Received: 29 August 2020 | Accepted: 9 September 2022

DOI: 10.1111/rssa.12965

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

ORIGINAL ARTICLE

Multivariate mixture model for small area estimation of poverty indicators

Agne Bikauskaite[1] | **Isabel Molina**[2] | **Domingo Morales**[3]

[1]Business Statistics Division, Eurostat, Luxembourg, Germany

[2]Department of Statistics, University of Carlos III de Madrid, Madrid, Spain

[3]Operations Research Center, University of Miguel Hernández de Elche, Elche, Spain

Correspondence
Domingo Morales, Operations Research Center, University of Miguel Hernández de Elche, Elche, Spain.
Email: domingo.morales@umh.es

Funding information
Generalitat Valenciana, Grant/Award Number: prometeo/2021/063; Ministerio de Economía y Competitividad, Grant/Award Numbers: PGC2018-096840-B-I00, PID2020-115598RB-I00

Abstract

When disaggregation of national estimates in several domains or areas is required, direct survey estimators, which use only the domain-specific survey data, are usually design-unbiased even under complex survey designs (at least approximately) and require no model assumptions. Nevertheless, they are appropriate only for domains or areas with sufficiently large sample size. For example, when estimating poverty in a domain with a small sample size (small area), the volatility of a direct estimator might make that area seems like very poor in one period and very rich in the next one. Small area (or indirect) estimators have been developed in order to avoid such undesired instability. Small area estimators borrow strength from the other areas so as to improve the precision and therefore obtain much more stable estimators. However, the usual model-based assumptions, which include some kind of area homogeneity, may not hold in real applications. A more flexible model based on multivariate mixtures of normal distributions that generalises the usual nested error linear regression model is proposed for estimation of general parameters in small areas. This flexibility makes the model adaptable to more general situations, where there may be areas with a different behaviour from the other ones, making the model less restrictive (hence, more close to nonparametric) and more robust to outlying areas. An expectation-maximisation (E-M) method is designed for fitting the proposed mixture model. Under the proposed mixture model, two different new predictors of general small area indicators are proposed. A parametric bootstrap method is used to estimate the mean squared errors of the proposed predictors. Small sample properties of the new predictors and of the bootstrap procedure are analysed by simulation studies and the new methodology is illustrated with an application to poverty mapping in Palestine.

KEYWORDS

empirical best estimator, expectation-maximisation algorithm, nested error model, normal mixture model, parametric bootstrap

1 | INTRODUCTION

There is no doubt of the importance of monitoring poverty and inequality. Not in vain the eradication of poverty is the first of the Sustainable Development Goals of the United Nations (UN), preceded by the Millennium Development Goals, where poverty was also in the first place. The UN are fostering and encouraging countries National Statistical Institutes (NSIs) to produce poverty and inequality maps and, more generally, to develop procedures for transparently monitoring the achievement of each goal.

Moreover, demand of disaggregated figures is also continuously increasing. Effective regional plans or programmes aimed to enhance cohesion of the territories by reducing the most extreme differences, as well as allocation of regional development funds, require from reliable statistical information at local level. Regional poverty maps are currently produced for this aim by the World Bank (WB) for many countries. However, the official surveys conducted by the NSIs to monitor poverty and inequality do not cover adequately all the local areas or population subgroups of interest in terms of sample size. Hence, direct survey estimates, obtained using only the observations collected from the corresponding area or population subgroup, are not reliable when sample sizes are insufficient. Small area estimation (SAE) techniques have been developed to enhance these estimates by increasing the effective sample size through the use of models that link all the areas, see Rao and Molina (2015) or Morales et al. (2021) for a comprehensive introduction to SAE.

Models are broadly classified into two types, although it is possible to specify models in between these groups. In the first group, we find the area-level models, which only require aggregated area-level data. The second group is composed by unit-level models, which specify a relation (model) between the variable of interest and the auxiliary variables at the unit level. As already said, the model can be also specified for subareas or subdomains within the areas of interest, making use of aggregated auxiliary information at that level. Here we focus on unit-level models. Among those, the nested error linear regression model with random area effects was the first unit-level model considered for SAE to estimate the county crop production of corn and soybeans using satellite data (Battese et al., 1988). This model has been widely used to obtain estimates of small area means in many types of applications.

In poverty estimation, most often a model is considered for income or expenditure or, more exactly, for a transformation of them. This means that, even if the target indicators were the area means of income/expenditure, we will need to express them in terms of the response variable in the model, which would be the transformed income/expenditure. This will lead to indicators defined as non-linear functions of the response variables for the individuals in the area, which we call non-linear area parameters. For the estimation of poverty and inequality indicators, the Word Bank traditionally used the ELL method due to Elbers et al. (2003), which was based on a nested error model with random cluster (or primary sampling units) effects. However, the appearance of the empirical best (EB) method of Molina and Rao (2010) has led to an update of the WB methods in accordance with the EB procedure (Corral, Molina and Nguyen, 2020). Further, the new approach has been extended to nested populations by Marhuenda et al. (2017), giving explicit estimators under lognormal models by Molina and Martín (2018), to complex sampling designs by Guadarrama et al. (2018), to cut-off sampling by Guadarrama et al. (2020) and to data-driven transformations by Rojas-Perilla et al. (2020). It has been applied to data from Uruguay by Molina (2019) and to Palestinian data by Molina Peralta and García-Portugués (2020).

In many applications, area heterogeneity might make the nested error model assumptions fail for certain areas. For example, bordering regions might behave more similar to the countries with which they share the border than to the remaining regions of their own country. In those cases, a more flexible model might be needed to represent this type of heterogeneity. For these situations, a number of alternative robust formulations to account for heterogeneity of the domain structure have been proposed in the literature. For example, area-level models that employ various types of mixtures of normals include Datta and Ghosh (1991), Bell and Huang (2006), Ghosh et al. (2018), Tang et al. (2018), Chakraborty et al. (2016) or Gershunskaya and Savitsky (2020). These papers introduce predictors that are robust against outliers in either the model errors or the survey errors, most of them under the Bayesian framework.

As a complement of the above cited contributions, this paper proposes a multivariate mixture (MM) model that generalises the nested error model, allowing for a latent clustering structure of the areas. Our model accounts for area heterogeneity in all the model parameters, representing a more flexible model than those with constant regression coefficients for all the areas or models with either the random effects variance or the error variance constant. Other types of mixture models considered in the literature are those used to model unit-level outliers, in which a two component mixture is specified for the individual errors (Gershunskaya, 2010). In this model, heterogeneity is introduced in the individual error variances. Elbers and Van der Weide (2014) consider a model that includes a mixture for the individual errors and another for the random effects. Our MM model is designed for representing heterogeneity across areas (e.g. due to outlying areas), but this heterogeneity might be present in the regression coefficients and/or one of the variance components, and it is identifiable just by restricting the space of the mixture probabilities.

We present the traditional nested error model in Section 3, describing the EB methodology for prediction of general indicators in small areas. Then, we propose the MM extension in Section 4. Based on that model, we introduce two new types of predictors. We give analytical expressions of these predictors for common area parameters of interest and a Monte Carlo simulation procedure for the case of more complex indicators without closed-form expressions. In Section 5, we give a parametric bootstrap procedure for estimation of the mean squared error of the proposed small area estimators. Section 6 analyses the small sample properties of the predictors compared to those based on the basic nested error model and the direct estimators. We illustrate the procedures with an application to the prediction of mean expenditure and poverty indicators in Palestinian localities. Some conclusions are drawn in Section 8. Finally, Appendix A describes the E-M fitting method for the proposed MM model and Appendix B lists the resulting parameter estimates from the considered models. This paper has supporting information with additional results of Sections 6 and 7.

2 | ESTIMATION OF POVERTY IN PALESTINIAN LOCALITIES: DATA DESCRIPTION

We illustrate the methodology proposed in this paper to the estimation of mean expenditure, poverty rates and gaps in Palestinian localities, which are regions nested within governorates. We use data from the 2016/2017 Palestinian Expenditure Consumption Survey (PECS) and the 2017 Census, which was used previously in Molina Peralta and García-Portugués (2020). We estimate in those $D = 162$ sampled localities in the PECS out of the 319 appearing in the census. The welfare measure w_{dj} is the monthly expenditure per adult equivalent in Israeli Shekels (ILS), which is measured in the PECS but not in the census.

The census size is $N = 4{,}266{,}953$ after cleaning missing data records, distributed as 2,395,774 for West Bank and 1,871,179 for Gaza. The total survey sample size, after appropriate cleaning, was $n = 18{,}363$, with 13,216 observations for West Bank and 5147 for Gaza. When disaggregating by localities, the minimum sample size is 30, and first quartile is 54, the median is 73.5, the third quartile is 124.8 and the maximum is 869. As we can see, the sample size is small for some of the localities. Note that for estimation of poverty rates and gaps, a larger sample size is required to have acceptable accuracy than when estimating the average of a continuous variable. The poverty line, measured in terms of expenditure, is $z = 10{,}027$ ILS. Based on this poverty line, approximately 26% of the Palestinian population is below the line.

The PECS sampling design is a stratified cluster systematic random sampling with two stages. First stage selects 192 enumeration areas (buildings and housing units with an average of about 120 households) from the whole population with a systematic random sampling, which implies a proportional allocation. Second stage selects 24 households from each selected enumeration area, using systematic random sampling. The strata are defined by governorate and locality type (urban, rural, refugee camp). Within each stratum, the sampling weights are later adjusted to compensate for non-response of households. This is done by multiplying the primary weights by the number of households in the stratum sample and by dividing them by the number of households in the whole stratum. Direct estimators of mean expenditure, poverty rates and gaps can be unweighted (sample means) or weighted (Hájeck estimators) using the survey weights. Because of the proportional allocation in the first stage and the non-response adjustment, the unweighted direct estimators turn out to be really very similar to the weighted ones for the three domain indicators of interest, which suggests no evidence of informativeness of the sampling design. Hence, we conclude that the PECS sample is essentially non informative, in the sense of Pfeffermann and Sverchkov (2007), so that we develop unit-level model-based SAE methodology under the assumption that the sampled and non-sampled parts of the target census y-vector follow the same statistical model. As response variable in the models, due to skewness of the welfares w_{dj}, we take the same transformation as in Molina Peralta and García-Portugués (2020), $y_{dj} = \log(w_{dj} + c)$, for $c = 1000$.

The census contains several auxiliary variables that are measured in a similar way in the PECS and which are potentially related with the above welfare measure. We consider in principle the same ones as in Molina Peralta and García-Portugués (2020), but in the final model for the transformed welfare y_{dj} we removed the non-significant covariates and those with a variance inflation factor (VIF) greater than 5. The selected covariates include location descriptors, household characteristics, attributes of the household head, dwelling characteristics, and types of supplies and amenities, where some of the categories of these variables that were observed only for few households have been merged in order to have a more stable model fitting. Concretely, the location descriptors are dummy indicators for the region (Gaza, West Bank) and the type of locality (rural and urban, camp). Concerning the household characteristics, we included size, proportion of females and employed ratio. The attributes of the household head are indicators of being unemployed, of ever been employed in Israel/settlement, of ever been employed for the national government, of refugee status, of having some difficulty, of never attended school and of having education level higher than secondary. Dwelling characteristics included type of dwelling (villa, other), type of tenure (rented, other) and number of rooms. Within commodities or supplies, we included the type of water and heating systems. Finally, we included indicators of owing washing machine, freezer, microwave, dishwasher, LED/LCD TV, electricity fan, air conditioning, central heating, solar boiler, phone line, home library, computer, iPad/tablet and smartphone.

3 | EB PREDICTORS BASED ON A NESTED ERROR MODEL

Let the population of interest be partitioned into D areas or domains, of sizes $N_1, \dots, N_D$. Although the methods described hereafter are general and can be applied in different fields, we will illustrate them through the application to estimation of monetary poverty indicators. Hence, we can think of the characteristic of interest for individual j within domain d as a welfare variable w_{dj} such as income or expenditure (as in our application), which may be transformed to achieve an approximately normal distribution. Hence, as response variable in the models we will consider the transformed welfare denoted as $y_{dj} = T(w_{dj})$, $d = 1, \dots, D$, $j = 1, \dots, N_d$, where $T(\cdot)$ is one-to-one. We wish to predict domain parameters of the form $\eta_d = h(y_{d1}, \dots, y_{dN_d})$, $d = 1, \dots, D$. Let $\mathbf{x}_{dj}$ be the vector of auxiliary variables for unit j within domain d. The basic nested error (NE) linear regression model considered by Molina and Rao (2010) for the prediction of η_d assumes that

$$y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad d = 1, \dots, D,\ j = 1, \dots, N_d, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. Domain effects u_d and individual errors e_{dj} are assumed to be all mutually independent, with $u_d \overset{iid}{\sim} N(0, \tau^2)$ and $e_{dj} \overset{iid}{\sim} N(0, \sigma^2)$.

In practice, we have a sample s_d of size $n_d \le N_d$ from each area d, for $d = 1, \dots, D$, where $n = \sum_{d=1}^{D} n_d$ is the total sample size. Let us define the domain vector $\mathbf{y}_d = (y_{d1}, \dots, y_{dN_d})'$ and split its elements according to whether the corresponding unit is in the sample or not, as $\mathbf{y}_d = (\mathbf{y}'_{ds}, \mathbf{y}'_{dr})'$, where the subscript s stands for sample units. The best predictor of η_d is the function of the sample data $\mathbf{y}_{ds}$ that minimises the MSE, among unbiased predictors, under the assumed model, and it is given by the conditional expectation

$$\hat{\eta}_d^B = E_{\mathbf{y}_{dr}}(\eta_d | \mathbf{y}_{ds}; \boldsymbol{\theta}), \tag{2}$$

which depends on the vector of unknown model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \tau^2, \sigma^2)'$. Under the nested error model (1), assuming that there is no sample selection bias, the conditional distribution required to calculate the above best predictor is given by

$$\mathbf{y}_{dr}|\mathbf{y}_{ds} \sim N\left(\boldsymbol{\mu}_{dr|s}(\boldsymbol{\theta}), \mathbf{V}_{dr|s}(\boldsymbol{\theta})\right), \tag{3}$$

where

$$\begin{aligned}\boldsymbol{\mu}_{dr|s}(\boldsymbol{\theta}) &= \mathbf{X}_{dr}\boldsymbol{\beta} + \tau^2 \mathbf{1}_{N_d-n_d}\mathbf{1}'_{n_d}\mathbf{V}^{-1}_{ds}(\boldsymbol{\theta})(\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta}),\\ \mathbf{V}_{dr|s}(\boldsymbol{\theta}) &= \tau^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}'_{N_d-n_d} + \sigma^2\mathbf{I}_{N_d-n_d},\end{aligned} \tag{4}$$

for $\mathbf{V}_{ds}(\boldsymbol{\theta}) = \tau^2\mathbf{1}_{n_d}\mathbf{1}'_{n_d} + \sigma^2\mathbf{I}_{n_d}$. Applying (A3) to calculate the inverse of $\mathbf{V}_{ds}(\boldsymbol{\theta})$ and after some algebra, the conditional mean for out-of-sample unit $j \in r_d$ from area d, denoted $\mu_{dj|s}(\boldsymbol{\theta})$, can be expressed as the linear regression plus the predicted area effect, that is,

$$\mu_{dj|s}(\boldsymbol{\theta}) = \mathbf{x}'_{dj}\boldsymbol{\beta} + \tilde{u}_d(\boldsymbol{\theta}), \tag{5}$$

where $\tilde{u}_d(\boldsymbol{\theta}) = \gamma_d(\bar{y}_{ds} - \bar{\mathbf{x}}'_{ds}\boldsymbol{\beta})$, for $\gamma_d = \tau^2(\tau^2 + \sigma^2/n_d)^{-1}$, and where $\bar{y}_{ds} = n_d^{-1}\sum_{j\in s_d} y_{dj}$ and $\bar{\mathbf{x}}_{ds} = n_d^{-1}\sum_{j\in s_d}\mathbf{x}_{dj}$ are the domain sample means of the transformed income and of the vector of auxiliary variables, respectively. The conditional variance for that same unit obtained from (4) is

$$\sigma^2_{dj|s}(\boldsymbol{\theta}) = \tau^2(1-\gamma_d) + \sigma^2. \tag{6}$$

If the target parameter is the sample mean of the response variable in the model for area d, that is, $\eta_d = \overline{Y}_d = N_d^{-1}\sum_{j=1}^{N_d} y_{dj}$, the best predictor given in (2) reduces to

$$\hat{\eta}_d^B = \hat{\overline{Y}}_d^B = N_d^{-1}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d}\mu_{dj|s}(\boldsymbol{\theta})\right),$$

where here r_d denotes the set of observations that are out of the sample from area d and $\mu_{dj|s}(\boldsymbol{\theta}) = \mathbf{x}'_{dj}\boldsymbol{\beta} + \tilde{u}_d(\boldsymbol{\theta})$, $j \in r_d$. In this special case, the best predictor of $\overline{Y}_d$ coincides with the best linear unbiased predictor (BLUP) of $\overline{Y}_d$ that uses the true $\boldsymbol{\theta}$.

Nonlinear indicators of interest are the family of FGT poverty indicators introduced by Foster et al. (1984) and defined in terms of the welfare variable w_{dj} and the poverty line z. For $\alpha \geq 0$, the FGT poverty indicator of order α for area d is defined as

$$F_{\alpha,d} = \frac{1}{N_d}\sum_{j=1}^{N_d} F_{\alpha,dj}, \quad F_{\alpha,dj} = \left(\frac{z - w_{dj}}{z}\right)^{\alpha} I(w_{dj} < z), \tag{7}$$

where $I(w_{dj} < z)$ is the indicator function taking value 1 if $w_{dj} < z$ and 0 otherwise. The poverty rate corresponds to $\alpha = 0$, whereas the poverty gap is obtained for $\alpha = 1$.

Since the response variable in the models is a one-to-one transformation of the welfare, $y_{dj} = T(w_{dj})$, we can express the variables $F_{\alpha,dj}$ involved in (7) in terms of the model response variables y_{dj} by using the inverse transformation $w_{dj} = T^{-1}(y_{dj})$, as

$$F_{\alpha,dj} = \left(\frac{z - T^{-1}(y_{dj})}{z}\right)^{\alpha} I(T^{-1}(y_{dj}) < z). \tag{8}$$

The best predictor of $F_{\alpha,d}$ is then given by

$$\hat{F}^{B}_{\alpha,d}(\boldsymbol{\theta}) = \frac{1}{N_d}\left(\sum_{j\in s_d} F_{\alpha,dj} + \sum_{j\in r_d} \hat{F}_{\alpha,dj}(\boldsymbol{\theta})\right), \tag{9}$$

where $\hat{F}_{\alpha,dj}(\boldsymbol{\theta}) := E_{\mathbf{y}_r}[F_{\alpha,dj}|\mathbf{y}_s;\boldsymbol{\theta}]$. For $\alpha = 0, 1$ and for certain transformations T, the expectation defining $\hat{F}_{\alpha,dj}(\boldsymbol{\theta})$ can be calculated analytically. Concretely, for the poverty incidence ($\alpha = 0$), if $T(\cdot)$ is a nondecreasing monotonous function, we obtain

$$\hat{F}_{0,dj} = E_{\mathbf{y}_r}\left[F_{0,dj}|\mathbf{y}_s;\boldsymbol{\theta}\right] = E_{\mathbf{y}_{dr}}\left[I\left(y_{dj} < T(z)\right)|\mathbf{y}_{ds};\boldsymbol{\theta}\right] = \Phi(\alpha_{dj}(\boldsymbol{\theta})),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal variable and

$$\alpha_{dj}(\boldsymbol{\theta}) = \frac{T(z) - \mu_{dj|s}(\boldsymbol{\theta})}{\sigma_{dj|s}(\boldsymbol{\theta})},$$

for $\mu_{dj|s}(\boldsymbol{\theta})$ and $\sigma^2_{dj|s}(\boldsymbol{\theta})$ given respectively in (5) and (6). For the poverty gap ($\alpha = 1$), considering the log transformation after a shift, $y_{dj} = \log(w_{dj} + c)$, with inverse $w_{dj} = T^{-1}(y_{dj}) = \exp(y_{dj}) - c$, we obtain

$$\begin{aligned}\hat{F}_{1,dj}(\boldsymbol{\theta}) &= E_{\mathbf{y}_{dr}}\left[F_{1,dj}|\mathbf{y}_{ds};\boldsymbol{\theta}\right] = E_{\mathbf{y}_{dr}}\left[\frac{z - T^{-1}(y_{dj})}{z}\, I\left(y_{dj} < T(z)\right)\Big|\mathbf{y}_{ds};\boldsymbol{\theta}\right] \\ &= \frac{z + c}{z}\,\Phi(\alpha_{dj}(\boldsymbol{\theta})) - \frac{1}{z}\exp\left(\mu_{dj|s}(\boldsymbol{\theta}) + \sigma^2_{dj|s}(\boldsymbol{\theta})/2\right)\Phi\left(\alpha_{dj}(\boldsymbol{\theta}) - \sigma_{dj|s}(\boldsymbol{\theta})\right).\end{aligned}$$

Finally, consider that we wish to estimate the average welfare of domain d under the latter transformation of w_{dj}. Then, we can express the average welfare as

$$\overline{W}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} w_{dj} = \frac{1}{N_d}\sum_{j=1}^{N_d} \exp(y_{dj}) - c,$$

and the best predictor of $\overline{W}_d$ becomes

$$\hat{\overline{W}}^{B}_d = \frac{1}{N_d}\left(\sum_{j\in s_d} w_{dj} + \sum_{j\in r_d} \hat{w}_{dj}(\boldsymbol{\theta})\right),$$

where, for out-of-sample unit $j \in r_d$, we have

$$\hat{w}_{dj}(\boldsymbol{\theta}) := E_{\mathbf{y}_{dr}}\left[w_{dj}|\mathbf{y}_s;\boldsymbol{\theta}\right] = E_{\mathbf{y}_{dr}}[\exp(y_{dj})|\mathbf{y}_{ds};\boldsymbol{\theta}] - c = \exp\left(\mu_{dj|s}(\boldsymbol{\theta}) + \sigma^2_{dj|s}(\boldsymbol{\theta})/2\right) - c.$$

The EB predictor of $\eta_d = h(\mathbf{y}_d)$ is finally obtained replacing $\boldsymbol{\theta}$ by a consistent estimator $\hat{\boldsymbol{\theta}}$ in the best predictor (2) and is denoted as $\hat{\eta}^{\mathrm{EB}}_d$. Usual estimators of $\boldsymbol{\theta}$ are obtained by maximum-likelihood (ML) or restricted/residual ML (REML) methods. For indicators $\eta_d = h(\mathbf{y}_d)$ that are more complex functions of $\mathbf{y}_d$ than the above ones, the expected value involved in the EB predictor might not have a closed form. When this is the case, that expected value can be approximated by the Monte Carlo (MC) simulation procedure described in Molina and Rao (2010).

Except for linear indicators $\eta_d = h(\mathbf{y}_d)$ (where $h(\cdot)$ is linear) or when the auxiliary variables do not vary across the units, apart from the corresponding survey sample data $(y_{dj}, \mathbf{x}_{dj}), j \in s_d$, the best predictor

$$\hat{\eta}_d^B = \frac{1}{N_d}\left(\sum_{j\in s_d} h(y_{dj}) + \sum_{j\in r_d} E_{\mathbf{y}_d}\left[h_j(y_{dj})|\mathbf{y}_{ds};\boldsymbol{\theta}\right]\right), \tag{10}$$

of a non-linear indicator η_d, requires the values of the auxiliary variables $\mathbf{x}_{dj}$ for each out-of-sample unit $j \in r_d$ (microdata) in the target area d. This is typically obtained from census or administrative records. In fact, the original EB procedure requires to identify the survey units in the census data file, which may not be possible in practice. Sometimes, the survey sample is not obtained from the considered census and, in that case, the available sample survey data $\mathbf{y}_{ds}$ is not really a sub-vector of the census vector $\mathbf{y}_d$. In that case, instead of the original best predictor of η_d, we can use an approximation of it, called Census best (CB) predictor, introduced by Correa et al. (2012) and described also in Molina (2019). For indicators, such as the FGT indicators (7), which may be expressed in the form $\eta_d = N_d^{-1}\sum_{j=1}^{N_d} h_j(y_{dj})$ (called hereafter additive indicators), the CB predictor is defined as

$$\hat{\eta}_d^{\text{CB}} = \frac{1}{N_d}\sum_{j=1}^{N_d} E_{\mathbf{y}_d}\left[h_j(y_{dj})|\mathbf{y}_{ds};\boldsymbol{\theta}\,\right], \tag{11}$$

where now the survey vector $\mathbf{y}_{ds}$ follows the same distribution as before, but it is not necessarily a sub-vector of $\mathbf{y}_d$. Hence, we may assume that the distribution of $\mathbf{y}_d|\mathbf{y}_{ds}$ is the same as that of $\mathbf{y}_{dr}|\mathbf{y}_{ds}$ given in (3), but with $\mathbf{X}_{dr}$ replaced by $\mathbf{X}_d$ and $\mathbf{1}_{N_d-n_d}$ replaced by $\mathbf{1}_{N_d}$. We note that the difference between (10) and (11) is

$$\frac{1}{N_d}\sum_{j\in s_d}\left\{h(y_{dj}) - E_{\mathbf{y}_d}\left[h_j(y_{dj})|\mathbf{y}_{ds};\boldsymbol{\theta}\,\right]\right\},$$

which is negligible if $n_d \ll N_d$, as it happens for the Palestinian poverty data. The Census EB (CEB) predictor $\hat{\eta}_d^{\text{CEB}}$ is then obtained by replacing the unknown $\boldsymbol{\theta}$ by the estimate $\hat{\boldsymbol{\theta}}$. When the area sampling fraction n_d/N_d is negligible, which is typically the case in small area applications, the CEB predictor is very similar to the original EB.

4 | MIXED BEST PREDICTORS BASED ON A MULTIVARIATE NORMAL MIXTURE MODEL

The NE model (1) 'borrows strength' from the other areas thanks to the assumption of a homogeneous relation between the response variables y_{dj} and the auxiliary variables $\mathbf{x}_{dj}$ for all the areas, represented by the common regression coefficients $\boldsymbol{\beta}$ and variances τ^2 and σ^2. Nevertheless, in real-life applications, such homogeneity assumption might not hold for all the areas; for example, certain areas might have different idiosyncracy (e.g., those at the border with other countries), displaying different regression parameters and/or variances. To account for these potential differences without knowing a priori which are the outlying areas, we consider the following mixture model, which is a generalisation of the NE model. This model considers that there is an underlying clustering structure of the areas in K clusters, with cluster specific regression coefficients β_k and/or variances of random effects τ_k^2, and/or error variances σ_k^2, where

$\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k', \tau_k^2, \sigma_k^2)'$, $k = 1, \dots, K$. The non-constant error variances σ_k^2 allow for heteroscedasticy across clusters. More formally, let π_k be the probability that an area belongs to cluster $k = 1, \dots, K$, where $\pi_1 + \pi_2 + \dots + \pi_K = 1$ and let $\boldsymbol{\psi} = (\boldsymbol{\pi}', \boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_K')'$ be the vector of model parameters. Then, we consider the following MM model for the area vectors $\mathbf{y}_d$:

$$\mathbf{y}_d \overset{ind.}{\sim} \sum_{k=1}^{K} \pi_k N(\mathbf{X}_d \boldsymbol{\beta}_k, \mathbf{V}_{kd}), \quad d = 1, \dots, D, \tag{12}$$

where the covariance matrices are given by $\mathbf{V}_{kd} = \tau_k^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}' + \sigma_k^2 \mathbf{I}_{N_d}$, $k = 1, \dots, K$, and where $\boldsymbol{\theta}_k \neq \boldsymbol{\theta}_\ell$ for each pair of clusters (k, ℓ) with $k \neq \ell$. Even for a fixed number of components $K \in I\ N$, model (12) is not identifiable for $\boldsymbol{\psi} = (\boldsymbol{\pi}', \boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_K')'$, because there always exist two different vectors of parameters $\boldsymbol{\psi}$ for which we obtain the same density just by switching the component labels. The condition $\pi_1 > \pi_2 > \dots > \pi_K$ avoids this label switching problem, making the MM model (12) with a fixed number of components $K \in I\ N$ identifiable for $\boldsymbol{\psi} = (\boldsymbol{\pi}', \boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_K')'$, see for example McLachlan and Peel (2000), p. 27. Note that the NE model (1) is obtained from the MM model (12) by setting $\pi_1 = 1$ and $\pi_k = 0$, $k = 2, \dots, K$. Moreover, setting $\boldsymbol{\beta}_k = \boldsymbol{\beta}$ and $\tau_k^2 = \tau^2$ for all $k = 1, \dots, K$, we obtain a mixture model where the mixture changes only the error variances σ_k^2. If we wish to consider a mixture model where only the area effects variances change across clusters, we can set $\boldsymbol{\beta}_k = \boldsymbol{\beta}$ and $\sigma_k^2 = \sigma^2$, for all $k = 1, \dots, K$ instead. Hence, the considered MM model includes some of the particular models that may be of interest, but it incorporates additional between-area heterogeneity, by allowing the regression coefficients and/or any of the variance components to vary across the (latent) clusters of areas.

The above MM model can be alternatively defined in terms of latent variables defining the clusters. Let $\mathbf{z}_d = (z_{d1}, \dots, z_{dK})'$ be a multinomial random vector of size parameter 1, where $z_{dk} \in \{0, 1\}$, for all $k = 1, \dots, K$, with $z_{d1} + \dots + z_{dK} = 1$. Concretely, we consider that $\mathbf{z}_d = (z_{d1}, \dots, z_{dK})' \overset{iid}{\sim} \text{Multin}(1, (\pi_1, \dots, \pi_K))$, $d = 1, \dots, D$. Let $\mathbf{e}_k = (\mathbf{0}_{k-1}', 1, \mathbf{0}_{K-k}')'$ be a vector of zeros except for a one in position k. Then, we have $P(\mathbf{z}_d = \mathbf{e}_k) = \pi_k$, $k = 1, \dots, K$. The MM model assumes that the vectors $(\mathbf{y}_d, \mathbf{z}_d)$ are independent for all $d = 1, \dots, D$, and that, given $\mathbf{z}_d = \mathbf{e}_k$, it holds

$$\mathbf{y}_d | \mathbf{z}_d = \mathbf{e}_k \sim N(\mathbf{X}_d \boldsymbol{\beta}_k, \mathbf{V}_{kd}), \quad k = 1, \dots, K, \ d = 1, \dots, D. \tag{13}$$

Concerning the prediction of the target domain parameter η_d, in this case the best predictor given in (2) has no interest, since the conditional distribution of $\mathbf{y}_{dr} | \mathbf{y}_{ds}$ would be averaging over all the possible clusters where area d could belong to; that is, over the possible values $\mathbf{e}_1, \dots, \mathbf{e}_K$ of the vector of latent variables $\mathbf{z}_d$. However, since the indicator of interest η_d is defined for area d, which belongs to a particular cluster, the conditional distribution $\mathbf{y}_{dr} | \mathbf{y}_{ds}, \mathbf{z}_d = \mathbf{e}_k$ is the one of interest in this case, with the only problem that $\mathbf{z}_d$ is unobservable. According to this, we define the following theoretical mixed best (MB) predictor of η_d, which will be used to define the final predictors in this setup:

$$\hat{\eta}_d^{\text{MBk}} = \hat{\eta}_d^{\text{MB}}(\mathbf{e}_k) = E_{\mathbf{y}_{dr}}(\eta_d | \mathbf{y}_{ds}, \mathbf{z}_d = \mathbf{e}_k; \boldsymbol{\theta}_k), \quad k = 1, \dots, K. \tag{14}$$

Note that the MB predictor of η_d given in (14) is optimal (or 'best'), in the sense of having minimum mean squared error with respect to the conditional distribution in (13), given a fixed value for the latent cluster indicator $\mathbf{z}_d$ for area d. Here we consider that the clustering structure defined by $\mathbf{z}_k$ is not known, in which case the MB predictor cannot be calculated. Then, we consider two different MB predictors based on (14). The first one, called mixed best 1 (MB1) predictor, is based on classifying the area d to the cluster k for which the posterior probability $p_{kd} = P(\mathbf{z}_d = \mathbf{e}_k|\mathbf{y}_{ds})$ is the largest for that area; that is,

$$\hat{\eta}_d^{\text{MB1}} = \hat{\eta}_d^{\text{MB}k} \quad \text{if} \quad p_{kd} = \max_{1\le\ell\le K}(p_{\ell d}), \tag{15}$$

where the posterior probabilities of area d belonging to each cluster are given by

$$p_{kd} = P(\mathbf{z}_d = \mathbf{e}_k|\mathbf{y}_{ds}) = \frac{f(\mathbf{y}_{ds}|\mathbf{z}_d = \mathbf{e}_k)\pi_k}{\sum_{\ell=1}^K f(\mathbf{y}_{ds}|\mathbf{z}_d = \mathbf{e}_\ell)\pi_\ell}, \quad k = 1, \dots, K.$$

Here, $f(\mathbf{y}_{ds}|\mathbf{z}_d = \mathbf{e}_k)$ is the p.d.f. of the sample domain vector $\mathbf{y}_{ds}$, which is obtained marginalising from (13) and has the same distribution as the population vector but for the sample elements only, provided that sample selection bias is absent.

The second predictor of η_d, called MB2, is the weighted average of the theoretical MB predictors for the K clusters, using as weights the above posterior probabilities,

$$\hat{\eta}_d^{\text{MB2}} = \sum_{k=1}^K p_{kd}\, \hat{\eta}_d^{\text{MB}k} \tag{16}$$

Now according to the MM model (13) and by the independency of the area vectors $(\mathbf{y}'_d, \mathbf{z}'_d)'$, $d = 1, \dots, D$, it is easy to see that, given $\mathbf{y}_{ds}$ and $\mathbf{z}_d = \mathbf{e}_k$, the out-of-sample area vectors $\mathbf{y}_{dr}$ are independent for all d, satisfying

$$\mathbf{y}_{dr}|\mathbf{y}_{ds}, \mathbf{z}_d = \mathbf{e}_k \sim N\left(\boldsymbol{\mu}_{dr|s}(\boldsymbol{\theta}_k), \mathbf{V}_{dr|s}(\boldsymbol{\theta}_k)\right), \; k = 1, \dots, K, \tag{17}$$

for $\boldsymbol{\mu}_{dr|s}(\boldsymbol{\theta}_k)$ and $\mathbf{V}_{dr|s}(\boldsymbol{\theta}_k)$ obtained, respectively, from the conditional mean vector and covariance matrix given in (4), replacing $\boldsymbol{\theta}$ by the parameter vector $\boldsymbol{\theta}_k$ of cluster k.

If an area d_0 has no sample in the survey ($n_{d_0} = 0$), then p_{kd_0} is not defined. We then replace it by the average posterior probabilities for the sampled areas, $\tilde{\pi}_k = D^{-1}\sum_{d=1}^D p_{kd}$, where the sum is done in the sampled domains. Further, we set $\tau_k = 0$ for all k in the conditional distribution (17), and then use the same MB1 and MB2 estimators.

For the sample mean of the response variable in the model for area d, $\eta_d = \overline{Y}_d$, the MB predictor given in (14) reduces to

$$\hat{\eta}_d^{MBk} = \hat{\overline{Y}}_d^{MBk} = N_d^{-1}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \mu_{dj|s}(\boldsymbol{\theta}_k)\right),$$

where $\mu_{dj|s}(\boldsymbol{\theta}_k) = \mathbf{x}'_{dj}\boldsymbol{\beta}_k + \tilde{u}_d(\boldsymbol{\theta}_k), j \in r_d$, which equals the BLUP of $\overline{Y}_d$ that uses the true value of the parameter vector for group k, $\boldsymbol{\theta}_k$. The MB1 predictor is then given by

$$\hat{\overline{Y}}_d^{\text{MB1}} = \hat{\overline{Y}}_d^{\text{MB}k} \quad \text{if} \quad p_{kd} = \max_{1\le\ell\le K}(p_{\ell d}), \tag{18}$$

and the MB2 predictor reads

$$\hat{\overline{Y}}_d^{\mathrm{MB2}} = \sum_{k=1}^{K} p_{kd}\,\hat{\overline{Y}}_d^{\mathrm{MB}k}. \tag{19}$$

For the FGT poverty indicator of order $\alpha \geq 0$, $F_{\alpha,d}$, the MB predictor is

$$\hat{F}_{\alpha,d}^{\mathrm{MB}k}(\boldsymbol{\theta}_k) = \frac{1}{N_d}\left(\sum_{j\in s_d} F_{\alpha,dj} + \sum_{j\in r_d} \hat{F}_{\alpha,dj}(\boldsymbol{\theta}_k)\right). \tag{20}$$

For the poverty rate ($\alpha = 0$) and for nondecreasing monotonous transformation $T(\cdot)$, $\hat{F}_{0,dj}(\boldsymbol{\theta}_k) = \Phi(\alpha_{dj}(\boldsymbol{\theta}_k))$. For the poverty gap ($\alpha = 1$), considering the log transformation after a shift, $y_{dj} = \log(w_{dj} + c)$, we have

$$\hat{F}_{1,dj}(\boldsymbol{\theta}_k) = \frac{z+c}{z}\,\Phi(\alpha_{dj}(\boldsymbol{\theta}_k)) - \frac{1}{z}\exp\left(\mu_{dj|s}(\boldsymbol{\theta}_k) + \sigma^2_{dj|s}(\boldsymbol{\theta}_k)/2\right)\Phi\left(\alpha_{dj}(\boldsymbol{\theta}_k) - \sigma_{dj|s}(\boldsymbol{\theta}_k)\right).$$

For the average welfare $\overline{W}_d$ under the same transformation of w_{dj}, the MB predictor becomes

$$\hat{\overline{W}}_d^{\mathrm{MB}k} = \frac{1}{N_d}\left(\sum_{j\in s_d} w_{dj} + \sum_{j\in r_d} \hat{w}_{dj}(\boldsymbol{\theta}_k)\right),$$

where $\hat{w}_{dj}(\boldsymbol{\theta}_k) = \exp\left(\mu_{dj|s}(\boldsymbol{\theta}_k) + \sigma^2_{dj|s}(\boldsymbol{\theta}_k)/2\right) - c$.

The two alternative predictors MB1 and MB2 for $F_{\alpha,d}$ and $\overline{W}_d$ are then obtained similarly as $\hat{\overline{Y}}_d^{\mathrm{MB1}}$ and $\hat{\overline{Y}}_d^{\mathrm{MB2}}$, based on the corresponding theoretical MB predictors $\hat{F}_{\alpha,d}^{\mathrm{MB}k}$ and $\hat{\overline{W}}_d^{\mathrm{MB}k}$, either selecting the cluster or making a weighted average based on the posterior probabilities p_{kd}, $k = 1, \ldots, K$.

Similarly as for the best predictor, empirical versions of MB1 and MB2, called EMB1 and EMB2, respectively, are defined replacing $\boldsymbol{\psi} = (\boldsymbol{\pi}', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')'$ by an estimator $\hat{\boldsymbol{\psi}}$ such as the ML one. Appendix A describes the expectation-maximisation (E-M) algorithm yielding the ML estimator of $\boldsymbol{\psi}$. The resulting E-M updating equations for the model parameters of cluster k given in (A5) involve sums of terms for each area, with weights p_{kd}, $d = 1, \ldots, D$. This means that data from the areas with higher posterior probability of belonging to group k are given more weight in the updating equations for group k than data from those areas with lower posterior probabilities of belonging to that group. When the survey data $\mathbf{y}_{ds}$ is not a sub-vector of the census counterpart $\mathbf{y}_d$, Census versions of MB1 and MB2 predictors for additive indicators are defined analogously to the Census EB predictor in (11).

Mixture models with $K = 2$ components are commonly used for modelling outliers. Since $\pi_1 > \pi_2$, the first cluster corresponds to the non-outliers and the second to the outliers. In our context, the considered MM may be used to represent outlying areas. According to the E-M algorithm described in Appendix A, in the presence of outlying areas modelled by the MM, the ML estimators $\hat{\boldsymbol{\beta}}_1$, $\hat{\tau}_1^2$ and $\hat{\sigma}_1^2$ for the first cluster ($k = 1$) will be giving less weight to the outlying areas, resulting in more robust estimators. Hence, the model parameter estimators for the first cluster and their corresponding MB predictors can be used as robust estimators in the usual NE model (1).

For complex indicators $\eta_d = h(\mathbf{y}_d)$ without closed form for the expectation in (14), the EMB1 and EMB2 predictors can be obtained using the following MC simulation procedure, which extends that one in Molina and Rao (2010):

(a) Fit the MM to the sample data $\mathbf{y}_s = (\mathbf{y}'_{1s}, \ldots, \mathbf{y}'_{Ds})'$ and obtain model parameter estimates $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\pi}}', \hat{\boldsymbol{\theta}}'_1, \ldots, \hat{\boldsymbol{\theta}}'_K)'$.
(b) For each $k = 1, \ldots, K$, draw I copies of the out-of-sample vector $\mathbf{y}_{dr}$ from its conditional distribution given the sample data $\mathbf{y}_{ds}$ and the latent vector $\mathbf{z}_d = \mathbf{e}_k$ given in (17), with $\boldsymbol{\theta}_k = (\boldsymbol{\beta}'_k, \tau_k^2, \sigma_k^2)'$ replaced by the estimate $\hat{\boldsymbol{\theta}}_k = (\hat{\boldsymbol{\beta}}'_k, \hat{\tau}_k^2, \hat{\sigma}_k^2)'$ obtained in step a). We denote the generated vectors as $\mathbf{y}_{drk}^{(i)}$, $i = 1, \ldots, I$.
(c) For each MC replicate $i = 1, \ldots, I$, with I large, attach the sample data $\mathbf{y}_{ds}$ to the generated out-of-sample vector $\mathbf{y}_{drk}^{(i)}$ to get the census vector $\mathbf{y}_{dk}^{(i)} = (\mathbf{y}'_{ds}, (\mathbf{y}_{drk}^{(i)})')'$, and calculate the indicator of interest based on it, $\eta_{dk}^{(i)} = h(\mathbf{y}_{dk}^{(i)})$, $k = 1, \ldots, K$.
(d) Let $\hat{p}_{dk} = p_{dk}(\hat{\boldsymbol{\psi}})$, $k = 1, \ldots, K$. The MC approximation of the MB predictor for each cluster k is then

$$\hat{\eta}_d^{\mathrm{MB}k} = \frac{1}{I}\sum_{i=1}^{I} \eta_{dk}^{(i)}, \quad k = 1, \ldots, K.$$

Based on the above MB predictors for each cluster $k = 1, \ldots, K$, we obtain a MC approximation to EMB1 as

$$\hat{\eta}_d^{\mathrm{EMB1}} = \hat{\eta}_d^{\mathrm{MB}k}, \quad \text{if} \quad \hat{p}_{kd} = \max_{1\le \ell \le K}(\hat{p}_{\ell d}),$$

and similarly for EMB2,

$$\hat{\eta}_d^{\mathrm{EMB2}} = \sum_{k=1}^{K} \hat{p}_{kd}\hat{\eta}_d^{\mathrm{MB}k}.$$

Step (b) of the above MC procedure requires to generate multivariate normal vectors $\mathbf{y}_{drk}^{(i)}$ that may be of very large dimension, since N_d is typically very large and n_d is small. To avoid generation of such large multivariate vectors, they can be generated from the following auxiliary models, which require generation of univariate random variables only,

$$\mathbf{y}_{drk} = \boldsymbol{\mu}_{dr|s}(\boldsymbol{\theta}_k) + v_{kd}\mathbf{1}_{N_d-n_d} + \mathbf{e}_{kdr}, \quad k = 1, \ldots, K,$$

where $v_{kd} \overset{ind.}{\sim} N(0, \tau_k^2(1-\gamma_{kd}))$ and $\mathbf{e}_{kdr} \overset{ind.}{\sim} N(\mathbf{0}_{N_d-n_d}, \sigma_k^2\mathbf{I}_{N_d-n_d})$, $k = 1, \ldots, K$. For the Census EMB1 and EMB2, the whole census vector for area d of size N_d, $\mathbf{y}_{dk}^{(i)}$, is generated in Step b) from the model

$$\mathbf{y}_{dk} = \boldsymbol{\mu}_{d|s}(\boldsymbol{\theta}_k) + v_{kd}\mathbf{1}_{N_d} + \mathbf{e}_{kd}, \quad k = 1, \ldots, K,$$

where $\boldsymbol{\mu}_{d|s}(\boldsymbol{\theta}_k)$ is the same as $\boldsymbol{\mu}_{dr|s}(\boldsymbol{\theta}_k)$, with $\mathbf{X}_{dr}$ replaced by $\mathbf{X}_d$ and $\mathbf{1}_{N_d-n_d}$ by $\mathbf{1}_{N_d}$, and $\mathbf{e}_{kd} \overset{ind.}{\sim} N(\mathbf{0}_{N_d}, \sigma_k^2\mathbf{I}_{N_d})$, $k = 1, \ldots, K$. Then, true values of the indicators are calculated in Step (c) from $\mathbf{y}_{dk}^{(i)}$ without attaching the sample data $\mathbf{y}_{ds}$.

5 | PARAMETRIC BOOTSTRAP FOR MEAN SQUARED ERROR ESTIMATION

Even for the EB predictor, an analytical expression of the MSE is not available in the case of complex indicators and bootstrap procedures are typically applied. Here we extend the parametric bootstrap procedure given in Molina and Rao (2010) for the EB predictor, based on González-Manteiga et al. (2008), to the case of the MM model and the proposed MB predictors. The extended bootstrap MSE estimator of the EMB1 or EMB2 predictor of η_d is obtained as follows:

(a) Estimate the MM model parameters $\boldsymbol{\psi} = (\boldsymbol{\pi}', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')'$ using the sample data $\mathbf{y}_s$, yielding $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\pi}}', \hat{\boldsymbol{\theta}}_1', \ldots, \hat{\boldsymbol{\theta}}_K')'$, for $\hat{\boldsymbol{\theta}}_k = (\hat{\boldsymbol{\beta}}_k', \hat{\tau}_k^2, \hat{\sigma}_k^2)'$, $k = 1, \ldots, K$.
(b) For a large number of bootstrap replicates B $(b = 1, \ldots, B)$, generate bootstrap census vectors for each area $\mathbf{y}_d^{*(b)}$, $d = 1, \ldots, D$, from the MM model (12) with parameters $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\pi}}', \hat{\boldsymbol{\theta}}_1', \ldots, \hat{\boldsymbol{\theta}}_K')'$. Calculate the target indicator $\eta_d^{*(b)} = h(\mathbf{y}_d^{*(b)})$.
(c) Take the bootstrap y^* and $\mathbf{x}$-values corresponding to the units in the original sample $s = \cup_{d=1}^{D} s_d$, that is, take $(y_{dj}^{*(b)}, \mathbf{x}_{dj})$, $j \in s_d$, $d = 1, \ldots, D$. With these data, fit the MM model, obtaining bootstrap parameter estimates $\hat{\boldsymbol{\psi}}^{*(b)} = ((\hat{\boldsymbol{\pi}}^{*(b)})', (\hat{\boldsymbol{\theta}}_1^{*(b)})', \ldots, (\hat{\boldsymbol{\theta}}_K^{*(b)})')'$, for $\hat{\boldsymbol{\theta}}_k^{*(b)} = ((\hat{\boldsymbol{\beta}}_k^{*(b)})', \hat{\tau}_k^{2*(b)}, \hat{\sigma}_k^{2*(b)})'$, $k = 1, \ldots, K$. This can be done using the E-M algorithm with starting values of the model parameters taken as the estimates obtained in step (a).
(d) Using the bootstrap sample data in (c) and the out-of-sample auxiliary data $\mathbf{x}_{dj}$, $j \in r_d$, calculate the EMB1 and EMB2 predictors, $\hat{\eta}_d^{\text{EMB1}*(b)}$ and $\hat{\eta}_d^{\text{EMB2}*(b)}$. Then, the bootstrap MSE estimators of the EMB1 and EMB2 predictors of η_d are given by

$$\text{mse}(\hat{\eta}_d^{\text{EMB1}}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\eta}_d^{\text{EMB1}*(b)} - \eta_d^{*(b)})^2, \quad \text{mse}(\hat{\eta}_d^{\text{EMB2}}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\eta}_d^{\text{EMB2}*(b)} - \eta_d^{*(b)})^2.$$

When the available census does not contain the survey units and therefore linking the two data sets is not possible, we consider a slight variation of this bootstrap procedure, which is an extension of the parametric bootstrap for the Census EB estimator described in Molina (2019) for the MM model. In this bootstrap procedure, instead of extracting the bootstrap sample data $\mathbf{y}_{ds}^{*(b)}$ in Step (c) from the generated census data $\mathbf{y}_d^{*(b)}$ in Step (b), the sample data $\mathbf{y}_{ds}^{*(b)}$ are generated separately within each bootstrap replicate, from the corresponding marginal model of the census vector $\mathbf{y}_d^{*(b)}$ in Step b), using exactly the same values of the model parameter estimates.

Section 4 of Supplementary Material presents simulation results that investigate empirically the behaviour of the proposed parametric bootstrap MSE estimator. Proving consistency of this MSE estimator is a challenge that deserves further research, since there is lack of the literature on bootstrap MSE methods for small area predictors based on mixture models.

6 | SIMULATION EXPERIMENTS

This section describes the results of simulation experiments performed to compare the properties, in terms of bias and MSE, of direct estimators, EB estimators based on the NE model and EMB1 and EMB2 estimators based on the MM model, for both linear parameters (area means) and complex parameters (area poverty rates). Similarly as in Molina and Rao (2010), we consider a population of size $N = \sum_{d=1}^{D} N_d$ with $D = 80$ areas, with area population sizes $N_d = 250$, $d = 1, \dots, D$. As auxiliary variables, we consider two dummy indicators as in Molina and Rao (2010), but in this case we consider that they have interaction between each other, defining a grouping of the $N_d = 250$ observations in each area in four groups, corresponding to the observations with values for the two dummies $(x_{1dj}, x_{2dj}) \in \{(0,0), (1,0), (0,1), (1,1)\}$. In this simulation experiment, data will be generated under the MM model, which defines a latent grouping, so it seems more realistic to consider covariates that also define a natural grouping of the observations. In order to consider heterogeneous areas also with respect to the configuration of the values of the covariates, we make the number of observations N_{d00} in the first group, with $(x_{1dj}, x_{2dj}) = (0,0)$, and the N_{d01} observations in the third group, with $(x_{1dj}, x_{2dj}) = (0,1)$, to decrease in the area index d. On the other hand, the number of observations N_{d10} in the second group, with $(x_{1dj}, x_{2dj}) = (1,0)$, as well as those in the last group N_{d11}, taking values $(x_{1dj}, x_{2dj}) = (1,1)$, increase in the area index d. As a consequence, the observations in the first areas (smaller d) take the values (0, 0) and (0, 1) more often, whereas the observations in the last areas (larger d) have larger proportion of (1, 0) and (1, 1) values than those in the first areas. More specifically, the values of (x_{1dj}, x_{2dj}) are generated as follows: first, generate success probabilities as

$$p_{1d} = 0.3 + \frac{0.5d}{D}, \quad p_{2d} = 0.2, \quad d = 1, \dots, D.$$

Then, defining $N_{d00} = \lfloor (1 - p_{1d})(1 - p_{2d}) N_d \rfloor$, $N_{d10} = \lfloor p_{1d}(1 - p_{2d}) N_d \rfloor$, $N_{d01} = \lfloor (1 - p_{1d}) p_{2d} N_d \rfloor$ and $N_{d11} = N_d - N_{d00} - N_{d10} - N_{d01}$, where $\lfloor \cdot \rfloor$ denotes the closest lower integer, generate the values of the two dummy variables as follows:

$$(x_{1dj}, x_{2dj}) = \begin{cases} (0,0) & \text{if } j = 1, \dots, N_{d00}, \\ (1,0) & \text{if } j = N_{d00} + 1, \dots, N_{d00} + N_{d10}, \\ (0,1) & \text{if } j = N_{d00} + N_{d10} + 1, \dots, N_{d00} + N_{d10} + N_{d01}, \\ (1,1) & \text{if } j = N_{d00} + N_{d10} + N_{d01} + 1, \dots, N_d, \end{cases}$$

Area effects are drawn as $u_{kd} \overset{iid}{\sim} N(0, \tau_k^2)$, with $\tau_0^2 = 0.20^2$ and $\tau_1^2 = 0.15^2$, and errors are drawn as $e_{kdj} \sim N(0, \sigma_k^2)$, $k = 0, 1$, for $\sigma_0^2 = 0.75^2$ and $\sigma_1^2 = 0.5^2$.

The values of the response variable y_{dj} for the population units are then generated from the MM model with $K = 2$ mixture components. For this, first we draw random values of the mixture component indicators $Z_d \sim \text{Bern}(\pi)$, with $\pi = 3/4$. Then, when $Z_d = k$, we generate the target variables y_{dj} from the model

$$y_{dj} = \beta_{k0} + \beta_{k1} x_{1dj} + \beta_{k2} x_{2dj} + u_{kd} + e_{kdj}, \quad d = 1, \dots, D, \ j = 1, \dots, N_d, \ k = 0, 1,$$

where $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \beta_{12}) = (2, 0.05, -0.06)$. Now for $\boldsymbol{\beta}_0$, in Case 1 we take $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \beta_{02}) = (3, 0.03, -0.04)$; in Case 2, we take the vector $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \beta_{02}) = (5, 0.03, -0.04)$ with a further apart intercept and, finally, in Case 3, we take the vector with opposite signs for all the elements, $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \beta_{02}) = (-2, -0.05, 0.06)$.

As target indicators for the areas η_d, we consider the means of the response variable $\overline{Y}_d = N_d^{-1} \sum_{j=1}^{N_d} y_{dj}$ and the poverty rates $F_{0d} = N_d^{-1} \sum_{j=1}^{N_d} I(\exp(y_{dj}) < z)$, $d = 1, \dots, D$, where the poverty line z is taken as $z = 0.6 \cdot \text{Median}\left\{y_{dj};\ d = 1, \ldots, D,\ j = 1, \ldots, N_d\right\}$ for one generated population and then kept fixed.

The sample is obtained by simple random sampling within each area, of size $n_d = 20, 50$, $d = 1, \ldots, D$. The sample indices and the values of the auxiliary variables for the population units are kept fixed across simulations. We consider three estimators: (1) direct estimator (sample mean), (2) EB based on the NE model and (3) EMB1 and EMB2 based on the MM model. ELL method is not included here because of its clearly poorer performance compared to EB method, as observed in many publications, see for example Corral et al. (2021) for a comparison of these methods in many simulation experiments under very different (including realistic) scenarios.

The simulations are repeated $R = 10{,}000$ times. With each simulated population $\{y_{dj}^{(r)};\ d = 1, \ldots, D,\ j = 1, \ldots, N_d\}$, we calculate the true values of the target indicators for the areas (means, poverty rates), then we select the sample values and fit the NE model by REML and the MM model by the E-M algorithm described in Appendix A. Then, we obtain direct, EB and EMB1 and EMB2 estimators of the area means and of the poverty rates. For a generic estimator $\hat{\eta}_d$ of an indicator η_d and for the indicator itself, we calculate empirical expectations as follows:

$$E(\hat{\eta}_d) = \frac{1}{R}\sum_{r=1}^{R}\hat{\eta}_d^{(r)}, \quad E(\eta_d) = \frac{1}{R}\sum_{r=1}^{R}\eta_d^{(r)}, \quad d = 1, \ldots, D.$$

Similarly, we compute the empirical MSEs of the generic estimator $\hat{\eta}_d$ as

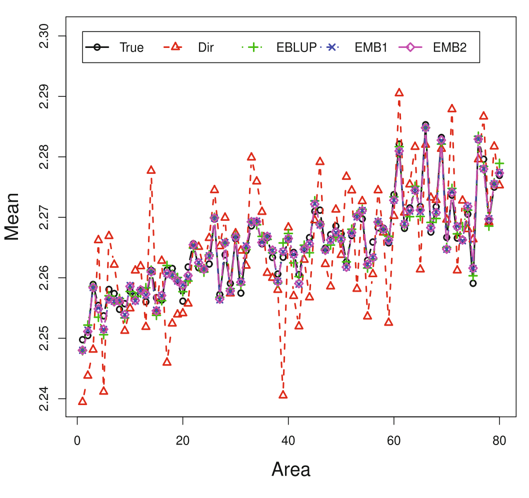
$$\text{MSE}(\hat{\eta}_d) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\eta}_d^{(r)} - \eta_d^{(r)})^2, \quad d = 1, \ldots, D.$$

Firstly, the proposed E-M fitting method (see Appendix A) was analysed by fitting the MM for area sample sizes $n_d = 20, 50$, number of areas $D = 80{,}300$, for different sets of model parameters, different starting values and for number of components $K = 2$ and $K = 3$. A selection of the results for $K = 2$ components are reported in Table 1 of Supplementary Material. See that the empirical expectations of the estimators of the model parameters delivered by the algorithm are very close to the true values and vary very little when using starting values that are far from the true values, although the relative root MSEs (RRMSEs) clearly increase. The absolute relative bias of the resulting estimators (ARB) and the RRMSEs tend to decrease when increasing the area sample sizes n_d or the number of areas D. For $K = 3$ mixture components, a selection of the results are shown in Table 2 of the Supplementary Material. We can see that, again, the empirical expectations remain similar to the true values of the model parameters when using the true values as starting values. Now when using starting values that are different from the true values, the empirical expectations remain close to the true value, except that the mixture components 1 and 3 are mixed up; that is, the empirical expectations of the estimators of the parameters of the first mixture component are close to the true values of the third component and reversely for the third component. This occurs because the fitting algorithm does not actually restrict the mixture component probabilities, π_k, but just by sorting the original components and the estimated ones by increasing π_k would lead to correct results. Hence, these results indicate the stability of the E-M fitting method.

Let us now analyse the performance of the small area estimators. Figure 1 left displays the empirical expectations of each estimator of the area means compared with those of the true means, for regression coefficients under Case 1 (closer values for the two components), and for $n_d = 20$. No systematic bias can be observed for any of the estimators, although direct ones seem more unstable than the other ones. However, looking at Figure 1 right, showing the empirical MSEs ($\times 10^3$), we can see a considerable MSE reduction for the two estimators based on the MM, with EMB2 performing very similarly, but just slightly better than EMB1. Note that for area means, EB estimators that use the weighted least squares estimator of the regression coefficients reduce to the EBLUPs. We can see that EBLUPs perform clearly better than direct estimators in terms of MSE, but they do not achieve the MSE reductions of the estimators obtained under the correct MM. Under Case 3 with strong differences in the regression coefficients for the two mixture components, results in Figure 2 indicate an overwhelming gain of the estimators based on the MM compared to the other ones, with the EBLUP becoming no better than the direct estimators. Results for Case 2 are omitted for brevity but, as can be guessed, they are in between those of Case 1 and 3.

As commented in Section 4, mixture models with $K = 2$ components are common models for outliers; in this case, the MM with $K = 2$ may represent outlying areas (those coming from the less-frequent mixture component). For this reason, Section 2 of Supplementary Material compares the EMB1 predictor with the robust EBLUP (REBLUP) of Sinha and Rao (2009), which is implemented in the rsae R package (Schoch, 2014). We use the Huber's psi function with the standard value of the tunning parameter 1.345, as suggested in Sinha and Rao (2009) for their robustified ML equations (15) and (16). Results show clear underestimation of the true area means for this estimator and much lower efficiency in terms of MSE than the other estimators. Note that here, the area vectors coming from the second mixture component (with mixture probability equal to 1/4) can be seen as representative outliers, in the sense that they are correct area vectors that should be considered in the estimation of the model parameters and in the area means, and this type of outlying areas may be better represented in the estimators based on the MM model.

For the poverty rates, results are displayed in Figures 3 and 4 for Cases 1 and 3, respectively. For this indicator, the EB estimators turn out to be clearly biased, as can be seen in the two left plots. Of course, direct estimators are unbiased, but their MSEs are much larger than those of the



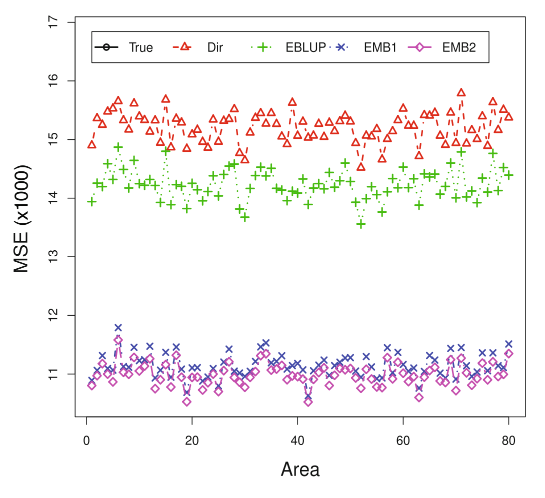


FIGURE 1 Empirical expectations of the true area means and of the corresponding direct, EBLUP, EMB1 and EMB2 estimators (left) and empirical MSEs ($\times 10^3$) of the estimators (right), for Case 1 and $n_d = 20$ [Colour figure can be viewed at wileyonlinelibrary.com]

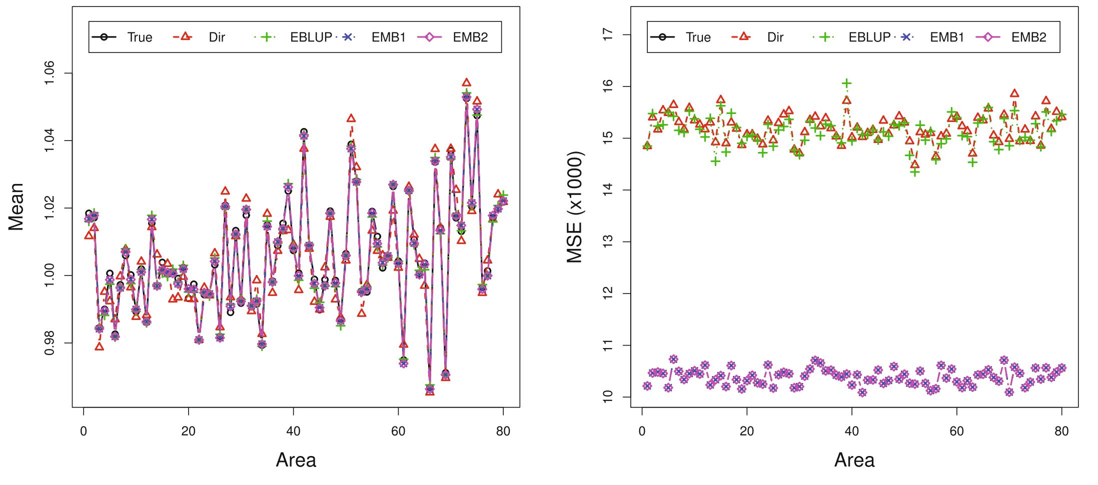


FIGURE 2 Empirical expectations of the true area means and of the corresponding direct, EBLUP, EMB1 and EMB2 estimators (left) and empirical MSEs ($\times 10^3$) of the estimators (right) for Case 3 and $n_d = 20$ [Colour figure can be viewed at wileyonlinelibrary.com]

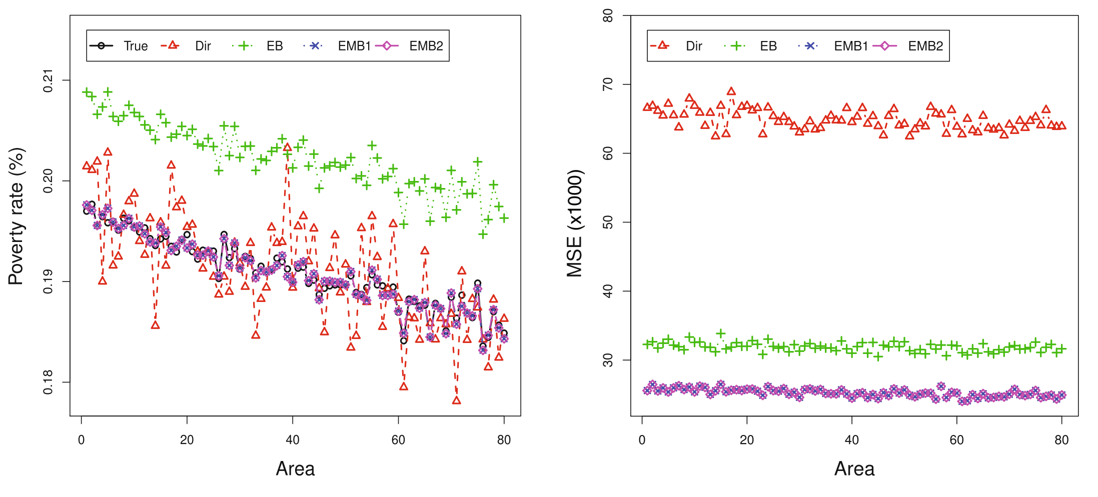


FIGURE 3 Empirical expectations of the true poverty rates (%) and of the corresponding direct, EBLUP, EMB1 and EMB2 estimators (left) and empirical MSEs ($\times 10^4$) of the estimators (right), for Case 1 and $n_d = 20$ [Colour figure can be viewed at wileyonlinelibrary.com]

model-based estimators (right plots). Despite the observed bias, EB predictors of poverty rates still have significantly lower MSE than direct estimators, but EMB1 and EMB2 obtain substantially lower empirical MSEs for the two scenarios of regression coefficients and also perform the best in terms of bias according to the left plots.

We also repeated the latter simulation experiment for estimation of poverty rates by considering now that data come from a model very close to the NE model. For this, we considered that all the model parameters in the two mixture components are exactly the same, except for the intercept. Specifically, we took as true values of the model parameters, $\boldsymbol{\beta}_1 = (3, 0.03, -0.04)$, $\boldsymbol{\beta}_0 = (2, 0.03, -0.04)$, $\tau_1^2 = \tau_0^2 = (0.15)^2$, $\sigma_1^2 = \sigma_0^2 = (0.5)^2$, and finally $\pi_1 = 3/4$ and $\pi_0 = 1/4$ as before. Results are shown in Section 3 of Supplementary Material. These results still show a clearly better

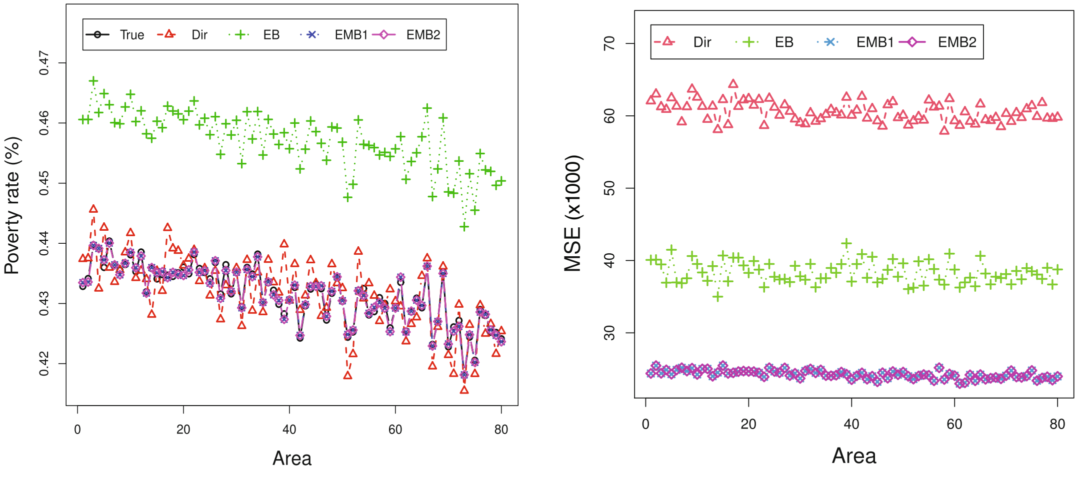


FIGURE 4 Empirical expectations of the true poverty rates (%) and of the corresponding direct, EBLUP, EMB1 and EMB2 estimators (left) and empirical MSEs ($\times 10^4$) of the estimators (right), for Case 3 and $n_d = 20$ [Colour figure can be viewed at wileyonlinelibrary.com]

performance of the EMB1 and EMB2 estimators of the poverty rates based on the MM model than the EB ones based on the NE model, in terms of MSE (and slightly better performance in terms of bias), even if the model that generates the data is very close to the NE model. We do not generate the target variable under a NER model, because this model is a non-identifiable MM model and, in this case, the MM fitting algorithm tends to encounter numerical problems, see Appendix A.

Section 4 of Data S1 presents an additional simulation experiment to study the performance of the parametric bootstrap MSE estimator proposed in Section 5. Results indicate that the empirical expectations of the bootstrap MSE estimates are aligned with the true MSEs.

We have seen that, when data comes from the MM model, EB estimators based on the NE model that ignores the mixture do not show appreciable bias in the case of linear indicators, but their bias is clear for complex indicators such as the poverty rate. On the other hand, EMB1 and EMB2 estimators based on the correct MM model show no bias and are much more efficient than EB and direct estimators.

7 | ESTIMATION OF POVERTY IN PALESTINIAN LOCALITIES: RESULTS

Molina Peralta and García-Portugués (2020) obtained estimates of mean expenditure, poverty rates and gaps for the 136 Palestinian localities in the PECS, based on a NE model with locality random effects as described in Section 3. This model included also the region indicator as a fixed effect. But even after accounting for the region effect, their results show strong regional differences. To check whether these differences are driven only by the region, we fitted a separate NE model for each region. QQ-normal plots of area level residuals (predicted locality effects) from the two fits are shown in Figure 5 for West Bank (left) and Gaza (right). The plot for West Bank indicates no serious departure from the normality assumption, but the one for Gaza shows clear outlying localities. This suggests that there is certain heterogeneity among the Palestinian localities that is not driven only by the region where they belong to, but there must be other factor/s determining this heterogeneity. In order to let the data speak, we consider a model that

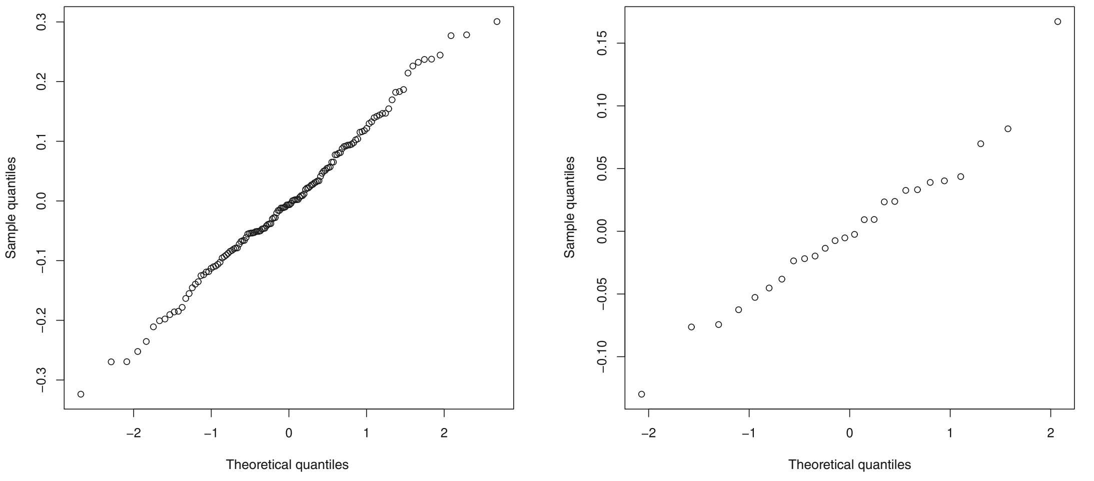


FIGURE 5 QQ-normal plots of area level residuals (predicted locality effects) from separate fits of the nested error model to the Palestinian Expenditure Consumption Survey data from West Bank (left) and Gaza (right)

accounts for heterogeneity by considering a latent grouping structure of the localities, in which the localities in the same group share the same values of the model parameters, but different groups of localities may have potentially different model parameters. This may be represented by the MM presented in Section 4, obtained as an extension of the NE model described in Section 3.

The results from fitting the NE and the MM model with $K = 2$ components to the survey data are shown in Table B1, accompanied with the corresponding SEs. These SEs have been obtained using the same bootstrap procedure as for MSE estimation described in Section 5, with the difference that, in this case, only the sample data (instead of full populations) need to be generated. The resulting regression coefficients for the $K = 2$ mixture components are in line with the ones obtained by fitting the NE model without mixture, but the latter are in between those for each of the $K = 2$ mixture components for most of the covariates, which seems a reasonable result. This occurs for the estimated variances of area effects and errors as well.

Table B1 reports the resulting 95% normality-based confidence intervals for all the parameters in the two models. Even though the interval for the two mixture components do overlap, see that the intervals for the variances σ_1^2 and σ_2^2, in the two mixture components, do not overlap. Neither do the analogous 95% confidence intervals for 23 of the regression coefficients, which indicates that localities are certainly heterogeneous, supporting the selection of the MM instead of the NE model. Note that the fitted mixing probabilities are not far from 0.5, whereas the number of localities in Gaza is 26, which represents 16% of the total number of sampled localities $D = 162$ (all other ones are in West Bank). This result, and a grouping of the localities according to the posterior probabilities p_{kd}, suggest no clear relationship between the latent clusters induced by the MM and the two Palestinian regions.

We calculate direct estimators, Census EB estimators based on the NE model as in Molina Peralta and García-Portugués (2020) (survey units cannot be identified in the census), and similarly the Census versions of EMB1 and EMB2 estimators based on the MM model of mean expenditure, poverty rates and gaps by localities. The only difference between the Census EB estimators here and those in Molina Peralta and García-Portugués (2020) is that here do not fit separately models for men and women and that a non-significant covariate and another with VIF > 5 were removed from the final model.

As the unweighted direct estimators are very similar to the weighted ones for the three target indicators, we discuss results for the unweighted ones only. In fact, EMB1 and EMB2 are also very similar in all our results, except for slightly smaller estimated MSE of EMB2, so we will sometimes restrict to EMB2 (based on the weighted average) when discussing the results. MSEs of all the estimators are estimated by the same procedure, which is the bootstrap procedure described at the end of Section 5 for the case when the survey cannot be linked to the census data.

Figure 6 shows scatterplots of Census EB estimates (left) and Census EMB2 ones (right) against direct estimates for mean expenditure, with region initials as point labels. In the left plot we can see that most of the points corresponding to localities from West Bank appear above the equality line. Since direct estimators are approximately design unbiased, this indicates an upward design bias for the Census EB estimators of mean expenditure for localities in the West Bank. On the other hand, the right plot, showing Census EMB2 estimates against direct ones, shows points more uniformly distributed at both sides of the equality line, which suggests no systematic bias for the EMB2 estimators. As expected, the analogous plots for the poverty rates show exactly the opposite: some underestimation of the Census EB estimates for the poverty rates in some localities from West Bank, but no systematic bias appears for the Census EBM2, see Figure 7. In fact, these plots show the zero values of the direct estimates of poverty rates for some localities. This occurs for 29 of the localities where no welfare is found below the poverty line, even if their sample sizes are not zero. The corresponding estimated variances of these direct estimators are also zero, which is obviously misleading. Neither Census EB estimators based on the NE model, nor Census EMB1 and EMB2 obtained with the MM model, where zero for any of the localities.

Another point supporting the new estimators is that, when aggregating to the national level, the direct estimators have good quality since the total sample size is large. The direct estimator of mean expenditure at the population level results in 14,698.6 ILS, whereas the Census EB estimate is 15,709.8 and the Census EMB1 and EMB2 are respectively 14,600.9 and 14,613.6. Disaggregating by region, one can see that the overestimation of the Census EB occurs in West Bank.

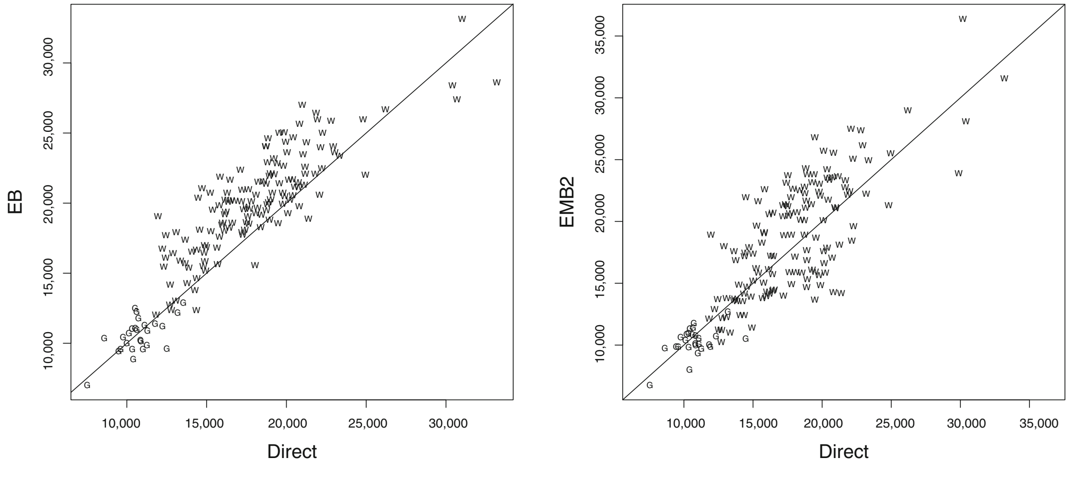


FIGURE 6 Census empirical best estimates (left) and Census EMB2 ones (right) against direct estimates for mean expenditure, with region initials as point labels (W, West Bank; G, Gaza)

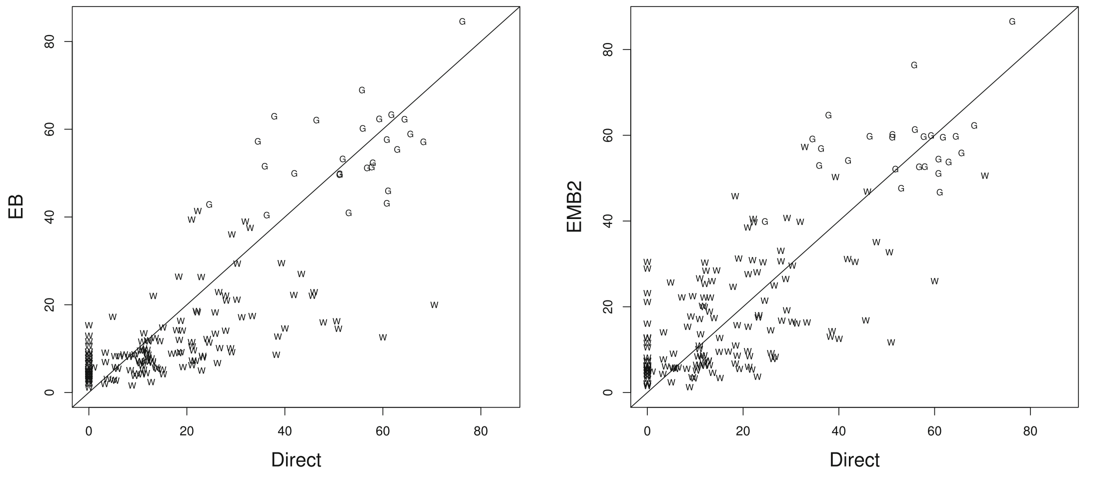


FIGURE 7 Census empirical best estimates (left) and Census EMB2 ones (right) against direct estimates for poverty rate, with region initials as point labels (W, West Bank, G=Gaza)

Let us now show detailed results for each locality. Estimates of poverty rates for West Bank and Gaza are shown in Figure 8 left and right respectively, where localities in both plots (in x axis) are sorted by the increasing order of the sample sizes and these sample sizes are indicated in the x axis labels. See that direct estimators are much more unstable than the two types of model-based estimators in both regions but, as can be appreciated better in the left plot, Census EB ones perhaps tend to be too stable, with EMB2 representing a compromise. This is sensible, since direct estimators are completely non-parametric, Census EB based on the NE model with normality are completely parametric, and EMB2 based on the MM model are more flexible than the usual fully parametric normality-based estimators, since the mixtures represent a large family of distributions. Similar observations are obtained for the mean expenditure and the poverty gap, although plots are not shown for brevity.

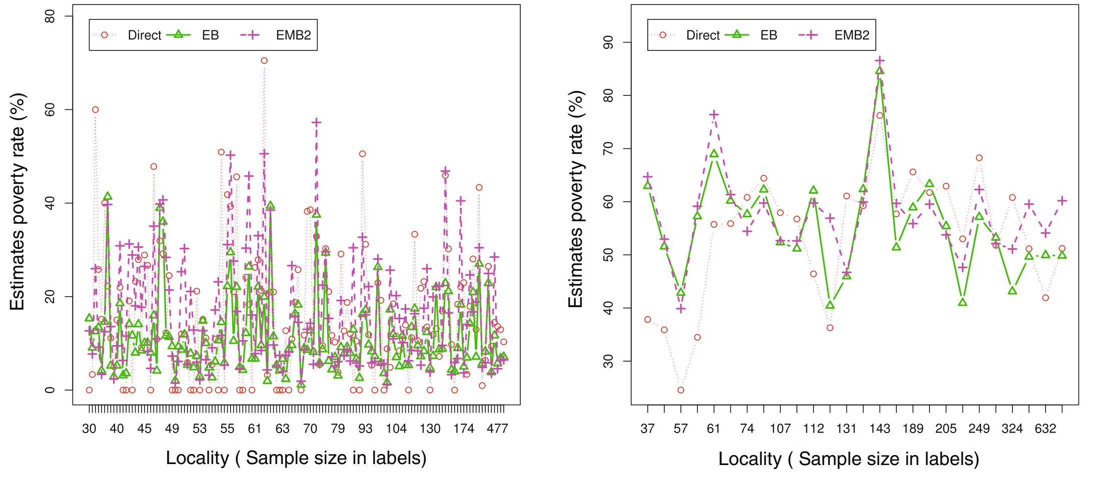


FIGURE 8 Direct, Census empirical best and Census EMB2 estimates of poverty rates for West Bank (left) and Gaza (right) [Colour figure can be viewed at wileyonlinelibrary.com]

Let us now analyse the quality of the estimates. Note that the MSE estimates of model-based estimators are obtained under the assumed model; in this case, they are obtained under the MM. Table 1 reports summaries of the estimated CVs of the direct, Census EB and Census EMB1 and EMB2 estimators of mean expenditure, obtained under the assumed MM model, and also the estimated CVs of the Census EB estimators, obtained under the NER model (using the analogous parametric bootstra procedure). Here, CVs are actually equal to the root MSEs (RMSEs) divided by the corresponding estimate of the mean expenditure (×100). Table 2 reports the analogous summaries for the RRMSEs of the estimators of the poverty rates.

From these tables, we can draw the following conclusions. If data follows the MM but we ignore it and obtain the usual Census EB estimators based on the NE model, their correctly estimated CVs or RMSEs are larger than one would expect if the NE model was true (compare column for 'CV(EB)' versus 'CV(EB) under NER' in Table 1 and the analogous columns in Table 2), and these values are clearly larger than those obtained for the Census EMB1 and EMB2 estimators. The latter estimators based on the MM model are both more efficient than the Census EB ones in this case, and EMB2 ones, based on the weighted average, are slightly better than EMB1. In fact, the average percent MSE reduction of the EMB2 estimator with respect to the EB one is 75.5% for the mean expenditure, 72.9% for the poverty rate and 70.8% for the poverty gap. Hence, the normality-based NE model should be carefully checked when using the Census EB

TABLE 1 Summaries of estimated CVs (%) of Direct, Census empirical best (EB), EMB1 and EMB2 estimators of locality mean expenditure under the multivariate mixture model, and of estimated CV of Census EB under the NER model

	CV (Direct)	CV (EB)	CV (EB) under NER	CV (EMB1)	CV (EMB2)
Min.	3.426	2.450	1.692	1.760	1.760
First Qu.	13.312	7.505	3.543	3.867	3.845
Median	18.265	13.830	4.247	5.038	4.938
Mean	18.232	13.470	4.333	6.031	5.694
Third Qu.	23.496	18.787	5.143	7.356	6.659
Max.	40.197	25.772	8.279	19.578	17.315

TABLE 2 Summaries of estimated relative root MSEs (RRMSEs) (×100) of Direct, Census empirical best (EB), EMB1 and EMB2 estimators of locality poverty rates under the MM model, and of estimated RRMSE of Census EB under the NER model

	RRMSE (Direct)	RRMSE (EB)	RRMSE (EB) under NER	RRMSE (EMB1)	RRMSE (EMB2)
Min.	2.946	1.215	0.545	0.718	0.718
First Qu.	7.101	4.304	1.415	2.103	2.093
Median	9.552	6.760	1.807	2.859	2.830
Mean	9.966	6.800	1.926	3.185	3.024
Third Qu.	11.939	8.742	2.249	3.927	3.660
Max.	25.071	21.502	4.157	12.157	10.173

estimators and their usual parametric bootstrap MSEs, since these are calculated assuming that the NE model actually holds. If the correct model is the MM one, then the true MSEs will be underestimated. Hence, EMB1 or EMB2 estimators and their corresponding MSE estimators based on the MM model are a more safe alternative.

Detailed CV estimates for each locality are shown in Figure 9 for mean expenditure, in West Bank (left) and Gaza (right). We can see that the estimated CVs for the Census EMB2 estimators are smaller than those of Census EB and the direct estimators, for almost all the localities in Gaza and in West Bank. Moreover, the tendency shown by the estimated CVs of direct and Census EB estimates as the locality sample size increases is somehow erratic, whereas the estimated CVs of the EMB2 decrease with the locality sample size in a more smooth and reasonable way, for both regions. Figure 10 shows estimated RMSE for the poverty rates, and the conclusions obtained

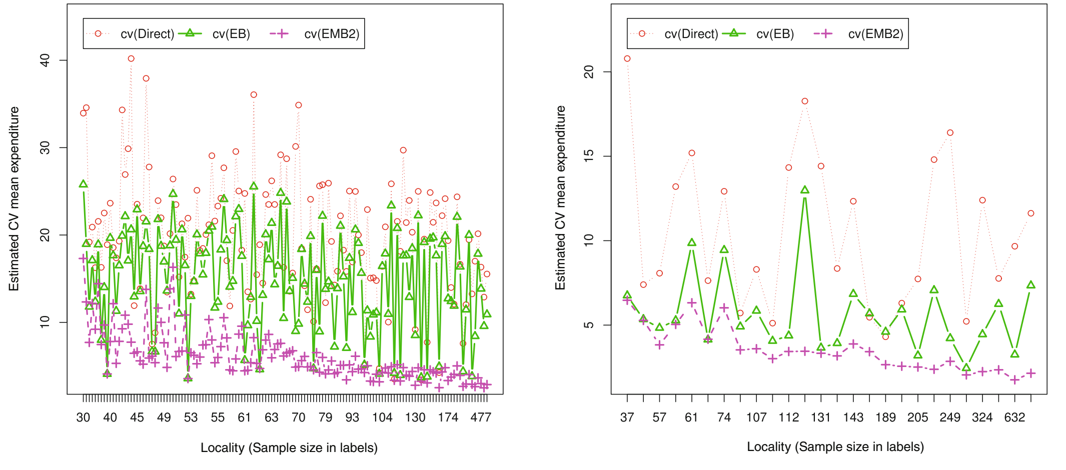


FIGURE 9 Estimated CV (%) of Direct, Census empirical best and Census EMB2 estimators of mean expenditure for localities in West Bank (left) and Gaza (right) [Colour figure can be viewed at wileyonlinelibrary.com]

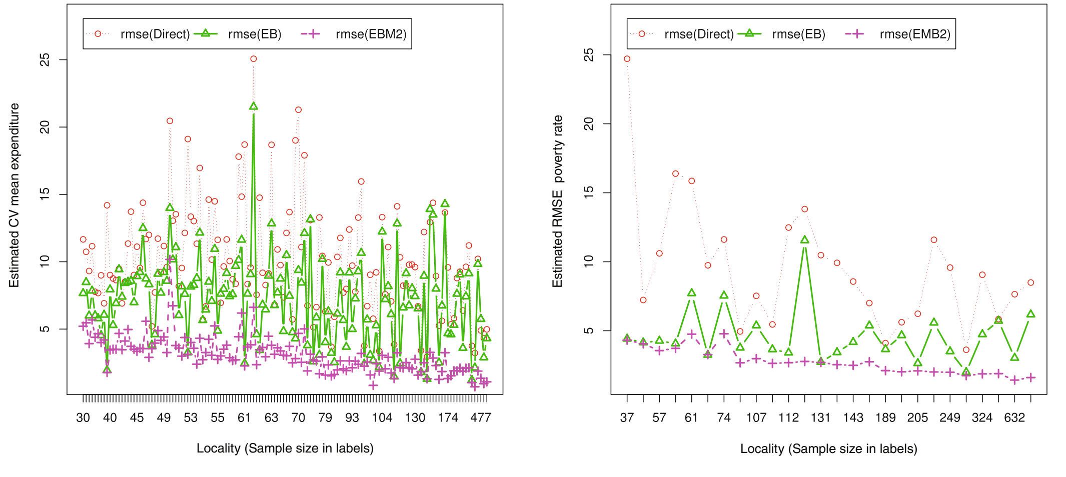


FIGURE 10 Estimated RMSE (×100) of Direct, Census empirical best and Census EMB2 estimators of mean expenditure for localities in West Bank (left) and Gaza (right) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Summaries of EMB2 estimates of mean expenditure (ILS), poverty rate (%) and poverty gap (%) for the localities in West Bank and Gaza

		Min.	First Qu.	Median	Mean	Third Qu.	Max.
Mean	West Bank	10,241	15,014	18,774	18,971	22,304	36,383
expenditure	Gaza	6,753	9,887	10,259	10,265	10,814	12,695
Poverty	West Bank	1.24	6.13	12.58	16.07	24.71	57.25
rate	Gaza	39.89	52.75	58.02	57.68	59.89	86.56
Poverty	West Bank	0.20	1.10	2.47	3.62	5.51	17.26
gap	Gaza	10.74	16.39	18.51	18.90	20.13	36.04

from them are similar. They are also similar for the poverty gaps (plots not shown for brevity), although differences seem smaller for this indicator.

Finally, Table 3 reports summaries of the resulting EMB2 estimates of mean expenditure, poverty rates and gaps in the Palestinian localities. We can see the great gap between the estimates for the two regions, with a median estimated mean expenditure of 18,774 ILS for West Bank, compared with 10,259 ILS for Gaza, median estimated poverty rate of 12.58% for West Bank compared to 58.02% for Gaza and 2.47% estimated poverty gap for West Bank, compared to 18.51% for Gaza.

8 | CONCLUSIONS

The proposed predictors based on the MM model are a compromise between the fully parametric EB estimators based on the NE model and the completely nonparametric direct estimators. Even though the MM model accounts for heterogeneity of the areas (but not of the units within the areas) and may then be applied to model area-level outliers, it is more general than mixture models that account for heterogeneity only in some of the variances, because it considers area heterogeneity in both variance components and in the regression coefficients as well. Hence, the MM represents a flexible mixture model and it is identifiable.

Even though the mixture components are based on the Normal distribution, mixtures can represent a large family of distributions. We have described the specific E-M fitting algorithm for the proposed MM model and defined two new small area estimators based on this model, which achieve promising results. This methodology has been developed for general number of mixtures K, but in many applications, $K = 2$ might be enough. Practitioners may consider two strategies to select the number of mixture components and obtain a model that fits the data. One strategy is going from the simplest to the most complex model; that is, starting with the NER model (fitted with the traditional fitting algorithms), a model with increasing K is successively fitted (with the proposed algorithm), applying diagnostic tools in each step, until the model with the smallest number of components whose diagnostics are adequate is selected. Conversely, the strategy complex-to-simple can be applied. Starting from a large K, the selection of the number of mixture components might be based on goodness-of-fit measures, on the significance of the estimated probabilities of the mixture components and on diagnostic tools.

In our simulation experiments, we have also seen that EB estimators based on the NE model will not perform as expected if the normality and homogeneity assumptions fail, perhaps due to the presence of outlying areas. If data follow the MM model, then EB estimators based on the NE model will be less efficient than expected, and might be substantially biased for complex indicators. In that case, the proposed estimators based on the MM do not show bias and are considerably more efficient. These conclusions seem to be supported by the results of our application to poverty mapping in Palestinian localities, where EB estimates for the localities in West Bank seem to be slightly biased, whereas the estimators based on the MM do not display systematic bias, and achieve smaller estimated MSEs for practically all the localities, with considerable average gains in (estimated) efficiency with respect to EB.

Such as the original EB method of Molina and Rao (2010), the EMB1 and EMB2 estimators do not account for the survey design. Hence, they may be biased under complex survey designs, specially when the design is informative, in the sense that the probability of selection of the sample depends on the particular outcomes, even after conditioning on the values of the available auxiliary variables. However, as in the definition of the Pseudo EB estimator of Guadarrama et al. (2018), it holds that the distribution of $\mathbf{y}_{dr}|\mathbf{y}_{ds}, \mathbf{z}_d = \mathbf{e}_k$ equals that of $\mathbf{y}_{dr}|\bar{y}_{ds}, \mathbf{z}_d = \mathbf{e}_k$, where $\bar{y}_{ds} = n_d^{-1}\sum_{j\in s_d} y_{dj}$ is the sample mean in area d. Hence, the EMB1 and EMB2 predictors may be extended to account for the survey design similarly as the Pseudo EB predictor, by defining them in terms of the conditional distribution of $\mathbf{y}_{dr}|\bar{y}_{d\omega}, \mathbf{z}_d = \mathbf{e}_k$, where $\bar{y}_{d\omega} = \left\{\sum_{j\in s_d}\omega_{dj}\right\}^{-1}\sum_{j\in s_d}\omega_{dj}y_{dj}$ is the weighted sample mean that uses the survey weights ω_{dj} of the sample units. It remains to find a design-based alternative to the posterior probabilities p_{dk}, we leave this for further research.

ACKNOWLEDGEMENTS

We wish to thank Marwan Kawaja from the United Nations Economic and Social Commission for Western Asia (UN-ESCWA) and Jawad AL-Saleh from the Palestinian Central Bureau of Statistics (PCBS), for providing and preparing the data used in the application. This work was partially supported by the grants PGC2018-096840-B-I00 from the Spanish Ministerio de Economía y Competitividad, PID2020-115598RB-I00 from the Spanish Ministerio de Ciencia e Innovación and by the grant prometeo/2021/063 from the Generalitat Valenciana.

ORCID

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

REFERENCES

Battese, G.E., Harter, R.M. & Fuller, W.A. (1988) An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28–36.

Bell, W.R. & Huang, E.T. (2006) Using the t-distribution to deal with outliers in small area estimation. *Survey Methodology*, 44(1), 101–115.

Chakraborty, A., Datta, G.S. & Mandal, A. (2016) A two-component normal mixture alternative to the Fay-Herriot model. *Statistics in Transition new series and Survey Methodology Joint Issue: Small Area Estimation*, 2014(17), 67–90.

Corral, P., Molina, I. & Nguyen, M. (2021) Pull your small area estimates up by the bootstraps. *Journal of Statistical Computation and Simulation*, 91(16), 3304–3357. https://doi.org/10.1080/00949655.2021.1926460

Correa, L., Molina, I. & Rao, J.N.K. (2012) *Comparison of methods for estimation of poverty indicators in small areas*. Unpublished report.

Datta, G.S. & Ghosh, M. (1991) Bayesian prediction in linear models: applications to small area estimation. *Annals of Statistics*, 19, 1748–1770.

Elbers, C., Lanjouw, J.O. & Lanjouw, P. (2003) Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355–364.

Elbers, C. & Van der Weide, R. (2014) *Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality*. World Bank Policy Research Working Paper: 6962.

Foster, J., Greer, J. & Thorbecke, E. (1984) A class of decomposable poverty measures. *Econometrica*, 52, 761–766.

Gershunskaya, J. (2010) *Robust small area estimation using a mixture model*. Section on Survey Research Methods - JSM.

Gershunskaya, J. & Savitsky, T.D. (2020) Model-based screening for robust estimation in the presence of deviations from linearity in small domain models. *Journal of Survey Statistics and Methodology*, 8(2), 181–205.

Ghosh, M., Myung, J. & Moura, F.A.S. (2018) Robust Bayesian small area estimation. *Survey Methodology*, 44(1), 101–115.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. & Santamaría, L. (2008) Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78(5), 443–462.

Guadarrama, M., Molina, I. & Rao, J.N.K. (2018) Small area estimation of general parameters under complex sampling designs. *Computational Statistics and Data Analysis*, 121, 20–40.

Guadarrama, M., Molina, I. & Tillé, Y. (2020) Small area estimation methods under cut-off sampling. *Survey Methodology*, 46(1), 51–75.

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.

McLachlan, G. & Peel, D. (2000) *Finite mixture models*. New York: Wiley.

Molina, I. (2019) *Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas*. Series of the Economic Commission for Latin America and the Caribbean (ECLAC) from United Nations, Estudios Estadísticos LC/TS.2018/82/Rev.1, CEPAL.

Molina, I. & Martín, N. (2018) Empirical best prediction under a nested error model with log transformation. *The Annals of Statistics*, 46(5), 1961–1993.

Molina, I. & Rao, J.N.K. (2010) Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38, 369–385.

Molina Peralta, I. & García-Portugués, E. (2020) *A short guide for small area estimation in household surveys*. Illustration to poverty mapping in Palestine with expenditure survey and census data, Series of the UN Economic and Social Commission for Western Asia (ESCWA), to appear.

Morales, D., Esteban, M.D., Pérez, A. & Hobza, T. (2021) *A course on small area estimation and mixed models, methods, theory and applications in R*. Cham: Springer.

Pfeffermann, D. & Sverchkov, M. (2007) Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102(480), 1427–1439.

Rao, J.N.K. & Molina, I. (2015) *Small area estimation*, 2nd edition. Hoboken, NJ: Wiley.

Rojas-Perilla, N., Pannier, S., Schmid, T. & Tzavidis, N. (2020) Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society, Series A*, 183(1), 121–148.

Schoch, T. (2014) *rsae: robust small area estimation*. R package version 0.1-5.

Sinha, K.S. & Rao, J.N.K. (2009) Robust small area estimation. *The Canadian Journal of Statistics*, 37(3), 381–399.

Tang, X., Ghosh, M., Ha, N.S. & Sedransk, J. (2018) Modeling random effects using global-local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 113(1), 1476–1489.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Bikauskaite, A., Molina, I. & Morales, D. (2022) Multivariate mixture model for small area estimation of poverty indicators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(Suppl. 2), S724–S755. Available from: https://doi.org/10.1111/rssa.12965

APPENDIX A. E-M FITTING ALGORITHM FOR THE MM MODEL

This appendix describes the calculations and final steps of the E-M algorithm (see e.g. McLachlan & Peel, 2000) that gives ML estimators of the parameters of the MM proposed in Section 4, $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\pi}}', \hat{\boldsymbol{\theta}}_1', \dots, \hat{\boldsymbol{\theta}}_K')'$. The MM model is fit to the sample vector composed by the sample parts of each area, $\mathbf{y}_s = (\mathbf{y}_{1s}', \dots, \mathbf{y}_{Ds}')'$. In absence of sample selection bias, $\mathbf{y}_{ds}$ follows the same mixture model (12) as $\mathbf{y}_d$, but for the sample elements, that is,

$$\mathbf{y}_{ds} \overset{ind.}{\sim} \sum_{k=1}^{K} \pi_k N(\mathbf{X}_{ds}\boldsymbol{\beta}_k, \mathbf{V}_{kds}), \quad d = 1, \dots, D,$$

where $\mathbf{V}_{kds} = \tau_k^2 \mathbf{1}_{n_d}\mathbf{1}_{n_d}' + \sigma_k^2 \mathbf{I}_{n_d}$, $k = 1, \dots, K$. Turning to the latent cluster indicator variables $\mathbf{z}_d$, we have that $(\mathbf{y}_{ds}', \mathbf{z}_d')'$ are also independent for $d = 1, \dots, D$, with $\mathbf{y}_{ds}|\mathbf{z}_d = \mathbf{e}_k \sim N(\mathbf{X}_{ds}\boldsymbol{\beta}_k, \mathbf{V}_{kds})$, $k = 1, \dots, K$. Then, the marginal p.d.f. of $\mathbf{y}_{ds}$ is given by

$$f_d(\mathbf{y}_{ds}) = \sum_{k=1}^{K} \pi_k f_{kd}(\mathbf{y}_{ds}).$$

where $f_{kd}(\mathbf{y}_{ds})$ denotes the (multivariate normal) density of $\mathbf{y}_{ds}|\mathbf{z}_d = \mathbf{e}_k$.

The complete data for area d is $(\mathbf{y}_{ds}', \mathbf{z}_d')'$, where $\mathbf{z}_d = (z_{d1}, \dots, z_{dK})'$. The complete data likelihood for area d is then obtained as the conditional given $\mathbf{z}_d$ times the marginal of $\mathbf{z}_d$, which yields

$$f_d(\mathbf{y}_{ds}, \mathbf{z}_d) = \prod_{k=1}^{K} \{\pi_k f_{kd}(\mathbf{y}_{ds})\}^{z_{dk}},$$

Now consider the vector with the sample elements for all areas, $\mathbf{y}_s = (\mathbf{y}_{1s}', \dots, \mathbf{y}_{Ds}')'$ and the vector with the latent variables $\mathbf{z} = (\mathbf{z}_1', \dots, \mathbf{z}_D')'$. Then, the complete data likelihood for all areas is

$$f(\mathbf{y}_s, \mathbf{z}) = \prod_{d=1}^{D} f(\mathbf{y}_{ds}, \mathbf{z}_d) = \prod_{d=1}^{D}\prod_{k=1}^{K} \{\pi_k f_{kd}(\mathbf{y}_{ds})\}^{z_{dk}}.$$

The complete data log-likelihood is then $\log f(\mathbf{y}_s, \mathbf{z}) = \sum_{d=1}^{D} \log f(\mathbf{y}_{ds}, \mathbf{z}_d)$, where

$$\log f(\mathbf{y}_{ds}, \mathbf{z}_d) = \sum_{k=1}^{K} z_{kd}\{\log \pi_k + \log f_{kd}(\mathbf{y}_{ds})\}.$$

In practice, $f_{kd}(\mathbf{y}_{ds})$ depends on $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \tau_k^2, \sigma_k^2)$, that is, $f_{kd}(\mathbf{y}_{ds}) = f_{kd}(\mathbf{y}_{ds}; \boldsymbol{\theta}_k)$, $k = 0, 1$, and the mixing proportions, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)'$, are also unknown. For the vector of unknown model parameters $\boldsymbol{\psi} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_K')'$, let $\boldsymbol{\psi}^{(r)} = (\boldsymbol{\pi}^{(r)}, \boldsymbol{\theta}_1^{(r)'}, \dots, \boldsymbol{\theta}_K^{(r)'})'$ be the estimates at the rth step of the fitting algorithm. The variables $\mathbf{z}_d$, given $\mathbf{y}_{sd}$, are independent for all areas $d = 1, \dots, D$, and the probability of $\mathbf{z}_d = \mathbf{e}_k$ given $\mathbf{y}_{sd}$ (posterior probability of $\mathbf{z}_d = \mathbf{e}_k$), is given for $k = 1, \dots, K$, and step r, by

$$p_{kd}^{(r)} = P_{\boldsymbol{\psi}^{(r)}}(\mathbf{z}_d = \mathbf{e}_k | \mathbf{y}_{ds}) = \frac{f(\mathbf{y}_{ds}|\mathbf{z}_d = \mathbf{e}_k)P(\mathbf{z}_d = \mathbf{e}_k)}{\sum_{\ell=1}^K f(\mathbf{y}_{ds}|\mathbf{z}_d = \mathbf{e}_\ell)P(\mathbf{z}_d = \mathbf{e}_\ell)} = \frac{\pi_k^{(r)} f_{kd}(\mathbf{y}_{ds}; \boldsymbol{\theta}_k^{(r)})}{\sum_{\ell=1}^K \pi_\ell^{(r)} f_{\ell d}(\mathbf{y}_{ds}; \boldsymbol{\theta}_k^{(r)})}.$$

At the $(r+1)$th iteration, the E-step consists of evaluating the conditional expectation of $\log f(\mathbf{y}_s, \mathbf{z})$ given the sample data $\mathbf{y}_s$, evaluated at the values of the parameters in the rth iteration. This conditional expectation is given by

$$Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(r)}) = \sum_{d=1}^D \sum_{k=1}^K \left[\log \pi_k + \log f_{kd}(\mathbf{y}_{ds})\right] p_{kd}^{(r)}.$$

In the M-step, we maximise $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(r)})$ with respect to $\boldsymbol{\psi}$. Noting that $\pi_K = 1 - \pi_1 - \cdots - \pi_{K-1}$ and $p_{Kd}^{(r)} = 1 - p_{1d}^{(r)} - \cdots - p_{K-1,d}^{(r)}$, and taking derivative with respect to π_k, we obtain

$$\frac{\partial Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(r)})}{\partial \pi_k} = \frac{\sum_{d=1}^D p_{kd}^{(r)}}{\pi_k} - \frac{D - \sum_{d=1}^D \sum_{\ell=1}^K p_{\ell d}^{(r)}}{1 - \sum_{\ell=1}^K \pi_\ell} = 0, \quad k = 1, \dots, K,$$

which leads to the updating equation for π_k,

$$\pi_k^{(r+1)} = \frac{1}{D} \sum_{d=1}^D p_{kd}^{(r)} = \frac{1}{D} \sum_{d=1}^D E_{\boldsymbol{\psi}^{(r)}}(z_{kd}|\mathbf{y}_{ds}).$$

Observe that this result is intuitive. If z_{kd} were observable, $d = 1, \dots, D$, then the ML estimator of π_k would be $\hat{\pi}_k = \bar{z}_k = D^{-1}\sum_{d=1}^D z_{kd}$. In the E-M method, $\bar{z}_k$ is replaced by its current conditional expectation. For the remaining parameters, we get

$$\frac{\partial Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(r)})}{\partial \boldsymbol{\theta}_k} = \sum_{d=1}^D p_{kd}^{(r)} \frac{\partial \log f_{kd}(\mathbf{y}_{ds}; \boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k}, \quad k = 1, \dots, K. \tag{A1}$$

Solving for $\boldsymbol{\theta}_k$ in the system $\partial Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(r)})/\partial \boldsymbol{\theta}_k = \mathbf{0}$, we obtain the new $\boldsymbol{\theta}_k^{(r+1)}$, $k = 1, \dots, K$. The equation system $\partial Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(r)})/\partial \boldsymbol{\theta}_k = \mathbf{0}$ is typically non-linear, so the solution $\boldsymbol{\theta}_k^{(r+1)}$, $k = 1, \dots, K$, can be found using iterative algorithms for solving non-linear equations.

We now spell out the expressions of the derivatives $\partial \log f_{kd}(\mathbf{y}_{ds}; \boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k$ involved in the updating equations. The multivariate normal density function of $\mathbf{y}_{ds}|\mathbf{z}_d = \mathbf{e}_k$ is

$$f_{kd}(\mathbf{y}_{ds}; \boldsymbol{\theta}_k) = (2\pi)^{-\frac{n_d}{2}} |\mathbf{V}_{kds}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta}_k)'\mathbf{V}_{kds}^{-1}(\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta}_k)\right\},$$

where $\mathbf{V}_{kds} = \tau_k^2 \mathbf{1}_{n_d}\mathbf{1}_{n_d}' + \sigma_k^2 \mathbf{I}_{n_d}$, $k = 1, \dots, K$. The corresponding log-density is then

$$\log f_{kd}(\mathbf{y}_{ds}; \boldsymbol{\theta}_k) = -\frac{n_d}{2}\log 2\pi - \frac{1}{2}\log |\mathbf{V}_{kds}| - \frac{1}{2}(\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta}_k)'\mathbf{V}_{kds}^{-1}(\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta}_k).$$

Applying the following formulas

$$\frac{\partial \log|\mathbf{V}(\theta)|}{\partial\theta} = \text{tr}\left\{\mathbf{V}(\theta)^{-1}\frac{\partial\mathbf{V}(\theta)}{\partial\theta}\right\}, \quad \frac{\partial\mathbf{V}^{-1}(\theta)}{\partial\theta} = -\mathbf{V}^{-1}(\theta)\frac{\partial\mathbf{V}(\theta)}{\partial\theta}\mathbf{V}^{-1}(\theta),$$

we obtain the derivatives of $\log f_{kd}$ with respect to $\boldsymbol{\beta}_k$, τ_k^2 and σ_k^2, which are given by

$$\begin{aligned}
\frac{\partial \log f_{kd}}{\partial\boldsymbol{\beta}_k} &= \mathbf{X}'_{ds}\mathbf{V}^{-1}_{kds}(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k),\\
\frac{\partial \log f_{kd}}{\partial\tau_k^2} &= -\frac{1}{2}\text{tr}\left(\mathbf{V}^{-1}_{kds}\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\right) + \frac{1}{2}(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k)'\mathbf{V}^{-1}_{kds}\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\mathbf{V}^{-1}_{kds}(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k),\\
\frac{\partial \log f_{kd}}{\partial\sigma_k^2} &= -\frac{1}{2}\text{tr}\left(\mathbf{V}^{-1}_{kds}\right) + \frac{1}{2}(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k)'\mathbf{V}^{-1}_{kds}\mathbf{V}^{-1}_{kds}(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k).
\end{aligned} \tag{A2}$$

The inverse of $\mathbf{V}_{kds}$ can be obtained applying the inversion formula

$$(\mathbf{A}+\mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}, \tag{A3}$$

with $\mathbf{A} = \sigma_k^2\mathbf{I}_{n_d}$, $\mathbf{u} = \tau_k^2\mathbf{1}_{n_d}$ and $\mathbf{v}' = \mathbf{1}'_{n_d}$, leading to

$$\mathbf{V}^{-1}_{kds} = \frac{1}{\sigma_k^2}\left(\mathbf{I}_{n_d} - \frac{\gamma_{kd}}{n_d}\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\right). \tag{A4}$$

Using this expression for the inverse, we can spell out the following terms involved in (A2),

$$\begin{aligned}
\text{tr}\left(\mathbf{V}^{-1}_{kds}\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\right) &= \mathbf{1}'_{n_d}\mathbf{V}^{-1}_{kds}\mathbf{1}_{n_d} = \sigma_k^{-2}\left(\mathbf{1}'_{n_d}\mathbf{1}_{n_d} - \frac{\gamma_{kd}}{n_d}\mathbf{1}'_{n_d}\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\mathbf{1}_{n_d}\right) = \frac{\gamma_{dk}}{\tau_k^2},\\
\mathbf{1}'_{n_d}\mathbf{V}^{-1}_{kds} &= \frac{1}{\sigma_k^2}(\mathbf{1}'_{n_d} - \gamma_{kd}\mathbf{1}'_{n_d}) = \frac{1-\gamma_{kd}}{\sigma_k^2}\mathbf{1}'_{n_d} = \frac{1}{\sigma_k^2+n_d\tau_k^2}\mathbf{1}'_{n_d},\\
\mathbf{V}^{-1}_{kds}\mathbf{V}^{-1}_{kds} &= \frac{1}{\sigma_k^4}\left\{\mathbf{I}_{n_d} + \frac{\gamma_{kd}}{n_d}(\gamma_{kd}-2)\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\right\},\\
\mathbf{V}^{-1}_{kds}\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\mathbf{V}^{-1}_{kds} &= \frac{(1-\gamma_{kd})^2}{\sigma_k^4}\mathbf{1}_{n_d}\mathbf{1}'_{n_d} = \frac{1}{(\sigma_k^2+n_d\tau_k^2)^2}\mathbf{1}_{n_d}\mathbf{1}'_{n_d} = \frac{\gamma_{dk}^2}{n_d^2\tau_k^4}\mathbf{1}_{n_d}\mathbf{1}'_{n_d},\\
\text{tr}\left(\mathbf{V}^{-1}_{kds}\right) &= \text{tr}\left\{\frac{1}{\sigma_k^2}\left(\mathbf{I}_{n_d} - \frac{\gamma_{kd}}{n_d}\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\right)\right\} = \frac{n_d-\gamma_{kd}}{\sigma_k^2} = \frac{n_d\left\{\sigma_k^2+(n_d-1)\tau_k^2\right\}}{\sigma_k^2(\sigma_k^2+n_d\tau_k^2)}.
\end{aligned}$$

Replacing these formulas in (A2), we finally obtain the following expressions:

$$\begin{aligned}
\frac{\partial \log f_{kd}}{\partial\boldsymbol{\beta}_k} &= \mathbf{X}'_{ds}\frac{1}{\sigma_k^2}\left(\mathbf{I}_{n_d} - \frac{\gamma_{kd}}{n_d}\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\right)(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k),\\
\frac{\partial \log f_{kd}}{\partial\tau_k^2} &= -\frac{1}{2}\frac{\gamma_{dk}}{\tau_k^2} + \frac{1}{2}\frac{\gamma_{dk}^2}{n_d^2\tau_k^4}(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k)'\mathbf{1}_{n_d}\mathbf{1}'_{n_d}(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k),\\
\frac{\partial \log f_{kd}}{\partial\sigma_k^2} &= -\frac{n_d-\gamma_{kd}}{2\sigma_k^2} + \frac{1}{2}(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k)'\frac{1}{\sigma_k^4}\left(\mathbf{I}_{n_d} + \frac{\gamma_{kd}}{n_d}(\gamma_{kd}-2)\mathbf{1}_{n_d}\mathbf{1}'_{n_d}\right)(\mathbf{y}_{ds}-\mathbf{X}_{ds}\boldsymbol{\beta}_k).
\end{aligned}$$

Replacing these derivatives in (A1) and equating to zero, we obtain the updating equations given in (A5) of the E-M algorithm described below.

E-M algorithm for ML fitting of the MM model:

1. Set the initial values $r = 0$ and $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\pi}^{(0)}, \boldsymbol{\theta}_1^{(0)'}, \dots, \boldsymbol{\theta}_K^{(0)'})'$.
2. Repeat the following steps till convergence
   a. Update the estimates of $\boldsymbol{\beta}_k$, τ_k^2, σ_k^2 and π_k, $k = 1, \dots, K$, with the equations

$$
\begin{aligned}
\boldsymbol{\beta}_k^{(r+1)} &= \left\{ \sum_{d=1}^{D} p_{kd}^{(r)} \mathbf{X}_{ds}' \frac{1}{\sigma_k^{2(r)}} \left(\mathbf{I}_{n_d} - \frac{\gamma_{kd}^{(r)}}{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}' \right) \mathbf{X}_{ds} \right\}^{-1} \\
&\quad \times \sum_{d=1}^{D} p_{kd}^{(r)} \mathbf{X}_{ds}' \frac{1}{\sigma_k^{2(r)}} \left(\mathbf{I}_{n_d} - \frac{\gamma_{kd}^{(r)}}{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}' \right) \mathbf{y}_{ds}, \\
\tau_k^{2(r+1)} &= \frac{\sum_{d=1}^{D} p_{kd}^{(r)} n_d^{-2} \gamma_{kd}^{2(r)} \left(\mathbf{y}_{ds} - \mathbf{X}_{ds} \boldsymbol{\beta}_k^{(r)} \right)' \mathbf{1}_{n_d} \mathbf{1}_{n_d}' \left(\mathbf{y}_{ds} - \mathbf{X}_{ds} \boldsymbol{\beta}_k^{(r)} \right)}{\sum_{d=1}^{D} p_{kd}^{(r)} \gamma_{kd}^{(r)}}, \\
\sigma_k^{2(r+1)} &= \frac{\sum_{d=1}^{D} p_{kd}^{(r)} \left(\mathbf{y}_{ds} - \mathbf{X}_{ds} \boldsymbol{\beta}_k^{(r)} \right)' \left\{ \mathbf{I}_{n_d} + \frac{\gamma_{kd}^{(r)}}{n_d} \left(\gamma_{kd}^{(r)} - 2 \right) \mathbf{1}_{n_d} \mathbf{1}_{n_d}' \right\} \left(\mathbf{y}_{ds} - \mathbf{X}_{ds} \boldsymbol{\beta}_k^{(r)} \right)}{\sum_{d=1}^{D} p_{kd}^{(r)} (n_d - \gamma_{kd}^{(r)})}, \\
\pi_k^{(r+1)} &= \frac{1}{D} \sum_{d=1}^{D} p_{kd}^{(r)}.
\end{aligned}
\tag{A5}
$$

   b. Update the iteration index $r \leftarrow r + 1$.
3. Output: $\boldsymbol{\psi}^{(r+1)}$.

The ML estimate of $\boldsymbol{\psi}$ is the value $\hat{\boldsymbol{\psi}}$ obtained in the last iteration after convergence.

We finally remark that $\boldsymbol{\beta}_k = \boldsymbol{\beta}$, $\tau_k^2 = \tau^2$ and $\sigma_k^2 = \sigma^2$, $k = 1, \dots, K$, with $\pi_1 > \pi_2 > \cdots > \pi_K$, yields the NER model; hence, the NER model is a non-identifiable MM model. Due to non-identifiability, the E-M algorithm encounters numerical problems when trying to fit a MM model to data generated by a NER model. For this reason, we do not recommend to use the SAE methodology based on a MM introduced here when the data obey a NER model instead of an identifiable MM model, this information should be taken into account by practitioners.

This kind of 'nonidentifiability due to overfitting', and the related numerical problems of the E-M algorithm, may occur in a more general situation, where a mixture of K components is fitted to data generated from a mixture of $G < K$ components. As this more general situation was not investigated in our simulation experiments, we suggest applying the proposed E-M algorithm in a 'supervised' manner with a plausible value of K to uncover the actual latent structure, assuming some knowledge about the number of components. One way of obtaining information about the possible number of components K is to check the distribution of the fitted area effects of the NER model (e.g. with histograms or q-q normal plots), trying to guess the number of latent clusters of the area effects.

APPENDIX B. MODEL FITTING RESULTS IN THE APPLICATION

TABLE B1 Estimated parameters from fitting the nested error (NE) and the multivariate mixture (MM) model, with SEs

	NE		MM ($k = 1$)		MM ($k = 2$)	
Parameter	**Coef.**	**SE**	**Coef.**	**SE**	**Coef.**	**SE**
(Intercept)	9.3917	0.0394	9.3874	0.0475	9.3837	0.0583
regionWest Bank	0.1637	0.0301	0.1513	0.0380	0.2343	0.0468
loctype2.f3	−0.1003	0.0332	−0.0960	0.0442	−0.0937	0.0493
hhsize_ae	−0.1237	0.0029	−0.1279	0.0025	−0.1144	0.0035
femalep	−0.1043	0.0180	−0.0910	0.0195	−0.1247	0.0237
head_age	0.0015	0.0003	0.0025	0.0003	−0.0001	0.0004
head_refugstatRegistered refugee	−0.0167	0.0079	−0.0205	0.0082	−0.0257	0.0108
head_refugstatUn-registered refugee	0.0858	0.0597	−0.2052	0.0308	0.1814	0.0427
head_diffYes	−0.1264	0.0161	−0.0926	0.0160	−0.1790	0.0202
head_neverschoolYes	−0.0961	0.0208	−0.0581	0.0192	−0.2077	0.0251
head_secondaboveYes	0.0434	0.0099	0.0959	0.0081	−0.0246	0.0105
employed_ratio	0.1878	0.0178	0.1879	0.0199	0.2061	0.0262
head_unemployed.f1	−0.1136	0.0174	−0.0739	0.0173	−0.1643	0.0227
head_employisrasett.f1	0.1064	0.0128	0.0322	0.0092	0.1708	0.0132
head_employnatgov.f1	0.0280	0.0095	0.0644	0.0080	−0.0246	0.0107
dwelltype2.fVilla	0.2415	0.0658	0.4815	0.0378	0.0023	0.0485
tenure2.fRented	−0.0791	0.0143	−0.0548	0.0143	−0.0848	0.0182
rooms	0.0483	0.0039	0.0270	0.0030	0.0771	0.0041
water_bottled	0.2203	0.0589	−0.0879	0.0252	0.3655	0.0315
heating2.fDiesel	0.4899	0.0814	0.4994	0.0830	0.6355	0.1021
heating2.fElectricity	−0.0466	0.0234	−0.1286	0.0175	0.0747	0.0235
heating2.fGas	−0.0418	0.0233	−0.1287	0.0180	0.0792	0.0246
heating2.fOther/NA	−0.1348	0.0215	−0.1665	0.0186	−0.0798	0.0253
heating2.fWood	−0.0609	0.0211	−0.0868	0.0190	0.0006	0.0243
freezer_ysno.f1	0.0732	0.0101	0.0839	0.0099	0.0498	0.0124
microwave_ysno.f1	0.0253	0.0074	0.0250	0.0076	0.0388	0.0094
dishwasher_ysno.f1	0.0612	0.0198	−0.0178	0.0176	0.1620	0.0256
tv_ledlcd_ysno.f1	0.1222	0.0085	0.1490	0.0072	0.0649	0.0095
electric_fan_ysno.f1	0.0711	0.0113	0.0893	0.0101	0.0363	0.0133

(Continues)

TABLE B1 (Continued)

	NE		MM ($k = 1$)		MM ($k = 2$)	
Parameter	**Coef.**	**SE**	**Coef.**	**SE**	**Coef.**	**SE**
air_conditioner_ysno.f1	0.1342	0.0090	0.1573	0.0089	0.1152	0.0106
central_heating_ysno.f1	−0.0954	0.0776	0.1980	0.0399	−0.1838	0.0567
solar_boiler_ysno.f1	0.0274	0.0069	0.0204	0.0069	0.0275	0.0092
phone_line_ysno.f1	0.0723	0.0072	0.0738	0.0076	0.0740	0.0098
home_library_ysno.f1	0.0722	0.0140	0.1209	0.0100	0.0048	0.0135
computer_ysno.f1	0.0685	0.0070	0.0802	0.0074	0.0518	0.0089
ipad_tablet_ysno.f1	0.0816	0.0079	0.0669	0.0076	0.1048	0.0095
smartphone_ysno.f1	0.1223	0.0161	0.1863	0.0097	−0.0176	0.0124
washing_machine_ysno.f1	0.0421	0.0138	0.0437	0.0160	0.0569	0.0198
τ^2, τ_k^2	0.0156	0.0020	0.0157	0.0023	0.0155	0.0029
σ^2, σ_k^2	0.1055	0.0019	0.0895	0.0012	0.1104	0.0017
π_k			0.5687	0.0434	0.4313	0.0434