SMALL AREA ESTIMATION OF POVERTY INDICATORS

Molina, Isabel[1] and Rao, J. N. K.[2]

Abstract

We propose to estimate non-linear small area population quantities by using Empirical Best (EB) estimators based on a nested error model. EB estimators are obtained by Monte Carlo approximation. We focus on poverty indicators as particular non-linear quantities of interest, but the proposed methodology is applicable to general non-linear quantities. Small sample properties of EB estimators are analyzed by model-based and design-based simulation studies. Results show large reductions in mean squared error relative to direct estimators and estimators obtained by simulated censuses. An application is also given to estimate poverty incidences and poverty gaps in Spanish provinces by sex with mean squared errors estimated by parametric bootstrap. In the Spanish data, results show a significant reduction in coefficient of variation of the proposed EB estimators over direct estimators for most domains.

Keywords: Empirical best estimator; Parametric bootstrap; Poverty mapping; Small area estimation.

[1] Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain. E-mail: isabel.molina@uc3m.es

[2] School of Mathematics and Statistics, Carleton University, 125 Colonel By Drive, Ottawa, ON CANADA K1S 5B6. E-mail: jrao@math.carleton.ca

Small Area Estimation of Poverty Indicators*

Isabel Molina†and J. N. K. Rao‡

Abstract

We propose to estimate non-linear small area population quantities by using Empirical Best (EB) estimators based on a nested error model. EB estimators are obtained by Monte Carlo approximation. We focus on poverty indicators as particular non-linear quantities of interest, but the proposed methodology is applicable to general non-linear quantities. Small sample properties of EB estimators are analyzed by model-based and design-based simulation studies. Results show large reductions in mean squared error relative to direct estimators and estimators obtained by simulated censuses. An application is also given to estimate poverty incidences and poverty gaps in Spanish provinces by sex with mean squared errors estimated by parametric bootstrap. In the Spanish data, results show a significant reduction in coefficient of variation of the proposed EB estimators over direct estimators for most domains.

Key words: Empirical best estimator; Parametric bootstrap; Poverty mapping; Small area estimation.

1 Introduction

The first of the Millennium Development Goals established by the United Nations is the eradication of extreme poverty and hunger. The availability of the most possible accurate information concerning the living conditions of people is a basic instrument for targeting policies and programs aiming at the reduction of poverty. Thus, there are no doubts about the importance of the estimation of poverty measures. In many cases the information collected from national surveys is limited and allows estimation only for larger regions or larger population subgroups. Therefore, small area estimation techniques are required that "borrow strength" across areas through linking models and auxiliary information such as

censuses and administrative data; see Rao (2003) for a comprehensive account of these techniques.

Many measures of poverty and inequality are non-linear functions of a welfare variable for the population units. This makes many of the current small area estimation methods, typically developed for the estimation of linear characteristics, such as means, not applicable. Here we propose the use of empirical best predictors (EBPs) obtained through Monte Carlo approximation. This method provides estimators that are "best" in the sense of minimizing the mean squared error (MSE) under assumed small area models. It is useful for the estimation of any function of a welfare variable, when this variable or some transformation of it follows a linear model. We show by simulations that empirical best predictors behave good in terms of bias and mean squared error. We also propose a parametric bootstrap method for MSE estimation and study its bias in simulations.

In the U.S. the need for small area poverty estimates has given rise to the SAIPE program (Small Area Income & Poverty Estimates) of the U.S. Census Bureau; for further details see http://www.census.gov/hhes/www/saipe. The main objective of this program is to provide updated estimates of income and poverty statistics for the administration of federal programs and the allocation of federal funds to local jurisdictions. The county level methodology, summarized by Bell (1997), basically uses a Fay-Herriot area level model (Fay & Herriot, 1979) to produce model-based county estimates of school-age children under poverty.

The World Bank (WB) has been releasing small area poverty and income inequality estimates for some countries, using the methodology of Elbers et al. (2003). This methodology is currently widely used, see e.g. Neri et al. (2005), Ballini et al. (2006), Tarozzi and Deaton (2007) and Haslett and Jones (2005). Elbers et al. (2003) assumed a unit level model that combines both census and survey data. Using that model, they produce disaggregated maps that describe the spatial distribution of poverty and inequality. Section 7 shows that this method, when used to estimate small area means, is basically equal to a synthetic regression estimator.

Measures of inequality include the Gini coefficient, the Sen index, the general entropy and the Theil index (see e.g. Neri et al., 2005). Although the method developed in this paper allows the estimation of these inequality measures as well, for the sake of brevity we will focus on the estimation of a class of poverty measures called FGT poverty measures due to Foster et al (1984), see Section 2, and used in the WB method.

A common definition of poverty classifies a person as "under poverty" when the selected welfare variable for this person is below a threshold called poverty line, which is usually a given percentage of the median welfare for the population. For instance, the Spanish National Statistical Institute (INE) established the poverty line at 60% of the median per capita income. Here we assume that the poverty line is fixed at some quantity established by the corresponding authority and we avoid discussions concerning the definition of this threshold.

The paper is organized as follows. Section 2 defines the family of FGT poverty measures and Section 3 introduces two basic types of direct estimators, which make use only of the sample data from the target area. Section 4 describes the best prediction methodology for finite populations and Section 5 applies this methodology to the estimation of FGT poverty measures through the use of a nested error linear regression model. Section 6 describes a parametric bootstrap method for MSE estimation. Section 7 makes a theoretical comparison of the different methods in the context of estimating small area means. Sections 8 and 9 present the results of simulation studies on the performance of the proposed method relative to the WB method and direct estimation, in terms of bias and MSE. Performance of bootstrap MSE estimator is also studied. Finally, in Section 10, the proposed method is applied to Spanish data to estimate poverty incidences and poverty gaps in Spanish provinces by gender.

2 FGT poverty measures for small areas

Consider a finite population of size N that is partitioned into D small areas of sizes $N_1, \ldots, N_D$. Let E_{dj} be a suitable quantitative measure of welfare for individual j in small area d, such as income or expenditure, and let z be the given poverty line; that is, the threshold for E_{dj} under which a person is considered as "under poverty". The family of poverty measures introduced by Foster et al. (1984) and called throughout the paper as FGT poverty measures, for each small area d, is defined as

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{z - E_{dj}}{z} \right)^\alpha I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D,$$

where $I(E_{dj} < z) = 1$ if $E_{dj} < z$ (person under poverty) and $I(E_{dj} < z) = 0$ if $E_{dj} \geq z$ (person not under poverty). For $\alpha = 0$ we get the proportion of individuals under poverty in small area d, also called poverty incidence or head count ratio. The FGT measure for $\alpha = 1$ is called poverty gap, and measures the area mean of the relative distance to non-poverty (the poverty gap) of each individual. For $\alpha = 2$ the measure is called poverty severity. This measure squares the poverty gaps, and thus emphasizes extreme poverty.

3 Direct estimators of poverty measures

In the inference process, a random sample of size $n < N$ is drawn from the population according to a specified sampling design. Let Ω denote the set of indexes of the population units. Let s be the set of units selected in the sample and r the set of indexes of the units that are not selected (with size $N - n$). The restrictions of Ω, s, N and n to area d are denoted by Ω_d, s_d, N_d and n_d respectively, where $n = n_1 + \cdots + n_D$. The unweighted sample FGT poverty

measures are given by

$$f_{\alpha d} = \frac{1}{n_d} \sum_{j \in s_d} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D. \quad (1)$$

A direct estimator for a small area, as a sample estimator, uses only the sample data from the target small area and it is usually design-based. Let w_{dj} be the sampling weight (inverse of the probability of inclusion) of individual j from area d. Direct estimators of the FGT measures are given by

$$f_{\alpha d}^{w} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_{dj} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D, \quad (2)$$

where $\hat{N}_d = \sum_{j \in s_d} w_{dj}$ is the direct estimator of the population size of small area d, N_d. If the sampling weights w_{dj} do not depend on the unit j, for example $w_{dj} = n_d/N_d$ under simple random sampling within areas, then (2) reduces to the unweighted mean (1).

The limited sample sizes n_d within some of the areas prevent the use of estimators such as (1) or (2). Indeed, a common definition of poverty classifies a person as "under poverty" when the selected welfare variable for this person is below 60% of the median. Under this definition the outcome of being under poverty is likely to have low frequency. Then, to obtain reliable estimators for small domains or geographical areas it is necessary to appeal to small area techniques (Rao, 2003). These techniques improve the estimation procedures by using models that establish some relationships between the areas, based on auxiliary information (census and/or administrative variables) related to the welfare variables of interest. These models provide "indirect" estimators that make use of related data from other areas, and which might reduce drastically the estimation errors as long as model assumptions hold. Model checking should be an integral part of indirect estimation methods.

4   Best prediction under a finite population

Section 4 introduces the best predictor (BP) of a function of a random vector in a finite population. Application of the BP methodology for estimating FGT poverty measures in small areas is described in Section 5.

Consider a random vector $\mathbf{y} = (Y_1, \ldots, Y_N)'$ containing the values of a random variable associated with N units of a finite population. Let $\mathbf{y}_s$ be the sub-vector of $\mathbf{y}$ corresponding to sample elements and $\mathbf{y}_r$ the sub-vector of out-of-sample elements; that is, $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$. The target is to predict the value of a real-valued function $\delta = h(\mathbf{y})$ of the random vector $\mathbf{y}$ using the sample data $\mathbf{y}_s$. For a particular predictor $\hat{\delta}$, the mean squared error is defined as

$$\text{MSE}(\hat{\delta}) = E_{\mathbf{y}}\{(\hat{\delta} - \delta)^2\}, \quad (3)$$

4

where $E_{\mathbf{y}}$ denotes expectation with respect to the joint distribution of the population vector $\mathbf{y}$. The BP of δ is the function of $\mathbf{y}_s$ that minimizes (3). Consider the conditional expectation $\delta^0 = E_{\mathbf{y}_r}(\delta|\mathbf{y}_s)$, where the expectation is taken with respect to the conditional distribution of $\mathbf{y}_r$ given $\mathbf{y}_s$ and the result is a function of sample data $\mathbf{y}_s$. Subtracting and adding δ^0 in the mean squared error, we obtain

$$\text{MSE}(\hat{\delta}) = E_{\mathbf{y}}\{(\hat{\delta} - \delta^0)^2\} + 2\, E_{\mathbf{y}}\{(\hat{\delta} - \delta^0)(\delta^0 - \delta)\} + E_{\mathbf{y}}\{\delta^0 - \delta)^2\}.$$

In this expression, the last term does not depend on $\hat{\delta}$. For the second term, observe that

$$
\begin{aligned}
E_{\mathbf{y}}\{(\hat{\delta} - \delta^0)(\delta^0 - \delta)\} &= E_{\mathbf{y}_s}\left[E_{\mathbf{y}_r}\left\{(\hat{\delta} - \delta^0)(\delta^0 - \delta)|\mathbf{y}_s\right\}\right] \\
&= E_{\mathbf{y}_s}\left[(\hat{\delta} - \delta^0)\left\{\delta^0 - E_{\mathbf{y}_r}(\delta|\mathbf{y}_s)\right\}\right] = 0.
\end{aligned}
$$

Thus, the BP is the value $\hat{\delta}$ that minimizes $E_{\mathbf{y}}\{(\hat{\delta} - \delta^0)^2\}$. Since this quantity is non-negative and its minimum value is zero, the BP is

$$\hat{\delta}^B = \delta^0 = E_{\mathbf{y}_r}(\delta|\mathbf{y}_s). \tag{4}$$

Typically, the joint distribution of $\mathbf{y}$ depends on some unknown model parameters. Then an empirical BP (EBP) of δ can be obtained by replacing all unknown parameters by suitable estimators and then evaluating the expectation (4) from the estimated distribution.

Remark 1. *Suppose that $\mathbf{y}$ follows a Normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, for a known matrix $\mathbf{X}$, and positive definite covariance matrix $\mathbf{V}$. Let the target quantity δ be a linear function of $\mathbf{y}$, $\delta = \mathbf{a}'\mathbf{y}$. In this case, the best linear unbiased predictor (BLUP) of δ, obtained by Royall (1976), is equal to the BP when the true $\boldsymbol{\beta}$ is replaced by the generalized least squares estimator.*

5 Empirical best prediction of FGT poverty measures

In this section we describe how to obtain BPs of FGT poverty measures for small areas. For a given $\alpha \in \{0, 1, 2\}$, let us define the random variables

$$F_{\alpha dj} = \left(\frac{z - E_{dj}}{z}\right)^\alpha I(E_{dj} < z), \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D.$$

Then the FGT poverty measure for area d is the mean

$$F_{\alpha d} = \frac{1}{N_d}\sum_{j=1}^{N_d} F_{\alpha dj}, \quad d = 1, \ldots, D. \tag{5}$$

Suppose that there is a one-to-one transformation $Y_{dj} = T(E_{dj})$ of the welfare variables, E_{dj}, such that the vector $\mathbf{y}$ containing the values of the transformed variables Y_{dj} for all the population units satisfies $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$. Then we can express the random variables $F_{\alpha dj}$ in terms of Y_{dj} as

$$F_{\alpha dj} = \left(\frac{z - T^{-1}(Y_{dj})}{z} \right)^{\alpha} I\{T^{-1}(Y_{dj}) < z\} = h_{\alpha}(Y_{dj}), \quad j = 1, \ldots, N_d.$$

Thus, the FGT poverty measure (5) is a non-linear function of the vector $\mathbf{y}$. By the results of Section 4, taking $\delta = F_{\alpha d}$ the BP of $F_{\alpha d}$ is

$$\hat{F}_{\alpha d}^{B} = E_{\mathbf{y}_r}(F_{\alpha d} | \mathbf{y}_s). \tag{6}$$

Using the decomposition of the mean (5) in terms of sample and out-of-sample elements, we have

$$F_{\alpha d} = \frac{1}{N_d} \left\{ \sum_{j \in s_d} F_{\alpha dj} + \sum_{j \in r_d} F_{\alpha dj} \right\}, \tag{7}$$

where r_d denotes the set of out-of-sample elements belonging to area d. Now taking conditional expectation of (7) and introducing the conditional expectation inside the sum, the BP becomes

$$\hat{F}_{\alpha d}^{B} = \frac{1}{N_d} \left\{ \sum_{j \in s_d} F_{\alpha dj} + \sum_{j \in r_d} \hat{F}_{\alpha dj}^{B} \right\}, \tag{8}$$

where $\hat{F}_{\alpha dj}^{B} = E_{\mathbf{y}_r}(F_{\alpha dj} | \mathbf{y}_s)$ is the BP of the out-of-sample variable $F_{\alpha dj} = h_{\alpha}(Y_{dj})$, which is defined as

$$\hat{F}_{\alpha dj}^{B} = E_{\mathbf{y}_r}[h_{\alpha}(Y_{dj}) | \mathbf{y}_s] = \int_{\mathbb{R}} h_{\alpha}(Y_{dj}) f(Y_{dj} | \mathbf{y}_s) \, dY_{dj}, \quad j \in r_d, \tag{9}$$

where $f(Y_{dj} | \mathbf{y}_s)$ is the conditional density of Y_{dj} given the data vector $\mathbf{y}_s$. The expectation in (9) cannot be calculated explicitly due to the complexity of h_{α}. However, since $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ is Normally distributed with mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_s', \boldsymbol{\mu}_r')'$ and covariance matrix partitioned conformably as

$$\mathbf{V} = \left(\begin{array}{cc} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{array} \right),$$

the conditional distribution of $\mathbf{y}_r$ given $\mathbf{y}_s$ is

$$\mathbf{y}_r | \mathbf{y}_s \sim N(\boldsymbol{\mu}_{r|s}, \mathbf{V}_{r|s}), \tag{10}$$

where

$$\boldsymbol{\mu}_{r|s} = \boldsymbol{\mu}_r + \mathbf{V}_{rs} \mathbf{V}_s^{-1} (\mathbf{y}_s - \boldsymbol{\mu}_s), \quad \mathbf{V}_{r|s} = \mathbf{V}_r - \mathbf{V}_{rs} \mathbf{V}_s^{-1} \mathbf{V}_{sr}. \tag{11}$$

Then, we propose to use an empirical approximation by Monte Carlo simulation of a large number L of vectors $\mathbf{y}_r$ generated from (10). Let $Y_{dj}^{(\ell)}$ be the value

6

of the out-of-sample observation Y_{dj}, $j \in r_d$, obtained in the ℓ-th simulation, $\ell = 1, \ldots, L$. A Monte Carlo approximation to the best predictor of Y_{dj} is then given by

$$\hat{F}^B_{\alpha dj} = E_{\mathbf{y}_r}[h_\alpha(Y_{dj})|\mathbf{y}_s] \approx \frac{1}{L} \sum_{\ell=1}^{L} h_\alpha(Y_{dj}^{(\ell)}), \quad j \in r_d. \tag{12}$$

In practice, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{V}$ usually depend on an unknown vector of parameters $\boldsymbol{\theta}$. Thus the conditional density $f(Y_{dj}|\mathbf{y}_s)$ depends on $\boldsymbol{\theta}$ and we make this explicit in the notation using $f(Y_{dj}|\mathbf{y}_s, \boldsymbol{\theta})$. We can take an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such as the maximum likelihood (ML) estimator or the residual ML (REML) estimator. Then the expectation can be approximated by generating values $Y_{dj}^{(\ell)}$ from the estimated density $f(Y_{dj}|\mathbf{y}_s, \hat{\boldsymbol{\theta}})$. The resulting predictor, denoted $\hat{F}^{EB}_{\alpha dj}$, is called the empirical best predictor (EBP). Finally, the EBP of the poverty measure $F_{\alpha d}$ is given by

$$\hat{F}^{EB}_{\alpha d} = \frac{1}{N_d} \left[\sum_{j \in s_d} F_{\alpha dj} + \sum_{j \in r_d} \hat{F}^{EB}_{\alpha dj} \right].$$

In this paper, we consider the nested error linear regression model for the Y_{dj} (Battese et al., 1988). This model relates the transformed variables Y_{dj} (e.g., log-earnings) to a vector of p explanatory variables $\mathbf{x}_{dj}$ for all areas, and includes a random area-specific effect u_d along with the usual residual errors e_{dj}:

$$Y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad u_d \sim \text{iid } N(0, \sigma_u^2),$$
$$e_{dj} \sim \text{iid } N(0, \sigma_e^2), \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D, \tag{13}$$

where the area effects u_d and the errors e_{dj} are independent. Defining vectors and matrices obtained by stacking the elements for area d as

$$\mathbf{y}_d = \operatorname*{col}_{1 \leq j \leq N_d}(Y_{dj}), \quad \mathbf{e}_d = \operatorname*{col}_{1 \leq j \leq N_d}(e_{dj}), \quad \mathbf{X}_d = \operatorname*{col}_{1 \leq j \leq N_d}(\mathbf{x}'_{dj}),$$

the vectors $\mathbf{y}_d$, $d = 1, \ldots, D$, under model (13), are independent with

$$\mathbf{y}_d \sim N(\boldsymbol{\mu}_d, \mathbf{V}_d),$$

where

$$\boldsymbol{\mu}_d = \mathbf{X}_d\boldsymbol{\beta} \quad \text{and} \quad \mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{I}_{N_d}.$$

We assume that the population model (13) holds for the sample, i.e., sample selection bias is absent (Pfeffermann et al. 1998).

Consider the decomposition of $\mathbf{y}_d$ into sample and out-of-sample elements $\mathbf{y}_d = (\mathbf{y}'_{dr}, \mathbf{y}'_{ds})'$, and the corresponding decomposition of $\boldsymbol{\mu}_d = E(\mathbf{y}_d)$ and $\mathbf{V}_d = \text{Var}(\mathbf{y}_d)$. Then the distribution of the out-of-sample vector $\mathbf{y}_{dr}$ given the sample data $\mathbf{y}_{ds}$ is

$$\mathbf{y}_{dr}|\mathbf{y}_{ds} \sim N(\boldsymbol{\mu}_{dr|s}, \mathbf{V}_{dr|s}), \tag{14}$$

where the conditional mean vector and covariance matrix are given by

$$\boldsymbol{\mu}_{dr|s} = \mathbf{X}_{dr}\boldsymbol{\beta} + \sigma_u^2 \mathbf{1}_{N_d-n_d}\mathbf{1}'_{n_d}\mathbf{V}_{ds}^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) \qquad (15)$$

$$\mathbf{V}_{dr|s} = \sigma_u^2(1 - \gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}'_{N_d-n_d} + \sigma_e^2 \mathbf{I}_{N_d-n_d}. \qquad (16)$$

Note that $\mathbf{y}_{dr}|\mathbf{y}_{ds}$ and $\mathbf{y}_{dr}|\mathbf{y}_s$ have the same distribution (14) due to the independence of the vectors $\mathbf{y}_d$.

Observe that the application of the Monte Carlo approximation (12) involves simulation of D multivariate Normal vectors of sizes $N_d - n_d$, $d = 1, \ldots, D$, from (14). Then this process has to be repeated L times, something computationally very intensive. This can be avoided in the following way. Observe that the matrix $\mathbf{V}_{dr|s}$ corresponds to the covariance matrix of a vector $\mathbf{y}_{dr}$ generated by the model

$$\mathbf{y}_{dr} = \boldsymbol{\mu}_{dr|s} + v_d \mathbf{1}_{N_d-n_d} + \boldsymbol{\epsilon}_{dr}, \qquad (17)$$

with new random effects v_d and errors $\boldsymbol{\epsilon}_{dr}$ that are independent and satisfy

$$v_d \sim N(0, \sigma_u^2(1 - \gamma_d)) \quad \text{and} \quad \boldsymbol{\epsilon}_{dr} \sim N(\mathbf{0}_{N_d-n_d}, \sigma_e^2 \mathbf{I}_{N_d-n_d}).$$

Using these relations, instead of generating a multivariate normal vector of size $N_d - n_d$, we just need to generate univariate normal variables $v_d \sim N(0, \sigma_u^2(1-\gamma_d))$ and $\epsilon_{dj} \sim N(0, \sigma_e^2)$ independently, for $j \in r_d$, and then obtain the responses Y_{dj} from (17) using the known value of $\boldsymbol{\mu}_{dr|s}$. As mentioned before, in practice all the unknown model parameters $\boldsymbol{\beta}$, σ_u^2 and σ_e^2 are replaced by suitable estimators, and then the variables Y_{dj} are generated from the corresponding estimated normal distributions.

Summarizing, the proposed EBP method to estimate poverty measures works in the following way:

(a) Fit model (13) to the initial (transformed) data $\mathbf{y}_s$.

(b) Draw L out-of-sample vectors $\mathbf{y}_r^{(\ell)}$, $\ell = 1, \ldots, L$ from (14), or equivalently from (17), but with the unknown parameters replaced by the estimators obtained in (a).

(c) With the L generated vectors $\mathbf{y}_r^{(\ell)}$, $\ell = 1, \ldots, L$, and using the sample data $\mathbf{y}_s$, compute EBPs of the poverty measures from (8) using the Monte Carlo approximation (12).

Instead of introducing the expectation inside the sum as in (8), we can approximate by Monte Carlo directly the expectation of the poverty measure (6). In this way, the proposed procedure can be used to predict any other function $h(\mathbf{y})$ and it requires only the assumption of a model for some transformation of the welfare variable E_{dj}.

6 Parametric bootstrap for MSE estimation

Here we propose to use an extension of the parametric bootstrap method for finite populations (González-Manteiga et al., 2008) to estimate the MSE of the empirical best predictors $\hat{F}_{\alpha d}^{EB}$, $d = 1, \ldots, D$. For a given α, this method works as follows:

1. Fit model (13) to sample data $\mathbf{y}_s$ and obtain model parameter estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$.

2. Generate bootstrap random area effects as $u_d^* \sim$ iid $N(0, \hat{\sigma}_u^2)$, $d = 1, \ldots, D$.

3. Generate, independently of the random effects u_d^*, bootstrap random errors $e_{dj}^* \sim$ iid $N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_d$, $d = 1, \ldots, D$.

4. Construct a bootstrap population using the estimated model,

$$Y_{dj}^* = \mathbf{x}_{dj}'\hat{\boldsymbol{\beta}} + u_d^* + e_{dj}^*, \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D, \qquad (18)$$

and calculate the FGT measures for this population, that is, calculate first $F_{\alpha dj}^* = h_\alpha(Y_{dj}^*)$, $j = 1, \ldots, N_d$, and then take the small area means as

$$F_{\alpha d}^* = \frac{1}{N_d} \sum_{j=1}^{N_d} F_{\alpha dj}^*, \quad d = 1, \ldots, D.$$

5. Take the elements Y_{dj}^* with indices contained in the sample s, denoted $\mathbf{y}_s^*$. Fit the model again to $\mathbf{y}_s^*$ obtaining new model parameter estimates $\hat{\boldsymbol{\beta}}^*$, $\hat{\sigma}_u^{2*}$ and $\hat{\sigma}_e^{2*}$.

6. Using the bootstrap sample data $\mathbf{y}_s^*$ and the known matrix $\mathbf{X}$, apply the EB method as described in Section 5 and calculate bootstrap EBPs, $\hat{F}_{\alpha d}^{EBP*}$, $d = 1, \ldots, D$.

Observe that the bootstrap elements, given the original sample data, preserve properties of the original population model. Random effects and errors are iid with

$$E_*(u_d^*) = 0, \quad Var_*(u_d^*) = \hat{\sigma}_u^2, \quad E_*(e_{dj}^*) = 0, \quad Var_*(e_{dj}^*) = \hat{\sigma}_e^2, \qquad (19)$$

where E_* and Var_* denote expectation and variance with respect to the distribution defined by the bootstrap model (18) given sample data $\mathbf{y}_s$. Consider the vector $\mathbf{y}^*$ with all the population bootstrap elements defined analogously to $\mathbf{y}$ for the original population. Then the mean vector and covariance matrix of this bootstrap vector are

$$E_*(\mathbf{y}^*) = \mathbf{X}\hat{\boldsymbol{\beta}}, \quad Var_*(\mathbf{y}^*) = \hat{\sigma}_u^2 \mathbf{Z}\mathbf{Z}' + \hat{\sigma}_e^2 \mathbf{I}_N.$$

Thus, the distribution of the bootstrap population $\mathbf{y}^*$ (given sample data $\mathbf{y}_s$) imitates that of the original population $\mathbf{y}$. Then an estimator of $\text{MSE}(\hat{F}_{\alpha d}^{EB})$ is the bootstrap MSE of the bootstrap EBP, defined as

$$\text{MSE}_*(\hat{F}_{\alpha d}^{EBP*}) = E_*\{(\hat{F}_{\alpha d}^{EBP*} - F_{\alpha d}^*)^2\}.$$

In practice, this quantity can be approximated through a Monte Carlo procedure, by repeating steps 2–6 a large number of times, B, and then taking the mean over the B replicates. More specifically, let $F_{\alpha d}^{*(b)}$ and $\hat{F}_{\alpha d}^{EB*(b)}$ be the poverty measure and its corresponding EBP for the bootstrap replicate b, for $b = 1, \ldots, B$. Then, the estimator of the MSE is calculated as

$$\text{mse}(\hat{F}_{\alpha d}^{EB}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{F}_{\alpha d}^{EB*(b)} - F_{\alpha d}^{*(b)})^2. \tag{20}$$

It is possible to obtain a better MSE estimator, in terms of relative bias, by using a double bootstrap method (Hall and Maiti, 2006) applied to the transformed sample data.

7   Estimation of small area means

In this section we restrict ourselves to the estimation of the small area means

$$\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} Y_{dj}, \quad d = 1, \ldots, D,$$

since means are common target quantities which also deserve attention; moreover, means allow for the study of some theoretical properties of estimators.

Consider first the case of population elements Y_{dj} following a linear model without area effects,

$$Y_{dj} = \mathbf{x}_{dj}'\boldsymbol{\beta} + e_{dj}, \quad e_{dj} \sim \text{iid } N(0, \sigma_e^2), \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D. \tag{21}$$

Let us assume for simplicity of exposition that all model parameters are known. Taking the average of (21) over the elements in area d, we can express the true mean as

$$\bar{Y}_d = \bar{\mathbf{x}}_d'\boldsymbol{\beta} + \bar{e}_d,$$

where $\bar{\mathbf{x}}_d = N_d^{-1} \sum_{j=1}^{N_d} \mathbf{x}_{dj}'$ and $\bar{e}_d = N_d^{-1} \sum_{j=1}^{N_d} e_{dj}$. Common small area estimators derived under model (21) are the synthetic estimators, obtained by predicting all the Y_{dj} and then taking the area mean, that is, taking

$$\hat{\bar{Y}}_d^{SYN} = \frac{1}{N_d} \sum_{j=1}^{N_d} \hat{Y}_{dj},$$

as the synthetic estimator, where $\hat{Y}_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta}$ is the predictor of Y_{dj} for $j = 1, \ldots, N_d$. However, under model (21), the BP of $\bar{Y}_d$ is obtained by predicting only the out-of-sample observations and keeping the sample data, i.e.,

$$\hat{\bar{Y}}_d^B = \frac{1}{N_d}\left\{\sum_{j \in s_d} Y_{dj} + \sum_{j \in r_d} \hat{Y}_{dj}\right\}.$$

Let us compare the MSEs of the synthetic and the BP estimators. Writing the synthetic estimator as $\hat{\bar{Y}}_d^{SYN} = \bar{\mathbf{x}}'_d\boldsymbol{\beta}$, we obtain

$$\hat{\bar{Y}}_d^{SYN} - \bar{Y}_d = \bar{\mathbf{x}}'_d\boldsymbol{\beta} - (\bar{\mathbf{x}}'_d\boldsymbol{\beta} + \bar{e}_d) = \bar{e}_d,$$

and then the MSE becomes

$$\text{MSE}(\hat{\bar{Y}}_d^{SYN}) = E\{(\hat{\bar{Y}}_d^{SYN} - \bar{Y}_d)^2\} = E(\bar{e}_d^2) = \frac{Var(e_{dj})}{N_d} = \frac{\sigma_e^2}{N_d}.$$

However, observe that the difference between the BP and the true mean is equal to

$$\hat{\bar{Y}}_d^B - \bar{Y}_d = \frac{1}{N_d}\sum_{j \in r_d} e_{dj},$$

which implies that the MSE of $\hat{\bar{Y}}_d^B$ is given by

$$\text{MSE}(\hat{\bar{Y}}_d^B) = E\{(\hat{\bar{Y}}_d^B - \bar{Y}_d)^2\} = \frac{\sigma_e^2}{N_d}\left(1 - \frac{n_d}{N_d}\right) < \frac{\sigma_e^2}{N_d} = \text{MSE}(\hat{\bar{Y}}_d^{SYN}).$$

Thus, under model (21) with known parameters, the BP has always smaller MSE due to the more efficient use of available information, namely the sample data. When the sampling fraction n_d/N_d is negligible both estimators have a similar MSE.

Now consider the case of extra area variation not explained by the auxiliary variables; that is, the true model is (13), but we fit the model (21). The true mean for area d is then given by

$$\bar{Y}_d = \bar{\mathbf{x}}'_d\boldsymbol{\beta} + u_d + \bar{e}_d.$$

Then the MSE of the synthetic estimator under the true model is

$$\text{MSE}(\hat{\bar{Y}}_d^{SYN}) = E[(u_d + \bar{e}_d)^2] = \sigma_u^2 + \frac{\sigma_e^2}{N_d}. \tag{22}$$

For the BP, the MSE under the true model is

$$\text{MSE}(\hat{\bar{Y}}_d^B) = \sigma_u^2 + \frac{\sigma_e^2}{N_d}\left(1 - \frac{n_d}{N_d}\right),$$

which is again smaller than $\text{MSE}(\hat{\bar{Y}}_d^{SYN})$.

11

The method of Elbers et al. (2003), henceforth called the ELL method, assumes a linear model with random cluster effects, where clusters may be different from the areas. They computed small area estimators by (a) fitting a linear model with cluster random effects, (b) generating bootstrap random cluster effects, (c) generating bootstrap random errors, (d) constructing a population from the bootstrap model

$$Y_{dj}^* = \mathbf{x}_{dj}'\boldsymbol{\beta} + u_c^* + e_{dj}^*, \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D, \tag{23}$$

where u_c^* is the bootstrap random effect of cluster c and (e) calculating the average of the bootstrap elements from area d:

$$\bar{Y}_d^* = N_d^{-1} \sum_{j=1}^{N_d} Y_{dj}^*. \tag{24}$$

The ELL estimator of the small area mean $\bar{Y}_d$ is then the bootstrap mean

$$\hat{\bar{Y}}_d^{ELL} = E_*(\bar{Y}_d^*),$$

which, in practice, is obtained from a Monte Carlo approximation by generating a large number, L, of populations, calculating the area means (24) for each population and later averaging over the L populations. The bootstrap variance is used as the mean squared error of the ELL estimator; that is, the ELL method uses

$$\text{MSE}(\hat{\bar{Y}}_d^{ELL}) = Var_*(\bar{Y}_d^*),$$

which again is approximated by Monte Carlo simulations.

In the case of no clusters (as in some establishment surveys), the ELL method fits the linear model (21) and uses this model to construct bootstrap populations. Then the bootstrap mean for area d becomes

$$\bar{Y}_d^* = N_d^{-1} \sum_{j=1}^{N_d} Y_{dj}^* = \frac{1}{N_d} \sum_{j=1}^{N_d} (\mathbf{x}_{dj}'\boldsymbol{\beta} + e_{dj}^*) = \hat{\bar{Y}}_d^{SYN} + \bar{e}_d^*,$$

and the ELL estimator is given by

$$\hat{\bar{Y}}_d^{ELL} = E_*(\bar{Y}_d^*) = E_*(\hat{\bar{Y}}_d^{SYN} + \bar{e}_d^*) = \hat{\bar{Y}}_d^{SYN} + E_*(\bar{e}_d^*) = \hat{\bar{Y}}_d^{SYN},$$

due to property (19) of the bootstrap method and the fact that the expectation E_* is conditional on the sample. The ELL estimator of the MSE is

$$\text{MSE}(\hat{\bar{Y}}_d^{ELL}) = Var_*(\bar{Y}_d^*) = E_*[(\bar{Y}_d^* - E_*(\bar{Y}_d^*))^2] = E_*[(\bar{e}_d^*)^2] = \frac{Var_*(e_{dj}^*)}{N_d} = \frac{\sigma_e^2}{N_d}.$$

Thus, when fitting a model without cluster effects, the ELL estimator of a small area mean is essentially the synthetic estimator, which is a good estimator when the true model is (21). When the true model is (13), the ELL method is not

12

accounting for the area effects and the MSE estimator has a bias equal to σ_u^2, see (22). Thus, the ELL estimator of MSE can lead to serious underestimation of MSE when the area effects have a substantial variance σ_u^2.

Now consider the ideal case for ELL, in which clusters are the same as the small areas. In this case we use the ELL method by fitting the correct model (13). Then the ELL estimator is again

$$\hat{\bar{Y}}_d^{ELL} = E_*(\hat{\bar{Y}}_d^{SYN} + u_d^* + \bar{e}_d^*) = \hat{\bar{Y}}_d^{SYN},$$

and the ELL estimator of the MSE becomes

$$\mathrm{MSE}(\hat{\bar{Y}}_d^{ELL}) = Var_*(\bar{Y}_d^*) = E_*[(\bar{Y}_d^* - E_*(\bar{Y}_d^*))^2] = E_*[(u_d^* + \bar{e}_d^*)^2] = \sigma_u^2 + \frac{\sigma_e^2}{N_d},$$

which is the MSE of the synthetic estimator under model (13). Thus, when the clusters are equal to the small areas, the ELL estimator remains essentially equal to the synthetic estimator, but in this case the ELL variance estimator is unbiased. Actually, under model (13), the difference between the ELL and EB methods is that the target quantities are not the same. EB method tries to estimate (or better predict) the actual area means $\bar{Y}_d$, while the ELL method is estimating instead the marginal expectations $E(\bar{Y}_d)$.

8 Model-based simulation experiment

A model-based simulation study has been carried out to study the performance of the proposed EB predictors of small area FGT poverty measures with $\alpha = 0$ (proportion of people under poverty) and $\alpha = 1$ (poverty gap). For this, we simulated populations of size $N = 20000$, composed of $D = 80$ areas with $N_d = 250$ elements in each area $d = 1, \ldots, D$. The response variables for the population units Y_{dj} were generated from model (13) taking as auxiliary variables two dummies $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1\}$ plus an intercept. The values of these two dummies for the population units were generated from Bernouilli distributions with success probabilities increasing with the area index for X_1 and constant for X_2; that is,

$$p_{1d} = 0.3 + 0.5 \, d/20; \quad p_{2d} = 0.2, \quad d = 1, \ldots, D,$$

respectively. Here the welfare variables E_{dj} are the exponential of the responses Y_{dj}; that is, the transformation $T(\cdot)$ defined in Section 5 is $T(x) = \log(x)$. A set of sample indices s was drawn from the population indices $\{1, \ldots, N\}$, by a stratified design with areas as strata and simple random sampling without replacement within each area. The values of the auxiliary variables for the population units and the sample indices were kept fixed over all Monte Carlo simulations.

The intercept and the regression coefficients associated with the two auxiliary variables used to generate populations were $\boldsymbol{\beta} = (3, 0.03, -0.04)'$. In this way,

the mean welfare increases when moving from the case $(X_1 = 0, X_2 = 0)$ to $(X_1 = 1, X_2 = 0)$, but decreases when moving from $(X_1 = 0, X_2 = 0)$ to $(X_1 = 0, X_2 = 1)$. This implies that the "richest" individuals are those with values $X_1 = 1$ and $X_2 = 0$. Since the probability p_{1d} of $X_1 = 1$ increases with the area index but that of $X_2 = 1$ is constant, then the last areas will have more individuals with larger Y_{dj} and then the FGT poverty measures will decrease with the area index. The random area effects variance was taken as $\sigma_u^2 = (0.15)^2$ and the error variance as $\sigma_e^2 = (0.5)^2$. The poverty line z was fixed as $z = 12$, which is roughly equal to 0.6 times the median of the welfare variables E_{dj} for a population generated as mentioned above. In this way, the poverty incidence for the simulated populations is approximately 16%.

Under this setup, $I = 1000$ populations $\mathbf{y}^{(i)}$ were generated from the true model. For each population i, we carried out the following tasks:

(a) The true area poverty incidences and gaps (FGT measures for $\alpha = 0$ and $\alpha = 1$ respectively) were obtained for each population as

$$F_{\alpha d}^{(i)} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{z - E_{dj}^{(i)}}{z} \right)^\alpha I(E_{dj}^{(i)} < z), \quad \alpha = 0, 1, \quad d = 1, \ldots, D,$$

where $E_{dj}^{(i)} = \exp(Y_{dj}^{(i)})$, $j = 1, \ldots, N_d$.

(b) Sample estimators of the same quantities were also calculated as

$$f_{\alpha d}^{(i)} = \frac{1}{n_d} \sum_{j \in s_d} \left(\frac{z - E_{dj}^{(i)}}{z} \right)^\alpha I(E_{dj}^{(i)} < z), \quad \alpha = 0, 1, \quad d = 1, \ldots, D.$$

In this simulation study sample estimators are equal to direct estimators.

(c) Model (13) was fitted to the sample data. Then, substituting the estimated model parameters in (15) and (16), $L = 50$ out-of-sample vectors $\mathbf{y}_r^{(i\ell)}$, $\ell = 1, \ldots, L$ were generated from the conditional distribution (14) using (17). Then using the sample data $\mathbf{y}_s^{(i)}$ and the generated out-of-sample data $\mathbf{y}_r^{(i\ell)}$, EB predictors of the area poverty incidences and gaps were calculated. For this, first, for each sample unit j we considered the sample values

$$F_{\alpha dj}^{(i)} = \left(\frac{z - E_{dj}^{(i)}}{z} \right)^\alpha I(E_{dj}^{(i)} < z), \quad j \in s_d, \quad d = 1, \ldots, D,$$

while for each out-of-sample unit j, we computed the EB predictors of $F_{\alpha dj}^{(i)}$ using the L simulated vectors,

$$\hat{F}_{\alpha dj}^{EB(i)} = \frac{1}{L} \sum_{\ell=1}^{L} \left(\frac{z - E_{dj}^{(i\ell)}}{z} \right)^\alpha I(E_{dj}^{(i\ell)} < z), \quad j \in r_d, \quad d = 1, \ldots, D,$$

14

for $\alpha = 0, 1$. Then the EB predictor of the FGT measure of order $\alpha = 0, 1$ was obtained as

$$\hat{F}_{\alpha d}^{EB(i)} = \frac{1}{N_d} \left[\sum_{j \in s_d} F_{\alpha dj}^{(i)} + \sum_{j \in r_d} \hat{F}_{\alpha dj}^{EB(i)} \right], \quad d = 1, \ldots, D.$$

(d) The ELL estimators of the poverty measures were calculated. For this, first the model (13) was fitted to the sample data and then $L = 50$ populations were generated using the parametric bootstrap described in Section 6. For each population, the poverty measures were obtained and finally, the results were averaged over the $L = 50$ populations.

Means over Monte Carlo populations $i = 1, \ldots, I$ of the true values of the FGT measures of order $\alpha = 0, 1$ were computed as

$$E(F_{\alpha d}) = \frac{1}{I} \sum_{i=1}^{I} F_{\alpha d}^{(i)}, \quad d = 1, \ldots, D.$$

These means are the reference values for comparison. Finally, for all estimators, namely, EBP, sample and ELL, the mean, variance and mean squared error over Monte Carlo populations $i = 1, \ldots, I$ were obtained for each area $d = 1, \ldots, D$.

Figure 1 a) plots the means of the true poverty incidences $F_{0d}^{(i)}$, EB estimates $\hat{F}_{0d}^{EB(i)}$, direct estimates $f_{0d}^{(i)}$ and ELL estimates $\hat{F}_{0d}^{ELL(i)}$ for each area $d = 1, \ldots, D$. Figure 1 b) plots the MSEs of EB, direct and ELL estimators. Similarly, Figures 2 a) and 2 b) plot the analogous quantities for the poverty gap F_{1d}. Observe that the means of EB estimates are for most areas close to the mean true values. On the other hand, the direct estimates seem to be too unstable (due to small sample sizes) and the ELL estimators seem too stable, that is, they do not seem to reproduce the between area variation. In terms of MSE, EB estimators are clearly much better than the other two estimators. Surprisingly, the MSEs of ELL estimators are larger than those of direct estimators.

Turning to MSE estimation, the parametric bootstrap procedure described in Section 6 was implemented with $B = 500$ replicates and the results are plotted in Figure 3. The number of Monte Carlo simulations was $I = 500$ and the true values of MSE were independently computed with $I = 50000$ Monte Carlo simulations. Observe that the bootstrap MSE estimator seems to track the pattern of the true MSE values.

9   Design-based simulation experiment

A design-based simulation experiment was also carried out to study the performance of estimators over repeated samples drawn from a fixed population. Only one population was generated as described in Section 8, with the same population and sample sizes, and using the same values of model parameters. Then,

Figure 1: a) Means ($\times 100$) and b) MSEs ($\times 10^4$) over simulated populations of true values, EB, sample and ELL estimators of the poverty incidence F_{0d} for each area d.

in each replication out of $I = 1000$, a new sample was drawn from this fixed population according to SRS without replacement within each area. From each sample, the three types of estimators of poverty measures, namely EBP, direct and ELL were obtained.

Results for the poverty incidence are displayed in Figures 4 and 5. Figure 4 shows the true values and the means over the Monte Carlo samples of the EB, direct and ELL estimators. Observe that the ELL estimators remain practically constant across the areas. On the other hand, EB estimators track the true values well, even though these estimators are supposed to have good theoretical properties with respect to the model. Of course, direct estimators perform better because they are design-unbiased. In terms of MSE, Figure 5 shows that ELL estimators have small MSEs for some of the areas and large for the other areas, while the MSE of EB and direct estimators remain small for all areas. For most areas, the MSE of EB estimators is smaller than that of the direct estimators.

10 Application

The EB method was applied to compute poverty incidences and poverty gaps in Spanish provinces crossed with gender. For this, data from the European Survey on Income and Living Conditions (EUSILC) from the year 2006 has been used. The welfare variable for the individuals is the normalized annual income calculated following the standard procedure of the Spanish Statistical Institute (INE). This variable has been transformed by adding a fixed quantity to make it always positive and then taking logarithm. This transformed variable acts as the response in the nested-error regression model. As auxiliary variables we have considered the indicators of the 5 quinquennial groupings of the variable age,
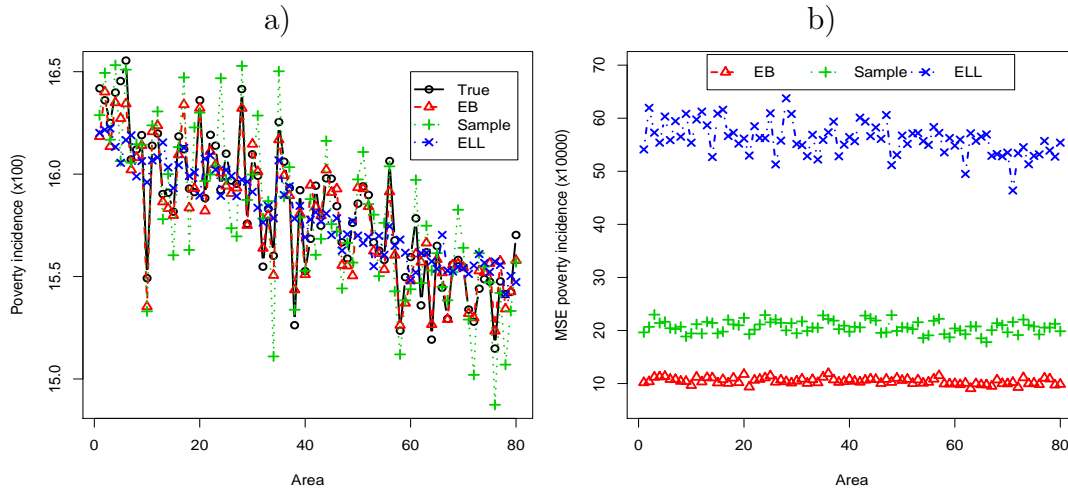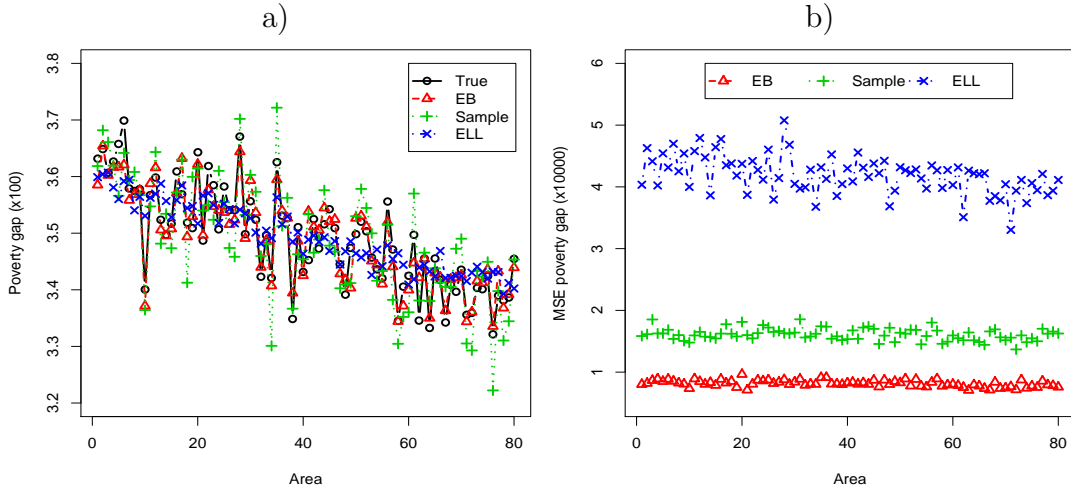
Figure 2: a) Means ($\times 100$) and b) MSEs ($\times 10^4$) over simulated populations of true values, EB, sample and ELL estimators of the poverty gaps $\hat{F}_{1d}^{E}$ for each area d.

the indicator of having Spanish nationality, the indicators of the 3 levels of the variable education level, and the indicators of the 3 categories of the variable employment, with categories "unemployed", "employed" and "inactive". From each variable, one of the categories was considered as base reference, omitting the corresponding indicator and then including an intercept in the model.

The values of the dummy indicators are not known for the out-of-sample units, but the EB method requires only the knowledge of the total number of people with the same x-values. These totals were estimated using the sampling weights attached to the sample units in the EUSILC.

The MSEs of the poverty measures were estimated by using the bootstrap procedure described in Section 6. Values of EB estimates and coefficients of variation (CVs) for the poverty incidence and the poverty gap are listed in Tables 1 and 2 respectively. In these tables we show also the direct estimates (2) and their estimated variances following standard formulas in sampling theory, but taking as observations the quantities $F_{\alpha dj}$, $j \in s_d$ and using the EUSILC sampling weights. Observe that the CVs of EB estimates are much smaller than those of direct estimators for all except few domains, in which the CVs are similar for both estimators. This improvement in efficiency is even more dramatic for the poverty gap than for the poverty incidence. Moreover, the reduction in CV tends to be greater for domains with small sample sizes. National statistical offices usually establish a maximum publishable CV. For these data, the estimated CVs of direct estimators of poverty incidences exceeded the level of 10% for 78 (out of the 104) domains while those of the EB estimators exceeded this level for only 28 domains. If we increase the level to 20%, then the direct estimators have greater CV for 17 domains but the CV of EB estimators exceeded 20% only for the first domain.

Cartograms of the estimated poverty incidences and the poverty gaps in Span-
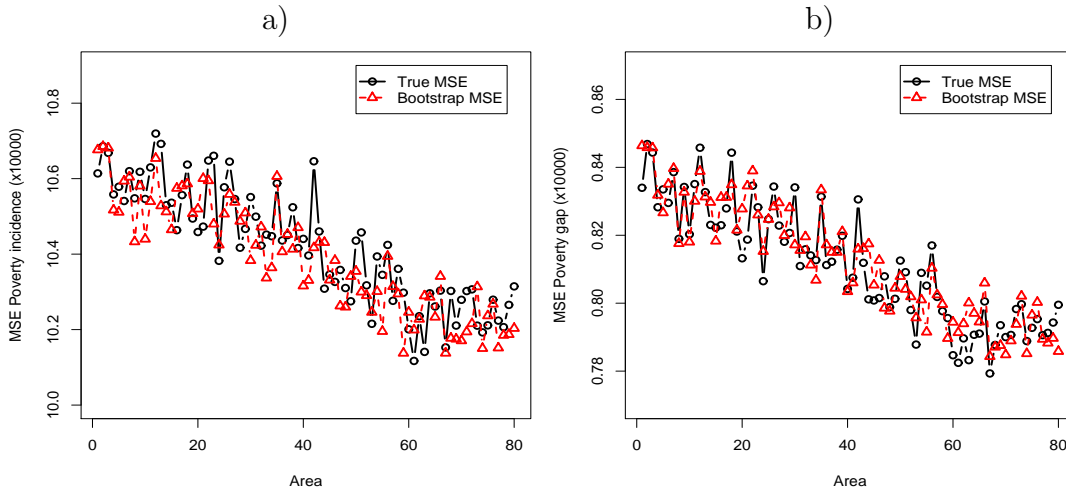
Figure 3: True MSEs ($\times 10^4$) of EB predictors and bootstrap MSE estimates with $B = 500$ for each area d: a) Poverty incidence, b) Poverty gap.

ish provinces for males and females have been constructed using the EB estimates, see Figures 6 and 7. In these maps we can see that the poorer provinces concentrate mainly in the south and west parts of Spain. Provinces with critical poverty incidences (over 30%) for men are, in the south: Almería and Córdoba; west: Badajoz, Ávila, Salamanza and Zamora and then Cuenca, situated east of Madrid. For women the poverty incidences increase in most provinces, becoming critical also, in the south: Granada, Jaén, Albacete and Ciudad Real, and in the north: Palencia and Soria. The poverty level for Lérida (north-east) seems unexpected considering that this province belongs to the region of Catalonia, which is commonly considered as a rich region.

The poverty gap measures the degree of poverty instead of the quantity of people under poverty. For a region with many people whose income is under the poverty line but very close to it, the poverty gap will be close to zero. Observe that the provinces with an income of over 12.5% under the poverty line are also among those provinces with critical values of poverty incidence, except for the northern provinces such as Lérida, which do not have significant gaps in comparison with the rest of the provinces.

11 Conclusions

In this paper Empirical Best (EB) methodology to estimate poverty measures is proposed. Parametric bootstrap is used for mean squared error (MSE) estimation. Simulation results show the good performance of EB estimators in comparison with the direct and the ELL estimators. Simulation results confirm the discussion that the ELL estimator is basically a synthetic-type estimator derived from a linear regression model.

Model (17) illustrates a parallelism between ELL and EB methods. When

18
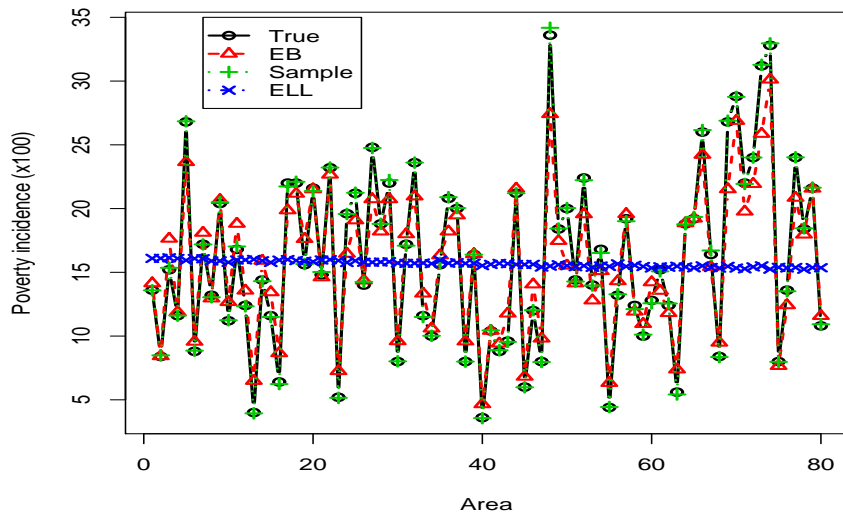
Figure 4: Means ($\times 100$) over Monte Carlo samples of true values, EB, sample and ELL estimators of the poverty incidence F_{0d} for each area d.
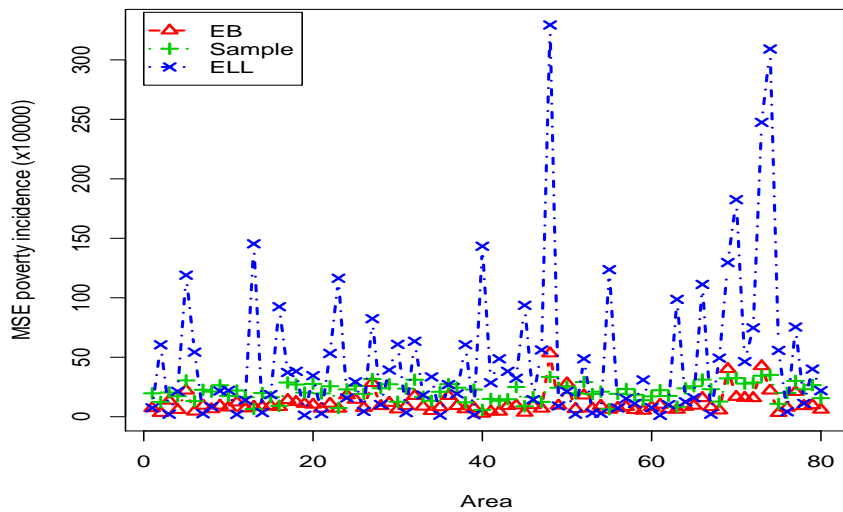
Figure 5: MSEs ($\times 10^4$) over Monte Carlo samples of EB, sample and ELL estimators of the poverty incidence F_{0d} for each area d.
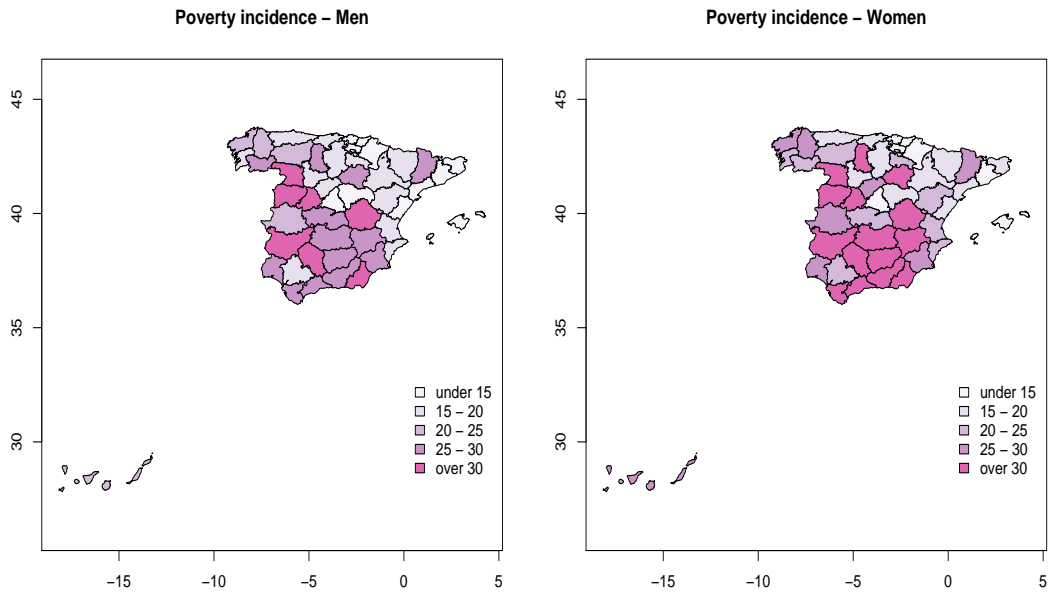
Figure 6: Cartograms of estimated percent poverty incidences in Spanish provinces for Men and Women.
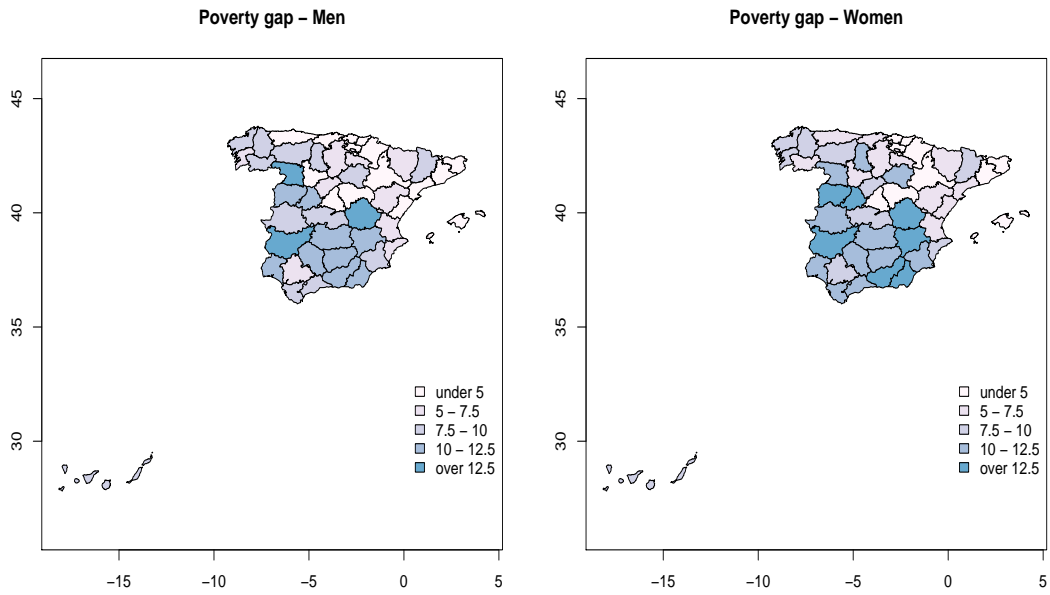
Figure 7: Cartograms of estimated percent poverty gaps in Spanish provinces for Men and Women.

the clusters in the ELL method are taken to be equal to the small areas, the ELL method generates a full population or census file of responses Y_{dj} from the bootstrap model (23) with $v_c^* = v_d^*$. Then the poverty measure is calculated from this census file. The procedure is repeated a large number of times and finally the computed poverty measures are averaged over bootstrap replications. The EB method also creates a new census file, but first plugging in the observed sample elements Y_{dj} in their corresponding place, and then generating only the non-sample values from the conditional model (17). The main difference between model (17) and bootstrap model (23) used for the ELL method is the term $\sigma_u^2 \mathbf{1}_{N_d - n_d} \mathbf{1}'_{n_d} \mathbf{V}_{ds}^{-1}(\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta})$ appearing in the conditional mean given in (16). The rest of the procedure is the same as in the ELL method. Thus, this term makes an improvement for areas that are not fully explained by auxiliary variables and therefore reduces the MSE of estimators significantly.

We remark that EB is a model-based method that relies on the validity of the model. Thus, model selection procedures and model diagnostics are essential in the practical application of this methodology.

References

Ballini, F., Betti, G. and Neri, L. (2006). Poverty and inequality mapping in the Commonwealth of Dominica. Preprint.

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.

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

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

Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

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

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008). *Journal of Statistical Computation and Simulation*, **75**, 443–462.

Hall, P. and Maiti, T. (2006). On Parametric Bootstrap Methods for Small Area Prediction. *Journal Royal Statistical Society, Series B*, **68**, 221–238.

Haslett, S. and Jones, G. (2005). Small area estimation using surveys and some practical and statistical issues. *Statistics in Transition*, **7**, 541–555.

Neri, L., Ballini, F. and Betti, G. (2005). Poverty and inequality in transition countries. *Statistics in Transition*, **7**, 135–157.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998). Weighting in Unequal Probabilities in Multilevel Models. *Journal of the Royal Statistical Society B*, **60**, 23–40.

Rao, J. N. K. (2003). *Small Area Estimation*. London: Wiley.

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

Tarozzi, A. and Deaton, A. (2007). Using census and survey data to estimate poverty and inequality for small areas. Preprint.

Application results

Table 1. Results on poverty incidence: Spanish SILC data.

Province	Sex	N_d	n_d	f_{0d}^w	$\hat{F}_{0d}^{EB}$	$\mathrm{var}(f_{0d}^w)$	$\mathrm{mse}(\hat{F}_{0d}^{EB})$	$\mathrm{CV}(f_{0d}^w)$	$\mathrm{CV}(\hat{F}_{0d}^{EB})$	Ratio
Álava	M	99354	95	7.10	12.84	0.6751	0.6805	36.60	20.32	1.80
Álava	F	108422	96	14.60	12.50	1.5637	0.6000	27.08	19.60	1.38
Albacete	M	184058	163	30.11	29.22	1.0801	0.4617	10.92	7.35	1.48
Albacete	F	186503	183	30.58	33.74	1.1558	0.4618	11.12	6.37	1.75
Alicante	M	929288	526	17.96	19.45	0.2793	0.1466	9.31	6.23	1.49
Alicante	F	931405	552	17.95	22.59	0.2482	0.1601	8.78	5.60	1.57
Almería	M	341228	204	35.34	32.88	1.3701	0.3642	10.47	5.80	1.80
Almería	F	318857	193	33.54	35.72	1.1329	0.5020	10.04	6.27	1.60
Ávila	M	56601	56	20.85	31.48	2.5112	1.2061	24.03	11.03	2.18
Ávila	F	61708	60	20.42	38.51	2.4398	1.3285	24.19	9.46	2.56
Badajoz	M	351985	472	31.97	36.56	0.5177	0.1703	7.12	3.57	1.99
Badajoz	F	346810	515	34.90	39.13	0.4958	0.1947	6.38	3.57	1.79
Baleares	M	477561	609	12.76	11.55	0.2297	0.1042	11.88	8.84	1.34
Baleares	F	472843	660	15.57	14.05	0.2228	0.1130	9.59	7.57	1.27
Barcelona	M	2617681	1358	9.82	10.49	0.0569	0.0524	7.68	6.90	1.11
Barcelona	F	2752431	1483	11.80	13.10	0.0619	0.0494	6.67	5.37	1.24
Burgos	M	215155	168	15.35	16.72	0.7776	0.3736	18.16	11.56	1.57
Burgos	F	211240	167	17.82	18.33	0.8419	0.4097	16.28	11.04	1.47
Cáceres	M	169833	261	33.20	24.69	0.8999	0.2099	9.04	5.87	1.54
Cáceres	F	184785	302	41.91	28.24	0.9402	0.2689	7.32	5.81	1.26
Cádiz	M	642053	373	34.75	26.88	0.5629	0.2013	6.83	5.28	1.29
Cádiz	F	681522	422	35.92	31.63	0.5660	0.2316	6.62	4.81	1.38
Castellón	M	201428	113	18.03	14.79	1.1401	0.6489	18.73	17.22	1.09
Castellón	F	197726	123	17.71	17.35	1.2807	0.6008	20.21	14.13	1.43
Ciudad Real	M	265393	260	27.97	28.39	0.8795	0.3060	10.60	6.16	1.72
Ciudad Real	F	254508	239	33.82	30.18	0.9979	0.3598	9.34	6.29	1.49
Córdoba	M	356218	217	37.15	30.16	1.0097	0.3636	8.55	6.32	1.35
Córdoba	F	364583	230	41.55	33.32	0.9769	0.5025	7.52	6.73	1.12
La Coruña	M	509141	457	22.89	24.66	0.3935	0.1549	8.67	5.05	1.72
La Coruña	F	563190	533	19.36	25.36	0.2888	0.1789	8.78	5.27	1.66
Cuenca	M	92275	96	35.91	35.26	3.2294	0.6676	15.82	7.33	2.16
Cuenca	F	86760	87	43.95	35.35	3.0522	0.9183	12.57	8.57	1.47
Gerona	M	307975	145	11.23	13.29	0.6006	0.4512	21.83	15.98	1.37
Gerona	F	245519	138	11.38	15.38	0.5847	0.5672	21.25	15.48	1.37
Granada	M	371735	188	31.53	29.16	0.8038	0.3423	8.99	6.35	1.42
Granada	F	424598	229	35.74	36.34	0.8358	0.3340	8.09	5.03	1.61
Guadalajara	M	87591	92	13.53	12.74	0.9309	0.6339	22.55	19.76	1.14
Guadalajara	F	79560	86	16.77	15.83	1.1738	0.9055	20.43	19.01	1.08
Guipúzcoa	M	323719	279	10.10	11.30	0.3226	0.2488	17.79	13.96	1.27
Guipúzcoa	F	348524	291	11.70	14.56	0.3297	0.2250	15.52	10.30	1.51
Huelva	M	223158	121	34.24	29.06	1.6956	0.4976	12.03	7.68	1.57
Huelva	F	214587	123	30.95	29.13	1.5825	0.5606	12.85	8.13	1.58
Huesca	M	96617	125	11.67	17.11	0.8957	0.6126	25.64	14.47	1.77
Huesca	F	91147	105	14.44	18.99	1.1769	0.8048	23.76	14.94	1.59
Jaén	M	380752	233	24.11	28.60	0.8679	0.2878	12.22	5.93	2.06
Jaén	F	356344	230	24.76	32.31	0.8523	0.4010	11.79	6.20	1.90
León	M	204462	209	18.96	22.60	0.7566	0.3772	14.51	8.59	1.69
León	F	225753	228	20.69	24.17	0.7933	0.3639	13.62	7.89	1.72

Lérida	M	214123	127	16.82	25.74	1.2317	0.6116	20.87	9.61	2.17
Lérida	F	218051	133	15.57	27.36	1.0677	0.5777	20.98	8.79	2.39
La Rioja	M	149238	519	16.75	18.57	0.3618	0.1383	11.36	6.33	1.79
La Rioja	F	147554	500	18.50	21.45	0.3613	0.1557	10.27	5.82	1.77
Lugo	M	175462	169	32.68	24.51	1.4726	0.3914	11.74	8.07	1.45
Lugo	F	167892	177	30.17	26.87	1.3821	0.4034	12.32	7.47	1.65
Madrid	M	2816184	893	7.95	12.06	0.0513	0.0619	9.01	6.52	1.38
Madrid	F	3011923	996	9.36	13.91	0.0527	0.0704	7.76	6.04	1.29
Málaga	M	693871	361	18.96	27.95	0.5823	0.2031	12.73	5.10	2.50
Málaga	F	702667	397	22.88	32.45	0.5375	0.2190	10.13	4.56	2.22
Murcia	M	668714	868	23.83	25.35	0.2471	0.1027	6.60	4.00	1.65
Murcia	F	660107	902	23.67	28.70	0.2371	0.1087	6.50	3.63	1.79
Navarra	M	286947	525	11.03	9.13	0.1812	0.1405	12.21	12.98	0.94
Navarra	F	289947	603	12.98	11.40	0.2185	0.1211	11.39	9.66	1.18
Orense	M	120257	118	24.94	25.07	1.3845	0.5993	14.92	9.77	1.53
Orense	F	137587	140	21.27	22.12	1.0968	0.4809	15.57	9.92	1.57
Oviedo	M	511169	742	12.26	16.01	0.1823	0.0824	11.01	5.67	1.94
Oviedo	F	546817	864	12.56	16.59	0.1464	0.0893	9.63	5.70	1.69
Palencia	M	75638	71	27.29	26.16	2.3375	1.0455	17.72	12.36	1.43
Palencia	F	72558	72	30.63	30.13	2.5426	1.0907	16.46	10.96	1.50
Las Palmas	M	592262	458	24.45	24.65	0.5337	0.1615	9.45	5.16	1.83
Las Palmas	F	580265	485	29.76	25.40	0.5325	0.1520	7.75	4.85	1.60
Pontevedra	M	494161	434	13.03	19.15	0.2430	0.1620	11.97	6.64	1.80
Pontevedra	F	525627	462	15.69	22.66	0.2803	0.1865	10.67	6.03	1.77
Salamanca	M	151335	166	26.92	31.46	1.2025	0.3862	12.88	6.25	2.06
Salamanca	F	152234	162	31.58	33.56	1.3392	0.4030	11.59	5.98	1.94
Tenerife	M	366253	370	17.63	24.14	0.3997	0.1590	11.34	5.22	2.17
Tenerife	F	376690	392	17.07	26.36	0.3078	0.2006	10.28	5.37	1.91
Santander	M	267290	424	8.79	16.00	0.1586	0.1398	14.33	7.39	1.94
Santander	F	279191	443	12.65	16.93	0.2339	0.1678	12.09	7.65	1.58
Segovia	M	62518	57	39.16	19.24	3.7441	1.0910	15.63	17.17	0.91
Segovia	F	63217	58	47.03	26.74	3.7211	1.2032	12.97	12.97	1.00
Sevilla	M	816795	472	22.06	19.61	0.2737	0.1575	7.50	6.40	1.17
Sevilla	F	853057	491	28.07	24.04	0.3173	0.1493	6.35	5.08	1.25
Soria	M	26431	24	23.68	26.33	8.0523	2.0666	37.89	17.26	2.19
Soria	F	17211	17	26.73	31.48	11.6416	2.7052	40.37	16.52	2.44
Tarragona	M	264627	129	12.96	14.86	0.6612	0.5761	19.85	16.15	1.23
Tarragona	F	255490	139	16.19	19.28	0.6499	0.5197	15.75	11.82	1.33
Teruel	M	53380	66	12.43	17.13	1.0073	0.8420	25.53	16.94	1.51
Teruel	F	65002	78	16.22	22.26	1.3288	1.0112	22.48	14.29	1.57
Toledo	M	288335	278	23.86	26.22	0.6157	0.1871	10.40	5.22	1.99
Toledo	F	305241	272	20.56	22.50	0.5188	0.2784	11.08	7.42	1.49
Valencia	M	1169258	686	17.53	17.89	0.1995	0.0940	8.06	5.42	1.49
Valencia	F	1197478	742	19.64	20.78	0.1978	0.1162	7.16	5.19	1.38
Valladolid	M	305496	292	14.98	15.34	0.4731	0.2216	14.52	9.70	1.50
Valladolid	F	322530	306	18.59	18.29	0.4771	0.2352	11.75	8.38	1.40
Vizcaya	M	576042	515	9.08	10.01	0.1458	0.1267	13.30	11.24	1.18
Vizcaya	F	590094	532	10.26	11.57	0.1753	0.1175	12.91	9.37	1.38
Zamora	M	101433	109	36.14	34.67	2.6728	0.7388	14.30	7.84	1.82
Zamora	F	98337	100	36.53	32.84	2.3562	0.7964	13.29	8.59	1.55
Zaragoza	M	466651	555	10.64	15.42	0.2081	0.1232	13.55	7.20	1.88
Zaragoza	F	462937	574	10.22	15.34	0.1435	0.0989	11.72	6.48	1.81
Ceuta	M	35705	223	40.87	30.26	1.5506	0.3482	9.63	6.17	1.56

Ceuta	F	40426	247	42.80	33.15	1.4583	0.3804	8.92	5.88	1.52
Melilla	M	30595	179	32.60	19.27	1.1714	0.3783	10.50	10.09	1.04
Melilla	F	27498	180	29.56	25.45	1.0739	0.5579	11.09	9.28	1.19

Columns respectively denote province, gender, population size, sample size, direct estimate of poverty incidence, EB estimate, estimated variance of direct estimator, estimated MSE of EB estimator, CV of direct estimator, CV of EB estimator and ratio of CVs of direct estimators over EB estimators. Estimated poverty incidences and CVs in percentage.

Table 2. Results on poverty gap: Spanish SILC data.

Province	Sex	N_d	n_d	f_{1d}^w	$\hat{F}_{1d}^{EB}$	$\text{var}(f_{1d}^w)$	$\text{mse}(\hat{F}_{1d}^{EB})$	$\text{CV}(f_{1d}^w)$	$\text{CV}(\hat{F}_{1d}^{EB})$	Ratio
Álava	M	99354	95	2.49	3.75	1.0904	1.1907	41.94	29.09	1.44
Álava	F	108422	96	1.53	3.65	0.4942	1.1315	45.97	29.12	1.58
Albacete	M	184058	163	9.63	10.53	2.9626	0.9286	17.87	9.15	1.95
Albacete	F	186503	183	11.72	12.68	3.5333	0.9250	16.03	7.59	2.11
Alicante	M	929288	526	5.00	6.29	0.5269	0.2937	14.53	8.62	1.68
Alicante	F	931405	552	5.89	7.55	0.6127	0.3407	13.28	7.73	1.72
Almería	M	341228	204	10.81	12.30	2.3507	0.6839	14.19	6.73	2.11
Almería	F	318857	193	11.18	13.64	2.8714	1.0880	15.16	7.65	1.98
Ávila	M	56601	56	10.82	11.64	7.0443	2.3237	24.54	13.09	1.87
Ávila	F	61708	60	12.30	15.40	6.1424	2.8775	20.15	11.01	1.83
Badajoz	M	351985	472	12.59	14.11	1.2979	0.3086	9.05	3.94	2.30
Badajoz	F	346810	515	12.15	15.46	1.0543	0.4007	8.45	4.09	2.06
Baleares	M	477561	609	2.88	3.34	0.4130	0.1955	22.28	13.24	1.68
Baleares	F	472843	660	2.94	4.23	0.2716	0.2227	17.72	11.17	1.59
Barcelona	M	2617681	1358	3.07	3.00	0.1224	0.0997	11.38	10.53	1.08
Barcelona	F	2752431	1483	3.60	3.92	0.1297	0.1027	10.00	8.17	1.22
Burgos	M	215155	168	4.22	5.21	2.2735	0.6704	35.72	15.70	2.27
Burgos	F	211240	167	3.50	5.81	1.4983	0.8377	34.93	15.75	2.22
Cáceres	M	169833	261	7.54	8.52	1.2188	0.3704	14.65	7.14	2.05
Cáceres	F	184785	302	9.33	10.13	1.2620	0.5620	12.03	7.40	1.63
Cádiz	M	642053	373	7.24	9.38	0.9284	0.4024	13.31	6.76	1.97
Cádiz	F	681522	422	10.95	11.65	1.4154	0.5101	10.87	6.13	1.77
Castellón	M	201428	113	3.97	4.48	2.8120	1.2242	42.19	24.68	1.71
Castellón	F	197726	123	3.86	5.51	2.0386	1.3159	36.97	20.83	1.77
Ciudad Real	M	265393	260	7.30	10.07	0.9995	0.6338	13.70	7.91	1.73
Ciudad Real	F	254508	239	7.15	10.86	0.9134	0.7758	13.36	8.11	1.65
Córdoba	M	356218	217	8.22	10.82	1.2822	0.6983	13.77	7.72	1.78
Córdoba	F	364583	230	8.01	12.26	1.1694	1.1819	13.50	8.87	1.52
La Coruña	M	509141	457	7.34	8.47	0.7480	0.2867	11.78	6.32	1.86
La Coruña	F	563190	533	8.33	8.72	0.8716	0.3791	11.20	7.06	1.59
Cuenca	M	92275	96	8.83	13.41	2.4195	1.4071	17.62	8.84	1.99
Cuenca	F	86760	87	10.73	13.36	3.0724	2.0791	16.33	10.80	1.51
Gerona	M	307975	145	1.87	3.95	0.5700	0.7954	40.35	22.56	1.79
Gerona	F	245519	138	2.15	4.67	0.7537	1.0857	40.30	22.29	1.81
Granada	M	371735	188	13.55	10.56	4.0423	0.6923	14.84	7.88	1.88
Granada	F	424598	229	16.81	14.02	4.8343	0.7568	13.08	6.20	2.11
Guadalajara	M	87591	92	1.52	3.80	0.2823	1.2206	34.88	29.10	1.20
Guadalajara	F	79560	86	2.55	4.90	0.4615	1.7868	26.63	27.28	0.98
Guipúzcoa	M	323719	279	2.60	3.25	0.9591	0.4277	37.69	20.09	1.88
Guipúzcoa	F	348524	291	4.42	4.38	1.3093	0.4294	25.90	14.95	1.73
Huelva	M	223158	121	10.46	10.37	7.2743	0.9412	25.78	9.36	2.75

Huelva	F	214587	123	9.13	10.40	4.2187	1.0980	22.49	10.07	2.23
Huesca	M	96617	125	2.56	5.39	1.2775	1.2615	44.18	20.86	2.12
Huesca	F	91147	105	3.04	6.06	1.7064	1.5781	42.92	20.72	2.07
Jaén	M	380752	233	9.63	10.28	1.8186	0.5968	14.01	7.51	1.86
Jaén	F	356344	230	11.41	11.94	2.1644	0.8385	12.89	7.67	1.68
León	M	204462	209	7.14	7.58	2.2850	0.7474	21.16	11.41	1.85
León	F	225753	228	7.56	8.31	2.2879	0.8014	20.00	10.77	1.86
Lérida	M	214123	127	9.22	9.08	4.8531	1.2797	23.88	12.46	1.92
Lérida	F	218051	133	9.34	9.77	4.5156	1.2979	22.75	11.66	1.95
La Rioja	M	149238	519	4.05	5.97	0.3139	0.2546	13.83	8.46	1.64
La Rioja	F	147554	500	4.34	7.14	0.2958	0.3245	12.52	7.98	1.57
Lugo	M	175462	169	8.64	8.40	6.9390	0.7199	30.50	10.10	3.02
Lugo	F	167892	177	5.26	9.40	1.3626	0.8026	22.20	9.54	2.33
Madrid	M	2816184	893	3.37	3.58	0.3812	0.1145	18.33	9.45	1.94
Madrid	F	3011923	996	3.59	4.26	0.3350	0.1442	16.14	8.92	1.81
Málaga	M	693871	361	8.95	9.90	1.9024	0.4162	15.41	6.52	2.37
Málaga	F	702667	397	10.80	12.04	1.9554	0.4561	12.95	5.61	2.31
Murcia	M	668714	868	7.54	8.74	0.4296	0.2175	8.69	5.34	1.63
Murcia	F	660107	902	8.30	10.31	0.4393	0.2373	7.99	4.73	1.69
Navarra	M	286947	525	2.99	2.53	0.3732	0.2389	20.45	19.28	1.06
Navarra	F	289947	603	2.73	3.31	0.2752	0.2450	19.23	14.96	1.29
Orense	M	120257	118	7.28	8.66	3.6924	1.1440	26.41	12.36	2.14
Orense	F	137587	140	4.77	7.44	2.0973	0.9954	30.34	13.41	2.26
Oviedo	M	511169	742	2.54	4.95	0.2335	0.1618	19.02	8.12	2.34
Oviedo	F	546817	864	3.11	5.14	0.2300	0.1857	15.41	8.38	1.84
Palencia	M	75638	71	5.65	9.10	2.9335	2.1179	30.32	15.99	1.90
Palencia	F	72558	72	6.08	10.92	3.1612	2.5262	29.27	14.56	2.01
Las Palmas	M	592262	458	7.63	8.40	1.5179	0.3192	16.15	6.73	2.40
Las Palmas	F	580265	485	8.46	8.78	1.6326	0.3139	15.11	6.38	2.37
Pontevedra	M	494161	434	3.00	6.09	0.1949	0.3100	14.73	9.14	1.61
Pontevedra	F	525627	462	4.40	7.57	0.3114	0.3888	12.68	8.24	1.54
Salamanca	M	151335	166	9.87	11.50	2.3273	0.7451	15.46	7.51	2.06
Salamanca	F	152234	162	8.85	12.74	2.0322	0.8510	16.11	7.24	2.22
Tenerife	M	366253	370	8.07	8.20	1.0073	0.2909	12.44	6.58	1.89
Tenerife	F	376690	392	9.35	9.18	1.2484	0.4136	11.95	7.00	1.71
Santander	M	267290	424	2.59	4.94	0.3045	0.2634	21.30	10.38	2.05
Santander	F	279191	443	2.95	5.29	0.3210	0.3526	19.20	11.22	1.71
Segovia	M	62518	57	7.01	6.30	4.5293	2.1717	30.36	23.41	1.30
Segovia	F	63217	58	10.90	9.64	5.5058	2.9114	21.52	17.71	1.22
Sevilla	M	816795	472	3.42	6.34	0.1609	0.2819	11.72	8.38	1.40
Sevilla	F	853057	491	4.53	8.14	0.3045	0.3112	12.19	6.85	1.78
Soria	M	26431	24	15.28	9.13	76.9805	3.9189	57.42	21.68	2.65
Soria	F	17211	17	23.46	11.84	122.9756	5.5980	47.27	19.99	2.37
Tarragona	M	264627	129	1.95	4.53	0.2800	1.0997	27.15	23.14	1.17
Tarragona	F	255490	139	2.79	6.16	0.4304	1.1317	23.52	17.26	1.36
Teruel	M	53380	66	4.48	5.49	5.9649	1.6368	54.54	23.29	2.34
Teruel	F	65002	78	5.16	7.38	3.8559	1.9972	38.09	19.15	1.99
Toledo	M	288335	278	7.69	9.18	1.3151	0.3438	14.92	6.39	2.33
Toledo	F	305241	272	5.85	7.58	0.8997	0.6082	16.21	10.28	1.58
Valencia	M	1169258	686	5.08	5.70	0.9538	0.1722	19.24	7.28	2.64
Valencia	F	1197478	742	4.26	6.78	0.3187	0.2456	13.25	7.31	1.81
Valladolid	M	305496	292	6.38	4.71	1.1430	0.4269	16.75	13.87	1.21
Valladolid	F	322530	306	7.45	5.82	1.3767	0.4782	15.76	11.89	1.33

Vizcaya	M	576042	515	2.57	2.80	0.2783	0.2338	20.49	17.27	1.19
Vizcaya	F	590094	532	2.26	3.35	0.1756	0.2177	18.56	13.92	1.33
Zamora	M	101433	109	12.58	13.10	5.5333	1.5147	18.71	9.40	1.99
Zamora	F	98337	100	9.86	12.18	4.7252	1.7185	22.04	10.76	2.05
Zaragoza	M	466651	555	4.29	4.77	0.7891	0.2377	20.69	10.23	2.02
Zaragoza	F	462937	574	5.08	4.72	0.9837	0.1956	19.53	9.38	2.08
Ceuta	M	35705	223	14.79	11.09	3.3694	0.7296	12.41	7.70	1.61
Ceuta	F	40426	247	20.68	12.52	5.5107	0.8832	11.35	7.50	1.51
Melilla	M	30595	179	11.87	6.22	7.3207	0.7442	22.80	13.86	1.64
Melilla	F	27498	180	12.47	8.82	3.5770	1.1392	15.16	12.10	1.25

Columns respectively denote province, gender, population size, sample size, direct estimate of poverty gap, EB estimate, estimated variance of direct estimator, estimated MSE of EB estimator, CV of direct estimator, CV of EB estimator and ratio of CVs of direct estimators over EB estimators. Estimated poverty gaps and CVs in percentage.