Policy Research Working Paper 10828

Frontiers in Small Area Estimation Research

Application to Welfare Indicators

Isabel Molina

WORLD BANK GROUP
Poverty and Equity Global Practice
June 2024

Policy Research Working Paper 10828

Abstract

This paper reviews the main methods for small area estimation of welfare indicators. It begins by discussing the importance of small area estimation methods for producing reliable disaggregated estimates. It mentions the baseline papers and describes the contents of the different sections. Basic direct estimators obtained from area-specific survey data are described first, followed by simple indirect methods, which include synthetic procedures that do not account for the area effects and composite estimators obtained as a composition (or weighted average) of a synthetic and a direct estimator. The previous estimators are design-based, meaning that their properties are assessed under the sampling replication mechanism, without assuming any model to be true. The paper then turns to proper model-based estimators that assume an explicit model. These models allow obtaining optimal small area estimators when the assumed model holds. The first type of models, referred to as area-level models, use only aggregated data at the area level to fit the model. However, unit-level survey data were previously used to calculate the direct estimators, which act as response variables in the most common area-level models. The paper then switches to unit-level models, describing first the usual estimators for area means, and then moving to general area indicators. Semi-parametric, non-parametric, and machine learning procedures are described in a separate section, although many of the procedures are applicable only to area means. Based on the previous material, the paper identifies gaps or potential limitations in existing procedures from a practitioner's perspective, which could potentially be addressed through research over the next three to five years.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The author may be contacted atisabelmolina@ucm.es.

Produced by the Research Support Team

Frontiers in Small Area Estimation Research: Application to Welfare Indicators[1]

Isabel Molina

Institute of Interdisciplinary Mathematics,

Department of Statistics and Operations Research,

Complutense University of Madrid, Spain, isabelmolina@ucm.es

Keywords: EBLUP, ELL, Empirical best, Poverty mapping, Poverty map, Review, Small area estimation, Welfare estimation.

JEL Classification: C55, C87, C15.

[1]This work was done under the Contract num. 7209970 between Universidad Complutense de Madrid and The World Bank Group, Poverty & Equity GP, TTL Utz J. Pape.

Contents

Chapter 1

Introduction

The recent natural disasters and current armed conflicts are having devastating effects on the lives of millions of people. In these times of great turbulence, it becomes even more urgent to help those in need. We cannot forget that social and economic development helps prevent future conflicts and mitigates the impact of adverse events. However, especially in the current circumstances, humanitarian aid and development funds are limited, making it even more important to direct them to the places where they are most needed.

Elbers et al. (2004) investigated the impact on poverty alleviation of transferring an exogenously given budget to geographically defined subgroups of the population according to their relative poverty status. They observed substantial gains from targeting smaller administrative units, such as districts or villages, as opposed to larger regions. However, for effective targeting, it is crucial to utilize the most reliable statistical information concerning the living conditions of people in those small areas.

Small area estimation (SAE) methods offer reliable statistical figures at the local level or for population subgroups. These techniques are well-developed for various contexts, but there are still certain limitations in the existing methodology under realistic circumstances. This paper first

reviews the existing SAE procedures, starting with the very simple direct estimates obtained from area-specific survey data, progressing through the simple past indirect methods, and concluding with the more sophisticated and recent model-based procedures, some of which can incorporate various, possibly heterogeneous, data sources.

The review is based on several documents, including other recent review papers. Specifically, it heavily relies on Molina (2019), Corral Rodas, Molina and Nguyen (2021), Molina, Corral and Nguyen (2022), Corral et al. (2022) and Molina and Rao (2023). Drawing upon this literature review, this paper then identifies challenges or gaps in the current SAE methodology that are relevant for practitioners and could potentially be addressed through research in the next 3-5 years.

Chapters 2 to 6 review a significant portion of the existing SAE literature. Chapter 2 describes direct estimators based on area-specific survey data and related methods. Chapter 3 goes over early indirect estimators, including the first synthetic estimators as well as composite estimators. Then, Chapters 4–6 provide an overview of modern model-based SAE procedures, describing many recent extensions of these models and methods. We emphasize the main ideas behind each method, focusing on model types and on the estimation methods based on them. The description includes the pros and cons of each procedure from a practical standpoint, while details such as fitting procedures and other technical issues are omitted due to space restrictions. Finally, Chapter 8 enumerates open problems or topics deserving further research, once again considering the perspective of practitioners. It suggests potential topics of interest that could be addressed through research in the next 3-5 years.

Chapter 2

Direct estimation

The earliest estimates for subpopulations based on sample surveys were "direct", in the sense that they used only the survey data from the subpopulation of interest without "borrowing strength" from other subpopulations. These estimates are based on the sampling design; that is, they exhibit good properties across all possible samples of units drawn from the target population using the specified sampling design (random mechanism used to draw the samples). For a detailed account of the sampling theory, refer to the well-known books by Cochran (1977), Särndal, Swensson and Wretman (1992), Thompson (1997), or Lohr (1999). The guidelines by Corral et al. (2022) briefly introduce the design-based setup for inference in the context of SAE for poverty mapping in Section 2.1.

This section introduces common direct estimators based on sampling theory, as they serve as the benchmark for any comparison. These estimators make no model assumptions and exhibit good properties for areas or domains with sufficiently large sample sizes. However, they are inefficient (with large variances) in areas with small sample sizes. SAE methods aim to improve the efficiency (reduce the variance) of direct estimators, typically at the cost of increased bias. For example, refer to Figure 2.1 in Corral et al. (2022), which illustrates the design bias of model-based

small area estimators compared with direct ones in a design-based validation study. It is crucial to note that this design bias must always be kept small for small area estimators to be useful (say, relative bias not exceeding 10%).

Let U be a finite population of N units, assumed to be partitioned into D subpopulations, referred to as areas or domains, denoted $U_1, \ldots, U_D$, with sizes $N_1, \ldots, N_D$, where $N = \sum_{d=1}^{D} N_d$. Let y_{di} be the target variable for unit i within area d, where $i = 1, \ldots, N_d$ and $d = 1, \ldots, D$. In survey sampling, model assumptions are not made for y_{di} and these values are assumed to be fixed constants (measured without error). Therefore, randomness arises only from the mechanism used to draw the sample (the sampling design).

In this paper, target indicators for the areas are defined as general (real) functions of the values of the target variable in all the units from area d, that is, $\delta_d = h_d(y_{d1}, \ldots, y_{dN_d})$. Simple indicators are obtained when $h_d(\cdot)$ is a linear function. Let us define the vector $\mathbf{y}_d = (y_{d1}, \ldots, y_{dN_d})'$. Then, a linear indicator has the form $\delta_d = \mathbf{a}_d'\mathbf{y}_d$, where $\mathbf{a}_d = (a_{d1}, \ldots, a_{dN_d})'$ is a vector of known constants. Sampling theory was traditionally developed for linear indicators. Here we review traditional direct estimators for area means

$$\bar{Y}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} y_{di}, \quad d = 1, \ldots, D.$$

which are special cases of linear indicators $\delta_d = \mathbf{a}_d'\mathbf{y}_d$, for $\mathbf{a}_d = (1/N_d, \ldots, 1/N_d)'$.

A sample s_d of size n_d is supposed to be drawn independently from each area U_d, and $s = s_1 \cup \cdots \cup s_D$ is the overall sample, with size $n = \sum_{d=1}^{D} n_d$. An unbiased estimator of $\bar{Y}_d$ across all the possible samples s_d drawn from U_d using the specified sampling design is the Horvitz-Thompson (HT) estimator. Let π_{di} be the inclusion probability of unit i in the sample s_d from area d, and $w_{di} = \pi_{di}^{-1}$ be the survey weight (or elevation factor) of that unit. The survey weight is interpreted as the number of population units that the i-th sample unit represents. The Horvitz-

Thompson estimator of $\bar{Y}_d$ is then given by

$$\hat{\bar{Y}}_d = \frac{1}{N_d} \sum_{i \in s_d} w_{di} y_{di}.$$

Although the HT direct estimator $\hat{\bar{Y}}_d$ is unbiased for $\bar{Y}_d$, it is not a weighted average unless $N_d = \sum_{i \in s_d} w_{di}$, and it may have larger variance than the ratio HT (or Hájek) estimator of $\bar{Y}_d$, which is defined as the weighted average

$$\hat{\bar{Y}}_d^R = \frac{1}{\hat{N}_d} \sum_{i \in s_d} w_{di} y_{di} = \bar{y}_{dw}, \tag{2.1}$$

where $\hat{N}_d = \sum_{i \in s_d} w_{di}$.

If $\pi_{di} > 0$ for all $i = 1, \ldots, N_d$, the variance across the possible samples s_d (hereafter design variance) of the HT estimator of $\bar{Y}_d$ is given by

$$\mathrm{var}_\pi(\hat{\bar{Y}}_d) = \frac{1}{N_d^2} \left\{ \sum_{i=1}^{N_d} \frac{y_{di}^2}{\pi_{di}} (1 - \pi_{di}) + 2 \sum_{i=1}^{N_d} \sum_{j>i}^{N_d} \frac{y_{di} y_{dj}}{\pi_{di} \pi_{dj}} \left(\pi_{d,ij} - \pi_{di} \pi_{dj} \right) \right\}. \tag{2.2}$$

The design variance of the Hájek estimator $\hat{\bar{Y}}_d^R$ may be obtained by applying the Taylor linearization method to the ratio of HT estimators $\hat{N}_d$ and $\hat{Y}_d^R = \sum_{i \in s_d} w_{di} y_{di}$ of the totals N_d and $Y_d = \sum_{i=1}^{N_d} y_{di}$, respectively.

If the sample s_d is drawn with simple random sampling (SRS), then $w_{di} = N_d / n_d$ for all $i = 1, \ldots, N_d$, and both the HT direct estimator and the ratio HT estimator reduce to the usual (unweighted) sample mean,

$$\hat{\bar{Y}}_d = \hat{\bar{Y}}_d^R = \frac{1}{n_d} \sum_{i \in s_d} y_{di} = \bar{y}_d.$$

Under SRS without replacement, the design variance (2.2) reduces to the usual formula,

$$\mathrm{var}_\pi(\hat{\bar{Y}}_d) = \left(1 - \frac{n_d}{N_d} \right) \frac{\sigma_y^2}{n_d}, \quad \sigma_y^2 = \frac{1}{N_d - 1} \sum_{i=1}^{N_d} \left(y_{di} - \bar{Y}_d \right)^2. \tag{2.3}$$

Direct estimators based on the sampling design offer several advantages, particularly when applied to areas with large sample sizes. Specifically, they avoid making model assumptions for the study variable and have good properties across all possible samples s_d drawn from U_d. As mentioned earlier, the HT estimator is design unbiased for $\bar{Y}_d$, and the ratio HT estimator is design consistent as the area sample size n_d grows. This means that they are more likely to be close to the true mean $\bar{Y}_d$ as n_d increases. Another important advantage of direct estimators is their use of "all-purpose" expansion weights. This means that the same expansion weights w_{di} are employed for the estimation of totals or means of any variable of interest, facilitating the automatic production of large amounts of statistical information. However, challenges arise for areas with small sample sizes, where the design variance becomes unacceptable. Note that $\text{var}_\pi(\hat{\bar{Y}}_d) = O(n_d^{-1})$, indicating that it decreases as the area-specific sample size n_d grows at the same rate as $1/n_d$. However, it grows unboundedly as n_d decreases, as illustrated clearly for SRS in (2.3).

Generalized Regression (GREG) estimators and more general calibration estimators (Deville and Särndal, 1992; Lehtonen, Särndal and Veijanen, 2003) applied to areas are designed to enhance the efficiency of direct domain estimators, owing to the knowledge of the totals of certain auxiliary variables. These procedures adjust the sampling weights w_{di} so that the expansion estimates of the known totals of the auxiliary variables, based on the final weights, become equal to the known true totals. The adjusted weights may be similarly used to estimate totals or means of other variables of interest.

Nowadays, expansion weights are typically calibrated using the known totals of certain auxiliary variables and are also adjusted for non-response. However, the resulting expansion estimators may still be inefficient for areas with a small sample size n_d, as their design variances are $O(n_d^{-1})$. One way to ameliorate the SAE problem at the design stage of the survey,

which is always advisable, is to allocate the total survey sample size n more efficiently among the different areas. However, as stated by Fuller (1999), p. 344, "the client will always require more than is specified at the design stage", and hence SAE techniques might still be needed.

Chapter 3

Basic indirect estimators

Indirect estimators overcome data scarcity in an area by "borrowing strength" across areas. This is done by imposing homogeneity assumptions that connect the areas through shared parameters. These common parameters are estimated with a larger sample size, leading to significantly more efficient small area estimators.

The first indirect estimators were "synthetic", a term used for estimators that assume common characteristics for all the areas, without allowing for area heterogeneity. According to Gonzalez (1973), "An estimator is called a synthetic estimator if a reliable direct estimator for a large area, covering several small areas, is used to derive an indirect estimator for a small area under the assumption that the small areas have the same characteristics as the large area".

Very simple synthetic estimators are post-stratified synthetic estimators, which assume that the population is partitioned into J groups $U^1, \ldots, U^J$, known as post-strata, each with sufficiently large sample sizes $n^1, \ldots, n^J$, that cut across the areas. Hence, the area U_d is also partitioned in J groups, $U_d^1, \ldots, U_d^J$ of population sizes $N_d^1, \ldots, N_d^J$, and means $\bar{Y}_d^1, \ldots, \bar{Y}_d^J$, where $\bar{Y}_d^j = \sum_{i \in U_d^j} y_{di}/N_d^j$, $j = 1, \ldots, J$. The area mean $\bar{Y}_d$ can be expressed as a weighted average of the means for the J

post-strata within that area, as

$$\bar{Y}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} y_{di} = \frac{1}{N_d}\sum_{j=1}^{J} N_d^j \bar{Y}_d^j. \tag{3.1}$$

Assuming that the mean $\bar{Y}^j$ of the study variable in post-stratum j is constant across the D areas and only varies between post-strata, that is, assuming that $\bar{Y}_d^j = \bar{Y}^j$, for all $d = 1, \ldots, D$, and for each $j = 1, \ldots, J$, we can replace $\bar{Y}_d^j$ by $\bar{Y}^j$ in (3.1), resulting in

$$\bar{Y}_d = \frac{1}{N_d}\sum_{j=1}^{J} N_d^j \bar{Y}^j. \tag{3.2}$$

Then, the post-stratified synthetic (PS-SYN) estimator of $\bar{Y}_d$ is obtained by replacing the mean in each post-stratum $\bar{Y}^j$ in (3.2) with the direct estimator $\hat{\bar{Y}}^{j,R}$, as follows

$$\hat{\bar{Y}}_d^{PS-SYN} = \frac{1}{N_d}\sum_{j=1}^{J} N_d^j \hat{\bar{Y}}^{j,R}.$$

Since post-strata sample sizes are assumed to be sufficiently large, estimators $\hat{\bar{Y}}^{j,R}$, $j = 1, \ldots, J$, have small design variances, resulting in a small design variance for the PS-SYN estimator of $\bar{Y}_d$. However, the assumption of a constant mean across the areas within the same post-stratum is hardly true, leading to typically large design bias in PS-SYN estimators. It is important to note that a reduction of design variance at the expense of a large design bias is not an acceptable estimation strategy.

Hansen, Hurwitz and Madow (1953), p. 483, were the first to use a model for synthetic estimation, in an application based on the 1945 Radio Listening Survey. The objective was to estimate the median number of radio stations heard during the day in family houses from $D = 500$ US counties. They had estimates x_d, $d = 1, \ldots, D$, obtained from a mail survey conducted in the 500 counties, which were biased due to only 20% response rates and incomplete coverage. Additionally, they had estimates

y_d obtained for an intensive survey conducted in $m = 85$ of the counties. By considering those counties as the first 85 and regarding y_d as the true county median, they assumed the linear regression model

$$y_d = \beta_0 + \beta_1 x_d + e_d, \quad d = 1, \ldots, D, \tag{3.3}$$

with the usual regression assumptions, namely independent errors, $E(e_d) = 0$ and $\text{var}(e_d) = \sigma_e^2$, $d = 1, \ldots, D$. They fitted the regression model using the available y_d values for the $m = 85$ counties. The fitted regression parameters were then applied to predict the number of radio stations heard during the day in the remaining $500 - 85 = 415$ counties, for which the mail survey estimates x_d were available. The resulting predicted values are taken as the small area estimators for the remaining 415 counties where y_d was not available. Estimators derived from a regression model with common regression coefficients for all the areas, and where area effects are not estimated, are referred to as *regression-synthetic estimators*. Note that the regression coefficients in (3.3) are constant for all the $D = 500$ US counties and are estimated with the data (x_d, y_d) from the $m = 85$ counties where y_d was available, and county effects are not estimated. Moreover, the estimators derived from (3.3) do not account for the fact that y_d are subject to sampling error, unlike the SAE procedures described in Chapter 4.

As mentioned earlier, synthetic estimators can have relatively small design variances, but their design bias can be substantial due to the unrealistic assumptions underlying synthetic estimators. Since their design bias is not negligible, design mean squared error (MSE) estimates, which account for both, bias and variance, should be used to complement the synthetic point estimates. Aside from the potentially large bias, a challenge lies in obtaining efficient and area-specific design MSE estimates for these estimators.

Composite estimators, defined as a weighted average of a synthetic and a direct estimator for the same area, were proposed to shrink direct estimators toward the synthetic ones, reducing the design variances of the

direct estimators at the cost of slightly increasing their design bias. It is worth noting that averaging different predictors is one of the main ideas behind modern machine learning procedures.

In composite SAE estimators, optimal weights are sought from a design-based standpoint. Unfortunately, the optimal weight depends on the true design MSE estimates of the two estimators involved, encountering once again the problem of estimating the design MSE for synthetic estimators. Griffiths (1996) studied composite estimators and applied them to the estimation of labor force characteristics for US congressional districts.

Purcell and Kish (1979) considered a common weight for all the areas and obtained the optimal weight that minimized the total design MSE for all the D small areas. The resulting composite estimators have good overall efficiency for the D areas, but not necessarily for each small area. In SAE, it is desirable to reduce the largest MSEs, which typically correspond to the areas with the smaller sample sizes, and this is not ensured by these composite estimators.

Actually, in SAE, it is much more appealing to consider composite estimators with area-specific weights, such that the weight attached to the synthetic estimator grows for areas with small sample sizes and decreases for areas with large sample sizes, giving more weight to the direct estimator. Following this idea, Drew, Singh and Choudhry (1982) proposed the sample-size-dependent (SSD) estimators. These are composite estimators defined with simple weights that depend on the area sample size. They applied these estimators to produce estimates for Census Divisions from the Canadian Labor Force Survey. In practice, as observed in the application by Drew, Singh and Choudhry (1982), SSD estimators borrow little or no strength, because the weights attached to the direct estimators frequently turn out to be either equal or close to one.

Chapter 4 describes the first SAE model, which also leads to a composite estimator, but with optimal properties under model assumptions made for the study variable. These model-based estimators outperform the

composite estimators described above by borrowing significant strength from other areas. They can achieve substantial efficiency gains, provided that the model assumptions hold.

Chapter 4

Area level models

A proper SAE model that incorporates the sampling errors into the linear regression (3.3), was introduced by Fay and Herriot (1979) with the purpose of estimating the mean per capita income in US areas with fewer than 1,000 inhabitants. This model utilizes only aggregated data at the area level, which is more easily available. Due to its broad applicability, simplicity, interpretability, and the favorable properties of the estimators derived from it, it is possibly the most popular SAE model.

The Fay-Herriot (FH) model is defined in two stages. In the first stage, Fay and Herriot (1979) assume that the true area indicators δ_d vary linearly with a $p \times 1$ vector $\mathbf{x}_d = (x_{d1}, x_{d2}, \ldots, x_{dp})'$ of area-level covariates, as follows:

$$\delta_d = \mathbf{x}_d'\boldsymbol{\beta} + u_d, \quad d = 1, \ldots, D. \tag{4.1}$$

This model is known as *linking model*, because the vector of regression coefficients, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$, is assumed to be constant for all the areas, thus "linking" the areas. The regression errors, u_d, are assumed to be independent for all $d = 1, \ldots, D$, following $u_d \sim N(0, \sigma_u^2)$, with $\sigma_u^2 > 0$ unknown, and are referred to as *area effects*, because they model the between-area heterogeneity that is not explained by the area-level covariates in $\mathbf{x}_d$. Typically, $x_{d1} = 1$ to allow for an intercept in the

regression and $x_{d2}, \ldots, x_{dp}$ are the population means of $p-1$ auxiliary variables, obtained from census data.

Note that the true indicators δ_d are not available and hence model (4.1), as it is, cannot be fit. However, direct estimators $\hat{\delta}_d^{DIR}$, $d = 1, \ldots, D$, are assumed to be available from the unit-level data in a (current) survey, where the areas may be identified. Traditional indicators of interest are the area means $\delta_d = \bar{Y}_d$, for which basic direct estimators $\hat{\delta}_d^{DIR}$ and corresponding design variances $\text{var}_\pi(\hat{\delta}_d^{DIR})$, are described in Chapter 2. Note that the direct estimators $\hat{\delta}_d^{DIR}$ are subject to sampling error, which might be substantial, since they are based on the n_d observations from area d, which is supposed to be too small for certain areas. Fay and Herriot (1979) proposed to account for the (important) sampling errors of the direct estimators $\hat{\delta}_d^{DIR}$ of δ_d, by considering the following model, known as *sampling model*, in the second stage:

$$\hat{\delta}_d^{DIR} = \delta_d + e_d, \quad d = 1, \ldots, D, \tag{4.2}$$

where e_d, $d = 1, \ldots, D$, are supposed to be independent, and independent of the area effects u_d, following $e_d \sim N(0, \psi_d)$, with variances $\psi_d = \text{var}_\pi(\hat{\delta}_d^{DIR})$, $d = 1, \ldots, D$, assumed to be known.

Replacing the linking model (4.1) in the sampling model (4.2) results in the linear mixed model

$$\hat{\delta}_d^{DIR} = \mathbf{x}_d'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D, \tag{4.3}$$

with known and heteroscedastic error variances ψ_d, $d = 1, \ldots, D$. The statistical theory behind linear mixed models is well described in Searle (1971), Searle, Casella and McCulloch (1997) and Jiang (2007). For σ_u^2 known, Henderson (1950) obtained the best linear unbiased predictor (BLUP) of a linear combination of the vector of fixed effects $\boldsymbol{\beta}$ and the vector random effects $\mathbf{u}_d = (u_1, \ldots, u_D)'$ in a linear mixed model. The BLUP of $\delta_d = \mathbf{x}_d'\boldsymbol{\beta} + u_d$ is the predictor $\tilde{\delta}_d$ that is a linear combination of $\mathbf{y} = (\hat{\delta}_1^{DIR}, \ldots, \hat{\delta}_D^{DIR})'$, which is unbiased in the sense $E(\tilde{\delta}_d - \delta_d)$, and

minimizes the model MSE, $E(\tilde{\delta}_d - \delta_d)^2$. Here, the expectations are taken with respect to the distribution induced by the FH model (4.3), where the normality assumptions are not necessary. Note that the inference under a model (model-based inference) is completely different from the inference under the sampling design (design-based inference) described in Chapter 2. In model-based inference, the true values δ_d are assumed to follow a model, and are thus random variables, unlike in the design-based inference. That is the reason why the term "estimator" of δ_d is replaced by "predictor" of (the realized value of) δ_d.

Based on the FH model (4.1)–(4.2), the BLUP of $\delta_d = \mathbf{x}_d'\boldsymbol{\beta} + u_d$ is obtained simply by fitting the linear mixed model (4.3) assuming σ_u^2 known, and then using the predicted value of δ_d through the model, that is,

$$\tilde{\delta}_d^{FH} = \mathbf{x}_d'\tilde{\boldsymbol{\beta}} + \tilde{u}_d. \tag{4.4}$$

Here, $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \mathbf{x}_d'\tilde{\boldsymbol{\beta}})$ is also the BLUP of the area effect u_d, where $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$, and $\tilde{\boldsymbol{\beta}}$ is the weighted least squares (WLS) estimator of $\boldsymbol{\beta}$, which equals the maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ under normality of random effects and errors, given by

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \gamma_d \mathbf{x}_d \mathbf{x}_d'\right)^{-1} \sum_{d=1}^{D} \gamma_d \mathbf{x}_d \hat{\delta}_d^{DIR}. \tag{4.5}$$

Note that it is possible to use only the regression part of the BLUP in (4.4) to estimate δ_d, which is the regression-synthetic estimator $\tilde{\delta}_d^{RSYN} = \mathbf{x}_d'\tilde{\boldsymbol{\beta}}$. This is customarily taken for areas d that are not observed in the survey or for which no positive variance estimates ψ_d are available (for which γ_d is not defined, so its limiting value $\gamma_d = 1$ as $\psi_d \to 0$ is taken in the BLUP).

Replacing $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \mathbf{x}_d'\tilde{\boldsymbol{\beta}})$ in (4.4) and re-arranging the terms, the BLUP can be seen as a composite estimator between the direct estimator $\hat{\delta}_d^{DIR}$ used as response variable, and the above regression synthetic

estimator, that is,

$$\tilde{\delta}_d^{FH} = \gamma_d \hat{\delta}_d^{DIR} + (1 - \gamma_d)\mathbf{x}_d'\tilde{\boldsymbol{\beta}}, \tag{4.6}$$

where the weight attached to the direct estimator, $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d) \in (0, 1)$, depends on the estimated variance ψ_d of the direct estimator (which grows as the area sample size n_d decreases), relative to the unexplained area heterogeneity measured by σ_u^2. For an area d where the direct estimator $\hat{\delta}_d^{DIR}$ is efficient, $\tilde{\delta}_d^{FH}$ attaches more weight to the direct estimator $\hat{\delta}_d^{DIR}$ than to the regression-synthetic counterpart, and more weight is given to the latter for the areas whose direct estimator has bad quality in terms of ψ_d, relative to the unexplained between-area variability σ_u^2. Note that the regression-synthetic estimator "borrows strength" from all the areas through $\tilde{\boldsymbol{\beta}}$. Hence, the BLUP (4.6) based on the FH model (4.3) automatically "borrows strength" for the areas where it is needed, but gets close to the direct estimator when it is efficient enough, which is a desirable property, given that direct estimators are design unbiased (or design consistent) without making any model assumptions. As a consequence, the BLUP (4.6) inherits the good (design) properties of the direct estimator for sufficiently large n_d, such as design consistency. The expression 4.6 indicates that the BLUP may be employed for all the areas $d = 1, \ldots, D$, regardless of whether they have large or small sample size n_d; or equivalently, whether their corresponding direct estimator has acceptable quality or not.

Note that using the regression-synthetic estimators $\tilde{\delta}_d^{RSYN} = \mathbf{x}_d'\tilde{\boldsymbol{\beta}}$ for all the areas is not recommendable for several reasons: first, area effects might be significant; we can never be completely sure that the available covariates are explaining all the area heterogeneity; second, the sample size n_d might be large in some or even many of the areas, and hence those observations on the actual target variable obtained from the survey, which are expensive to get, would be waisted; third, we can never be completely sure that the assumed model is correct, and we would give zero weight to the direct estimator (which makes no model assumptions) for all the

areas, even for those with large sample size. Finally, design consistency is lost; that is, given the true value of the indicator δ_d, the estimator $\tilde{\delta}_d^{RSYN}$ is not more likely close to it as the area sample size n_d grows.

The BLUP of δ_d depends on the area effects variance, σ_u^2, which is unknown in practice, and must be estimated. The usual estimation methods are ML and restricted/residual ML (REML), both using the Normal likelihood, or the FH method proposed by Fay and Herriot (1979), which is a moments method. The REML method corrects the ML estimator by accounting for the degrees of freedom associated with estimating the vector of regression coefficients $\boldsymbol{\beta}$, resulting in an estimator of σ_u^2 with smaller bias for finite number of areas D. Under certain regularity assumptions, the three estimators are consistent as the number of areas D grows.

Let $\hat{\sigma}_u^2$ be a consistent estimator of σ_u^2. By replacing σ_u^2 with $\hat{\sigma}_u^2$ in the BLUP given in (4.4), we obtain the empirical BLUP (EBLUP) of δ_d,

$$\hat{\delta}_d^{FH} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1 - \hat{\gamma}_d)\mathbf{x}_d'\hat{\boldsymbol{\beta}}, \tag{4.7}$$

where now $\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \psi_d)$ and $\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^D \hat{\gamma}_d \mathbf{x}_d \mathbf{x}_d'\right)^{-1} \sum_{d=1}^D \hat{\gamma}_d \mathbf{x}_d \hat{\delta}_d^{DIR}$. We will refer to the EBLUP (4.7) based on the FH model (4.3) as the FH estimator.

When $\boldsymbol{\beta}$ and σ_u^2 are known, the MSE of the BLUP, $\tilde{\delta}_d^{FH}$, under the FH model (4.3) is given by

$$\mathrm{MSE}(\tilde{\delta}_d^{FH}) = \gamma_d \psi_d \leq \psi_d = \mathrm{var}_\pi(\hat{\delta}_d^{DIR}). \tag{4.8}$$

As a consequence, given δ_d, if σ_u^2 and $\boldsymbol{\beta}$ are known, the FH estimator, $\tilde{\delta}_d^{FH}$, cannot be less efficient than the direct estimator. In practice, σ_u^2 and $\boldsymbol{\beta}$ are estimated, and the error due to the estimation of those parameters is added to the MSE. Under the normality of random effects u_d and sampling errors e_d and certain regularity assumptions, Prasad and Rao (1990) obtained an approximation with $o(D^{-1})$ error for large number of areas D for the MSE of the FH estimator, see Rao and Molina (2015), eqn. (6.2.1). They further obtained an estimator of the MSE with bias of

error $o(D^{-1})$, see Rao and Molina (2015), pp. 136-137. The terms added to the MSE in (4.8) due to the estimation of σ_u^2 and $\boldsymbol{\beta}$ are $O(D^{-1})$, meaning that they tend to zero as D grows at a rate of $1/D$. Therefore, for a large number of areas D, the FH estimator is likely to improve the direct estimator in terms of MSE. As a consequence, these estimators usually improve in most of the areas, as long as the number of areas D is large enough, but the efficiency gains might be small or even non-existing for small D.

The FH model is probably the most applied SAE method, because it requires only aggregated data at the area level and corrects the potentially large bias of regression-synthetic estimators, by accounting for the unexplained between-area heterogeneity, while preserving good design properties for areas with large sample sizes n_d. Still, the unit-level models described in Sections 5 and 6 use unit-level data, of size $n = \sum_{d=1}^{D} n_d$, typically much larger than D. Consequently, they can achieve substantially greater efficiency gains than FH estimators, estimated with a sample size equal to D, provided that the unit-level covariates are useful.

An issue with FH estimators is that the error variances in the sampling model, $\psi_d = \text{var}_\pi(\hat{\delta}_d^{DIR})$, $d = 1 \ldots, D$, are assumed to be known. These variances are not known and are usually estimated with the n_d observations in the survey for area d, which are also "direct". Therefore, the estimated variances $\widehat{\text{var}}_\pi(\hat{\delta}_d^{DIR})$ have a significant error, leading to FH estimators that are poorer than those obtained with the known true variances. Additionally, the error due to the estimation of these sampling variances, which depends on the area-specific sample size n_d, is typically ignored in the MSE of the FH estimator; hence, gains with respect to direct estimators might be overstated in the areas with small sample sizes n_d.

Bell (2008), as well as Rivest and Vandal (2003), studied the effect in the MSE of the BLUP for $\boldsymbol{\beta}$ and σ_u^2 known, of using the direct estimators $\widehat{\text{var}}_\pi(\hat{\delta}_d^{DIR})$ of $\psi_d = \text{var}_\pi(\hat{\delta}_d^{DIR})$ in place of the true variances. Using

estimated variances $\hat{\psi}_d = S_d^2/n_d$, where $S_d^2 = (n_d - 1)^{-1} \sum_{i \in s_d} (y_{di} - \bar{y}_d)^2$, Rivest and Vandal (2003) proposed an estimator of the MSE that accounts for the uncertainty due to the estimation of these variances. Wang and Fuller (2003) gave another estimator of the MSE when $\boldsymbol{\beta}$ and σ_u^2 are estimated, using certain moment estimators of these parameters. However, more general results for other fitting methods and estimators $\hat{\psi}_d$ of ψ_d are lacking.

The Generalized variance function (GVF) proposed by Vaillant (1987), or a more general preliminary regression model for the estimated variances $\widehat{\text{var}}_\pi(\hat{\delta}_d^{DIR})$, is often applied to smooth these estimated variances. The smoothed variances obtained as predicted values from this model are then treated as the true variances ψ_d. However, the number of variables (or number of parameters) in that model, which tunes the level of smoothing of these variances, is not clear. Note that the aim of the preliminary model is not to reproduce the direct variance estimates. Moreover, when assessing the gains of FH estimators compared to direct estimators in applications, it is unclear whether the comparison should be done based on the raw estimated sampling errors or on the smoothed versions.

The SAIPE project (see e.g. Bell, 1997) within the US Census Bureau regularly employs the FH model. Ericksen and Kadane (1985) and Cressie (1989) used the FH model to estimate the decennial census undercounts in each US state, and Dick (1995) used it to estimate Canadian census undercounts. A historical application of FH model is given by You and Chapman (2006), who used the same data set as in Arora and Lahiri (1997) to estimate small-area average expenditure on fresh milk.

The FH model has been utilized by various authors to estimate welfare indicators. Let us mention just a few examples. Molina and Morales (2009) estimated poverty rates and gaps in Spanish provinces by gender. Jedrzejczak and Kubacki (2013) estimated income inequality and poverty rates by regions and family type in Poland. Casas-Cordero Valencia, Encina and Lahiri (2015) estimated poverty rates in Chilean comunas

based on the FH model with arcsin transformation. Corral and Cojocaru (2019) estimated poverty indicators in Moldova by districts and Seitz (2019) estimated poverty at the district level in Central Asian countries.

The FH model has been extended in many different ways. For the case where the sampling errors e_d in (4.2) are correlated, Isaki, Huang and Tsay (1991), an later Isaki, Tsay and Fuller (2000), extended the FH model and used it to estimate census undercounts by different post-strata. Multivariate versions used to estimate several dependent area indicators were proposed by Fay (1987), and were utilized by Datta, Fay and Ghosh (1991) and Datta et al. (1996) to improve the direct estimates of median income for four-person families. A bivariate FH model with a t-distribution was considered by Bell and Huang (2006) to account for outliers in the poverty ratios for school-aged (5-17) children for US states in 2002. Later, Benabent and Morales (2016) used a bivariate FH model to estimate poverty proportions and gaps at the province level for the years 2005 and 2006.

A subarea-level model, used to produce estimates at two different nested aggregation levels, was introduced by Fuller and Goyeneche (1998), and a similar model was studied by Torabi and Rao (2014). Unmatched sampling and linking models, where a one-to-one transformation of the target indicator $g(\delta_d)$ is taken as response variable in the linking model (4.1) but retaining the sampling model (4.2), were originally studied by You and Rao (2002b). A FH model with measurement error in the covariates was introduced by Ybarra and Lohr (2008). This model was used by Marchetti et al. (2015) using big data covariates to estimate poverty rates and mean income in the ten provinces of the Tuscany region in Italy.

The FH model has been extended to "borrow strength" over time, as well as across areas, by Rao and Yu (1992; 1994). They added AR(1) time effects nested within the area effects in the FH model. You, Rao and Gambino (2003) applied the Rao-Yu model to estimate monthly un-

employment rates for cities with population over 100,000 inhabitants in Canada. Ghosh, Nangia and Kim (1996) proposed a different time series cross-sectional model and used it to estimate the median income of four-person families for US states and the District of Columbia. Datta et al. (1999), You (1999) and Datta, Lahiri and Maiti (2002) replaced the AR(1) time effects in Rao-Yu model by a random walk, and Datta et al. (1999) applied the model, including seasonal variation, to estimate monthly unemployment rates for 49 US states and the District of Columbia. Pfeffermann and Burck (1990) proposed a more complex model with area-by-time random effects and regression coefficients varying by area and time. Esteban et al. (2010; 2012) applied the Rao-Yu model to estimate small area poverty indicators in Spanish provinces by gender.

Various extensions of the FH model have been propose to "borrow strength" from space. A model with area effects following a Conditionally Autoregressive (CAR) process (Besag, 1974) was proposed by Cressie (1991) to estimate the census undercount in US states. Petrucci and Salvati (2006) proposed a model with area effects following a Simultaneously Autoregressive (SAR) process to estimate the amount of erosion delivered to streams in the Rathbun Lake Watershed in Iowa by $D = 61$ sub-watersheds. Pratesi and Salvati (2008) used the same model to estimate mean PCI in $D = 43$ sub-regions of Tuscany. A similar spatial FH model was used by Giusti, Masserini and Pratesi (2017) to estimate mean income and poverty rates for the 57 Labor Local Systems of the Tuscany region in Italy for the year 2011. Chandra, Salvati and Chambers (2017) proposed an SAE model for non-stationary spatial data. Marhuenda, Molina and Morales (2013) considered a spatio-temporal model, including simultaneously area effects following a SAR process and time effects following an AR(1) process nested within the area effects. They applied this model to the estimation of poverty indicators for Spanish provinces in 2008, using survey data from 2004-2008.

Generalized Linear Models (GLMs) (Nelder and Wedderburn 1972; McCullagh and Nelder 1989) have also been used in SAE. Although most of the applications are related to disease mapping, there are also contributions to SAE estimation of wealth. For example, a hierarchical Beta mixed regression model was employed by Fabrizi and Trivisano (2016) to estimate of the Gini coefficient. A Poisson mixed model with normal area effects was applied by Boubeta, Lombardía and Morales (2016) to estimate poverty rates in counties of the Spanish region of Galicia by gender. This work was extended in Boubeta, Lombardía and Morales (2017) by including AR(1) time effects nested within the area effects, and in Boubeta et al. (2020) to a model with spatio-temporal correlation. Franco and Bell (2013; 2015) used a bivariate Binomial Logit Normal model to estimate county poverty rates of school-aged (5-17) children.

We now discuss extensions of the EBLUP under the FH model that are robust to certain model misspecifications. Jiang, Nguyen and Rao (2011) proposed the observed best predictor (OBP) as a more robust alternative under misspecification in the linking model. The OBP is obtained by using estimators of the model parameters obtained from a predictive point of view, without appealing to the linking model. The OBP was extended to models for counts by Chen, Jiang and Nguyen (2015).

The presence of area-level outliers in the FH model was studied by Datta and Lahiri (1995) and Bell and Huang (2006) under the Hierarchical Bayesian (HB) setup. The first robust proposal for the FH model seems to be by Ghosh, Maiti and Roy (2008), based on Huberized residuals, obtained by applying the Huber's ψ function (Huber, 1964) to model residuals.

Other practical issues related to the FH model studied in the literature are benchmarking the small area estimates so that they add up to the estimate in a larger area covering the small areas, see Pfeffermann and Barnard (1991), Wang, Fuller and Qu (2008), Datta et al. (2011), Steorts and Ghosh (2013) and Bell, Datta and Ghosh (2013).

The EBLUP based on the basic FH model is implemented in several R packages, namely the `sae` package (Molina and Marhuenda, 2015), which includes also spatial and spatio-temporal extensions of the FH model, together with either analytical or resampling-based functions for MSE estimation; the `Josae` package (Breidenbach, 2018), which allows to include heteroscedasticity in the model; `hbsae` (Boonstra, 2012), which includes also Hierarchical Bayesian methods and `BayesSAE` (Shi, 2018), which includes functions for unmatched models and also spatial models. The temporal Rao and Yu (1992) model is implemented in the R package `saery` (Esteban Lefler, Morales González and Pérez Martin, 2014); multivariate FH models can be fit with the R package `msae` (Permatasari and Ubaidillah, 2020) and, finally, measurement error models can be applied with the package `saeME` (Mubarak and Ubaidillah, 2020). The EBLUP based on the basic FH model is also available as a command in Stata `fhsae` (Corral et al. 2018). The Stata command `fhsae` also allows for the aggregation of estimators and obtains the mean cross-product error detailed in Rao and Molina (2015), Section 6.2.6. Additionally, Halbmeier et al. (2019) developed the `fayherriot` Stata command, which includes transformations to address violations of the model assumptions and adjustments of the non-positive random effect variance estimates.

Chapter 5

Unit level models: Linear indicators

Battese, Harter and Fuller (1988) proposed a linear regression model with random area effects for data at the unit level, known as the nested error model, to estimate small area means. The model is defined as

$$\begin{aligned} &y_{di} = \mathbf{x}'_{di}\boldsymbol{\beta} + u_d + e_{di},\ u_d \overset{iid}{\sim} (0, \sigma_u^2), \\ &e_{di} \overset{ind}{\sim} (0, \sigma_e^2 k_{di}^2), \quad i = 1, \ldots, N_d,\ d = 1, \ldots, D, \end{aligned} \tag{5.1}$$

where $\mathbf{x}_{di}$ is a $1 \times p$ vector of auxiliary variables for unit i from area d. The first component of $\mathbf{x}_{di}$ is typically set to one to allow for a common intercept, and the remaining elements are the values of auxiliary variables that may vary by units or only by areas, $\boldsymbol{\beta}$ is the $p \times 1$ vector or regression coefficients, common for all the areas, u_d represents the random effect of area d and e_{di} is the unit-level regression error. Area effects u_d and errors e_{di} are all independent, and k_{di} are known heteroscedasticity constants, $i = 1, \ldots, N_d$, $d = 1, \ldots, D$.

Battese, Harter and Fuller (1988) utilized the nested error model mentioned above to estimate county means of crop areas under corn and under soybeans. They obtained unit-level data on corn and soybeans production from farm-interview data and auxiliary information from LANDSAT

satellite images. Therefore, their application represents a very early example of integration of different data sources for SAE.

This model has traditionally been used to estimate linear indicators such as area means, employing the EBLUPs defined by Henderson (1950) under the "infinite" population setup, or those derived by Royall (1970, 1976) under a finite population scheme, which agree when the area sampling fractions are negligible. We start describing the latter. Let us decompose the mean of area d into a sum of values for the units observed in the sample s_d and for the units in the sample complement $r_d = U_d - s_d$, as follows:

$$\bar{Y}_d = N_d^{-1} \left(\sum_{i \in s_d} y_{di} + \sum_{i \in r_d} y_{di} \right).$$

The values y_{di} for sample units $i \in s_d$ are observed, and only those for non-sample units $i \in r_d$ are unknown. The BLUP of $\bar{Y}_d$ under the nested error model (5.1) is obtained by fitting the model to the sample data using the WLS estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, and then predicting the values y_{di} for units outside the sample from area d. Specifically, the BLUP is given by

$$\tilde{\bar{Y}}_d^{BLUP} = N_d^{-1} \left(\sum_{i \in s_d} y_{di} + \sum_{i \in r_d} \tilde{y}_{di} \right), \tag{5.2}$$

with predicted values given by

$$\tilde{y}_{di} = \mathbf{x}'_{di}\tilde{\boldsymbol{\beta}} + \tilde{u}_d, \quad \tilde{u}_d = \gamma_d(\bar{y}_{da} - \bar{\mathbf{x}}'_{da}\tilde{\boldsymbol{\beta}}), \quad \gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/a_{d\cdot}}. \tag{5.3}$$

Here, $\bar{y}_{da} = a_{d\cdot}^{-1} \sum_{i \in s_d} a_{di} y_{di}$ and $\bar{\mathbf{x}}_{da} = a_{d\cdot}^{-1} \sum_{i \in s_d} a_{di} \mathbf{x}_{di}$ are respectively the weighted means of the response and the auxiliary variables, with weights given by $a_{di} = k_{di}^{-2}$, for $a_{d\cdot} = \sum_{i \in s_d} a_{di}$, $\tilde{u}_d$ is the BLUP of u_d and $\tilde{y}_{di}$ is the BLUP of y_{di} for $i \in r_d$, under the model (5.1).

Similar to the vector of response variables for area d, $\mathbf{y}_d = (y_{d1}, \ldots, y_{dN_d})'$, we construct the corresponding matrix of auxiliary variables for that same area, $\mathbf{X}_d = (\mathbf{x}_{d1}, \ldots, \mathbf{x}_{dN_d})'$. Under the nested error model (5.1), it holds

that $\mathbf{y}_d \overset{ind}{\sim} N(\mathbf{X}_d\boldsymbol{\beta}, \mathbf{V}_d)$, $d = 1, \ldots, D$, where

$$\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{A}_d,$$

where $\mathbf{1}_k$ denotes a vector of ones of size k and $\mathbf{A}_d = \text{diag}(k_{di}^2; i = 1, \ldots, N_d)$. Consider the partition of the vector $\mathbf{y}_d$ of area d and the matrices $\mathbf{X}_d$ and $\mathbf{V}_d$ into sub-vectors and matrices corresponding to the sample units $i \in s_d$, and for the non-sample units $i \in r_d = U_d - s_d$, as follows:

$$\mathbf{y}_d = \begin{pmatrix} \mathbf{y}_{ds} \\ \mathbf{y}_{dr} \end{pmatrix}, \quad \mathbf{X}_d = \begin{pmatrix} \mathbf{X}_{ds} \\ \mathbf{X}_{dr} \end{pmatrix}, \quad \mathbf{V}_d = \begin{pmatrix} \mathbf{V}_{ds} & \mathbf{V}_{dsr} \\ \mathbf{V}_{drs} & \mathbf{V}_{dr} \end{pmatrix}.$$

With this notation, when σ_u^2 and σ_e^2 are known, the WLS estimator of $\boldsymbol{\beta}$, which equals the ML estimator under normality, is given by

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \mathbf{X}_{ds}\mathbf{V}_{ds}^{-1}\mathbf{X}'_{ds} \right)^{-1} \sum_{d=1}^{D} \mathbf{X}_{ds}\mathbf{V}_{ds}^{-1}\mathbf{y}_{ds}. \tag{5.4}$$

For areas with negligible sampling fraction, i.e., with $n_d/N_d \approx 0$, the BLUP of the area mean $\bar{Y}_d$ can be written as

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d \left\{ \bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})'\tilde{\boldsymbol{\beta}} \right\} + (1 - \gamma_d)\bar{\mathbf{X}}'_d\tilde{\boldsymbol{\beta}}. \tag{5.5}$$

As $\gamma_d \in (0, 1)$, similar to the case of the FH estimator, the BLUP under the nested error model is a composite estimator, where the role of the direct estimator is now played by $\bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})'\tilde{\boldsymbol{\beta}}$, known as the survey regression (SR) estimator. This SR estimator is composed in (5.5) with the regression-synthetic estimator, $\bar{\mathbf{X}}'_d\tilde{\boldsymbol{\beta}}$. Note that the SR estimator is obtained by fitting the same model (5.1), but by treating the effects of areas u_d as fixed rather than random. This estimator is not efficient for areas with small sample sizes n_d. Therefore, for those areas, the BLUP borrows strength from all the other areas through the regression-synthetic estimator.

To clarify the comment above, let us consider the homoscedastic case where $k_{di} = 1$ for all i and d. In this scenario, $\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_d)$. Hence, for an area with a small sample size n_d, γ_d is close to zero, and the BLUP approaches the regression-synthetic estimator, which borrows information from other areas. However, for an area with a large sample size n_d, γ_d approaches one, and the BLUP approaches the "survey regression" estimator. Note that γ_d also depends on the heterogeneity among areas measured by σ_u^2. If the areas are highly heterogeneous (σ_u^2 is large compared to σ_e^2/n_d), or equivalently, if the considered auxiliary variables do not explain a significant portion of the variability, then γ_d approaches one, and more weight is attached to the "survey regression" estimator, which is similar to a direct estimator. Conversely, if the areas are homogeneous, or in other words, the auxiliary variables are powerful predictors, then more weight is given to the synthetic estimator, obtained through regression with these auxiliary variables.

Once again, the BLUP given in (5.2) depends on the true values of the variance components of the model (5.1), $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$. Replacing the true $\boldsymbol{\theta}$ with a consistent estimator $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ in the BLUP (5.2), we obtain the EBLUP, given by

$$\hat{\bar{Y}}_d^{EBLUP} = N_d^{-1}\left(\sum_{i\in s_d} y_{di} + \sum_{i\in r_d} \hat{y}_{di}\right), \tag{5.6}$$

where, denoting $\hat{\boldsymbol{\beta}}$ to the result of substituting $\boldsymbol{\theta}$ with the estimator $\hat{\boldsymbol{\theta}}$ in $\tilde{\boldsymbol{\beta}}$ given in (5.4), the predicted values are now

$$\hat{y}_{di} = \mathbf{x}_{di}'\hat{\boldsymbol{\beta}} + \hat{u}_d, \quad \hat{u}_d = \hat{\gamma}_d(\bar{y}_{da} - \bar{\mathbf{x}}_{da}'\hat{\boldsymbol{\beta}}), \quad \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/a_d}. \tag{5.7}$$

For a non-sampled area, i.e., with sample size $n_d = 0$, by taking the limiting value $\gamma_d = 0$, we report the regression-synthetic estimator $\hat{\bar{Y}}_d = \bar{\mathbf{X}}_d'\hat{\boldsymbol{\beta}}$.

An analytical estimator of the MSE of the EBLUP, $\hat{\bar{Y}}_d^{EBLUP}$, was obtained under the assumption of normality for random effects and errors

by Prasad and Rao (1990) for negligible sampling fraction (see Rao and Molina, 2015), Section 7.2.2). For the MSE estimator under nonnegligible sampling fraction, see Section 7.2.3 of the same book.

The BLUP is unbiased under the model (5.1) and is optimal in terms of minimum MSE, among linear estimators in the sample data $\mathbf{y}_s$ that are unbiased for $\bar{Y}_d$. When substituting $\boldsymbol{\theta}$ with the estimator $\hat{\boldsymbol{\theta}}$, the EBLUP $\hat{\bar{Y}}_d^{EBLUP}$ remains unbiased under the model (5.1), under certain conditions on the estimator $\hat{\boldsymbol{\theta}}$. Common estimation methods, specifically ML, REML and Henderson's III method, satisfy these conditions. EBLUPs under the nested error model (5.1) may significantly increase efficiency compared to direct estimators and even compared to FH estimators, as they utilize much more detailed information (without reducing the data to means).

A clear disadvantage of small area estimators based on a nested error model, compared to those based on the FH model with design-based direct estimators as response variables, is that the sampling design is not accounted for. Hence, complex and, specially, informative (non-ignorable) sampling might produce a substantial design bias in the resulting EBLUPs. As noted already by Kott (1990) and Prasad and Rao (1999), it is appealing to have design-consistent estimators, which provide protection against model failures for areas with larger sample sizes.

You and Rao (2002a) proposed the Pseudo EBLUP of an area mean $\bar{Y}_d$, which incorporates the sampling weights and therefore accounts for the sampling design. Pseudo EBLUPs are obtained as follows. Taking the weighted average of y_{di} over the sample units in area d, in the homoscedastic nested error model (5.1) obtained taking $k_{di} = 1$, for $i = 1, \ldots, N_d$, we obtain

$$\bar{y}_{dw} = \bar{\mathbf{x}}'_{dw}\boldsymbol{\beta} + u_d + \bar{e}_{dw}, \quad d = 1, \ldots, D, \tag{5.8}$$

where $E(e_{dw}) = 0$ and $\text{var}(e_{dw}) = \sigma_e^2 \hat{N}_d^{-1} \sum_{i \in s_d} w_{di}^2$. On the other hand, taking the average over the population units in area d in the same model, we obtain that $\bar{Y}_d \approx \bar{\mathbf{X}}'_d\boldsymbol{\beta} + u_d$, since by the SLLN, the population mean

of the errors tends to the expected value of the errors, which is zero. The Pseudo BLUP is defined as Henderson (1950)'s BLUP of the linear combination $\bar{\mathbf{X}}_d'\boldsymbol{\beta} + u_d$ under the aggregated nested error model (5.8), that is,

$$\tilde{Y}_{dw} = \bar{\mathbf{X}}_d'\tilde{\boldsymbol{\beta}}_w + \tilde{u}_{dw}, \quad \tilde{u}_{dw} = \gamma_{dw}\left(\bar{y}_{dw} - \bar{\mathbf{x}}_{dw}'\tilde{\boldsymbol{\beta}}_w\right), \quad \gamma_{dw} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 \hat{N}_d^{-1}\sum_{i\in s_d} w_{di}^2}, \tag{5.9}$$

but replacing the WLS estimator of $\boldsymbol{\beta}$ given in (4.5) by

$$\tilde{\boldsymbol{\beta}}_w = \left\{\sum_{d=1}^{D}\sum_{i\in s_d} w_{di}\mathbf{x}_{di}\left(\mathbf{x}_{di} - \gamma_{dw}\bar{\mathbf{x}}_{dw}\right)\right\}^{-1} \sum_{d=1}^{D}\sum_{i\in s_d} w_{di}\left(\mathbf{x}_{di} - \gamma_{dw}\bar{\mathbf{x}}_{dw}\right) y_{di} = \tilde{\boldsymbol{\beta}}_w(\sigma_u^2, \sigma_e^2).$$

The Pseudo EBLUP is obtained by replacing the unknown variance components σ_u^2 and σ_e^2 with consistent estimators in the Pseudo BLUP (5.9). You and Rao (2002a) considered the method of fitting constants, also known as Henderson's method III, and, using this method, provided an analytical estimator for the MSE of the Pseudo EBLUP under the assumption of normality of random effects and errors, following Prasad and Rao (1990). Based on Henderson's method III, Huang and Hidiroglou (2003) derived estimators of the variance components σ_u^2 and σ_e^2 that incorporate the survey weights.

Pseudo EBLUPs are not optimal under the model, but reduce the design bias of EBLUPs under complex designs. They are, in fact, design-consistent as n_d goes to infinity. Moreover, the estimated totals satisfy the convenient benchmarking property of adding up to a design-based estimator of the total at a larger region covering the areas. Consequently, they represent a good alternative to FH estimators, taking into account the sampling design but utilizing much more detailed unit-level information to gain greater efficiency.

In model-based SAE based on unit-level data, a model is assumed for all the population units. However, once the sample is drawn from the population, the model for the sample part $\mathbf{y}_{ds}$ of the population vector

$\mathbf{y}_d = (\mathbf{y}'_{ds}, \mathbf{y}'_{dr})'$ is simply obtained by marginalization, i.e., integrating out with respect to the sample complement, $\mathbf{y}_{dr}$. The model for $\mathbf{y}_{ds}$ (sample model) then has the same shape as the population model for $\mathbf{y}_d$ when the sampling design is ignorable. However, this does not hold in the case of non-ignorable (informative) samplings, where the selection of the units, given the available covariates, depends on the values of the target variable.

Pfeffermann, Krieger and Rinott (1998) and Pfeffermann and Sverchkov (1999, 2003) studied model-based inference under non-ignorable informative sampling. The procedure derives the likelihood of the sample vector $\mathbf{y}_s$ (known as the "sample likelihood") from that of the population $\mathbf{y}$ and from a model assumed for the survey weights, based on Bayes Theorem. By fitting the assumed model for the survey weights and replacing the fitted parameters in the sample likelihood of $\mathbf{y}_s$, this likelihood can then be maximized to obtain the sample model parameters. Pfeffermann and Sverchkov (2007) applied the sample likelihood approach to SAE, obtaining adjusted EBLUPs of area means $\bar{Y}_d$ under informative sampling of areas and of units within areas. Sverchkov and Pfeffermann (2018) extend the sample likelihood approach to correct also for non-ignorable non-response when estimating small area means. The Pseudo EBLUP of You and Rao (2002b), despite not being based on any optimality criteria, also protects against informative sampling, avoiding model assumptions for the survey weights. Verret et al. (2015) proposed another simple approach to deal with informative sampling in SAE, based on adding the survey weight as a covariate in the small area model.

For two-stage sampling, or when estimation is intended at two different (nested) aggregation levels, Stukel and Rao (1999) used a two-fold nested error model with area and subarea random effects. Let C_d be the number of subareas in the population for area d, $d = 1, \ldots, D$. The homoscedastic model is then

$$y_{dci} = \mathbf{x}'_{dci}\boldsymbol{\beta} + u_d + v_{dc} + e_{dci},\ u_d \overset{iid}{\sim} (0, \sigma_u^2),\ v_{dc} \overset{iid}{\sim} (0, \sigma_v^2)$$

$$e_{dci} \overset{iid}{\sim} (0, \sigma_e^2), \quad i = 1, \ldots, N_{dc},\ c = 1, \ldots, C_d,\ d = 1, \ldots, D \tag{5.10}$$

where area effects u_d, subarea effects v_{dc} and errors e_{dci} are all independent. Under this model, they provided EBLUPs of area means $\bar{Y}_d$ and subarea means $\bar{Y}_{dc}$, for sampled and non-sampled subareas.

A nested error model with random slopes was considered by Moura and Holt (1999). A general linear mixed model that includes the nested error model as a special case was studied by Datta and Ghosh (1991). To model simultaneously several dependent response variables, a multivariate extension of the nested error model was introduced by Fuller and Harter (1987), and this model was also studied by Datta, Day and Basawa (1999). Different multivariate extensions were also considered by Lohr and Prasad (2003) and Baíllo and Molina (2009).

The nested error model is designed for continuous response variables, such as income or expenditure. However, some poverty indicators are functions of binary variables, such as having income or expenditure below the poverty line or not. Hence, regression models for binary responses are also of interest. MacGibbon and Tomberlin (1989) proposed a model for binary response variables following Bernoulli distribution, assuming a linear regression model with normally distributed random effects apart from covariates, for the logit of the probability of success. This model belongs to the class generalized linear mixed models (GLMMs). Malec et al. (1997) considered a similar model with random coefficients. An extension of the logistic GLMM to the multinomial distribution for a categorical response variable was proposed originally by Molina, Saei and Lombardía (2007). López-Vizcaíno, Lombardía and Morales (2013) considered an area-level model for multinomial counts, with independent area effects for each multinomial category, and López-Vizcaíno, Lombardía and Morales (2015) extended the latter model by adding time-correlated random effects following an AR(1) process.

Ghosh et al. (1998) proposed generalized linear mixed models (GLMMs) within the natural exponential family, including area and unit random

effects. GLMMs were also studied by Hobza, Marhuenda and Morales (2020). Ghosh et al. (1999) extended the GLMM to include spatial correlation and applied it to disease mapping. Jiang and Larihi (2001) obtained empirical best (EB) predictors under a logistic GLMM at the unit level. The optimal EB predictors under a unit-level logistic mixed model do not have closed-form expressions and hence require computationally intensive methods such as Monte Carlo simulation. Given that the likelihood does not either have a closed form and complex numerical methods are also needed for model fitting, a resampling procedure for MSE estimation of EB predictors might become too computationally intensive. A way to reduce this computational burden is to consider suboptimal predictors such as plug-in predictors. EB and plug-in predictors are compared by Hobza and Morales (2016) and by Molina and Strzalkowska-Kominiak (2020). Hobza, Morales and Santamaría (2018) estimated poverty proportions under unit-level temporal Binomial-Logit mixed models. Berg and Chandra (2014) studied small area prediction for a unit-level log-normal model. Berg (2022, 2023) proposed a unit-level Poisson-Gamma model for counts and Tzavidis et al. (2015) proposed robust small area prediction for counts.

GLMs are also used by Isidro, Haslett and Jones (2016) in the so called extended structure preserving estimation (extended SPREE) method for updating small area estimates of poverty in intercensal years. A couple of methods for SAE estimation under GLMMs for counts are implemented in the R package `saeeb` (Fauziah and Wulansari, 2020).

Models with measurement error in the covariates are classified in Fuller (1987) into functional measurement error models, where the true value of the covariate is assumed to be fixed, and structural models, when it is regarded as random. The former were developed by Ghosh and Sinha (2007) and Datta, Rao and Torabi (2010), while the latter were firstly studied by Fuller and Harter (1987), Ghosh, Sinha and Kim (2006) and Torabi, Datta and Rao (2009).

EBLUPs are affected by misspecification of the model. Jiang, Nguyen and Rao (2014) proposed the OBP for an area mean as a robust alternative to misspecification of a nested error model. EBLUPs are also affected by unit and area-level outliers. Robust versions that are less affected by outliers have been also proposed. A robust EBLUP (REBLUP) under the nested error model was first obtained by Sinha and Rao (2009). Robust bias corrections for the presence of unit and area-level outliers were proposed by Chambers et al (2014), Jiongo, Haziza and Duchesne (2013) following the robust approach for finite populations by Chambers (1986), and Beaumont, Haziza and Ruiz-Gazen (2013) based on conditional bias to measure the influence. Based on the M-quantiles defined by Breckling and Chambers (1988), Chambers and Tzavidis (2006) proposed M-quantile models for SAE of area means, and a bias-adjusted M-quantile estimator was proposed by Tzavidis, Marchetti and Chambers (2010). Fabrizi et al. (2012) studied benchmarking under M-quantile models. Lahiri and Salvati (2023) have recently extended the nested error model to the case of area-specific regression slopes, as well as heteroscedastic error variances, estimating these parameters with robust estimating equations that pool the sample data from all the areas.

Chapter 6

Unit level models: General indicators

The models described in Chapter 5 were in principle intended to estimate only linear indicators, specially, area means. However, most poverty and inequality indicators are non-linear. Note that, even if the target indicators are simple area means, once a non-linear transformation of the target variable is taken as the response variable in the SAE model (such as log, routinely taken for monetary variables to reduce skewness and heteroscedasticity), EBLUPs are not useful anymore. Taking exponentials of EBLUPs in a nested error model with log transformation might result in substantial bias, as shown in Molina and Martín (2018).

Elbers, Lanjouw and Lanjouw (2002, 2003) proposed the first method, known as the ELL, designed to estimate general indicators defined in terms of a continuous monetary variable. This method became very popular and was employed by default within the World Bank until 2020. For example, Bedi, Coudouel and Simler (2007) describe poverty mapping applications in Albania, Bolivia, Bulgaria, Cambodia, Yunnan Province of China, Ecuador, Indonesia, Mexico, Morocco, Sri Lanka, Thailand and Vietnam. For a later application of ELL method, see Farris et al. (2017),

who estimate poverty in Uganda.

The ELL method assumes the nested error model of Battese, Harter and Fuller (1988) given in (5.1) for the log of a welfare measure, but with random effects for the sampling clusters (or primary sampling units), which are typically nested within the areas. Assume there are C_d clusters in area d, and that the number of population units in cluster c from area d is N_{dc}, for $c = 1, \ldots, C_d$ and $d = 1, \ldots, D$. Let y_{dci} denote the natural log of the welfare measure z_{dci} for unit i within cluster c from area d. Then, the original ELL method assumes the model

$$\begin{aligned} &y_{dci} = \mathbf{x}'_{dci}\boldsymbol{\beta} + u_{dc} + e_{dci},\ u_{dc} \overset{iid}{\sim} (0, \sigma_u^2), \\ &e_{dci} \overset{ind}{\sim} (0, \sigma_{dci}^2), \quad i = 1, \ldots, N_{dc},\ c = 1, \ldots, C_d,\ d = 1, \ldots, D, \end{aligned} \tag{6.1}$$

where $\mathbf{x}_{dci}$ is a $1 \times p$ vector of auxiliary variables for unit i within cluster c from area d and $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients. Again, the first component of $\mathbf{x}_{dci}$ is customarily set to one, and the remaining values may vary by units, clusters, or areas. In this model, u_{dc} and e_{dci} are respectively cluster and unit-specific idiosyncratic errors, assumed to be independent of each other. Error variances σ_{dci}^2 are previously estimated using an additional regression model for squared unit-level residuals, in terms of possibly different covariates, called the *alpha model*.

In the case where clusters are equal to areas, or equivalently, a single cluster within each area d, we have $C_d = 1$ and $N_{dc} = N_d$, for all the areas $d = 1, \ldots, D$. In that case, we can remove the subscript c, and the model (6.1) reduces to the BHF model given in (5.1), with area effects u_d and heteroscedastic model errors e_{di}, that is, with $\text{var}(e_{di}) = \sigma_{di}^2$.

The ELL estimator of a general indicator $\delta_d = \delta_d(\mathbf{y}_d)$ is obtained using a bootstrap procedure that approximates the marginal expectation $\hat{\delta}_d^{ELL} = E[\delta_d]$ under the model (6.1). In contrast, as will be seen later, the EB predictor is approximately the optimal predictor in terms of MSE, given by the conditional expectation $E[\delta_d|\mathbf{y}_{ds}]$, where $\mathbf{y}_{ds}$ is the vector of sample observations of the response variable in area d. The ELL

bootstrap procedure delivers also an estimate of the MSE of the ELL estimator.

We outline here the original bootstrap procedure of Elbers, Lanjouw and Lanjouw (2003). A later implemented version of the ELL bootstrap procedure is described in Corral, Molina and Nguyen (2021), Section 2. For simplicity, here we describe the procedure for the case in which all the clusters are sampled. In this procedure, we denote by s_{dc} the set of units sampled from cluster c within area d, of size n_{dc}, $c = 1, \ldots, C_d$, $d = 1, \ldots, D$. We also use $\mathbf{y}_s = (\mathbf{y}'_{1s}, \ldots, \mathbf{y}'_{Ds})'$ for the vector with the sample data from the response variable for all the areas, and $\mathbf{X}_s = (\mathbf{X}'_{1s}, \ldots, \mathbf{X}'_{Ds})'$ for the corresponding matrix with the sample data on the auxiliary variables.

ELL bootstrap procedure:

1. Fit the model (6.1) by ordinary LS to the sample data, leading to the OLS estimator $\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{y}_s$ of $\boldsymbol{\beta}$. Then, obtain marginal residuals $y_{dci} - \mathbf{x}'_{dci}\hat{\boldsymbol{\beta}}_0$ for each sample unit from cluster c in area d, i.e., for each $i \in s_{dc}$, $c = 1, \ldots, C_d$, $d = 1, \ldots, D$.

2. Draw regression coefficients from their estimated distribution as follows
$$\boldsymbol{\beta}^* \sim N_p\left(\hat{\boldsymbol{\beta}}_0, \widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_0)\right),$$

3. From the marginal residuals obtained in step 1, predict the cluster effects as $\hat{u}_{dc0} = n_{dc}^{-1}\sum_{i\in s_{dc}}(y_{dci} - \mathbf{x}'_{dci}\hat{\boldsymbol{\beta}}_0)$, $c = 1, \ldots, C_d$, $d = 1, \ldots, D$. Then, generate randomly bootstrap cluster effects u^*_{dc} for all the clusters $c = 1, \ldots, C_d$, $d = 1, \ldots, D$, from the empirical distribution of $\{\hat{u}_{dc0},\ c = 1, \ldots, C_d,\ d = 1, \ldots, D\}$.

4. Construct now the conditional residuals for each sample unit as $\hat{e}_{dci} = y_{dci} - \mathbf{x}'_{dci}\hat{\boldsymbol{\beta}}_0 - \hat{u}_{dc0}$, $i \in s_{dc}$, $c = 1, \ldots, C_d$, $d = 1, \ldots, D$. Then, draw individual errors e^*_{dci}, for each population unit $i = 1, \ldots, N_{dc}$, $c = 1, \ldots, C_d$, $d = 1, \ldots, D$, from the empirical distribution of the sample conditional residuals $\hat{e}_{dci}$.

5. Using the bootstrap conditional residuals from step 4, along with the generated $\boldsymbol{\beta}^*$ in step 2, and making use of the values of the auxiliary variables for all the population units, generate bootstrap values of the response variable for all individuals in the population, as follows:

$$y^*_{dci} = \mathbf{x}'_{dci}\boldsymbol{\beta}^* + u^*_{dc} + e^*_{dci}, \quad i = 1, \ldots, N_{dc},\ c = 1, \ldots, M_d,\ d = 1, \ldots, D.$$

This provides us with a complete census of the response variable, from which we calculate the true value of the indicator for area d,

$$\delta^*_d = h_d(y^*_{d11}, \ldots, y^*_{d1N_{d1}}, \ldots, y^*_{dC1}, \ldots, y^*_{dCN_{dC}}).$$

6. Repeat steps 1–5 for $r = 1, \ldots, R$, resulting in R complete censuses. For each census r, we compute the indicator of interest $\delta^{*(r)}_d$. The ELL estimator of δ_d is then obtained by averaging the true values $\delta^{*(r)}_d$ over the R censuses, as follows:

$$\hat{\delta}^{ELL}_d = \frac{1}{R}\sum_{r=1}^{R} \delta^{*(r)}_d.$$

As an estimator of the MSE of the ELL estimator, the ELL method uses the bootstrap variance, given by

$$\text{mse}_{ELL}(\hat{\delta}^{ELL}_d) = \frac{1}{R}\sum_{r=1}^{R} (\delta^{*(r)}_d - \hat{\delta}^{ELL}_d)^2.$$

Consider now that we wish to estimate the area mean $\delta_d = \bar{Y}_d$ using the ELL bootstrap procedure, keeping $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}_0$ fixed across the bootstrap replicates for simplicity. Let $y^{*(r)}_{dci}$ be the values generated in step 5, in the r-th bootstrap replicate of the ELL bootstrap procedure. According to the bootstrap generating process, for $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}_0$, the bootstrapped value of the area mean in the r-th replicate is

$$\bar{Y}^{*(r)}_d = \frac{1}{N_d}\sum_{c=1}^{C_d}\sum_{i=1}^{N_{dc}} y^{*(r)}_{dci} = \bar{\mathbf{X}}'_d\hat{\boldsymbol{\beta}}_0 + \frac{1}{N_d}\sum_{c=1}^{C_d} N_{dc}u^{*(r)}_{dc} + \frac{1}{N_d}\sum_{c=1}^{C_d}\sum_{i=1}^{N_{dc}} e^{*(r)}_{dci}.$$

Then, the ELL estimator of $\delta_d = \bar{Y}_d$ is obtained by taking the average of $\bar{Y}_d^{*(r)}$ over the R bootstrap replicates,

$$\hat{\bar{Y}}_d^{ELL} = \frac{1}{R}\sum_{r=1}^{R} \bar{Y}_d^{*(r)} = \bar{\mathbf{X}}_d'\hat{\boldsymbol{\beta}}_0 + \frac{1}{N_d}\sum_{c=1}^{C_d} N_{dc}\frac{1}{R}\sum_{r=1}^{R} u_{dc}^{*(r)} + \frac{1}{N_d}\sum_{c=1}^{C_d}\sum_{i=1}^{N_{dc}}\frac{1}{R}\sum_{r=1}^{R} e_{dci}^{*(r)}. \tag{6.2}$$

However, note that, for large R, the averages over the R bootstrap replicates of the cluster random effects $u_{dc}^{*(r)}$ and of the errors $e_{dci}^{*(r)}$ are zero by the strong law of large numbers,

$$\frac{1}{R}\sum_{r=1}^{R} u_{dc}^{*(r)} \approx E(u_{dc}) = 0, \quad \frac{1}{R}\sum_{r=1}^{R} e_{dci}^{*(r)} \approx E(e_{dci}) = 0. \tag{6.3}$$

This means that the ELL bootstrap procedure makes cluster effects u_{dc} vanish, even if they are actually included in ELL model (6.1). Then, replacing (6.3) in (6.2), the resulting ELL estimator reduces to the regression-synthetic estimator, which does not account for the cluster effect,

$$\hat{\bar{Y}}_d^{ELL} \approx \bar{\mathbf{X}}_d'\hat{\boldsymbol{\beta}}_0. \tag{6.4}$$

This issue occurs because the ELL estimator approximates the marginal mean $E[\delta_d]$, without conditioning on the available sample data $\mathbf{y}_{ds}$. Hence, it does not utilize the sample observations $\mathbf{y}_{ds}$ available from area d, relying solely on the linear regression without the area effects. Consequently, the ELL estimator faces the same issues as the regression-synthetic estimator; specifically, it can be highly design-biased if the regression model is not correctly specified, and has substantial MSE when the considered auxiliary variables do not sufficiently explain the between-area heterogeneity.

To reduce between-area heterogeneity, Elbers, Lanjouw and Lanjouw (2002) recommended the inclusion of contextual (aggregated) covariates, such as location means of household characteristics and those derived from Geographic Information Systems (GIS). Haslett (2016) also noted the importance of including such contextual covariates. Although any

type of auxiliary information is welcome as long as it is useful to predict welfare, in practice there is no guarantee that area heterogeneity has been completely accounted for by the available covariates, unless a formal statistical test for $\sigma_u^2 = 0$ was conducted in each application. In any case, estimators that do incorporate area effects converge to their synthetic counterparts whenever σ_u^2 becomes small, so there is no reason for excluding area effects.

An additional problem with the traditional ELL bootstrap method is that the noise measures delivered do not correctly estimate the true error measures of ELL estimators. The reason is that, unlike in usual bootstrap methods, the ELL bootstrap procedure does not re-fit the model and re-estimate the target indicators with each bootstrap sample (which should be drawn from each bootstrap census). Hence, the real-world estimation process is not adequately replicated in the bootstrap world. As a result, the estimated MSE through this method does not accurately replicate the error incurred in estimation in the real world. Finally, when the sampling clusters are not the areas of interest, the MSE of the ELL estimator for an area indicator δ_d can be severely underestimated if the area effects are not well explained (i.e., when their variance σ_u^2 is significant).

The ELL method is implemented in the `PovMap` software (Zhao, 2006), which offers a simple point-and-click user interface and is considerably fast and efficient in terms of memory. The `sae` Stata command developed by Nguyen et al. (2018) replicates most of the procedures implemented in the `PovMap` software.

Banerjee et al. (2006) conducted a review of the research at the World Bank and raised several concerns about the ELL method. First of all, they suggested that ELL was not accounting for potential area effects. This diagnosis was correct even if clusters in the ELL model are equal to the areas, because the cluster/area effects u_{dc} vanish when applying ELL bootstrap procedure according to (6.3)–(6.4). They also realized that ELL estimated sampling errors were not accounting for the correlation

between observations in different clusters within the same area. These two problems were eventually solved by the Empirical Best (EB) method and the bootstrap MSE estimator proposed by Molina and Rao (2010).

The EB method is similar to ELL in that it combines survey data with auxiliary data, obtained typically (but not exclusively) from a census or administrative records, uses a unit-level model and is designed for general (and perhaps several) indicators that depend on a welfare measure, based on a single model. This original EB method of Molina and Rao (2010) assumes that a one-to-one transformation of the welfare variable, $y_{di} = T(z_{di})$, follows the nested error model (5.1) with normality for the area effects u_d and the errors e_{di}. Under this model, the area vectors $\mathbf{y}_d = (y_{d1}, \ldots, y_{dN_d})'$, $d = 1, \ldots, D$, are independent and satisfy $\mathbf{y}_d \overset{ind}{\sim} N(\boldsymbol{\mu}_d, \mathbf{V}_d)$, with mean vectors $\boldsymbol{\mu}_d = \mathbf{X}_d\boldsymbol{\beta}$, and covariance matrices $\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{A}_d$, where $\mathbf{A}_d = \text{diag}(k_{di}^2; i = 1, \ldots, N_d)$.

The best predictor of a general indicator $\delta_d = h_d(\mathbf{y}_d)$ is defined as the function of the sample observations, $\tilde{\delta}_d = \tilde{\delta}_d(\mathbf{y}_s)$, which minimizes the MSE under the model, given by $\text{MSE}(\tilde{\delta}_d) = E\left[(\tilde{\delta}_d - \delta_d)^2\right]$. The minimizer is given by

$$\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\mathbf{y}_{dr}}[h_d(\mathbf{y}_d)|\mathbf{y}_{ds}; \boldsymbol{\theta}], \tag{6.5}$$

where the expectation is taken with respect to the distribution of the out-of-sample vector $\mathbf{y}_{dr}$ given the in-sample values $\mathbf{y}_{ds}$ from area d. Since under the nested error model (5.1), the area vector $\mathbf{y}_d$ follows a normal distribution, the distribution of $\mathbf{y}_{dr}|\mathbf{y}_{ds}$, required to calculate the best predictor (6.5), is also normal, given by

$$\mathbf{y}_{dr}|\mathbf{y}_{ds} \overset{ind}{\sim} N(\boldsymbol{\mu}_{dr|s}, \mathbf{V}_{dr|s}), \quad d = 1, \ldots, D, \tag{6.6}$$

where the vector of conditional means and the corresponding covariance matrix take the form

$$\boldsymbol{\mu}_{dr|s} = \mathbf{X}_{dr}\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\mathbf{x}}'_{da}\boldsymbol{\beta})\mathbf{1}_{N_d - n_d}, \tag{6.7}$$

$$\mathbf{V}_{dr|s} = \sigma_u^2(1 - \gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}'_{N_d-n_d} + \sigma_e^2 \text{diag}_{i \in r_d}(k_{di}^2). \tag{6.8}$$

For a single out-of-sample unit $i \in r_d$, we have

$$y_{di}|\mathbf{y}_{ds} \sim N(\mu_{di|s}, \sigma^2_{di|s}), \tag{6.9}$$

where the conditional mean and variance are given by

$$\mu_{di|s} = \mathbf{x}'_{di}\boldsymbol{\beta} + \tilde{u}_d, \quad \tilde{u}_d = \gamma_d(\bar{y}_{da} - \bar{\mathbf{x}}'_{da}\boldsymbol{\beta}), \tag{6.10}$$

$$\sigma^2_{di|s} = \sigma^2_u(1-\gamma_d) + \sigma^2_e k^2_{di}. \tag{6.11}$$

This conditional distribution depends on the vector of model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2_u, \sigma^2_e)'$. The empirical best predictor (EB) is obtained by replacing $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2_u, \sigma^2_e)'$ with a consistent estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}^2_u, \hat{\sigma}^2_e)'$ in the best predictor (6.5), that is, $\hat{\delta}^{EB}_d = \tilde{\delta}^B_d(\hat{\boldsymbol{\theta}})$. Standard estimation methods that provide consistent estimators even if normality does not hold under certain regularity assumptions, are ML and REML methods using the normal likelihood, and Henderson's III method.

The conditional expectation defining the best predictor accepts a closed form for certain indicators $\delta_d = h_d(\mathbf{y}_d)$, such as for the FGT indicators for $\alpha \in \{0, 1\}$, but not in general. A general procedure to approximate the EB predictor is by using Monte Carlo (MC) simulation. The MC simulation procedure proposed by Molina and Rao (2010) is described next.

Monte Carlo simulation procedure for EB estimator:

1. Fit the model (5.1) to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$, obtaining an estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}^2_u, \hat{\sigma}^2_e)'$ of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2_u, \sigma^2_e)'$.
2. Generate, for $\ell = 1, \ldots, L$, vectors of response variables for out-of-sample individuals in area d, $\mathbf{y}^{(\ell)}_{dr}$, based on the distribution of $\mathbf{y}_{dr}|\mathbf{y}_{ds}$ given in (6.6)–(6.8), with $\boldsymbol{\theta}$ replaced by its estimator $\hat{\boldsymbol{\theta}}$ obtained in step 1.
3. Augment the generated vector $\mathbf{y}^{(\ell)}_{dr}$ with the in-sample data $\mathbf{y}_{ds}$ to form a census vector for area d, $\mathbf{y}^{(\ell)}_d = (\mathbf{y}'_{ds}, (\mathbf{y}^{(\ell)}_{dr})')'$. Using the census vector $\mathbf{y}^{(\ell)}_d$, calculate the indicator of interest $\delta^{(\ell)}_d = h_d(\mathbf{y}^{(\ell)}_d)$.

4. Repeat steps 2-3 for $\ell = 1, \ldots, L$. The MC approximation of the EB predictor of δ_d is obtained by averaging the indicators over the L simulated censuses, that is,

$$\hat{\delta}_d^{EB} = \frac{1}{L} \sum_{\ell=1}^{L} \delta_d^{(\ell)}. \tag{6.12}$$

In step 2, it is necessary to simulate L vectors $\mathbf{y}_{dr}^{(\ell)}$ from a multivariate Normal distribution of size $N_d - n_d$ (which can be huge in many real applications) with covariance matrix $\mathbf{V}_{dr|s}$, making the process computationally unfeasible. Molina and Rao (2010) proposed a procedure that avoids the generation of huge multivariate normal vectors, by noting that $\mathbf{V}_{dr|s}$ given in (6.8) is the covariance matrix of a random vector $\mathbf{y}_{dr}^{(\ell)}$, whose elements arise from the following nested error model,

$$y_{di}^{(\ell)} = \mathbf{x}_{di}'\hat{\boldsymbol{\beta}} + \hat{u}_d + v_d^{(\ell)} + \epsilon_{di}^{(\ell)}, \quad i \in r_d,\ d = 1, \ldots, D, \tag{6.13}$$

where $v_d^{(\ell)}$ and $\epsilon_{di}^{(\ell)}$ are all independent and satisfy, respectively,

$$v_d^{(\ell)} \overset{ind}{\sim} N(0, \hat{\sigma}_u^2(1 - \hat{\gamma}_d)),\ \epsilon_{di}^{(\ell)} \overset{ind}{\sim} N(0, \hat{\sigma}_e^2 k_{di}^2). \tag{6.14}$$

Using the model (6.13)–(6.14) to generate the non-sample elements $y_{di}^{(\ell)}$, $i \in r_d$, instead of generating a multivariate Normal vector of size $N_d - n_d$, we only need to generate $1 + N_d - n_d$ independent Normal random variables, $v_d^{(\ell)} \overset{ind}{\sim} N(0, \sigma_u^2(1 - \gamma_d))$ and $\epsilon_{di}^{(\ell)} \overset{ind}{\sim} N(0, \sigma_e^2 k_{di}^2)$, for $i \in r_d$. By using the vector $\mathbf{y}_{dr}^{(\ell)}$ with these generated out-of-sample elements $y_{di}^{(\ell)}$, $i \in r_d$, in step 3, we construct the census vector $\mathbf{y}_d^{(\ell)} = (\mathbf{y}_{ds}', (\mathbf{y}_{dr}^{(\ell)})')'$ and calculate the indicator of interest $\delta_d^{(\ell)} = h_d(\mathbf{y}_d^{(\ell)})$.

For an area d that is not sampled (i.e., with $n_d = 0$), as there is no in-sample part in this case, we generate the whole census vector $\mathbf{y}_d^{(\ell)} = \mathbf{y}_{dr}^{(\ell)}$ from the model (6.13) by setting $\gamma_d = 0$ (limiting value as $n_d \to 0$).

Obtaining an analytical estimator for the MSE of the EB predictor $\hat{\delta}_d^{EB}$ of δ_d in general is challenging. Molina and Rao (2010) proposed a parametric

bootstrap method, based on the bootstrap procedure for finite populations introduced by González-Manteiga et al. (2008). This method is described below.

Bootstrap procedure for MSE estimation:

1. Fit model (5.1) to the sample data $\mathbf{y}_s = (\mathbf{y}'_{1s}, \ldots, \mathbf{y}'_{Ds})'$, obtaining estimates of the model parameters, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$.

2. Generate bootstrap area effects as follows:

$$u_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2), \quad d = 1, \ldots, D.$$

   Generate, independently of $u_1^{*(b)}, \ldots, u_D^{*(b)}$, bootstrap errors as follows:

$$e_{di}^{*(b)} \overset{ind}{\sim} N(0, \hat{\sigma}_e^2 k_{di}^2), \quad i = 1, \ldots, N_d,\ d = 1, \ldots, D.$$

   Then, generate a bootstrap population (or census) of the response variable through the model,

$$y_{di}^{*(b)} = \mathbf{x}'_{di}\hat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i = 1, \ldots, N_d,\ d = 1, \ldots, D.$$

3. Define the census vector of response variables for area d, $\mathbf{y}_d^{(b)} = (y_{d1}^{(b)}, \ldots, y_{dN_d}^{(b)})'$, and calculate from it the indicator of interest, $\delta_d^{(b)} = h_d(\mathbf{y}_d^{(b)})$, $d = 1, \ldots, D$.

4. For the original sample $s = s_1 \cup \cdots \cup s_D$, let $\mathbf{y}_s^{*(b)} = ((\mathbf{y}_{1s}^{*(b)})', \ldots, (\mathbf{y}_{Ds}^{*(b)})')'$ be the vector containing the bootstrap observations whose indices are in the sample, i.e., it contains the variables $y_{di}^{*(b)}$, $i \in s_d$, $d = 1, \ldots, D$. Fit the model (5.1) again to the bootstrap sample data $\mathbf{y}_s^{(b)}$ and obtain the bootstrap EB predictors $\hat{\delta}_d^{EB*(b)}$, $d = 1, \ldots, D$.

5. Repeat steps 2–4 for $b = 1, \ldots, B$, and obtain the true values, $\delta_d^{*(b)}$, and the corresponding bootstrap EB predictors, $\hat{\delta}_d^{EB*(b)}$, for each area $d = 1, \ldots, D$, and for each bootstrap replicate, $b = 1, \ldots, B$.

6. The "naive bootstrap" estimator of the MSE of the EB predictor, $\hat{\delta}_d^{EB}$, is given by

$$\text{mse}_B(\hat{\delta}_d^{EB}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_d^{EB*(b)} - \delta_d^{*(b)}\right)^2, \quad d = 1, \ldots, D.$$

If the model is correctly specified, the best predictor is exactly model unbiased and has minimum MSE under the model. The "empirical" version obtained by replacing the unknown model parameters with estimators are neither exactly unbiased nor optimal. Nevertheless, when using consistent estimators of the model parameters as the total sample size n grows, EB is expected to be nearly unbiased and nearly optimal for large total sample size n. The gains in efficiency of EB with respect to ELL may be remarkable when the nested error model assumptions hold and area effects are significant (equivalent to a poor explanatory power of auxiliary variables), as illustrated by Molina and Rao (2010) and corroborated by the results of many different authors, see, for example, Corral Rodas, Molina and Nguyen (2021).

The EB method under a homoscedastic model is implemented in the `sae` R package by Molina and Marhuenda (2015), in the `emdi` by Kreutzmann et al. (2019) and the more recent updating of the later, the `povmap` R package (Edochie et al., 2023). It is also implemented in the Stata `sae` command by Nguyen et al. (2018), see https://github.com/pcorralrodas/SAE-Stata-Package. The EB method has been applied, for example, to estimate poverty indicators in Spanish provinces by gender in 2006 (Molina and Rao, 2010), mean income in Mexican municipalities (Molina and Martín, 2018), mean income and (non-extreme) poverty rates for census tracts by gender in Montevideo, Uruguay, and poverty rates and gaps in Palestinian localities by gender (Molina Peralta and García Portugués, 2020).

Note that to estimate non-linear indicators, both ELL and EB methods require two data sources: a survey with unit-level data of the variable of

interest and of the auxiliary variables for all areas, $\{(z_{di}, \mathbf{x}_{di}); i \in s_d,\ d = 1, \ldots, D\}$, as well as a census with the values of the same auxiliary variables for all the population units, $\{\mathbf{x}_{di}; i = 1, \ldots, N_d,\ d = 1, \ldots, D\}$. The original EB estimator proposed by Molina and Rao (2010) additionally requires identifying the units in the census that are also in the sample for each area, a process that was not needed in the ELL approach. Linking the survey and census data files for each area is not always feasible in practice. Actually, the original EB method assumes that, for each area d, the survey sample s_d is contained in the set of census units, which is regarded as U_d. In practice, these two data files might come from different time instants, and the census set of units does not necessarily include the set of survey units. Hence, $\mathbf{y}_{ds}$ is not necessarily a sub-vector of $\mathbf{y}_d$, $d = 1 \ldots, D$.

Correa, Molina, and Rao (2012) proposed the *Census* EB (CEB) predictor, which does not require identifying the survey sample units in the census file and delivers practically the same point estimate as EB, if the sampling fraction n_d/N_d is small. The CEB predictor of $\delta_d = h_d(\mathbf{y}_d)$ is the EB predictor of $\delta_d = h_d(\mathbf{y}_d)$, assuming that the augmented vectors $\mathbf{y}_{da} = (\mathbf{y}_{ds}, \mathbf{y}_d)'$, composed by the survey and the census vectors of model responses, $d = 1, \ldots, D$, follow the nested error model (5.1). Hence, the CEB predictor of δ_d is the EB predictor $\hat{\delta}_d^{CEB} = E_{\mathbf{y}_d}[\delta_d | \mathbf{y}_{ds}; \hat{\theta}]$ obtained under the nested error model for $\mathbf{y}_{da}$, $d = 1, \ldots, D$. A general MC simulation procedure that approximates the CEB predictors is then adapted from that for EB predictors. This procedure is described next.

Monte Carlo simulation procedure for Census EB estimator:

1. Fit the model (5.1) to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$, obtaining an estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$.
2. Generate, for $\ell = 1, \ldots, L$, a census vector $\mathbf{y}_d^{(\ell)}$, whose elements $y_{di}^{(\ell)}$, are generated as follows

$$y_{di}^{(\ell)} = \mathbf{x}_{di}'\hat{\boldsymbol{\beta}} + \hat{u}_d + v_d^{(\ell)} + \epsilon_{di}^{(\ell)}, \quad i \in r_d,\ d = 1, \ldots, D, \tag{6.15}$$

where $v_d^{(\ell)}$ and $\epsilon_{di}^{(\ell)}$ are generated independently, as follows

$$v_d^{(\ell)} \sim N(0, \hat{\sigma}_u^2(1-\hat{\gamma}_d)),\ \epsilon_{di}^{(\ell)} \sim N(0, \hat{\sigma}_e^2 k_{di}^2). \tag{6.16}$$

3. With the generated census vector $\mathbf{y}_d^{(\ell)}$, calculate the indicator $\delta_d^{(\ell)} = h_d(\mathbf{y}_d^{(\ell)})$.
4. Repeat step 3 for $\ell = 1, \ldots, L$. The MC approximation of the EB predictor of δ_d is obtained by averaging the indicators over the L simulated censuses:

$$\hat{\delta}_d^{EB} = \frac{1}{L}\sum_{\ell=1}^{L} \delta_d^{(\ell)}. \tag{6.17}$$

The MSEs of the Census EB estimators may be estimated through a slight modification of the bootstrap procedure described for EB. The difference between the EB and Census EB estimators lies in the fact that the units from the survey cannot be identified in the census. Therefore, in each bootstrap repetition, we cannot generate census vectors $\mathbf{y}_d^{*(b)}$, $d = 1, \ldots, D$, and then take the sample part $\mathbf{y}_s^{(b)}$ from them. For the Census EB predictors, in step 2, we generate the bootstrap census vector $\mathbf{y}_d^{(b)}$, $d = 1, \ldots, D$, using the values of the auxiliary variables from the census, $\mathbf{X}_d$, $d = 1, \ldots, D$. In step 3, the true values of the bootstrap indicators are obtained from the bootstrap census vectors, $\delta_d^{*(b)} = h_d(\mathbf{y}_d^{*(b)})$. In step 4, the bootstrap sample vector $\mathbf{y}_{ds}^{*(b)}$ is not extracted from $\mathbf{y}_d^{*(b)}$; instead, it is generated again using the values of the same auxiliary variables, but this time taken from the survey, $\mathbf{X}_{ds}$, $d = 1, \ldots, D$.

Corral Rodas, Molina and Nguyen (2021) extended the model-based simulation experiment of Molina and Rao (2010) to more realistic scenarios with a much better explanatory power of the model, including contextual variables, using much larger area population sizes and much smaller sampling fractions, generating errors from a Student's t_5 distribution instead of a normal distribution, and also decreasing the overall sample size and the area sample sizes. Additionally, Corral et al. (2021) performed a design-based validation study, using the Mexican Intracensal Survey

as a fixed census and rawing 500 samples from it using a realistic sampling method. In these simulation experiments, the CEB estimates were practically identical to EB ones, and both performed generally better, in terms of MSE, than the traditional ELL.

As the EBLUP, neither the original ELL estimators, nor the EB or the Census EB estimators based on the nested error model, account for the sampling design, and might be biased under informative sampling. Similar to the Pseudo EBLUP of You and Rao (2002a), Guadarrama, Molina and Rao (2014) defined Pseudo EB predictors that incorporate the survey weights in the EB procedure. They noticed that the conditional distribution (6.9) defining the EB predictor under the homoscedastic nested error model (5.1) with $k_{di} = 1$ for all i and d, depends on the sample data $\mathbf{y}_{ds}$ only through the (unweighted) area sample mean $\bar{y}_d$, that is, $\hat{\delta}_d^{EB} = E_{\mathbf{y}_{dr}}(\delta_d|\bar{y}_d)$. Then they defined the Pseudo EB predictor of δ_d as $\hat{\delta}_d^{PEB} = E_{\mathbf{y}_{dr}}(\delta_d|\bar{y}_{dw})$, where the expectation is now taken with respect to the distribution of out-of-sample units, given the weighted area means $\bar{y}_{dw}$. The resulting conditional distribution, for each out-of-sample unit y_{di}, $i \in r_d$, is now

$$y_{di}|\bar{y}_{dw} \overset{ind}{\sim} N(\mu^w_{di|s}, \sigma^{2w}_{di|s}), \quad i \in r_d, \tag{6.18}$$

where the conditional mean and variance are given by

$$\mu^w_{di|s} = \mathbf{x}'_{di}\boldsymbol{\beta} + \tilde{u}_{dw}, \quad \tilde{u}_{dw} = \gamma_{dw}(\bar{y}_{dw} - \bar{\mathbf{x}}'_{dw}\boldsymbol{\beta}), \tag{6.19}$$

$$\sigma^{2w}_{di|s} = \sigma^2_u(1 - \gamma_{dw}) + \sigma^2_e, \tag{6.20}$$

for γ_{dw} defined in (5.9). Guadarrama, Molina and Rao (2014) adapted the Monte Carlo simulation procedure for EB estimation to the Pseudo EB and proposed a parametric bootstrap procedure to estimate the MSE of the Pseudo EB predictor. A Census version of the Pseudo EB predictor, similar to the Census EB, can be easily defined.

For an area mean $\bar{Y}_d$, the Pseudo EB reduces to the Pseudo EBLUP of You and Rao (2002a) given in (5.9), when using the same estimators of

the model parameters. Hence, it inherits the good design properties of the Pseudo EBLUP for $\bar{Y}_d$; that is, it is design consistent as the area sample size n_d becomes large and satisfies the self-benchmarking property.

Van der Weide (2014) modified the methodology implemented in the `PovMap` sofware, attempting to imitate the EB approach, but under the original ELL model given in (6.1). In the modified methodology, the generation of censuses of the response variable is done by retaining the predicted cluster effects obtained under normality, rather than generating cluster effects from the empirical distribution of average marginal residuals, as in ELL bootstrap procedure. Instead of the raw best predictors of the cluster effects, they considered the predicted cluster effects from the Pseudo EB of You and Rao (2002a), analogous to $\tilde{u}_{dw}$ given in (6.19) but for the clusters, and previously extended to the heteroscedastic case of $\text{var}(e_{dci}) = \sigma^2_{dci}$, $i = 1, \ldots, N_d$, $d = 1, \ldots, D$. These extended predicted cluster effects are given by

$$\tilde{u}_{dcw,2} = \gamma_{dcw,2}(\bar{y}_{dcw,2} - \bar{\mathbf{x}}'_{dcw,2}\boldsymbol{\beta}), \quad \gamma_{dcw,2} = \frac{\sigma^2_u}{\sigma^2_u + \sum_{i\in s_{dc}} w^2_{dci}\left(\sum_{i\in s_{dc}} w_{dci} \sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}}\right)^{-1}}, \tag{6.21}$$

where

$$\bar{y}_{dcw,2} = \left\{\sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}}\right\}^{-1} \sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}} y_{dci}, \quad \bar{\mathbf{x}}_{dcw,2} = \left\{\sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}}\right\}^{-1} \sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}} \mathbf{x}_{dci}.$$

To estimate the regression parameter $\boldsymbol{\beta}$, they extended the survey-weighted estimator $\tilde{\boldsymbol{\beta}}_w$ of You and Rao (2002a) to the heteroscedastic model. This estimator is given by

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{w,2} &= \left\{\sum_{d=1}^{D}\sum_{c\in C}\left(\sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}}\mathbf{x}_{dci}\mathbf{x}'_{dci} - \gamma_{dcw,2}\sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}}\bar{\mathbf{x}}_{dcw}\bar{\mathbf{x}}'_{dcw}\right)\right\}^{-1} \\ &\times \sum_{d=1}^{D}\sum_{c\in C}\left(\sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}}\mathbf{x}_{dci}y_{dci} - \gamma_{dcw,2}\sum_{i\in s_{dc}} \frac{w_{dci}}{\sigma^2_{dci}}\bar{\mathbf{x}}_{dcw}\bar{y}_{dcw}\right). \end{aligned} \tag{6.22}$$

Expressions for the variances of $\tilde{u}_{dcw,2}$ and $\tilde{\boldsymbol{\beta}}_{w,2}$ were also provided.

Under homoscedasticity and with clusters equal to the areas, the predicted area effects $\tilde{u}_{dcw,2}$ are identical to those of the Pseudo EB in Guadarrama, Molina and Rao (2014), $\tilde{u}_{dw}$, given in (6.19). Additionally, if no survey weights are considered ($w_{di} = 1$ for all i and d), $\tilde{u}_{dcw,2}$ reduces to the best predictor of the area effect $\tilde{u}_d$ of Molina and Rao (2010), given in (6.10).

The new bootstrap procedure proposed by Van der Weide (2014) draws a sample of clusters from the present in the survey. Using the data from this bootstrap sample of clusters, the model (6.1) is fitted, and predicted cluster effects $\tilde{u}_{dcw,2}$ are calculated for the clusters c that appear in the bootstrap sample of clusters. For those that do not appear, predicted cluster effects are set to zero. These predicted cluster effects are then retained in the generation of the census data of the response variable. In all the simulation experiments conducted by Corral et al. (2021), this new procedure, referred to as clustered bootstrap-EB (CB-EB), appeared to be seriously biased, while EB and Census EB performed better than ELL and CB-EB even in simulation scenarios without normality.

In an attempt to combine the positive aspects of each procedure, Corral et al. (2021) extended the Census Pseudo EB estimators of Guadarrama, Molina and Rao (2018), to the heteroscedastic nested error model by using (6.21) and (6.22), for the case when the clusters are equal to the areas ($C_d = 1$ for all d). They also incorporated the survey weights in the estimates of the variance components, as proposed by Van der Weide (2014). Furthermore, they adapted the parametric bootstrap procedure for MSE estimation of the Census EB, as described above, to the extended Census Pseudo EB estimators. The Stata `sae` command was updated by incorporating these extensions; for more details on the implemented computational procedures, refer to see Section 6 of Corral et al. (2021).

The Pseudo EB by Guadarrama, Molina and Rao (2014) accommodates the survey weights and is expected to mitigate the design bias of EB in cases of informative sampling (or sample selection bias). However, this

predictor is not based on any optimality criteria, so good properties are not guaranteed. As discussed in Chapter 5, Pfeffermann and Sverchkov (2007) proposed the sample likelihood approach that adjusts the likelihood for the selection process and obtained adjusted EBLUPs of area means. Cho et al. (2024) extended the approach of Pfeffermann and Sverchkov (2007) to the case of general area indicators, obtaining EB predictors under informative sampling. In simulation experiments, Cho et al. (2024) show how the adjusted EB predictors reduce design bias under informative selection of units within areas, but also showcased a surprisingly good behavior of the Pseudo EB predictor, which does not make model assumptions about the survey weights. Cho et al. (2024) also propose a parametric bootstrap procedure to estimate the MSE of adjusted Pseudo EB predictors, which performs well in simulations for sampled areas.

In the case of very complex non-linear indicators, where the EB/CEB predictor has no closed form, the complete MC simulation procedure for EB/CEB, along with the bootstrap procedure for MSE estimation described above, are computationally demanding. Hierarchical Bayesian procedures avoid the application of resampling procedures for MSE estimation, since they provide samples from the posterior distribution of the target indicators, from which posterior variance or credible intervals can be obtained directly. Hence, Bayesian SAE procedures can be computationally faster for large populations. However, when estimating living conditions of people, Bayesian procedures need to be applied with caution. To minimize the influence of priors on the final results, only non-informative priors are recommended for this kind of application. Molina, Nandram and Rao (2014) proposed a hierarchical Bayes (HB) version of the EB procedure based on non-informative priors. This HB procedure yields practically the same point estimates as EB and avoids the convergence issues of Markov Chain MC (MCMC) procedures. A Census HB counterpart is straightforward.

Marhuenda et al. (2017) extended the EB procedure based on the nested error model to the two-fold nested error model of Stukel and Rao (1999) given in (5.10), which includes subarea (e.g. cluster) effects nested within the area effects. This procedure has now been incorporated into the `sae` Stata package. Marhuenda et al. (2017) obtained clear losses in efficiency of EB estimators of poverty indicators when the random effects are specified only for the subareas (e.g. clusters), but estimation is desired for areas. Recently, Guadarrama, Morales and Molina (2020) have further extended the EB procedure by considering a two-fold model with time effects nested within the area effects, following an AR(1) process.

EB procedure under the nested error model (5.1) requires normality of area effects and unit-level errors, and this may be achieved by using a transformation of the welfare as the model response. In practice, the most common transformation for monetary variables, typically right-skewed and displaying heteroscedasticity, is the natural logarithm. When estimating area means of the original variables in a model with log transformation, Molina and Martín (2018) studied the analytical EB predictors and obtained second-order correct MSE estimators with a closed form. Often, a positive constant is added to the monetary variable before taking log transformation. Adding this constant is not only needed to have positive values; it is often necessary to shift the values far from zero, where the log function is less steep. Note that values z_{di} close to zero have $\log(z_{di})$ tending to minus infinity, which causes outliers on the left tail of the distribution, see Figure 4.7 of Corral et al. (2022). Hence, the constant is required to achieve approximate normality. Other transformations different from log, such as those from the Box-Cox or power families, can be taken, but making the shift previously is likely needed as well. The constant might be chosen by fitting the model to a grid of values in the range of the monetary variable, and selecting the value for which a skewness measure of model residuals is closest to zero. Rojas-Perilla et al. (2020) study data-driven transformations of the target variable in the EB method.

Note that, theoretically, a transformation to achieve normality always exists, but the problem is how to find it. Searching for an adequate transformation to achieve normality may be avoided by considering instead a skewed distribution for the original welfare variable. Diallo and Rao (2018) extended the EB method to the skew normal distribution and Graf, Marín and Molina (2019) to the generalized beta of the second kind (GB2), a flexible distribution that accommodates different types of unimodal skewness.

When the deviation from normality appears only in isolated observations and/or areas, which may occur even after transformation of the target variable, another possible approach is to consider models for outliers, such as models based on mixtures of normal distributions. A mixture may be specified for the model errors only, as proposed by Gershunskaya (2010) to estimate area means, or for both the random effects and errors, as considered by Elbers and Van der Weide (2014) to estimate poverty and inequality indicators. Bikauskaite, Molina and Morales (2021) extend the EB procedure to a multivariate mixture model that accounts for area-level outliers and incorporates heterogeneity in the regression coefficients apart from the variance components. Based on this model, they estimate poverty rates and gaps in Palestinian localities by gender.

Finally, robust procedures have been applied to reduce the effect of outliers in the final predictors of poverty indicators by Marchetti, Tzavidis and Pratesi (2012), using the unit-level M-quantile approach introduced by Chambers and Tzavidis (2006).

Chapter 7

Semi-parametric and machine-learning methods

This section describes model-based methods that are more flexible than the usual parametric linear or generalized linear regression models, by removing certain assumptions from the latter models. Some procedures remove the distributional assumptions but preserve linearity, while others preserve distributional assumptions, but consider more general semi-parametric specifications of the regression model. Therefore, in all cases, certain assumptions are retained.

We include here a wide range of machine learning procedures because of their potential flexibility, although all of the model-based procedures described in the previous chapters can be considered also machine learning procedures. Note that they all "learn" from the available data to predict the values (or improve poorer predictions) of the target variable/s. However, from the methods described hereafter, those based on unit-level data are designed to estimate only area means and cannot be extended directly to more complex area indicators, and those based on models/methods for aggregated data are specific for the target indicator used in the particular application.

Without assuming any probability distribution, Ghosh and Lahiri (1987) obtained estimators of small area means by specifying only the first two moments of the response variable. Their work was initially restricted to the case of no covariates, but was further extended to the case of area-level covariates by Raghunathan (1993).

Avoiding the specification of a linear regression model for the conditional expectation of the study variable, but assuming normality, Opsomer et al. (2008) proposed a nested error model with non-parametric mean function, using splines (Ruppert, Wand and Carrol, 2003). Specifically, Opsomer et al. (2008) derived EBLUPs of small area means in a model with penalized spline regression models. Robust estimators of small area means have been developed under the penalized spline model by Rao, Sinha and Dumitrescu (2013). Ugarte et al. (2009) used a unit-level model with penalized B-splines to estimate average prices per square meter of used dwellings in neighborhoods of the city of Vitoria, Spain. Wagner et al. (2017) used constrained penalized B-splines to estimate means of spruce timber reserves in forest districts of the German federal state Rhineland-Palatinate.

Spline models are designed for quantitative covariates. However, in many SAE applications with unit-level data, most of the covariates available in surveys and the corresponding census (if not all) are actually qualitative. Machine-learning techniques based on non-parametric regression methods, such as regression and classification trees (CART) described by Breiman et al. (1977), are currently very popular. Reasons for their popularity include a more automatized process of model construction that may be applied to large number of predictors, avoiding the application of model selection procedures, and the similar treatment of qualitative and quantitative predictors. The high variability of predictions obtained from a single regression tree has led to the development of ensemble methods that combine the predictions from multiple trees, such as the random forests introduced by Breiman (2001) or tree boosting techniques (Fried-

man, 2001).

In the context of SAE for poverty mapping, there are numerous recent applications of machine-learning techniques, and we will discuss just a sample of them. For example, Blumenstock et al. (2015) employed the gradient boosting techniques by Friedman (2001) to predict poverty and wealth in small areas of Rwanda, using covariates obtained from mobile phone metadata. Pokhriyal and Jacques (2017) applied Bayesian Gaussian Process (GP) regression, which assumes a linear regression with a non-parametric additive term, where a GP prior is assumed for the non-parametric part, and utilized elastic net regularization for model selection. They predicted the Global Multidimensional Poverty Index (MPI) at the commune level in Senegal, by combining the predictions obtained from two separate Bayesian GP regression models, each using a different data source. Steele et al. (2017) obtained highly granular poverty maps for Bangladesh using Bayesian Geostatistical Models with auxiliary information obtained from mobile phone and satellite data. Hersh et al. (2021) employed four different machine learning techniques (ridge and elastic net linear regression, random forests and extreme gradient boosted trees) to map poverty by enumeration districts of Belize, using open satellite derived features. Chi et al. (2022) utilized gradient boosting techniques to obtain estimates of average asset-based relative wealth indexes (RWIs) for 2.4-km populated microregions in the 135 low and middle income countries, using aggregated satellite imagery and big data sources, namely mobile phone data, topographic maps, and aggregated and deidentified connectivity data from Facebook. The latter procedure is specific for the asset-based RWIs, as these indexes are obtained from household assets and amenities, recorded by household surveys from the Demographic and Health Survey (DHS) Program, which include subregional geomarkers. Hence, the procedure may not be easily extendable to other types of poverty indicators.

The aforementioned machine-learning SAE procedures may be applied

in off-census years, as they are not based on census data. However, they seem to be all synthetic, similar to the traditional ELL method, in the sense of not accounting for area effects. If that is the case, their performance in an application is going to depend on the ability of the auxiliary information at hand to explain the between-area heterogeneity of the variable of interest. They might achieve great efficiency gains in an application where the covariates fully explain the between-area heterogeneity, but result in very poor estimates when the covariates are week in that sense. Note that to evaluate small area estimators, we need to look at their true MSE (accounting for both, bias and variance), since the produced MSE estimates might be biased low, giving a misleading picture of the true efficiency of the estimators. In real applications as those reviewed above, true MSEs are not available. The only way of having the true MSE is conducting simulation experiments.

Corral, Henderson and Segovia (2023) conducted design-based simulations based on a real census that included an income variable, by drawing multiple samples from it. They showed that, in the case of weak auxiliary information, such as the geo-referenced area-level data in their study, gradient boosting techniques applied at the area or cluster (PSU) level perform very poorly compared to other SAE procedures, both in terms of design bias and MSE. However, when using much more powerful census auxiliary area-level information, the same technique performs well in terms of design bias and MSE, close to the gold standard EB method based on the richer unit-level census data and including area effects. The EB method has a more consistent performance in terms of design bias and MSE than the machine learning procedures, even when the auxiliary information is week. Hence, their results point out the need to include area effects in modern machine learning procedures as well.

Moreover, the methods for aggregated data in certain locations that use survey estimates as a response variable in the models do not take into account the sampling errors of these survey estimates. These sampling

errors actually vary across the considered locations, so heteroscedasticity should be considered across locations. Note that sampling errors do not occur in epidemiological studies, where disease mapping procedures are typically applied. In disease mapping, most often all of the individuals with a particular disease are registered, and hence the data are not from sample surveys. In official statistics, where data on the study variable comes from surveys of limited location sample sizes, the sampling errors by location can be rather large, and these sampling errors should be incorporated in the estimation procedures. The above machine learning procedures seem not to include sampling errors, neither in the estimation procedures, nor in the accompanying noise measures. Noise measures that ignore sampling errors might severely understate the true error measures of the poverty estimates shown in the resulting maps.

In a machine-learning context different from SAE, Hajem et al. (2014) proposed mixed-effects random forests (MERF) obtained by adding normally distributed random effects to the random forests by Breiman (2001). They used an iterative algorithm that starts fitting a random forest, then adds predicted area effects under normality, and then iterates these two steps until convergence. Recently, Krennmair and Schmid (2022) have used the MERF introduced by Hajem et al. (2014) to estimate small area means of the variable of interest. The exact fitting algorithm, based on the EM algorithm is not detailed, and the theoretical properties of the resulting estimators are not known. Convergence of the iterative algorithm is neither ensured. They do not provide details of how the bootstrap samples are drawn in the construction of the random forest (since the sample comes from different areas) or how to set the tuning parameters, namely the number of variables for the splits in the tree branches and the number of bootstrap samples. Nevertheless, simulation results seem promising, as they indicate robustness to skewness on the distribution of the model errors and to misspecification of the regression function, at the expense of small efficiency losses with respect to the optimal estimators obtained under normality of model errors and with a correct specification

of the regression.

The MERF proposed by Hajem et al. (2014) and applied to SAE by Krennmair and Schmid (2022) are actually based on the normality of the random area effects. Similarly, although more flexible than fully parametric procedures, semi-parametric and other machine-learning methods still make certain model assumptions (although sometimes not mentioned), and their properties will depend on the extent to which these assumptions hold. Actually, perhaps the only purely non-parametric estimators are direct ones, which unfortunately are inefficient in small areas, and increasing the level of flexibility typically results in less efficient estimators.

Certainly, practically all small area estimation methods require a proper check of the underlying model assumptions. This should be carried out with the available data, e.g. through customary residual plots. Model checking is especially important for those models that do not include area effects, which lead to synthetic estimators, since they rely very strongly on the regression model for all of the areas, even for those with plenty of observations. Note that the methods that account for area effects give a positive weight to the direct estimators, which are approximately design-unbiased, regardless of whether model assumptions hold. In the case of clear evidences of model departure, the model should be modified to accommodate the existing data features. Otherwise, the resulting small area estimates should be taken with a lot of caution.

It is important to emphasize that, as in Statistics in general, in SAE, there is no panacea. That is, there is no universal procedure that works well for all possible datasets. Instead, depending on the data features, we need to choose a method with the level of flexibility/complexity that accommodates the real features reasonably well, without losing too much efficiency compared to less flexible/complex procedures. Furthermore, the challenging aspect is that there is no automatic procedure for correctly choosing between procedures with different levels of complexity for all possible datasets.

As mentioned in the previous comment, to ensure the reliability of model-based estimators, model diagnostics based on the available data should not provide evidence against the model assumptions. However, it is impossible to check the model for non-sampled areas. Therefore, we cannot be certain that these areas truly satisfy the model assumptions and are not outliers, unless additional information is available. Furthermore, as discussed earlier, synthetic estimators typically used for those areas are inefficient if area effects are significant. Thus, it is not advisable to produce estimates for non-sampled areas.

Chapter 8

Challenges and potential research topics

The previous chapters reviewed SAE methods, ranging from basic direct and indirect methods to modern model-based procedures designed to estimate general non-linear area indicators, defined in terms of a single continuous variable. For other important SAE topics such as model fitting methods and their properties, MSE estimation, prediction intervals or HB procedures, we refer the reader to Rao and Molina (2015).

This chapter outlines limitations in the current SAE literature encountered in certain practical situations. The issues or limitations listed below highlight research topics that might be of interest for practitioners and could potentially be addressed in the next 3-5 years.

1. *Estimation of general indicators with unit-level models in off-census years:* A common limitation of model-based SAE procedures using unit-level data for estimating non-linear indicators, is the requirement of a census containing the values of the auxiliary variables for all population units. Census files are available every 10 years, and might be severely outdated in off-census years. Outdated census auxiliary information can result in significant biases in the result-

ing small area estimators. Various proposals exist in the literature for SAE in off-census years. Applying a FH model or a machine-learning procedure that uses only aggregated auxiliary information are possible solutions, but aggregated models lose efficiency compared to unit-level models. Other authors suggest applying ELL or EB methods with the so-called unit-context models, which use unit-level survey data on welfare as the model response but with aggregated auxiliary information. For ELL applications using a unit-context model, see Nguyen (2012) or Lange, Pape and Pütz (2018). For application of EB using a unit-context model, see Masaki et al. (2020). When estimating small area means, the unit-context model is nearly equivalent to the FH model that uses the same aggregated covariates, but this is not the case for non-linear area indicators. Unfortunately, simulations experiments conducted by Corral et al. (2021), both at the model- and design-based frameworks, showed that estimators of poverty indicators obtained using unit-context models can be significantly biased. Hence, further research is needed to obtain estimators of poverty indicators in off-census years based on unit-level data.

2. *Variable selection in SAE:* Typically, covariates for an SAE model are selected by applying model selection procedures to the analogous regression model without the area effects. The usual model selection methods include exhaustive search for the case of small number of potential covariates, forward, backward and stepwise selection procedures, or the least absolute shrinkage and selection operator (LASSO) by Tibshirani (1996). However, the resulting set of covariates is not necessarily the best for the analogous regression model that includes area effects. Specific model selection procedures for mixed models, such as the fence method by Jiang et al. (2008), have been proposed. Nevertheless, computationally efficient model selection procedures specific to SAE deserve further research.

3. *Estimation of error variances in the FH model and associated MSE estimates:* Error variances in the FH model are assumed to be known, because they cannot be estimated with area-level data. Typically, these error variances are replaced with the estimated sampling variances of the direct estimators acting as model response variables. However, these variance estimators are also based in the area-specific survey data, making them inefficient in small areas. Moreover, the usual estimators of the MSE for FH estimators do not account for the uncertainty due to the estimation of these error variances, which might be substantial. More research is needed to obtain efficient error variances and derive MSE estimators that account for the uncertainty due to the estimation of these variances.

4. *The Alpha model in unit-level models:* The traditional ELL method includes heteroscedastic idiosyncratic errors, and EB or Census EB under the nested error model allows for the inclusion of known heteroscedasticity factors k_{di}^2, although this possibility is not implemented in the `sae` R package. In the ELL method, the error variances σ_{dci}^2 are firstly predicted through the so-called "Alpha model", which uses a regression model in logistic form for the squared residuals e_{dci}^2, in terms of possibly different covariates than those included later in the regression of the welfare variable. Heteroscedasticity is likely to be needed in applications where the statistical units are households instead of individuals, due to the different households sizes. In general, before applying the Alpha model, it would be recommendable to check for evidences of heteroscedasticity in the particular data set using a formal test such as Wald test (Greene 2000, Section 12.5.3), and consider the Alpha model only when necessary. In any case, further research is needed on proper variable selection methods in models for heteroscedasticity within the SAE context.

5. *Machine learning techniques:* Semi- or non-parametric machine learn-

ing procedures are more flexible than their linear or generalized linear model counterparts, and some of them avoid the application of variable selection methods. Procedures used for aggregated data might be applied in off-census years, but sampling errors (varying across locations) need to be incorporated into the models, as is done in the FH model. Except for the mixed random forest for unit-level data employed by Krennmair and Schmid (2022), the machine-learning procedures reviewed here are synthetic, not accounting for area effects. Hence, when applying these methods, one needs to cross fingers for the available covariates to explain all the between-area heterogeneity, as illustrated by Corral, Henderson and Segovia (2023). The mixed-effects random forests used by Krennmair and Schmid (2022) include area effects and show promising results in simulations. However, normality is still required, convergence of the fitting algorithm is not ensured, and theoretical properties of the estimators are not known. In general, non-parametric regression techniques are flexible enough to accommodate complex trends, but they are expected to lose efficiency compared to a model with an approximately correct parametric specification. Moreover, additional research is needed for estimating general non-linear area indicators and obtaining reliable estimates of MSE under machine learning procedures.

6. *Estimation of design MSE:* Conventional model MSE estimates of model-based estimators are obtained assuming that the corresponding model assumptions hold. However, we know that "All models are wrong, but some are useful". Hence, model MSE estimators might be understating the real uncertainty when the model does not hold exactly. The MSE across the possible samples of units that may be drawn from the population with the selected sampling design, known as design MSE, does not assume that the model holds, and hence accounts for model uncertainty. Accordingly, design MSE might be a more objective error measure of small area estimators. When

estimating small area means, Molina and Strzalkowska-Kominiak (2020) proposed using the same idea of "borrowing strength" behind SAE, to estimate the design MSE of small area estimators. Reliable design MSE estimators for general non-linear indicators are still a challenge.

7. *Correction for non ignorable non-response:* As already discussed, under sample selection bias (or non-ignorable informative sampling), the selection of units depends on the values of the target variable. In that case, the model for the sample units differs from the model assumed for the population units and hence the sample model should be fitted to the sample data. Pfeffermann and Sverchkov (2007) obtained adjusted EBLUPs of small area means under non-ignorable informative sampling, and Cho et al. (2024) have extended the procedure to general non-linear parameters and more general models for the sampling weights. Similarly, in the case of non-ignorable non-response, the indicators of responding depend on the values of the target variable, causing the model followed by the respondents (which are a subset of the sample units) to differ from the model for the sample units. Sverchkov and Pfeffermann (2018) studied SAE of small area means, accounting for non-ignorable informative sampling and non-ignorable non-response. Extension to non-linear indicators is currently under study.

8. *Estimation of multidimensional poverty indicators:* Multidimensional poverty indicators assess poverty from a wider perspective, but there are multidimensional indicators of different natures. Some of them are based on constructing a single index by applying principal components or other dimension reduction procedure to a set of variables measuring different deprivations. Since the resulting index is quantitative, unit-level SAE procedures such as Census EB might be applied similarly to estimate linear or non-linear functions of that index in the area units. Other multidimensional poverty indicators

define a person as multidimensionally poor when its welfare measure falls below a certain threshold and at the same time has a number of deprivations from a given set. Estimating this multidimensional poverty indicator with area-level models is straightforward, since direct estimators may be obtained from the survey and used as response variable in a FH model. However, using unit-level models entails simultaneously modeling the welfare measure, which is a continuous variable, and the required deprivations, which are dummy indicators. All these variables are dependent, and employing an SAE model for several dependent response variables of different natures is a challenge, at least in the frequentist setup.

Acknowledgments

This work was supported by the Contract num. 7209970 between The World Bank Group and Universidad Complutense de Madrid. The author wishes to thank Utz J. Pape and Jed Friedman for their constructive comments, which have contributed to the improvement of the paper.

References

Arora, V., and Lahiri, P. (1997), On the Superiority of the Bayes Method over the BLUP in Small Area Estimation Problems, *Statistica Sinica*, **7**, 1053–1063.

Baíllo, A. and Molina, I. (2009), Mean-squared errors of small-area estimators under a unit-level multivariate model, *Statistics*, **43**, 553–569.

Banerjee, A. V., Deaton, A., Lustig, N., Rogoff, K. and Hsu, E. (2006). An evaluation of World Bank research, 1998-2005. Available at SSRN 2950327.

Battese, G.E., Harter, R.M. and 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.

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013), A Unified Approach to Robust Estimation in Finite Population Sampling, *Biometrika*, **100**, 555–569.

Bedi, T., Coudouel, A., and Simler, K. (Eds.). (2007). More than a pretty picture: using poverty maps to design better policies and interventions. World Bank Publications.

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

Bell, W. (1997). Models for county and state poverty estimates. Preprint, Statistical Research Division, U.S. Census Bureau.

Bell, W.R. (2008), Examining Sensitivity of Small Area Inferences to Uncertainty About Sampling Error Variances, *Proceedings of the Survey Research Section*, American Statistical Association, pp. 327–334.

Bell, W.R., Datta, G.S. and Ghosh, M. (2013), Benchmarking Small Area Estimators, *Biometrika*, **100**, 189–202.

Bell, W.R. and Huang, E.T. (2006), Using the t-distribution to Deal with Outliers in Small Area Estimation, *Proceedings of Statistics*, Canada Symposium on Methodological Issues in Measuring Population Health, Statistics Canada, Ottawa, Canada.

Berg, E. (2022), Empirical best prediction of small area means based on a unit-level gamma-poisson model, *Journal of Survey Statistics and Methodology*, https://doi.org/10.1093/jssam/smac026.

Berg, E. (2023), Small area prediction of seat-belt use rates using a bayesian hierarchical unit-level poisson model with multivariate random effects, *Stat*, **12**(1), https://doi.org/10.1002/sta4.544.

Berg, E., and Chandra, H. (2014), Small area prediction for a unit-level lognormal model, *Computational Statistics & Data Analysis*, **78**, 159–175.

Cho, Y., Guadarrama-Sanz, M., Molina, I., Eideh, A. and Berg, E. (2023), Optimal predictors of general small area parameters under an informative sample design using parametric sample distribution models, *Journal of Survey Statistics and Methodology*, https://doi.org/10.1093/jssam/smae0

Besag, J.E. (1974), Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion), *Journal of the Royal Statistical Society, Series B*, **35**, 192-236.

Bikauskaite, A., Molina, I. and Morales, D. (2022). Multivariate mixture model for small area estimation of poverty indicators, *Journal of the*

Royal Statistical Society, Series A, https://doi.org/10.1111/rssa.12965

Blumenstock, J., Cadamuro, G. and On, R. (2015). Predicting poverty and wealth from mobile phone metadata, *Science*, **350**(6264), 1073–1076.

Boubeta, M., Lombardía, M. J. and Morales, D. (2016), Empirical Best Prediction under Area-Level Poisson Mixed Models, **Test**, **25**(3), 548–569.

Boubeta, M., Lombardía, M. J. and Morales, D. (2017), Poisson Mixed Models for Studying the Poverty in Small Areas, *Computational Statistics & Data Analysis*, **107**, 32–47.

Boubeta, M., Lombardía, M. J., Marey-Pérez, F. and Morales, D. (2020), Area-Level Spatio-Temporal Poisson Mixed Models for Predicting Domain Counts and Proportions, arXiv:2012.00069. DOI: 10.48550/arXiv.2012.00069.

Breckling, J., and Chambers, R. (1988), M-quantiles, *Biometrika*, **75**, 761–771.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984), *Classification and Regression Trees*, New York: Chapman & Hall/CRC.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**(1), 5–32.

Casas-Cordero Valencia, C., Encina, J. and Lahiri, P. (2016). Poverty Mapping for the Chilean Comunas, In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, New York: Wiley.

Chambers, R.L. (1986), Outier Robust Finite Population Estimation. *Journal of the American Statistical Association*, **81**, 1063–1069.

Chambers, R., and Clark, R. (2012), *An Introduction to Model-Based Survey Sampling with Applications*, Oxford: Oxford University Press.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014), Outlier Robust Small Area Estimation, *Journal of the Royal Statistical Society, Ser. B*, **76**, 47–69.

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**, 255 –268.

Chandra, H., Salvati, N., and Chambers, R. (2017), Small Area Prediction of Counts under a Non-Stationary Spatial Model, *Spatial Statistics*, **20**, 30–56.

Chaudhuri, A. (2012), *Developing Small Domain Statistics: Modelling in Survey Sampling*, Saarbrücken: LAP LAMBERT Academic Publishing GMbH & Co. KG.

Chen, S., Jiang, J., and Nguyen, T. (2015), Observed Best Prediction for Small Area Counts, *Journal of Survey Statistics and Methodology*, **3**(2), 136–161.

Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J.E. (2022). Microestimates of wealth for all low-and middle-income countries, *Proceedings of the National Academy of Sciences*, **119**(3), 1–11.

Clayton, D., and Kaldor, J. (1987), Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping, *Biometrics*, **43**, 671–681.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed., New York: Wiley.

Corral, P. and Cojocaru, A. (2019). Moldova Poverty Map: An Application of Small Area Estimation, Unpublished report.

Corral, P., Henderson, H. and Segovia, S. (2023). Poverty mapping in the age of machine learning, The World Bank, https://doi.org/10.1596/1813-9450-10429

Corral, P., Himelein, K., McGee, K., and Molina, I. (2021). A Map of the Poor or a Poor Map?, **Mathematics**, **9**(21), 2780; https://doi.org/10.3390/math9212780

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

Corral, P., Molina, I., Cojonaru, A. and Segovia, S. (2022), Guidelines to small area estimation for poverty mapping. The World Bank.

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

Cressie, N. (1989), Empirical Bayes Estimation of Undercount in the Decennial Census, *Journal of the American Statistical Association*, **84**, 1033–1044.

Cressie, N. (1991), Small-Area Prediction of Undercount Using the General Linear Model, *Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality*, Ottawa: Statistics Canada, pp. 93–105.

Datta, G.S. (2009), Model-Based Approach to Small Area Estimation, in *Sample Surveys: Inference and Analysis*, D. Pfeffermann and C.R. Rao, (Eds.), *Handbook of Statistics*, Volume 29B, Amsterdam: North-Holland, pp. 251–288.

Datta, G.S., Day, B. and Basawa, I. (1999), Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation, *Journal of Statistical Planning and Inference*, **75**, 169–179.

Datta, G.S., Fay, R.E. and Ghosh, M. (1991), Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation, in *Proceedings of Bureau of the Census 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, pp. 63–79.

Datta, G.S., and Ghosh, M. (1991), Bayesian Prediction in Linear Models: Applications to Small Area Estimation, *Annals of Statistics*, **19**, 1748–1770.

Datta, G.S., Ghosh, M., Nangia, N. and Natarajan, K. (1996), Estimation of Median Income of Four-Person Families: A Bayesian Approach, in W.A. Berry, K.M. Chaloner, and J.K. Geweke (Eds.), *Bayesian Analysis in Statistics and Econometrics*, New York: Wiley, pp. 129-140.

Datta, G.S., Ghosh, M., Steorts, S., and Maples, J.J. (2011), Bayesian Benchmarking with Applications to Small Area Estimation, *Test*, **20**, 574–588.

Datta, G.S., Ghosh, M. and Waller, L.A. (2000), Hierarchical and Empirical Bayes Methods for Environmental Risk Assessment, in P.K. Sen and C.R. Rao (Eds.), *Handbook of Statistics*, Volume 18, Amsterdam: Elsevier Science B.V., pp. 223–245.

Datta, G.S., and Lahiri, P. (1995), Robust Hierarchical Bayes Estimation of Small Area Characteristics in the Presence of Covariates, *Journal of Multivariate Analysis*, **54**, 310–328.

Datta, G.S., Lahiri, P., and Maiti, T. (2002), Empirical Bayes Estimation of Median Income of Four-Person Families by State Using Time Series and Cross-Sectional Data, *Journal of Statistical Planning and Inference*, **102**, 83–97.

Datta, G.S., Rao, J.N.K., and Torabi, M. (2010), Pseudo-empirical Bayes Estimation of Small Area Means Under a Nested Error Linear Regression Model with Functional Measurement Errors, *Journal of Statistical Planning and Inference*, **140**, 2952–2962.

DeSouza, C.M. (1992), An Appropriate Bivariate Bayesian Method for Analysing Small Frequencies, *Biometrics*, **48**, 1113–1130.

Deville, J.C., and Särndal, C.E. (1992), Calibration Estimation in Survey Sampling, *Journal of the American Statistical Association*, **87**, 376–382.

Diallo, M. and Rao, J. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, **45**, 1092–1116.

Dick, P. (1995), Modelling Net Undercoverage in the 1991 Canadian Census, *Survey Methodology*, **21**, 45–54.

Drew, D., Singh, M.P., and Choudhry, G.H. (1982), Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey,

Survey Methodology, **8**, 17–47.

Edochie, I., Newhouse, D., Würz, N. and Schmid, T. (2023). povmap: extension to the 'emdi' Package, Version 1.0.0.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2002). Micro-level estimation of welfare. World Bank Policy Research Working Paper 2911.

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

Elbers, C. and 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.

Elbers, C. Fujii, T., Lanjouw, P., Özler, B. and Yin, W. (2004). Poverty Alleviation through Geographic Targeting: How Much Does Disaggregation Help? World Bank Policy Research Working Paper 3419.

Ericksen, E.P., and Kadane, J.B. (1985), Estimating the Population in Census Year: 1980 and Beyond (with discussion), *Journal of the American Statistical Association*, **80**, 98–131.

Estevan, M.D., Morales, D., Pérez, A. and Santamaría, L. (2010), Area-Level Time Models for Small Area Estimation of Poverty Indicators, In: Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Intelligent and Soft Computing, vol 77. Springer, Berlin, Heidelberg.
https://doi.org/10.1007/9783642147463_29

Estevan, M.D., Morales, D., Pérez, A. and Santamaría, L. (2012), Small Area Estimation of Poverty Proportions under Area-level Time Models, *Computational Statistics and Data Analysis*, **56**, 2840–2855.

Fabrizi, E., Salvati, N. and Pratesi, M. (2012), Constrained small area estimators based on M-quantile methods, *Journal of Official Statistics*, **28**, 89–106.

Fabrizi, E. and Trivisano, E. (2016). Small area estimation of the Gini concentration coefficient. *Computational Statistics and Data Analysis*, **99**, 223–234.

Farris, J., Larochelle, C., Alwang, J., Norton, G.W. and King, C. (2017). Poverty analysis using small area estimation: an application to conservation agriculture in Uganda. *Agricultural Economics*, **48**(6), 671–681.

Fauziah, R.A. and Wulansari, I.Y. (2020). saeeb: Small Area Estimation for Count Data, R package 0.1.0.

Fay, R.E. (1987), Application of Multivariate Regression to Small Domain Estimation, in R. Platek, J.N.K. Rao, C.E. Särndal, and M.P. Singh (Eds.), *Small Area Statistics*, New York: Wiley, pp. 91–102.

Fay, R.E. and Herriot, R.A. (1979), Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association, 85, 398–409.

Foster, J., Greer, J., and Thorbecke, E. (1984). Accounting for dependent informative sampling in model-based finite population inference. *Econometrica*, **52**, 761-–766.

Franco, C. and Bell, W.R. (2013). Applying Bivariate Binomial/Logit Normal Models to Small Area Estimation. Proceedings of the American Statistical Association, Section on Survey Research Methods, 690–702,
URL http://ww2.amstat.org/sections/srms/ Proceedings/.

Franco, C. and Bell, W.R. (2015). Borrowing Information Over Time in Binomial/Logit Normal Models for Small Area Estimation. Statistics in Transition (new series) and Survey Methodology, joint issue on Small Area Estimation, 16, 563–584, available at http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/volume-16-number-4-december-2015/.

Friedman, J.H. (2001). Greedy function approximation: A gradient

boosting machine, *Annals of Statistics*, **29**(5), 1189–1232.

Fuller, W.A. (1987), *Measurement Error Models*, New York: Wiley.

Fuller, W.A. (1999), Environmental Surveys Over Time, *Journal of Agricultural, Biological and Environmental Statistics*, **4**, 331–345.

Fuller, W.A. (2009), *Sampling Statistics*, New York: Wiley.

Fuller, W.A., and Goyeneche, J.J. (1998), Estimation of the State Variance Component, Unpublished manuscript.

Fuller, W.A. and Harter, R.M. (1987), The Multivariate Components of Variance Model for Small Area Estimation, in R. Platek, J.N.K. Rao, C.E. Särndal, and M.P. Singh (Eds.), *Small Area Statistics*, New York: Wiley, pp. 103–123.

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

Ghosh, M. (2020), Small area estimation: its evolution in five decades, *Statistics in Transition*, **21** (4), 40–44 (with discussion).

Ghosh, M. and Lahiri, P. (1987), Robust Empirical Bayes Estimation of Means from Stratified Samples, *Journal of the American Statistical Association*, **82**, 1153–1162.

Ghosh, M., Maiti, T. and Roy, A. (2008), Influence Functions and Robust Bayes and Empirical Bayes Small Area Estimation, *Biometrika*, **95**, 573–585.

Ghosh, M., Nangia, N., and Kim, D. (1996), Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach, *Journal of the American Statistical Association*, **91**, 1423–1431.

Ghosh, M., Natarajan, K., Stroud, T.W.F., and Carlin, B.P. (1998), Generalized Linear Models for Small Area Estimation, *Journal of American Statistical Association*, **93**, 273–282.

Ghosh, M., Natarajan, K., Waller, L.A., and Kim, D.H. (1999), Hierarchical Bayes GLMs for the Analysis of Spatial Data: An Application

to Disease Mapping, *Journal of Statistical Planing and Inference*, **75**, 305–318.

Ghosh, M., and Sinha, K. (2007), Empirical Bayes Estimation in Finite Population Sampling Under Functional Measurement Error Models, *Scandinavian Journal of Statistics*, **33**, 591–608.

Ghosh, M., Sinha, K. and Kim, D. (2006), Empirical and Hierarchical Bayesian Estimation in Finite Population Sampling Under Structural Measurement Error Models, *Journal of Statistical Planning and Inference*, **137**, 2759–2773.

Ghosh, M., and Rao, J.N.K. (1994), Small Area Estimation: an Appraisal (with Discussion), *Statistical Science*, **9**, 55–93.

Giusti, C., Masserini, L. and Pratesi, M. (2017). Local comparisons of small area estimates of poverty: an application within the Tuscany region in Italy. *Social Indicators Research*, **131** (1), 235–254.

Gonzalez, M.E. (1973), Use and Evaluation of Synthetic Estimates, *Proceedings of the Social Statistics Section*, American Statistical Association, 33–36.

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

Graf, M., Marín, J. M., and Molina, I. (2019). A generalized mixed model for skewed distributions applied to small area estimation. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, **28**(2), 565–597.

Griffiths, R. (1996), Current Population Survey Small Area Estimations for Congressional Districts, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 314–319.

Guadarrama, M., Molina, I. and Rao, J.N.K. (2014). A comparison of small area estimation methods for poverty mapping. *Statistics in*

Transition, **1** (17), 41–66.

Guadarrama, M., Molina, I., and 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., Morales, D. and Molina, I. (2020). Time stable small area estimates of general parameters under a unit-level model. *Computational Statistics and Data Analysis*, **160**, 107226.

Hajjem, A., Bellavance, F. and Larocque, D. (2014) Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, **84**(6), 1313–1328.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), *Sample Survey Methods and Theory I*, New York: Wiley.

Henderson, C.R. (1950), Estimation of Genetic Parameters (Abstract), *Annals of Mathematical Statistics*, **21**, 309–310.

Hersh, J., Engstrom, R. and Mann, M. (2021). Open data for algorithms: Mapping poverty in Belize using open satellite derived features and machine learning, *Information Technology for Development*, **27**(2), 263–292.

Hobza, T., and Morales, D. (2016), Empirical Best Prediction under Unit-Level Logit Mixed Models, *Journal of Official Statistics*, **32**(3), 661–692.

Hobza, T., Marhuenda, Y., and Morales, D. (2020), Small Area Estimation of Additive Parameters under Unit-Level Generalized Linear Mixed Models, *SORT-Statistics and Operations Research Transactions*, **44**(1), 3–38.

Hobza, T., Morales, D., and Santamaría, L. (2018), Small Area Estimation of Poverty Proportions under Unit-Level Temporal Binomial-Logit Mixed Models, *Test*, **27**(2), 270–294.

Huber, P.J.(1964). Robust estimation of a location parameter, *The Annals of Mathematical Statistics*, **35**,73–101.

Isaki, C.T., Huang to the case of informative sampling, E.T., and Tsay, J.H. (1991), Smoothing Adjustment Factors from the 1990 Post Enumeration Survey, *Proceedings of the Social Statistics Section*, American Statistical Association, Washington, DC, pp. 338–343.

Isaki, C.T., Tsay, J.H., and Fuller, W.A. (2000), Estimation of Census Adjustment Factors, *Survey Methodology*, **26**, 31–42.

Isidro, M., Haslett, S. and Jones, G. (2016). Extended Structure Preserving Estimation (ESPREE) for updating small area estimates of poverty. *The Annals of Applied Statistics*, **10** (1), 451–476.

Jiang, J. (2007). Linear and generalized linear mixed models and their applications. New York: Springer-Verlag

Jiang, J. and Larihi, P. (2001). Empirical best prediction for small area inference with binary data, *Annals of the Institute of Statistical Mathematics*, **53**, 217–243.

Jiang, J., and Lahiri, P. (2006), Mixed Model Prediction and Small Area Estimation, *Test*, **15**, 1–96.

Jiang, J., Nguyen, T. and Rao, J.S. (2009), A Simplified Adaptive Fence Procedure, *Statistics and Probability Letters*, **79**, 625–629.

Jiang, J., Nguyen, T. and Rao, J.S. (2015), Observed Best Prediction Via Nested-error Regression with Potentially Misspecified Mean and Variance, *Survey Methodology*, **41** (1), 37–55.

Jiang, J., Rao, J.S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection, *Annals of Statistics*, **36**, 1669–1692.

Jiongo, V.D., Haziza, D. and Duchesne, P. (2013), Controlling the Bias of Robust Small-area Estimation, *Biometrika*, **100**, 843–858.

Jedrzejczak, A. and Kubacki, J. (2013). Estimation of Income Inequality and the Poverty Rate in Poland, by Region and Family Type,". Statistics in Transition-New Series, 14(3).

Kott, P.S. (1990), Robust Small Domain Estimation Using Random Effects Modelling, *Survey Methodology*, **15**, 3–12.

Kreutzmann, A., Pannier, S., Rojas-Perilla, N., Schmid, T. Templ, M. and Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators, *Journal of Statistical Software*, **91**(7), 1–33, DOI:10.18637/jss.v091.i07.

Lahiri, P. and Salvati, N. (2023), A nested error regression model with high-dimensional parameter for small area estimation, *Journal of the Royal Statistical Society Series, https://doi.org/10.1093/jrsssb/qkac010*.

Lange, S., Pape, U. J. and Pütz, P. (2018). Small Area Estimation of Poverty under Structural Change, World Bank Policy Research Working Paper 9383.

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2003), The Effect of Model Choice in Estimation for Domains Including Small Domains, *Survey Methodology*, **29**, 33–44.

Lehtonen,R. and Veijanen, A. (2009), Design-Based Methods of Estimation for Domains and Small Areas, in *Sample Surveys: Inference and Analysis*, D. Pfeffermann and C.R. Rao, (Eds.), *Handbook of Statistics*, Volume 29B, Amsterdam: North-Holland, pp. 219–249.

Longford, N.T. (2005), *Missing Data and Small-Area Estimation*, New York: Springer.

Lohr, S.L. (2010), *Sampling: Design and Analysis*, Pacific Grove; CA: Duxbury.

Lohr, S.L., and Prasad, N.G.N. (2003), Small Area Estimation with Auxiliary Survey Data, *Canadian Journal of Statistics*, **31**, 383–396.

López-Vizcaíno, E., Lombardía, M.J. and Morales, D. (2013), Multinomial-based Small Area Estimation of Labour Force Indicators, *Statistical Modelling*, **13**, 153–178.

López-Vizcaíno, E., Lombardía, M.J. and Morales, D. (2015), Small Area Estimation of Labour Force Indicators Under a Multinomial

Model with Correlated Time and Area Effects, *Journal of the Royal Statistical Society, Series A*, **178** (3), 535–565.

MacGibbon, B., and Tomberlin, T.J. (1989), Small Area Estimation of Proportions Via Empirical Bayes Techniques, *Survey Methodology*, **15**, 237–252.

Maiti, T. (1998), Hierarchical Bayes Estimation of Mortality Rates for Disease Mapping, *Journal of Statistical Planning and Inference*, **69**, 339–348.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. and Gabrielli, L. (2015). Small area model-based Estimators using big data sources, *Journal of Official Statistics*, **31** (2), 263–281,

Marchetti, S., Tzavidis, N. and Pratesi, M. (2012). Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. *Computational Statistics and Data Analysis*, **56**, 2889–2902.

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, **58**, 308–325.

Marhuenda, Y., Molina, I., Morales, D., and Rao, J. N. K. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A*, **180**(4), 1111–1136.

Masaki, T., Newhouse, D., Silwal, A.R., Bedada, A. and Engstrom, R. (2020). Small area estimation of non-monetary poverty with geospatial Data, World Bank Policy Research Working Paper 9383.

Mauro, F., Molina, I., García-Abril, A., Valbuena, R. and Ayuga-Télleza, E. (2015). Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels, *Environmetrics*, **27** (4), 225–238.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Chapman & Hall.

Militino, A. F., Ugarte, M. D., Goicoa, T. and González-Aud'icana, M. (2006), Using Small Area Models to Estimate the Total Area Occupied by Olive Trees, *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 450–461.

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. (2020). Discussion on "Small area estimation: its evolution in five decades", by M. Ghosh, *Statistics in Transition*, **21** (4), 40–44.

Molina, I., Corral, P. and Nguyen, M. (2022). Estimation of poverty and inequality in small areas: review and discussion, *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, **31**(4), 1143–1166, DOI: 10.1007/s11749-022-00822-1

Molina, I. and Marhuenda, Y. (2015). sae: An R package for small area estimation. *The R Journal*, **7(1)**, 81–98.

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

Molina, I. and Morales, D. (2009). Small area estimation of poverty indicators. Boletín de Estadística e Investigación Operativa, **25**, 218-225.

Molina, I., Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. *The Annals of Applied Statistics*, **8**(2), 852–885.

Molina, I. and Rao, J.N.K. (2010). Small Area Estimation of Poverty Indicators. *The Canadian Journal of Statistics*, **38**, 369–385.

Molina, I. and Rao, J.N.K. (2023). Historical Overview of Small Area Estimation in the 50th Birthday of the IASS. *The Survey Statistician*, **88**, 23–35.

Molina, I., Rao, J.N.K. and Guadarrama, M. (2019), Small Area Estimation Methods for Poverty Mapping: A Selective Review, *Statistics and Applications*, **17** (1), 11–22.

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

Molina, I. and Strzalkowska-Kominiak, E. (2020). Estimation of proportions in small areas: application to the labour force using the Swiss Census Structural Survey. *Journal of the Royal Statistical Society, Series A*, **183** (1), 281—310.

Molina Peralta I. and García Portugués E. (2020), Short guide for small-area estimation using household survey data: illustration to poverty mapping in Palestine with expenditure survey and census data. UN Economic and Social Commission for Western Asia (ESCWA), E/ESCWA/SD/2019/TP.4

Morales, D., Esteban, M. D., Perez, A. and Hobza, T. (2021), A Course on Small Area Estimation and Mixed Models: Methods, Theory and Applications in R (1st ed.), Cham, Switzerland: Springer.

Moura, F.A.S., and Holt, D. (1999), Small Area Estimation Using Multilevel Models, *Survey Methodology*, **25**, 73–80.

Mukhopadhyay, P. (1998), *Small Area Estimation in Survey Sampling*, New Delhi: Narosa Publishing House.

Nandram, B., Sedransk, J. and Pickle, L. (1999), Bayesian Analysis of Mortality Rates for U.S. Health Service Areas, *Sankhyā, Series B*, **61**, 145–165.

Nelder, J. A. and Wedderburn, R. W. M. (1972), *Journal of the Royal Statistical Society, Series A*, **135** (3), 370–384.

Nguyen, V. C. (2012). A Method to Update Poverty Maps, *The Journal of Development Studies*, **48** (12), 1844-–1863

Nguyen, M. C., Corral, P., Azevedo, J. P., and Zhao, Q. (2018). Sae: A stata package for unit level small area estimation. World Bank Policy Research Working Paper 8630.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008), Nonparametric small area estimation using penalized spline regression, *Journal of the Royal Statistical Society, Series B*, **70**, 265–286.

Pfeffermann, D. (2002). Small area estimation-new developments and directions. *International Statistical Review*, **70** (1), 125–143.

Pfeffermann, D. (2013), New Important Developments in Small Area Estimation, *Statistical Science*, **28**(1), 40–68.

Pfeffermann, D. and Barnard, C. (1991), Some New Estimators for Small Area Means with Applications to the Assessment of Farmland Values, *Journal of Business and Economic Statistics*, **9**, 73–84.

Pfeffermann, D., and Burck, L. (1990), Robust Small Area Estimation Combining Time Series and Cross-Sectional Data, *Survey Methodology*, **16**, 217–237.

Pfeffermann, D., and Sverchkov, M. (1999), Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data, *Sankhya*, Ser. B, **61**, 166–186.

Pfeffermann, D., and Sverchkov, M. (2003), Fitting Generalized Linear Models Under Informative Probability Sampling, in *Analysis of Survey Data*, eds. C.J. Skinner and R.L. Chambers, New York: Wiley, pp. 175–195.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998), Parametric Distributions of Complex Survey Data Under Informative Probability

Sampling, *Statistica Sinica*, **8**, 1087–1114.

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

Pokhriyal, N. and Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping, *Proceedings of the National Academy of Sciences*, **114**(46), E9783–E9792.

Prasad, N.G.N., and Rao, J.N.K. (1999), On Robust Small Area Estimation Using a Simple Random Effects Model, *Survey Methodology*, **25**, 67–72.

Pratesi, M. and Salvati, N. (2016). Introduction on measuring poverty at local level using small area estimation methods. In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, New York: Wiley, 1–18.

Raghunathan, T.E. (1993), A Quasi-Empirical Bayes Method for Small Area Estimation, *Journal of the American Statistical Association*, **88**, 1444–1448.

Rao, J.N.K. and Molina, I. (2015). Small Area Estimation, 2nd. Ed., Hoboken, NJ:Wiley.

Rao, J.N.K. and Molina, I. (2016). Empirical Bayes and hierarchical Bayes estimation of poverty measures for small areas. In M. Pratesi (Ed.) Analysis of Poverty Data by Small Area Estimation, New York: Wiley, 315–324.

Rao, J.N.K., Sinha, S.K. and Dumitrescu, L. (2014), Robust Small Area Estimation under Semi-parametric Mixed Models, *The Canadian Journal of Statistics*, **42**, 126–141.

Rao, J.N.K., and Yu, M. (1992), Small Area Estimation by Combining Time Series and Cross-Sectional Data, *Proceedings of the Section on Survey Research Method*, American Statistical Association, pp. 1–9.

Rao, J.N.K., and Yu, M. (1994), Small Area Estimation by Combining Time Series and Cross-Sectional Data, *Canadian Journal of Statistics*, **22**, 511–528.

Rivest, L-P., and Vandal, N. (2003), Mean Squared Error Estimation for Small Areas when the Small Area Variances are Estimated, *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Technical Report No. 386, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, Canada.

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

Royall, R.M. (1970), On Finite Population Sampling Theory Under Certain Linear Regression, *Biometrika*, **57**, 377–387.

Royall, R.M. (1976), The Linear Least-Squares Prediction Approach to Two-stage Sampling, *Journal of the American Statistical Association*, **71**, 657–664.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003), *Semiparametric Regression*, New York: Cambridge University Press.

Särdnal, C.E., Swensson, B., and Wretman, J.H. (1989), The Weighted Regression Technique for Estimating the Variance of Generalized Regression Estimator, *Biometrika*, **76**, 527–537.

Scott, A. and Smith, T.M.F. (1969), Estimation in Multi-Stage Surveys, *Journal of the American Statistical Association*, **64**, 830–840.

Searle, S.R. (1971). Linear Models, Wiley.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). Variance Components, Wiley.

Schaible, W.A. (1978), Choosing Weights for Composite Estimators for Small Area Statistics, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 741–746.

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

Seitz, W. H. (2019). Where they live: district-level measures of poverty, average consumption, and the middle class in Central Asia, World Bank Policy Research Working Paper 8940.

Steorts, R., and Ghosh, M. (2013), On Estimation of Mean Squared Error of Benchmarked Empirical Bayes Estimators, *Statistica Sinica*, **23**, 749–767.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J. and Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data, *Journal of The Royal Society Interface*, **14**(127), 20160690.

Stukel, D. and Rao, J.N.K. (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference*, **78**, 131–147.

Sverchkov, M. and Pfeffermann, D. (2018). Small area estimation under informative sampling and not missing at random non-response. *Journal of the Royal Statistical Society: Series A*, **181** (4), https://doi.org/10.1111/rssa.1236

Thompson, M.E. (1997), Theory of Sample Surveys, Chapman & Hall.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso, *Journal of the Royal Statistical Society, Series B (methodological)*, **58** (1), 267–88.

Torabi, M., Datta, G.S. and Rao, J.N.K. (2009), Empirical Bayes Estimation of Small Area Means Under a Nested Error Linear Regression Model with Measurement Errors in the Covariates, *Scandinavian Journal of Statistics*, **36**, 355–368.

Torabi, M., and Rao, J.N.K. (2014), On Small Area Estimation under a Sub-area Model, *Journal of Multivariate Analysis*, **127**, 36–55.

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E., and Chambers, R. (2015), Robust Small Area Prediction for Counts, *Statistical Methods in Medical Research*, **24**(3), 373–395.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust prediction of small area means and distributions. *Australian and New Zealand Journal of Statistics*, **52**, 167–186.

Ugarte, M.D., Goicoa, T., Militino, A.F. and Durban, M. (2009), Spline Smoothing in Small Area Trend Estimation and Forecasting, *Computational Statistics and Data Analysis*, **53**, 3616–3629.

Valliant, R.L. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling, *Journal of the American Statistical Association*, **82** (398), 499–508.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2001), *Finite Population Sampling and Inference: A Prediction Approach*, New York: Wiley.

Verret, F., Rao, J. N. K. and Hidiroglou, M. A. (2015), Model-based small area estimation under informative sampling, *Survey Methodology*, **41**(2), 333–348.

Wagner, J., Münich, R., Hill, J. and Stoffels, J. (2017). Non-parametric small area models using shape-constrained penalized B-splines, *Journal of the Royal Statistical Society Series A*, **180**(4), DOI:10.1111/rssa.12295.

Wang, J., and Fuller, W.A. (2003), The Mean Squared Error of Small Area Predictors Constructed with Estimated Area Variances, *Journal of the American Statistical Association*, **98**, 718–723.

Wang, J., Fuller, W.A., and Qu, Y. (2008), Small Area Estimation Under Restriction, *Survey Methodology*, **34**, 29–36.

Ybarra, L.M.R. and Lohr, S.L. (2008), Small area estimation when auxiliary information is measured with error, *Biometrika*, **95**, 919–931.

You, Y. (1999), *Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation*, Unpublished Ph.D. Thesis, Carleton University, Ottawa, Canada.

You, Y. and Chapman, B. (2006), Small Area Estimation Using Area Level Models and Estimated Sampling Variances, *Survey Methodology*, **32**, 97–103.

You, Y. and Rao, J.N.K. (2002a), A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights, *Canadian Journal of Statistics*, **30**, 431–439.

You, Y., and Rao, J.N.K. (2002b), Small Area Estimation Using Unmatched Sampling and Linking Models, *Canadian Journal of Statistics*, **30**, 3–15.

Zhao, Q. (2006). User manual for povmap. World Bank. http://siteresources. worldbank. org/INTPGI/Resources/342674-1092157888460/Zhao_ManualPovMap.pdf.