UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working Papers
Statistics and Econometrics
16-05
ISSN 2387-0303
April 2016

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

Small area estimation of general parameters under complex sampling designs

María Guadarrama, Isabel Molina[a] and J.N.K. Rao[b]

Abstract

When the probabilities of selecting the individuals for the sample depend on the outcome values, we say that the selection mechanism is informative. Under informative selection, individuals with certain outcome values appear more often in the sample and therefore the sample is not representative of the population. As a consequence, usual model-based inference based on the actual sample without appropriate weighting might be strongly biased. For estimation of general non-linear parameters in small areas, we propose a model-based pseudo empirical best (PEB) method that incorporates the sampling weights and reduces considerably the bias of the unweighted empirical best (EB) estimators under informative selection mechanisms. We analyze the properties of this new method in simulation experiments carried out under complex sampling designs, including informative selection. Our results confirm that the proposed weighted PEB estimators perform significantly better than the unweighted EB estimators in terms of bias under informative sampling, and compare favorably under non-informative sampling. In an application to poverty mapping in Spain, we compare the proposed weighted PEB estimators with the unweighted EB analogues.

Keywords: Empirical best estimator; Nested-error model; Poverty mapping; Pseudo empirical best estimator; Unit level models.

[a] Department of Statistics, Universidad Carlos III de Madrid, Addriess: C/Madrid 126, 28903 Getafe (Madrid), Spain. Tf: +34 916249859, +34 916249887. E-mail: maria.guadarrama@uc3m.es, isabel.molina@uc3m.es

[b] School of Mathematics and Statistics, Carlenton University. E-mail: jrao@math.crlenton.ca

Acknowledgements: the second author acknowledge financial support from the Spanish Ministry of Education and Science, research project MTM2015-64842-P.

Small area estimation of general parameters under complex sampling designs*

María Guadarrama, Isabel Molina †
Department of Statistics, Universidad Carlos III de Madrid,
and J.N.K. Rao
School of Mathematics and Statistics, Carleton University.

Abstract: When the probabilities of selecting the individuals for the sample depend on the outcome values, we say that the selection mechanism is informative. Under informative selection, individuals with certain outcome values appear more often in the sample and therefore the sample is not representative of the population. As a consequence, usual model-based inference based on the actual sample without appropriate weighting might be strongly biased. For estimation of general non-linear parameters in small areas, we propose a model-based pseudo empirical best (PEB) method that incorporates the sampling weights and reduces considerably the bias of the unweighted empirical best (EB) estimators under informative selection mechanisms. We analyze the properties of this new method in simulation experiments carried out under complex sampling designs, including informative selection. Our results confirm that the proposed weighted PEB estimators perform significantly better than the unweighted EB estimators in terms of bias under informative sampling, and compare favorably under non-informative sampling. In an application to poverty mapping in Spain, we compare the proposed weighted PEB estimators with the unweighted EB analogues.

*Supported by the Spanish grant MTM2015-64842-P.

†Corresponding author. Department of Statistics, Universidad Carlos III de Madrid. Address: C/Madrid 126, 28903 Getafe (Madrid), Spain, Tf: +34 916249887. E-mail: isabel.molina@uc3m.es

Keywords: Empirical best estimator; Nested-error model; Poverty mapping; Pseudo empirical best estimator; Unit level models.

1 Introduction

In many applications, individuals with certain outcome values are more likely selected for the sample. For example, in forestry the largest trees may be more likely to be selected; in case-control studies, cases are typically selected with larger probability than controls. In such situations, we say that the selection mechanism (or sampling design) is informative. The result of an informative selection is a sample that is not representative of the target population and a weighting procedure is needed to downweight the outcomes of individuals that appear more often in the sample. This is the idea of design-based estimation, where the Horvitz-Thompson expansion estimator or the weighted sample mean are the basic estimators of a population mean. Design-based estimators are consistent when the sample size is large and do not require model assumptions. However, when estimating at highly disaggregated levels of a population (e.g. in counties), the sample sizes in some of these disaggregated areas might be very small, leading to unreliable design-based estimators for those small areas. This occurs because design-based estimators are *direct* in the sense of using only the sample observations from the corresponding target area. Small area estimation techniques obtain *indirect* estimators based on implicit or explicit models that link the data from all the areas through common parameters. These models increase the "effective" sample size considerably, leading to more efficient small area estimators, see Rao and Molina (2015) for an updated monograph on small area estimation.

For the estimation of general non-linear parameters for small areas, Molina and Rao (2010) introduced the empirical best (EB) method based on the unit level nested error model of Battese, Harter and Fuller (1977). Non-linear parameters of great interest are poverty or inequality indicators, which can be used to obtain poverty or inequality maps showing the regional distribution of poverty in a certain population or country. The World Bank has been producing poverty maps for many countries all over the world using traditionally the method of Elbers, Lanjouw and Lanjouw (2003), called here ELL method. Under the same model assumptions, EB method for poverty mapping outperforms ELL method when the area effects are significant, see Molina and Rao (2010). Both methods assume that the model for the sampled units is exactly the same as the model

considered for the population; in other words, the sample selection mechanism is not affecting the distribution of the outcomes (non-informative selection). In the case of informative selection, using the sample to obtain EB estimators of poverty indicators without any weighting will lead to biased estimators.

In the literature we can find two approaches to handle informative selection in small area estimation. The approach of Pfeffermann and Sverchkov (2007) is to calculate the sample likelihood as the usual likelihood conditional on the selected sample, where the inclusion probabilities are modeled in terms of the observed outcomes and covariates. In contrast, the approach of Verret et al. (2015) is to model the outcomes in terms of the sampling weights or inclusion probabilities and covariates, that is, to augment the assumed population model for the outcomes by including the weights or inclusion probabilities as an additional covariate. Both methods are used to estimate small area means and are not directly applicable to non-linear parameters. In fact, applying the augmenting model approach of Verret et al. (2015) for non-linear parameters would require to have the inclusion probabilities or sampling weights not only for the sample units, but for the non-sample units as well. In this paper we propose a very simple procedure that reduces the bias due to an informative selection mechanism based on combining the ideas of conditioning on the sample of the EB method with the correct weighting of design-based estimators. Instead of conditioning on the sample mean of the target area as EB method does, we propose to condition on the weighted sample mean using as weights the inverses of the inclusion probabilities. This leads to a weighted EB approach called here pseudo EB.

The paper is organized as follows. Section 2 introduces the assumed population model. Section 3 defines informative/non-informative selection. EB method is reviewed in Section 4 and our proposal is described in Section 5. A bootstrap procedure for mean squared error estimation is included in Section 6. Results of simulation experiments carried out under both informative and non informative selection are described in Section 7. Finally, Section 8 applies the proposed method to poverty mapping in Spanish provinces by gender and compares the resulting estimates with the unweighted EB estimates of Molina and Rao (2010).

2 Population model

In this paper, we wish to estimate a certain characteristic in each of m domains or areas U_i, $i = 1, \ldots, m$, into which our finite population U is partitioned. Each

domain U_i has population size N_i, $i = 1, \ldots, m$, where $N = \sum_{i=1}^{m} N_i$ is the total population size. We denote by Y_{ij} the measurement of the study variable for j-th unit within i-th domain. We wish to estimate possibly non-linear domain parameters that are separable, in the sense that they can be expressed as

$$H_i = \frac{1}{N_i} \sum_{j=1}^{N_i} h(Y_{ij}), \quad i = 1, \ldots, m, \tag{1}$$

where $h(\cdot)$ is a real measurable function. For the special case $h(y) = y$, we obtain the mean of domain i, that is, $H_i = \bar{Y}_i$.

We assume that the population measurements Y_{ij} follow the nested error model introduced by Battese et al. (1988),

$$\begin{aligned} &Y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij},\ v_i \overset{iid}{\sim} N(0, \sigma_v^2), \\ &e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2),\ j = 1, \ldots, N_i,\ i = 1, \ldots, m, \end{aligned} \tag{2}$$

where $\mathbf{x}_{ij}$ is a $p \times 1$ vector of auxiliary variables, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, v_i is the effect of domain i and e_{ij} is the individual regression error, where domain effects and errors are all mutually independent. Let us write the model in matrix notation by defining the domain vectors and matrices

$$\mathbf{y}_i = (Y_{i1}, \ldots, Y_{iN_i})',\ \mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i})',\ \mathbf{e}_i = (e_{i1}, \ldots, e_{iN_i})',\ i = 1, \ldots, m.$$

Then, model (2) becomes

$$\mathbf{y}_i \overset{ind}{\sim} N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i),\ \mathbf{V}_i = \sigma_v^2 \mathbf{1}_{N_i}\mathbf{1}_{N_i}' + \sigma_e^2 \mathbf{I}_{N_i},\ i = 1, \ldots, m, \tag{3}$$

where $\mathbf{1}_k$ denotes a vector of ones of size k and $\mathbf{I}_k$ is the $k \times k$ identity matrix. Additionally, we denote by $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_m')'$ the population vector of measurements, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_m')'$ is the population design matrix and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2)'$ is the vector of unknown model parameters.

3 Sample selection mechanism

The target domain parameters H_i, $i = 1, \ldots, m$, are estimated based on a sample s drawn from the population U using a given selection mechanism or sampling design. The sample s is composed of subsamples s_i, drawn independently from each domain U_i, $i = 1, \ldots, m$. Let n_i be the sample size of domain i, $i = 1, \ldots, m$. The total sample size is then $n = \sum_{i=1}^{m} n_i$. We denote by $r_i = U_i - s_i$ the set of out-of-sample units from domain i, of size $N_i - n_i$, $i = 1, \ldots, m$.

In this paper, we assume that the population matrix $\mathbf{X}$ of auxiliary variables is available from a census or a register. Then, all the probability distributions involved in this paper are conditional on $\mathbf{X}$ but we will omit this dependence in the notation for simplicity.

Traditional model-based inference assumes that the selection mechanism is noninformative. This means that the probability of the sample is not related with the outcome values. More formally, let $P(s|\mathbf{y})$ be the probability of sample s according to the selected sampling mechanism given $\mathbf{y}$. We say that the sampling design is noninformative when

$$P(s|\mathbf{y}) = P(s), \quad \forall \mathbf{y} \in I\!R^N, \forall s.$$

Equivalently, using Bayes Theorem, the sampling is noninformative when

$$f(\mathbf{y}|s) = f(\mathbf{y}), \quad \forall \mathbf{y} \in I\!R^N, \forall s.$$

Otherwise, we say that the sampling design is informative. Under noninformative sampling, $f(\mathbf{y}_s|s) = f(\mathbf{y}_s)$ and then inference based on the usual likelihood $f(\mathbf{y}_s)$ is valid. This means that the selection process does not affect the distribution of the outcomes for selected units.

4 EB method

This method assumes that the sampling design is noninformative. Then, the outcomes corresponding to sampled units preserve the same distribution as the outcomes for out-of-sample units, given by (2) under the considered nested error model. Let us decompose the domain vector $\mathbf{y}_i$ into subvectors corresponding to sample and out-of-sample elements as $\mathbf{y}_i = (\mathbf{y}'_{is}, \mathbf{y}'_{ir})'$, where the subscript s denotes the sample units and r the out-of-sample units. The sample data is then $\mathbf{y}_s = (\mathbf{y}'_{1s}, \ldots, \mathbf{y}'_{ms})'$. For a general domain parameter $H_i = H_i(\mathbf{y}_i)$, the best predictor is defined as the function of the sample observations $\mathbf{y}_s$ that minimizes the mean squared error (MSE) and is given by

$$\tilde{H}_i^B(\boldsymbol{\theta}) = E_{\mathbf{y}_{ir}}(H_i|\mathbf{y}_{is}; \boldsymbol{\theta}),$$

where the expectation is taken with respect to the distribution of $\mathbf{y}_{ir}|\mathbf{y}_{is}$, which depends on the true value of $\boldsymbol{\theta}$. For a domain parameter H_i that is separable as in (1), the best predictor reduces to

$$\tilde{H}_i^B(\boldsymbol{\theta}) = \frac{1}{N_i}\left[\sum_{j\in s_i} h(Y_{ij}) + \sum_{j\in r_i} \tilde{H}_{ij}^B(\boldsymbol{\theta})\right], \tag{4}$$

where $\tilde{H}_{ij}^B(\boldsymbol{\theta}) = E[h(Y_{ij})|\mathbf{y}_{is};\boldsymbol{\theta}]$ is also the best predictor of the out-of-sample element $H_{ij} = h(Y_{ij})$. The best predictor $\tilde{H}_i^B(\boldsymbol{\theta})$ is exactly model unbiased for H_i regardless of the complexity of the function $h(\cdot)$. However, it cannot be calculated in practice since model parameters $\boldsymbol{\theta}$ are typically unknown. An empirical best predictor (EB) of H_i, denoted as $\hat{H}_i^{EB}$, is then obtained by replacing $\boldsymbol{\theta}$ in $\tilde{H}_i^B(\boldsymbol{\theta})$ by a consistent estimator $\hat{\boldsymbol{\theta}}$, that is, $\hat{H}_i^{EB} = \tilde{H}_i^B(\hat{\boldsymbol{\theta}})$. The EB predictor is not exactly unbiased, but the bias arising from the estimation of $\boldsymbol{\theta}$ is typically negligible when the overall sample size n is large. For $h(\cdot)$ linear and under normality of $\mathbf{y}$, the EB predictor of H_i equals the empirical best linear unbiased predictor (EBLUP) of H_i.

Given the nested error model specified in (2) and assuming non-informative selection, the out-of-sample vectors $\mathbf{y}_{ir}$ given the sample data vectors $\mathbf{y}_{is}$ are independent and follow exactly the same distribution as $\mathbf{y}_{ir}|\bar{y}_{is}$, where $\bar{y}_{is}$ is the unweighted sample mean for area i. Thus, the best predictor of $H_{ij} = h(Y_{ij})$ is $\tilde{H}_i^B(\boldsymbol{\theta}) = E[h(Y_{ij})|\bar{y}_{is};\boldsymbol{\theta}]$. For an out-of-sample observation Y_{ij}, $j \in r_i$, we have $Y_{ij}|\bar{y}_{is} \sim N(\mu_{ij|s}, \sigma^2_{ij|s})$, where the conditional mean and variance are given respectively by

$$\mu_{ij|s} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_{is}(\bar{y}_{is} - \bar{\mathbf{x}}'_{is}\boldsymbol{\beta}), \quad \sigma^2_{ij|s} = \sigma^2_v(1-\gamma_{is}) + \sigma^2_e, \quad j \in r_i, \tag{5}$$

for $\bar{\mathbf{x}}_{is} = n_i^{-1}\sum_{j\in s_i}\mathbf{x}_{ij}$ and $\gamma_{is} = \sigma^2_v/(\sigma^2_v + \sigma^2_e/n_i)$.

Foster, Greer and Thorbecke (1984) introduced a family of poverty indicators, called here FGT poverty indicators, which contain several widely-used poverty measures and which are separable in the sense described above. In particular, the poverty maps released by World Bank are traditionally based on members of this family. Let E_{ij} be a welfare measure for individual j in area i and z be the poverty line. The family of FGT poverty indicators for domain i is given by

$$F_{\alpha i} = \frac{1}{N_i}\sum_{j=1}^{N_i} F_{\alpha ij}, \quad F_{\alpha ij} = \left(\frac{z - E_{ij}}{z}\right)^{\alpha} I(E_{ij} < z), \quad j = 1,\ldots,N_i, \alpha \geq 0, \tag{6}$$

where $I(E_{ij} < z) = 1$ if $E_{ij} < z$, and $I(E_{ij} < z) = 0$ otherwise. For $\alpha = 0$, we obtain the poverty incidence, measuring the frequency of income-based poverty. For $\alpha = 1$, we get the poverty gap, measuring the poverty depth. Both indicators together give a good description of poverty.

Consider that the population model (2) holds for $Y_{ij} = \log(E_{ij} + c)$, for a positive constant c. Then, we can express $F_{\alpha ij}$ in terms of the response variable Y_{ij} as

$$F_{\alpha ij} = \left[\frac{z - \exp(Y_{ij}) + c}{z}\right]^{\alpha} I[\exp(Y_{ij}) - c < z] =: h_\alpha(Y_{ij}),$$

which shows that $F_{\alpha i}$ is a separable parameter. According to (4), the best predictor of $H_i = F_{\alpha i}$ is given by

$$\tilde{F}^B_{\alpha i}(\boldsymbol{\theta}) = \frac{1}{N_i}\left(\sum_{j\in s_i} F_{\alpha ij} + \sum_{j\in r_i} \tilde{F}^B_{\alpha ij}(\boldsymbol{\theta})\right), \tag{7}$$

where $\tilde{F}^B_{\alpha ij}(\boldsymbol{\theta}) = E[h_\alpha(Y_{ij})|\bar{y}_{is};\boldsymbol{\theta}]$ is the best predictor of $F_{\alpha ij} = h_\alpha(Y_{ij})$. For $\alpha = 0, 1$, the best predictor $\tilde{F}^B_{\alpha ij}(\boldsymbol{\theta})$ can be calculated analytically. Let us define $\alpha_{ij} = [\log(z+c) - \mu_{ij|s}]/\sigma_{ij|s}$. Then, the best predictors of F_{0ij} and F_{1ij} are respectively given by

$$\tilde{F}^B_{0ij}(\boldsymbol{\theta}) = \Phi(\alpha_{ij}), \tag{8}$$

$$\tilde{F}^B_{1ij}(\boldsymbol{\theta}) = \Phi(\alpha_{ij})\left\{1 - \frac{1}{z}\left[\exp\left(\mu_{ij|s} + \frac{\sigma^2_{ij|s}}{2}\right)\frac{\Phi(\alpha_{ij} - \sigma_{ij|s})}{\Phi(\alpha_{ij})} - c\right]\right\}, \tag{9}$$

where $\Phi(\cdot)$ is the c.d.f. of a standard Normal random variable, $N(0,1)$.

For separable area parameters $H_i = N_i^{-1}\sum_{j=1}^{N_i} h(Y_{ij})$ with more complex $h(\cdot)$, analytical expressions may not be available. In any case, the EB predictor $\hat{H}^{EB}_{ij} = E[h(Y_{ij})|\bar{y}_{is};\hat{\boldsymbol{\theta}}]$ of a general $H_{ij} = h(Y_{ij})$ can be approximated by Monte Carlo, similarly as in Molina and Rao (2010). This is done by simulating L replicates $\{Y^{(\ell)}_{ij}; \ell = 1,\ldots,L\}$ of Y_{ij}, $j \in r_i$, from the estimated conditional distribution of $Y_{ij}|\bar{y}_{is}$ and then averaging over the L replicates as $\hat{H}^{EB}_{ij} = L^{-1}\sum_{\ell=1}^{L} h(Y^{(\ell)}_{ij})$.

A variation of EB method, called census EB, was defined by Guadarrama, Molina and Rao (2016) to handle the case when the sample units cannot be identified in the census of auxiliary variables, in which case the EB estimators, given by (7) with $\boldsymbol{\theta}$ replaced by a consistent estimator $\hat{\boldsymbol{\theta}}$, cannot be calculated. The census EB estimator is obtained by predicting the sample values H_{ij}, $j \in s_i$, as well as the out-of-sample ones H_{ij}, $j \in r_i$ as

$$\hat{H}^{CEB}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \hat{H}^{EB}_{ij}. \tag{10}$$

Typically the sampling fraction n_i/N_i is very small, and in that case the census EB estimator of H_i is approximately equal to the EB estimator.

5 Pseudo EB method

As stated above, under the nested error model (2), $\mathbf{y}_{ir}|\bar{y}_{is}$ follows exactly the same distribution as $\mathbf{y}_{ir}|\mathbf{y}_{is}$ and the best predictor of $H_{ij} = h(Y_{ij})$, $j \in r_i$ can

be expressed as $\tilde{H}_{ij}^{B} = E[h(Y_{ij})|\bar{y}_{is}]$. When the sample selection mechanism is informative, to avoid a bias due to a non-representative sample, the estimation procedure should incorporate the sampling weights. Let w_{ij} be the sampling weight of j-th unit within i-th domain and $w_{i\cdot} = \sum_{j\in s_i} w_{ij}$. We consider the same conditioning idea of the EB estimator, but now we condition on the weighted sample mean $\bar{y}_{iw} = w_{i\cdot}^{-1} \sum_{j\in s_i} w_{ij} y_{ij}$ instead of on the unweighted sample mean $\bar{y}_{is}$. Thus, we define the pseudo best (PB) estimator of $H_{ij} = h(Y_{ij})$ as

$$\tilde{H}_{ij}^{PB}(\boldsymbol{\theta}) = E[h(Y_{ij})|\bar{y}_{iw}; \boldsymbol{\theta}]. \tag{11}$$

The PB estimator of the separable area parameter H_i is then

$$\tilde{H}_{i}^{PB}(\boldsymbol{\theta}) = \frac{1}{N_i}\left[\sum_{j\in s_i} h(Y_{ij}) + \sum_{j\in r_i} \tilde{H}_{ij}^{PB}(\boldsymbol{\theta})\right]. \tag{12}$$

Jiang and Lahiri (2006) used a similar approach in the special case of area means under the nested error model and also in the case of a binary response variable and a logit linking model. However, their method is applicable only for area level covariates in the unit level models, unlike our method. For example, the area mean vector $\bar{\mathbf{X}}_i = N_i^{-1} \sum_{i=1}^{N_i} \mathbf{x}_{ij}$ may be used as area level covariates in the unit level model.

Similarly as in the EB method, the PB estimator (12) depends on the true values of the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2)'$, which need to be estimated. We define the pseudo EB (PEB) predictor as the PB predictor with $\boldsymbol{\theta}$ replaced by a consistent estimator such as maximum likelihood (ML), restricted ML (REML) estimators or estimators based on the method of moments (You and Rao, 2002).

For an out-of-sample variable Y_{ij}, $j \in r_i$, under the nested error population model (2), we have $Y_{ij}|\bar{y}_{iw} \overset{ind.}{\sim} N(\mu_{ij|s}^{w}, \sigma_{ir|s}^{2w})$, with conditional mean and variance given respectively by

$$\mu_{ij|s}^{w} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \gamma_{iw}(\bar{y}_{iw} - \bar{\mathbf{x}}_{iw}'\boldsymbol{\beta}), \quad \sigma_{ij|s}^{2w} = \sigma_v^2(1-\gamma_{iw}) + \sigma_e^2, \tag{13}$$

where $\bar{\mathbf{x}}_{iw} = w_{i\cdot}^{-1} \sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}$ and $\gamma_{iw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2\delta_i^2)$, for $\delta_i^2 = w_{i\cdot}^{-2} \sum_{j\in s_i} w_{ij}^2$. Observe that the mean $\mu_{ij|s}^{w}$ is obtained from $\mu_{ij|s}$ given in (5) by replacing the unweighted best predictor $\tilde{v}_{is} = \gamma_{is}(\bar{y}_{is} - \bar{\mathbf{x}}_{is}'\boldsymbol{\beta})$ of the domain effect v_i by its weighted version, given by $\tilde{v}_{iw} = \gamma_{iw}(\bar{y}_{iw} - \bar{\mathbf{x}}_{iw}'\boldsymbol{\beta})$.

For the FGT poverty indicators of order $\alpha = 0, 1$, the best predictors are given by (8) and (9) with $\mu_{ij|s}$ and $\sigma_{ij|s}^2$ replaced by the weighted versions $\mu_{ij|s}^{w}$ and $\sigma_{ij|s}^{2w}$. For more complex separable parameters, such as the FGT indicators for $\alpha > 1$, we can apply a Monte Carlo procedure to approximate the PEB predictor

of $H_{ij} = h(Y_{ij})$ similarly as done for the EB predictor. We generate L replicates $\{Y_{ij}^{(\ell)}; \ell = 1, \ldots, L\}$ of Y_{ij}, $j \in r_i$, from the estimated conditional distribution of $Y_{ij}|\bar{y}_{iw}$ and then average over the L replicates as $\hat{H}_{ij}^{PEB} = L^{-1} \sum_{\ell=1}^{L} h(Y_{ij}^{(\ell)})$.

Similarly as in the census EB estimator given in (10), we define the census PEB estimator as

$$\hat{H}_i^{CPEB} = \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{H}_{ij}^{PEB}. \tag{14}$$

Note that the census PEB estimator (14) is obtained by predicting all the population values, $H_{ij} = h(Y_{ij})$, $j \in U_i$.

For the special case of a domain mean $H_i = \bar{Y}_i$, if $\boldsymbol{\beta}$ is estimated by the weighted regression estimator $\hat{\boldsymbol{\beta}}_w$ given in You and Rao (2002), the census PEB estimator of $H_i = \bar{Y}_i$ equals the pseudo EBLUP of You and Rao (2002). Similarly, the PEB estimator obtained from (12) tends to the pseudo EBLUP as the domain sampling fraction $f_i = n_i/N_i$ becomes small. Thus, for a domain mean $\bar{Y}_i$, the census PEB estimator (and PEB for small domain sampling fraction) preserves the good properties of the pseudo EBLUP, which are: a) design consistency as n_i becomes large, and b) automatic benchmarking to the survey regression estimator of the overall population total, provided the sampling weights are calibrated to agree with the known population total $w_{i\cdot} = N_i$. Stefan (2005) and Verret et al. (2015) showed that the pseudo EBLUP of the area mean $\bar{Y}_i$ performs well under informative sampling in terms of bias and mean squared error (MSE).

6 Parametric bootstrap MSE estimator

The PEB estimators proposed in the previous section are essentially model-based even though they incorporate the sampling weights. For this reason, here we propose estimators of the MSE of PEB estimators under the model. For this, we consider a similar bootstrap procedure as in Molina and Rao (2010), based on the parametric bootstrap method for finite populations introduced by González-Manteiga et al. (2008). The parametric bootstrap estimator of the MSE of $\hat{H}_i^{PEB}$ is obtained as follows: i) Fit the model (2) to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$ and obtain estimators $\hat{\boldsymbol{\beta}}_w$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ of $\boldsymbol{\beta}$, σ_u^2 and σ_e^2 respectively. ii) For $b = 1, \ldots, B$, with B large, generate $v_i^{*(b)} \sim N(0, \hat{\sigma}_v^2)$ and $e_{ij}^{*(b)} \sim N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$, independently. iii) Construct B iid bootstrap population vectors $\mathbf{y}^{*(b)}$, $b = 1, \ldots, B$, with elements $Y_{ij}^{*(b)}$ generated as

$$Y_{ij}^{*(b)} = \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}_w + v_i^{*(b)} + e_{ij}^{*(b)}, \quad j = 1, \ldots, N_i, \ i = 1, \ldots, m.$$

From each bootstrap population b, calculate the true value of the domain parameter $H_i^{*(b)} = N_i^{-1} \sum_{j=1}^{N_i} h(Y_{ij}^{*(b)})$, $b = 1, \ldots, B$. iv) From each bootstrap population b, take the sample with the same indexes as the initial sample s and, using the sample elements $\mathbf{y}_s^{*(b)}$ of $\mathbf{y}^{*(b)}$ and the known population vectors $\mathbf{x}_{ij}$, $j \in U_i$, calculate the bootstrap pseudo EB predictors of H_i, denoted $\hat{H}_i^{PEB*(b)}$, $b = 1, \ldots, B$. v) A bootstrap estimator of MSE($\hat{H}_i^{PEB}$) is then

$$\text{mse}(\tilde{H}_i^{PEB}) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{H}_i^{PEB*(b)} - H_i^{*(b)} \right)^2. \tag{15}$$

7 Simulation experiments

We carried out simulation experiments to analyze the performance of the PEB estimators $\hat{F}_{\alpha i}^{PEB}$ of poverty incidences and gaps $F_{\alpha i}$, $\alpha = 0, 1$, compared to EB estimators $\hat{F}_{\alpha i}^{EB}$. We also compare with two types of direct estimators, namely the usual (unweighted) sample means (SMs) and the weighted sample means (WSMs), given respectively by

$$\bar{F}_{\alpha i} = \frac{1}{n_i} \sum_{j \in s_i} F_{\alpha ij}, \quad \bar{F}_{\alpha i,w} = \frac{1}{w_{i\cdot}} \sum_{j \in s_i} w_{ij} F_{\alpha ij}. \tag{16}$$

Since we are dealing with informative selection mechanisms but we are obtaining model-based estimators, our simulation experiments will be with respect to the joint distribution of the population vector $\mathbf{y}$ and the sample s; that is, under a model-design setup, where, in each Monte Carlo (MC) simulation, a population vector $\mathbf{y}$ is generated and a sample s is drawn according to a given selection mechanism. Subsections 7.1 and 7.2 describe two simulation experiments where the sample is drawn by (complex but) non-informative and informative selection mechanisms respectively.

7.1 Simulation study with non-informative selection

We consider the same simulation setup as in Molina and Rao (2010), where the population contains $N = 20,000$ units distributed into $m = 80$ domains, with $N_i = 250$ units in each domain $i = 1, \ldots, m$. We consider two dummy auxiliary variables, $x_q \in \{0, 1\}$, $q = 1, 2$, whose values are generated as $x_{q,ij} \sim \text{Bern}(p_{qi})$, $q = 1, 2$, with success probabilities given by $p_{1i} = 0.3 + 0.5\, i/m$ and $p_{2i} = 0.2$, $i = 1, \ldots, m$. The $x_{q,ij}$ values are kept fixed across simulations. The vector of true regression coefficients is taken as $\boldsymbol{\beta} = (3, 0.03, -0.04)'$ and the domain effects variance and error variance are respectively $\sigma_v^2 = 0.15^2$ and $\sigma_e^2 = 0.5^2$.

In each MC simulation out of $K = 1,000$, we generate a population vector $\mathbf{y}^{(k)}$, whose elements $Y_{ij}^{(k)}$ are generated from the nested error model (2). Using the population vector $\mathbf{y}^{(k)}$, we calculate the true values of the domain parameters $F_{\alpha i}^{(k)}$, $i = 1, \ldots, m$. We fixed the poverty line at $z = 12$, which is approximately 0.6 times the median of a population of incomes $\{E_{ij}; j = 1, \ldots, N_i,\ i = 1, \ldots, m\}$, where $E_{ij} = \exp(Y_{ij})$ with Y_{ij} generated as mentioned above. For each Monte Carlo population $k = 1, \ldots, K$, we draw a sample $s^{(k)}$. We use independent Poisson sampling within each domain i, with inclusion probability for individual j in the sample from domain i taken as $\pi_{ij} \sim \text{Beta}(\alpha_1, \alpha_2)$. We set $\alpha_1 = 2.5$ and select α_2 to achieve a specified expected domain sample size, $\bar{n}_i = K^{-1} \sum_{k=1}^{K} n_i^{(k)}$, where $n_i^{(k)}$ is the realized sample size in domain i in the k-th MC simulation replicate. We consider three expected domain sample sizes: $\bar{n}_i = 25, 50, 75$. To achieve approximately those domain sample sizes, we take $\alpha_2 = 25$, $\alpha_2 = 10$ and $\alpha_2 = 5$ respectively.

With the sample data from the k-th Monte Carlo population $\mathbf{y}_s^{(k)}$, we compute direct estimators of $F_{\alpha i}^{(k)}$, namely SM and also WSM as in (16), using as weights $w_{ij} = \pi_{ij}^{-1}$. We also compute EB and pseudo EB estimates of $F_{\alpha i}^{(k)}$, for $\alpha = 0, 1$ and $i = 1, \ldots, m$, using the population values of the auxiliary variables. For the EB estimator, we computed $\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$ and $\hat{\boldsymbol{\beta}}$ by the REML method. For the pseudo EB estimator, we used the weighted estimator $\hat{\boldsymbol{\beta}}_w$ given in You and Rao (2002) and the REML estimators of σ_v^2 and σ_e^2. We evaluate the performance of estimators in terms of relative bias (RB) and relative root MSE (RRMSE). Let $\hat{F}_{\alpha i}^{(k)}$ be one of the obtained estimates (SM, WSM, EB or pseudo EB) in MC replicate k. RB and RRMSE are approximated empirically as

$$\text{RB}(\hat{F}_{\alpha i}) = \frac{K^{-1} \sum_{k=1}^{K} (\hat{F}_{\alpha i}^{(k)} - F_{\alpha i}^{(k)})}{K^{-1} \sum_{k=1}^{K} F_{\alpha i}^{(k)}}, \quad \text{RRMSE}(\hat{F}_{\alpha i}) = \frac{\sqrt{K^{-1} \sum_{k=1}^{K} (\hat{F}_{\alpha i}^{(k)} - F_{\alpha i}^{(k)})^2}}{K^{-1} \sum_{k=1}^{K} F_{\alpha i}^{(k)}}.$$

Averages across domains of absolute RB ($\overline{\text{ARB}}$) and of RRMSE ($\overline{\text{RRMSE}}$) are also calculated as

$$\overline{\text{ARB}}_\alpha = m^{-1} \sum_{i=1}^{m} |\text{RB}(\hat{F}_{\alpha i})|, \quad \overline{\text{RRMSE}}_\alpha = m^{-1} \sum_{i=1}^{m} \text{RRMSE}(\hat{F}_{\alpha i}).$$

Figures 1, 2 and 3 display, respectively for $\bar{n}_i = 25, 50$ and 75, percent RB (left) and RRMSE (right) of the estimators of the poverty gap, F_{1i}, for each

domain $i = 1, \ldots, m$ (x-axis). These figures show that all the estimators display a small RB for the three expected sample sizes, although the WSM appears to be more unstable across domains than the other ones. This estimator also performs the worst in terms of RRMSE, followed by the SM. Thus, model-based estimators (EB and pseudo EB) appear to be significantly more efficient than the two types of direct estimators (SM and WSM) for all the domains. In this simulation experiment with non-informative sampling, weighted estimators (WSM and pseudo EB) loose efficiency with respect to the respective unweighted ones, but the efficiency loss of the pseudo EB turns out to be much smaller than the loss of the WSM with respect to the SM. As expected, the gain in efficiency of the model-based estimators compared to the direct estimators decreases as the expected sample size increases, with SMs becoming close to model-based estimators for the largest expected domain sample size $\bar{n}_i$ (Figure 3). Conclusions for the poverty incidence, F_{0i}, are similar and hence figures are not shown.

Table 1 displays averages of absolute RB and RRMSE across domains for the considered expected domain sample sizes. This table shows that $\overline{\text{ARB}}$ is small ($< 2\%$) for all the considered estimators and sample sizes. EB and pseudo EB estimators have considerably smaller $\overline{\text{RRMSE}}$ than direct estimators for small $\bar{n}_i$ and preserve smaller $\overline{\text{RRMSE}}$ even for the largest value of $\bar{n}_i$. Since the sample selection mechanism is in this case non-informative, the $\overline{\text{RRMSE}}$ of pseudo EB estimator turns out to be between 3% and 4% larger than that of EB estimator. This suggests that EB estimators work well under unequal probability sampling as long as the inclusion probabilities do not depend on the outcomes. Nevertheless, in this case pseudo EB estimator does not loose too much.

Table 1: Averages across domains of percent absolute RB and RRMSE for SM, WSM, EB and pseudo EB estimators of poverty incidence, F_{0i}, and poverty gap, F_{1i}, under non-informative selection with $\bar{n}_i = 25, 50, 75$.

	$\bar{n}_i = 25$				$\bar{n}_i = 50$				$\bar{n}_i = 75$			
	$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$		$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$		$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$	
Method	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}
SM	1.34	1.65	46.27	58.69	0.69	0.87	29.03	36.85	0.54	0.66	21.41	27.93
WSM	1.65	1.94	56.46	71.59	0.83	1.12	36.26	45.95	0.68	0.82	26.98	34.34
EB	0.74	0.89	28.21	35.60	0.46	0.60	20.99	26.73	0.40	0.47	17.58	22.29
PEB	0.88	1.04	31.25	39.29	0.54	0.72	24.13	30.43	0.49	0.61	20.07	25.39

Figure 1: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under non-informative selection with $\bar{n}_i = 25$.

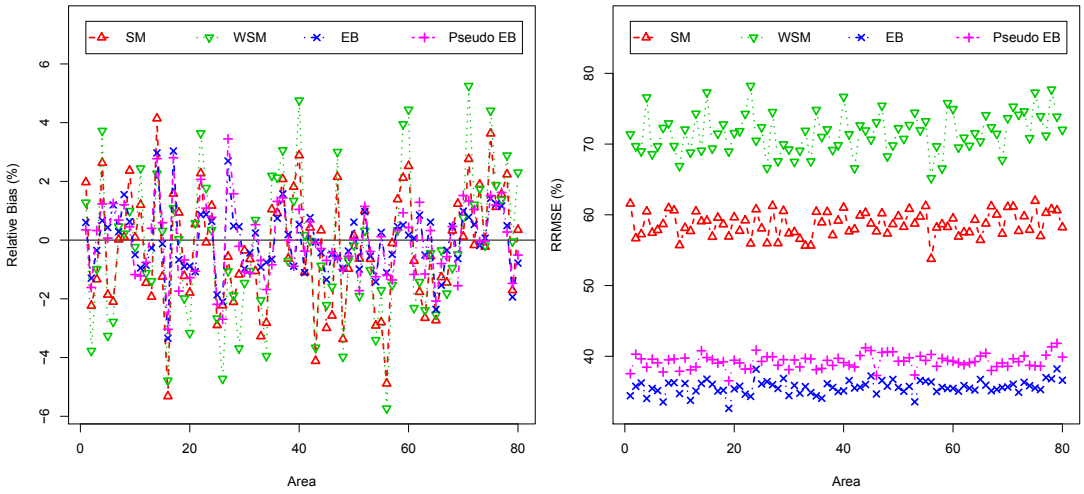


Figure 2: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under non-informative selection with $\bar{n}_i = 50$.

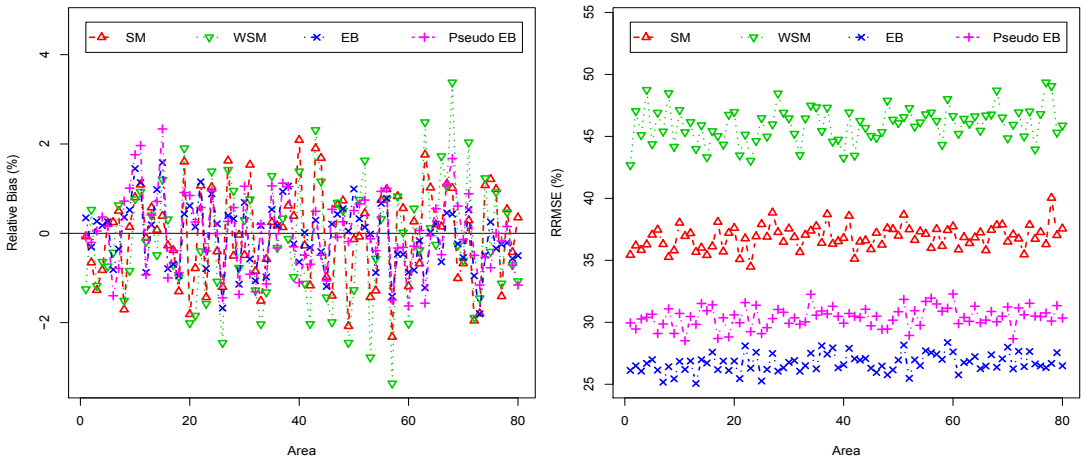
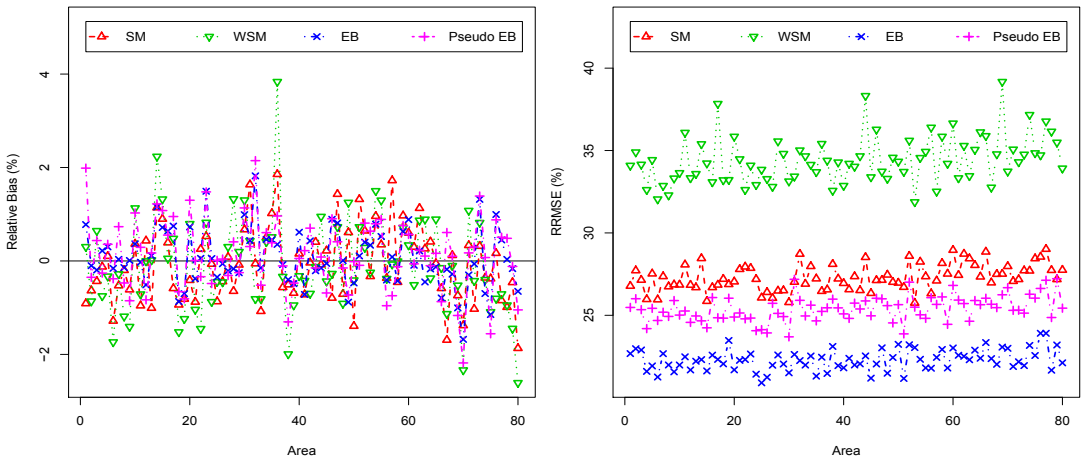


Figure 3: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under non-informative selection with $\bar{n}_i = 75$.



7.2 Simulation study with informative selection

A simulation experiment was carried out under the same setup as in Section 7.1, that is, with the same population structure and the same model that generates the population values. However, in this experiment, for each MC replicate, we draw the sample using an informative selection mechanism, where the probability of selecting a unit from a given domain depends on the outcome for that unit. Thus, again, we generate $K = 1,000$ population vectors $\mathbf{y}^{(k)}$, $k = 1, \ldots, K$ from the true nested error model (2). For each MC replicate k, we draw a sample $s^{(k)}$. The sample $s^{(k)}$ is drawn independently for each domain using Poisson sampling as in the previous experiment. However, in this case the inclusion probability, π_{ij}, for individual j in the sample from domain i depends on a random variable Z_{ij} that is correlated with the unexplained part of Y_{ij}, i.e, the model error e_{ij}. More concretely, each population unit j comes to the sample s_i from domain i according to a Bernoulli random value $Q_{ij} \sim \text{Bern}(\pi_{ij})$, with $\pi_{ij} = b^{-1}\exp(-aZ_{ij})$, for $a > 0$, $b > 0$, where $Z_{ij} \sim \text{Gamma}(\tau_{ij}, \theta_{ij})$, with model parameters τ_{ij} and θ_{ij} depending on the model error e_{ij}. Here, the degree of informativeness can be measured by the size of the correlation coefficient between Z_{ij} and e_{ij}. A 40% correlation coefficient is approximately achieved by taking $\tau_{ij} = 5 \times (2 + 0.25e_{ij})$ and $\theta_{ij} = 0.25 \times (2 + 0.25e_{ij})$. To make this simulation experiment comparable with the previous one, we take the same expected domain sample sizes $\bar{n}_i = 25, 50, 75$, which can be approximately obtained by fixing

$a = 0.15$ and then taking $b = 5.5$ for $\bar{n}_i = 25$, $b = 2.5$ for $\bar{n}_i = 50$ and $b = 1.5$ for $\bar{n}_i = 75$. From each sample $s^{(k)}$, the four estimators (SM, WSM, EB and pseudo EB) are computed.

Figures 4, 5 and 6 depict percent RB (left) and RRMSE (right) of the poverty gap, F_{1i}, for $\bar{n}_i = 25, 50$ and 75 respectively. These figures show how, when the inclusion probabilities are related with the outcome values, the two unweighted estimators (SM and EB) exhibit a substantial positive RB (about 15%). Comparing EB and pseudo EB estimators in terms of RRMSE, the situation is exactly the opposite of the previous simulation study, with pseudo EB estimators having smaller RRMSE than EB estimators for all the domains. For the poverty incidence, F_{0i}, plots are not shown because conclusions are similar.

Again, in Table 2 we can see $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$ of the estimators. This table confirms that the weighted estimators (WSM and pseudo EB) preserve a small $\overline{\text{ARB}}$ for the three considered expected domain sample sizes, whereas the unweighted estimators (SM and EB) have $\overline{\text{ARB}}$ over 13% for the poverty incidence, F_{0i}, and over 15% for the poverty gap, F_{1i}. In terms of $\overline{\text{RRMSE}}$, pseudo EB is more efficient than all the other estimators for the three considered expected domain sample sizes, but the WSM becomes close to the pseudo EB estimator for the largest $\bar{n}_i$. In terms of $\overline{\text{RRMSE}}$, the improvement of the pseudo EB over the unweighted EB estimator is not striking, but it is in terms of $\overline{\text{ARB}}$.

Table 2: Averages across domains of percent absolute RB and RRMSE for SM, WSM, EB and pseudo EB estimators of poverty incidence, F_{0i}, and poverty gap, F_{1i}, under informative selection with $\bar{n}_i = 25, 50, 75$.

	$\bar{n}_i = 25$				$\bar{n}_i = 50$				$\bar{n}_i = 75$			
	$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$		$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$		$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$	
Method	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}
SM	13.35	15.93	51.14	66.13	13.08	15.66	33.47	42.96	13.12	15.99	25.38	32.61
WSM	1.39	1.72	46.13	56.98	0.83	1.04	28.69	35.11	0.53	0.65	20.15	24.66
EB	13.25	16.15	31.27	39.27	13.09	15.83	24.80	30.98	13.16	16.04	21.53	26.94
PEB	0.79	0.99	29.06	36.59	0.47	0.63	21.94	27.71	0.44	0.55	17.95	22.75

We also studied the performance of the parametric bootstrap procedure described in Section 6 for estimation of the MSE of the pseudo EB estimator. We considered the same simulation setup as above, considering an informative sample, but since the proposed bootstrap procedure gives a model-based MSE, in this case we carry simulations only under the model (given the selected sample).

Figure 4: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under informative selection, $\bar{n}_i = 25$.

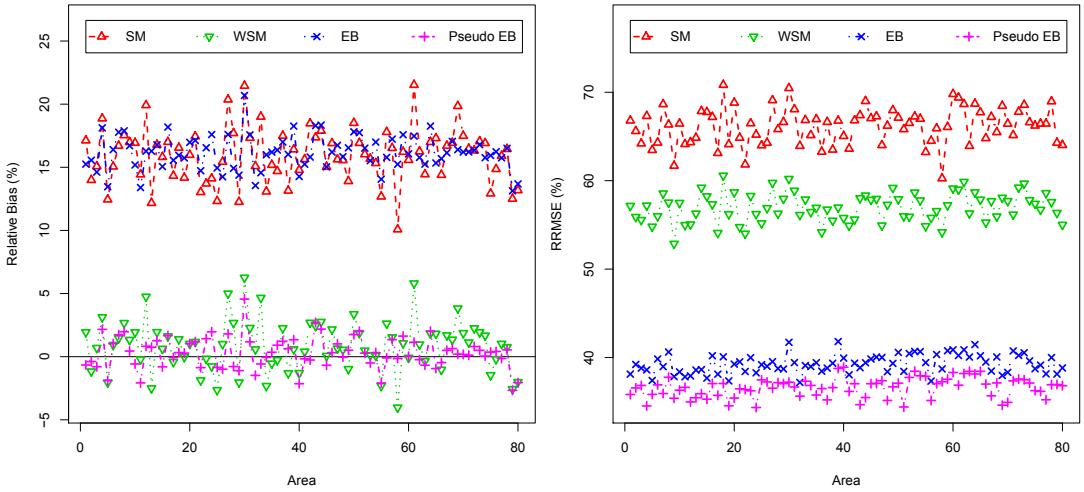


Figure 5: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under informative selection, $\bar{n}_i = 50$.

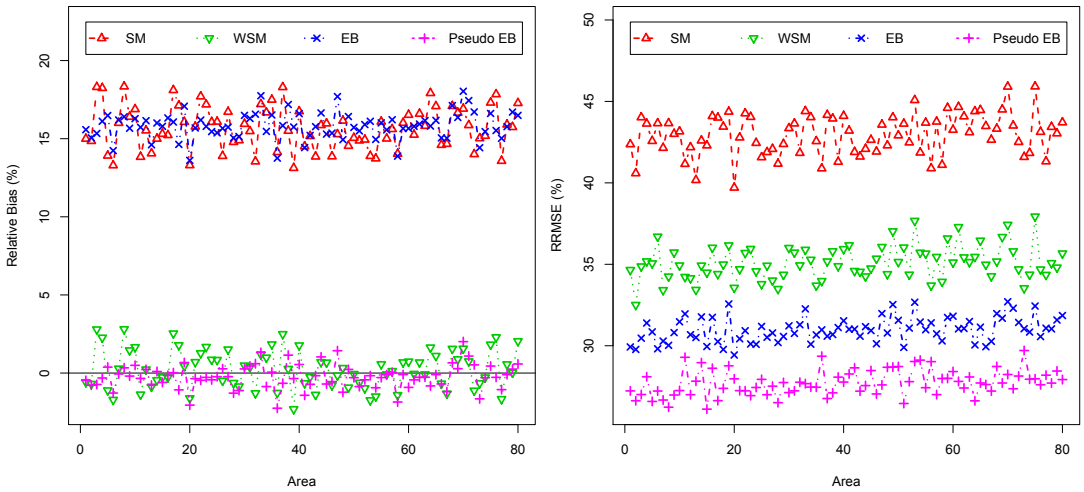
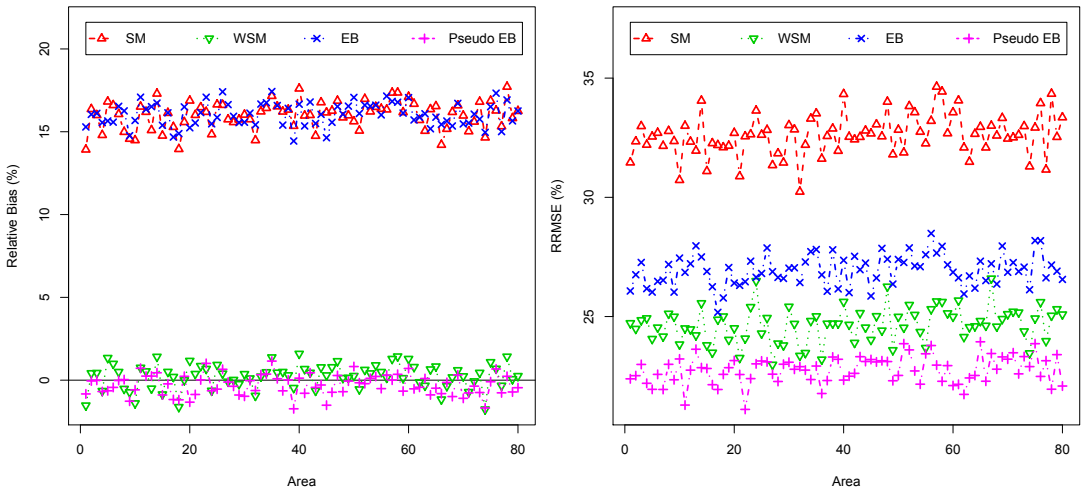


Figure 6: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under informative selection, $\bar{n}_i = 75$.

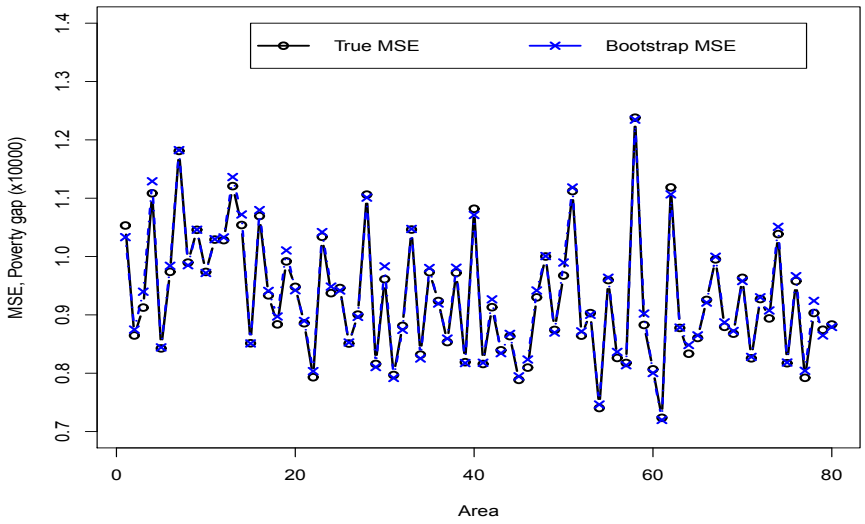


The true MSEs were previously approximated with $K = 50,000$ MC replicates. Then, we perform other $K = 500$ MC simulation replicates, and in each we calculate the bootstrap MSE estimators (15) with $B = 500$ bootstrap replicates. The expected values of the bootstrap MSE estimators across the $K = 500$ MC replicates are shown in Figure 7 together with the empirical MSEs for the poverty gap, F_{1i}, with $\bar{n}_i = 50$. This figure shows that the expected values of the bootstrap MSE estimator are almost equal to the true MSE values. Similar results were observed for the poverty incidence, F_{0i}, (not reported).

8 Application to poverty mapping in Spain

In this section we compare the performance of pseudo EB and EB estimators. For this, we consider the same application of Molina and Rao (2010), dealing with estimation of poverty incidences and gaps for the Spanish provinces by gender using the 2006 Spanish Survey on Income and Living Conditions (SILC). The SILC collects microdata on income and living conditions in a timely and comparable way across EU countries. The results obtained from the SILC are used for the structural index of social cohesion. The SILC survey provides reliable estimates for the overall Spain and for large Spanish regions (Autonomus Communities), but it does not allow reliable estimation for Spanish provinces by gender because of the small SILC sample sizes in some of these domains. Thus,

Figure 7: True MSEs of pseudo EB estimators of poverty gap, F_{1i}, and expected values of bootstrap MSE estimators with $B = 500$ bootstrap replicates, for each domain.



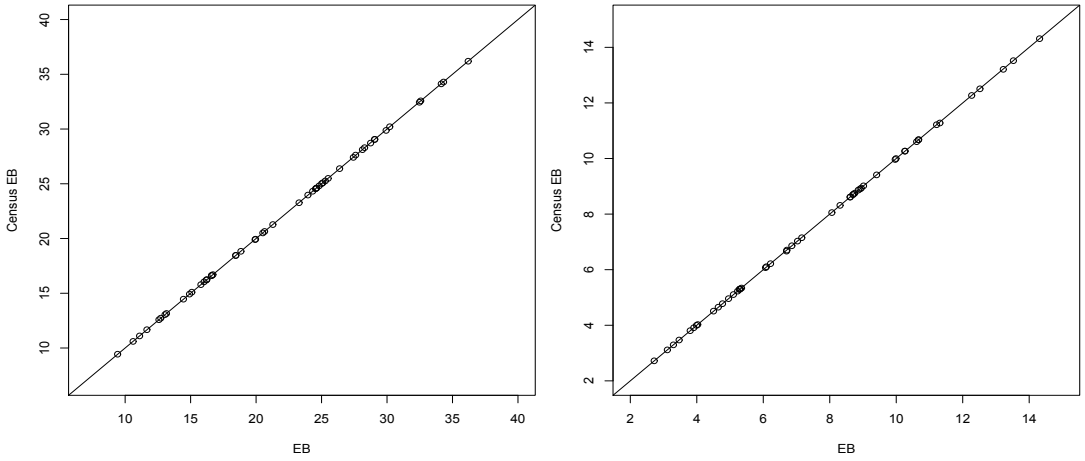
the small areas here are the $m = 52$ Spanish provinces for each gender. The overall sample size is 17,739 for women and 16,650 for men. The population size is 22,077,565 for women and 21,509,962 for men.

As auxiliary variables in the nested error model, we considered the same as in Molina and Rao (2010), namely the indicators of quinquennial age groups, of having Spanish nationality, of the three levels of the variable education level and of the three categories of the variable labor force status. Similarly as in Molina and Rao (2010), full census matrices $\mathbf{X}_i$ were constructed by replicating each record in the Spanish Labor Force Survey (LFS) a number of times equal to its LFS sampling weight. These matrices $\mathbf{X}_i$ were treated as the census matrices because the LFS has a very large sample size.

The welfare measure E_{ij} considered here is the equivalent annual net income, which is defined as the household annual net income divided by a measure of household size calculated according to the scale defined by OCDE. The poverty line was also computed as $z = 0.6 \times \text{Median}(E_{ij})$. Finally, due to the right skewness of the equivalent annual net income, we consider the same transformation as in Molina, Nandram and Rao (2014), given by $Y_{ij} = T(E_{ij}) = \log(E_{ij} + c)$, where c is selected such that the residuals obtained from the model fit, $\hat{e}_{ij} = Y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} - \hat{v}_i$, are approximately symmetric. We fitted separate models for women and men.

We compare the estimates obtained using the EB and pseudo EB methods and their estimated coefficients of variation (estimated RRMSEs). Instead of the original EB and pseudo EB methods, since here the sampling fractions are very small for all provinces, we applied the census EB and census PEB respectively. Figure 8 confirms that census EB estimates are approximately equal to EB estimates for all provinces in this application. As noted above, the same occurs for pseudo EB estimates.

Figure 8: Census EB estimates of poverty incidence, F_{0i}, (left) and poverty gap, F_{1i}, (right) against EB estimates for each province, i, for men.



In model-based inference, it is important to check the fitted model. Figure 9 shows a scatterplot of pseudo EB residuals, $\hat{e}_{ijw} = Y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_w - \hat{v}_{iw}$, against predicted values $\hat{Y}_{ij} = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_w + \hat{v}_{iw}$ in the model for men (left) and women (right). Plots look acceptable without any visible pattern. Figure 10 shows that even after the considered log-transformation, the distribution of pseudo EB residuals has slightly heavier tails than the normal distribution. These plots are practically identical to those obtained for EB residuals. Figure 11 shows Q-Q plots of estimated area effects $\hat{v}_{iw}$ under pseudo EB approach for each province, again for men (left) and for women (right). In this case, estimated random effects seem to follow a normal distribution.

Molina, Nandram and Rao (2014) analyzed graphically if the sampling weights are related with the response variables and no relation was observed. This indicates that, in this application, the sampling design is at most weakly informative. Thus, we expect only small differences between pseudo EB and EB estimators and their estimated CVs.

First of all we compare EB and pseudo EB estimates with usual direct esti-

Figure 9: Pseudo EB residuals against predicted values obtained from the model for men (left) and women (right).

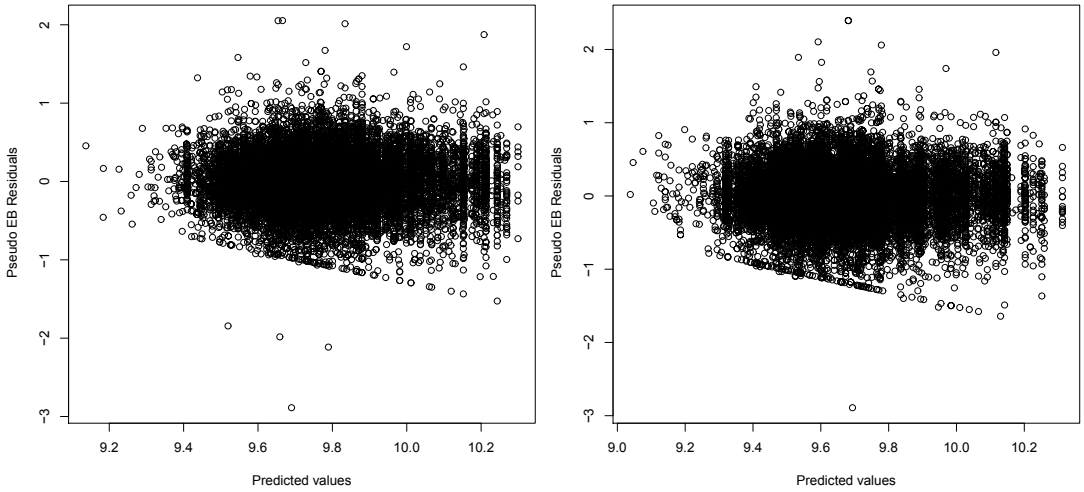


Figure 10: Q-Q plot of pseudo EB residuals obtained fro the model for men (left) and women (right).

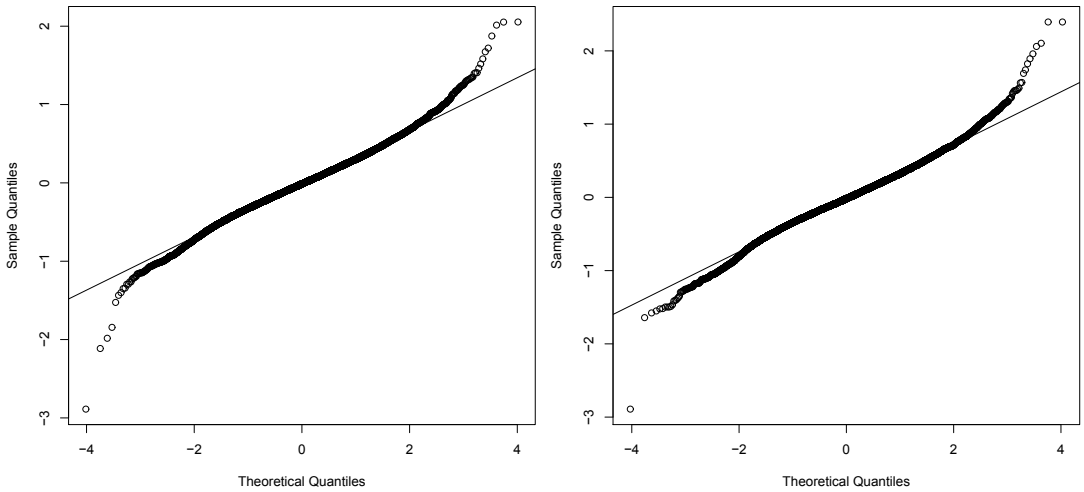
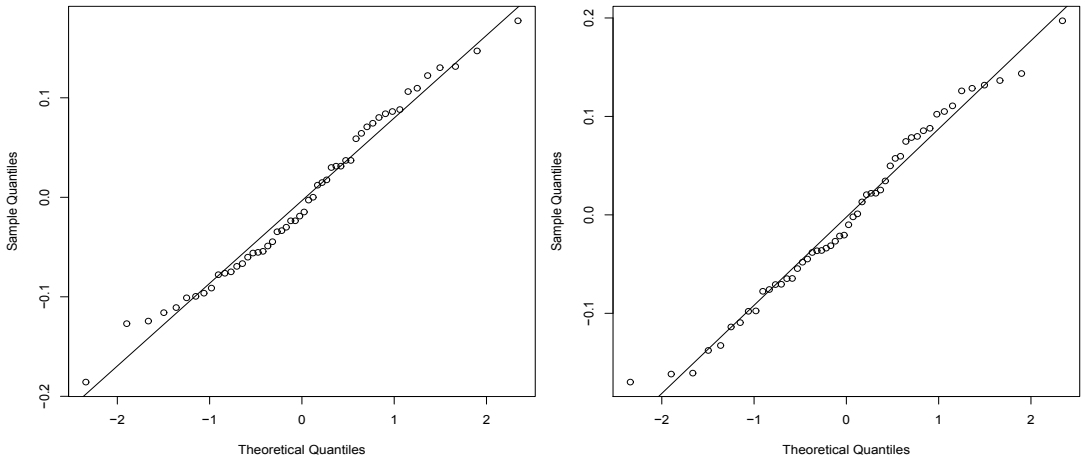


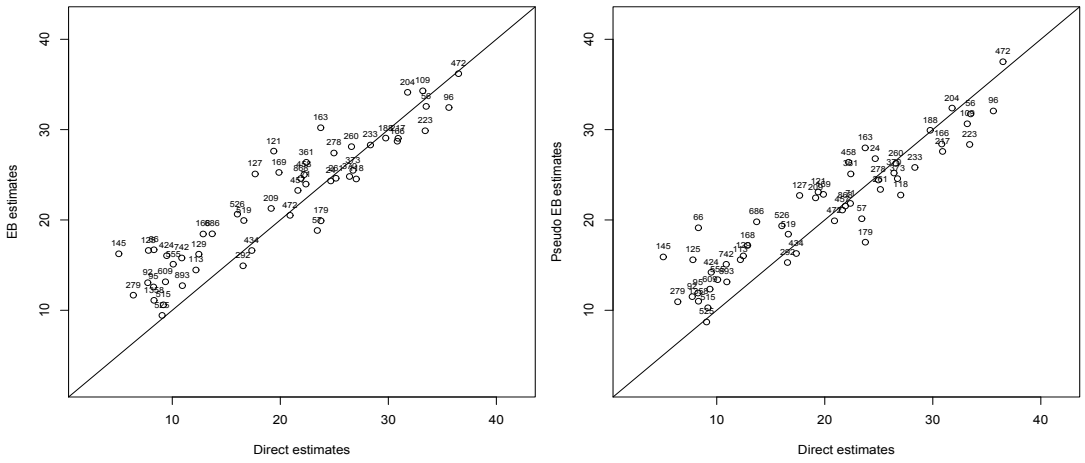
Figure 11: Q-Q plot of estimated random effects by pseudo EB for men (left) and women (right) for each province i.



mates (WSMs). Figure 12 displays EB estimates (left) and pseudo EB estimates (right) of poverty incidence for men against WSMs, with province sample sizes indicated in the point labels. On the left plot, we can see that most of the points are on the top-left side of the line, with only few points on the other side. The considered direct estimators (WSMs) are design-unbiased because sampling weights are calibrated so that $w_{i\cdot} = N_i$. Then, the fact that EB estimates for most domains are above direct estimates suggests that EB estimators are slightly biased upwards, and this bias could be in part due to a (weakly) informative sampling. Looking now at the right plot showing pseudo EB estimates against WSMs, this plot shows more points distributed at both sides of the line, which indicates that pseudo EB estimates have a smaller design bias than EB estimates. Results are similar for the poverty gap and also for women, so plots are not shown.

Tables 3 and 4 report obtained estimates with estimated bootstrap CVs for a selection of domains. CVs are in fact estimated RRMSEs. Since the considered direct estimators (WSMs) are ratio estimators, the MSE was calculated by using the Taylor linearization method. For EB estimators, the MSE was obtained using the parametric bootstrap approach of Molina and Rao (2010). Finally, for pseudo EB estimators, the MSE was approximated by the bootstrap procedure of Section 6. These tables report the results for a selection of domains. Concretely, we show the domains with sample sizes closest to minimum, maximum, first, second and third quartiles. In these tables, the three types of estimates agree to a some extent for the domains with larger sample sizes. However, direct estimates differ significantly for the two domains with smaller sample sizes, giving a much

Figure 12: EB estimates of poverty incidence for men in left panel and pseudo EB in right panel against ratio direct estimates for each province i.



larger estimate for Soria:Females and a much smaller estimate for Gerona:Males. As expected, since the selection is at most weakly informative, estimated CVs of pseudo EB estimators for these selected domains are slightly larger than those of EB estimators except for the domain with the smallest sample size. However, PEB estimators lead to large reduction in CV relative to direct estimators, while preserving a small bias under the design as shown in simulations.

Table 3: Results for poverty incidence F_{0i}: Direct, EB and pseudo EB estimates together with estimated coefficients of variation, cv, (%) for the Spanish provinces by gender with sample sizes closest to minimum, quantiles 0.25, 0.5, 0.75 and maximum.

Province	Gen	Dom	n_i	$\hat{F}_{0i}^{DIR}$	$\hat{F}_{0i}^{EB}$	$\hat{F}_{0i}^{PEB}$	$cv(\hat{F}_{0i}^{DIR})$	$cv(\hat{F}_{0i}^{EB})$	$cv(\hat{F}_{0i}^{PEB})$
Soria	F	42	17	55.62	32.70	36.93	42.69	15.36	14.61
Gerona	M	17	145	5.05	16.25	15.90	36.44	13.28	15.11
Jaén	F	23	230	33.86	32.84	30.83	11.78	5.60	6.92
Sevilla	M	41	472	20.90	20.51	19.90	10.64	6.10	6.36
Barcelona	F	8	1483	10.87	13.80	13.25	7.86	5.27	6.56

Let us now look at the estimates for each province. Figure 13 displays cartograms of EB (left) and pseudo EB (right) estimates of poverty incidence F_{0i} in Spanish provinces for women. Figure 14 shows the analogous estimates for the poverty gap. It is clear from these figures that the provinces with larger poverty incidence and poverty gap are those at the south and west of Spain. Neverthe-

Table 4: Results for poverty gap F_{1i}: Direct, EB and pseudo EB estimates together with estimated coefficients of variation, cv, (%) for Spanish provinces with sample sizes closest to minimun, quantiles 0.25, 0.5, 0.75 and maximum.

Province	Gen	Dom	n_i	$\hat{F}_{1i}^{DIR}$	$\hat{F}_{1i}^{EB}$	$\hat{F}_{1i}^{PEB}$	$cv(\hat{F}_{1i}^{DIR})$	$cv(\hat{F}_{1i}^{EB})$	$cv(\hat{F}_{1i}^{PEB})$
Soria	F	42	17	24.97	12.28	14.46	60.52	18.77	16.86
Gerona	M	17	145	1.87	5.31	5.26	40.74	19.56	21.80
Jaén	F	23	230	11.42	11.97	11.05	14.35	7.01	8.70
Sevilla	M	41	472	3.42	6.86	6.67	12.39	8.16	8.29
Barcelona	F	8	1483	3.62	4.11	3.96	10.26	8.13	10.10

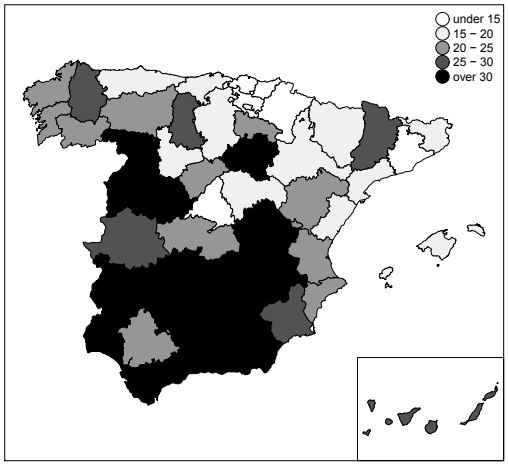
less, EB estimates give more provinces with largest poverty incidence (over 30%). For the poverty gap, the colors also tend to be darker for EB estimates. Maps for EB method are not exactly the same as those obtained in Molina and Rao (2010) in some of the provinces because here a separate model is fitted for men and women. Figure 15 shows the analogous plots for men. Again, EB estimates seem to give a larger number of very poor provinces than pseudo EB estimates according to both poverty incidence and gap. All these results indicate that EB estimates might be slightly biased upwards and pseudo EB estimates seem to be correcting this bias to some extent.

9 Conclusions

To handle informative selection when estimating separable non-linear small area parameters, we proposed pseudo EB estimators obtained as expected values with respect to the distribution of out-of-sample variables given the weighted sample means. This method combines the conditioning idea of the EB method for small area estimation of general parameters of Molina and Rao (2010) with the weighting approach of design-based inference. In our simulation studies, pseudo EB estimators reduce considerably the bias of EB estimators when the selection mechanism is informative. On the other hand, under a non-informative complex selection mechanism, the loss of efficiency is small. In the application, we obtained evidences of small upward bias of EB estimates, which seems to be reduced by pseudo EB estimates. Thus, pseudo EB estimates represent a compromise between model-based and design-based inference, reducing the design bias of purely model-based estimators but at the same time gaining efficiency with respect to direct estimators with the use of a model that represents the

Figure 13: Cartograms of estimated percent poverty incidences, F_{0i}, in Spanish provinces for women obtained with EB (left) and pseudo EB (right) methods.

EB Pov Incidence, Women



Pseudo EB Pov Incidence, Women

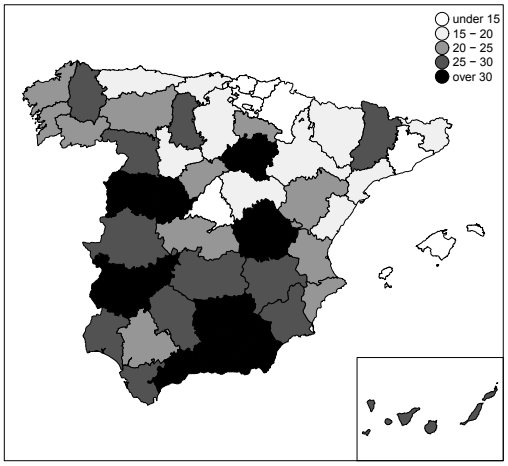
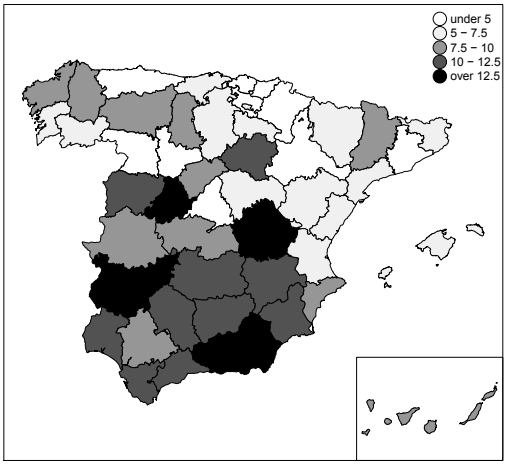


Figure 14: Cartograms of estimated percent poverty gap, F_{1i}, in Spanish provinces for women obtained with EB (left) and pseudo EB (right) methods.

EB Pov Gap, Women



Pseudo EB Pov Gap, Women

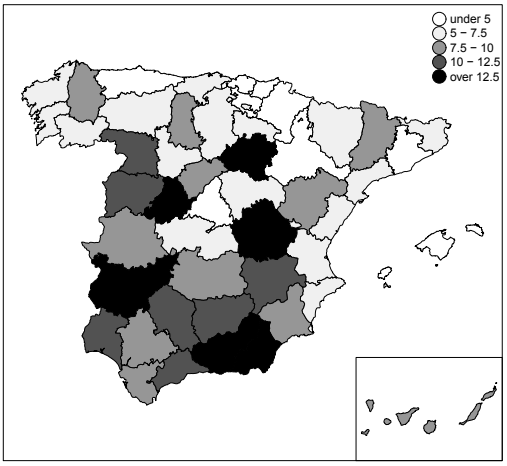
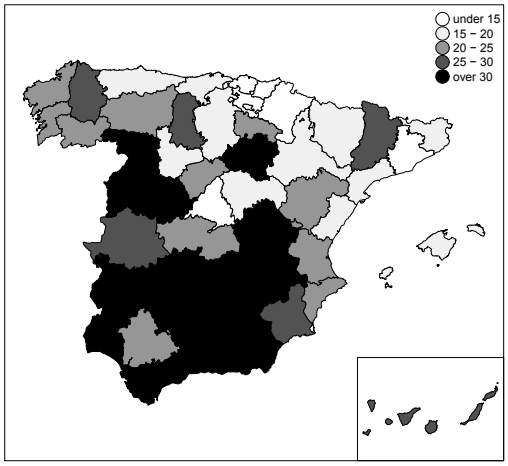


Figure 15: Cartograms of estimated percent poverty incidences, F_{0i}, in Spanish provinces for men obtained with EB (left) and Pseudo EB (right) methods.

EB Pov Incidence, Men



Pseudo EB Pov Incidence, Men

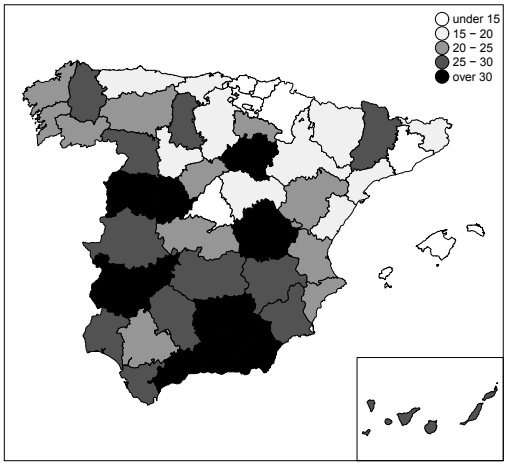
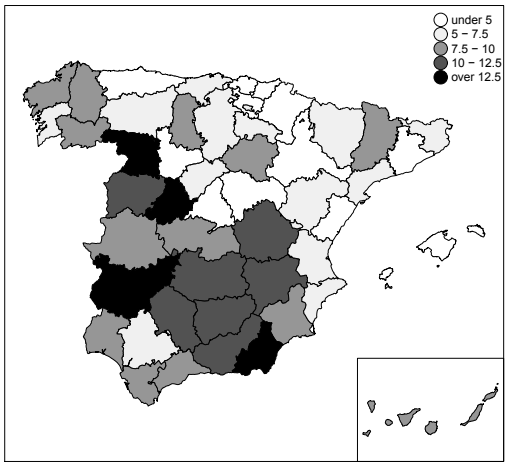
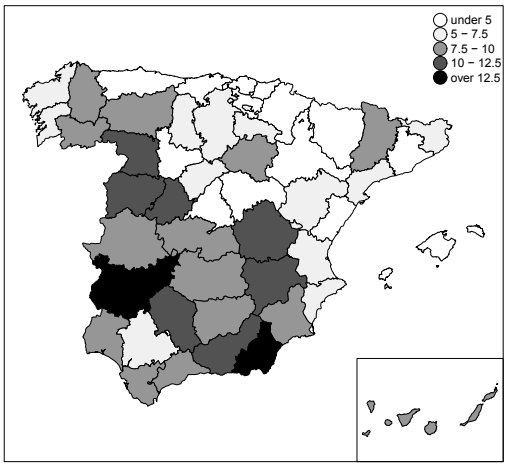


Figure 16: Cartograms of estimated percent poverty gap, F_{1i}, in Spanish provinces for men obtained with EB (left) and pseudo EB (right) methods.

EB Pov Gap, Men



Pseudo EB Pov Gap, Men



common factors that affect the outcomes in all the areas.

R codes of simulation studies are available under request.

Acknowledgements

We would like to express our gratitude to Prof. Yves Tillé for his valuable comments on this paper and interesting discussions on the topic.

10 References

Bates, D., Maechler, M. Bolker, B. and Walker, S. (2014) R package version 1.1-7, http://CRAN.R-project.org/package=lme4.

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, *J. Am. Statist. Ass.*, **83**, 28-36.

Elbers, C., Lanjouw, J. O. and Lanjouw, P. (2003) Micro-level Estimation of Poverty and Inequality. *Econometrica*, **71** (1), 355-364.

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

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

Guadarrama, M., Molina, I. and Rao, J.N.K. (2016) A comparison of small area estimation methods for poverty mapping. *Statist. Trans. new series and Surv. Methodol.*, **17**, (1), 41-66.

Jiang, J., and Lahiri, P. (2006). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *J. Am. Statist. Ass.*, **101**, 301-311.

Molina, I., and Mahuenda, Y. (2015) sae: An R Package for Small Area Estimation, *R Journal*, **7** (1), 81–98.

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. *Ann. Appl Statist*, **8** (2), 852-885.

Molina, I., and Rao, J.N.K. (2010) Small area estimation of poverty indicators. *Can. J. Statist.*, **38**, 369-385.

Pinheiro J, Bates D, DebRoy S, Sarkar D and R Core Team (2014) *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-117, http://CRAN.R-project.org/package=nlme.

Pfeffermann, D. and Sverchkov, M. (2007) Small-area estimation under informative probability sampling of areas and within the selected areas. *J. Am. Statist. Ass.*, **102** (480), 1427-1439.

Rao, J.N.K., and Molina, I. (2015) Small Area Estimation. *Hoboken, NY: Wiley*

Stefan, M. (2005) Contributions à l'estimation pour petits domaines. Ph.D. Thesis, Université Libre de Bruxelles.

Verret, F., Rao, J.N.K. and Hiridoglou, M.A. (2015) Model-based small area estimation under informative sampling. *Surv. Methodol.*, **41**, 333-347.

You, Y. and Rao, J.N.K. (2002) A pseudo-empirical best linear unbiased predictor approach to small area estimation using survey weights. *Can. J. Statist.*, **30** (3), 431-439.