LISER

LUXEMBOURG INSTITUTE OF SOCIO-ECONOMIC RESEARCH

n° 2019-01 January 2019

WORKING PAPERS

Small area estimation methods under cut-off sampling

Maria **GUADARRAMA**[1]
Isabel **MOLINA**[2]
Yves **TILLÉ**[3]

[1] Luxembourg Institute of Socio-Economic Research (LISER), Luxembourg
[2] Universidad Carlos III de Madrid, Spain
[3] Institut de Statistique, Université de Neuchâtel, Switzerland

Small area estimation methods under cut-off sampling*

María Guadarrama
Luxemoburg Institute of Socio-Economic Research (LISER), Luxembourg

Isabel Molina
Universidad Carlos III de Madrid, Spain

Yves Tillé
Institut de Statistique, Université de Neuchâtel, Switzerland

December, 2018

Abstract

Cut-off sampling is applied when there is a subset of units from the population from which getting the required information is too expensive or difficult and, therefore, those units are deliberately excluded from sample selection. If those excluded units are different from the sampled ones in the characteristics of interest, naïve estimators obtained by ignoring the cut-off sampling may be severely biased. Calibration estimators have been proposed to reduce the mentioned design-bias. However, the resulting estimators may have large variance when estimating in small domains. Similarly as calibration, model-based small area estimation methods using auxiliary information might decrease this bias if the assumed model holds for the whole population. At the same time, these methods provide more efficient estimators than calibration methods for small domains. We analyze the properties of calibration and model-based procedures for estimation of small domain characteristics under cut-off sampling. Our results confirm that the model-based estimators reduce the bias due to cut-off sampling and perform significantly better in terms of mean squared error.

Keywords: Calibration estimators; Cut-off sampling; EBLUP; EBP; Nested-error model; Unit level models.

* This work has been supported by the grant MTM2015-69638-R (MINECO/FEDER, UE)

Small area estimation methods under cut-off sampling

María Guadarrama, Isabel Molina and Yves Tillé*

Abstract: Cut-off sampling is applied when there is a subset of units from the population from which getting the required information is too expensive or difficult and, therefore, those units are deliberately excluded from sample selection. If those excluded units are different from the sampled ones in the characteristics of interest, naïve estimators obtained by ignoring the cut-off sampling may be severely biased. Calibration estimators have been proposed to reduce the mentioned design-bias. However, the resulting estimators may have large variance when estimating in small domains. Similarly as calibration, model-based small area estimation methods using auxiliary information might decrease this bias if the assumed model holds for the whole population. At the same time, these methods provide more efficient estimators than calibration methods for small domains. We analyze the properties of calibration and model-based procedures for estimation of small domain characteristics under cut-off sampling. Our results confirm that the model-based estimators reduce the bias due to cut-off sampling and perform significantly better in terms of mean squared error.

Keywords: Calibration estimators; Cut-off sampling; EBLUP; EBP; Nested-error model; Unit level models.

*María Guadarrama, Luxembourg Institute of Socio-Economic Research (LISER), 11, Porte des Sciences, Campus Belval L-4366 Esch-sur-Alzette, Luxembourg. Email: maria.guadarrama@liser.lu; Isabel Molina, Universidad Carlos III de Madrid, C/Madrid 126, 28903, Getafe, Madrid, Spain. Email: isabel.molina@uc3m.es; Yves Tillé, Institut de statistique, Université de Neuchâtel, 51, Av. de Bellevaux, 2000 Neuchâtel, Switzerland. Email: yves.tille@unine.ch

1 Introduction

Haziza et al. (2010) describes cut-off sampling as a technique in which a set of units is deliberately excluded from possible selection in a sample. For the OECD, it is a sampling procedure in which a threshold is established and all units at or above (below) the threshold are excluded from the possible selection in a sample. According to Särndal et al. (1992), pp. 531-533, this sampling technique is typically used when the distribution of the study variable is highly skewed and there is not a reliable frame covering the small elements. Benedetti et al. (2010) recognizes the advantage of cut-off sampling in terms of survey reduction cost. This procedure is often used in business surveys, where small firms are deliberately excluded from the sample due to difficulty of getting information from them. The cost of obtaining and maintaining the frame covering the whole population of firms is much higher than the gain in accuracy obtained from a sample drawn from this frame. The monthly survey of manufacturing performed by Statistics Canada is an example of cut-off sampling (Benedetti et al., 2010). In Spain, the monthly survey of industrial production index (IPI) performed by the Spanish National Statistical Institute (in Spanish, INE) collects data from those firms which produce a significant volume of products according to the annual industrial survey of products (in Spanish EIAP), see INE (2018). Related surveys, e.g. the index of industrial prices (IIP) and the index of business turnover (IBT) also use this sampling technique. This procedure leads to biased estimates since the inclusion probabilities for the excluded units are zero, see e.g. Särndal et al. (1992), Haziza et al. (2010) among others. Haziza et al. (2010) propose to use auxiliary information either at the design or at the estimation stage in order to reduce the bias when estimating population totals; more concretely, they propose to use balanced sampling and/or calibration.

In this work, we restrict ourselves to the estimation stage and study how cut-off sampling affects the estimation of domain (or area) parameters. We analyze some of the calibration methods proposed by Haziza et al. (2010) to reduce this problem. For domains with small sample size (small domains or areas), calibration estimators might suffer from large sampling variances. Alternatively, we consider small area estimation

methods. Concretely, for estimation of linear area parameters, we consider the empirical best linear unbiased predictor (EBLUP) and, for general non-linear parameters, the empirical best/Bayes predictor (EBP). We apply the methods studied in this work to the estimation of the total sales of certain tobacco product in the provinces from Spain.

The material is organized as follows. Section 2 describes the set-up of this paper. The following four sections describe the considered estimation methods, namely the basic direct estimators (Section 3), different approaches to calibration (Section 4), the EBLUP for estimation of linear parameters (Section 5) and the EBP for estimation of more general parameters in small domains (Section 6). Section 7 describes a bootstrap procedure for estimation of the mean squared error of the proposed small area estimators. Section 8 compares, through simulation experiments, the performance of the considered calibration small area estimators under cut-off sampling. Section 9 describes the application and, finally, Section 10 draws some conclusions.

2 Cut-off sampling in small areas

We consider a population U partitioned into m subsets U_i, $i = 1, \ldots, m$, called hereafter domains or areas, of sizes N_i, $i = 1, \ldots, m$, with $N = \sum_{i=1}^{m} N_i$. Independent samples are drawn from the different domains, where the sample s_i of size n_i from domain i is supposed to be drawn by cut-off sampling, $i = 1, \ldots, m$. This is done by excluding from the selection a subset of units $U_{iE} \subseteq U_i$. In other words, the domain U_i is partitioned into two subsets, U_{iI} and U_{iE}, called hereafter strata, of known sizes N_{iI} and N_{iE} respectively, with $N_i = N_{iI} + N_{iE}$. Stratum U_{iI} contains the units that can be potentially selected for the sample, called here the set of included units, whereas stratum U_{iE} contains those units that are excluded.

In the next three sections, we focus on estimation of domain totals or means of a variable of interest,

$$Y_i = \sum_{i=1}^{N_i} y_{ij}, \quad \bar{Y}_i = Y_i/N_i, \quad i = 1, \ldots, m,$$

where y_{ij} denotes the value of this variable for the j-th unit within the i-th domain. Under cut-off sampling within each domain, the sample s_i is supposed to be drawn from the subset of included individuals, U_{iI}, from domain i. Then, the inclusion probabilities for the included individuals ($j \in U_{iI}$) are $\pi_{ij} = \Pr(j \in s_i) > 0$ and $w_{ij} = \pi_{ij}^{-1}$ are the corresponding sampling weights. However, for the excluded units ($j \in U_{iE}$), the inclusion probabilities are zero and, therefore, sampling weights are not defined. As a consequence, for domains i with $U_{iE} \neq \emptyset$, design-unbiased estimators of Y_i or $\bar{Y}_i$ do not exist.

3 Basic direct estimators

We first consider basic direct estimators, obtained using only the n_i observations of the variable of interest from the target area. In absence of cut-off sampling, these estimators are design consistent as the domain sample size n_i increases. Moreover, they are nonparametric in the sense that do not require any model assumption. However, they may have unacceptable sampling errors in small domains. Moreover, as we shall see below, under cut-off sampling, their design bias might be substantial.

Note that, under cut-off sampling, the usual expansion estimator (Horvitz & Thompson, 1952) of Y_i obtained ignoring that the sample s_i is drawn only from U_{iI}, $\hat{Y}_i = \sum_{j \in s_i} w_{ij} y_{ij}$, actually estimates the total in the included strata, $Y_{iI} = \sum_{i \in U_{iI}} y_{ij}$, rather than the overall total $Y_i = Y_{iI} + Y_{iE}$, where $Y_{iE} = \sum_{i \in U_{iE}} y_{ij}$. Indeed, $E_\pi(\hat{Y}_i) = Y_{iI}$, where E_π denotes expectation under the sampling-replication mechanism, since the sampling weights $w_{ij} = \pi_{ij}^{-1}$ in $\hat{Y}_i$ expand to U_{iI} and not to U_i. No one would use this estimator since its bias, $B_\pi(\hat{Y}_i) = E_\pi(\hat{Y}_i) - Y_i = -Y_{iE}$ might be huge. However, in absence of any additional information, it would make much more sense to use Hájek estimator (Hájek, 1971) of the mean $\bar{Y}_i$, given by $\hat{\bar{Y}}_i^{HA} = \hat{Y}_i / \hat{N}_i$, where $\hat{N}_i = \sum_{j \in s_i} w_{ij}$, as estimator of the overall mean $\bar{Y}_i$. Then, one could estimate the total in domain i in terms of the mean estimator, $\hat{Y}_i^{HA} = N_i \hat{\bar{Y}}_i^{HA}$, considering that the means in the included and excluded strata are equal. Indeed, ignoring the ratio bias (of lower order) and noting that

$E_\pi(\hat{N}_i) = N_{iI}$, the approximate design-biases of $\hat{Y}_i^{HA}$ and $\hat{\bar{Y}}_i^{HA}$ are respectively given by

$$B_\pi(\hat{Y}_i^{HA}) \cong N_{iE}(\bar{Y}_{iI} - \bar{Y}_{iE}), \qquad B_\pi(\hat{\bar{Y}}_i^{HA}) \cong N_i^{-1} N_{iE}(\bar{Y}_{iI} - \bar{Y}_{iE}),$$

where $\bar{Y}_{iI} = Y_{iI}/N_{iI}$ and $\bar{Y}_{iE} = Y_{iE}/N_{iE}$ are the true means of the sets of included and excluded units from area i respectively (Haziza et al., 2010). For a domain i with $U_{iE} \neq \emptyset$, this bias vanishes only when these two means coincide ($\bar{Y}_{iI} = \bar{Y}_{iE}$), which is unlikely in the real cases where cut-off sampling is applied, see e.g. Haziza et al. (2010) or Section 9.

In the next section, we briefly describe calibration techniques as means to obtain estimators of reduced design bias in the context of estimation in small domains under cut-off sampling.

4 Calibration estimators

Calibration is applied when the true totals of certain auxiliary variables that are potentially correlated with the study variable are known. The idea of calibration is then adjusting the design weights applied in the expansion estimator of Y_i, so that the corresponding expansion estimators of the totals of the auxiliary variables match their known true values (calibration constraints). If the adjusted weights provide estimators of the available totals of the auxiliary variables that are absent of error, then one expects that they will also decrease the error in the estimation of the total of the study variable, provided that it is linearly related with the auxiliary variables. Even if there is an underlying linear model, calibration estimators are design-consistent as the area sample size n_i increases even if the model does not hold.

As we shall see below, under cut-off sampling, calibration estimators reduce the design-bias if the underlying linear model holds for the whole population (included and excluded units). However, for small domains, they might still have unacceptably large sampling errors.

Let us denote by $\mathbf{x}_{ij}$ the vector of auxiliary variables for unit j within domain i. Depending on whether the domain totals or only the population totals of these auxiliary

variables are available, we can apply different calibration approaches. First, consider that the domain total $\mathbf{X}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is available. In that case, one approach to calibration, using for illustration the chi-squared distance, is to find the calibration weights h_{ij} for the sample units in the domain, $j \in s_i$, that minimize the sum of distances to the original weights, $G_{ij}(h, w) = (h_{ij} - w_{ij})^2/w_{ij}$, for those units $j \in s_i$, subject to a set of calibration constraints for the same domain i. In this case, the calibration weights for the sample units in domain i, h_{ij}, $j \in s_i$, are the solution of the domain-specific problem

$$\begin{aligned} \min_{\{h_{ij}: j \in s_i\}} \quad & \sum_{j \in s_i} (h_{ij} - w_{ij})^2/w_{ij} \\ \text{s.t.} \quad & \sum_{j \in s_i} h_{ij}\mathbf{x}_{ij} = \mathbf{X}_i. \end{aligned} \tag{1}$$

Typically, this problem is solved by the method of Lagrange multipliers. Defining the lagrangian domain-specific function

$$L_i = \sum_{j \in s_i} (h_{ij} - w_{ij})^2/w_{ij} + 2\boldsymbol{\lambda}_i' \left(\sum_{j \in s_i} h_{ij}\mathbf{x}_{ij} - \mathbf{X}_i \right),$$

$\boldsymbol{\lambda}_i$ is the vector of Lagrange multipliers for that domain, taking derivatives of L_i with respect to h_{ij}, $j \in s_i$, and $\boldsymbol{\lambda}_i$ and equating them to zero, we obtain the calibration weights that solve the above problem. Denoting $\hat{\mathbf{X}}_i = \sum_{j \in s_i} w_{ij}\mathbf{x}_{ij}$ to the usual expansion estimator of $\mathbf{X}_i$, these calibration weights are given by

$$\begin{aligned} h_{ij} &= w_{ij}(1 + \mathbf{x}_{ij}'\boldsymbol{\lambda}_i), \ j \in s_i, \\ \boldsymbol{\lambda}_i &= \left(\sum_{j \in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}' \right)^{-1} (\mathbf{X}_i - \hat{\mathbf{X}}_i), \end{aligned} \tag{2}$$

provided that $\sum_{j \in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}'$ is non-singular. Note that the new weights in (2) are obtained as an adjustment of the ordinary design weights, $h_{ij} = w_{ij}a_{ij}$, with adjustment factors

$$a_{ij} = 1 + (\mathbf{X}_i - \hat{\mathbf{X}}_i)' \left(\sum_{j \in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}' \right)^{-1} \mathbf{x}_{ij}. \tag{3}$$

The calibration estimator of the domain total Y_i is then given by the expansion estimator based on the adjusted weights h_{ij}, that is,

$$\hat{Y}_i^{LCAL} = \sum_{j \in s_i} h_{ij} y_{ij}. \tag{4}$$

Now replacing the adjusted weights $h_{ij} = w_{ij} a_{ij}$ in (4) for a_{ij} given in (3), the calibration estimator of Y_i turns out to be equal to the generalized regression (GREG) estimator,

$$\hat{Y}_i^{LCAL} = \hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)' \hat{\mathbf{B}}_i =: \hat{Y}_i^{GREG}, \tag{5}$$

where we have used the notation $\hat{\mathbf{B}}_i = (\sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}')^{-1} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} y_{ij}$. Note that $\hat{\mathbf{B}}_i$ is the weighted least squares (WLS) estimator of the vector of regression coefficients $\boldsymbol{\beta}_i$ in the following linear regression model for the units in domain i:

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta}_i + \epsilon_{ij}, \ E_m(\epsilon_{ij}) = 0, \ E_m(\epsilon_{ij}^2) = \sigma_\epsilon^2, \quad j = 1, \ldots, N_i. \tag{6}$$

Thus, in (5), the regression corrects the bias of the basic expansion estimator $\hat{Y}_i$ as estimator of Y_i with the help of the known domain totals in $\mathbf{X}_i$.

In the above procedure, estimating for all the domains involves solving the corresponding m calibration problems and requires availability of the m vectors of totals $\mathbf{X}_i$, $i = 1, \ldots, m$. In the case that only the overall population total $\mathbf{X} = \sum_{i=1}^{m} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is known, a different calibration estimator can be applied by minimizing the sum of distances at the population level subject to a calibration constraint for the population total. In this case, calibration weights are obtained at once for all the sample units, g_{ij}, $j \in s_i$, $i = 1, \ldots, m$, by solving the following calibration problem:

$$\begin{aligned} \min_{\{g_{ij}: j \in s_i\}} \quad & \sum_{i=1}^{m} \sum_{j \in s_i} (g_{ij} - w_{ij})^2 / w_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^{m} \sum_{j \in s_i} g_{ij} \mathbf{x}_{ij} = \mathbf{X}. \end{aligned} \tag{7}$$

Defining now the Lagrangian function

$$L = \sum_{i=1}^{m} \sum_{j \in s_i} (g_{ij} - w_{ij})^2 / w_{ij} + 2\boldsymbol{\lambda}' \left(\sum_{i=1}^{m} \sum_{j \in s_i} g_{ij} \mathbf{x}_{ij} - \mathbf{X}_i \right),$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers, taking derivatives with respect to g_{ij} and $\boldsymbol{\lambda}$ and equating to zero, we obtain the new calibration weights for all the sample units. These weights are given by

$$\begin{aligned} g_{ij} &= w_{ij}(1 + \mathbf{x}'_{ij}\boldsymbol{\lambda}), \quad j \in s_i,\ i = 1, \ldots, m, \\ \boldsymbol{\lambda} &= \left(\sum_{i=1}^{m} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \mathbf{x}'_{ij} \right)^{-1} (\mathbf{X} - \hat{\mathbf{X}}), \end{aligned} \tag{8}$$

provided that $\sum_{i=1}^{m} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \mathbf{x}'_{ij}$ is non-singular. The resulting calibration estimator of the domain total Y_i is then obtained as

$$\hat{Y}_i^{LCALN} = \sum_{j \in s_i} g_{ij} y_{ij} = \hat{Y}_i + (\mathbf{X} - \hat{\mathbf{X}})' \hat{\mathbf{B}}_i^N, \tag{9}$$

where, in this case, $\hat{\mathbf{B}}_i^N = (\sum_{\ell=1}^{m} \sum_{j \in s_\ell} w_{\ell j} \mathbf{x}_{\ell j} \mathbf{x}'_{\ell j})^{-1} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} y_{ij}$. Note that the regression correction in $\hat{Y}_i^{LCALN}$ uses the national total $\mathbf{X}$ and its corresponding expansion estimator unlike the GREG estimator given in (5).

The calibration (or GREG) estimator (5) is expected to have smaller design-bias than (9) because it fits a different regression model for each domain i. On the other hand, for domains with small sample sizes n_i, its variance may be large since it uses only the domain-specific data. The alternative calibration estimator given in (9) is expected to have slightly larger design-bias because the calibration problem is solved at the national level, but its design-variance should be smaller. Let us now study more formally these properties. For this, we consider the theoretical version of the LCAL estimator of the domain total (5), given by

$$\tilde{Y}_i^{LCAL} = \hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)' \mathbf{B}_{iI}, \tag{10}$$

where here $\mathbf{B}_{iI} = (\sum_{j\in U_{iI}} \mathbf{x}_{ij}\mathbf{x}'_{ij})^{-1} \sum_{j\in U_{iI}} \mathbf{x}_{ij}y_{ij}$ is the census LS estimator of $\boldsymbol{\beta}_i$ in model (6) based on the included units from domain i. Note that the sample s_i is drawn only from U_{iI} and thus $\hat{\mathbf{B}}_i$ estimates $\mathbf{B}_{iI}$. Using the facts that $E_\pi(\hat{Y}_i) = Y_{iI}$ and $E_\pi(\hat{\mathbf{X}}_i) = \mathbf{X}_{iI}$, where $\mathbf{X}_{iI} = \sum_{j\in U_{iI}} \mathbf{x}_{ij}$ and noting that $\mathbf{X}_i - \mathbf{X}_{iI} = \mathbf{X}_{iE}$, for $\mathbf{X}_{iE} = \sum_{j\in U_{iE}} \mathbf{x}_{ij}$, we obtain the design-bias of this LCAL theoretical estimator, given by

$$B_\pi(\tilde{Y}_i^{LCAL}) = -(Y_{iE} - \mathbf{X}'_{iE}\mathbf{B}_{iI}). \tag{11}$$

Now since the calibration estimator $\hat{Y}_i^{LCAL}$ is intended to estimate actually Y_i for the overall population rather than for the included units, for the domain mean $\bar{Y}_i = Y_i/N_i$, we consider the LCAL estimator given simply by $\hat{\bar{Y}}_i^{LCAL} = \hat{Y}_i^{LCAL}/N_i$. The bias of the corresponding theoretical LCAL estimator of the mean, $\tilde{\bar{Y}}_i^{LCAL} = \hat{\bar{Y}}_i + (\bar{\mathbf{X}}_i - \hat{\bar{\mathbf{X}}}_i)'\mathbf{B}_{iI}$, is then given by

$$B_\pi(\tilde{\bar{Y}}_i^{LCAL}) = -\frac{N_{iE}}{N_i}(\bar{Y}_{iE} - \bar{\mathbf{X}}'_{iE}\mathbf{B}_{iI}). \tag{12}$$

This bias is small when either the proportion of excluded units is small, or when the model for the included individuals also holds for the excluded ones. In fact, if the linear regression model (6) actually holds for all the units in the domain (included and excluded), then $E_m(\mathbf{B}_{iI}) = \boldsymbol{\beta}_i$, which is constant for the included and excluded units, where here E_m denotes expectation under model (6). Taking now expectation of (12) under the model (6), we obtain the bias under the model and the sampling replication mechanism (model-design bias), given by

$$\begin{aligned} B_{m,\pi}(\tilde{\bar{Y}}_i^{LCAL}) &= -\frac{N_{iE}}{N_i}\left\{E_m(\bar{Y}_{iE}) - \bar{\mathbf{X}}'_{iE}E_m(\mathbf{B}_{iI})\right\} \\ &= -\frac{N_{iE}}{N_i}\left(\bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i - \bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i\right) = 0. \end{aligned}$$

In contrast, assuming exactly the same regression model, the bias of the basic direct estimator $\hat{\bar{Y}}_i^{HA}$ under cut-off sampling is not zero unless the means of the auxiliary

variables for the excluded and included units are equal. Indeed,

$$\begin{aligned} B_{m,\pi}(\hat{\bar{Y}}_i^{HA}) &= \frac{N_{iE}}{N_i} E_m(\bar{Y}_{iI} - \bar{Y}_{iE}) \\ &= \frac{N_{iE}}{N_i} (\bar{\mathbf{X}}_{iI} - \bar{\mathbf{X}}_{iE})' \boldsymbol{\beta}_i. \end{aligned} \tag{13}$$

We have seen that the condition under which the LCAL estimator is design-unbiased, namely that the linear model (6) holds for all the units in the domain, is much weaker than the requirements for the basic direct estimator to be design-unbiased. This means that calibration estimators will tend to be less biased than the considered basic direct estimator.

For the alternative calibration estimator (9), we define similarly its theoretical version

$$\tilde{Y}_i^{LCALN} = \hat{Y}_i + (\mathbf{X} - \hat{\mathbf{X}})' \mathbf{B}_{iI}^N, \tag{14}$$

where here, $\mathbf{B}_i^N = (\sum_{\ell=1}^m \sum_{j \in U_{\ell I}} \mathbf{x}_{\ell j} \mathbf{x}'_{\ell j})^{-1} \sum_{j \in U_{iI}} \mathbf{x}_{ij} y_{ij}$. Using the decomposition $\mathbf{X} = \mathbf{X}_I + \mathbf{X}_E$, where $\mathbf{X}_I$ and $\mathbf{X}_E$ are the national totals for the included and excluded units respectively, we obtain the design bias of $\tilde{Y}_i^{LCALN}$, given by

$$B_\pi(\tilde{Y}_i^{LCALN}) = -\left(Y_{iE} - \mathbf{X}'_E \mathbf{B}_{iI}^N\right). \tag{15}$$

Consider now the linear model with constant regression coefficients for all the population units, called model m_2,

$$y_{ij} = \mathbf{x}'_{ij} \boldsymbol{\beta} + \epsilon_{ij},\ E_{m_2}(\epsilon_{ij}) = 0,\ E_{m_2}(\epsilon_{ij}^2) = \sigma_\epsilon^2, \quad j = 1, \ldots, N_i,\ i = 1, \ldots, m. \tag{16}$$

Note that, under this model, $E_{m_2}(\mathbf{B}_{iI}^N) \neq \boldsymbol{\beta}$ in general, but considering instead the sum $\mathbf{B}_I = \sum_{i=1}^m \mathbf{B}_{iI}^N$, we have $E_{m_2}(\mathbf{B}_I) = \boldsymbol{\beta}$. This means that the LCALN estimator for particular domain, $\tilde{Y}_i^{LCALN}$, is not model-design unbiased, because

$$B_{m_2,\pi}(\tilde{Y}_i^{LCALN}) = -\left\{\mathbf{X}'_{iE} \boldsymbol{\beta} - \mathbf{X}'_E E_{m_2}(\mathbf{B}_{iI}^N)\right\},$$

is not necessarily equal to zero. However, the national estimator obtained adding those of the domains, $\tilde{Y}^{LCALN} = \sum_{i=1}^{m} \tilde{Y}_i^{LCALN} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})'\mathbf{B}_I$, is actually model-design unbiased, because

$$B_{m_2,\pi}(\tilde{Y}^{LCALN}) = -\{\mathbf{X}'_E\boldsymbol{\beta} - \mathbf{X}'_E E_{m_2}(\mathbf{B}_I)\} = 0.$$

Hence, under the model (16), the LCALN estimator is not model-design unbiased for a particular domain, but it is unbiased when aggregating for all the domains, provided that the same model holds for the included and excluded units in all domains. For the mean $\bar{Y}_i$, the bias of the estimator $\tilde{\bar{Y}}_i^{LCALN} = \tilde{Y}_i^{LCALN}/N_i$ is the same as that for the total Y_i, but dividing by N_i.

Let us now study the variances. For the theoretical LCAL estimator (10), the design-variance is given by

$$V_\pi(\tilde{Y}_i^{LCAL}) = V_\pi(\hat{Y}_i - \hat{\mathbf{X}}'_i\mathbf{B}_{iI}). \tag{17}$$

This variance can be easily estimated by expressing it as the variance of an expansion estimator, $V_\pi\left(\sum_{j\in s_i} w_{ij}\varepsilon_{ij}\right)$, for $\varepsilon_{ij} = y_{ij} - \mathbf{x}'_{ij}\mathbf{B}_{iI}$, $j \in U_{iI}$, and then applying the usual variance estimators for these expansion estimators. In the case of the LCALN estimator given in (14), the variance is given by

$$V_\pi(\tilde{Y}_i^{LCALN}) = V_\pi(\hat{Y}_i - \hat{\mathbf{X}}'\mathbf{B}_{iI}^N).$$

The contribution of $\hat{\mathbf{X}}$ to this variance is much smaller than the contribution of $\hat{\mathbf{X}}_i$ in (17), because $\hat{\mathbf{X}}$ is calculated with the n sample units, unlike $\hat{\mathbf{X}}_i$ which uses only the n_i units in domain i. This means that, provided that the domain and national regression lines are similar, the variance of the LCALN estimator, obtained from the calibration at the national level, will be much smaller than that of the LCAL estimator, based on the domain calibration.

5 EBLUP under the nested error model

The estimators described so far use mainly the information coming from the domain. When the domain sample size n_i is small, these estimators might be inefficient. Small area (or indirect) estimation methods are designed to reduce the variance by increasing the effective sample size, see the book by Rao & Molina (2015) for a comprehensive account of small area estimation methods. In this section, we focus on model-based methods, which provide estimators with good properties under the distribution induced by the model. However, here we are also interested in their design properties.

We consider a very popular unit level model introduced by Battese et al. (1988) and often called nested error model. Similary as model m_2 in (16), this model assumes a constant linear regression for all the population units, but allows for unexplained heterogeneity between the domains by including random domain effects u_i apart from model errors e_{ij}. This model, denoted model m_3, assumes

$$\begin{aligned} &y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + e_{ij},\ u_i \overset{iid}{\sim} N(0, \sigma_u^2), \\ &e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2),\ j = 1, \ldots, N_i,\ i = 1, \ldots, m, \end{aligned} \tag{18}$$

where area effects u_i and errors e_{ij} are assumed to be mutually independent. We will denote by $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ be the vector of unknown parameters. Note that setting $\sigma_u^2 = 0$, we obtain model m_2 given in (16).

Define the vector of random variables for domain i, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iN_i})'$, and the corresponding design matrix $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i})'$. Then, in matrix notation the model is

$$\mathbf{y}_i \overset{ind}{\sim} N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i),\ \mathbf{V}_i = \sigma_u^2 \mathbf{1}_{N_i}\mathbf{1}'_{N_i} + \sigma_e^2 \mathbf{I}_{N_i},\ i = 1, \ldots, m, \tag{19}$$

where $\mathbf{1}_k$ denotes a vector of ones of size k and $\mathbf{I}_k$ is the $k \times k$ identity matrix. Define also the population vector $\mathbf{y} = (\mathbf{y}'_1, \ldots, \mathbf{y}'_m)'$ and the matrix $\mathbf{X} = (\mathbf{X}'_1, \ldots, \mathbf{X}'_m)'$. Then, the model for the population vector is $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, for $\mathbf{V} = \sigma_u^2 \mathbf{Z}\mathbf{Z}' + \sigma_e^2 \mathbf{I}_N = \mathrm{diag}_{1 \le i \le m}(\mathbf{V}_i)$, where $\mathbf{Z} = \mathrm{diag}_{1 \le i \le m}(\mathbf{1}_{N_i})$.

In this section, we consider general population parameters defined as linear functions

of the vector $\mathbf{y}$, that is, we consider parameters of the type $H = \mathbf{b}'\mathbf{y}$, where $\mathbf{b}$ is a non-stochastic vector of known elements. Let us decompose the population vector $\mathbf{y}$ into the sample part $\mathbf{y}_s$, where the sample s is composed of the samples s_i drawn from the sets of included units in each area U_{iI}, and out-of-sample elements $\mathbf{y}_r$, and we decompose accordingly the design matrix $\mathbf{X}$ and the covariance matrix $\mathbf{V}$, that is,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{pmatrix}.$$

The linear parameter $H = \mathbf{b}'\mathbf{y}$ can be also decomposed as $H = \mathbf{b}_s'\mathbf{y}_s + \mathbf{b}_r'\mathbf{y}_r$. Under the model (18), the best linear unbiased predictor (BLUP) of H is the model-unbiased linear function of the sample data $\tilde{H} = \boldsymbol{\alpha}_s'\mathbf{y}_s$ that minimizes the model mean squared error (MSE), $\mathrm{MSE}_{m_3}(\tilde{H}) = E_{m_3}(\tilde{H} - H)^2$. The BLUP of $H = \mathbf{b}_s'\mathbf{y}_s + \mathbf{b}_r'\mathbf{y}_r$ is then given by

$$\hat{H}^{BLUP}(\boldsymbol{\theta}) = \mathbf{b}_s'\mathbf{y}_s + \mathbf{b}_r'[\mathbf{X}_r\tilde{\boldsymbol{\beta}}_s + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s'\tilde{\boldsymbol{\beta}}_s)], \tag{20}$$

where $\tilde{\boldsymbol{\beta}}_s$ is the weighted least squares estimator of $\boldsymbol{\beta}$, given by

$$\tilde{\boldsymbol{\beta}}_s = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s. \tag{21}$$

The BLUP of H given in (20) depends on the true values of the variance components σ_u^2 and σ_e^2, which are typically unknown. Replacing them by their respective consistent estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$, we obtain the so called empirical BLUP (EBLUP), and denoted here as $\hat{H}^{EBLUP}$.

For the special case of a domain mean $H = \bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$, the vector $\mathbf{b}$ is given by $\mathbf{b} = (\mathbf{0}_{N_1}', \ldots, \mathbf{0}_{N_{i-1}}', N_i^{-1}\mathbf{1}_{N_i}', \mathbf{0}_{N_i+1}', \ldots, \mathbf{0}_{N_m}')'$, where $\mathbf{0}_k$ denotes a vector of zeros of size k. If the domain sampling fraction, n_i/N_i, is negligible, the BLUP estimator of $\bar{Y}_i$ may be expressed as the weighted average

$$\hat{\bar{Y}}_i^{BLUP} \cong \gamma_{is}[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\tilde{\boldsymbol{\beta}}_s] + (1 - \gamma_{is})\bar{\mathbf{X}}_i'\tilde{\boldsymbol{\beta}}_s, \tag{22}$$

where $\gamma_{is} = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_i)$ (Rao & Molina, 2015). Thus, for domains with large sample size n_i, $\hat{\bar{Y}}_i^{BLUP}$ approaches the survey regression estimator $\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\tilde{\boldsymbol{\beta}}_s$, whereas for domains with small sample size n_i, $\hat{\bar{Y}}_i^{BLUP}$ borrows strength from the other domains by approaching the regression-synthetic estimator $\bar{\mathbf{X}}_i'\tilde{\boldsymbol{\beta}}_s$. Replacing σ_u^2 and σ_e^2 by consistent estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ in the BLUP, we obtain the EBLUP of $\bar{Y}_i$, given by

$$\hat{\bar{Y}}_i^{EBLUP} \cong \hat{\gamma}_{is}[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\hat{\boldsymbol{\beta}}_s] + (1 - \hat{\gamma}_{is})\bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_s, \tag{23}$$

where $\hat{\gamma}_{is} = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i)$ and $\hat{\boldsymbol{\beta}}_s$ is obtained by replacing respectively σ_u^2 and σ_e^2 by $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ in $\tilde{\boldsymbol{\beta}}_s$.

The BLUP is unbiased under model m_3 and optimal in the sense of minimizing the MSE under that model. Let us now study its design properties. For this, we consider the census regression parameter for the included units defined as $\mathbf{B}_I = (\mathbf{X}_I'\mathbf{V}_I^{-1}\mathbf{X}_I)^{-1}\mathbf{X}_I'\mathbf{V}_I^{-1}\mathbf{y}_I$, where $\mathbf{y}_I$, $\mathbf{X}_I$ and $\mathbf{V}_I$ are the corresponding sub-vector and sub-matrices of $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{V}$, for the included units in the population. Again, we consider the theoretical version of the BLUP defined in terms of $\mathbf{B}_I$,

$$\tilde{\bar{Y}}_i^{BLUP} = \gamma_{is}[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\mathbf{B}_I] + (1 - \gamma_{is})\bar{\mathbf{X}}_i'\mathbf{B}_I.$$

If each sample s_i is drawn from the corresponding U_{iI} by simple random sampling without replacement (SRSWOR), then $E_\pi(\bar{y}_{is}) = \bar{Y}_{iI}$ and $E_\pi(\bar{\mathbf{x}}_{is}) = \bar{\mathbf{X}}_{iI}$. Using these facts, it is easy to calculate the design-bias of $\tilde{\bar{Y}}_i^{BLUP}$, which is given by

$$B_\pi(\tilde{\bar{Y}}_i^{BLUP}) = \gamma_{is}\frac{N_{iE}}{N_{iI}}\left[(\bar{Y}_i - \bar{\mathbf{X}}_i'\mathbf{B}_I) - (\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}'\mathbf{B}_I)\right] + (1 - \gamma_{is})(\bar{\mathbf{X}}_i'\mathbf{B}_I - \bar{Y}_i).$$

This bias will be small if the same model (18) holds for the whole population or if the ratio of excluded over included individuals is small. Indeed, if model (18) holds for all the population units, then $E_{m_3}(\mathbf{B}_I) = \boldsymbol{\beta}$, $E_{m_3}(\bar{Y}_i) = \bar{\mathbf{X}}_i\boldsymbol{\beta}$ and $E_{m_3}(\bar{Y}_{iE}) = \bar{\mathbf{X}}_{iE}\boldsymbol{\beta}$. Using these results when taking expectation under the model m_3 in (24), we get $B_{m_3,\pi}(\tilde{\bar{Y}}_i^{BLUP}) = 0$. In fact, the same result holds also under model m_2.

Finally, if s_i is obtained by SRSWOR within U_{iI}, the design-variance of the theoretical

BLUP estimator is given by

$$V_\pi(\tilde{\bar{Y}}_i^{BLUP}) = \gamma_{is}^2 V_\pi(\bar{y}_{is} - \bar{\mathbf{x}}_{is}\mathbf{B}_I) = \frac{\gamma_{is}^2}{N_i^2} V_\pi(\hat{Y}_i - \hat{\mathbf{X}}_i'\mathbf{B}_I).$$

Hence, if the census LS regression lines for the domains from model (6) are similar to the national census WLS regression line from model (18), that is, if $\mathbf{B}_I \approx \mathbf{B}_{iI}$, then the variance of the EBLUP for $\bar{Y}_i$ decreases that of the LCAL estimator obtained from (17) by the factor γ_{is}^2.

6 EBP under the nested error model

For estimation of non-linear parameters, the BLUP has no meaning and we need to resort to methods dealing with more general parameters, such as the best/Bayes predictor (BP), see Molina & Rao (2010). Special non-linear parameters are poverty and inequality indicators defined in terms of a welfare measure such as the FGT family of poverty indicators due to Foster et al. (1984). The best predictor can also be used for estimation of other characteristics such as median, quantiles or even the empirical distribution function of the variable of interest, see Pratesi (2016). Additionally, it can be used for estimation of totals and means of a target variable, when the dependent variable in the model is a one-to-one transformation of this target variable (e.g. log or more general Box-Cox transformations). Those transformations are typically applied in cases of non normality or heteroscedasticity.

In this section, the target variable (e.g. the welfare measure) for the j-th unit in domain i is denoted as v_{ij} and $y_{ij} = T(v_{ij})$ is a one-to-one transformation of it. We consider that y_{ij} follows the nested error model (18). By the inverse transformation $v_{ij} = T^{-1}(y_{ij})$, we can express our target parameter (defined originally in terms of the target variables v_{ij}) as a function of the vector $\mathbf{y}$ of model responses y_{ij} for the population units, $H = h(\mathbf{y})$. The best predictor (BP) of $H = h(\mathbf{y})$ is defined as the function of the sample observations

$\mathbf{y}_s$ that minimizes the model MSE, and it turns out to be

$$\tilde{H}^{BP}(\boldsymbol{\theta}) = E_{m_3}[h(\mathbf{y})|\mathbf{y}_s; \boldsymbol{\theta}], \tag{24}$$

where the expectation is taken with respect to the model distribution of $\mathbf{y}_r|\mathbf{y}_s$, which depends on the true value of $\boldsymbol{\theta}$. $\tilde{H}^{BP}(\boldsymbol{\theta})$ is unbiased with respect to the model (18), regardless of the complexity of the function $h(\cdot)$ defining the target parameter. However, it cannot be calculated in practice since model parameters $\boldsymbol{\theta}$ are typically unknown. An empirical best predictor (EBP) of H, denoted as $\hat{H}^{EBP}$, is then obtained by replacing $\boldsymbol{\theta}$ in $\tilde{H}^{BP}(\boldsymbol{\theta})$ by a consistent estimator $\hat{\boldsymbol{\theta}}$ as $\hat{H}^{EBP} = \tilde{H}^{BP}(\hat{\boldsymbol{\theta}})$. The EBP 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 a linear parameter $H = \mathbf{b}'\mathbf{y}$, the EBP under the nested error model with normality, obtained using as estimator of $\boldsymbol{\beta}$ the WLS estimator in (21) equals the BLUP given in (20).

For some non-linear parameters where $h(\cdot)$ is too complex, the expectation defining the EBP in (24) cannot be calculated analytically; in those cases, $\hat{H}^{EBP}$ can be approximated by Monte Carlo as proposed in Molina & Rao (2010). This is done by simulating, from the model (18) fitted to the original sample, L replicates $y_{ij}^{(\ell)}; \ell = 1, \ldots, L$ of y_{ij}, $j \in r_i$, where r_i are the non-sample units of area i, attaching the sample elements y_{ij}, $j \in s_i$ to form the population vector $\mathbf{y}^{(\ell)}$, calculating the corresponding target parameter $H^{(\ell)} = h(\mathbf{y}^{(\ell)})$ for each $\ell = 1, \ldots, L$ and, finally, averaging over the L replicates as $\hat{H}^{EBP} = L^{-1} \sum_{\ell=1}^{L} H^{(\ell)}$. Note that the EBP requires the values $\mathbf{x}_{ij}$ for all units in the population, and not only for the included units. For further details on the calculation of EBP, see Molina & Rao (2010).

7 MSE estimation

The EBP of Section 6 or the EBLUP in Section 5 are based on the nested error model (18). Calibration estimators described in Section 4 are also assisted by linear regression models. If we wish to have comparable accuracy measures, it seems reasonable to obtain

the MSEs of all the estimators under a given regression model (model MSE), assuming that the model holds for all the population units (included and excluded). Here we estimate the model MSE using the bootstrap method proposed in Molina & Rao (2010), which is based on the original parametric bootstrap method for finite populations of González-Manteiga et al. (2008). According to this procedure, the bootstrap MSE of $\hat{H}^{EBP}$ under the nested error model (18) is obtained as follows: i) Fit the model (18) to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$, obtaining estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ 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 independently $u_i^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$ and $e_{ij}^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. iii) For $b = 1, \ldots, B$, construct a bootstrap population vector $\mathbf{y}^{*(b)}$ with elements $y_{ij}^{*(b)}$ generated as

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

From the bootstrap population vector $\mathbf{y}^{*(b)}$, calculate the target bootstrap parameter $H^{*(b)} = h(\mathbf{y}^{*(b)})$, for $b = 1, \ldots, B$. iv) From each bootstrap population vector $\mathbf{y}^{*(b)}$, take the sample part $\mathbf{y}_s^{*(b)}$, where s is the original sample composed of the sub-samples s_i from each domain $i = 1, \ldots, m$. Using also the population vectors $\mathbf{x}_{ij}$, $j = 1, \ldots, N_i$, assumed to be known for all population units, calculate the bootstrap EBP of H, denoted as $\hat{H}^{EBP*(b)}$, $b = 1, \ldots, B$. v) A bootstrap MSE estimator for the EBP under model (18), $\text{MSE}_{m_3}(\hat{H}^{EBP})$, is obtained as

$$\text{mse}_B(\hat{H}^{EBP}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{H}^{EBP*(b)} - H^{*(b)})^2. \tag{25}$$

Bootstrap estimators of the MSE under the same model of the calibration estimators can be obtained similarly. For the special case of a linear parameter, $H = \mathbf{b}'\mathbf{y}$, if $\hat{\boldsymbol{\beta}}$ is the WLS estimator (21), then (25) is an estimator of $\text{MSE}_{m_3}(\hat{H}^{EBLUP})$. This naïve bootstrap estimator of the model MSE is first-order unbiased in the sense that its model bias is $O(m^{-1})$, but not $o(m^{-1})$. Bias corrections existing in the literature increase the variance and may yield negative MSE estimates. In the literature, we cannot find bootstrap estimators of the MSE that are strictly positive and also second-order unbiased. Thus,

we consider the naive bootstrap estimator (25), which cannot yield negative estimators and performs well for moderate number of areas m.

8 Simulation experiments

8.1 Aims and general description

The purpose of these simulation experiments is to compare the performance of the considered calibration and model-based methods for estimation in small domains when the sample is drawn by cut-off sampling. Specifically, for the domain means $\bar{Y}_i$, $i = 1, \ldots, m$, we will compare the two calibration estimators $\hat{\bar{Y}}_i^{LCAL}$ and $\hat{\bar{Y}}_i^{LCALN}$, the naïve direct estimator $\hat{\bar{Y}}_i^{HA}$ and the EBLUP under the nested error model $\hat{\bar{Y}}_i^{EBLUP}$, under two different scenarios. In the first scenario, we consider that the values of the target variable for all units in the population are generated from the same model and, in the second, included and excluded units are generated from different models.

Calibration estimators are design-consistent as the domain size n_i increases even if the corresponding model does not hold, but this is not the case for model-based estimators. On the other hand, under the corresponding model, the EBLUP of a linear parameter is approximately the most efficient linear and unbiased estimator, so making simulations under a model would not provide any additional knowledge. The purpose here is to see whether the considered model-based estimators also perform well with respect to the (cut-off sampling) design. For this reason, we run design-based simulations by generating one population vector $\mathbf{y}$, keeping it fixed and repeatedly drawing cut-off samples. The population vector $\mathbf{y}$ is generated from the nested error model in (18). To allocate the units into the set of included and excluded units, we generate a random binary variable c_{ij} for each unit $j = 1, \ldots, N_i$ and $i = 1, \ldots, m$. The units j with $c_{ij} = 1$ are assigned to U_{iI} and those with $c_{ij} = 0$ to U_{iE}. In each Monte Carlo (MC) replicate, samples are drawn, independently for each domain i, from the U_{iI} units, $i = 1, \ldots, m$.

8.2 Common regression model

We consider a population of $N = 20,000$ individuals divided into $m = 80$ domains with the same size $N_i = 250$, $i = 1, \ldots, m$. We consider three auxiliary variables, with values generated as $x_{ij\kappa} \overset{iid}{\sim} N(3, 2)$, $\kappa = 1, 2, 3$. The binary variables determining the allocation of units in U_{iI} or U_{iE} for each domain i are generated independently as $c_{ij} \overset{ind}{\sim} \text{Bern}(p_{ij})$. The probabilities $p_{ij} = \Pr(c_{ij} = 1)$ are considered to be related to the vector of auxiliary variables $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})'$ through a logit model, that is,

$$p_{ij} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\zeta})}{1 + \exp(\mathbf{x}'_{ij}\boldsymbol{\zeta})}, \quad j = 1, \ldots, N_i, \ i = 1, \ldots, m,$$

taking $\boldsymbol{\zeta} = (0.75, 1, 1)'$. With these model parameters, the units in U_{iI}, that is, those with $c_{ij} = 1$, for all $i = 1, \ldots, m$, represent roughly half of the population.

The values of the target variable y_{ij} are generated from the nested error model (18) using $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})'$ and taking $\boldsymbol{\beta} = (1, 1.5, 1)'$, $\sigma_u^2 = (0.75)^2$ and $\sigma_e^2 = 4^2$, which leads to a determination coefficient $R^2 \approx 0.5$. Then, we draw $K = 1,000$ Monte Carlo samples $s^{(k)}$, $k = 1, \ldots, K$. Each of these samples is obtained by drawing independent domain sub-samples $s_i^{(k)}$ of size n_i from the units in U_{iI} by SRSWOR, $i = 1, \ldots, m$. The domain sample sizes are $n_i = 5$, $i = 1, \ldots, 20$, $n_i = 10$, $i = 21, \ldots, 40$, $n_i = 30$, $i = 41, \ldots, 60$ and $n_i = 50$, $i = 61, \ldots, 80$. With the data from the k-th sample, we compute the basic direct estimator, the estimators with calibration at the domain level (LCAL) and at the population level (LCALN) and the EBLUP estimators of $\bar{Y}_i$. Weights, h_{ij} and g_{ij}, in the calibration estimators (4) and (9) respectively are obtained using the function `calib` from package `sampling` (Tillé & Matei, 2016) of `R` (R development core team 2016). EBLUPs are obtained using `R` package `sae` (Molina & Marhuenda, 2015), which by default estimates the model parameters σ_u^2, σ_e^2 and $\boldsymbol{\beta}$ using the restricted maximum likelihood (REML) method.

Let $\hat{\bar{Y}}_i$ be a generic estimator of $\bar{Y}_i$ and $\hat{\bar{Y}}_i^{(k)}$ its value obtained with k-th sample. We evaluate the performance of estimators in terms of relative bias (RB) and relative root

MSE (RRMSE) under the design, approximated empirically as

$$\mathrm{RB}_\pi(\hat{\bar{Y}}_i) = 100\frac{K^{-1}\sum_{k=1}^{K}(\hat{\bar{Y}}_i^{(k)} - \bar{Y}_i)}{\bar{Y}_i}, \quad \mathrm{RRMSE}_\pi(\hat{\bar{Y}}_i) = 100\frac{\sqrt{K^{-1}\sum_{k=1}^{K}(\hat{\bar{Y}}_i^{(k)} - \bar{Y}_i)^2}}{\bar{Y}_i}.$$

Averages across domains of absolute RB and of RRMSE are also calculated as

$$\overline{\mathrm{ARB}} = m^{-1}\sum_{i=1}^{m}|\mathrm{RB}_\pi(\hat{\bar{Y}}_i)|, \quad \overline{\mathrm{RRMSE}} = m^{-1}\sum_{i=1}^{m}\mathrm{RRMSE}_\pi(\hat{\bar{Y}}_i).$$

Figure 8.1 displays the percent RB (left) and RRMSE (right) for the considered estimators of the mean $\bar{Y}_i$ for each domain i (x-axis). These two plots show large design bias and MSE for the basic direct estimators uniformly for all the domains. LCALN estimator shows large bias for some domains, probably because this estimator does not account for the domain effects. LCAL estimator performs globally the best in terms of design bias, because it does account for the domain effects. However, the two calibration estimators, but specially LCAL, obtain very large RRMSEs for the domains with the smallest sample sizes ($n_i \leq 20$). EBLUP exhibits the best results in terms of design RRMSE and at the same time keeping a small design bias. In fact, the difference between EBLUP and LCAL estimators in terms of bias is pretty small.

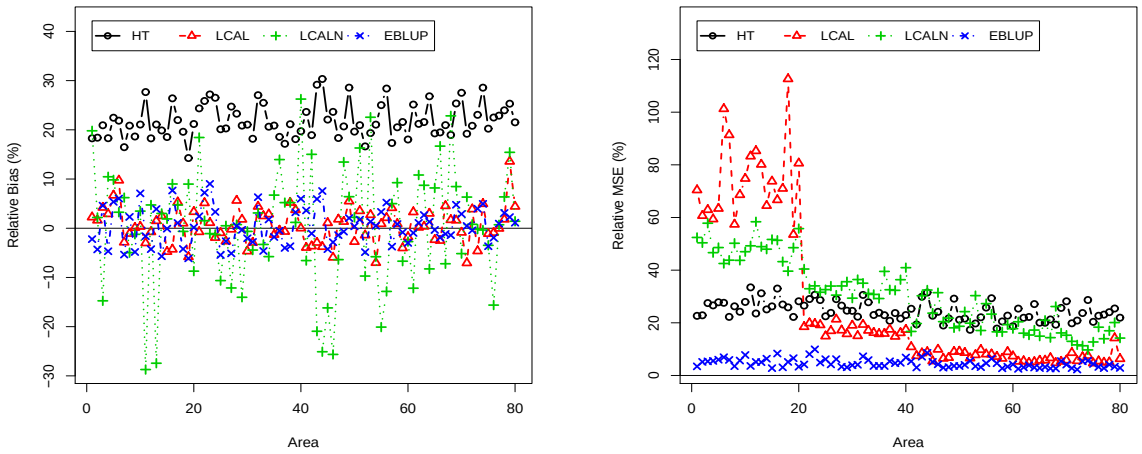

Figure 8.1: Percent RB (left) and RRMSE (right) of basic direct, LCAL, LCALN and EBLUP estimators of the domain means, for each area.

Table 8.1 displays $\overline{\text{ARB}}$, $\overline{\text{RRMSE}}$ and % share of squared bias from the total design MSE for the considered estimators. In this table, the basic direct estimator exhibits a large design-bias, with a bias share of $\text{B}_\pi^2/\text{MSE}_\pi \approx 100\%$, whereas the considered calibration estimators and EBLUP have a considerably smaller bias. LCAL estimator has the smallest average ARB. LCALN performs the best in terms of $\text{B}_\pi^2/\text{MSE}_\pi$ because of its large MSE. Thus, we consider that LCAL performs better. On the other hand, EBLUP clearly performs the best when accounting for both MSE and bias.

Table 8.1: Averages across areas of percent absolute RB and RRMSE, and average $\text{B}_\pi^2/\text{MSE}_\pi$ for basic direct, LCAL, LCALN and EBLUP (in percentage).

Method	$\overline{\text{ARB}}$	$\overline{\text{RRMSE}}$	$\text{B}_\pi^2/\text{MSE}_\pi$
DIR	21.82	24.45	98.32
LCAL	2.96	27.33	2.48
LCALN	8.97	30.44	0.04
EBLUP	3.13	4.56	0.18

8.3 Different regression models

In this simulation study, we preserve the same population values and sampling scheme as before. However, in this case, we consider that the values of the target variable for the included and excluded units are generated from different models. Of course, this is not a favorable scenario for the considered model-based estimators, but it may be realistic taking into account that the assumed model cannot be checked for the excluded units. Thus, instead of a constant $\boldsymbol{\beta}$ for all population units, we take $\boldsymbol{\beta}_I = (1, 1.5, 1)'$ for the included units and $\boldsymbol{\beta}_E = (0.5, 1.6, 0.5)'$ for the the excluded ones. The values of the explanatory variables, domain effects and error variances are generated the same as before, with σ_u^2 and σ_e^2. Again, we draw $K = 1,000$ samples $s^{(k)}$ from those units with $c_{ij} = 1$ for each domain i, by independent SRSWOR, with the same domain sample sizes n_i. With the sample data from the k-th sample, we compute basic direct, LCAL, LCALN and EBLUP estimates of $\bar{Y}_i$.

Figure 8.2 shows the corresponding results. In this case, all the estimators are biased, but the bias of the basic direct estimator becomes huge, exceeding 60% for some of the domains. The bias of LCAL and EBLUP is kept relatively small for all the domains, but

that of LCALN estimator is still very large in absolute value for some of the domains. In absence of cut-off sampling, the calibration estimators are asymptotically design-unbiased as the domain sample size n_i increases, even if the considered model does not hold. However, this is not true under cut-off sampling. Even under this unfavorable scenario, EBLUP shows a moderate bias and performs clearly the best in terms of RRMSE.

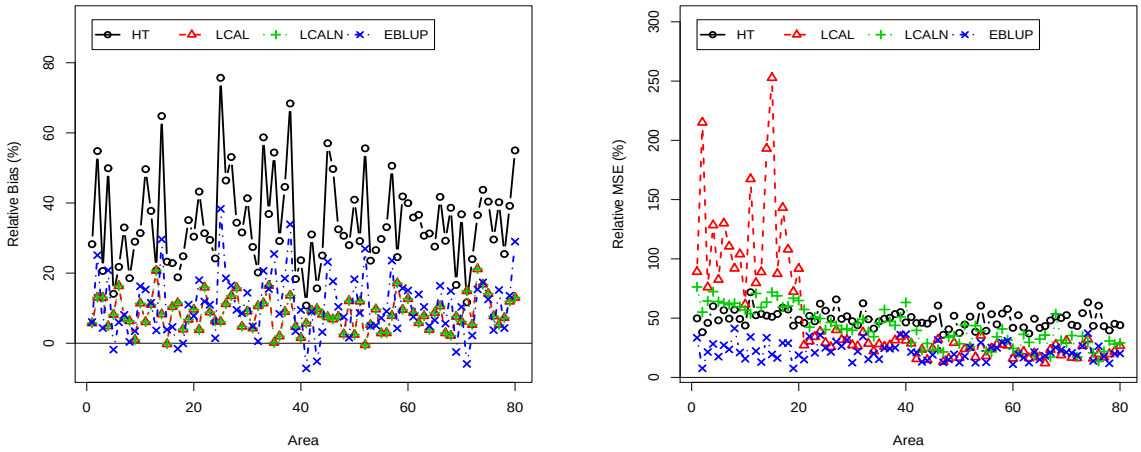

Figure 8.2: Percent RB (left) and RRMSE (right) of basic direct, LCAL, LCALN and EBLUP estimators of the domain means, under different models for included and excluded units.

Table 8.2 reports $\overline{\text{ARB}}$, $\overline{\text{RRMSE}}$, together with the % share of squared bias from the total design MSE. As already noted, the basic direct estimator exhibits a considerably large design bias compared to the other estimators, whereas LCAL and EBLUP estimators keep an $\overline{\text{ARB}}$ below 10%. LCALN displays the lowest $\text{B}_\pi^2/\text{MSE}_\pi$ because of a larger MSE. Again, EBLUP shows the best performance in terms of efficiency, with an average $\overline{\text{RRMSE}}$ also below 10%, while keeping a relatively small $\overline{\text{ARB}}$.

Table 8.2: Averages across areas of percent absolute RB and RRMSE and average $\text{B}_\pi^2/\text{MSE}_\pi$ for basic direct, LCAL, LCALN and EBLUP (in percentage), under different models for included and excluded units.

Method	$\overline{\text{ARB}}$	$\overline{\text{RRMSE}}$	$\text{B}_\pi^2/\text{MSE}_\pi$
DIR	31.78	34.11	99.87
LCAL	8.47	30.83	77.43
LCALN	12.75	34.49	29.56
EBLUP	8.73	9.48	75.78

The simulation experiment was repeated taking a value of $\boldsymbol{\beta}_E$ further away from

$\boldsymbol{\beta}_I$, making the two regression models differ substantially. Results are not included due to space constraints but, as one would expect, RB and RRMSE values increase for all estimators, but conclusions are similar to the last experiment. The basic direct estimator gets the largest RB, calibration estimators and EBLUP clearly reduce the bias of the basic direct estimator due to cut-off sampling and, for the domains with the smallest sample sizes, EBLUP gets the lowest RRMSE.

9 Estimation of total sales in Spanish provinces

Here we describe an application to the estimation of the total sales of a certain tobacco product in the Spanish provinces. The available data set contains, for $N = 12,791$ tobacco establishments (practically all of them) in $m = 48$ provinces from Spain (the Canary Islands, Ceuta and Melilla are not included), the volume of purchases made by each establishment of this product during the three months previous to November 2016 (z_{ij}, in euros). It also contains a variable indicating whether the establishment is supplied with a device recording all the required information about each sale. Only the establishments with larger sales are supplied with such a device. Those establishments (in total $n = 1,842$) are able to report proper data on sales and therefore the volume of sales (v_{ij}, in euros) of the considered product in November 2016 is also included in the data for those establishments.

We estimate the total sales $V_i = \sum_{i=1}^{N_i} v_{ij}$ in each of the $m = 48$ provinces included in the data using the basic direct, the selected calibration estimators and a model-based estimator. Establishments j with both z_{ij} and v_{ij} available for a province i compose the set of included units U_{iI}, which equals the sample s_i in this case (there is no sampling within U_{iI}). Then, here the basic direct estimators are given by

$$\hat{V}_i^{HA} = N_i \bar{V}_{iI}, \quad i = 1, \ldots, m, \tag{26}$$

which have actually zero variance, but might be severely biased. However, the bias cannot be estimated because there is no information from U_{iE}. Since true values in

real applications are not available and therefore real biases cannot be calculated, here we will compare the estimators considering the set of establishments with sales recorded from each province as a SRSWOR from that province. Note that this is the best scenario for the basic direct estimator. Thus, for the basic direct estimator given in (26), considering that the actual sample $s_i = U_{iI}$ is a SRSWOR from U_i, the variance equals the MSE (we ignore the bias). A design-unbiased estimator of the MSE is then

$$\text{mse}_\pi(\hat{V}_i) = N_i^2 \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right), \quad i = 1, \ldots, m,$$

where $s_i^2 = (n_i - 1)^{-1} \sum_{j \in s_i} (v_{ij} - \bar{v}_{is})^2$ is the sample variance of the sales for province i and here $n_i = N_{iI}$, $i = 1, \ldots, m$.

For the estimators that consider a regression model, we first make a preliminary descriptive analysis of the variables. Histograms of sales v_{ij} and of purchases z_{ij} show right-skewed distributions for both variables. Moreover, a scatterplot of ordinary LS residuals from a linear model for v_{ij} in terms of z_{ij}, against z_{ij} reveals a mild pattern of heteroscedasticity. Transforming the sales with the squared root, that is, taking $y_{ij} = v_{ij}^{1/2}$ as response variable and $\mathbf{x}_{ij} = (1, x_{ij})'$, with $x_{ij} = z_{ij}^{1/2}$ as covariate seems to minimize the problem. Accordingly, we will consider a nested error model (18) for the transformed sales y_{ij} in terms of the transformed purchases x_{ij}, and EBPs of the total sales in each province, $V_i = \sum_{j=1}^{N_i} v_{ij}$, will be computed based on this model. Note that, in terms of the model responses y_{ij}, the total sales are given by $V_i = \sum_{j=1}^{N_i} y_{ij}^2 = h(\mathbf{y}_i)$. Then, the EBP of $V_i = h(\mathbf{y}_i)$ is given by

$$\hat{V}_i^{EBP} = E_{m_3}[h(\mathbf{y}_i)|\mathbf{y}_{is}; \hat{\boldsymbol{\theta}}], \quad i = 1, \ldots, m,$$

which can be calculated analytically or approximated by Monte Carlo simulation. We estimate the model MSE of the EBP using the parametric bootstrap described in Section 7, taking $H^{*(b)} = V_i^{*(b)}$ and $\hat{H}^{EBP*(b)} = \hat{V}_i^{EBP*(b)}$ and considering that the assumed model holds for included and excluded units. Residuals from this model are described below.

Note that the LCAL (or GREG) estimator is not defined for a non-linear function

of the values of the response variable in the population units, such as the total sales $V_i = \sum_{j=1}^{N_i} y_{ij}^2$ after the square root transformation. Hence, here we calculate the GREG based on the linear model (6) for the untransformed sales v_{ij} in terms of purchases z_{ij} according to (5). As a measure of uncertainty of the GREG, to make it comparable with that of the EBP, we estimated its model MSE through the same bootstrap procedure, taking instead $\hat{H}^{EBP*(b)} = \hat{V}_i^{GREG*(b)}$. The obtained bootstrap MSE estimator actually includes the error due to the fact that the correct model is the one with transformed variables.

Before comparing the estimates, let us analyze the residuals from the nested error model (18), given by $\hat{e}_{ij} = y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} - \hat{u}_i$. Figure 9.1 shows a scatterplot of those residuals against predicted values $\hat{y}_{ij} = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \hat{u}_i$ (left) and a histogram of residuals (right). We can see a few negative outliers on the left plot, which agrees with a slightly larger left tail in the histogram. Apart from that, residuals do not exhibit any remarkable pattern. In fact, they appear to be very much concentrated around zero, which indicates a high predictive power of the model.

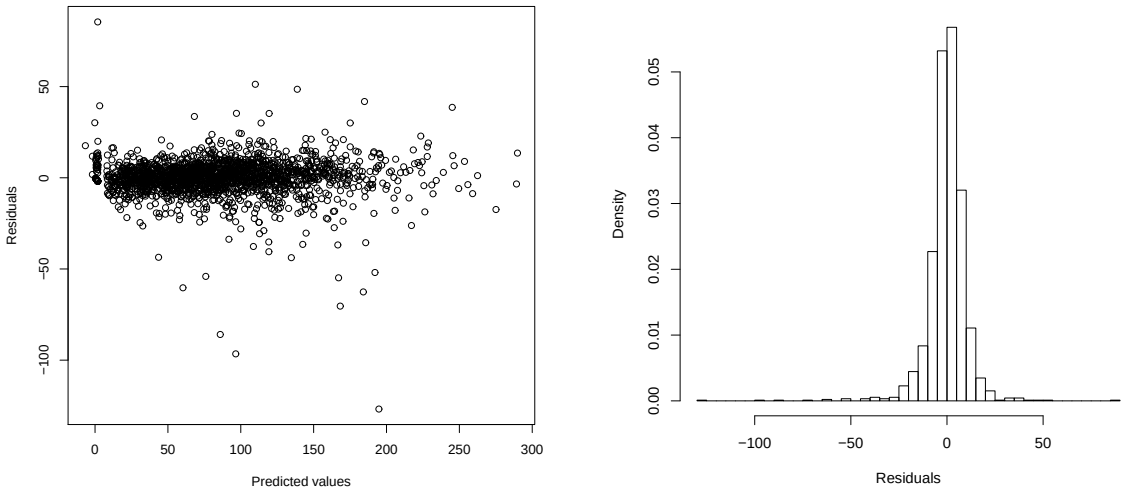

Figure 9.1: EBP residuals against predicted values (left), and histogram of EBP residuals (right).

Figure 9.2 shows the normal Q-Q plot of predicted area effects $\hat{u}_i$. This plot supports the normality of $\hat{u}_i$ except for one outlier appearing at the left tail of the distribution. This point corresponds to the province with the smallest sample size ($n_i = 3$ observations), which suggests that the estimated random effect for that province, $\hat{u}_i$, is not very reliable.

Thus, we consider that the nested error model fits reasonably well the available data.

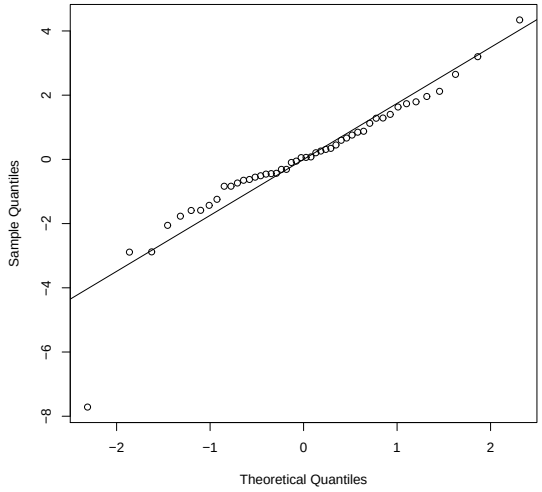

Figure 9.2: Normal Q-Q plot of predicted province effects $\hat{u}_i$.

We proceed now to compare the obtained estimates. Figure 9.3 left shows EBPs of the total sales of the considered tobacco product for each province against direct estimates. Province sample sizes are used as point labels. This plot indicates that the two types of estimates are very similar for the provinces with small sample sizes. However, for the two provinces with the largest sample sizes, the EBPs are slightly larger than the corresponding direct estimates, which could be due to bias of the direct estimator. Figure 9.3 right displays EBPs against GREG estimates. The great similarity of GREG and EBP estimates shown by this plot supports the fact that direct estimators might be the ones that actually underestimate the total sales in this application.

Finally, we compare the three types of estimates of the total sales for each province in Figure 9.4 left, showing the point estimates for each province (x axis), with provinces sorted from smaller to larger sample sizes, and with sample sizes indicated in the x-axis labels. The conclusions are the same as before; that is, the three types of estimates take very similar values for all provinces except for a couple of provinces with the larger sample sizes, where the basic direct estimator takes slightly smaller values (possibly understating the total sales). Figure 9.4 (right) shows the estimated coefficients of variation (CV) obtained ignoring the bias due to cut-off sampling. EBP estimators perform uniformly better than the other estimators in terms of estimated CV, keeping the CV values below

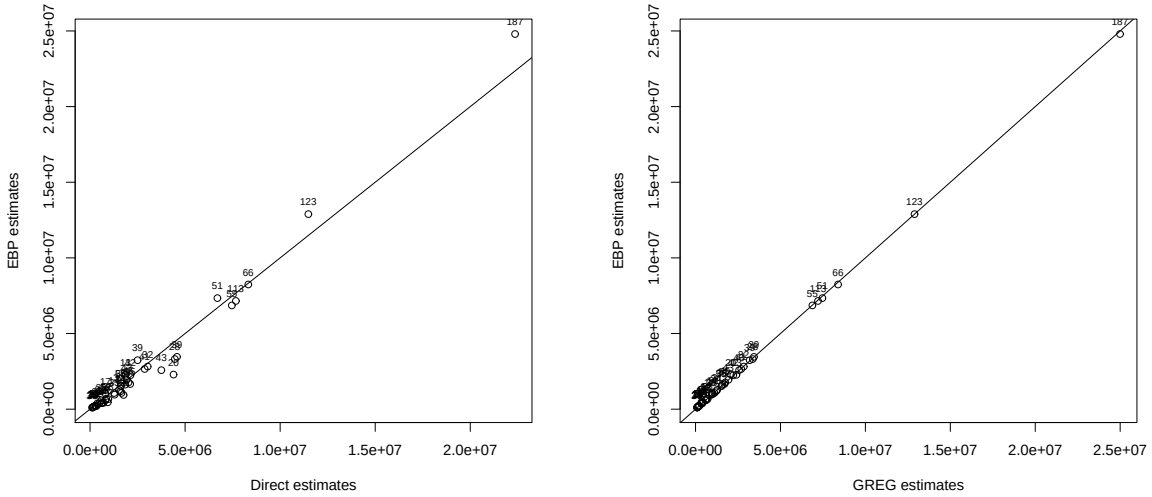

Figure 9.3: EBPs of total sales for each province against direct estimates (left) and against GREG estimates (right).

10% for practically all provinces, whereas GREG estimator obtains CV values above 10% for the provinces with the smallest sample sizes. We can see some peaks in the estimated CVs for some provinces with not necessarily the smallest sample sizes. These larger CV values are due to the presence of zero purchases and sales of the considered product in many tobacco shops for those particular provinces (that particular product is not acquired every month). Finally, it is clear that the direct estimator performs clearly the worst in terms of efficiency.

Table 10 in Appendix 10 lists the resulting direct, calibration and EBP estimates of the total sales of the product for each province accompanied with their estimated CVs. This table confirms the better performance of EBP in terms of estimated CV under the nested error model, specially for those provinces with small sample sizes. Finally, the direct estimator performs poorly in terms of CV even if the bias due to cut-off sampling is not accounted for.

10 Conclusions

Cut-off sampling is frequently used in business surveys, when drawing a representative sample from the whole population entails a cost that does not really compensate the

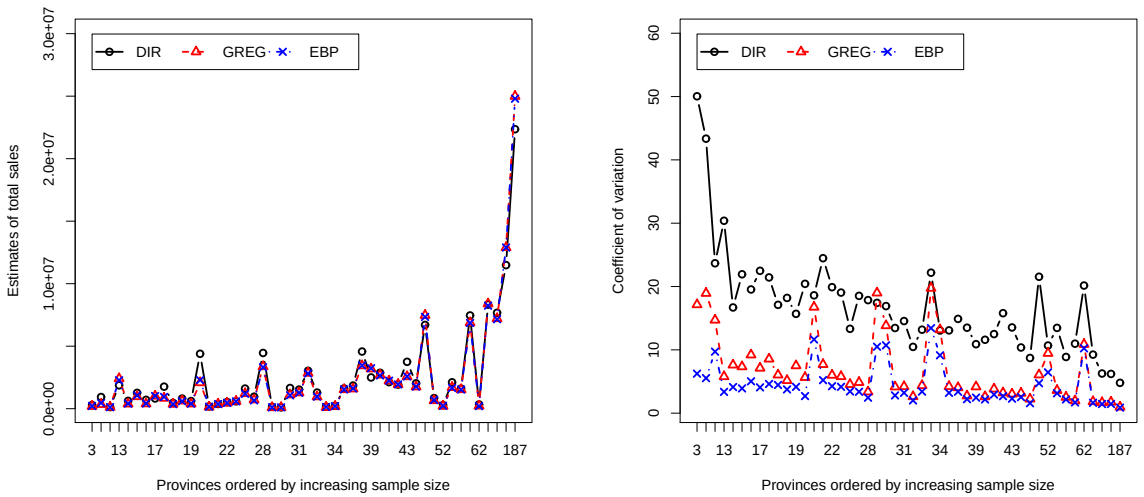

Figure 9.4: Direct, calibration and EBP estimates of total sales for each province (left) and corresponding estimated coefficients of variation (right).

subsequent gain in accuracy. On the other hand, in some surveys, part of the target population may not be actually available for sampling; that is, there may be population sectors that cannot be represented in the sample. These situations appear more often than desired, providing biased direct estimates as we have seen along this work.

We have studied the theoretical design properties of basic direct, calibration and model-based estimators under cut-off sampling in small areas. Our results show that EBLUP for a linear parameter, similarly as calibration estimators, reduce considerably the bias due to cut-off sampling if the models for the included and excluded individuals are reasonably similar. In terms of MSE, EBLUP performs significantly better than calibration estimators for domains with small sample size.

In our simulation studies and in the application, we compared the proposed methods by assuming that the model is the same for all units in the population (included or excluded). The model assumption could be arguable because there is no way of checking the model for the excluded units. In the case that estimation for the overall domain (and not only for U_{iI}) is required as is the case in this work, one will need to rely on subjective prior information concerning the validity of the assumed model for the excluded units. In any case, estimates can be considered just as indicatives of what could be the true values in the case that the same model holds for all the domain units. In fact, the case

of different models for included and excluded units was also analyzed in simulations. In this case, model-based estimators remained to be the most efficient, with not much larger bias than that of calibration estimators.

Finally, MSEs of calibration and model-based estimators are obtained under the model, whereas for the direct estimator we have considered the design MSE. Design MSEs are preferred by National Statistical Institutes because they do not assume that a model is correct and therefore account for model failures. There is ongoing research on finding stable design MSE estimates of model-based small area estimators, see Strzalkowska & Molina (2017). We plan to use their ideas to find design MSE estimators of the considered small area estimators in the context of cut-off sampling.

Acknowledgements

The work of M. Guadarrama and I. Molina is supported by the Spanish Ministry of Economy and Competitiveness [Ministerio de Economía y Competitividad, grants MTM2015-69638-R (MINECO/FEDER, UE) and MTM2015-72907-EXP].

Appendix: Estimates of total sales by provinces

Table .1: Basic direct, GREG and EBP estimates of total sales for the selected product and estimated coefficients of variation (%) for each Spanish province (by increasing sample size).

PROVINCE	n_i	$\hat{V}_i^{HA}$	$\hat{V}_i^{GREG}$	$\hat{V}_i^{EBP}$	$cv(\hat{V}_i^{HA})$	$cv(\hat{V}_i^{GREG})$	$cv(\hat{V}_i^{EBP})$
SORIA	3	293020.0	187824.9	213325.0	50.0	17.1	6.2
ZAMORA	7	932520.0	345095.8	454657.0	43.3	18.9	5.5
ALAVA	11	130083.6	119918.5	118835.3	23.7	14.7	9.7
ALMERIA	13	1870104.6	2407333.1	2272051.4	30.4	5.8	3.4
PALENCIA	14	626340.0	380367.4	409775.4	16.7	7.6	4.1
SALAMANCA	14	1265580.0	966094.1	1068230.6	21.9	7.3	3.9
AVILA	15	708696.0	392474.1	418917.2	19.5	9.2	5.0
LERIDA	17	817817.6	1011032.3	1014770.2	22.5	7.1	4.1
CIUDAD REAL	18	1764000.0	841228.2	939994.9	21.4	8.6	4.6
GUADALAJARA	18	463047.8	362148.3	363856.9	17.1	6.0	4.5
RIOJA	18	809900.0	622488.3	595178.6	18.2	5.2	3.7
SEGOVIA	19	610370.5	386734.4	402324.0	15.7	7.5	4.2
CACERES	20	4391826.0	2081619.7	2286462.0	20.4	5.6	2.7
GUIPUZCOA	20	181634.0	136700.0	156311.8	18.6	16.7	11.6

Continued on next page

Table .1 – *Continued from previous page*

PROVINCE	n_i	$\hat{Y}_i^{DIR}$	$\hat{Y}_i^{GREG}$	$\hat{Y}_i^{EB}$	$cv(\hat{Y}_i^{DIR})$	$cv(\hat{Y}_i^{GREG})$	$cv(\hat{Y}_i^{EB})$
HUESCA	22	377954.5	372101.3	371246.5	24.5	7.7	5.2
TERUEL	22	534417.3	446565.7	465643.3	19.9	6.0	4.3
CUENCA	23	588464.3	587005.5	586347.5	19.0	5.8	4.2
VALLADOLID	24	1609875.0	1210132.8	1188336.1	13.3	4.5	3.4
BURGOS	28	961645.7	708510.0	666698.1	18.5	4.9	3.4
CORDOBA	28	4457614.3	3367169.5	3312801.5	17.9	3.4	2.4
ORENSE	28	148577.1	88104.6	108428.9	17.4	19.0	10.5
LUGO	30	107213.3	92938.7	104233.7	16.9	13.8	10.7
ALBACETE	31	1654606.5	1115182.2	1073719.8	13.4	4.2	2.8
LEON	31	1528254.2	1274531.6	1270341.6	14.5	4.2	3.2
HUELVA	32	3031328.1	2838874.0	2816281.3	10.5	2.6	2.0
NAVARRA	33	1291343.0	956737.9	957660.4	13.2	4.4	3.4
PONTEVEDRA	33	159229.1	107198.9	138367.4	22.2	19.7	13.4
VIZCAYA	34	228618.8	183267.3	206304.6	13.1	13.2	9.1
TOLEDO	35	1619939.4	1529104.8	1539799.3	13.1	4.2	3.2
CADIZ	38	1851521.1	1585755.9	1620844.2	14.9	4.0	3.4
BADAJOZ	39	4571743.6	3439625.5	3457692.5	13.5	2.7	2.2
MALAGA	39	2499392.3	3188031.1	3237081.8	10.9	4.2	2.5
TARRAGONA	41	2872882.0	2690969.7	2656117.8	11.6	2.6	2.2
GRANADA	42	2123693.3	2221155.1	2241916.2	12.5	3.8	2.9
JAEN	43	1928229.8	1940379.2	1943101.0	15.8	3.2	2.7
ZARAGOZA	43	3750210.7	2564909.0	2578011.3	13.5	3.0	2.3
GERONA	45	2029222.2	1748165.7	1767490.3	10.4	3.2	2.5
MURCIA	51	6700070.6	7467465.0	7341434.6	8.7	2.2	1.6
BALEARES	52	849950.8	650012.6	694416.3	21.5	6.1	4.7
CANTABRIA	52	285632.3	204947.7	226163.1	10.7	9.5	6.4
ASTURIAS	55	2113034.5	1702020.8	1661932.8	13.5	3.6	3.1
CASTELLON	55	1605604.4	1526618.1	1530394.2	8.9	2.5	2.2
SEVILLA	55	7458078.2	6878368.2	6857368.8	11.0	2.0	1.7
CORUNA	62	340200.0	217028.5	206041.8	20.2	10.9	10.2
ALICANTE	66	8324589.1	8390895.3	8240996.9	9.2	1.8	1.6
VALENCIA	113	7671137.7	7209128.2	7153290.2	6.3	1.7	1.4
MADRID	123	11483342.8	12892853.8	12892305.0	6.2	1.7	1.5
BARCELONA	187	22356500.5	24990558.9	24797372.9	4.8	1.0	0.9

References

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

Benedetti, R., Bee, M. & Espa, G. (2010). A framework for cut-off sampling in business survey design. *Journal of Official Statistics* **26**, 651–671.

Foster, J., Greer, J. & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society* , 761–766.

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

Hájek, J. (1971). Comment on a paper by D. Basu. *Foundations of statistical inference* **236**.

Haziza, D., Chauvet, G. & Deville, J.-C. (2010). Sampling estimation in presence of cut-off sampling. *Australian & New Zealand Journal of Statistics* **52**, 303–319.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* **47**, 663–685.

INE (2018). Índices de producción industrial (ipi) base 2015. Tech. rep., Instituto Nacional de Estadística, España.

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

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

Pratesi, M. (2016). *Analysis of poverty data by small area estimation.* Hoboken, New Jersey: John Wiley & Sons.

Rao, J. N. K. & Molina, I. (2015). *Small area estimation.* Hoboken, New Jersey: John Wiley & Sons.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Strzalkowska, E. & Molina, I. (2017). Estimation of proportions in small areas: application to the labor force using the Swiss Census Structural survey. *Unpublished work* .

Tillé, Y. & Matei, A. (2016). *sampling: Survey Sampling*. R package version 2.8.

LUXEMBOURG INSTITUTE OF
SOCIO-ECONOMIC RESEARCH

11, Porte des Sciences
Campus Belval
L-4366 Esch-sur-Alzette

T +352 58 58 55-1
F +352 58 58 55-700

www.liser.lu