Noname manuscript No.
(will be inserted by the editor)

Model-assisted estimation of small area poverty measures: an application within the Valencia Region in Spain

Domingo Morales · María del Mar Rueda · Dolores Esteban

the date of receipt and acceptance should be inserted later

Abstract This paper introduces small area estimators of some poverty indexes, with special attention to the poverty rate or Head Count Index. The estimators are assisted by nested error regression models and they are model-assisted counterparts of model-based empirical best predictors. The paper studies the sampling-design consistency and the asymptotic normality of the introduced estimators. The results of simulation studies show that the new estimators present a good balance between sampling bias and mean squared error. We use data of the 2013 Spanish living conditions survey from the region of Valencia to explore the performance of the new estimation methods of the poverty rate.

Keywords small area estimation · poverty index · model-assisted estimation · nested error regression model · empirical best predictor

1 Introduction

The analysis of poverty and social exclusion measures is a topic of increased interest to society. For governments is of high interest the estimation of poverty, inequality and life condition indicators. The official poverty rate and the number of people in poverty are important measures of the country's economic

D. Morales
Operations Research Center, Miguel Hernández University of Elche, Spain
E-mail: d.morales@umh.es

M. Rueda
Department of Statistics and Operations Research and IEMath-GR, University of Granada, Spain
E-mail: mrueda@ugr.es

M.D. Esteban
Operations Research Center, Miguel Hernández University of Elche, Spain
E-mail: md.esteban@umh.es

wellbeing. The media and the policy-making circles pay a lot of attention to these sorts of statistics and important policy decisions may be influenced by them.

The Laeken European Council in December 2001 endorsed a first set of 18 common statistical indicators for poverty and social inclusion, which will allow monitoring in a comparable way the progress of Member States towards the agreed European Union objectives. These indicators need to be considered as a consistent whole reflecting a balanced representation of European Union social concerns. They cover four important dimensions of social inclusion (financial poverty, employment, health and education), which highlight the multidimensionality of the phenomenon of social exclusion. The common characteristic of these poverty measures is their complexity. They are nonlinear functions of the observations that cannot be expressed as regular functions of totals (that is, continuously differentiable up to the second order). Some relevant poverty measures are based on a threshold defined in relation to the income distribution. Among these, the *at-risk-of-poverty rate*, also known as the Head Count Index, *the relative mean at-risk-of-poverty gap*, also known as Poverty Gap and the *persistent at risk-of-poverty rate* are included among the primary indicators.

The Head Count Index (HCI) gives a picture of the incidence of poverty and can be calculated as the proportion of persons (or households) with an equivalised disposable income below the 60% of the national median equivalised income. The popularity of this indicator is mostly due to its ease of construction and interpretation. Indeed, the HCI is widely used by institutions to elaborate their reports on poverty. In the literature, numerous references discuss about the HCI and related poverty indicators. For instance, some references are Medeiros (2006), Crettaz and Suter (2013) and Navicke et al. (2014).

The HCI and other poverty indicators at national and regional level are estimated from complex surveys with many thousands of observations, conducted in a harmonized manner over many countries. Usually the method for estimating this index is by using direct estimators. Such estimators depend only on the sample data and are usually obtained by applying standard weighted design-based procedures without using auxiliary information.

Official surveys on income and living conditions generally contain additional variables related to the variable of interest. Such additional variables can be used as auxiliary variables to improve the estimation of poverty indicators by means of regression and calibration procedures (Lehtonen and Veijanen, 2016) and of pseudo empirical likelihood approaches (Rueda and Muñoz, 2011). Muñoz et al. (2015) propose alternative estimation methods for the HCI by using the known ratio and regression techniques after transforming an auxiliary variable related to the variable of interest into a dummy variable.

In the last years there was a worldwide increase in the demand for poverty and living conditions estimates at the local level, since these quantities can help in planning local policies aimed at decreasing poverty and social exclusion. Surveys on income and living conditions are currently conducted in many

countries, but their sample sizes are not large enough to obtain reliable estimates at local level. The direct estimators are appropriate when the sample size in the municipalities or counties is reasonably large but they could be inaccurate when the sample size is small. The small area estimation (SAE) theory deals with this kind of estimation settings. See the monographs of Rao (2003) and Rao and Molina (2015) for an introduction to SAE.

The unit-level model-based approach is commonly used in SAE. The basic SAE unit-level model is the nested error regression (NER) model. Battese et al. (1988) applied this model to the prediction of United States county crop areas using survey and satellite data. Since then, the empirical best linear unbiased predictors (EBLUP) of domains means based on the NER model are being widely applied.

The use of small area estimation methods for the analysis of poverty at the local level has a great potentiality. Without beng exhaustive, we cite some related papers. Molina and Rao (2010) derive empirical best predictors (EBP) of poverty incidences and gaps based on a NER model. Hobza and Morales (2016) proposes EBPs of poverty incidences based on unit-level logit mixed models. Tzavidis et al. (2008) and Marchetti et al. (2012) give M-quantile estimators for poverty mapping. Marchetti and Secondi (2016) use small area methods for obtaining reliable provincial estimates of household consumption expenditure in Italy. Giusti et al. (2016) compute the mean household equivalised income and the head count ratio for the Tuscany region in Italy. Tzavidis et al. (2015) propose a semiparametric approach to model-based small area prediction for estimating the average number of visits to physicians for Health Districts in Central Italy.

Concerning area-level models, Esteban et al. (2012a, 2012b), Marhuenda et al. (2013) and Morales et al. (2015) give EBLUPs of Spanish poverty proportions based on temporal and spatio-temporal linear mixed models. Boubeta et al. (2016, 2017) and López-Vizcaíno et al. (2013, 2015) introduce EBPs of counts and proportions based on logit and Poisson area-level mixed models with applications to Spanish data.

The model-based estimators, when the assumed model is correct, tend to be better than other estimators. However, when the assumed model is incorrect, the model-based estimators are biased and they can do worse than even the naïve estimators. Särndal et al. (1992) presented the model-assisted approach to inference in finite populations, where the superpopulation model is not the basis of the inferences. The model-assisted methodology considers the properties under the design-based distribution, but employs the model to motivate the choice of estimators. Important examples of model-assisted estimators are the generalized regression (GREG) estimator and the calibration estimator introduced by Deville and Särndal (1992). GREG estimation was introduced for domain estimation in Särndal (1981, 1984), Hidiroglou and Särndal (1985) and Särndal and Hidiroglou (1989) and were developed further (including computational tools) in Estevao et al. (1995).

More recently, Lehtonen and Veijanen (2009, 2016) discussed GREG estimators of domain means and proportions and presented empirical studies

based on simulation experiments. Calibration techniques are used in the context of small area for Estevao and Särndal (2004), Chambers (2005) and Chandra and Chambers (2005, 2009). The calibration and regression approaches are built on different arguments. Both are sound in that they yield design-consistent and very nearly design unbiased estimators of the parameter, but they can differ considerably with respect to variance.

It is well-known that calibration, although apparently completely model free, implicitly assumes that a linear regression model well describes the relationship between the variable of interest and the auxiliary variables (see e.g. discussion on this in Wu and Sitter, 2001, and in Montanari and Ranalli, 2005). In the case of poverty measures the variable of interest involves indicator variables and this assumption doesn't seem to be adequate. Lethonen and Veijanen (2012) introduce model calibration methods for estimation of poverty rate for small area. In this approach a logistic regression model is first fitted to the sample and calibration weights are determined using this fitted values instead of the original auxiliary variables. A comparison between generalized regression and model-calibration estimation for domains is given in Lehtonen, Särndal and Veijanen (2008). A model-assisted approach is also used in Fabrizi et al. (2014) which assume a linear M-quantile model at developing design-consistent small area estimators.

This paper introduces a new GREG type estimator of small area HCIs inspired by the Monte Carlo estimation procedure proposed by Molina and Rao (2010). The new estimators are obtained by summing up model-based predicted values and adjusting by design-based weighted sum of residuals. Thus, the model and the sampling design are used in the definition of the estimators.

The article is arranged as follows. Section 2 gives the notation and a brief description of the direct estimation of poverty measures. Section 3 discusses some aspects of the model-based small area estimation approach and presents the EBP of a class of poverty measures under a NER superpopulation model. Section 4 introduces the new model-assisted counterpart of the EBP, studies the design-based properties of the proposed estimator and gives design-based variance estimators. Sections 5 and 6 reports design-based and model-based Monte Carlo simulation experiments to empirically investigate the behavior of the new poverty estimator and of a jackknife variance estimator. The simulation results are in agreement with theoretical findings of Section 4. Section 7 applies the model-assisted estimator of the HCI to a survey data set from the Spanish living conditions survey. Section 8 gives some concluding remarks.

2 Direct estimators of poverty indicators

Let U be a population of size N partitioned into D domains or small areas $U_1, \ldots, U_D$ of sizes $N_1, \ldots, N_D$. Let z_{dj} be a quantitative measure of welfare, such as income or expenditure, for individual j in small area d. The poverty line, κ, is commonly used by many statistical agencies to classify the population

into poor and not poor. This is to say, an individual is considered as poor if its income or expenditure is less than the poverty line. We assume that the poverty line is established by the corresponding authority, i.e. κ is fixed at some official quantity. In some national statistical agencies (as the Spanish Statistical Office) the poverty threshold is set at 60% of the national median equivalised disposable income.

Foster, Greer and Thorbecke (1984) introduced the class of FGT poverty measures

$$\delta_{\alpha d} = \frac{1}{N_d}\sum_{j=1}^{N_d} h_\alpha(z_{dj}), \quad h_\alpha(z_{dj}) = (\frac{\kappa - z_{dj}}{\kappa})^\alpha I(z_{dj} < \kappa), \quad d = 1, \ldots, D, \ (1)$$

where $I(z_{dj} < \kappa) = 1$ if $z_{dj} < \kappa$ and $I(z_{dj} < \kappa) = 0$ otherwise. For $\alpha = 0$ we get the proportion of individuals in small area d, $p_d = \delta_{0d}$, with an equivalised disposable income below the at-risk-at-poverty threshold. This measure is called poverty incidence, poverty proportion, poverty risk or HCI. For $\alpha = 1$ we have the poverty gap measuring the relative average distance to the poverty line for individuals of a given domain d. For $\alpha = 2$ the FGT measure is called poverty severity.

The poverty measures (1) belongs to the more general class of additive parameters

$$\delta_d = \frac{1}{N_d}\sum_{j=1}^{N_d} h(y_{dj}), \quad y_{dj} = T(z_{dj}), \quad d = 1, \ldots, D, \qquad (2)$$

where T is a one-to-one increasing transformation of the income variable and h is a known measurable function. The HCI can be written in the form

$$p_d = \frac{1}{N_d}\sum_{j=1}^{N_d} I(z_{dj} < \kappa) = \frac{1}{N_d}\sum_{j=1}^{N_d} I(y_{dj} < T(\kappa)), \ \ d = 1, \ldots, D. \qquad (3)$$

In practice, the poverty measure $\delta_{\alpha d}$ is unknown and statistical agencies use survey data for estimating it. In the inference process, a random sample s of size $n < N$ is drawn from the population according to a specified sampling design $\pi(s)$. Let $s_1, \ldots, s_D$ be the corresponding domain subsamples of sizes $n_1, \ldots, n_D$, where $n = n_1 + \ldots + n_D$ (note that $n_d = 0$ if an area d is not sampled.) The first and second-order inclusion probabilities are the probabilities of obtaining the unit j and the units j and k of domain d, respectively, while sampling from the population according to the sampling design. They are $\pi_{dj} = \sum_{j \in s_d} \pi(s)$ and $\pi_{djk} = \sum_{j,k \in s_d} \pi(s)$ respectively. A direct estimator of $\delta_{\alpha d}$ for a sampled domain is the unweighted sample mean

$$\overline{\delta}_{\alpha d} = \frac{1}{n_d}\sum_{j=1}^{s_d} h_\alpha(z_{dj}),$$

which uses only the sample data from the target small area. The estimator $\overline{\delta}_{\alpha d}$ is biased. A design-unbiased estimator of $\delta_{\alpha d}$ is the weighted sample mean

$$\hat{\delta}_{\alpha d}^{dir} = \frac{1}{N_d} \sum_{j=1}^{s_d} \pi_{dj}^{-1} h_\alpha(z_{dj}).$$

3 Model-based estimators of poverty indicators

Consider a random vector $\boldsymbol{y} = (y_1, ..., y_N)$ containing the values of a random variable associated with the N units of a finite population. The model-based approach assumes that $\boldsymbol{y}$ follows a superpopulation model that incorporates the auxiliary information $\boldsymbol{x}_{dj} = (x_{dj1}, \ldots, x_{djp})$, $j \in U_d$, $d = 1, \ldots, D$. By using the column and mean operator, we define

$$\boldsymbol{y} = \underset{1\le d\le D}{\text{col}}(\boldsymbol{y}_d),\ \boldsymbol{y}_d = \underset{1\le j\le N_D}{\text{col}}(y_{dj}),\ \ \boldsymbol{X} = \underset{1\le d\le D}{\text{col}}(\boldsymbol{X}_d),$$
$$\boldsymbol{X}_d = \underset{1\le j\le N_d}{\text{col}}(\boldsymbol{x}_{dj}),\quad \boldsymbol{\beta} = \underset{1\le k\le p}{\text{col}}(\beta_k).$$

The NER superpopulation model is

$$y_{dj} = \boldsymbol{x}_{dj}\boldsymbol{\beta} + u_d + e_{dj},\quad d = 1, \ldots, D,\ j = 1, \ldots, N_d, \tag{4}$$

where the random effects $\{u_d\}$ and the errors $\{e_{dj}\}$ are mutually independent with $u_d \sim N(0, \sigma_u^2)$ and $e_{dj} \sim N(0, \sigma_e^2)$. Let us define $\boldsymbol{e}_d = \underset{1\le j\le N_D}{\text{col}}(e_{dj})$. Then, the model (4) can be written as

$$\boldsymbol{y}_d = \boldsymbol{X}_d\boldsymbol{\beta} + u_d + \boldsymbol{e}_d,\quad d = 1, \ldots, D.$$

The vectors $\boldsymbol{y}_d$ are independent with $\boldsymbol{y}_d \sim N(\boldsymbol{\mu}_d, \boldsymbol{V}_d)$, $\boldsymbol{\mu}_d = \boldsymbol{X}_d\boldsymbol{\beta}$ and $\boldsymbol{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}'_{N_d} + \sigma_e^2 \boldsymbol{I}_{N_d}$, where $\mathbf{1}_K = \underset{1\le j\le K}{\text{col}}(1)$ and $\boldsymbol{I}_K = \underset{1\le j\le K}{\text{diag}}(1)$ are the 1-column vector and the identity matrix of sizes K and $K \times K$ respectively.

Let $\boldsymbol{y}_{ds}$ be the sub-vector of $\boldsymbol{y}_d$ corresponding to sample elements and $\boldsymbol{y}_{dr}$ the sub-vector of $\boldsymbol{y}_d$ corresponding to the out-of-sample elements. Without lack of generality, we can write $\boldsymbol{y}_d = \text{col}(\boldsymbol{y}_{ds}, \boldsymbol{y}_{dr}) = (\boldsymbol{y}'_{ds}, \boldsymbol{y}'_{dr})'$, where $r = U - s$ is the set of indexes of the units that are not sampled (with size $N - n$). Define also $\text{col}(\boldsymbol{X}_{ds}, \boldsymbol{X}_{dr})$ and $\text{diag}(\boldsymbol{V}_{ds}, \boldsymbol{V}_{dr})$ as the corresponding decompositions of $\boldsymbol{X}_d$ and $\boldsymbol{V}_d$. The sample vector $\boldsymbol{y}_{ds}$ follows the corresponding submodel of (4), i.e.

$$y_{dj} = \boldsymbol{x}_{dj}\boldsymbol{\beta} + u_d + e_{dj},\quad d = 1, \ldots, D,\ j = 1, \ldots, n_d, \tag{5}$$

where we change N and N_d by the sample counterparts n and n_d respectively. When $\sigma_e^2 > 0$ and $\sigma_u^2 > 0$ are known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) of u_d, $d = 1, \ldots, D$, are

$$\tilde{\boldsymbol{\beta}} = \Big(\sum_{d=1}^{D} \boldsymbol{X}'_{ds}\boldsymbol{V}_{ds}^{-1}\boldsymbol{X}_{ds}\Big)^{-1} \sum_{d=1}^{D} \boldsymbol{X}'_{ds}\boldsymbol{V}_{ds}^{-1}\boldsymbol{y}_{ds},\ \tilde{u}_d = \sigma_u^2 \mathbf{1}'_{n_d}\boldsymbol{V}_{ds}^{-1}\left(\boldsymbol{y}_{ds} - \boldsymbol{X}_{ds}\tilde{\boldsymbol{\beta}}\right). \tag{6}$$

Replacing σ_e^2 and σ_u^2 by estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ in (6), the empirical BLUE (EBLUE) of $\boldsymbol{\beta}$ and the EBLUP of u_d, $d = 1, \ldots, D$, are

$$\hat{\boldsymbol{\beta}} = \Big(\sum_{d=1}^{D} \boldsymbol{X}'_{ds} \hat{\boldsymbol{V}}_{ds}^{-1} \boldsymbol{X}_{ds} \Big)^{-1} \sum_{d=1}^{D} \boldsymbol{X}'_{ds} \hat{\boldsymbol{V}}_{ds}^{-1} \boldsymbol{y}_{ds}, \;\; \hat{u}_d = \hat{\sigma}_u^2 \mathbf{1}'_{n_d} \hat{\boldsymbol{V}}_{ds}^{-1} \Big(\boldsymbol{y}_{ds} - \boldsymbol{X}_{ds} \hat{\boldsymbol{\beta}} \Big), \tag{7}$$

where $\hat{\boldsymbol{V}}_d = \hat{\sigma}_u^2 \mathbf{1}_{n_d} \mathbf{1}'_{n_d} + \hat{\sigma}_e^2 \boldsymbol{I}_{n_d}$. The distribution of $\boldsymbol{y}_{dr}$, given the sample data $\boldsymbol{y}_s$, is

$$\boldsymbol{y}_{dr} | \boldsymbol{y}_s \sim \boldsymbol{y}_{dr} | \boldsymbol{y}_{ds} \sim N(\boldsymbol{\mu}_{dr|s}, \boldsymbol{V}_{dr|s}), \tag{8}$$

where

$$\boldsymbol{\mu}_{dr|s} = \boldsymbol{X}_{dr} \boldsymbol{\beta} + \sigma_u^2 \mathbf{1}_{N_d - n_d} \mathbf{1}'_{n_d} \boldsymbol{V}_{ds}^{-1} (\boldsymbol{y}_{ds} - \boldsymbol{X}_{ds} \boldsymbol{\beta}),$$

$$\boldsymbol{V}_{dr|s} = \sigma_u^2 (1 - \gamma_d) \mathbf{1}_{N_d - n_d} \mathbf{1}'_{N_d - n_d} + \sigma_e^2 \boldsymbol{I}_{N_d - n_d}, \quad \gamma_d = \frac{n_d \sigma_u^2}{n_d \sigma_u^2 + \sigma_e^2}.$$

For any $j \in r_d = U_d - s_d$, the components of $\boldsymbol{\mu}_{dr|s}$ and the diagonal elements of $\boldsymbol{V}_{dr|s}$ are

$$\mu_{dj|s} = \begin{cases} \boldsymbol{x}_{dj} \boldsymbol{\beta} + \gamma_d (\bar{y}_{ds} - \bar{\boldsymbol{x}}_{ds} \boldsymbol{\beta}) & \text{if } n_d \neq 0, \\ \boldsymbol{x}_{dj} \boldsymbol{\beta} & \text{if } n_d = 0, \end{cases} \quad v_{d|s} = \begin{cases} \sigma_u^2 (1 - \gamma_d) + \sigma_e^2 & \text{if } n_d \neq 0, \\ \sigma_u^2 + \sigma_e^2 & \text{if } n_d = 0, \end{cases}$$

where $\bar{y}_{ds} = n_d^{-1} \sum_{j=1}^{n_d} y_{dj}$ and $\bar{\boldsymbol{x}}_{ds} = n_d^{-1} \sum_{j=1}^{n_d} \boldsymbol{x}_{dj}$.

If we assume that a one-to-one increasing transformation of the income variable, $y_{dj} = T(z_{dj})$, follows the NER model (4), then the best predictor (BP) of an additive parameter δ_d, defined in (2), is its expectation with respect to the distribution (8) of the non sample data $\boldsymbol{y}_{dr}$, given the sample data $\boldsymbol{y}_s$, i.e.

$$\hat{\delta}_d^B = E_{\boldsymbol{y}_{dr}} \Big[\frac{1}{N_d} \sum_{j=1}^{N_d} h(y_{dj}) \big| \, \boldsymbol{y}_s \Big] = \frac{1}{N_d} \Big\{ \sum_{j \in s_d} h(y_{dj}) + \sum_{j \in r_d} E_{\boldsymbol{y}_{dr}} \big[h(y_{dj}) | \boldsymbol{y}_s \big] \Big\}.$$

The conditional distribution (8) depends on the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ of unknown model parameters, which must be estimated, that is,

$$E_{\boldsymbol{y}_{dr}} \left[h(y_{dj}) | \boldsymbol{y}_s \right] = E_{\boldsymbol{y}_{dr}} \left[h(y_{dj}) | \boldsymbol{y}_s ; \boldsymbol{\theta} \right].$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ be an estimator based on the sample data $\boldsymbol{y}_s$. The EBP of δ_d, introduced by Molina and Rao (2010), is

$$\hat{\delta}_d^{eb} = \frac{1}{N_d} \Big\{ \sum_{j \in s_d} h(y_{dj}) + \sum_{j \in r_d} E_{\boldsymbol{y}_{dr}} \big[h(y_{dj}) | \boldsymbol{y}_s ; \hat{\boldsymbol{\theta}} \big] \Big\}. \tag{9}$$

Let $\hat{\mu}_{dj|s} = \boldsymbol{x}_{dj} \hat{\boldsymbol{\beta}} + \hat{\gamma}_d (\bar{y}_{ds} - \bar{\boldsymbol{x}}_{ds} \hat{\boldsymbol{\beta}})$, $\hat{v}_{d|s} = \hat{\sigma}_u^2 (1 - \hat{\gamma}_d) + \hat{\sigma}_e^2$ and $\hat{\gamma}_d = \frac{n_d \hat{\sigma}_u^2}{n_d \hat{\sigma}_u^2 + \hat{\sigma}_e^2}$ be the plug-in estimators of $\hat{\mu}_{dj|s}$, $\hat{v}_{d|s}$ and $\hat{\gamma}_d$ respectively. The EBP of p_d is

$$\hat{p}_d^{eb} = \frac{1}{N_d} \Big\{ \sum_{j \in s_d} h(y_{dj}) + \sum_{j \in r_d} \hat{p}_{dj}^{eb} \Big\} = \frac{1}{N_d} \Big\{ \sum_{j \in s_d} I(y_{dj} < T(\kappa)) + \sum_{j \in r_d} \Phi(t_{dj}) \Big\}, \tag{10}$$

where $t_{dj} = v_{d|s}^{-1/2}(T(\kappa) - \mu_{dj|s})$, $\Phi(\cdot)$ is the c.d.f. of a standard normal random variable and

$$\hat{p}_{dj}^{eb} = E_{\boldsymbol{y}_{dr}}\left[I(y_{dj} < T(\kappa))|\boldsymbol{y}_s; \hat{\boldsymbol{\theta}}\right] =$$

$$P_{\boldsymbol{y}_{dr}}\big(y_{dj} < T(\kappa)|\boldsymbol{y}_s; \hat{\boldsymbol{\theta}}\big) = P\big(N(0,1) < t_{dj}\big) = \Phi(t_{dj}).$$

For a general function h, the expectation in (9) might be not tractable analytically. When this occurs, Molina and Rao (2015) applies the following Monte Carlo procedure.

(a) Estimate the unknown parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ using sample data $\boldsymbol{y}_s$.
(b) Replacing $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ by the estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ obtained in (a), draw L copies of each non-sample variable y_{dj} as

$$y_{dj}^{(\ell)} \sim N(\hat{\mu}_{dj|s}, \hat{v}_{d|s}), \quad j \in r_d,\ d = 1, \ldots, D,\ \ell = 1, \ldots, L.$$

(c) The Monte Carlo approximation of the expected value is

$$E_{\boldsymbol{y}_{dr}}\left[h(y_{dj})|\boldsymbol{y}_s; \hat{\boldsymbol{\theta}}\right] \approx \frac{1}{L}\sum_{\ell=1}^{L} h(y_{dj}^{(\ell)}), \quad j \in r_d,\ d = 1, \ldots, D.$$

The Monte Carlo approximation of the EBP of δ_d is

$$\hat{\delta}_d^{eb} \approx \frac{1}{L}\sum_{\ell=1}^{L} \delta_d^{(\ell)}, \quad \delta_d^{(\ell)} = \frac{1}{N_d}\left[\sum_{j \in s_d} h(y_{dj}) + \sum_{j \in r_d} h(y_{dj}^{(\ell)})\right]. \tag{11}$$

Based on the logistic mixed model

$$I(y_{dj} < T(\kappa)) \sim \text{Bin}(1, p_{dj}),\ \text{logit}(p_{dj}) = \boldsymbol{x}_{dj}\boldsymbol{\beta} + u_d,\ d = 1, \ldots, D,\ j = 1, \ldots, N_d,$$

where $u_1, \ldots, u_d$ are i.i.d. $N(0, \sigma_u^2)$, we may also consider the model-based domain proportion estimator

$$\hat{p}_d^L = \frac{1}{N_d}\Big\{\sum_{j \in s_d} I(y_{dj} < T(\kappa)) + \sum_{j \in r_d} \hat{p}_{dj}^L\Big\}, \quad \hat{p}_{dj}^L = \frac{\exp\{\boldsymbol{x}_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d\}}{1 + \exp\{\boldsymbol{x}_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d\}}, \tag{12}$$

where $\hat{\beta}$ and $\hat{u}_d$ are the ML estimator and predictor of $\boldsymbol{\beta}$ and u_d respectively obtained by applying the Laplace approximation algorithm. The estimator (12) was applied by Hobza and Morales (2016) to estimate poverty proportions of counties in the Spanish region of Valencia.

4 Model-assisted estimation of poverty measures

The model-assisted Monte Carlo estimator of δ_d can be calculated by the steps (a)-(c) given in Section 3. The model-assisted counterpart of $\hat{\delta}_d^{eb}$, defined in (11), is the Monte Carlo estimator

$$\hat{\delta}_d^{ma} \approx \frac{1}{L}\sum_{\ell=1}^{L} \delta_d^{ma(\ell)},\ \delta_d^{ma(\ell)} = \frac{1}{N_d}\left[\sum_{j\in U_d} h(y_{dj}^{(\ell)}) + \sum_{j\in s_d} \pi_{dj}^{-1}\{h(y_{dj}) - h(y_{dj}^{(\ell)})\}\right]. \tag{13}$$

As a counterpart of $\hat{p}_d^{eb}$, defined in (10), the model assisted estimators of p_d is

$$\hat{p}_d^{ma} = \frac{1}{N_d}\Big(\sum_{j\in U_d} \Phi(t_{dj}) + \sum_{j\in s_d} \pi_{dj}^{-1}\{I\big(y_{dj} < T(\kappa)\big) - \Phi(t_{dj})\}\Big). \tag{14}$$

Following Lehtonen and Veijanen (1998), a model-assisted counterpart of $\hat{p}_d^{L}$, defined in (12), is

$$\hat{p}_d^{LM} = \frac{1}{N_d}\Big(\sum_{j\in U_d} \hat{p}_{dj}^{L} + \sum_{j\in s_d} \pi_{dj}^{-1}\{I\big(y_{dj} < T(\kappa)\big) - \hat{p}_{dj}^{L}\}\Big). \tag{15}$$

For studying the design-based properties of properties of the model-assisted estimator (13), we rely on the properties of the Horvitz-Thompson (HT) design-unbiased estimator of a population mean $\overline{Y_d} = N_d^{-1}\sum_{j\in U_d} y_{dj}$. Horvitz and Thompson (1952) introduced the estimator

$$\hat{\overline{Y}}_d^{HT} = N_d^{-1}\sum_{j\in s_d} \pi_{dj}^{-1} y_{dj}.$$

The HT estimator is admissible within the class of all unbiased estimators, but makes no use of auxiliary information. From this property, it follows that the model assisted estimators (13)-(15) are also design-based unbiased.

Hájek, J. (1971) proposed the ratio estimator

$$\hat{\overline{Y}}_d^{dir} = \frac{1}{\hat{N}_d^{dir}}\sum_{j\in s_d} \pi_{dj}^{-1} y_{dj}, \quad \hat{N}_d^{dir} = \sum_{j\in s_d} \pi_{dj}^{-1},$$

and Särndal et al. (1992, p. 182) gave several reasons for regarding the Hájek as usually better than the HT estimator. However, it is common to assume that there exists some auxiliary variables (to be obtained from census results, administrative files, etc.) that can be essential for efficient estimation of domain means. The model-assisted estimators (13) incorporate the auxiliary information by using models and preserving good design-based properties.

For analyzing the design-based asymptotic behavior of estimator (13), we introduce some notation and we consider some new hypothesis. Apart from the mathematical interest, the asymptotic results of Theorem 4.1 might be applicable to domains where the sample sizes are large. Although domain sample sizes are generally small in the SAE framework, in many real data

cases we can find some domains with large enough sample sizes. This is why, the asymptotic theory still has some practical interest in SAE.

The study of the asymptotic design-based properties of small area estimator has been previously considered in other works as Estevao and Särndal (2004) or Fabrizi et al. (2014). These authors follow Isaki and Fuller (1982) where the properties of estimators are established under a fixed sequence of populations and a corresponding sequence of random sampling designs. We also used this asymptotic framework. In particular, the finite population U and the sampling design $\pi(\cdot)$ are embedded into a sequence of such populations and designs indexed by N, $\{U_N, \pi_N(\cdot)\}$, with $N \to \infty$. We will assume, thus, that N_N tend to infinity and that n_N also tend to infinity when $N \to \infty$.

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ be an estimator of the superpopulation parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2)$, $\boldsymbol{\beta} \in R^p$, $\sigma_u^2 > 0$, $\sigma_e^2 > 0$. Let $\boldsymbol{\theta}_N = (\boldsymbol{\beta}_N, \sigma_{Nu}^2, \sigma_{Ne}^2)$ be the corresponding estimator of $\boldsymbol{\theta}$ based on the data from the entire population. For $\boldsymbol{\eta} = (\boldsymbol{\gamma}, \tau_u, \tau_e)$, $\boldsymbol{\gamma} \in R^p$, $\tau_u > 0$, $\tau_e > 0$, we define

$$m_{dj}(\boldsymbol{\eta}) = \boldsymbol{x}_{dj}\boldsymbol{\gamma} + u_{dN}(\boldsymbol{\eta}),\ \sigma_{dj}^2(\boldsymbol{\eta}) = \tau_u\big(1-\gamma_d(\boldsymbol{\eta})\big)+\tau_e,\ \gamma_d(\boldsymbol{\eta}) = \tau_u\big(\tau_u+\tau_e/n_d\big)^{-1},$$

where

$$u_{dN}(\boldsymbol{\eta}) = \tau_u \mathbf{1}'_{N_d} \boldsymbol{V}_d^{-1}(\tau_u, \tau_e)\left(\boldsymbol{y}_d - \boldsymbol{X}_d\boldsymbol{\gamma}\right),\ \boldsymbol{V}_d(\tau_u, \tau_e) = \tau_u \mathbf{1}_{N_d}\mathbf{1}'_{N_d} + \tau_e \boldsymbol{I}_{N_d}.$$

We also define

$$g_{dj}(\boldsymbol{\eta}) = \int_{-\infty}^{+\infty} h(y) f_{N(0,1)}\big((y - m_{dj}(\boldsymbol{\eta}))/\sigma_{dj}(\boldsymbol{\eta})\big)\, dy,$$

$$\delta_d^{ma}(\boldsymbol{\eta}) = \frac{1}{N_d}\left[\sum_{j\in U_d} g_{dj}(\boldsymbol{\eta}) + \sum_{j\in s_d} \pi_{dj}^{-1}\big(h(y_{dj}) - g_{dj}(\boldsymbol{\eta})\big)\right].$$

In order to prove our results, we make a set of technical assumptions reported in the Appendix.

Theorem 4.1. Under Assumptions A1 to A8, the model-assisted estimator $\hat{\delta}_d^{ma}$ is design-consistent for δ_d and has an asymptotic normal distribution with asymptotic mean and variance

$$AE_\pi = \delta_d = \frac{1}{N_d}\sum_{j=1}^{N_d} h(y_{dj}), \quad AV_\pi(\hat{\overline{Y}}_d^{ma}) = \frac{1}{N_d^2}\sum_{j\in U_d}\sum_{k\in U_d} \Delta_{djk}\pi_{dj}^{-1}\pi_{dk}^{-1}\epsilon_{dj}\epsilon_{dk},$$

where $\Delta_{djk} = \pi_{djk} - \pi_{dj}\pi_{dk}$, $\epsilon_{dj} = h(y_{dj}) - g_{dj}(\boldsymbol{\theta})$.

Proof.

We may write the the model-assisted estimator $\hat{\delta}_d^{ma}$ as

$$\hat{\delta}_d^{ma} = \hat{\delta}_d^{ma}(\hat{\boldsymbol{\theta}}) = \frac{1}{N_d}\Big(\sum_{j\in s_d}\pi_{dj}^{-1}h(y_{dj}) - \sum_{j\in s_d}\pi_{dj}^{-1}g_{dj}(\hat{\boldsymbol{\theta}})\Big) + \frac{1}{N_d}\sum_{j\in U_d} g_{dj}(\hat{\boldsymbol{\theta}}).$$

Under Assumptions A1-A8 and according to Theorem 1 in Wang and Opsomer (2011), the sample estimator $\sum_{j\in s_d} \pi_{dj}^{-1} g_{dj}(\hat{\boldsymbol{\theta}})$ is design consistent for $\sum_{j\in U_d} g_{dj}(\boldsymbol{\theta})$ and asymptotically normally distributed. In the same way, the sample estimator $\sum_{j\in s_d} \pi_{dj}^{-1} h(y_{dj})$ is design consistent for $\sum_{j\in U_d} h(y_{dj})$ and asymptotically normally distributed under regularity conditions A3-A6 (see Mukhopadhyay (2001), chapter 6). Therefore the asymptotic distribution of $\hat{\delta}_d^{ma}$ follows. The asymptotic mean and variances of the estimator is thus obtained from A6.

We can obtain estimators of the asymptotic variance of $AV_\pi(\hat{\delta}_d^{ma})$ by using the properties of HT estimators. Letting $\widehat{\epsilon}_{dj} = h(y_{dj}) - g_{dj}(\hat{\boldsymbol{\theta}})$, we can define the simple estimator of the variance based on the weighted residual variance estimator

$$\widehat{V}\left(\hat{\delta}_d^{ma}\right) = \frac{1}{N_d^2} \sum_{j\in s_d} \sum_{k\in s_d} \frac{\Delta_{djk}}{\pi_{djk}} \pi_{dj}^{-1} \pi_{dk}^{-1} \widehat{\epsilon}_{dj} \widehat{\epsilon}_{dk} \tag{16}$$

for $AV_\pi(\hat{\delta}_d^{ma})$. The estimator is a first-order approximation because 16 does not take into account both of the variability due to the estimation of $\boldsymbol{\theta}$ and underestimates the design variance. Wang and Opsomer (2011) derive a better variance expression for nondifferentiable survey estimators with estimated parameters by using the differentiable limit for T_N, however, this expression is difficult to obtain in practice.

On the other hand, these variance estimators are based on asymptotic properties and require knowledge of the second order inclusion probabilities which often are impossible to compute or unavailable to data analysts for complex sampling designs. Therefore, they have limited use in the small area estimation setup. A simple alternative is the use of with replacement variance estimators (see Särndal et al. (1992), p. 99) or replicated sampling methods (see Wolter (2007) for a detailed description of these techniques in finite population sampling). The replicated methods, also referred to as resampling methods, include the balanced repeated replication, the jackknife repeated replication and the bootstrap method (see Tukey (1958) and Efron (1979)).

Quenouille (1949) introduced the jackknife method to estimate the bias of an estimator by deleting one datum each time from the original data set and recalculating the estimator based on the rest of the data. Tukey (1958) found that the jackknife can also be used to construct variance estimators. Miller (1964) gave the first theorem concerning the jackknife variance estimator. Rao and Tasui (2004) described jackknife variance estimators under stratified multistage sampling. Herrador et al. (2008) investigated resampling methods for estimating design-based variances of model-based and model-assisted small area estimators in a complex survey sampling setup. In survey sampling it is usual to use jackknife techniques due to their simplicity and because they are implemented in general purpose software packages, such as R. See for example the packages "sampling" by Tillé and Matei (2015), "samplingVarEst" by Escobar and Barrios (2016) or "samplingEstimates" by Escobar (2014).

In order to apply the jackknife for design-based variance estimation, we use the delete-one-cluster jackknife method (see e.g. Rao and Tausi, 2004). To

obtain the delete-one-cluster jackknife variance estimator of $\hat{\theta}_d$, we generate jackknife samples by deleting a cluster each time. There are as many jackknife samples as clusters are in the sample. Consider the jackknife sample, $s^*_{(d_*c_*)}$, obtained by excluding the cluster c_* of the domain d_* from the sample s and denote the corresponding domain subsample by $s^*_{d(d_*c_*)}$. Let D_s be the number of domains in s, m_d be the number of clusters in s_d and $C = \sum_{d=1}^{D} m_d$. The jackknife weight of individual j, cluster c and domain d in $s^*_{(d_*c_*)}$ is

$$w_{dcj(d_*,c_*)} = w_{dcj} b_{dc(d_*,c_*)}, \quad b_{dc(d_*,c_*)} = \begin{cases} w_{d.}/w^*_{d.} & \text{if } d = d_*, c \neq c_*, \\ 1 & \text{if } d \neq d_*, \end{cases}$$

where $w_{d.} = \sum_{c=1}^{m_d} \sum_{j \in s_d} w_{dcj}$ and $w^*_{d.} = \sum_{c=1, c\neq c_*}^{m_d} \sum_{j \in s^*_{d(d_*c_*)}} w_{dcj}$. Note that the case $d = d_*$ and $c = c_*$ does not appear in the jackknife sample $s^*_{(d_*c_*)}$. The jackknife resampling method is done as follows:

1. By using the procedure described above, use sample s to draw jackknife samples $s^*_{(d_*c_*)}$, $d_* = 1, \ldots, D_s$, $c_* = 1, \ldots, m_{d_*}$. For every jackknife sample calculate $\hat{\theta}^*_{d(d_*,c_*)}$ in the same way as $\hat{\theta}_d$ was calculated, but using the jackknife weights $w_{dcj(d_*,c_*)}$.
2. The observed distribution of $\{\hat{\theta}^*_{d(d_*c_*)} : d_* = 1, \ldots, D_s, c_* = 1, \ldots, m_d\}$ is expected to imitate the distribution of estimator $\hat{\theta}_d$.
3. The jackknife estimator of θ_d and of bias$(\hat{\theta}_d)$ are

$$\hat{\theta}_{Jd} = \frac{1}{C} \sum_{d_*=1}^{D_s} \sum_{c_*=1}^{m_{d_*}} \hat{\theta}^*_{d(d_*c_*)}, \quad \text{bias}_J(\hat{\theta}_d) = \sum_{d_*=1}^{D_s} (m_{d_*} - 1)\big(\hat{\theta}^*_{d(d_*c_*)} - \hat{\theta}_{Jd}\big). \tag{17}$$

4. The design-based variance of $\hat{\theta}_d$ can be approximated by

$$\text{var}_J(\hat{\theta}_d) = \sum_{d_*=1}^{D_s} \frac{m_{d_*} - 1}{m_{d_*}} \sum_{c_*=1}^{m_{d_*}} \big(\hat{\theta}^*_{d(d_*c_*)} - \hat{\theta}_{Jd}\big)^2. \tag{18}$$

5 Design-based Simulations

This section presents an artificial population and two desig-based simulation experiments. The target parameter is the domain HCI. Simulation 1 is designed to compare the model-assisted estimators $\hat{p}^{ma}_d$ and $\hat{p}^{ML}_d$ with their model-based counterparts $\hat{p}^{eb}_d$ and $\hat{p}^{L}_d$. Simulation 2 studies the behavior of the jackknife variance estimator 18. Our simulations are programmed in `R` and uses the *sae* package (Molina and Marhuenda, 2015).

Similarly to Santamaría et al. (2004), we construct a nested artificial population of size $N = 40000$ divided in $D = 50$ domains. Each domain is partitioned into 20 clusters and each cluster contains 40 units. Auxiliary variables are drawn from normal distributions $x_0 \sim N(10, 1)$ and $x_1 \sim N(\mu_d, 1)$, with means $\mu_d = 8 + 4(d-1)/D$, $d = 1, \ldots, D$.

The y-variables, y_0, y_1 and y_{01}, are drawn from the linear mixed models

$$M_k : y_{k,dj} = \beta_1 x_{k,dj} + u_d + e_{dj}, \quad k = 0, 1,\ d = 1, \ldots, D,\ j = 1, \ldots, N_d,$$
$$M_{01} : y_{01,dj} = \beta_1 x_{0,dj} + \beta_2 x_{1,dj} + u_d + e_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d,$$

where $\beta_1 = 2$, $\beta_2 = 2$, the random effects are i.i.d. $u_d \sim N(0, \sigma_u^2)$ with $\sigma_u^2 = 1$, the random errors are i.i.d. $e_{dj} \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = 1$, and they are all independent.

The income variables, z_0, z_1 and z_{01}, are derived from the transformations

$$I_k : z_{k,dj} = (y_{k,dj})^4,\ I_{01} : z_{01,dj} = (y_{01,dj})^4,\ k = 0, 1, d = 1, \ldots, D, j = 1, \ldots, N_d.$$

For each variable $z \in \{z_0, z_1, z_{01}\}$, the associated target variable is $\xi \in \{\xi_0, \xi_1, \xi_{01}\}$, with $\xi_{dj} = I(z_{dj} < \kappa)$, and their poverty lines and global poverty proportions,

$$\kappa = 0.6 \times \text{median}\{z_{dj} : d = 1, \ldots, D,\ j = 1, \ldots, N_d\} \quad \text{and} \quad p = \frac{1}{N} \sum_{d=1}^{D} \sum_{j=1}^{N_d} \xi_{dj},$$

are $\kappa = 127253$, $\kappa = 129595$, $\kappa = 1795079$ and $p = 0.1540$, $p = 0.2415$, $p = 0.1132$ respectively. For each variable $z \in \{z_0, z_1, z_{01}\}$, the target parameters are the domain proportions $p_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \xi_{dj}$.

Simulation 1 calculates empirical biases and MSEs of small area estimators of poverty proportions. The simulation is carried out under the Bernoulli sampling design with logistic inclusion probabilities

$$\pi_{dj} = \frac{\exp\{b_0 + b_1 x_{0,dj}\}}{1 + \exp\{b_0 + b_1 x_{0,dj}\}}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d, \tag{19}$$

where $b_1 = 1$ and $b_0 = 14$ (Scenario 1) or $b_0 = 13$ (Scenario 2) for "very small" or "small" domain sample sizes. Under the Bernoulli sampling design, each population unit j from any domain d enters in the sample, independently of any other, with probability π_{dj}. As the inclusion probabilities depends on the variable x_0, the Bernoulli sampling design is informative for the variables x_0, y_0, y_{01}, ξ_0 and ξ_{01}. Models containing the auxiliary variable x_0 transfer sampling design information to the model-based and model-assisted estimators. This is done through the regression equation. In this case, the survey-weight summands

$$\sum_{j \in s_d} \pi_{dj}^{-1} \big\{ I\big(y_{dj} < T(\kappa)\big) - \Phi(t_{dj}) \big\}, \quad \sum_{j \in s_d} \pi_{dj}^{-1} \big\{ I\big(y_{dj} < T(\kappa)\big) - \hat{p}_{dj}^{L} \big\}$$

appearing in the formulas of the model-assisted estimators, play in models with x_0 a less relevant role for decreasing the design-based bias than in models without x_0.

The expected sample size $n_E = \sum_{d=1}^{D} \sum_{j=1}^{N_d} \pi_{dj}$ for Scenarios 1 and 2 are 1126.95 and 2776.43 respectively. Table 5.1 presents the quartiles and the mean of the expected domain sample sizes $n_{Ed} = \sum_{j=1}^{N_d} \pi_{dj}$ and of the sampling weights $w_{dj} = 1/\pi_{dj}$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$.

Scenario	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	1.67	28.90	55.56	91.30	108.70	2500.00
2	1.24	11.27	21.10	34.21	40.65	1000.00
1	20.50	22.07	22.45	22.54	22.95	25.48
2	51.11	54.35	55.24	55.53	56.54	61.86

Table 5.1. Quartiles of $\{w_{dj}\}$ (top) and $\{n_{Ed}\}$ (bottom).

For better understanding the results of the simulation, we recall that the direct estimator $\hat{p}_d^{dir}$ of p_d is a ratio estimator such that

$$\hat{p}_d^{dir} = \frac{\hat{\xi}_d^{dir}}{\hat{N}_d^{dir}}, \quad \hat{\xi}_d^{dir} = \sum_{j \in s_d} w_{dj}\xi_{dj}, \quad \hat{N}_d^{dir} = \sum_{j \in s_d} w_{dj},$$

$$\mathrm{cov}_\pi(\hat{p}_d^{dir}, \hat{N}_d^{dir}) = -N_d\big(E_\pi[\hat{p}_d^{dir}] - p_d\big)$$

$$B_\pi\big(\hat{p}_d^{dir}\big) = E_\pi\big[\hat{p}_d^{dir}\big] - p_d = -\frac{\mathrm{cov}_\pi\big(\hat{p}_d^{dir}, \hat{N}_d^{dir}\big)}{N_d}.$$

Table 5.2 presents the empirical covariances $\mathrm{cov}_\pi\big(\hat{p}_{a,d}^{dir}, \hat{N}_d^{dir}\big)$ for $a = 0, 1, 01$, $d = 1, 20, 35, 50$, calculated by running the Simulation 1 under the Scenario 2. The column "bias" shows the sign of the direct estimator bias that we expect to obtain in the output of Simulation 1. The symbols "$-$" and "0" say that the direct estimator has negative and almost null bias respectively.

a	$d=1$	$d=20$	$d=35$	$d=50$	bias
$\hat{p}_{0,d}^{dir}$	8.813	8.769	6.413	7.612	$-$
$\hat{p}_{1,d}^{dir}$	-0.34	1.472	-0.012	-0.049	0
$\hat{p}_{01,d}^{dir}$	7.722	6.162	1.643	-0.021	$-$

Table 5.2. Covariances between $\hat{p}_{a,d}^{dir}$ and $\hat{N}_d^{dir}$.

Simulation 1 calculates the EB and MA estimators of domain poverty proportions based or assisted by the income variable models

$$I(z \,|\, x): \; z_{dj} = (y_{dj})^4, \quad y_{dj} \in M(y \,|\, x), \quad d = 1, \ldots, D, j = 1, \ldots, N_d.$$

If the target and the auxiliary variable are $z = z_k$ and $x = x_k$, $k \in \{0, 1\}$, then the EB and MA estimators are constructed by using the model that generates the population. This is to say, the true model $I_k = I(z_k|x_k)$ is employed. If $z = z_{k_1}$ or $z = z_{01}$ and $x = x_{k_2}$, with $k_1 \neq k_2$, then the estimators are based or assisted by incorrect or incomplete models. By using the same auxiliary information, Simulation 1 also calculates the L and LM estimators (12) and (15) based or assisted by the corresponding logistic mixed models fitted to the dichotomic variables ξ_{dj}'s. The steps of Simulation 1 are

1. For $i = 1, \ldots, I$ ($I = 10^3$), draw a Bernoulli sample from the population and calculate $\hat{p}_d^i \in \big\{\hat{p}_d^{dir,i}, \hat{p}_d^{eb,i}, \hat{p}_d^{ma,i}, \hat{p}_d^{L,i}, \hat{p}_d^{ML,i}\big\}$.
2. Calculate the empirical bias and MSE, i.e.

$$BIAS_d = \frac{1}{I}\sum_{i=1}^{I}\big\{\hat{p}_d^i - p_d\big\}, \;\; MSE_d = \frac{1}{I}\sum_{i=1}^{I}\big(\hat{p}_d^i - p_d\big)^2, \;\; d = 1, \ldots, D.$$

3. Calculate also $ABIAS = \frac{1}{D}\sum_{d=1}^{D}\left|BIAS_d\right|$ and $MSE = \frac{1}{D}\sum_{d=1}^{D} MSE_d$.

Tables 5.3-5.6 present the empirical average absolute biases, $ABIAS$, and the empirical average mean squared errors MSE. the lowest values are printed in bold characters. The tables are divided in three parts. The first part concerns the results for the correct models $I(z_0 \,|\, x_0)$ and $I(z_1 \,|\, x_1)$. The second part gives the simulation results of the incorrect model $I(z_1 \,|\, x_0)$ and the incomplete model $I(z_{01}|\, x_0)$. In these four cases, the EB estimator has the greatest bias and the L and EB estimators have the lowest MSE. The MA and LM estimators have the lowest bias and they have lower MSE than the direct estimator. See also Figures 5.1 and 5.2 (for the sake of brevity we only present boxplots for Scenario 2).

The third part gives the simulation results of the incorrect model $I(z_0 \,|\, x_1)$ and the incomplete model $I(z_{01}|\, x_1)$. In this part, the model-based and model-assisted estimators have greater MSEs than in parts 1 and 2. Their MSEs are similar to the ones of the direct estimator. Concerning bias, the MA and LM estimators present the best results. See also Figure 5.3.

All the considered estimators reduces their bias and MSEs when moving from Scenario 1 to Scenario 2. This is to say, by increasing the sample sizes the performance of the estimators improves. However their relative behavior remains basically the same. This is why, we only consider Scenario 2 in the the remaining simulations.

$I(z \,\|\, x)$	DIR	EB	MA	L	LM
$I(z_0 \,\|\, x_0)$	0.0238	0.0273	0.0034	0.0077	**0.0032**
$I(z_1 \,\|\, x_1)$	0.0033	0.0262	0.0025	0.0075	**0.0023**
$I(z_1 \,\|\, x_0)$	**0.0033**	0.0408	0.0034	0.0193	0.0035
$I(z_{01}\|\, x_0)$	0.0117	0.0188	**0.0033**	0.0298	0.0049
$I(z_0 \,\|\, x_1)$	0.0238	0.0916	**0.0054**	0.1223	0.0055
$I(z_{01}\|\, x_1)$	0.0117	0.0594	**0.0040**	0.0696	0.0040

Table 5.3. Design-based $ABIAS$ for Scenario 1.

$I(z \,\|\, x)$	DIR	EB	MA	L	LM
$I(z_0 \,\|\, x_0)$	0.0095	0.0252	**0.0020**	0.0079	0.0021
$I(z_1 \,\|\, x_1)$	0.0018	0.0251	0.0015	0.0072	**0.0014**
$I(z_1 \,\|\, x_0)$	0.0018	0.0315	**0.0018**	0.0106	0.0019
$I(z_{01}\|\, x_0)$	0.0047	0.0150	**0.0016**	0.0124	0.0020
$I(z_0 \,\|\, x_1)$	0.0095	0.0889	**0.0026**	0.1189	0.0027
$I(z_{01}\|\, x_1)$	0.0047	0.0566	**0.0019**	0.0666	0.0020

Table 5.5. Design-based $ABIAS$ for Scenario 2.

$I(z \mid x)$	DIR	EB	MA	L	LM
$I(z_0 \mid x_0)$	0.0248	0.0010	0.0166	**0.0007**	0.0153
$I(z_1 \mid x_1)$	0.0129	0.0011	0.0096	**0.0003**	0.0091
$I(z_1 \mid x_0)$	0.0129	**0.0052**	0.0149	0.0057	0.0146
$I(z_{01} \mid x_0)$	0.0135	**0.0016**	0.0132	0.0052	0.0118
$I(z_0 \mid x_1)$	0.0248	**0.0087**	0.0404	0.0152	0.0426
$I(z_{01} \mid x_1)$	0.0135	**0.0072**	0.0201	0.0091	0.0207

Table 5.4. Design-based MSE for Scenario 1.

$I(z \mid x)$	DIR	EB	MA	L	LM
$I(z_0 \mid x_0)$	0.0112	0.0008	0.0061	**0.0003**	0.0060
$I(z_1 \mid x_1)$	0.0051	0.0009	0.0035	**0.0002**	0.0034
$I(z_1 \mid x_0)$	0.0051	0.0025	0.0055	0.**0022**	0.0054
$I(z_{01} \mid x_0)$	0.0060	**0.0008**	0.0048	0.0022	0.0046
$I(z_0 \mid x_1)$	0.0112	**0.0081**	0.0150	0.0143	0.0160
$I(z_{01} \mid x_1)$	**0.0060**	0.0065	0.0074	0.0082	0.0077

Table 5.6. Design-based MSE for Scenario 2.

Figures 5.1-5.3 present the boxplots of $BIAS_d$ and MSE_d, $d = 1, \ldots, D$, under Scenario 2, for the target parameters $p_{0,d}$ and $p_{01,d}$, when the EB and MA estimators rely on the correct model $I(z_0 \mid x_0)$, the incomplete model $I(z_{01} \mid x_0)$ and the incorrect model $I(z_0 \mid x_1)$ respectively.

Tables 5.3-5.6 and Figures 5.1-5.3, give similar conclusions. Choosing a model with a bad fit to data, might produce bad model-based estimates. Model-assisted estimator are more robust with respect to model selection as they have a good behavior under the sampling distribution.

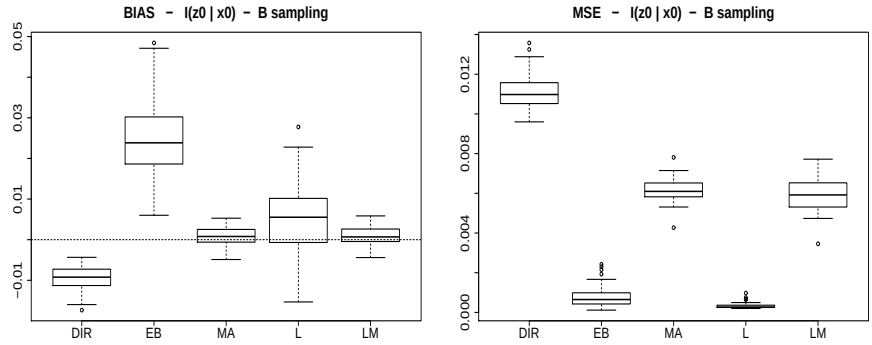

Figure 5.1: Boxplots of $BIAS_d$ (left) and MSE_d (right), $d = 1, \ldots, D$, for $I(z_0 \mid x_0)$ under Scenario 2.

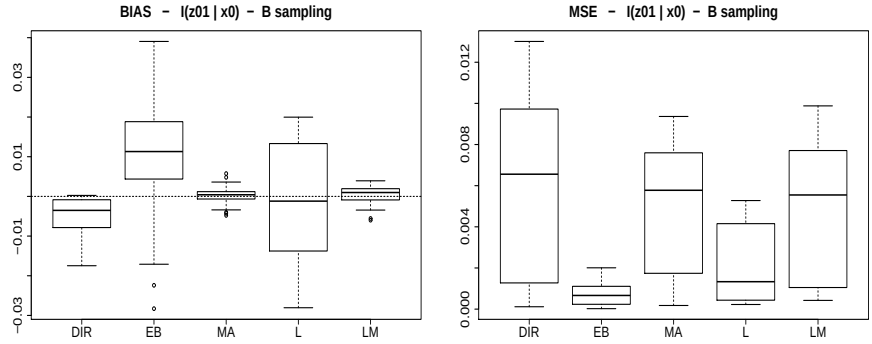

Figure 5.2: Boxplots of $BIAS_d$ (left) and MSE_d (right), $d = 1, \ldots, D$, for $I(z_{01} \,|\, x_0)$ under Scenario 2.

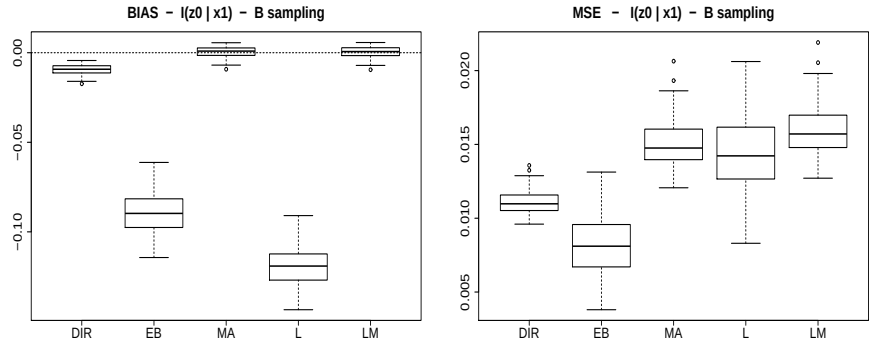

Figure 5.3: Boxplots of $BIAS_d$ (left) and MSE_d (right), $d = 1, \ldots, D$, for $I(z_0 \,|\, x_1)$ under Scenario 2.

Simulation 2 investigates the behavior of the jackknife variance estimator var_J given in (18) when it is employed for estimating the MSE of the five considered poverty proportion estimators. The steps of Simulation 2 are

1. Take $MSE_d \in \{MSE(\hat{p}_d^{dir}), MSE(\hat{p}_d^{eb}), MSE(\hat{p}_d^{ma}), MSE(\hat{p}_d^{L}), MSE(\hat{p}_d^{ML})\}$ from the output of Simulation 1.
2. For $i = 1, \ldots, I$ ($I = 10^2$), draw a Bernoulli sample from the population and apply (18) for calculating the jackknife variance estimators $mse_d^i \in \big\{\mathrm{var}_J(\hat{p}_d^{dir,i}), \mathrm{var}_J(\hat{p}_d^{eb,i}), \mathrm{var}_J(\hat{p}_d^{ma,i}), \mathrm{var}_J(\hat{p}_d^{L,i}), \mathrm{var}_J(\hat{p}_d^{ML,i})\big\}$.
3. Calculate the empirical bias and MSE of the jackknife variances, i.e.

$$B_d = \frac{1}{I}\sum_{i=1}^{I}\big\{mse_d^i - MSE_d\big\}, \quad E_d = \frac{1}{I}\sum_{i=1}^{I}\big(mse_d^i - MSE_d\big)^2, \quad d = 1, \ldots, D.$$

4. Calculate also $AB = \dfrac{1}{D}\sum_{d=1}^{D}\big|B_d\big|$ and $E = \dfrac{1}{D}\sum_{d=1}^{D} E_d$.

Tables 5.7-5.8 and Figures 5.4-5.6 summarize the results of Simulation 2. The general conclusion is that the delete-one-cluster jackknife variance estimator (18) underestimate the MSE of the model-based EB and L estimators, but it has a low bias for the direct and the model-assisted MA and LM estimators. Concerning the MSE of the jackknife estimator (18), it has lower values for the EB and L estimators and than for the remaining ones, but in all cases these values are very small (note that values of E are multiplicated by 10^4 in Table 5.8 and Figures 5.4-5.6). As in Simulation 1, under correct models the jackknife estimator of model-based and model assisted estimators presents better results than under incomplete or incorrect models.

$I(z \mid x)$	DIR	EB	MA	L	LM
$I(z_0 \mid x_0)$	0.0400	0.0725	0.0283	0.0114	0.0340
$I(z_1 \mid x_1)$	0.0529	0.0864	0.0641	0.0094	0.0687
$I(z_1 \mid x_0)$	0.0529	0.1405	0.0702	0.0310	0.0716
$I(z_{01} \mid x_0)$	0.0874	0.0307	0.0758	0.0322	0.0552
$I(z_0 \mid x_1)$	0.0400	0.8064	0.2708	1.4286	0.2864
$I(z_{01} \mid x_1)$	0.0874	0.6482	0.1725	0.8167	0.1805

Table 5.7. Design-based $10^2 AB$ of var_J for Scenario 2.

$I(z \mid x)$	DIR	EB	MA	L	LM
$I(z_0 \mid x_0)$	2.4950	0.0084	1.0880	0.0025	1.379
$I(z_1 \mid x_1)$	0.8065	0.0151	0.7560	0.0009	0.9402
$I(z_1 \mid x_0)$	0.8065	0.0892	1.4066	0.1348	1.2389
$I(z_{01} \mid x_0)$	1.5338	0.0105	0.9212	0.1430	0.6249
$I(z_0 \mid x_1)$	2.4950	0.6995	10.2401	2.1265	11.3638
$I(z_{01} \mid x_1)$	1.5338	1.2915	5.3282	1.7914	5.8642

Table 5.8. Design-based $10^4 E$, of var_J for Scenario 2.

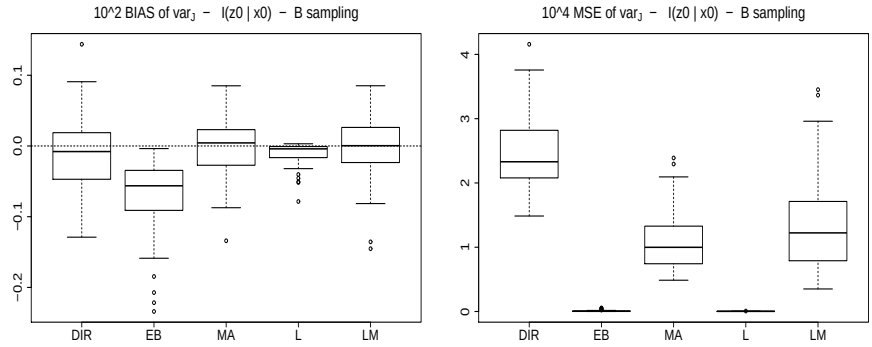

Figure 5.4: Boxplots of $10^2 B_d$ (left) and $10^4 E_d$ (right), $d = 1, \ldots, D$, for $I(z_0 \mid x_0)$ under Scenario 2.

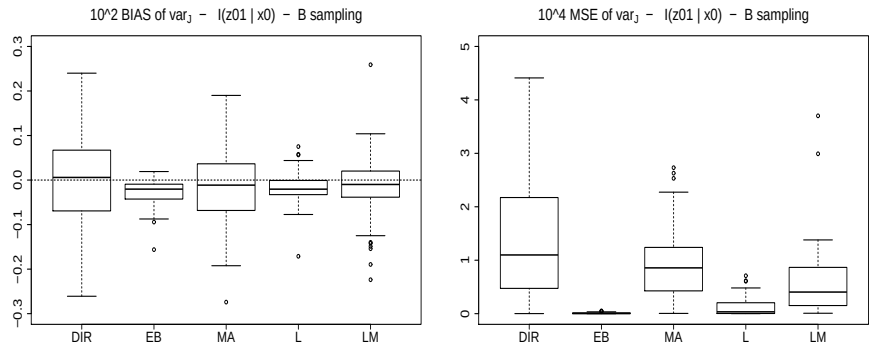

Figure 5.5: Boxplots of $10^2 B_d$ (left) and $10^4 E_d$ (right), $d = 1, \ldots, D$, for $I(z_{01} \,|\, x_0)$ under Scenario 2.

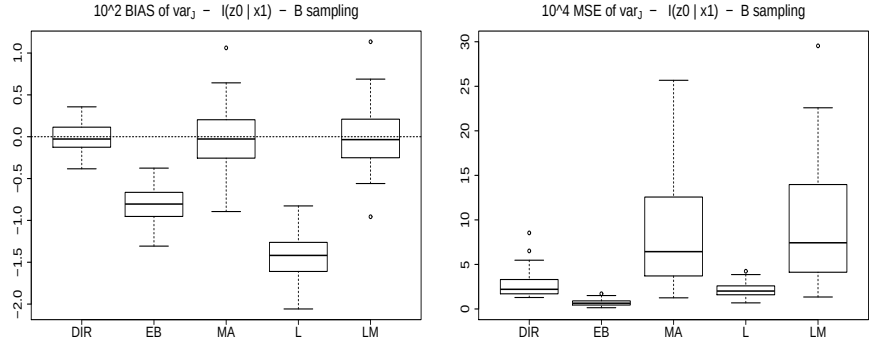

Figure 5.6: Boxplots of $10^2 B_d$ (left) and $10^4 E_d$ (right), $d = 1, \ldots, D$, for $I(z_0 \,|\, x_1)$ under Scenario 2.

6 Model-based simulation

This section presents a model-based simulation for comparing the performance of the domain HCI estimators $\hat{p}_d^{dir}$, $\hat{p}_d^{eb}$, $\hat{p}_d^{ma}$, $\hat{p}_d^{L}$ and $\hat{p}_d^{ML}$. Simulation 3 takes the same population data file as Simulation 1, but excluding the y-variables. The sample indexes (1 if unit j from domain d is sampled and 0 otherwise) are also included in the file. They are generated from Bernoulli distributions with the parameters π_{dj} defined in (19) and they remain fixed in the simulation.

In each iteration of Simulation 3 the variables y_0, y_1 and y_{01} are generated from models M_0, M_1 and M_{01} respectively. Simulation 3 is thus the model-based version of Simulation 1. In Simulation 1, the y-values are fixed and the sample indexes are drawn with logistic inclusion probabilities. In Simulation 3, the sample indexes are fixed and the y-values are generated from the models M_0, M_1 and M_{01}. The steps 2 and 3 of Simulation 3 are the same as the corresponding ones of Simulation 1. The step 1 of Simulation 3 is

1. For $i = 1, \ldots, I$ $(I = 10^3)$, generate a population target vector $\boldsymbol{y}$ under the model M_0, M_1 or M_{01}, select the fixed population subset (sample) and calculate $\hat{p}_d^i \in \{\hat{p}_d^{dir,i}, \hat{p}_d^{eb,i}, \hat{p}_d^{ma,i}, \hat{p}_d^{L,i}, \hat{p}_d^{ML,i}\}$.

Tables 6.1-6.2 present the empirical average absolute biases, $ABIAS$, and the empirical average mean squared errors MSE. The tables are divided in three parts with the same structure as Tables 5.3-5.6. In the first rows, the EB estimator has greater biases then the MA, L and LM estimators and the L and EB estimators have the lowest MSE. The MA and LM estimators have lowest bias and lower MSE than the direct estimator. See also Figures 6.1 and 6.2. In the third part, the model-based and model-assisted estimators have greater MSEs than in parts 1 and 2. Their MSEs are similar to the ones of the direct estimator. See also Figures 6.1-6.3. It is interesting to note that in the incomplete model $I(z_{01} \,|\, x_1)$ the model-assisted estimator MA present the best results.

Although Simulation 3 is model-based, the selection of the fixed sample indexes is highly related with the generating model and they influence the model simulation results. Simulation 3 extracts more population units with high values of x_0 and therefore with high values of y_0 and y_{01}. As expected, the model-based estimators have, in general, the lowest MSEs. However, the Bernoulli selection of the fixed sample indexes have a greater effect on the bias of the EB estimator than on the bias of the L estimator. This phenomenon might not happens in the case of selecting the indexes with a simple random sampling.

$I(z \,\|\, x)$	DIR	EB	MA	L	LM
$I(z_0 \,\|\, x_0)$	0.0571	0.0274	0.0133	**0.0064**	0.0103
$I(z_1 \,\|\, x_1)$	0.0344	0.0256	0.0095	**0.0034**	0.0043
$I(z_1 \,\|\, x_0)$	0.0344	0.0412	0.0349	**0.0236**	0.0349
$I(z_{01} \,\|\, x_0)$	0.0429	**0.0187**	0.0310	0.0267	0.0320
$I(z_0 \,\|\, x_1)$	**0.0571**	0.0890	0.0623	0.1197	0.0652
$I(z_{01} \,\|\, x_1)$	0.0429	0.0580	**0.0388**	0.0685	0.0394

Table 6.1. $ABIAS$ for Simulation 3 and Scenario 2.

$I(z \,\|\, x)$	DIR	EB	MA	L	LM
$I(z_0 \,\|\, x_0)$	0.0099	0.0010	0.0052	**0.0003**	0.0050
$I(z_1 \,\|\, x_1)$	0.0051	0.0010	0.0028	**0.0004**	0.0027
$I(z_1 \,\|\, x_0)$	0.0051	0.0033	0.0052	**0.0022**	0.0053
$I(z_{01} \,\|\, x_0)$	0.0056	**0.0009**	0.0039	0.0025	0.0042
$I(z_0 \,\|\, x_1)$	0.0099	**0.0081**	0.0114	0.0145	0.0121
$I(z_{01} \,\|\, x_1)$	0.0056	0.0068	**0.0048**	0.0087	0.0050

Table 6.2. MSE for Simulation 3 and Scenario 2.

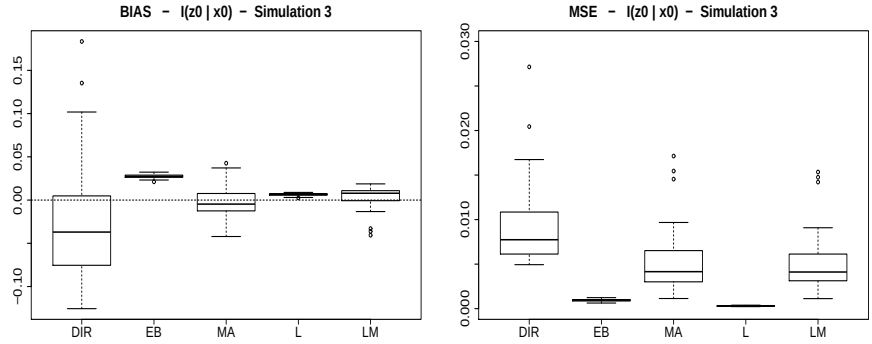

Figure 6.1: Boxplots of $BIAS_d$ (left) and MSE_d (right), $d = 1, \ldots, D$, for $I(z_0 \,|\, x_0)$, Simulation 3 and Scenario 2.

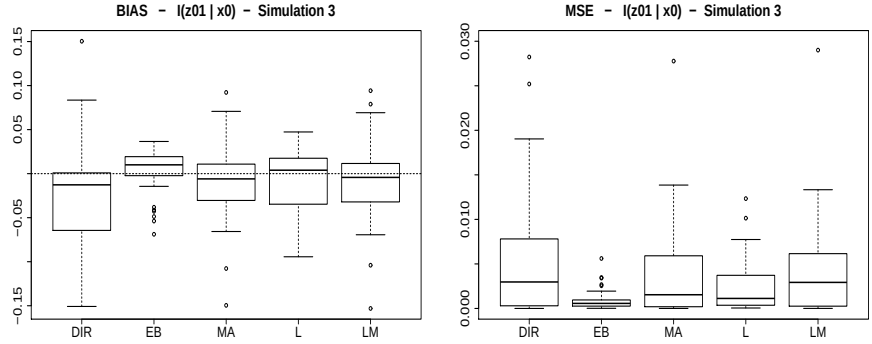

Figure 6.2: Boxplots of $BIAS_d$ (left) and MSE_d (right), $d = 1, \ldots, D$, for $I(z_{01} \,|\, x_0)$ Simulation 3 and Scenario 2.

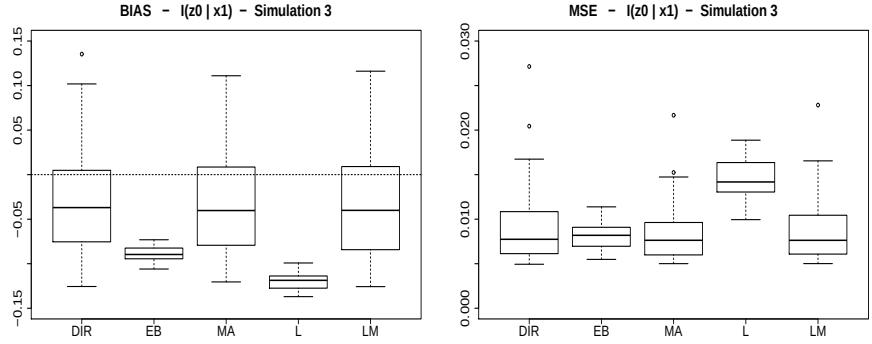

Figure 6.3: Boxplots of $BIAS_d$ (left) and MSE_d (right), $d = 1, \ldots, D$, for $I(z_0 \,|\, x_1)$ Simulation 3 and Scenario 2.

7 Application to the Spanish Living Conditions Survey

In addition to the simulation studies, we have used a set of real data to check the performance of the proposed estimators. We deal with data of the 2013 Spanish Living Conditions Survey (SLCS2013) from the region of Valencia (East of Spain).

7.1 Data, variables and sampling design

The Spanish Living Conditions Survey (SLCS) is carried out by the Instituto Nacional de Estadística (INE) of Spain. It provides information regarding the household income received during the year prior to that of the interview. This income includes income from work for others, benefits/losses from self-employed work, social benefits, income from private pension schemes not related to work, capital and property income, transfers between other households, income received by minors and the result of the income tax statement.

In order to select a sample inside each Spanish autonomous community (region), INE uses a two-stage design with first stage unit stratification. The first stage is formed by census sections and the second stage by main family dwellings (households). Within these no sub-sampling is carried out, investigating all dwellings that are their usual residence. The sections are selected within each stratum with a probability proportional to their size. The dwellings, in each section, are extracted with the same probability via random start systematic sampling. The inclusion probabilities are corrected because of non response and later calibrated to sex-age groups at the region level. This procedure leads to sampling weights that are constant inside each sampled household.

The income per household consumption unit (or equivalent personal income) is calculated in order to take into account scale economies in households. It is obtained by dividing the total household income by the number of consumption units. These are calculated using the modified OECD scale, which assigns a weight of 1 to the first adult, a weight of 0.5 to remaining adults, and a weight of 0.3 to children under 14 years of age. Once the household equivalent income is calculated, it is assigned to each of its members. The poverty threshold is calculated each year, using the distribution of the equivalent personal income for the previous year. Following the criteria recommended by Eurostat, this threshold is set at 60% of the median income per household consumption units.

The region of Valencia has three provinces, Alicante, Castellón and Valencia. The provinces are partitioned in comarcas. The target domains are the $D = 26$ comarcas appearing in SLCS2013. From the statistical office of the Valencian government, we have got a SLCS2013 subfile containing the following four variables: domain, calibrated sampling weight, equivalent personal income and labour status (employed, unemployed, inactive and below 15 years old). As the SLCS2013 calibrated sampling weights and the income are con-

stant within households, we construct a household 0-1 variable. Further, we generate a unit-level poverty indicator by comparing the equivalent personal income with the 2013 Valencian poverty threshold $\kappa = 7280$ (in euros).

Table 7.1 presents the quartiles and the mean of the domain sample sizes n_d and of the calibrated sampling weights w_{dj}, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$. Half of the domain sample sizes are lower than 74, so we are dealing with a small area estimation problem.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
354.5	1213.0	1637.0	1963.0	2280.0	15490.0
10.00	51.00	74.00	96.62	123.20	406.00

Table 7.1. Quartiles of $\{w_{dj}\}$ (top) and $\{n_d\}$ (bottom).

In addition to the SLCS2013 data, we take auxiliary aggregated data from the following statistical sources.

- The 2013 Spanish post-census data file that contains reliable demographic data. Domain sizes are taken from this file.
- The 2013 Labour Force Survey (SLFS2013) file that contains survey data about the labour market. The sizes of domains crossed by the labour status categories employed, unemployed and inactive are taken from this file. The sizes of the last category (age$\leq$ 15) are calculated by difference with the domain sizes.

The SLFS is designed for sampling within provinces with much greater sample sizes than the SLCS2013. This is why, the sizes of domains crossed by labour status (employed, unemployed, inactive and below 15 years old) are calculated by summing the corresponding SLFS calibrated sample weights. We treat these estimates as true sizes. The labour status variable is the available variable that we have got at the unit level (SLC2013) and at the aggregated level (SLFS2013).

In this paper, we focuss on the estimation of the HCI at the domain level. Let z_{dj} denote the equivalent personal income of individual j of domain d, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$. For estimating these parameters we use the direct estimators (DIR), the EBPs (EB) and the corresponding model-assisted estimators (MA) based on a nested error regression model (NER).

If a continuous auxiliary variable x is included in the NER model, then a census file containing the values of x is needed. This is a serious drawback for applying the EBP methodology.

If the set of selected auxiliary variables includes only the intercept and one factor with A categories, as it is the case in this application to real data, then the EBP estimator of a proportion takes the form

$$\hat{p}_d^{eb} = \frac{1}{N_d}\Big(\sum_{j \in s_d} I(z_{dj} < \kappa) + \sum_{a=1}^{A} (N_{da} - n_{da}) \Phi(t_{da}) \Big), \quad d = 1, \ldots, D, \qquad (20)$$

where N_{da} and n_{da} are the population and sample sizes of category a crossed by domain d and t_{da} is the common value of t_{dj} for all sampled units of

domain d and category a. A similar formula to (20) was derived by Hobza and Morales (2016) for EBP based on logistic regression mixed models. For the MA estimator, we have

$$\hat{p}_d^{ma} = \frac{1}{N_d}\Big(\sum_{j\in s_d} w_{dj} I(z_{dj} < \kappa) + \sum_{a=1}^{A} (N_{da} - \hat{N}_{da})\Phi(t_{da})\Big),$$

$$\hat{N}_{da} = \sum_{j\in s_{da}} w_{dj}, \quad d = 1, \ldots, D,$$

where w_{dj} is the calibrated sampling weight and s_{da} is the sample of domain d and category a.

7.2 The NER model

Let $x_{1,dj}$ and $x_{2,dj}$ the dichotomic variables indicating the labor status categories *employed* and *unemployed* respectively (1 if yes and 0 if no). As $\min\{z_{dj}\} = 8696.8$, we take $y_{dj} = \log(z_{dj} + k_0)$, $k_0 = 9000$, and we select the NER model

$$y_{dj} = \beta_0 + \beta_1 x_{1,dj} + \beta_2 x_{2,dj} + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \;\; d = 1, \ldots, D. \quad (21)$$

Table 7.2.1 presents the estimates of the NER model parameters and the corresponding p-values. We observe that the more people are employed ($\beta_1 > 0$) the greater is the equivalent personal income and the more people are unemployed ($\beta_2 < 0$) the smaller is the target variable. The estimated variances are $\hat{\sigma}_u^2 = 0.007216602$ and $\hat{\sigma}_e^2 = 0.1200654$.

		estimate	std.error	t-value	p-value
constant	β_0	9.94710	0.02013	494.0619	0.000
employed	β_1	0.14716	0.01534	9.5904	0.000
unemployed	β_2	-0.11338	0.02046	-5.5414	0.000

Table 7.2.1: Estimates of NER model parameters.

Figure 7.2.1 presents the dispersion graphs of model residuals versus labour status (left) and log(income+k_0) (right). The left plot shows some heterogeneity between the labour status categories. The right plot shows that residuals increases with the target variable, instead of being random distributed around zero. The model is overestimating y when y is small and underestimating y when y is large. Although the two 0-1 selected auxiliary variables are significative, Figure 7.2.1 shows that the NER model does not fit well to data. Unfortunately, we do not have any other auxiliary data available at the unit and at the domain level simultaneously. This situation happens in practice. The question is how the model-based and model-assisted estimators will behave when the selected model is not good but contains useful explanatory variables.

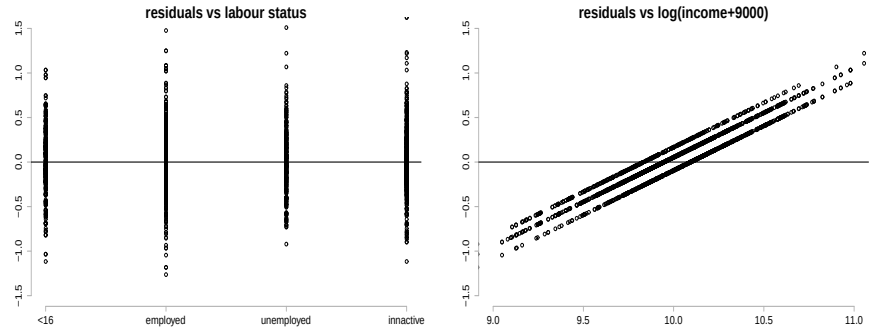

Figure 7.2.1: Residuals versus labour status (left) and income (right).

As the global sample size is $n = 2512$, the p-value of Shapiro-Wilk normality test on the residuals is 0.00. Therefore we do a graphical analysis about how close is the distribution of the residuals to normality. Figure 7.2.2 presents the density histogram of model residuals (left) and the empirical and estimated normal cumulative distribution functions (CDF) of residuals (right). The left plot shows that residual probability density function (PDF) has a right asymmetry with low probability in the interval (30000, 40000). Although the assumption of normality does not hold in strict sense, the right plot shows that the deviation is not highly severe.

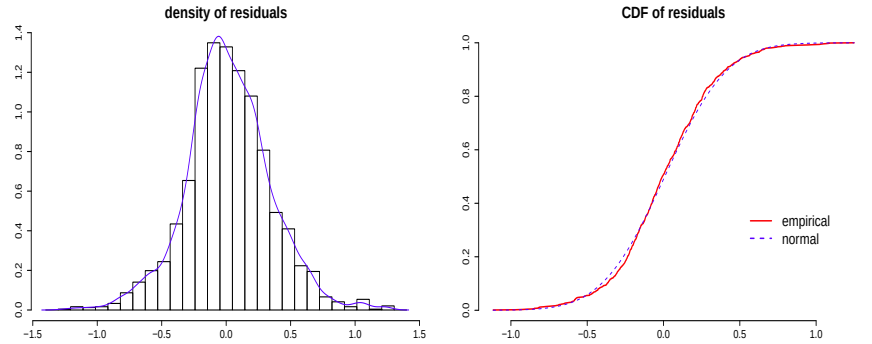

Figure 7.2.2: Histogram (left) and CDFs (right) of residuals.

This application to real data has also an illustrative purpose. We want to show how the EBP and the corresponding MA estimator behaves under a model that does not fulfils all theoretical requirements. This is why we present and discuss the results based on model (21).

7.3 Jackknife variance estimation

Based on the selected NER model, we calculate the direct, EB and MA estimators of the domain parameters θ_d, $d = 1, \ldots, D$. Their design-based jackknife

estimators $\hat{\theta}_{Jd}$, biases $\text{bias}_J(\hat{\theta}_d)$ and variances $\text{var}_J(\hat{\theta}_d)$, $d = 1, \ldots, D$, are calculated by applying (17) and (18) Let $\hat{\theta}_d^{dir}$, $\hat{\theta}_d^{ma}$ and $\hat{\theta}_d^{eb}$ be three estimators of θ_d. The comparable jackknife coefficient of variation of $\hat{\theta}_d \in \{\hat{\theta}_d^{dir}, \hat{\theta}_d^{ma}, \hat{\theta}_d^{eb}\}$ is defined as

$$\text{CCV}_d(\hat{\theta}_d) = 100\frac{\left(\text{var}_J(\hat{\theta}_d)\right)^{1/2}}{\frac{1}{3}\left(\hat{\theta}_{Jd}^{dir} + \hat{\theta}_{Jd}^{ma} + \hat{\theta}_{Jd}^{eb}\right)}.$$

7.4 Head Count Index

Figure 7.4.1 plots the HCI (left) and the jackknife CCVs (right) estimates sorted by sample size. Figure 7.4.2 presents the boxplots of the domain jackknife biases by estimators.

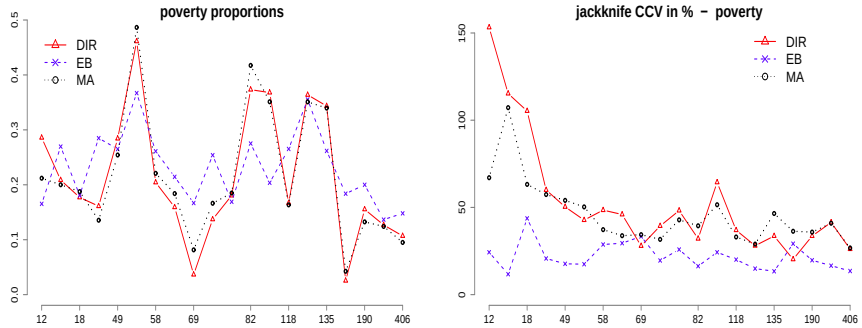

Figure 7.4.1: HCI (left) and CCV (right) estimates sorted by n_d.

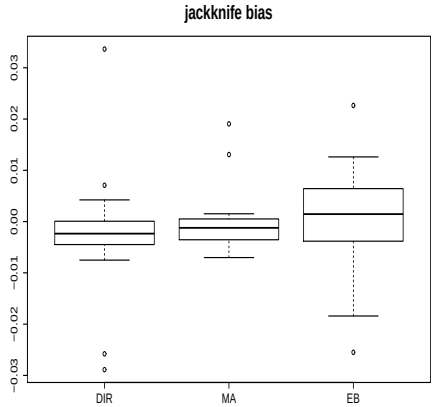

Figure 7.4.2: Boxplots of domain jackknife biases by estimators.

Figure 7.4.1 shows that the EBP estimator has in general lower design-based variance than the direct and MA estimators. The three estimators tend

to give more equal estimates as the sample size increases. Figure 7.4.2 shows that the EB estimator has in general larger biases than the MA or direct estimators.

8 Concluding remarks

In recent years there has been growing consensus among policy makers and public administrators at both national and local level concerning the need of accurate and reliable poverty, inequality and life condition indicators. This paper introduces small area estimators of the poverty measures. We introduce new estimators assisted by the NER models and they are the model-assisted counterparts of the model-based EBPs. The proposed estimators are design-consistent and asymptotically normally distributed under certain conditions. Simulation experiments are performed to investigate the behavior of the proposed estimators under the true model and their robustness against deviations from model specifications. We also compare our estimators to the competitors based or assisted by logistic regression mixed models. In addition to simulation studies, we have used real data to check the performance of the proposed estimators. Concerning the estimators DIR, EBP and MA, we summarize the results of these studies.

The direct (DIR) estimator uses only the considered domain information. It is basically an unbiased estimator with respect to the sampling design distribution but with a big variance in small area problems. Its estimated variance and coefficient of variation are generally greater than those of other estimators. Its estimated bias is generally close to zero according to the asymptotic theory. Although ideally the direct estimator is basically an unbiased estimator, there are cases in which it may has a considerable bias. Recall that the DIR estimator for a comarca is calculated as the sum of the values of the target variable multiplied by the calibrated weights. However, the calibrated weights depend on the stratum and are calculated at the region level in the SLCS. That can lead to a bias because the weights of sample units of a given domain does not "represent" well the strata structure. For example, a comarca with municipalities in h different strata could contains sampled municipalities in only one stratum.

The EBPs are based on the NER model. This model has one random effect that takes into account for the variability between domains not explained by the auxiliary variables. These estimators have good statistical properties under the distribution of the fitted model. If the model fit well to data, then they will be good estimators

The Spanish Statistical Office has not a permanently updated census file containing auxiliary variables for applying the EBP methodology to SLCS data files. The only possibility of using EBPs is by restricting to one-factor-ANOVA NER (ANOVA-NER) models. This is to say, NER models with only one auxiliary variables that is categorical. Further the sizes of the domains crossed by the variable categories should be known.

The ANOVA-NER models are useful tools for detecting differences between the target variable means by the categories. However, they are not good models for predicting the values of the target variable. As the number of values taken by the fixed effect part of ANOVA-NER models is upper bounded by the number of categories, the variability of the EBPs come basically from the variability of the predicted random effects. In this situation, these estimators are over-smoothing the pattern of the target parameter across domains.

The EBP does not employ the calibrated sampling weights at all. Therefore, they are not protected against the bias effect produced by non response. This is a serious drawback in real data applications.

The MA estimators try to collect the good properties of DIR and EBP estimators. On the one hand, they are constructed from the NER model and therefore they introduce the auxiliary information in the estimation process. On the other hand, they employ the calibrated weights in a design-based bias correction term. This term gives protection against the non response bias. In summary, the new estimators present a good balance between sampling bias and MSE.

Although theory and application have been centered on the HCI, the results are also valid to a broader class of poverty indices proposed by Foster, Greer and Thorbecke (1984).

Acknowledgements

The author thanks to the Office of Social, Demographic and Economic Statistics of the Valencian Government for providing the real data employed in the application of this paper. The authors also thanks the valuable comments and suggestions given by two anonymous reviewers. This study was partially supported by the Spanish grants MTM2015-64842-P and MTM2015-63609-R.

References

1. 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.
2. Boubeta, M., Lombardía, M.J. and Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. TEST, 25, 548-569.
3. Boubeta, M., Lombardía, M.J. and Morales, D. (2017) Poisson mixed models for studying the poverty in small areas. Computational Statistics and Data Analysis, 107, 32-47.
4. Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical assocoation, 87(418), 376-382.
5. Crettaz, E. and Suter, C. (2013). The impact of adaptive preferences on subjective indicators: An analysis of poverty indicators. Social Indicators Research, 114, 139-152.
6. Chambers, R.L. (2005). Calibrated weighting for small area estimation. Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, United Kingdom, Methodology Working Paper Series, M05/04.
7. Chandra, H., Chambers, R. (2005). Comparing EBLUP and CEBLUP for small area estimation. Statistics in Transition 7, 637-648.
8. Chandra, H., Chambers, R. (2009). Multipurpose weighting for small area estimation. Journal of Official Statistics 25 (3), 379-395.

9. Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1-26.
10. Escobar, E.L. (2014). sampling estimates: Sampling estimates. r package version 0.1-3.
11. Escobar, E.L. and Barrios, E. (2016). Samplingvarest: Sampling variance estimation. r package version 0.9-9.
12. Esteban, M.D., Morales, D., Pérez, A. and Santamaría, L. (2012a). Two Area-Level Time Models for Estimating Small Area Poverty Indicators. Journal of the Indian Society of Agricultural Statistics, 66(1), 75-89.
13. Esteban, M.D., Morales, D., Pérez, A. and Santamaría, L. (2012b). Small area estimation of poverty proportions under area-level time models. Computational Statistics and Data Analysis, 56, 2840-2855.
14. Estevao, V.M., Hidiroglou, M.A. and Särndal, C.E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. Journal of Official Statistics, 11, 181-204.
15. Estevao, V.E. and Särndal, C.E. (2004). Borrowing Strength Is Not the Best Technique Within a Wide Class of Design-Consistent Domain Estimators. Journal of Official Statistics, 20,4, 645–669.
16. Fabrizi, E., Salvati, N., Pratesi, M. and Tzavidis, N. (2014). Outlier robust model-assisted small area estimation. Biometrical Journal, 56, 157-175.
17. Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures, Econometrica, 52, 761-766.
18. Giusti, C., Masserini, L. and Pratesi, M. (2016). Local Comparisons of Small Area Estimates of Poverty: An Application Within the Tuscany Region in Italy. Social Indicators Research. In press. DOI 10.1007/s11205-015-1193-1.
19. Hájek, J. (1971) Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One". In: The Foundations of Survey Sampling, 236. Eds. Godambe, V.P. and Sprott, D.A., Holt, Rinehart and Winston.
20. Herrador, M.. Morales, D., Esteban, M.D., Sánchez, A., Santamaría, L., Marhuenda, Y. and Pérez, A. (2008). Sampling design variance estimation of small area estimators in the Spanish Labour Force survey. SORT, 32, 2, 177-198.
21. Hidiroglou, M.A. and Särndal, C.E. (1985). An empirical study of some regression estimators for small domains. Survey Methodology, 11, 65-77.
22. Hobza, T. and Morales, D. (2016). Empirical Best Prediction Under Unit-Level Logit Mixed Models. Journal of official statistics, 32, 3, 661-669.
23. 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.
24. Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. Journal of the American Statistical Association 77, 89–96.
25. Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. Survey Methodology, Vol. 24, N. 1, 51-55.
26. Lehtonen, R.,Särndal,C-E. and Veijanen, A. (2008) Generalized regression and model-calibration estimation for domains: Accuracy comparison. Paper presented at workshop on survey sampling theory and methodology.25–29 August 2008, Kuressaare, Estonia.Available at: http://www.ms.ut.ee/samp2008/present.html.
27. Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. In: Sample surveys - Inference and Analysis, Vol. 29B, 219-249. Ed. Pfeffermann, D. and Rao, C.R. Elsevier.
28. Lehtonen, R. and Veijanen, A. (2012). Small area povert estimation by model calibration. Journal of the Indian Society of Agricultural Statistics, 66, 1, 125-133.
29. Lehtonen, R. and Veijanen, A. (2016). Model-assisted method for small area estimation of poverty indicators. In: Analysis of Poverty Data by Small Area Estimation, 109-127. Ed. Monica Pratesi, John Wiley.
30. López-Vizcaíno, E., Lombardía, M. J. and Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. Statistical Modelling, 13 , 153-178.
31. López-Vizcaíno, E., Lombardía, M. J. and Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. Journal of the Royal Statistical Association, series A, 178 , 535-565.

32. Marchetti, S., Tzavidis, N. and Pratesi, M. (2012). Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. Computational Statistics and Data Analysis, 56, 2889–2902.
33. Marchetti, S. and Secondi, L. (2016). Estimates of Household Consumption Expenditure at Provincial Level in Italy by Using Small Area Estimation Methods: Real Comparisons Using Purchasing Power Parities. Social Indicators Research. In press. DOI 10.1007/s11205-016-1230-8.
34. Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics and Data Analysis, 58, 308-325.
35. Miller, R. G. (1964). A trust worthy jackknife. Annals of Mathematical Statistics, 35, 1594-1605.
36. Medeiros, M. (2006). The rich and the poor: The construction of an affluence line from the poverty line. Social Indicators Research, 78, 1-18.
37. Molina I. and Rao J.N.K. (2010). Small area estimation of poverty indicators. The Canadian Journal of Statistics, 38, 369-385.
38. Molina I. and Marhuenda, Y. (2015). sae: An R Package for Small Area Estimation. The R Journal, 7, 81-98.
39. Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. J. Amer. Statist. Assoc., 472(100), 1429–1442.
40. Morales, D., Pagliarella, M.C. and Salvatore, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. SORT-Statistics and Operations Research Transactions, 39, 1, 19-34.
41. Mukhopadhyay, P. (2001). Topics in Survey Sampling. Lecture Notes in Statistics. Springer.
42. Muñoz, J.F., Álvarez-Verdejo, E., García-Fernández, R. and Barroso, L.J.(2015). Efficient Estimation of the Headcount Index. Social Indicators Research, 123, 713-732.
43. Navicke, J., Rastrigina, O. and Sutherland, H. (2014). Nowcasting Indicators of Poverty Risk in the European Union: A Microsimulation Approach. Social Indicators Research, 119, 1, 101-119.
44. Quenouille, M. (1949). Approximation tests of correlation in time series. Journal of the Royal Statistical Society, Series B, 11, 18-84.
45. Rao, J.N.K. (2003). Small Area Estimation, New York: Wiley.
46. Rao, J. N. K. and Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. Communications in Statistic. Theory and methods, 33, 2087-2095.
47. Rao, J.N.K. and Molina, I. (2015). Small area estimation, Second Edition. John Wiley.
48. Rueda M., Muñoz J.F. (2011). Estimation of poverty measures with auxiliary information in sample surveys. Quality & Quantity, 45(3), 687-700
49. Santamaría, L., Morales, D. and Molina, I. (2004). A comparative study of small area estimators. SORT, 28, 2, 215-230.
50. Särndal, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. Bulletin of the International Statistical Institute, 49, 494-513.
51. Särndal, C.E. (1984). Design-consistent versus model-dependent estimation for small domains. Journal of the American Statistical Association, 79, 624-631.
52. Särndal, C.E. and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis. Journal of the American Statistical Association, 84, 266-275.
53. Särndal, C., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling, Springer-Verlag, New York.
54. Tillé, Y. and Matei, A. (2015). sampling: Survey sampling. r package version 2.6.
55. Tukey, J.W. (1958). Bias and confidence in not-quite large samples. The Annals of Mathematical Statistics, 29, 614.
56. Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008). M-quantile models with application to poverty mapping. Statistical Methods and Applications, 17, 393–411.
57. Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E. and Chambers, R. (2015). Robust small area prediction for counts. Statistical Methods in Medical Research, 24 , 373-395.
58. Wang, J.C., Opsomer, J. D (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators, Biometrika 98, 1, 91-106.

59. Wolter, K.M. (2007). Introduction to Variance Estimation, 2nd Edition, Springer
60. Wu, C. and Sitter, R. R. (2001) A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association, 96, 185–193.

9 Appendix - Assumptions of Theorem 4.1.

A 1 $\lim_{N\to\infty} \boldsymbol{\theta}_N = \boldsymbol{\theta} + O(N^{-1/2})$ *and* $\lim_{N\to\infty} \left(u_{dN}(\boldsymbol{\theta}_N) - u_{dN}(\boldsymbol{\theta})\right) = o(1)$.

A 2 $\lim_{N\to\infty}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) = O_\pi(n_N^{-1/2})$,

A 3 *The expected sample size* $n^* = E_\pi(n) = O(N^\delta)$, *with* $1/2 < \delta < 1$

A 4 $K_L \leq N\pi_j/n^* \leq K_U$ *for all* j, *where* K_L *and* K_U *are positive constants.*

A 5 *For any vector* z *with finite* $2 + \lambda$ *population moments with arbitrarily small* $\lambda > 0$, *let* $\bar{z}_{HT} = \frac{1}{N}\sum_{j\in s} z_j/\pi_j$ *we assume that* $V_\pi(\bar{z}_{HT}) \leq g_1 n^*(N-1)^{-1}\sum_{j\in U}(z_j - \bar{z}_N)(z_j - \bar{z}_N)'$ *for some constant* g_1

A 6 *For any* z *with finite fourth population moment the Horvitz-Thompson estimators satisfy a central limit theorem:*

$$(V_\pi(\bar{z}_{HT}))^{-1/2}(\bar{z}_{HT} - \bar{z}_N) \xrightarrow{\mathcal{L}} N(0, I_{p\times p})$$

and the estimated covariance matrix for the Horvitz-Thompson estimators is design consistent in the following sense:

$$(V_\pi(\bar{z}_{HT}))^{-1}\hat{V}_{HT}(\bar{z}_{HT}) - I_{p\times p} = O_\pi(n^{*-1/2})$$

where the design variance-covariance matrix of $\bar{z}_{HT}$ *denoted by* $V_\pi(\bar{z}_{HT})^{-1/2}$, *is positive definite, and* $\hat{V}_{HT}(\bar{z}_{HT})$ *is the Horvitz-Thompson estimator of* $V_\pi(\bar{z}_{HT}))^{-1/2}$.

A 7 *The population level function* $T_N(\boldsymbol{\eta}) = \frac{1}{N_d}\sum_{j\in U_d} g_{dj}(\boldsymbol{\eta})$ *converges to a limiting smooth function* $T(\boldsymbol{\eta})$, *uniformly in a neighborhood of* $\boldsymbol{\theta}$. *This limiting function is uniformly continuous for* $\boldsymbol{\eta}$ *in a neighborhood of* $\boldsymbol{\theta}$ *and has finite first and second derivatives with respect to* $\boldsymbol{\eta}$.

A 8 *The population quantity*

$$sup_{\boldsymbol{\eta}\in C} N^\alpha |T_N(\boldsymbol{\theta}_N + N^{-\alpha}\boldsymbol{\eta}) - T_N(\boldsymbol{\theta}_N) - T(\boldsymbol{\theta}_N + N^{-\alpha}\boldsymbol{\eta}) + T(\boldsymbol{\theta}_N)| \to 0$$

where C *is a large enough compact set in* R^{p+2} *and* $\alpha \in (\frac{1}{4}, \frac{1}{2}]$.

These assumptions are similar to those used in Wang and Opsomer (2011) and Fabrizi et al. (2014). Assumptions A1 and A2 ensure that the sample fit $\hat{\boldsymbol{\theta}}$ and the population fit $\boldsymbol{\theta}_N$ share a common limit. Assumptions A3, A4, A5 and A6 are satisfied for commonly used sample size designs in reasonably finite populations. However, it would not hold for systematic sampling or one-per-stratum designs. A7 assumption about the estimator allows us to use the limiting smooth function instead of nonsmooth population quantity in asymptotic expansion.