similar results are obtained for labour indicators of young women and mature working-age people, although the differences are smaller for the latter, which is consistent with the larger sample size in this population stratum.

Table 12 displays the mean relative MSE reduction, i.e.

$$RR = \frac{1}{D}\sum_{d=1}^{D} RR_d, \quad RR_d = 100\frac{\text{mse}_d^{*dir} - \text{mse}_d^{*in}}{\text{mse}_d^{*dir}}, \quad d = 1, \ldots, D,$$

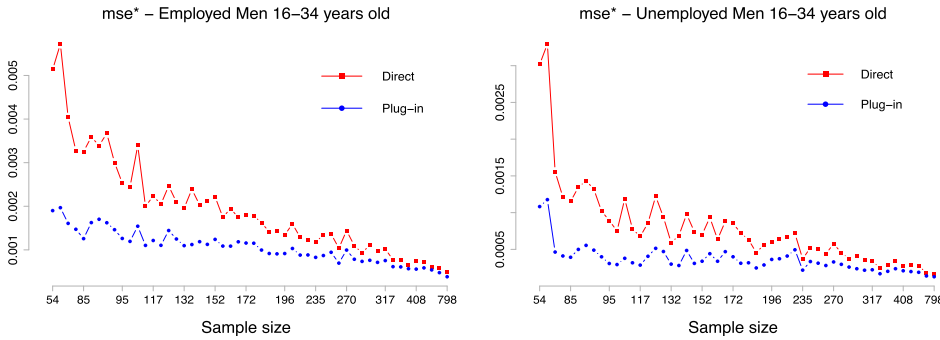

Figure 7. Bootstrap estimates of model-based MSEs of plug-in predictors and design-based estimates of variances of direct estimators of proportions of employed (left) and unemployed (right) young working-age men.

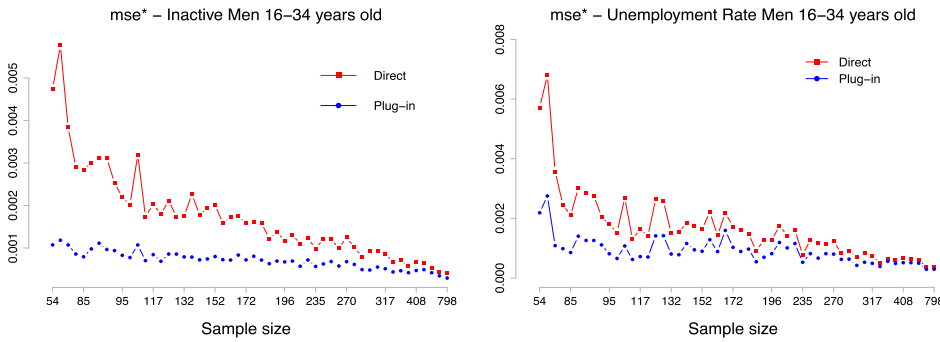

Figure 8. Bootstrap estimates of model-based MSEs of plug-in predictors and design-based estimates of variances of direct estimators of proportions of inactive people (left) and of unemployment rates (right) for young working-age men.

Table 12. Mean relative reduction of model-based MSEs between the direct estimator and the plug-in predictor by sex, age group and labour indicator.

Sex	Age Group	Employed	Unemployed	Inactive	Rate
Men	16–34	38.21%	47.54%	51.27%	42.29%
Men	35–64	24.28%	34.79%	40.16%	31.58%
Women	16–34	39.11%	44.64%	50.79%	39.32%
Women	35–64	18.28%	22.18%	30.87%	18.47%

between the direct estimator and the plug-in predictor by sex, age group and labour indicator. The bootstrap MSEs estimates of plug-in predictors (mse_d^{*in}) and of direct estimators (mse_d^{*dir}) are described in Section 4.3. We observe mean relative MSE reductions between 20% and 50%. These gains are especially remarkable for the young male and female groups, which typically have smaller sample sizes and thus benefit more from model-based borrowing of strength.

Figures 9 and 10 maps the plug-in predictions and the direct estimates of unemployment rates of young working-age men (left) and women (right) for Spanish provinces. A clear pattern of inequality in unemployment rates is evident across Spanish provinces among the youth, with those in the south,

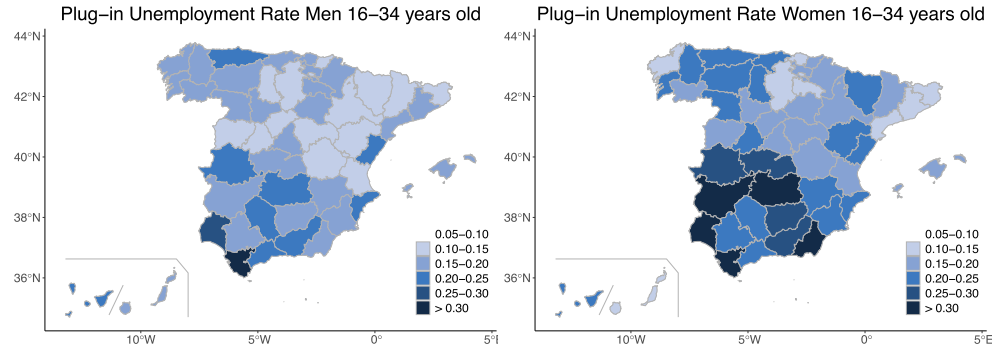

Figure 9. Plug-in predictions of unemployment rates of young working-age men (left) and women (right).

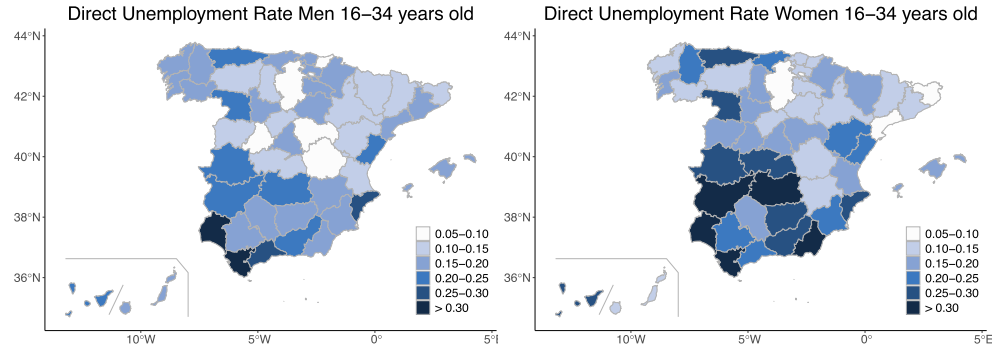

Figure 10. Direct estimates (bottom) of unemployment rates of young working-age men (left) and women (right).

southwest, and south-central regions being the most affected. We observe that, on average, women exhibit higher unemployment rates than men. Young men in the provinces of Huelva and Cádiz, exhibit unemployment rates above 25% (27% and 30%, respectively), placing them among the most affected regions. However, although scarcely discernible on the map, Ceuta records the highest unemployment rate among young men, at 35%. In comparison, as many as eleven provinces report female unemployment rates exceeding 25%: Granada (25.5%), Cáceres (26%), Toledo (26%), Jaén (26.5%), Badajoz (31%), Ciudad Real (31%), Almería (32%), Huelva (35%), Cádiz (35%), Ceuta (39%), and Melilla (40.0%). Online supplementary material, Figure D.15 of Appendix D plots the spatial distribution of the unemployment rates for mature working-age men and women. Similarly, higher unemployment rates are found in central and southern Spain, with mature working-age women being more affected again. By contrast, unemployment rates are lower among mature working-age people.

We conduct paired *t*-tests to analyse the mean difference in unemployment rates between men and women. Table 13 reports the results of the *t*-tests, confirming that unemployment rates in women are significantly higher than in men ($p < 0.001$).

We compute the bootstrap-estimated CVs of the plug-in predictor and of the direct estimator of the unemployment rate, for each province, as the bootstrap root-MSE divided by the plug-in predictor, i.e.

$$\widehat{\mathrm{CV}}_d^{in} = \frac{\left(\mathrm{mse}_d^{*in}\right)^{1/2}}{\hat{R}_d^{in}}, \quad \widehat{\mathrm{CV}}_d^{dir} = \frac{\left(\mathrm{mse}_d^{*dir}\right)^{1/2}}{\hat{R}_d^{in}}, \quad d = 1, \ldots, D. \tag{6.3}$$

Table 13. Mean difference *t*-tests of unemployment rates between men and women

Age group	*p*-value	95% CI
16–34	<0.001	(−0.057, −0.021)
35–64	<0.001	(− 0.071, −0.041)

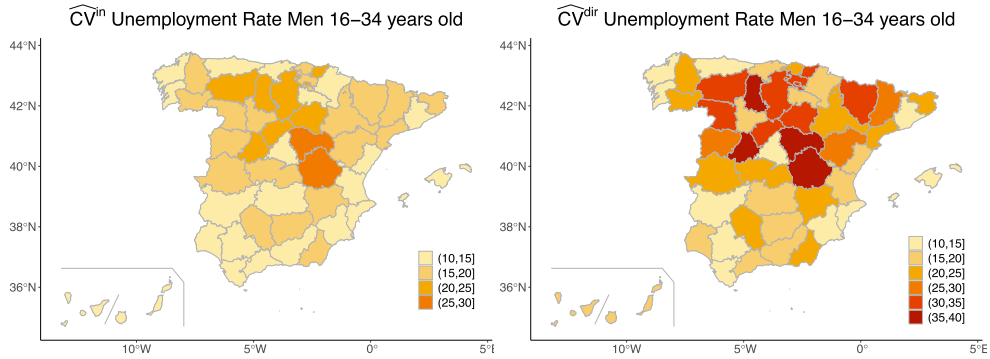

Figure 11. CVs estimates of plug-in predictors (left) and direct estimators (right) of unemployment rates of young working-age men.

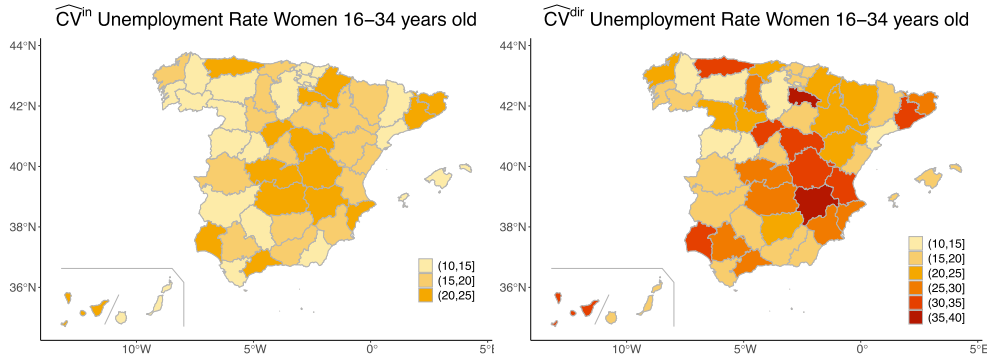

Figure 12. CVs estimates of plug-in predictors (left) and direct estimators (right) of unemployment rates of young working-age women.

Figures 11 and 12 map the CVs estimates of the plug-in predictors (left) and the direct estimators (right) of the unemployment rates for young working-age men and women in Spanish provinces. As expected, due to the imprecision of the direct estimator, $\widehat{CV}_d^{dir}$ are larger than $\widehat{CV}_d^{in}$ for both sexes. For $\widehat{CV}_d^{in}$, no province exceeds the 30% threshold; only Cuenca and Guadalajara for young men of working age fall within the 25–30% range, with values close to 25.5%. Notably, none of $\widehat{CV}_d^{in}$ exceeds 25% for the plug-in predictor. On the other hand, $\widehat{CV}_d^{dir}$ exceeds the 30% threshold in up to 13 provinces for young working-age men and up to 11 provinces for women. Online supplementary material, Appendix D provides maps of the CVs estimates of the unemployment rates for mature working-age men and women.

For selected provinces, approximately one out of every five provinces in sample size order, Table 14 presents the direct estimates, plug-in predictions and estimated CVs under the survey design ($\widehat{CV}_d^{\pi}$), of

Table 14. Direct estimates, plug-in predictions and CVs estimates of the unemployment rates of the young working-age people for one out of five Spanish provinces, sorted by sample size

	Men 16–34						Women 16–34					
Prov	n_d	$\hat{R}_d^{dir}$	$\hat{R}_d^{in}$	$\widehat{\text{CV}}_d^{\pi}$	$\widehat{\text{CV}}_d^{dir}$	$\widehat{\text{CV}}_d^{in}$	n_d	$\hat{R}_d^{dir}$	$\hat{R}_d^{in}$	$\widehat{\text{CV}}_d^{\pi}$	$\widehat{\text{CV}}_d^{dir}$	$\widehat{\text{CV}}_d^{in}$
Melilla	54	0.19	0.24	36.12	33.55	21.61	57	0.36	0.40	28.14	24.92	15.39
Palencia	85	0.11	0.15	41.51	35.04	23.35	91	0.17	0.21	33.91	27.62	19.78
Gipúzcoa	96	0.10	0.11	36.33	33.73	22.39	109	0.18	0.19	30.27	28.51	19.90
León	119	0.14	0.17	32.90	30.08	21.23	111	0.14	0.18	37.38	32.06	20.55
Salamanca	139	0.13	0.14	33.21	28.11	19.42	116	0.14	0.20	34.32	30.36	21.40
Almería	172	0.16	0.17	25.06	23.29	17.28	156	0.35	0.32	18.62	16.95	12.88
Toledo	196	0.14	0.15	23.04	22.36	17.96	192	0.27	0.26	17.32	17.97	15.50
Málaga	238	0.25	0.22	16.83	15.01	12.52	255	0.23	0.23	15.69	16.86	13.75
Badajoz	282	0.20	0.19	15.63	14.81	12.89	264	0.33	0.31	12.02	12.12	10.09
Pontevedra	384	0.18	0.17	15.51	14.32	12.50	351	0.20	0.20	16.90	15.29	13.14
Barcelona	500	0.17	0.16	13.19	12.16	10.77	486	0.13	0.14	15.50	15.72	14.09

the direct estimators ($\widehat{\text{CV}}_d^{dir}$) and under the model ($\widehat{\text{CV}}_d^{in}$) of the unemployment rates of the young working-age people. CVs estimates under the model exhibit much lower values than those of the direct estimator and under the design, especially in provinces and groups with smaller sample sizes. Online supplementary material, Appendix D contains analogous tables for the mature working-age people.

7 Conclusions

Compositional data frequently arise in the analysis of socio-economic indicators within public statistics. Compositions possess an inherent dependency structure, rendering them challenging to model with conventional statistical techniques. This study introduces a novel statistical methodology based on an area-level mixed Dirichlet model to estimate category proportions by taking into account the correlations between the components of the target vector. This approach offers two main advantages over fixed-effects Dirichlet regression: (1) capturing variability not explained by fixed effects through the random effects, and (2) the ability to deal with directly area-level proportions rather than unit-level data, thereby avoiding to aggregate unit-level predictions later. However, it is not possible to accommodate Dirichlet models for positive correlations between proportions. This is not a problem for small values of q, since positive correlations between the proportions would be extremely unlikely. Positive correlations may appear when q is large. In such cases, it is not recommended to use multivariate or Dirichlet models, and multivariate Fay-Herriot models applied to log-ratio transformations of the proportions should be used. This last approach does allow for modelling positive correlations. See Esteban et al. (2020).

The new ADMM is a multivariate GLMM, therefore the likelihood function involves a multiple integral, impeding direct maximization. Due to the lack of available software for computing maximum likelihood estimators for the model parameters, we have developed and implemented a maximum likelihood Laplace approximation algorithm to estimate the model parameters and derive model-based predictors. As a complementary approach, we also implement a Gauss-Hermite quadrature method in R. The gaussian quadrature is derivative-free and requires relatively little additional coding. The details for its implementation are provided. The Laplace approximation presents clear advantages over Gauss-Hermite quadrature. In simulation experiments, the ML-Laplace approximation is found to substantially outperforms Gaussian quadrature in terms of computational fitting times, and to be consistently more accurate. Finally, a practical limitation is that the Gaussian quadrature algorithm does not return predictions of the random effects, which is a clear drawback for small area prediction.

Although it is theoretically possible to approximate the EBPs of the random effects using a Monte Carlo simulation, this approach involves a level of computation that is comparable to directly

estimating the EBP of the target proportions. In contrast, the proposed ML-Laplace algorithm provides the calculation of the likelihood modal predictors of the random effects from the optimization process.

For these reasons, we adopt the ML-Laplace approximation in the real-data application.

Three main simulation experiments are carried out. First experiment studies and compares the behaviour of the fitting algorithms and the consistency of the model parameters estimates. Second experiment analyses the bias and the root-MSEs of the EBPs and plug-in predictors, showing that in terms of root-MSEs there are no differences between both predictors; therefore, we recommend the use of the plug-in predictor due to its simplicity and computational efficiency. Third experiment examines the performance of the bootstrap estimators of the MSE of the plug-in predictors, concluding that using at least $B = 400$ iterations achieves a good compromise between accuracy and computational time. Two complementary simulation experiments are provided in online supplementary material, Appendices B.4 and B.5. The first one is a model-based, data-driven simulation where datasets are simulated from the model fitted to real data and the model parameter values are varied to analyse their effect on the accuracy of the predictors. The second one is a design-based simulation that analyses the design-based properties of the predictors based on the ADMM.

The new methodology is applied to the Spanish Labour Force Survey for the last quarter of 2022 to estimate the proportions of employed, unemployed and inactive men and women in the Spanish provinces, as well as the unemployment rates. The model produces smoother predictions and yields lower root-MSEs than the direct estimator. As a result of the model analysis, we conclude that the highest unemployment rates are located in the central and southern zones of the peninsula. For both sexes and age groups, all Spanish provinces exhibit CVs below 30%, which is a highly satisfactory outcome in SAE problems.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the Generalitat Valenciana (Conselleria de Educación, Cultura, Universidades y Empleo) [CIPROM/2024/34] and by the Agencia Estatal de Investigación [PID2022-136878NB-I00].

Data availability

The R-developed codes are available on the GitHub repository: https://github.com/small-area-estimation/AreaLevel-Dirichlet-Mixed-Models.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Berg E. (2023). Empirical best prediction of small area means based on a unit-level Gamma-Poisson model. *Journal of Survey Statistics and Methodology*, *11*(4), 873–894. https://doi.org/10.1093/jssam/smac026

Berg E., & Fuller W. A. (2014). Estimators of error covariance matrices for small area prediction. *Computational Statistics & Data Analysis*, *56*, 2949–2962. https://doi.org/10.1016/j.csda.2012.02.030

Boubeta M., Lombardía M. J., & Morales D. (2016). Empirical best prediction under area-level Poisson mixed models. *Test*, *25*, 548–569. https://doi.org/10.1007/s11749-015-0469-8

Boubeta M., Lombardía M. J., & Morales D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics & Data Analysis*, *107*, 32–47. https://doi.org/10.1016/j.csda.2016.10.014

Bugallo M., Esteban M. D., Hobza T., Morales D., & Pérez A. (2024). Small area estimation of labour force indicators under unit-level multinomial mixed models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *188*(1), 241–270. https://doi.org/10.1093/jrsssa/qnae033

Burgard J. P., Esteban M. D., Morales D., & Pérez A. (2020). A Fay-Herriot model when auxiliary variables are measured with error. *Test*, *29*(1), 166–195. https://doi.org/10.1007/s11749-019-00649-3

Burgard J. P., Esteban M. D., Morales D., & Pérez A. (2021). Small area estimation under a measurement error bivariate Fay-Herriot model. *Statistical Methods & Applications*, *30*(1), 79–108. https://doi.org/10.1007/s10260-020-00515-9

Burgard J. P., Morales D., & Wölwer A. L. (2022). Small area estimation of socioeconomic indicators for sampled and unsampled domains. *Advances in Statistical Analysis: AStA: A Journal of the German Statistical Society*, *106*, 287–314. https://doi.org/10.1007/s10182-021-00426-4

Cabello E., Morales D., & Pérez A. (2024). Area-level model-based small area estimation of divergence indexes in the Spanish labour force survey. *Journal of Survey Statistics and Methodology*, *12*(5), 1531–1566. https://doi.org/10.1093/jssam/smae023

Cabello E., Morales D., & Pérez A. (2025). Predicting gender employment discrepancies: A multivariate Fay-Herriot model for transformed proportions. *The Annals of Applied Statistics*, *19*(2), 1753–1777. https://doi.org/10.1214/25-AOAS2020

Chambers R., Chandra H., Salvati N., & Tzavidis N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *76*(1), 47–69. https://doi.org/10.1111/rssb.12019

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. *Journal of the Royal Statistical Society, Series A*, *179*(2), 453–479. https://doi.org/10.1111/rssa.12123

Chandra H., Salvati N., & Chambers R. (2017). Small area prediction of counts under a non-stationary spatial model. *Spatial Statistics*, *20*, 30–56. https://doi.org/10.1016/j.spasta.2017.01.004

Chen S., & Lahiri P. (2012). Inferences on small area proportions. *Journal of the Indian Society of Agricultural Statistics: Indian Society of Agricultural Statistics*, *66*(1), 121. https://pmc.ncbi.nlm.nih.gov/articles/PMC3896051/

Chen X., & Nandram B. (2022). Bayesian inference for multinomial data from small areas incorporating uncertainty about order restriction. *Survey Methodology*, *48*(1), 145–177. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022001/article/00004-eng.pdf

De Nicolò S., Fabrizi E., & Gardini A. (2024). Extended beta models for poverty mapping. An application integrating survey and remote sensing data in Bangladesh. *The Annals of Applied Statistics*, *18*(4), 3229–3252. https://doi.org/10.1214/24-AOAS1934

Diz-Rosales N., Lombardía M. J., & Morales D. (2024). Poverty mapping under area-level random regression coefficient Poisson models. *Journal of Survey Statistics and Methodology*, *12*(2), 404–434. https://doi.org/10.1093/jssam/smad036

Dreassi E., Petrucci A., & Rocco E. (2014). Small area estimation for semicontinuous skewed spatial data: An application to the grape wine production in Tuscany. *Biometrical Journal*, *56*(1), 141–156. https://doi.org/10.1002/bimj.201200271

Erciulescu A. L., & Fuller W. A. (2013). Small area prediction of the mean of a binomial random variable. In *Joint Statistical Meeting 2013 Proceedings - Survey Research Methods Section, Session 37 Small-Area Estimation: Theory and Applications* (pp. 855–863). American Statistical Association.

Erciulescu A. L., & Fuller W. A. (2016). Small area prediction under alternative model specifications. *Statistics in Transition new Series*, *17*(1), 9–24. https://doi.org/10.59170/stattrans

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*, *29*(3), 793–818. https://doi.org/10.1007/s11749-019-00688-w

Esteban M. D., Lombardía M. J., Lṕez-Vizcaíno E., Morales D., & Pérez A. (2023). Small area estimation of average compositions under multivariate nested error regression models. *Test*, *32*(2), 651–676. https://doi.org/10.1007/s11749-023-00847-0

Esteban M. D., Morales D., Pérez A., & Santamaría L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics & Data Analysis*, *56*, 2840–2855. https://doi.org/10.1016/j.csda.2011.10.015

Fabrizi E., Ferrante M. R., & Trivisano C. (2016). Bayesian beta regression models for the estimation of poverty and inequality parameters in small areas. In M. Pratesi (Ed.), *Analysis of Poverty Data by Small Area Estimation* (pp. 299–314). John Wiley & Sons. https://doi.org/10.1002/9781118814963

Farrell P. J. (2000). Bayesian inference for small area proportions. *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002)*, *62*(3), 402–416.

Ferrante M. R., & Trivisano C. (2010). Small area estimation of the number of firms' recruits by using multivariate models for count data. *Survey Methodology*, *36*(2), 171–180. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010002/article/11379-eng.pdf?st=1A6HDS46

Ferrari S., & Cribari-Neto F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

González-Manteiga W., Lombardía M. J., Molina I., Morales D., & Santamaría L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics & Data Analysis*, *52*, 5242–5252. https://doi.org/10.1016/j.csda.2008.04.031

González-Manteiga W., Lombardía M. J., Molina I., Morales D., & Santamaría L. (2010). Small area estimation under Fay-Herriot models with nonparametric estimation of heteroscedasticity. *Statistical Modelling*, *10*(2), 215–239. https://doi.org/10.1177/1471082X0801000206

Guadarrama M., Morales D., & Molina I. (2021). Time stable empirical best predictors under a unit-level model. *Computational Statistics & Data Analysis*, *160*, Article 107226. https://doi.org/10.1016/j.csda.2021.107226

Hájek J. (1971). Comment on "An essay on the logical foundations of survey sampling, Part One." In V. P. Godambe & D. A. Sprott (Eds.), *The foundations of survey sampling* (p. 236). Holt, Rinehart and Winston.

Hall P., & Maiti T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *68*, 221–238. https://doi.org/10.1111/j.1467-9868.2006.00541.x

Herrador M., Morales D., Esteban M. D., Sánchez A., Santamaría L., Marhuenda Y., & Pérez A. (2008). Sampling design variance estimation of small area estimators in the Spanish labour force survey. *SORT (Barcelona)*, *32*(2), 177–198. https://upcommons.upc.edu/entities/publication/5138ae96-1b70-434c-b903-b2de9be77a23

Hijazi R. H., & Jernigan R. W. (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability Statistics*, *4*(1), 77–91. https://www.researchgate.net/publication/228576713_Modelling_compositional_data_using_Dirichlet_regression_models

Hobza T., & Morales D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, *32*(3), 661–692. https://doi.org/10.1515/jos-2016-0034

Hobza T., Morales D., & Santamaría L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *Test*, *27*, 270–294. https://doi.org/10.1007/s11749-017-0545-3

Horvitz D. G., & Thompson D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

Krause J., Burgard J. P., & Morales D. (2022a). L2-penalized approximate likelihood inference in logit mixed models for regional prevalence estimation under covariate rank-defficiency. *Metrika*, *85*, 459–489. https://doi.org/10.1007/s00184-021-00837-y

Krause J., Burgard J. P., & Morales D. (2022b). Robust prediction of domain compositions from uncertain data using isometric logratio transformations in a penalized multivariate Fay-Herriot model. *Statistica Neerlandica*, *76*(1), 65–96. https://doi.org/10.1111/stan.12253

Larsen M. D. (2003). Estimation of small-area proportions using covariates and survey data. *Journal of Statistical Planning and Inference*, *112*(1–2), 89–98. https://doi.org/10.1016/S0378-3758(02)00325-7

Liu B., & Lahiri P. (2017). Adaptive hierarchical Bayes estimation of small area proportions. *Calcutta Statistical Association Bulletin*, *69*(2), 150–164. https://doi.org/10.1177/0008068317722293

López-Vizcaíno E., Lombardía M. J., & Morales D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, *13*(2), 153–178. https://doi.org/10.1177/1471082X13478873

López-Vizcaíno E., Lombardía M. J., & Morales D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society, Series A*, *178*(3), 535–565. https://doi.org/10.1111/rssa.12085

Maples J. (2019). Small area estimates of the child population and poverty in school districts using Dirichlet multinomial models. *Proceedings, Survey Research Methods Section, American Statistical Association Alexandria*, *VA*, 2448–2456. http://www.asasrms.org/Proceedings/y2019/files/1199633.pdf

Marhuenda Y., Molina I., & Morales D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics & Data Analysis*, *58*, 308–325. https://doi.org/10.1016/j.csda.2012.09.002

Marhuenda Y., Molina I., Morales D., & Rao J. N. K. (2017). Poverty mapping in small areas under a two-fold nested error regression model. *Journal of the Royal Statistical Society, Series A*, *180*(4), 1111–1136. https://doi.org/10.1111/rssa.12306

Marhuenda Y., Morales D., & Pardo M. C. (2014). Information criteria for Fay-Herriot model selection. *Computational Statistics & Data Analysis*, *70*, 268–280. https://doi.org/10.1016/j.csda.2013.09.016

Marino M. F., Ranalli M. G., Salvati N., & Alfò M. (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, *13*(2), 1166–1197. https://doi.org/10.1214/18-AOAS1226

Martínez-Minaya J., Lindgren F., López-Quílez A., Simpson D., & Conesa D. (2023). The integrated nested laplace approximation for fitting models with multivariate response. *Journal of Statistical Software*, *32*(3), 805–823. https://doi.org/10.1080/10618600.2022.2144330

Molina I., & Rao J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369–385. https://doi.org/10.1002/cjs.10051

Molina I., Saei A., & Lombardía M. J. (2007). Small area estimates of labour force participation under multinomial logit mixed model. *The Journal of the Royal Statistical Society, Series A*, *170*, 975–1000. https://doi.org/10.1111/j.1467-985X.2007.00493.x

Morales D., Esteban M. D., Pérez A., & Hobza T. (2021). *A course on small area estimation and mixed models*. Springer.

Morales D., Krause J., & Burgard J. P. (2022). On the use of aggregate survey data for estimating regional major depressive disorder prevalence. *Psychometrika*, *87*(1), 344–368. https://doi.org/10.1007/s11336-021-09808-8

Morales D., 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. https://hdl.handle.net/2117/88517

Peña D., & Yohai V. (2006). A Dirichlet random coefficient regression model for quality indicators. *Journal of Statistical Planning and Inference*, *136*(3), 942–961. https://doi.org/10.1016/j.jspi.2004.07.012

Pratesi M. (2016). *Analysis of poverty data by small area estimation*, Wiley series in survey methodology. John Wiley.

Ranalli M. G., Montanari G. E., & Vicarelli C. (2018). Estimation of small area counts with the benchmarking property. *Metron*, *76*, 349–378. https://doi.org/10.1007/s40300-018-0146-2

Rao J. N. K., & Molina I. (2015). *Small area estimation* (2nd ed.). Wiley.

Saei A., & Taylor A. (2012). Labour force status estimates under a bivariate random components model. *Journal of the Indian Society of Agricultural Statistics: Indian Society of Agricultural Statistics*, *66*, 187–201.

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

Slud E., Franco C., & Hall A. (2024). Small area estimates for voting rights act section 203 (b) coverage determinations. *Calcutta Statistical Association Bulletin*, *76*(1), 137–159. https://doi.org/10.1177/00080683231215985

Souza D. F., & Moura F. A. (2016). Multivariate beta regression with application in small area estimation. *Journal of Official Statistics*, *32*(3), 747–768. https://doi.org/10.1515/jos-2016-0038

Sun H., Berg E., & Zhu Z. (2024). Multivariate small-area estimation for mixed-type response variables with item nonresponse. *Journal of Survey Statistics and Methodology*, *12*(2), 320–342. https://doi.org/10.1093/jssam/smad018

Tang Z. Z., & Chen G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, *20*(4), 698–713. https://doi.org/10.1093/biostatistics/kxy025

Tsagris M., & Stewart C. (2018). A Dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, *39*, 398–412. https://doi.org/10.1134/S1995080218030198

Tzavidis N., Ranalli M. G., Salvati N., Dreassi E., & Chambers R. (2015). Robust small area prediction for counts. *Statistical Methods in Medical Research*, *24*(3), 373–395. https://doi.org/10.1177/0962280214520731

Wang X., Philip V. M., Ananda G., White C. C., Malhotra A., Michalski P. J., & Carter G. W. (2018). A Bayesian framework for generalized linear mixed modeling identifies new candidate loci for late-onset Alzheimer's disease. *Genetics*, *209*(1), 51–64. https://doi.org/10.1534/genetics.117.300673

Ybarra L. M., & Lohr S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, *95*(4), 919–931. https://doi.org/10.1093/biomet/asn048

Zhang L., & Chambers R. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, *66*(2), 479–496. https://doi.org/10.1111/j.1369-7412.2004.05266.x