Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

# Small area estimation methods under cut-off sampling


by María Guadarrama, Isabel Molina and Yves Tillé

Release date: June 30, 2020

Statistics Canada Statistique Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

**Email at** STATCAN.infostats-infostats.STATCAN@canada.ca

**Telephone,** from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

**Depository Services Program**

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada



**An HTML version is also available.**

*Cette publication est aussi disponible en français*.

Survey Methodology, June 2020
Vol. 46, No. 1, pp. 51-75
Statistics Canada, Catalogue No. 12-001-X


# Small area estimation methods under cut-off sampling

**María Guadarrama, Isabel Molina and Yves Tillé**[1]

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 may be severely biased. Calibration estimators have been proposed to reduce the design-bias. However, when estimating in small domains, they can be inefficient even in the absence of cut-off sampling. Model-based small area estimation methods may prove useful for reducing the bias due to cut-off sampling if the assumed model holds for the whole population. At the same time, for small domains, these methods provide more efficient estimators than calibration methods. Since model-based properties are obtained assuming that the model holds but no model is exactly true, here we analyze the design properties of calibration and model-based procedures for estimation of small domain characteristics under cut-off sampling. Our results confirm that model-based estimators reduce the bias due to cut-off sampling and perform significantly better in terms of design mean squared error.

**Key Words:** Calibration estimators; Cut-off sampling; Empirical best linear unbiased predictor (EBLUP); Empirical best/Bayes predictor (EBP); Nested-error model; Unit level models.

## 1 Introduction

Haziza, Chauvet and Deville (2010) describe cut-off sampling as a technique in which a set of units is deliberately excluded from possible selection in the sample. For the Organisation for Economic Co-operation and Development (OECD), it is a sampling procedure in which a threshold is established such that all units above or below the threshold are excluded from selection in a sample. According to Särndal, Swensson and Wretman (1992, pages 531-533), this sampling technique is typically used when the distribution of the study variable is highly skewed and there is no reliable frame covering the small elements. Benedetti, Bee and Espa (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 keeping a reliable frame for the whole population does not compensate the subsequent gain in accuracy.

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 firms that 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 one form of cut-off sampling. Since the inclusion probabilities for the excluded units are zero, this procedure leads to biased design-based estimators, see e.g., Särndal et al. (1992) or Haziza et al. (2010) among others. To reduce the cut-off sampling bias, Haziza et al. (2010) propose to use

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

auxiliary information either at the design or at the estimation stage; 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), even in absence of cut-off sampling, calibration estimators might be inefficient. To improve efficiency, we consider small area estimation methods. For estimation of linear parameters, we consider the empirical best linear unbiased predictor (EBLUP) and, for general non-linear parameters, we consider 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 Spanish provinces.

In the absence of cut-off sampling, the considered model-based estimators are approximately optimal when the model holds for all the population units. However, since no model holds exactly, we wish to study whether model-based estimators still perform better than basic design-based estimators (which do not depend on models) and calibration estimators under the sampling replication mechanism; i.e., without model assumptions and when cut-off sampling is present.

The article is organized as follows. Section 2 describes the theoretical set-up. 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 estimating the mean squared error of the proposed small area estimators. Section 8 compares, through simulation experiments, the performance of several 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 domains

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 $N = \sum_{i=1}^{m} N_i$. We restrict ourselves to the case in which the domains act as sampling strata. Then, 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 a subset of units $U_{iE} \subseteq U_i$ from the selection. In other words, the domain $U_i$ is partitioned into two subsets, $U_{iI}$ and $U_{iE}$, of known sizes $N_{iI}$ and $N_{iE}$ respectively, with $N_i = N_{iI} + N_{iE}$. The set $U_{iI}$ contains the units that can be potentially selected for the sample, called here the set of included units, whereas $U_{iE}$ contains the excluded units.

Let $y_{ij}$ be the value of the target variable $y$ for the $j^{\text{th}}$ unit within the $i^{\text{th}}$ domain. We focus on estimation of domain totals $Y_i = \sum_{i=1}^{N_i} y_{ij}$ or means $\bar{Y}_i = Y_i / N_i$, $i = 1, \ldots, m$. 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 $\pi_{j|i} = \Pr(j \in s_i) > 0$ and $w_{j|i} = \pi_{j|i}^{-1}$ are the corresponding sampling weights. For the excluded units

$(j \in U_{iE})$, the inclusion probabilities are zero and, therefore, the corresponding sampling weights are not defined. As a consequence, for domains $i$ with $U_{iE} \neq \varnothing$, basic design-based estimators of $Y_i$ or $\bar{Y}_i$ are biased and a design-unbiased estimator does 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 the 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. In addition, as we shall see below, under cut-off sampling, their design-bias might be substantial.

The usual expansion estimator (Horvitz and Thompson, 1952) of $Y_i$ obtained ignoring that the sample $s_i$ is drawn only from $U_{iI}$ is given by $\hat{Y}_i = \sum_{j \in s_i} w_{ij} y_{ij}$. Under cut-off sampling, $\hat{Y}_i$ 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 $E_\pi$ denotes expectation under repeated sampling, since the sampling weights $w_{j|i} = \pi_{j|i}^{-1}$ in $\hat{Y}_i$ expand to $U_{iI}$ instead of $U_i$. No one would use this estimator since its bias, $B_\pi(\hat{Y}_i) = E_\pi(\hat{Y}_i) - Y_i = -Y_{iE}$, given in relative terms by the proportion of the total represented by the excluded population, $\mathrm{RB}_\pi(\hat{Y}_i) = -Y_{iE}/Y_i$, can be substantial.

When auxiliary information is not available, it makes more sense to use the Hájek estimator (Hájek, 1971) for the mean $\bar{Y}_i$, given by $\hat{\bar{Y}}_i^{\mathrm{HA}} = \hat{Y}_i / \hat{N}_i$, where $\hat{N}_i = \sum_{j \in s_i} w_{ij}$. The corresponding estimator for the total is $\hat{Y}_i^{\mathrm{HA}} = N_i \hat{\bar{Y}}_i^{\mathrm{HA}}$, 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 asymptotic (as $n_i \to \infty$) design-bias of $\hat{Y}_i^{\mathrm{HA}}$ is given in absolute and relative terms by

$$B_\pi(\hat{Y}_i^{\mathrm{HA}}) \cong N_{iE}(\bar{Y}_{iI} - \bar{Y}_{iE}), \qquad \mathrm{RB}_\pi(\hat{Y}_i^{\mathrm{HA}}) \cong \frac{N_{iE}}{N_i} \frac{\bar{Y}_{iI} - \bar{Y}_{iE}}{\bar{Y}_i}, \tag{3.1}$$

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 the mean, the bias of $\hat{\bar{Y}}_i^{\mathrm{HA}}$ is obtained dividing by $N_i$ in (3.1). For a domain $i$ with $U_{iE} \neq \varnothing$, the above bias vanishes only when $\bar{Y}_{iI} = \bar{Y}_{iE}$, 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 a mean of reducing the cut-off sampling bias.

**Remark 3.1.** The Hájek estimator of $\bar{Y}_i$ is a special case of the customary ratio estimator. In many monthly business surveys, parameters of interest are actually the changes over time of certain totals, such as $\theta_{it} = Y_i(t)/Y_i(t-1)$, where $Y_i(t)$ is the total of the target variable at time $t$ within domain $i$. The ratio estimates of change are actually reported instead of the actual totals because it is often believed that such ratios are not affected by cut-off sampling bias. Let $\hat{\theta}_{it} = \hat{Y}_i(t)/\hat{Y}_i(t-1)$ be the basic direct estimator of $\theta_{it}$. As we have seen above, the bias of the ratio estimator due to cut-off sampling tends to be much smaller than that of the absolute totals $\hat{Y}_i(t)$ and $\hat{Y}_i(t-1)$. However, as we have also seen, the

cut-off sampling bias of ratio estimators vanishes only under strong assumptions. Indeed, ignoring the ratio bias, which is negligible for large $n_i$, the bias of $\hat{\theta}_{it}$ is given by

$$B_\pi\left(\hat{\theta}_{it}\right) \cong \frac{Y_{iI}(t)}{Y_{iI}(t-1)} - \frac{Y_i(t)}{Y_i(t-1)},$$

where $Y_{iI}(t)$ denotes the corresponding total for the included units only. This bias is zero only if the ratios for the population $Y_i(t)/Y_i(t-1)$ are the same as those for the included units $Y_{iI}(t)/Y_{iI}(t-1)$.

# 4 Calibration estimators

Calibration is traditionally applied when the true totals of certain auxiliary variables, which are potentially correlated with the study variable, are known. The idea of calibration is to adjust the design weights $w_{j|i}$, so that the corresponding expansion estimators of the available true totals have zero error. 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, in the absence of cut-off sampling, calibration estimators are design-consistent as the area sample size $n_i$ increases even if the model does not hold. In this sense, they are model-assisted and their properties are typically evaluated under the design-based setup. However, if $n_i$ is small, the estimates may suffer from small sample bias.

As we shall see below, calibration estimators reduce the bias due to cut-off sampling if the underlying linear model holds for the whole population (included and excluded units). However, for small domains, they might have unacceptably large sampling errors, apart from non-negligible small sample bias.

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 the case whereby the vector of domain totals $\mathbf{X}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is available. Note that $\mathbf{X}_i$ is the total in the whole domain $U_i = U_{iI} \cup U_{iE}$. Then, one approach to calibration is to determine calibration weights $h_{j|i}$, $j \in s_i$, that minimize

$$\sum_{j\in s_i}\left(h_{j|i} - w_{j|i}\right)^2 \Big/ w_{j|i} \tag{4.1}$$

$$\text{s.t.} \sum_{j\in s_i} h_{j|i}\mathbf{x}_{ij} = \mathbf{X}_i.$$

The resulting calibration weights $h_{j|i}$ are given by

$$h_{j|i} = w_{j|i}\left\{1 + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)'\left(\sum_{j\in s_i} w_{j|i}\mathbf{x}_{ij}\mathbf{x}'_{ij}\right)^{-1}\mathbf{x}_{ij}\right\},\ j \in s_i, \tag{4.2}$$

provided that $\sum_{j\in s_i} w_{j|i}\mathbf{x}_{ij}\mathbf{x}'_{ij}$ is non-singular. The calibration estimator of the domain total $Y_i$ is then given by

$$\hat{Y}_i^{\mathrm{LCAL}} = \sum_{j \in s_i} h_{j|i} y_{ij} = \hat{Y}_i + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)' \hat{\mathbf{B}}_i, \tag{4.3}$$

which is the well-known generalized regression (GREG) estimator of $Y_i$, where

$$\hat{\mathbf{B}}_i = \left(\sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}_{ij}'\right)^{-1} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} y_{ij}.$$

The Hájek estimator $\hat{Y}_i^{\mathrm{HA}}$ is a special case of (4.3), with $\mathbf{x}_{ij} = 1$, $j = 1, \ldots, N_i$. In the absence of cut-off sampling, the above GREG estimator is design-consistent as the domain sample size $n_i$ increases, although it may suffer from small sample bias. It reduces the variance if the calibration variables are linearly correlated with the outcome and the correlation is strong. Under cut-off sampling, the second term on the right-hand side of (4.3) corrects for the bias of the basic expansion estimator $\hat{Y}_i$ as estimator of $Y_i$ with the help of the known domain totals in $\mathbf{X}_i$. However, for small domain sample size $n_i$, this reduction in cut-off sampling bias might be transferred to an increase in variance.

In the above procedure, we have a different calibration problem for each domain. In the case that only the overall population total $\mathbf{X} = \sum_{i=1}^{m} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is available, we may seek calibration weights for all the domains at once, $g_{j|i}$, $j \in s_i$, $i = 1, \ldots, m$, by solving only one calibration problem:

$$\min_{\{g_{j|i}:\, j \in s_i,\, i = 1, \ldots, m\}} \sum_{i=1}^{m} \sum_{j \in s_i} \left(g_{j|i} - w_{j|i}\right)^2 \Big/ w_{j|i} \tag{4.4}$$

$$\text{s.t.} \sum_{i=1}^{m} \sum_{j \in s_i} g_{j|i} \mathbf{x}_{ij} = \mathbf{X}.$$

In this case, the calibration weights $g_{j|i}$ are given by

$$g_{j|i} = w_{j|i} \left\{1 + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \left(\sum_{i=1}^{m} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}_{ij}'\right)^{-1} \mathbf{x}_{ij}\right\}, \quad j \in s_i,\ i = 1, \ldots, m, \tag{4.5}$$

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

$$\hat{Y}_i^{\mathrm{LCALN}} = \sum_{j \in s_i} g_{j|i} y_{ij} = \hat{Y}_i + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \hat{\mathbf{B}}_i^N, \tag{4.6}$$

where

$$\hat{\mathbf{B}}_i^N = \left(\sum_{\ell=1}^{m} \sum_{j \in s_\ell} w_{j|\ell} \mathbf{x}_{\ell j} \mathbf{x}_{\ell j}'\right)^{-1} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} y_{ij}.$$

In contrast with the GREG estimator, the correction of $\hat{Y}_i$ in $\hat{Y}_i^{\mathrm{LCALN}}$ uses the overall population total $\mathbf{X}$ and its corresponding expansion estimator.

The LCAL (or GREG) estimator (4.3) is expected to have smaller cut-off sampling bias than (4.6) because it uses auxiliary information from each particular domain $i$. On the other hand, for domains with

small sample sizes $n_i$, its variance (and small sample bias) may be large since it uses only domain-specific data. The alternative calibration estimator given in (4.6) is expected to have slightly larger cut-off sampling bias because it uses only aggregated auxiliary information at the national level, but its design-variance is expected to be smaller. We now study the properties of (4.3). To this end, consider the theoretical version of LCAL estimator (4.3), given by

$$\tilde{Y}_i^{\mathrm{LCAL}} = \hat{Y}_i + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)' \mathbf{B}_{iI}. \tag{4.7}$$

Here, $\mathbf{B}_{iI} = \left(\sum_{j\in U_{iI}} \mathbf{x}_{ij}\mathbf{x}_{ij}'\right)^{-1} \sum_{j\in U_{iI}} \mathbf{x}_{ij} y_{ij}$ is the census version of $\hat{\mathbf{B}}_i$ based on the set of included units from domain $i$. Note that the sample $s_i$ is drawn only from $U_{iI}$ and thus $\hat{\mathbf{B}}_i$ estimates $\mathbf{B}_{iI}$. We decompose the bias of $\hat{Y}_i^{\mathrm{LCAL}}$ as

$$\begin{aligned} B_\pi\left(\hat{Y}_i^{\mathrm{LCAL}}\right) &= E_\pi\left(\hat{Y}_i^{\mathrm{LCAL}} - \tilde{Y}_i^{\mathrm{LCAL}}\right) + B_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right), \\ &= E_\pi\left\{\left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)'\left(\hat{\mathbf{B}}_i - \mathbf{B}_{iI}\right)\right\} + B_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right). \end{aligned} \tag{4.8}$$

The term $E_\pi\left\{\left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)\left(\hat{\mathbf{B}}_i - \mathbf{B}_{iI}\right)\right\}/N_i$ tends to zero as $n_i \to \infty$ regardless of whether cut-off sampling is applied or not, since $\hat{\mathbf{B}}_i$ tends to $\mathbf{B}_{iI}$. However, for small $n_i$ this term may not be negligible; that is, the LCAL estimator has small sample bias even if $U_{iE} = \varnothing$. In the absence of cut-off sampling, the bias term $B_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right)$ in (4.8) is exactly equal to zero. Under cut-off sampling, we know that $E_\pi\left(\hat{Y}_i\right) = Y_{iI}$ and $E_\pi\left(\hat{\mathbf{X}}_i\right) = \mathbf{X}_{iI}$, where $\mathbf{X}_{iI} = \sum_{j\in U_{iI}} \mathbf{x}_{ij}$. 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 in absolute and relative terms by

$$B_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right) = -N_{iE}\left(\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}'\mathbf{B}_{iI}\right), \quad \mathrm{RB}_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right) = -\frac{N_{iE}}{N_i}\,\frac{\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}'\mathbf{B}_{iI}}{\bar{Y}_i}. \tag{4.9}$$

This bias is small when the same model holds for the included and excluded individuals.

Since the calibration estimator $\hat{Y}_i^{\mathrm{LCAL}}$ is intended to estimate $Y_i$ (and not $Y_{iI}$), for the domain mean $\bar{Y}_i = Y_i/N_i$ we consider the estimator obtained simply dividing $\hat{Y}_i^{\mathrm{CAL}}$ by $N_i$ (instead of $N_{iI}$), $\hat{\bar{Y}}_i^{\mathrm{LCAL}} = \hat{Y}_i^{\mathrm{LCAL}}/N_i$. The asymptotic bias of $\hat{\bar{Y}}_i^{\mathrm{LCAL}}$ is given by (4.9) divided by $N_i$.

We now analyze properties under the model and the sampling replication mechanism. Note that $\hat{\mathbf{B}}_i$ in the GREG estimator 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 $i$:

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_i + \varepsilon_{ij},\ E_m\left(\varepsilon_{ij}\right) = 0,\ E_m\left(\varepsilon_{ij}^2\right) = \sigma_\varepsilon^2, \quad j = 1,\ldots,N_i, \tag{4.10}$$

where model errors $\varepsilon_{ij}$ are all mutually independent. We wish to see the value added by the model to the design properties of the estimators; that is, how much would be gained if data were actually generated (at least approximately) by the assumed model. Let $E_m$ denote expectation under model (4.10). If the linear

regression model (4.10) actually holds for all the units in the domain (included and excluded), then $E_m(\mathbf{B}_{iI}) = \boldsymbol{\beta}_i$ and taking expectation of the bias term in (4.9) under model (4.10), we obtain the model-design bias,

$$B_{m,\pi}\left(\tilde{Y}_i^{\mathrm{LCAL}}\right) = -N_{iE}\left\{E_m\left(\bar{Y}_{iE}\right) - \bar{\mathbf{X}}'_{iE}E_m(\mathbf{B}_{iI})\right\} = -N_{iE}\left(\bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i - \bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i\right) = 0. \tag{4.11}$$

In contrast, assuming exactly the same regression model, the bias of the basic direct estimator $\hat{\bar{Y}}_i^{\mathrm{HA}}$ under cut-off sampling is not zero unless the means of the auxiliary variables for the excluded and included units are equal. Indeed,

$$B_{m,\pi}\left(\hat{Y}_i^{\mathrm{HA}}\right) = N_{iE}E_m\left(\bar{Y}_{iI} - \bar{Y}_{iE}\right) = N_{iE}\left(\bar{\mathbf{X}}_{iI} - \bar{\mathbf{X}}_{iE}\right)'\boldsymbol{\beta}_i. \tag{4.12}$$

Thus, the condition under which the LCAL estimator is design-unbiased, namely that the linear model (4.10) holds without error 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 basic direct estimator and can reduce substantially the cut-off sampling bias if the outcome is generated by the above domain-specific linear regression model.

Turning now to LCALN estimator (4.6), we define the corresponding theoretical version

$$\tilde{Y}_i^{\mathrm{LCALN}} = \hat{Y}_i + \left(\mathbf{X} - \hat{\mathbf{X}}\right)'\mathbf{B}_{iI}^N, \tag{4.13}$$

where $\mathbf{B}_i^N$ is the census version for the included units,

$$\mathbf{B}_i^N = \left(\sum_{\ell=1}^{m}\sum_{j\in U_{\ell I}}\mathbf{x}_{\ell j}\mathbf{x}'_{\ell j}\right)^{-1}\sum_{j\in U_{iI}}\mathbf{x}_{ij}y_{ij}.$$

Decomposing the bias similarly as in (4.8), we obtain

$$B_\pi\left(\hat{Y}_i^{\mathrm{LCALN}}\right) = E_\pi\left\{\left(\mathbf{X} - \hat{\mathbf{X}}\right)'\left(\hat{\mathbf{B}}_i^N - \mathbf{B}_{iI}^N\right)\right\} + B_\pi\left(\tilde{Y}_i^{\mathrm{LCALN}}\right). \tag{4.14}$$

Again, $E_\pi\left\{\left(\mathbf{X} - \hat{\mathbf{X}}\right)\left(\hat{\mathbf{B}}_i^N - \mathbf{B}_{iI}^N\right)\right\}/N_i$ is not zero for small $n_i$ but it tends to zero as $n_i \to \infty$ even under cut-off sampling, whereas $B_\pi\left(\tilde{Y}_i^{\mathrm{LCALN}}\right) = 0$ only in the absence of cut-off sampling bias. In general, 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, the design bias of $\tilde{Y}_i^{\mathrm{LCALN}}$ is given by

$$B_\pi\left(\tilde{Y}_i^{\mathrm{LCALN}}\right) = -\left(Y_{iE} - \mathbf{X}'_E\mathbf{B}_{iI}^N\right). \tag{4.15}$$

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

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij},\ E_{m_2}\left(\varepsilon_{ij}\right) = 0,\ E_{m_2}\left(\varepsilon_{ij}^2\right) = \sigma_\varepsilon^2,\quad j = 1, \ldots, N_i,\ i = 1, \ldots, m, \tag{4.16}$$

where again the model errors $\varepsilon_{ij}$ are mutually independent. Note that, under this model, $E_{m_2}\left(\mathbf{B}_{iI}^N\right) \neq \boldsymbol{\beta}$ in general, but if we consider the sum $\mathbf{B}_I = \sum_{i=1}^m \mathbf{B}_{iI}^N$ instead, we have $E_{m_2}(\mathbf{B}_I) = \boldsymbol{\beta}$. This means that the theoretical LCALN estimator for a particular domain, $\tilde{Y}_i^{\text{LCALN}}$, is not model-design unbiased, because

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

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

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

Hence, under model (4.16) with constant regression coefficients for all the population units, 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 theoretical estimator $\tilde{\bar{Y}}_i^{\text{LCALN}} = \tilde{Y}_i^{\text{LCALN}} / N_i$ is given by (4.15) divided by $N_i$.

We now study the variances. For the theoretical LCAL estimator (4.7), the design-variance is given by

$$V_\pi\left(\tilde{Y}_i^{\text{LCAL}}\right) = V_\pi\left(\hat{Y}_i - \hat{\mathbf{X}}_i'\mathbf{B}_{iI}\right) = V_\pi\left(\sum_{j\in s_i} w_{j|i} E_{ij}\right), \tag{4.17}$$

where $E_{ij} = y_{ij} - \mathbf{x}_{ij}'\mathbf{B}_{iI}$, $j \in U_{iI}$. We can then apply the usual variance estimators for expansion type estimators. In the case of LCALN given in (4.13), the variance is given by

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

Note that $\hat{\mathbf{X}}$ is based on the $n$ sample units, whereas $\hat{\mathbf{X}}_i$ uses only the $n_i$ units in domain $i$. As a consequence, the contribution of $\hat{\mathbf{X}}$ to the variance of LCALN should be much smaller than the contribution of $\hat{\mathbf{X}}_i$ in (4.17). This means that, provided that the domain and national regression lines are similar, the variance of LCALN estimator, obtained from the calibration at the national level, should be smaller than that of the domain-specific calibration estimator LCAL.

# 5 EBLUP under the nested error model

Estimators described so far use only the outcome information coming from the domain. This means that, when the domain sample size $n_i$ is small, these estimators might be inefficient even in the absence of cut-off sampling. Small area (or indirect) estimation methods are designed to reduce the variance by increasing the effective sample size; see Rao and 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. Since the model-based properties are

well known, we wish to analyze whether the estimators have good properties under the sampling-replication mechanism, which does not assume that the model actually holds.

We consider a very popular unit level model introduced by Battese, Harter and Fuller (1988) and often called nested error model. Similarly as for model $m_2$ in (4.16), 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

$$\begin{aligned} y_{ij} &= \mathbf{x}_{ij}'\boldsymbol{\beta} + u_i + e_{ij},\ u_i \overset{\text{iid}}{\sim} N\left(0, \sigma_u^2\right), \\ e_{ij} &\overset{\text{iid}}{\sim} N\left(0, \sigma_e^2\right),\ j = 1, \ldots, N_i,\ i = 1, \ldots, m, \end{aligned} \tag{5.1}$$

where area effects $u_i$ and errors $e_{ij}$ are all mutually independent. The vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = \left(\sigma_u^2, \sigma_e^2\right)'$ are unknown. Setting $\sigma_u^2 = 0$ in (5.1), we obtain model $m_2$ given in (4.16). If $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{iN_i}\right)'$ denotes the vector of outcomes for domain $i$ and $\mathbf{X}_i = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i}\right)'$ the corresponding design matrix, the model in matrix notation reads

$$\mathbf{y}_i \overset{\text{ind}}{\sim} N\left(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i\right),\ \mathbf{V}_i = \sigma_u^2\mathbf{1}_{N_i}\mathbf{1}_{N_i}' + \sigma_e^2\mathbf{I}_{N_i},\ i = 1, \ldots, m, \tag{5.2}$$

where $\mathbf{1}_k$ denotes a vector of ones of size $k$ and $\mathbf{I}_k$ is the $k \times k$ identity matrix.

We consider linear domain parameters defined as $H_i = \mathbf{b}_i'\mathbf{y}_i$, where $\mathbf{b}_i$ is a non-stochastic vector of known elements. The domain mean $H_i = \bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$ is obtained with $\mathbf{b}_i = N_i^{-1}\mathbf{1}_{N_i}$.

A sample $s_i$ is supposed to be drawn from the set of included units in domain $i$, that is, $s_i \subset U_{iI}$. We denote by $r_i = \left(U_{iI} - s_i\right) \cup U_{iE}$ the set of non-sampled units from domain $U_i$, which includes those non-sampled units from $U_{iI}$ and all the units in $U_{iE}$. Note that $U_i = s_i \cup r_i = U_{iI} \cup U_{iE}$. Then, the overall sample $s$ is composed of the samples $s_i$ drawn from the sets of included units in each area $U_{iI}$, $i = 1, \ldots, m$, that is, $s = s_1 \cup \cdots \cup s_m$.

We decompose the domain vector $\mathbf{y}_i$ and the design and covariance matrices $\mathbf{X}_i$ and $\mathbf{V}_i$ into the corresponding subvectors and submatrices for sample and out-of-sample units, indicated with subscripts $s$ and $r$ respectively, as follows

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_{is} \\ \mathbf{y}_{ir} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_{is} \\ \mathbf{X}_{ir} \end{pmatrix}, \quad \mathbf{V}_i = \begin{pmatrix} \mathbf{V}_{is} & \mathbf{V}_{isr} \\ \mathbf{V}_{irs} & \mathbf{V}_{ir} \end{pmatrix}.$$

The linear parameter $H_i = \mathbf{b}_i'\mathbf{y}_i$ can then be expressed as $H_i = \mathbf{b}_{is}'\mathbf{y}_{is} + \mathbf{b}_{ir}'\mathbf{y}_{ir}$. Under model (5.1), the best linear unbiased predictor (BLUP) of $H$ is the model-unbiased linear function of the sample data $\hat{H}_i = \boldsymbol{\alpha}_{is}'\mathbf{y}_{is}$, which minimizes the model mean squared error (MSE), $\text{MSE}_{m_3}\left(\hat{H}_i\right) = E_{m_3}\left(\hat{H}_i - H_i\right)^2$. The BLUP of $H_i = \mathbf{b}_{is}'\mathbf{y}_{is} + \mathbf{b}_{ir}'\mathbf{y}_{ir}$ is then

$$\hat{H}_i^{\text{BLUP}}\left(\boldsymbol{\theta}\right) = \mathbf{b}_{is}'\mathbf{y}_{is} + \mathbf{b}_{ir}'\left[\mathbf{X}_{ir}\tilde{\boldsymbol{\beta}}_s + \mathbf{V}_{irs}\mathbf{V}_{is}^{-1}\left(\mathbf{y}_{is} - \mathbf{X}_{is}'\tilde{\boldsymbol{\beta}}_s\right)\right], \tag{5.3}$$

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

$$\tilde{\boldsymbol{\beta}}_s = \tilde{\boldsymbol{\beta}}_s(\boldsymbol{\theta}) = \left(\sum_{i=1}^{m} \mathbf{X}_{is}' \mathbf{V}_{is}^{-1} \mathbf{X}_{is}\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_{is}' \mathbf{V}_{is}^{-1} \mathbf{y}_{is}. \tag{5.4}$$

The BLUP of $H_i$ given in (5.3) depends on the true values of the variance components $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$, which are typically unknown. Replacing them by corresponding model-consistent estimators $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$, we obtain the so-called empirical BLUP (EBLUP), denoted $\hat{H}_i^{\text{EBLUP}} = \hat{H}_i^{\text{BLUP}}(\hat{\boldsymbol{\theta}})$.

If the domain sampling fraction, $n_i / N_i$, is negligible, the BLUP of $\bar{Y}_i$ may be expressed as the weighted average

$$\hat{\bar{Y}}_i^{\text{BLUP}} \cong \gamma_{is}\left[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})' \tilde{\boldsymbol{\beta}}_s\right] + (1 - \gamma_{is}) \bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s, \tag{5.5}$$

where $\gamma_{is} = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 / n_i)$ is in the $(0, 1)$ interval and tends to 1 as $n_i \to \infty$ (Rao and Molina, 2015). Thus, for domains with large sample size $n_i$, $\hat{\bar{Y}}_i^{\text{BLUP}}$ 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 $n_i$, $\hat{\bar{Y}}_i^{\text{BLUP}}$ borrows strength from the other domains by approaching the regression-synthetic estimator $\bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s$. Replacing the variance components in $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$ by consistent estimators $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ in the BLUP, denoting $\hat{\gamma}_{is} = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i)$ and $\hat{\boldsymbol{\beta}}_s = \tilde{\boldsymbol{\beta}}_s(\hat{\boldsymbol{\theta}})$, we obtain the EBLUP of $\bar{Y}_i$, given by

$$\hat{\bar{Y}}_i^{\text{EBLUP}} \cong \hat{\gamma}_{is}\left[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})' \hat{\boldsymbol{\beta}}_s\right] + (1 - \hat{\gamma}_{is}) \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}}_s. \tag{5.6}$$

The BLUP is unbiased and optimal under model $m_3$ in the sense of minimizing the MSE under that model. We now study its design properties, which do not assume that the model is correct and hence account for bias under model departures. To that end, we consider the census regression parameter for the included units, defined as $\mathbf{B}_I = \left(\sum_{i=1}^{m} \mathbf{X}_{iI}' \mathbf{V}_{iI}^{-1} \mathbf{X}_{iI}\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_{iI}' \mathbf{V}_{iI}^{-1} \mathbf{y}_{iI}$, where $\mathbf{y}_{iI}$, $\mathbf{X}_{iI}$ and $\mathbf{V}_{iI}$ are the corresponding sub-vector and sub-matrices of $\mathbf{y}_i$, $\mathbf{X}_i$ and $\mathbf{V}_i$, for the included units $(j \in U_{iI})$. Again, we consider the theoretical version of the BLUP defined in terms of $\mathbf{B}_I$,

$$\tilde{\bar{Y}}_i^{\text{BLUP}} = \gamma_{is}\left[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})' \mathbf{B}_I\right] + (1 - \gamma_{is}) \bar{\mathbf{X}}_i' \mathbf{B}_I. \tag{5.7}$$

If each sample $s_i$ is drawn from the corresponding domain $U_{iI}$ 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 $\tilde{\bar{Y}}_i^{\text{BLUP}}$ under SRSWOR, which is given by

$$B_\pi\left(\tilde{\bar{Y}}_i^{\text{BLUP}}\right) = \gamma_{is} \frac{N_{iE}}{N_{iI}}\left[\left(\bar{Y}_i - \bar{\mathbf{X}}_i' \mathbf{B}_I\right) - \left(\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}' \mathbf{B}_I\right)\right] + (1 - \gamma_{is})\left(\bar{\mathbf{X}}_i' \mathbf{B}_I - \bar{Y}_i\right). \tag{5.8}$$

This bias will be small if (5.1) holds for the whole population, in which case $E_{m_3}(\bar{Y}_i) = \bar{\mathbf{X}}_i \boldsymbol{\beta}$ and $E_{m_3}(\bar{Y}_{iE}) = \bar{\mathbf{X}}_{iE} \boldsymbol{\beta}$. Using these results when taking expectation under model $m_3$ in (5.8), we get $B_{m_3,\pi}\left(\tilde{\bar{Y}}_i^{\text{BLUP}}\right) = 0$. In fact, the same result also holds under model $m_2$.

Concerning variance, if $s_i$ is obtained by SRSWOR within $U_{iI}$, the design-variance of the theoretical BLUP estimator is given by

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

Hence, if the census least squared (LS) regression lines for the domains from model (4.10) are similar to the national census weighted least squared (WLS) regression line from model (5.1), that is, if $\mathbf{B}_I \approx \mathbf{B}_{iI}$, then the variance of the BLUP for $\bar{Y}_i$ reduces to that of the LCAL estimator of $\bar{Y}_i$ obtained from (4.17), multiplied by the factor $\gamma_{is}^2 \in (0, 1)$.

Under more general sampling designs within $U_{iI}$, we consider the pseudo-EBLUP of $\bar{Y}_i$ proposed by You and Rao (2002) instead of the EBLUP. Defining the analogous theoretical estimator that uses the weighted sample means $\bar{y}_{iw} = \left(\sum_{j\in s_i} w_{j|i}\right)^{-1} \sum_{j\in s_i} w_{j|i} y_{ij}$ and $\bar{\mathbf{x}}_{iw} = \left(\sum_{j\in s_i} w_{j|i}\right)^{-1} \sum_{j\in s_i} w_{j|i}\mathbf{x}_{ij}$ instead or the unweighted ones $\bar{y}_{is}$ and $\bar{\mathbf{x}}_{is}$ in (5.7), we obtain the same expressions for the design bias and variance, with $\gamma_{is}$ changed to $\gamma_{iw} = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 \delta_{iw})$, for $\delta_{iw} = \left(\sum_{j\in s_i} w_{j|i}\right)^{-2} \sum_{j\in s_i} w_{j|i}^2$.

# 6 Empirical best predictor under the nested error model

Estimation of non-linear domain parameters requires more general small area estimation methods, such as the best/Bayes predictor (BP), see Molina and Rao (2010). Special non-linear parameters are poverty and inequality indicators defined in terms of a welfare measure, such as the family of poverty indicators introduced by Foster, Greer and Thorbecke (1984). The best predictor can also be used for the estimation of other characteristics such as median, quantiles or even the whole empirical distribution function of the variable of interest, see Pratesi (2016). Additionally, it can be used to estimate totals and means of a given target variable, when the dependent variable in the considered model is a one-to-one transformation (e.g., log or more general Box-Cox transformations) of this target variable. These transformations are typically applied in the case of non-normality or heteroscedasticity.

In this section, the target variable (e.g., the welfare measure) for the $j^{\text{th}}$ unit in $i^{\text{th}}$ domain is denoted as $v_{ij}$ and $y_{ij} = T\left(v_{ij}\right)$, where $T$ is a one-to-one transformation. We assume that $y_{ij}$ follows the nested error model (5.1). By the inverse transformation $v_{ij} = T^{-1}\left(y_{ij}\right)$, we can express our target parameter (defined originally in terms of the target variables $v_{ij}$) as a function of the vector $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{iN_i}\right)'$ of model responses for the domain units, $H_i = h\left(\mathbf{y}_i\right)$. The best predictor (BP) of $H_i = h\left(\mathbf{y}_i\right)$ is defined as the function of the sample data $\mathbf{y}_{is}$ that minimizes the model MSE, and it turns out to be

$$\hat{H}_i^{\text{BP}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = E_{m_3}\left[h\left(\mathbf{y}_i\right) \middle| \mathbf{y}_{is}; \boldsymbol{\beta}, \boldsymbol{\theta}\right], \tag{6.1}$$

where the expectation is taken with respect to the model distribution of $\mathbf{y}_{ir} \,|\, \mathbf{y}_{is}$, which depends on the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The BP of $H_i$ is unbiased with respect to the model (5.1), regardless of the complexity of the function $h(\cdot)$ defining the target parameter. However, it cannot be calculated in practice

since model parameters $\boldsymbol{\beta}$ and $\theta$ are typically unknown. An empirical best predictor (EBP) of $H_i$, denoted as $\hat{H}_i^{\mathrm{EBP}}$, is then obtained by replacing $\boldsymbol{\beta}$ and $\theta$ in $\hat{H}_i^{\mathrm{BP}}(\boldsymbol{\beta}, \theta)$ by consistent estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}$, as $\hat{H}_i^{\mathrm{EBP}} = \hat{H}_i^{\mathrm{BP}}(\hat{\boldsymbol{\beta}}, \hat{\theta})$. The EBP is not exactly unbiased, but the bias arising from the estimation of $\boldsymbol{\beta}$ and $\theta$ is typically negligible when the overall sample size $n$ is large. In the case of a linear parameter $H_i = \mathbf{b}_i'\mathbf{y}_i$, the EBP under the nested error model with normality obtained using $\hat{\boldsymbol{\beta}}_s = \tilde{\boldsymbol{\beta}}_s(\hat{\theta})$ to estimate $\boldsymbol{\beta}$ equals $\hat{H}_i^{\mathrm{EBLUP}}$.

When $h(\cdot)$ is so complex that the expectation defining the EBP in (6.1) cannot be calculated analytically, Monte Carlo methods can be applied to approximate $\hat{H}_i^{\mathrm{EBP}}$ as proposed in Molina and Rao (2010). This is done by simulating, from the model (5.1) fitted to the original sample data, $L$ 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 $\mathbf{y}_i^{(\ell)}$, calculating the corresponding target parameter $H_i^{(\ell)} = h(\mathbf{y}_i^{(\ell)})$ for each $\ell = 1, \ldots, L$ and, finally, averaging over the $L$ replicates as $\hat{H}_i^{\mathrm{EBP}} = L^{-1}\sum_{\ell=1}^{L} H_i^{(\ell)}$. 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, see Molina and Rao (2010).

# 7 MSE estimation

The EBLUP in Section 5 or the EBP described in Section 6 are based on the nested error model (5.1). Calibration estimators described in Section 4 are also assisted by a linear regression model. 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 and Rao (2010), which is based on the original parametric bootstrap method for finite populations of González-Manteiga, Lombardia, Molina, Morales and Santamaría (2008). According to this procedure, the bootstrap MSE of $\hat{H}_i^{\mathrm{EBP}}$ under the nested error model (5.1) is obtained as follows: i) Fit Model (5.1) to the sample data $\{(\mathbf{y}_{is}, \mathbf{X}_{is}); i = 1, \ldots, m\}$, to obtain the estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ of $\boldsymbol{\beta}$, $\sigma_u^2$ and $\sigma_e^2$ respectively. ii) For $b = 1, \ldots, B$, generate independently $u_i^{*(b)} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}_u^2)$ and $e_{ij}^{*(b)} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. iii) For $b = 1, \ldots, B$, construct bootstrap domain vectors $\mathbf{y}_i^{*(b)} = (y_{i1}^{*(b)}, \ldots, y_{iN_1}^{*(b)})'$, whose elements are generated as

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

From the bootstrap domain vector $\mathbf{y}_i^{*(b)}$, calculate the target bootstrap parameter $H_i^{*(b)} = h(\mathbf{y}_i^{*(b)})$, for $b = 1, \ldots, B$. iv) From each bootstrap population vector $\mathbf{y}_i^{*(b)}$, take the sample part $\mathbf{y}_{is}^{*(b)}$, where the sample indices $s_i$ are exactly those of the original sample drawn from $U_{iI}$, for $i = 1, \ldots, m$. Using the overall bootstrap sample data $\mathbf{y}_s^{*(b)} = (\mathbf{y}_{1s}^{*(b)}, \ldots, \mathbf{y}_{ms}^{*(b)})'$ and the population vectors $\mathbf{x}_{ij}$, $j = 1, \ldots, N_i$, assumed to be known for all population units, calculate the bootstrap EBP of $H_i$, denoted as $\hat{H}_i^{\mathrm{EBP}*(b)}$, $b = 1, \ldots, B$. v) A bootstrap MSE estimator for the EBP under model (5.1), $\mathrm{MSE}_{m_3}(\hat{H}_i^{\mathrm{EBP}})$, is obtained as

$$\text{mse}_B\left(\hat{H}_i^{\text{EBP}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{H}_i^{\text{EBP}*(b)} - H_i^{*(b)}\right)^2. \quad (7.1)$$

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, $H_i = \mathbf{b}_i'\mathbf{y}_i$, if $\hat{\boldsymbol{\beta}}_s$ is the WLS estimator (5.4), then (7.1) is actually an estimator of $\text{MSE}_{m_3}\left(\hat{H}_i^{\text{EBLUP}}\right)$. This naïve bootstrap estimator of the model MSE is first-order unbiased in the sense that its model bias is $O\left(m^{-1}\right)$, but not $o\left(m^{-1}\right)$. 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, for simplicity, we consider the naive bootstrap estimator (7.1), which cannot yield negative values and performs well for moderate number of areas $m$.

# 8 Simulation experiments

## 8.1 Aims and general description

In this section, we describe simulation experiments designed to compare the small sample properties of the estimators of $\overline{Y}_i$ discussed above in the context of cut-off sampling. Specifically, we compare the naïve direct estimator $\hat{\overline{Y}}_i^{\text{HA}}$, calibration estimators $\hat{\overline{Y}}_i^{\text{LCAL}}$ and $\hat{\overline{Y}}_i^{\text{LCALN}}$, and the EBLUP under the nested error model $\hat{\overline{Y}}_i^{\text{EBLUP}}$, under two different scenarios. In the first scenario, the values of the target variable for all the population units are generated from the same model; in the second, included and excluded units are generated from different models.

In the absence of cut-off sampling, 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 model-based predictors also perform well with respect to the (cut-off sampling) design. For this reason, we run design-based simulations by generating one population vector $\mathbf{y} = \left(\mathbf{y}_1', \ldots, \mathbf{y}_m'\right)'$ from the nested error model in (5.1), keeping it fixed and repeatedly drawing a new cut-off sample in each MC simulation. Allocation of units to the sets of included or excluded units is done by generating a random binary variable $c_{ij}$ for each unit $j = 1, \ldots, N_i$ and area $i = 1, \ldots, m$. 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 $x_{ij\kappa} \overset{\text{iid}}{\sim} N(3, 2)$, $\kappa = 1, 2, 3$. The binary variables $c_{ij}$ determining the allocation of units in $U_{iI}$ or $U_{iE}$ for each domain $i$

are generated independently as $c_{ij} \overset{\text{ind}}{\sim} \text{Bern}\left(p_{j|i}\right)$, where the probabilities $p_{j|i} = \Pr\left(c_{ij} = 1\right)$ are related to the vector of auxiliary variables $\mathbf{x}_{ij} = \left(x_{ij1}, x_{ij2}, x_{ij3}\right)'$ in the form

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

We take $\boldsymbol{\zeta} = (0.75, 1, 1)'$. Based on this value, the total number of included units (with $c_{ij} = 1$) from all the domains represents roughly half of the population.

The values of the target variable $y_{ij}$ are generated from the nested error model (5.1) using $\mathbf{x}_{ij} = \left(x_{ij1}, x_{ij2}, x_{ij3}\right)'$ 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, keeping the population values $\left\{\left(\mathbf{x}_{ij}, y_{ij}, c_{ij}\right);\ j = 1, \ldots, N_i,\ i = 1, \ldots, m\right\}$ fixed, 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 taken as $n_i \in \{5, 10, 30, 50\}$, with each sample size repeated for 20 subsequent domains. With the data from the $k^{\text{th}}$ sample, we compute the basic direct estimator, calibration estimators at the domain level (LCAL) and at the population level (LCALN), and EBLUP. Weights, $h_{j|i}$ and $g_{j|i}$, in the calibration estimators (4.3) and (4.6) respectively are obtained using the function calib from package `sampling` (Tillé and Matei, 2016) of R (R Development Core Team, 2016). EBLUPs are obtained using R package `sae` (Molina and Marhuenda, 2015), which by default estimates the model parameters $\sigma_u^2$, $\sigma_e^2$ and $\boldsymbol{\beta}$ using restricted maximum likelihood (REML).

Let $\hat{\bar{Y}}_i$ be a generic estimator of $\bar{Y}_i$ and $\hat{\bar{Y}}_i^{(k)}$ its value obtained with $k^{\text{th}}$ sample. We evaluate the performance of estimators in terms of relative bias (RB) and relative root MSE (RRMSE) under the design, approximated empirically as

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

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

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

Figure 8.1 displays boxplots of percent RB for the considered estimators of the mean $\bar{Y}_i$, where each boxplot is for the 20 domains in each group of sample sizes $n_i = 5, 10, 30, 50$. We can see the large cut-off sampling bias of the basic direct estimator, with median RB exceeding 20% for all the domain sample sizes. This cut-off sampling bias is corrected by all the other estimators. Nevertheless, the LCALN estimator shows wider boxplots. This estimator gets large bias for some domains probably because its assisting model is not accounting for the domain effects. The LCAL estimator is based on a model that accounts for domain effects and performs well in terms of design bias uniformly for all the domain sample sizes, although EBLUP also performs rather well in terms of design bias.

Looking now at the RRMSE in Figure 8.2, we can see the much smaller RRMSEs of EBLUPs for all the domain sample sizes. The LCAL estimator gets closer RRMSEs as the domain sample size grows, but for $n_i = 5$ it gets huge RRMSEs. We have seen that the LCALN can be substantially biased for some domains and it also has large RRMSEs for all the domain sample sizes. Thus, in summary, EBLUP exhibits the lowest design RRMSE and at the same time keeps the design bias under control.

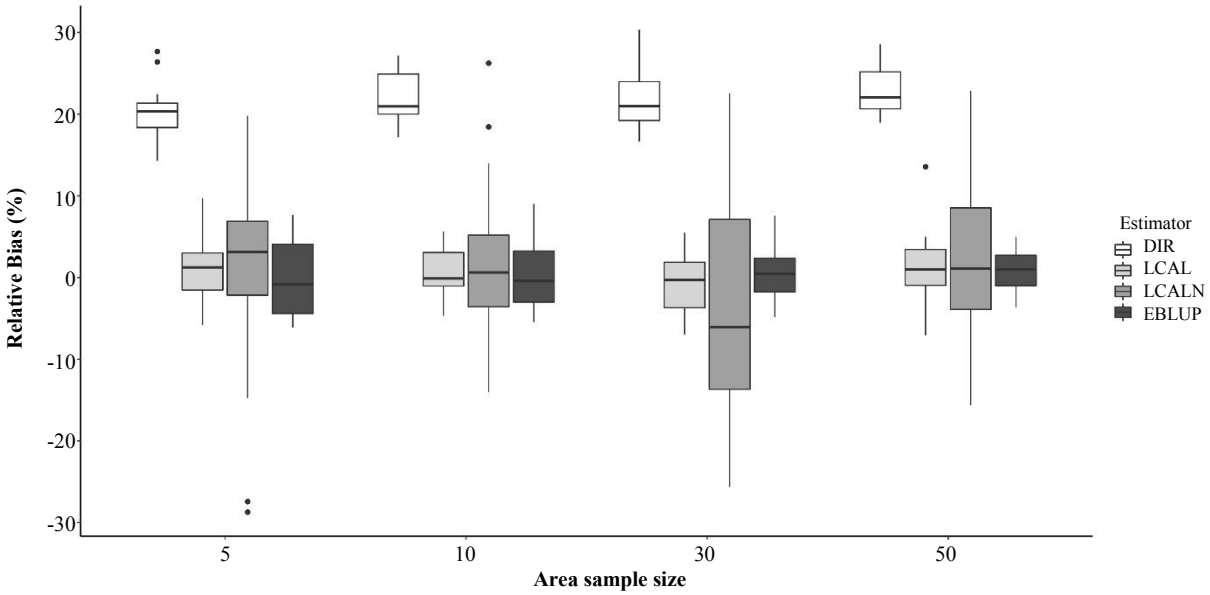


**Figure 8.1 Boxplots of domain RBs (%) of basic direct, LCAL, LCALN and EBLUP estimators for $n_i = 5, 10, 30, 50$.**

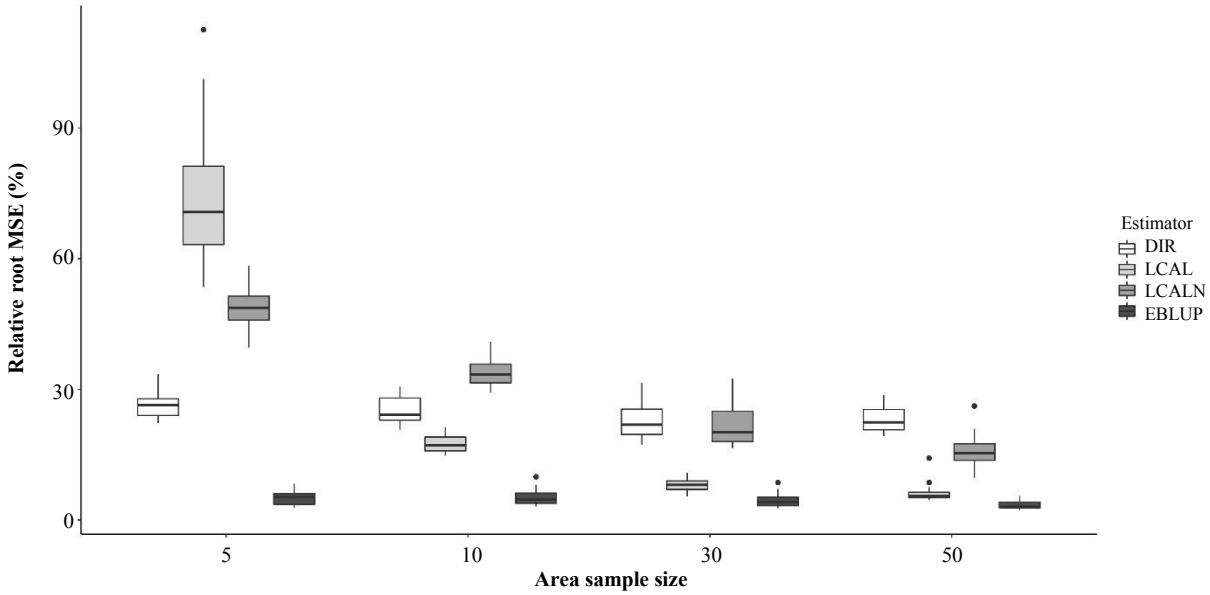


**Figure 8.2 Boxplots of domain RRMSEs (%) of basic direct, LCAL, LCALN and EBLUP estimators for $n_i = 5, 10, 30, 50$.**

Table 8.1 reports averages across all the domains of absolute RB and RRMSE, together with % share of squared bias from the total design MSE. We can see again the large cut-off sampling bias of the basic direct estimator, with a bias share of $B_\pi^2 / \text{MSE}_\pi \approx 100\%$, in contrast to all other estimators. The LCAL estimator has the smallest average ARB, followed closely by EBLUP. LCALN performs the best in terms of bias ratio because of its large MSE. Thus, we consider that LCAL performs better. As already said, EBLUP clearly performs the best when looking at both bias and MSE.

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

| Method | $\overline{\text{ARB}}$ | $\overline{\text{RRMSE}}$ | $B_\pi^2 / \text{MSE}_\pi$ |
|---|---|---|---|
| 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 experiment, we preserve the same population values and sampling scheme as before, but the values of the target variable for the included and excluded units are generated from models with different parameter values. Of course, this is not a favorable scenario for the considered model-based estimators, but it may be realistic since, in practice, the assumed model cannot be checked for the excluded units. Thus, instead of a constant $\boldsymbol{\beta}$ for all the 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 excluded ones. The values of the explanatory variables and variance components $\sigma_u^2$ and $\sigma_e^2$ are taken exactly as before. Again, we draw $K = 1{,}000$ samples $s^{(k)}$ by independent SRSWOR within the units in domain $i$ with $c_{ij} = 1$, with the same domain sample sizes $n_i$ as before. With the sample data from the $k^{\text{th}}$ sample, we compute basic direct, LCAL, LCALN and EBLUP estimates of $\bar{Y}_i$.

Figure 8.3 shows boxplots of the corresponding percent RBs for each domain sample size. In this case, all the estimators are biased, but the bias of the basic direct estimator becomes huge, exceeding 40% 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 and for this reason the RBs of calibration estimators do not decrease as $n_i$ grows. Even under this unfavorable scenario of different generating models for included and excluded units, EBLUP shows a

moderate bias, which is comparable to that of LCAL estimator, and performs clearly the best in terms of RRMSE.

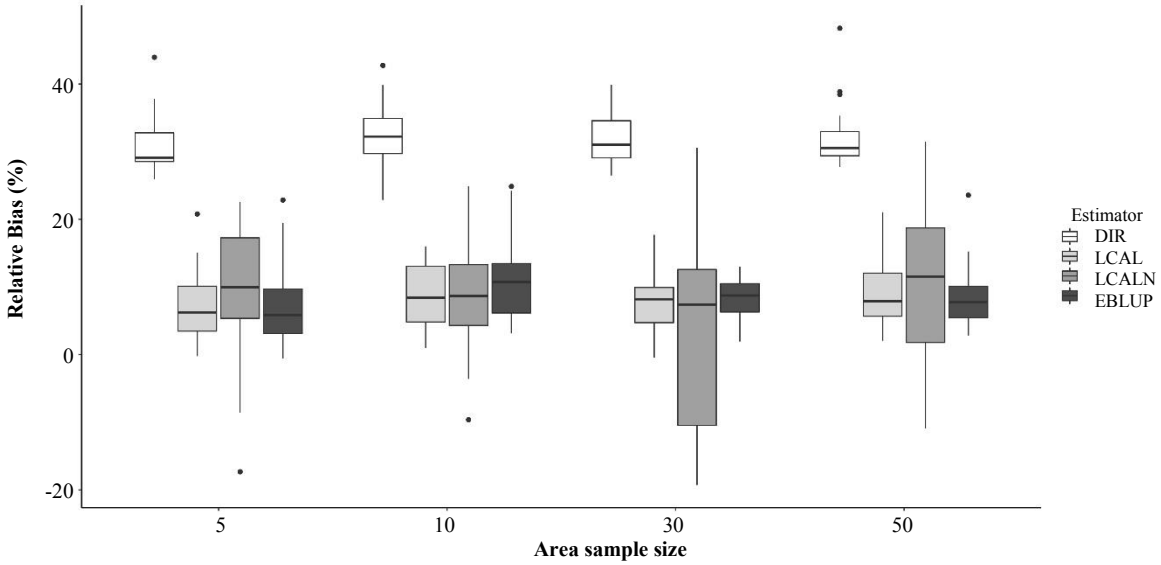


**Figure 8.3 Boxplots of domain RBs (%) of basic direct, LCAL, LCALN and EBLUP estimators for $n_i = 5, 10, 30, 50$, when $\beta_I = (1, 1.5, 1)'$ for included units and $\beta_E = (0.5, 1.6, 0.5)'$ for excluded ones.**

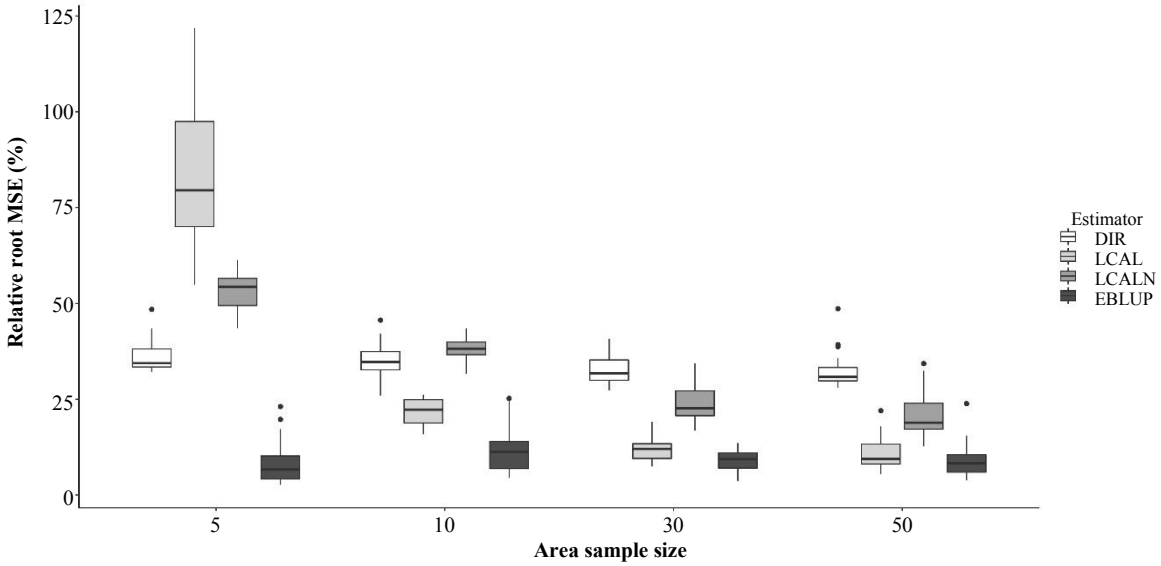


**Figure 8.4 Boxplots of domain RRMSEs (%) of basic direct, LCAL, LCALN and EBLUP estimators for $n_i = 5, 10, 30, 50$, when $\beta_I = (1, 1.5, 1)'$ for included units and $\beta_E = (0.5, 1.6, 0.5)'$ for excluded ones.**

Again, averages across all the domains of absolute RB and RRMSE are shown in Table 8.2, together with sq. bias ratio. As already noted, the basic direct estimator has a huge bias, whereas LCAL and EBLUP estimators keep an $\overline{\text{ARB}}$ below 10%. LCALN displays the lowest bias ratio because of a larger MSE. Again, EBLUP shows the best performance in terms of efficiency, with an average RRMSE also below 10%.

**Table 8.2**
**Averages across areas of absolute RB, RRMSE and $B_\pi^2/\text{MSE}_\pi$ for basic direct, LCAL, LCALN and EBLUP, when $\boldsymbol{\beta}_I = (1, 1.5, 1)'$ for included units and $\boldsymbol{\beta}_E = (0.5, 1.6, 0.5)'$ for excluded ones (in percentage)**

| Method | $\overline{\text{ARB}}$ | $\overline{\text{RRMSE}}$ | $B_\pi^2/\text{MSE}_\pi$ |
|---|---|---|---|
| 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 cut-off sampling bias of the basic direct estimator and EBLUP gets smaller RRMSE, specially for the domains with the smallest sample sizes.

# 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 $\hat{V}_i^{\text{HA}} = N_i \bar{V}_{iI}$, $i = 1, \ldots, m$, which have actually zero variance, but might be severely biased. Since true values in real applications are not available and therefore real biases cannot be evaluated (there is no information from $U_{iE}$), 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 $\hat{V}_i^{\text{HA}}$ 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

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

where $s_i^2 = (n_i - 1)^{-1} \sum_{j \in s_i} \left(v_{ij} - \bar{v}_{is}\right)^2$ 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 $\mathbf{x}_{ij} = \left(1, x_{ij}\right)'$, with $x_{ij} = z_{ij}^{1/2}$ as covariate seems to minimize the problem. Accordingly, we will consider a nested error model (5.1) 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 $\hat{V}_i^{\text{EBP}} = E_{m_3}\left[h(\mathbf{y}_i) \middle| \mathbf{y}_{is}; \hat{\boldsymbol{\theta}}\right]$, $i = 1, \ldots, m$, 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 for $H_i = V_i$, taking $H_i^{*(b)} = V_i^{*(b)}$ and $\hat{H}_i^{\text{EBP}*(b)} = \hat{V}_i^{\text{EBP}*(b)}$ and considering that the 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 according to (4.3) using $v_{ij}$ instead of $y_{ij}$ and $z_{ij}$ instead of $x_{ij}$, which is assisted by the linear model (4.10) for the untransformed sales $v_{ij}$ in terms of purchases $z_{ij}$. 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, replacing $\hat{H}_i^{\text{EBP}*(b)}$ by $\hat{V}_i^{\text{GREG}*(b)}$. 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, we analyze the residuals from the nested error model (5.1), 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, the residuals do not

exhibit any remarkable pattern. In fact, in the histogram they appear to be very much concentrated around zero, which indicates a high predictive power of the model.

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.

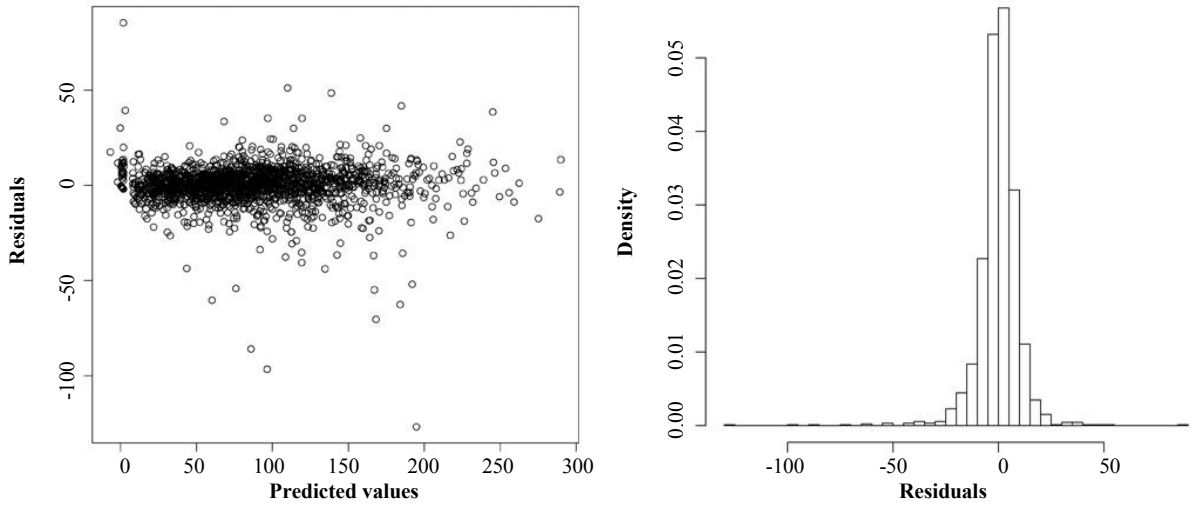


**Figure 9.1 EBP residuals against predicted values (left), and histogram of EBP residuals (right).**

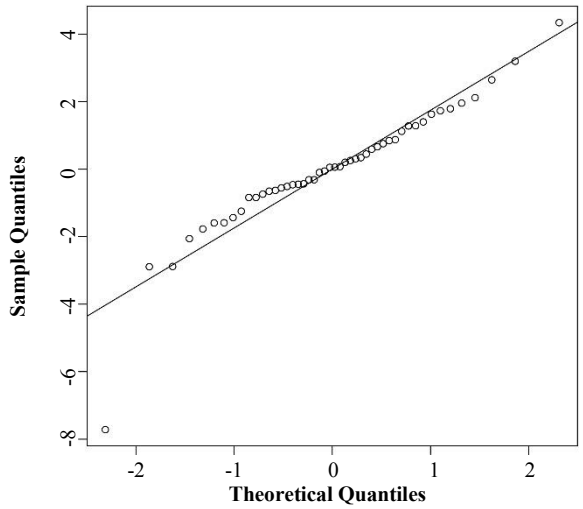


**Figure 9.2 Normal Q-Q plot of predicted province effects $\hat{u}_i$.**

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 a great similarity of the two types of estimates except for the two provinces with the largest sample sizes, where the EBPs are slightly larger than direct estimates, which could be due to cut-off sampling 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 actually understating 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 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). Clearly, the direct estimator performs the worst in terms of efficiency.

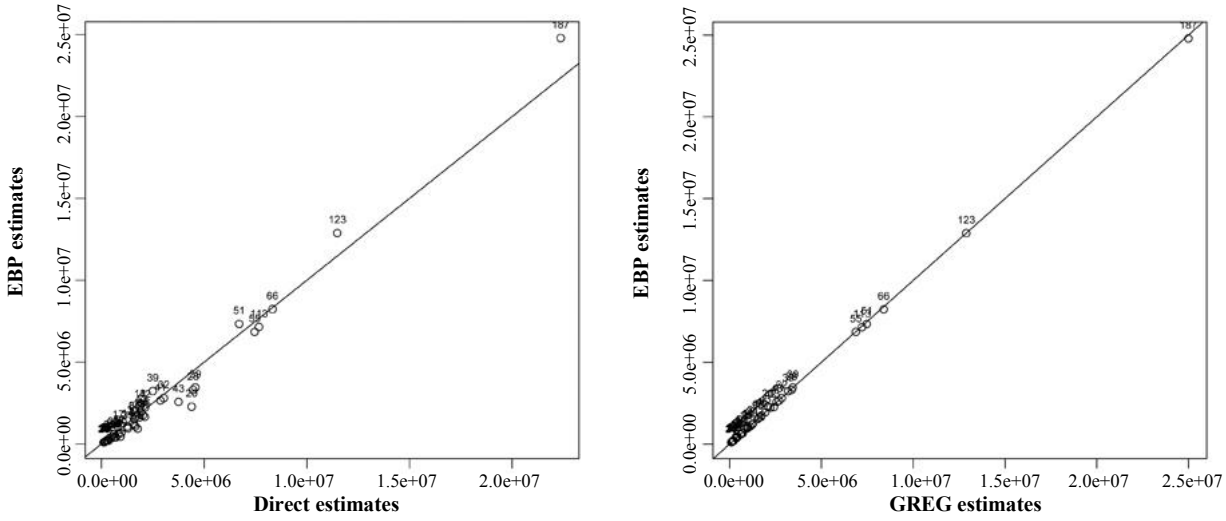


**Figure 9.3 EBPs of total sales for each province against direct estimates (left) and against GREG estimates (right).**

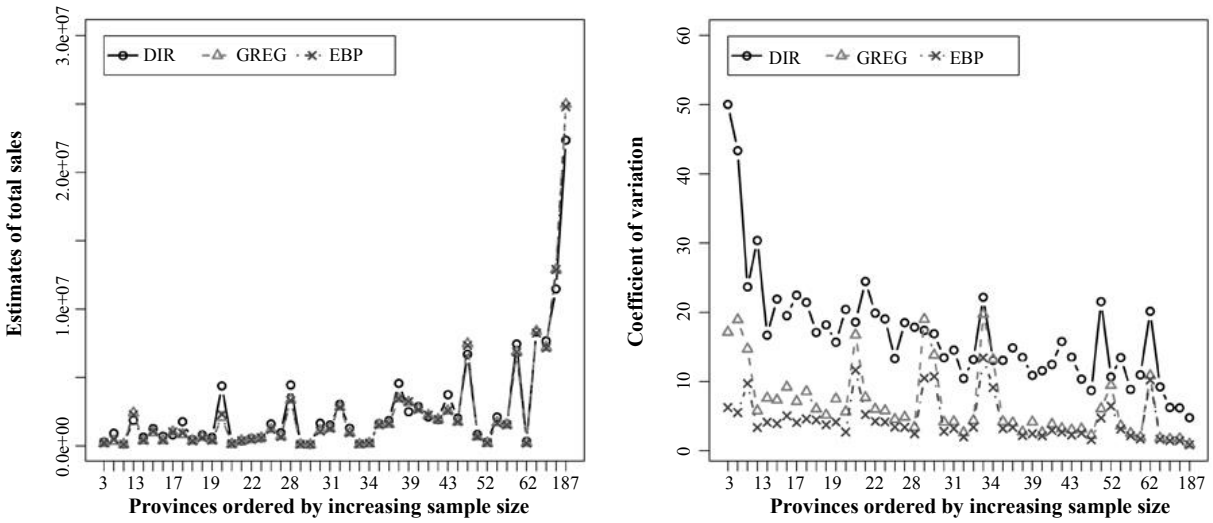


**Figure 9.4 Direct, calibration and EBP estimates of total sales for each province (left) and corresponding estimated coefficients of variation (right).**

Table A.1 in the Appendix reports direct, LCAL and EBP estimates of province total sales of the product supplemented 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 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, specially 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 $U_{il}$) 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.

MSEs of calibration and model-based estimators are obtained under the model. Design MSEs are preferred by National Statistical Institutes because they do not assume that a model is correct and therefore account for model failures. However, finding design-unbiased estimators for the design MSE under cut-off sampling encounters the same problems as finding design-unbiased estimators of the target domain indicators $H_i$. We plan to use the ideas of Strzalkowska-Kominiak and Molina (2019), based on borrowing strength from the other domains also for estimating the design MSE in a given domain, to find design MSE estimators with reduced cut-off sampling bias.

Finally, we have considered that the domains act as sampling strata and cut-off sampling is applied within each domain. Considering that the strata are different from the domains (typically cutting-across the domains) and applying cut-off sampling within each strata yields random domain sample sizes. Small area estimation is seldom studied under this case in the literature. Nevertheless, putting together the subsamples from the different strata corresponding to the same domain we get a sample from each domain. Inference could then be done conditionally on the observed domain sample sizes Rao (1985), which would reduce to the same problem considered here.

# Acknowledgements

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

# Appendix

## Estimates of total sales by provinces

**Table A.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 | $n_i$ | $\hat{V}_i^{\text{HA}}$ | $\hat{V}_i^{\text{GREG}}$ | $\hat{V}_i^{\text{EBP}}$ | cv($\hat{V}_i^{\text{HA}}$) | cv($\hat{V}_i^{\text{GREG}}$) | cv($\hat{V}_i^{\text{EBP}}$) |
|---|---|---|---|---|---|---|---|
| SORIA | 3 | 293,020.0 | 187,824.9 | 213,325.0 | 50.0 | 17.1 | 6.2 |
| ZAMORA | 7 | 932,520.0 | 345,095.8 | 454,657.0 | 43.3 | 18.9 | 5.5 |
| ALAVA | 11 | 130,083.6 | 119,918.5 | 118,835.3 | 23.7 | 14.7 | 9.7 |
| ALMERIA | 13 | 1,870,104.6 | 2,407,333.1 | 2,272,051.4 | 30.4 | 5.8 | 3.4 |
| PALENCIA | 14 | 626,340.0 | 380,367.4 | 409,775.4 | 16.7 | 7.6 | 4.1 |
| SALAMANCA | 14 | 1,265,580.0 | 966,094.1 | 1,068,230.6 | 21.9 | 7.3 | 3.9 |
| AVILA | 15 | 708,696.0 | 392,474.1 | 418,917.2 | 19.5 | 9.2 | 5.0 |
| LERIDA | 17 | 817,817.6 | 1,011,032.3 | 1,014,770.2 | 22.5 | 7.1 | 4.1 |
| CIUDAD REAL | 18 | 1,764,000.0 | 841,228.2 | 939,994.9 | 21.4 | 8.6 | 4.6 |
| GUADALAJARA | 18 | 463,047.8 | 362,148.3 | 363,856.9 | 17.1 | 6.0 | 4.5 |

**Table A.1 (continued)**
**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($\hat{V}_i^{EBP}$) |
|---|---|---|---|---|---|---|---|
| RIOJA | 18 | 809,900.0 | 622,488.3 | 595,178.6 | 18.2 | 5.2 | 3.7 |
| SEGOVIA | 19 | 610,370.5 | 386,734.4 | 402,324.0 | 15.7 | 7.5 | 4.2 |
| CACERES | 20 | 4,391,826.0 | 2,081,619.7 | 2,286,462.0 | 20.4 | 5.6 | 2.7 |
| GUIPUZCOA | 20 | 181,634.0 | 136,700.0 | 156,311.8 | 18.6 | 16.7 | 11.6 |
| HUESCA | 22 | 377,954.5 | 372,101.3 | 371,246.5 | 24.5 | 7.7 | 5.2 |
| TERUEL | 22 | 534,417.3 | 446,565.7 | 465,643.3 | 19.9 | 6.0 | 4.3 |
| CUENCA | 23 | 588,464.3 | 587,005.5 | 586,347.5 | 19.0 | 5.8 | 4.2 |
| VALLADOLID | 24 | 1,609,875.0 | 1,210,132.8 | 1,188,336.1 | 13.3 | 4.5 | 3.4 |
| BURGOS | 28 | 961,645.7 | 708,510.0 | 666,698.1 | 18.5 | 4.9 | 3.4 |
| CORDOBA | 28 | 4,457,614.3 | 3,367,169.5 | 3,312,801.5 | 17.9 | 3.4 | 2.4 |
| ORENSE | 28 | 148,577.1 | 88,104.6 | 108,428.9 | 17.4 | 19.0 | 10.5 |
| LUGO | 30 | 107,213.3 | 92,938.7 | 104,233.7 | 16.9 | 13.8 | 10.7 |
| ALBACETE | 31 | 1,654,606.5 | 1,115,182.2 | 1,073,719.8 | 13.4 | 4.2 | 2.8 |
| LEON | 31 | 1,528,254.2 | 1,274,531.6 | 1,270,341.6 | 14.5 | 4.2 | 3.2 |
| **PROVINCE** | $n_i$ | $\hat{Y}_i^{DIR}$ | $\hat{Y}_i^{GREG}$ | $\hat{Y}_i^{EBP}$ | cv($\hat{Y}_i^{DIR}$) | cv($\hat{Y}_i^{GREG}$) | cv($\hat{Y}_i^{EBP}$) |
| HUELVA | 32 | 3,031,328.1 | 2,838,874.0 | 2,816,281.3 | 10.5 | 2.6 | 2.0 |
| NAVARRA | 33 | 1,291,343.0 | 956,737.9 | 957,660.4 | 13.2 | 4.4 | 3.4 |
| PONTEVEDRA | 33 | 159,229.1 | 107,198.9 | 138,367.4 | 22.2 | 19.7 | 13.4 |
| VIZCAYA | 34 | 228,618.8 | 183,267.3 | 206,304.6 | 13.1 | 13.2 | 9.1 |
| TOLEDO | 35 | 1,619,939.4 | 1,529,104.8 | 1,539,799.3 | 13.1 | 4.2 | 3.2 |
| CADIZ | 38 | 1,851,521.1 | 1,585,755.9 | 1,620,844.2 | 14.9 | 4.0 | 3.4 |
| BADAJOZ | 39 | 4,571,743.6 | 3,439,625.5 | 3,457,692.5 | 13.5 | 2.7 | 2.2 |
| MALAGA | 39 | 2,499,392.3 | 3,188,031.1 | 3,237,081.8 | 10.9 | 4.2 | 2.5 |
| TARRAGONA | 41 | 2,872,882.0 | 2,690,969.7 | 2,656,117.8 | 11.6 | 2.6 | 2.2 |
| GRANADA | 42 | 2,123,693.3 | 2,221,155.1 | 2,241,916.2 | 12.5 | 3.8 | 2.9 |
| JAEN | 43 | 1,928,229.8 | 1,940,379.2 | 1,943,101.0 | 15.8 | 3.2 | 2.7 |
| ZARAGOZA | 43 | 3,750,210.7 | 2,564,909.0 | 2,578,011.3 | 13.5 | 3.0 | 2.3 |
| GERONA | 45 | 2,029,222.2 | 1,748,165.7 | 1,767,490.3 | 10.4 | 3.2 | 2.5 |
| MURCIA | 51 | 6,700,070.6 | 7,467,465.0 | 7,341,434.6 | 8.7 | 2.2 | 1.6 |
| BALEARES | 52 | 849,950.8 | 650,012.6 | 694,416.3 | 21.5 | 6.1 | 4.7 |
| CANTABRIA | 52 | 285,632.3 | 204,947.7 | 226,163.1 | 10.7 | 9.5 | 6.4 |
| ASTURIAS | 55 | 2,113,034.5 | 1,702,020.8 | 1,661,932.8 | 13.5 | 3.6 | 3.1 |
| CASTELLON | 55 | 1,605,604.4 | 1,526,618.1 | 1,530,394.2 | 8.9 | 2.5 | 2.2 |
| SEVILLA | 55 | 7,458,078.2 | 6,878,368.2 | 6,857,368.8 | 11.0 | 2.0 | 1.7 |
| CORUNA | 62 | 340,200.0 | 217,028.5 | 206,041.8 | 20.2 | 10.9 | 10.2 |
| ALICANTE | 66 | 8,324,589.1 | 8,390,895.3 | 8,240,996.9 | 9.2 | 1.8 | 1.6 |
| VALENCIA | 113 | 7,671,137.7 | 7,209,128.2 | 7,153,290.2 | 6.3 | 1.7 | 1.4 |
| MADRID | 123 | 11,483,342.8 | 12,892,853.8 | 12,892,305.0 | 6.2 | 1.7 | 1.5 |
| BARCELONA | 187 | 22,356,500.5 | 24,990,558.9 | 24,797,372.9 | 4.8 | 1.0 | 0.9 |

# References

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

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

Foster, J., Greer, J. and 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. and Santamaría, L. (2008). Bootstrap mean squared error of a small area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 443-462.

Hájek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, comment on a paper by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart, Winston.

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

Horvitz, D.G., and 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. Technical report, Instituto Nacional de Estadística, España.

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

Molina, I., and 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, Inc.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 1, 15-31. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1985001/article/14364-eng.pdf.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons, Inc.

R Development Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

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

Strzalkowska-Kominiak, E., and Molina, I. (2019). Estimation of proportions in small areas: Application to the labour force using the Swiss Census Structural Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

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

You, Y., and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30, 431-439.