The Canadian Journal of Statistics
Vol. 38, No. 3, 2010, Pages 369–385
La revue canadienne de statistique

Small area estimation of poverty indicators

Isabel MOLINA[1]* and J. N. K. RAO[2]

[1] *Department of Statistics, Universidad Carlos III de Madrid, Getafe, Madrid 28903, Spain*
[2] *School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6*

Key words and phrases: Empirical best estimator; Parametric bootstrap; Poverty mapping.

MSC 2000: Primary 62D05; secondary 62G09.

Abstract: The authors propose to estimate nonlinear small area population parameters by using the empirical Bayes (best) method, based on a nested error model. They focus on poverty indicators as particular nonlinear parameters of interest, but the proposed methodology is applicable to general nonlinear parameters. They use a parametric bootstrap method to estimate the mean squared error of the empirical best estimators. They also study small sample properties of these estimators by model-based and design-based simulation studies. Results show large reductions in mean squared error relative to direct area-specific estimators and other estimators obtained by "simulated" censuses. The authors also apply the proposed method to estimate poverty incidences and poverty gaps in Spanish provinces by gender with mean squared errors estimated by the mentioned parametric bootstrap method. For the Spanish data, results show a significant reduction in coefficient of variation of the proposed empirical best estimators over direct estimators for practically all domains. *The Canadian Journal of Statistics* 38: 369–385; 2010 © 2010 Statistical Society of Canada

Résumé: Les auteurs proposent d'estimer les paramètres non linéaires d'une population de petits domaines en utilisant une méthode bayésienne empirique. L'emphase est mise sur les indicateurs de pauvreté comme paramètres non linéaires d'intérêt particuliers, mais ils proposent une méthodologie qui s'applique à des paramètres non linéaires plus généraux. Ils utilisent une méthode de rééchantillonnage paramétrique pour estimer l'erreur quadratique moyenne du meilleur estimateur empirique. À l'aide de simulations basées sur le modèle et sur le plan de sondage, ils étudient les propriétés de ces estimateurs pour les petits échantillons. Les résultats obtenus montrent une grande réduction de l'erreur quadratique moyenne par rapport aux estimateurs propres aux régions et les autres estimateurs obtenus par recensements « simulés». Les auteurs ont aussi appliqué la méthodologie proposée à l'estimation des incidences de pauvreté et des disparités, en fonction du sexe, du niveau de la pauvreté des provinces espagnoles. Les erreurs quadratiques moyennes sont estimées en utilisant la méthode de rééchantillonnage paramétrique citée auparavant. Pour les données espagnoles, les résultats montrent une réduction substantielle du coefficient de variation des meilleurs estimateurs empiriques proposés par rapport aux estimateurs spécifiques pour pratiquement tous les domaines. *La revue canadienne de statistique* 38: 369–385; 2010 © 2010 Société statistique du Canada

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 at a regional level is a basic instrument for targeting policies and programs aiming at the reduction of poverty. However, 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 based on auxiliary information coming from censuses or administrative registers; see Rao (2003) for a comprehensive account of these techniques.

* *Author to whom correspondence may be addressed.*
E-mail: isabel.molina@uc3m.es

Many measures of poverty and inequality are nonlinear functions of a quantitative 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 the assumed small area model, and it can be applied to estimate practically any (linear or nonlinear) function of the values of a target associated with the units of a finite population, when this variable or some transformation of it follows a linear model. However, for illustration and due to the relevance of the application, we focus on the estimation of poverty indicators. We show by simulations that EBPs of poverty indicators perform well in terms of bias and MSE. We also propose a parametric bootstrap method for MSE estimation and study its bias through simulations.

In Europe, the project EURAREA developed methods for estimation of income characteristics in small areas, see http://www.statistics.gov.uk/eurarea. Their results are restricted to linear parameters. 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. 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; for further details see http://www.census.gov/hhes/www/saipe. 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 the number 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, Lanjouw & Lanjouw (2003). This methodology is currently widely used, see, for example, Neri, Ballini & Betti (2005), Ballini, Betti, Carrette & Neri (2006), Tarozzi and Deaton (2009), and Haslett and Jones (2005). Elbers, Lanjouw & Lanjouw (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.

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

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

2. DIRECT ESTIMATORS OF FGT POVERTY MEASURES

Consider a finite population of size N 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 a fixed poverty line; that is, the threshold for E_{dj} under which a person is considered as "under poverty." Then the family of FGT poverty measures for each small area d is defined as the area mean

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

of the values $F_{\alpha dj}$ defined as

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

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 the poverty line (the poverty gap) of each individual. For $\alpha = 2$ the measure is called poverty severity. This measure squares, and large values of F_{2d} point out to areas with severe level of poverty.

Remark 1. Observe that the FGT measure for $\alpha = 0$ is equal to the empirical distribution function of the population values $\{E_{dj};\ j = 1, \ldots, N_d\}$ evaluated at point z,

$$F_{1d} = \frac{1}{N_d} \sum_{j=1}^{N_d} I(E_{dj} < z).$$

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 $s \subset \Omega$ be the set of units selected in the sample and $r = \Omega - s$ the set of indexes of the units that are not selected (with size $N - n$). The restrictions of Ω, s, and n to area d are denoted by Ω_d, s_d, and n_d, respectively, where $n = n_1 + \cdots + n_D$. Note that $n_d = 0$ if an area d is not sampled. A direct estimator for a small area uses only the sample data from the target small area. A direct estimator of $F_{\alpha d}$ for a sampled domain is the unweighted sample mean

$$\hat{F}_{\alpha d} = \frac{1}{n_d} \sum_{j \in s_d} F_{\alpha dj}, \quad \alpha = 0, 1, 2, \ d = 1, \ldots, D. \tag{3}$$

Let w_{dj} be the sampling weight (inverse of the inclusion probability) of individual j from sampled area d. Then an approximately design-unbiased estimator of $F_{\alpha d}$ is the weighted sample mean

$$\hat{F}^{w}_{\alpha d} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_{dj} F_{\alpha dj}, \quad d = 1, \ldots, D, \ \alpha = 0, 1, 2, \tag{4}$$

where $\hat{N}_d = \sum_{j \in s_d} w_{dj}$ is a design-unbiased estimator of the population size N_d of sampled area d. If the sampling weights w_{dj} do not depend on the unit j, as it occurs for example under simple random sampling within areas where $w_{dj} = n_d/N_d$, $j = 1, \ldots, N_d$, then (4) reduces to the unweighted sample mean (3).

The limited sample sizes n_d within some of the sampled areas prevent the use of estimators such as (3) or (4). Indeed, a common definition of poverty classifies a person as "under poverty" when the selected welfare variable for this person is below a given percentage of the median (the Spanish National Statistical Institute uses 60%). Under this definition, the outcome of being under poverty is likely to have low frequency and, in this case, direct estimation becomes even more inefficient. Then, reliable estimation of poverty measures for small areas requires application of small area estimation techniques (Rao, 2003). These techniques improve the estimation procedures by using models that establish some relationships among 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.

3. EB PREDICTION OF FGT POVERTY MEASURES

Consider a random vector $\mathbf{y} = (Y_1, \ldots, Y_N)'$ containing the values of a random variable associated with the N units of a finite population. Let $\mathbf{y}_s$ be the sub-vector of $\mathbf{y}$ corresponding to sample elements s and $\mathbf{y}_r$ the sub-vector of out-of-sample elements r. By reordering the units of the population, we can write $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$. 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

$$\text{MSE}(\hat{\delta}) = E_{\mathbf{y}}\{(\hat{\delta} - \delta)^2\}, \tag{5}$$

where $E_{\mathbf{y}}$ denotes expectation with respect to the joint distribution of the population vector $\mathbf{y}$. The best predictor (BP) of δ is the function of $\mathbf{y}_s$ that minimizes (5) and it is given by the conditional expectation

$$\hat{\delta}^B = E_{\mathbf{y}_r}(\delta|\mathbf{y}_s), \tag{6}$$

where the expectation is taken with respect to the conditional distribution of $\mathbf{y}_r$. Note that the BP is unbiased because

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

Typically, $\hat{\delta}^B$ depends on a vector $\boldsymbol{\theta}$ of unknown model parameters. Then an empirical BP (EBP) of δ can be obtained by replacing $\boldsymbol{\theta}$ by a suitable estimator, $\hat{\boldsymbol{\theta}}$, and then evaluating the expectation (6) at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

We now describe how to obtain BPs of FGT poverty measures for small areas. 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 $F_{\alpha dj}$ given by (2) 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 (1) is a nonlinear function of the vector $\mathbf{y}$. By taking $\delta = F_{\alpha d}$, it now follows from (6) that the BP of $F_{\alpha d}$ is

$$\hat{F}_{\alpha d}^B = E_{\mathbf{y}_r}(F_{\alpha d}|\mathbf{y}_s). \tag{7}$$

Using the decomposition of $F_{\alpha d}$ defined in (1) in terms of sample and out-of-sample elements, we have

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

where r_d denotes the set of out-of-sample elements belonging to area d. Now taking conditional expectation of (8) and introducing the conditional expectation inside the sum, the BP becomes

$$\hat{F}^{B}_{\alpha d} = \frac{1}{N_d}\left\{\sum_{j\in s_d} F_{\alpha dj} + \sum_{j\in r_d} \hat{F}^{B}_{\alpha dj}\right\}, \tag{9}$$

where $\hat{F}^{B}_{\alpha dj}$ is the BP of $F_{\alpha dj} = h_\alpha(Y_{dj})$ given by

$$\hat{F}^{B}_{\alpha dj} = E_{\mathbf{y}_r}[h_\alpha(Y_{dj})|\mathbf{y}_s] = \int_{IR} h_\alpha(y) f_{Y_{dj}}(y|\mathbf{y}_s)\,\mathrm{d}y, \quad j \in r_d. \tag{10}$$

Here $f_{Y_{dj}}(y|\mathbf{y}_s)$ is the conditional (or predictive) density of Y_{dj} given the data vector $\mathbf{y}_s$. The expectation in (10) cannot be calculated explicitly due to the complexity of $h_\alpha(y)$. 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

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

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

$$\mathbf{y}_r|\mathbf{y}_s \sim N(\boldsymbol{\mu}_{r|s}, \mathbf{V}_{r|s}), \tag{11}$$

where

$$\boldsymbol{\mu}_{r|s} = \boldsymbol{\mu}_r + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_s) \quad \text{and} \quad \mathbf{V}_{r|s} = \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr}. \tag{12}$$

Formulae (11) and (12) are valid under the assumption that sample selection bias is absent; i.e., the population model holds for the sample (Pfeffermann et al., 1998).
We propose to use an empirical approximation to (10) by Monte Carlo simulation of a large number L of vectors $\mathbf{y}_r$ generated from (11). Let $Y_{dj}^{(\ell)}$ be the value of the out-of-sample observation Y_{dj}, $j \in r_d$, obtained in the ℓth simulation, $\ell = 1, \ldots, L$. A Monte Carlo approximation to the best predictor of Y_{dj} for $j \in r_d$ is then given by

$$\hat{F}^{B}_{\alpha dj} = E_{\mathbf{y}_r}[h_\alpha(Y_{dj})|\mathbf{y}_s] \approx \frac{1}{L}\sum_{\ell=1}^{L} h_\alpha(Y_{dj}^{(\ell)}), \quad j \in r_d. \tag{13}$$

The one-dimensional integral (10) can also be evaluated by numerical quadrature methods. In practice, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{V}$ usually depend on an unknown vector of parameters $\boldsymbol{\theta}$; see Section 4. Thus the conditional density $f_{Y_{dj}}(y|\mathbf{y}_s)$ depends on $\boldsymbol{\theta}$ and we make this explicit by writing it as $f_{Y_{dj}}(y|\mathbf{y}_s; \boldsymbol{\theta})$. 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 $f_{Y_{dj}}(y|\mathbf{y}_s;\hat{\boldsymbol{\theta}})$. The resulting predictor, denoted $\hat{F}_{\alpha dj}^{\mathrm{EB}}$, is called empirical best predictor (EBP) of $F_{\alpha dj}$. Finally, the EBP of the poverty measure $F_{\alpha d}$ is given by

$$\hat{F}_{\alpha d}^{\mathrm{EB}} = \frac{1}{N_d}\left[\sum_{j\in s_d} F_{\alpha dj} + \sum_{j\in r_d} \hat{F}_{\alpha dj}^{\mathrm{EB}}\right]. \tag{14}$$

Remark 2. Instead of introducing the expectation inside the sum as in (9), the expectation in (7) can be directly approximated by Monte Carlo. This allows to estimate practically any small area parameter $\delta_d = h(\mathbf{y}_d)$, not necessarily of the separable form $\sum_j h(Y_{dj})$. Examples of parameters of interest are the area quantiles of the welfare variables $E_{dj} = T^{-1}(Y_{dj})$. The EB method for estimating a general small area parameter $\delta_d = h(\mathbf{y}_d)$ is then:

(a) Estimate the unknown parameter $\boldsymbol{\theta}$ of the distribution of the transformed vector $\mathbf{y}$ using sample data $\mathbf{y}_s$.
(b) Draw L out-of-sample vectors $\mathbf{y}_r^{(\ell)}$, $\ell = 1, \ldots, L$ from (11) and (12), with $\boldsymbol{\theta}$ replaced by the estimator $\hat{\boldsymbol{\theta}}$ obtained in (a).
(c) Augment each of the L generated vectors $\mathbf{y}_r^{(\ell)}$ with the sample data $\mathbf{y}_s$ to form a population (or "census") vector $\mathbf{y}^{(\ell)} = (\mathbf{y}_s', (\mathbf{y}_r^{(\ell)})')'$, $\ell = 1, \ldots, L$. Using the elements of $\mathbf{y}^{(\ell)}$ for the dth area, $\mathbf{y}_d^{(\ell)} = (\mathbf{y}_{ds}', (\mathbf{y}_{dr}^{(\ell)})')'$, calculate the small area parameter of interest $\delta_d^{(\ell)} = h(\mathbf{y}_d^{(\ell)})$. The Monte Carlo approximation of the EBP of δ_d is obtained by averaging the small area parameters for the L simulated populations, that is,

$$\hat{\delta}_d^{\mathrm{EB}} = \frac{1}{L}\sum_{\ell=1}^{L} \delta_d^{(\ell)}.$$

The only requirement of this method is that the distribution of some transformation $Y_{dj} = T(E_{dj})$ of the welfare variables is known and that the conditional distribution of $\mathbf{y}_r|\mathbf{y}_s$ can be derived.

4. NESTED ERROR LINEAR REGRESSION MODEL

In this section we introduce a particular super-population model ξ, the nested error linear regression model (Battese, Harter & Fuller, 1988), which can be used to evaluate the EB predictor (14). This model relates linearly, for all areas, the transformed population variables Y_{dj} (e.g., log-earnings) to vectors $\mathbf{x}_{dj}$ containing the values of p explanatory variables, and includes a random area-specific effect u_d along with the usual residual errors e_{dj}:

$$\begin{aligned}\xi:\ Y_{dj} &= \mathbf{x}_{dj}'\boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D,\\ u_d &\sim \text{iid } N(0, \sigma_u^2), \quad e_{dj} \sim \text{iid } N(0, \sigma_e^2),\end{aligned} \tag{15}$$

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

$$\mathbf{y}_d = \underset{1\le j\le N_d}{\mathrm{col}}(Y_{dj}), \quad \mathbf{e}_d = \underset{1\le j\le N_d}{\mathrm{col}}(e_{dj}), \quad \mathbf{X}_d = \underset{1\le j\le N_d}{\mathrm{col}}(\mathbf{x}_{dj}').$$

Then, the vectors $\mathbf{y}_d$, $d = 1, \ldots, D$, are independent with $\mathbf{y}_d \sim N(\boldsymbol{\mu}_d, \mathbf{V}_d)$, where

$$\boldsymbol{\mu}_d = \mathbf{X}_d\boldsymbol{\beta} \quad \text{and} \quad \mathbf{V}_d = \sigma_u^2\mathbf{1}_{N_d}\mathbf{1}_{N_d}' + \sigma_e^2\mathbf{I}_{N_d}. \tag{16}$$

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

Consider the decomposition of $\mathbf{y}_d$ into sample and out-of-sample elements $\mathbf{y}_d = (\mathbf{y}_{ds}', \mathbf{y}_{dr}')'$ when $n_d > 0$, and the corresponding decomposition of $\mathbf{X}_d$, $\boldsymbol{\mu}_d$ and $\mathbf{V}_d$. Then the distribution of $\mathbf{y}_{dr}$ given the sample data $\mathbf{y}_{ds}$ is

$$\mathbf{y}_{dr}|\mathbf{y}_{ds} \sim N(\boldsymbol{\mu}_{dr|s}, \mathbf{V}_{dr|s}), \tag{17}$$

where

$$\boldsymbol{\mu}_{dr|s} = \mathbf{X}_{dr}\boldsymbol{\beta} + \sigma_u^2\mathbf{1}_{N_d-n_d}\mathbf{1}_{n_d}'\mathbf{V}_{ds}^{-1}(\mathbf{y}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta}), \tag{18}$$

$$\mathbf{V}_{dr|s} = \sigma_u^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}_{N_d-n_d}' + \sigma_e^2\mathbf{I}_{N_d-n_d}, \tag{19}$$

for $\mathbf{V}_{ds} = \sigma_u^2\mathbf{1}_{n_d}\mathbf{1}_{n_d}' + \sigma_e^2\mathbf{I}_{n_d}$ and $\gamma_d = \sigma_u^2(\sigma_u^2 + \sigma_e^2/n_d)^{-1}$. Note that $\mathbf{y}_{dr}|\mathbf{y}_{ds}$ and $\mathbf{y}_{dr}|\mathbf{y}_s$ have the same distribution due to the independence of $\mathbf{y}_d$, $d = 1, \ldots, D$. We have assumed that the partition of Ω_d into s_d and r_d is known and that the explanatory variables $\mathbf{x}_{dj}$ associated with $j \in r_d$ are known.

Observe that the application of the Monte Carlo approximation (13) involves simulation of D multivariate Normal vectors $\mathbf{y}_{dr}$ of sizes $N_d - n_d$, $d = 1, \ldots, D$, from (17). Then this process has to be repeated L times, something computationally very intensive and even unfeasible for large N_d. This can be avoided by noting that the matrix $\mathbf{V}_{dr|s}$, given by (19), corresponds to the covariance matrix of a vector $\mathbf{y}_{dr}$ generated by the model

$$\mathbf{y}_{dr} = \boldsymbol{\mu}_{dr|s} + v_d\mathbf{1}_{N_d-n_d} + \boldsymbol{\epsilon}_{dr}, \tag{20}$$

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

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

Using (20), instead of generating a multivariate normal vector $\mathbf{y}_{dr}$ of size $N_d - n_d$, we need to generate only 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 corresponding elements Y_{dj}, $j \in r_d$, from (20) using $\boldsymbol{\mu}_{dr|s}$ given by (18). As mentioned before, in practice the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ are replaced by suitable estimators $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$, and then the variables Y_{dj} are generated from the corresponding estimated normal distributions.

If a domain d is not sampled, then $Y_{dj}^{(\ell)}$, for $j = 1, \ldots, N_d$, are generated by bootstrap from $Y_{dj} = \mathbf{x}_{dj}'\hat{\boldsymbol{\beta}} + u_d^* + e_{dj}^*$ where $u_d^* \sim \text{iid } N(0, \hat{\sigma}_u^2)$ and $e_{dj}^* \sim \text{iid } N(0, \hat{\sigma}_e^2)$, and u_d^* is independent of e_{dj}^*. Formula (13) is then used to get the estimator $\hat{F}_{\alpha dj}^{\text{EB}}$ of $F_{\alpha dj}$ and the EB estimator $F_{\alpha d}$ as

$$\hat{F}_{\alpha d}^{\text{EB}} = N_d^{-1}\sum_{j=1}^{N_d}\hat{F}_{\alpha dj}^{\text{EB}} \tag{21}$$

The estimator (21) is essentially a synthetic estimator since no sample observations are available from domain d if $n_d = 0$.

5. PARAMETRIC BOOTSTRAP MSE ESTIMATOR

The model MSE of $\hat{F}_{\alpha d}^{\mathrm{EB}}$ is given by

$$\mathrm{MSE}(\hat{F}_{\alpha d}^{\mathrm{EB}}) = E_{\xi}(\hat{F}_{\alpha d}^{\mathrm{EB}} - F_{\alpha d})^2, \tag{22}$$

where E_{ξ} denotes expectation with respect to the super-population model ξ. Note that here the target parameter $F_{\alpha d}$ is a random variable, so the usual decomposition of the MSE in terms of squared bias and variance of $\hat{F}_{\alpha d}^{\mathrm{EB}}$ does not hold. However, (22) can be decomposed as

$$\mathrm{MSE}(\hat{F}_{\alpha d}^{\mathrm{EB}}) = V_{\xi}(\hat{F}_{\alpha d}^{\mathrm{EB}} - F_{\alpha d}) + \left\{E_{\xi}(\hat{F}_{\alpha d}^{\mathrm{EB}} - F_{\alpha d})\right\}^2, \tag{23}$$

where V_{ξ} denotes model variance and $E_{\xi}(\hat{F}_{\alpha d}^{\mathrm{EB}} - F_{\alpha d})$ is the model bias of $\hat{F}_{\alpha d}^{\mathrm{EB}}$. Since the model bias of the "best" estimator $\hat{F}_{\alpha d}^{B}$ is exactly zero, the squared bias of the "empirical best" estimator $\hat{F}_{\alpha d}^{\mathrm{EB}}$ in (23) is typically very small relative to the variance of the prediction error $\hat{F}_{\alpha d}^{\mathrm{EB}} - F_{\alpha d}$ when D is large. In this case, the MSE is dominated by the variance term in (23).

Analytical approximations to the MSE are difficult to derive in the case of complex parameters such as the FGT poverty measures. We therefore obtain a parametric bootstrap MSE estimator by following the bootstrap method for finite populations of González-Manteiga et al. (2008). This bootstrap method can be readily applied to other complex parameters not necessarily of the separable form as the FGT measures. Steps for implementing this method are now given.

Step 1. Fit model (15) to sample data $(\mathbf{y}_s, \mathbf{X}_s)$ and obtain estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ of $\boldsymbol{\beta}$, σ_u^2 and σ_e^2 respectively, using a suitable method; in Section 7 we used REML method.

Step 2. Generate $u_d^* \sim$ iid $N(0, \hat{\sigma}_u^2)$, $d = 1, \ldots, D$ and, independently, generate $e_{dj}^* \sim$ iid $N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_d$, $d = 1, \ldots, D$.

Step 3. Construct the bootstrap superpopulation model ξ^* using $u_d^*, e_{dj}^*, \mathbf{x}_{dj}, j = 1, \ldots, N_d$ and $\hat{\boldsymbol{\beta}}$:

$$\xi^* : \; Y_{dj}^* = \mathbf{x}_{dj}'\hat{\boldsymbol{\beta}} + u_d^* + e_{dj}^*, \quad j = 1, \ldots, N_d, \; d = 1, \ldots, D. \tag{24}$$

Step 4. Under the bootstrap superpopulation model (24), generate a large number B of independent and identically distributed bootstrap populations $\{Y_{dj}^{*(b)};\ j = 1, \ldots, N_d,\ d = 1, \ldots, D\}$ and calculate bootstrap population parameters $F_{\alpha d}^{*(b)} = N_d^{-1} \sum_{j=1}^{N_d} F_{\alpha dj}^{*(b)}$, where $F_{\alpha dj}^{*(b)} = h_{\alpha}(Y_{dj}^{*(b)})$, $b = 1, \ldots, B$.

Step 5. From each bootstrap population b generated in Step 4, take the sample with the same indices $s \subset \Omega$ as the initial sample, and calculate the bootstrap EBPs, $\hat{F}_{\alpha d}^{EB*(b)}$, $b = 1, \ldots, B$, as described in Sections 3 and 4 using the bootstrap sample data $\mathbf{y}_s^*$ and the known population values $\mathbf{x}_{dj}$.

Step 6. A Monte Carlo approximation to the theoretical bootstrap estimator $\mathrm{MSE}_*(\hat{F}_{\alpha d}^{EB*}) = E_{\xi^*}\{(\hat{F}_{\alpha d}^{EB*} - F_{\alpha d}^*)\}^2$ of $\hat{F}_{\alpha d}^{\mathrm{EB}}$ is calculated as

$$\mathrm{mse}_*(\hat{F}_{\alpha d}^{\mathrm{EB}}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{F}_{\alpha d}^{EB*(b)} - F_{\alpha d}^{*(b)})^2. \tag{25}$$

The estimator (25) is used to estimate $\mathrm{MSE}(\hat{F}_{\alpha d}^{\mathrm{EB}})$ given in (22).

The double bootstrap method of Hall and Maiti (2006) could provide a better MSE estimator in terms of relative bias, but for large populations this method might not be computationally feasible.

We now provide a heuristic justification for the proposed bootstrap MSE estimator (25), by showing that the bootstrap population model ξ^* in (24), given the original sample data $\mathbf{y}_s$, preserves the properties of the original population model ξ in (15). It follows from the properties of u_d^* and e_{dj}^* that the bootstrap vectors $\mathbf{y}_d^*$, analogous to $\mathbf{y}_d$ for the original model, are independent with $\mathbf{y}_d^* \sim N(\hat{\boldsymbol{\mu}}_d, \hat{\mathbf{V}}_d)$, where

$$\hat{\boldsymbol{\mu}}_d = \mathbf{X}_d\hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\mathbf{V}}_d = \hat{\sigma}_u^2\mathbf{1}_{N_d}\mathbf{1}'_{N_d} + \hat{\sigma}_e^2\mathbf{I}_{N_d}. \tag{26}$$

The consistency of the parameter estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ to the true values $\boldsymbol{\beta}$, σ_u^2 and σ_e^2 ensure the consistency of the moments in (26) of bootstrap vectors $\mathbf{y}_d^*$ to their counterparts of the original vectors $\mathbf{y}_d$ given in (16). Thus, the distribution of the bootstrap population vectors $\mathbf{y}_d^*$ (given the sample data $\mathbf{y}_s$) tracks the distribution of the original population vectors $\mathbf{y}_d$.

6. ELL METHOD

The method of Elbers, Lanjouw & Lanjouw (2003), called ELL method or World Bank (WB) method, assumes a nested error linear regression model on the transformed population values, Y_{dj}, similar to (15) but using random cluster effects, where the clusters may be different from the small areas. In fact, the small areas are not specified in advance. To make it comparable with the EB method described earlier, here we assume that the clusters are the same as the small areas. Then this method basically uses the bootstrap population model ξ^* given by (24) and generates A bootstrap "censuses" $\{Y_{dj}^{*(a)};\ j = 1, \ldots, N_d,\ d = 1, \ldots, D\}$, for $a = 1, \ldots, A$. Note that the ELL census values do not contain the observed sample data in contrast to the EB method described in Remark 2.

Then, similarly as in Section 3, the bootstrap population measures $F_{\alpha d}^{*(a)}$ are calculated from each bootstrap census a and the ELL estimator of $F_{\alpha d}$ is given by

$$\hat{F}_{\alpha d}^{\mathrm{ELL}} = \frac{1}{A}\sum_{a=1}^{A} F_{\alpha d}^{*(a)} =: F_{\alpha d}^{*(\cdot)}. \tag{27}$$

For a non-sampled area d, the ELL estimator (27) is essentially equivalent to our synthetic EB estimator (21). Regardless of whether area d is a sampled area or not, note that (27) is a Monte Carlo approximation to $E_{\xi*}(F_{\alpha d}^*)$.

The MSE of $\hat{F}_{\alpha d}^{\mathrm{ELL}}$ is then estimated as

$$\mathrm{mse}(\hat{F}_{\alpha d}^{\mathrm{ELL}}) = \frac{1}{A}\sum_{a=1}^{A}(F_{\alpha d}^{*(a)} - F_{\alpha d}^{*(\cdot)})^2,$$

which is an approximation to $V_{\xi*}(F_{\alpha d}^*) = E_{\xi^*}\{F_{\alpha d}^* - E_{\xi^*}(F_{\alpha d}^*)\}^2$.

To illustrate the ELL method, consider the special case of estimating the area mean of the transformed variables, $\bar{Y}_d = N_d^{-1}\sum_{d=1}^{N_d} Y_{dj}$, assuming that all the model parameters, $\boldsymbol{\beta}$, σ_u^2 and σ_e^2 are known. In this case $\bar{Y}_d^* = \bar{\mathbf{X}}_d\boldsymbol{\beta} + u_d^* + \bar{E}_d^*$, where $\bar{\mathbf{X}}_d = N_d^{-1}\sum_{j=1}^{N_d}\mathbf{x}'_{dj}$ and $\bar{E}_d^* = N_d^{-1}\sum_{j=1}^{N_d} e_{dj}^*$. Hence, $E_{\xi^*}(\bar{Y}_d^*) = \bar{\mathbf{X}}_d\boldsymbol{\beta}$ noting that $E_{\xi^*}(u_d^*) = 0$ and $E_{\xi^*}(e_{dj}^*) = 0$. Therefore, the ELL estimator of $\bar{Y}_d$ is essentially a regression synthetic estimator, $\bar{\mathbf{X}}_d\boldsymbol{\beta}$, which is considerably less efficient than the EBP of $\bar{Y}_d$ when σ_u^2 is not small relative to σ_e^2/n_d (Rao, 2003, Chapter 7). Moreover,

$$V_{\xi^*}(\bar{Y}_d^*) = E_{\xi^*}(\bar{Y}_d^* - \bar{\mathbf{X}}_d\boldsymbol{\beta})^2 = \sigma_u^2 + \frac{\sigma_e^2}{N_d},$$

which coincides with the model variance of the true area mean,

$$V_\xi(\bar{Y}_d) = E_\xi(u_d + \bar{E}_d)^2 = \sigma_u^2 + \frac{\sigma_e^2}{N_d},$$

where $\bar{E}_d = N_d^{-1}\sum_{d=1}^{D} e_{dj}$, so that the ELL estimator of the MSE is tracking $V_\xi(\bar{Y}_d)$.

7. SIMULATION EXPERIMENTS

7.1. Model-Based Simulation Experiment

A model-based simulation study has been carried out to study the performance of the proposed EBPs of small area FGT poverty measures with $\alpha = 0$ (poverty incidence) and $\alpha = 1$ (poverty gap). For this, we simulated populations of size $N = 20{,}000$, 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 the model (15) 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; more specifically, with probabilities

$$p_{1d} = 0.3 + \frac{0.5d}{80}; \quad p_{2d} = 0.2, \quad d = 1, \ldots, D,$$

respectively. Here the welfare variables E_{dj} are exponential functions of the responses Y_{dj}; that is, the transformation $T(\cdot)$ defined in Section 3 is $T(x) = \log(x)$. A set of sample indices s_d with $n_d = 50$ was drawn independently in each area d using simple random sampling without replacement. 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 "less poor" 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, $I = 10^4$ population vectors $\mathbf{y}^{(i)}$ 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

$$F_{\alpha d}^{(i)} = \frac{1}{N_d}\sum_{j=1}^{N_d} F_{\alpha dj}^{(i)}, \quad d = 1, \ldots, D,$$

where

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

(b) A direct estimator of $F_{\alpha d}^{(i)}$ for each area d and $\alpha = 0, 1$ was calculated as

$$\hat{F}_{\alpha d}^{(i)} = \frac{1}{n_d} \sum_{j \in s_d} F_{\alpha dj}^{(i)}, \quad d = 1, \ldots, D.$$

(c) Model (15) was fitted to the sample data $(\mathbf{y}_s^{(i)}, \mathbf{X}_s)$ for each population i. Then, substituting the estimated model parameters in (18) and (19), $L = 50$ out-of-sample vectors $\mathbf{y}_r^{(i\ell)}, \ell = 1, \ldots, L$ were generated from the conditional distribution (17) using (20). The sample data $\mathbf{y}_s^{(i)}$ was attached to the generated out-of-sample data $\mathbf{y}_r^{(i\ell)}$ to form a population vector $\mathbf{y}^{(i\ell)}$. The area FGT poverty measures for $\alpha = 0, 1$ were calculated from each population $\mathbf{y}^{(i\ell)}$ as

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

where for sample units $j \in s_d$, $F_{\alpha dj}^{(i)}$ was already obtained in (a), while for out-of-sample units $j \in r_d$, $F_{\alpha dj}^{(i\ell)}$ is calculated as

$$F_{\alpha dj}^{(i\ell)} = \left(\frac{z - E_{dj}^{(i\ell)}}{z} \right)^{\alpha} I(E_{dj}^{(i\ell)} < z), \qquad E_{dj}^{(i\ell)} = \exp(Y_{dj}^{(i\ell)}), \quad j \in r_d.$$

Then the EB predictor of $F_{\alpha d}$ was calculated for each d and $\alpha = 0, 1$ as

$$\hat{F}_{\alpha d}^{EB(i)} = \frac{1}{L} \sum_{\ell=1}^{L} F_{\alpha d}^{(i\ell)}.$$

(d) ELL estimators of the FGT poverty measures for $\alpha = 0, 1$ were also calculated. For this, first model (15) was fitted to the sample data $(\mathbf{y}_s^{(i)}, \mathbf{X}_s)$ and then $A = 50$ populations or censuses were generated using the parametric bootstrap described in Section 5. For each population, the poverty measures were calculated and finally, the results were averaged over the $A = 50$ populations, as described in Section 6, to calculate the ELL estimators $\hat{F}_{\alpha d}^{\mathrm{ELL}(i)}, d = 1, \ldots, D$, for each i.

Remark 3. We used $L = A = 50$ for the EB and ELL methods in the simulation study. In practical applications of the ELL method, the choice of A ranged from 50 to 100. As for the EB method, a limited comparison with the values for $L = 50$ and $L = 1{,}000$ showed that the choice $L = 50$ gives fairly accurate results. In practice, however, when applying the EB method to a given sample data set, it is advisable to use larger values of L such as $L \geq 200$.

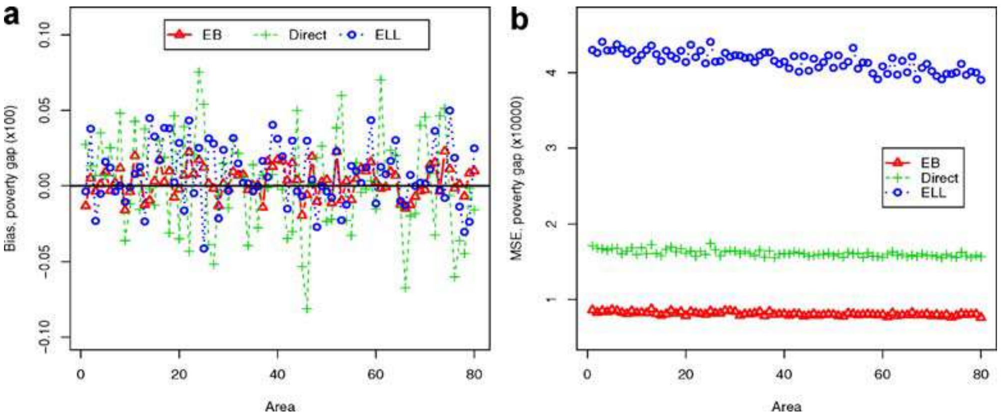

FIGURE 1: (a) Bias (× 100) and (b) MSE (× 10^4) over simulated populations of EB, direct and ELL estimators of the poverty gap F_{1d} for each area d. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

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

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

Similarly, biases over Monte Carlo populations of the three estimators, $E(\hat{F}_{\alpha d}^{\text{EB}}) - E(F_{\alpha d})$, $E(\hat{F}_{\alpha d}) - E(F_{\alpha d})$, and $E(\hat{F}_{\alpha d}^{\text{ELL}}) - E(F_{\alpha d})$, along with corresponding MSEs $E(\hat{F}_{\alpha d}^{\text{EB}} - F_{\alpha d})^2$, $E(\hat{F}_{\alpha d} - F_{\alpha d})^2$, and $E(\hat{F}_{\alpha d}^{\text{ELL}} - F_{\alpha d})^2$, were computed.

Figure 1a and b reports, respectively, the biases and the MSEs of the estimators for the poverty gap ($\alpha = 1$). Figure 1a shows that the EB estimator has the smallest absolute bias followed by ELL and the direct estimator, but compared to the corresponding values of MSE (Figure 1b)), the square of the model bias is negligible for all the three estimators. Hence, the MSE of the estimators considered here is dominated by the model variance of the prediction error, as explained in Section 5. It is clear from Figure 1b that the EB estimator is significantly more efficient than ELL and direct estimators. Surprisingly, Figure 1b also reveals that the ELL estimator is less efficient than the direct estimator, showing that the prediction error variance is larger for the ELL method. Results for the poverty incidence ($\alpha = 0$) were similar and are not reported here.

Turning to MSE estimation, the parametric bootstrap procedure described in Section 5 was implemented with $B = 500$ replicates and the results are plotted in Figure 2 for the poverty gap ($\alpha = 1$). The number of Monte Carlo simulations was $I = 500$ and the true values of the MSE were independently computed with $I = 50{,}000$ Monte Carlo simulations. Figure 2 shows that the bootstrap MSE estimator tracks the pattern of the true MSE values. Similar results were observed for the poverty incidence ($\alpha = 0$).

7.2. 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 in the same way as in Section 7.1, with the same population and sample sizes, and using the same values of model parameters. Then, in each replication out of $I = 1{,}000$, 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 EB, direct and ELL were obtained.

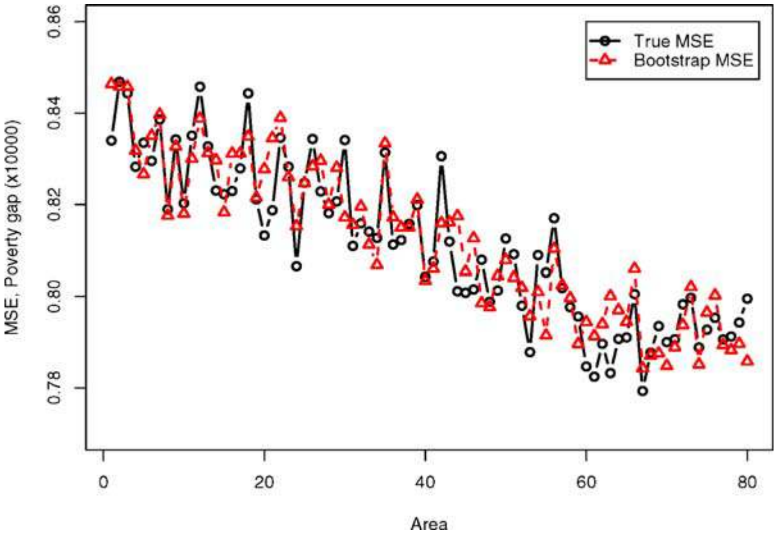

FIGURE 2: True MSE ($\times 10^4$) of EB estimators of poverty gap ($\alpha = 1$) and bootstrap MSE estimate with $B = 500$ for each area d. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

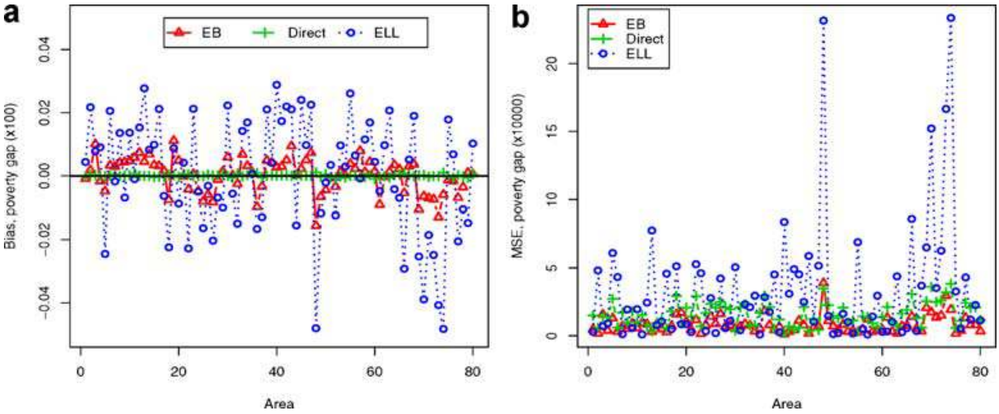

FIGURE 3: (a) Bias ($\times 100$) and (b) MSE ($\times 10^4$) of EB, direct and ELL estimators of the poverty gap F_{1d} for each area d under the design-based setup. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Results on the design bias and design MSE of the estimators for poverty gap ($\alpha = 1$) are reported in Figure 3a and b, respectively. As expected, Figure 3a shows that the Monte Carlo design bias of the direct estimator is practically zero, followed by EB and ELL estimators. In terms of MSE, Figure 3b shows that ELL estimators have small MSEs for some of the areas and very large for other areas, while the MSE of EB and direct estimators remain small for all areas. For most areas, the MSE of EB estimator is smaller than that of direct estimator.

8. APPLICATION

The EB method was applied to compute poverty incidences and poverty gaps by gender in Spanish provinces. For this, data from the European Survey on Income and Living Conditions (EUSILC)

TABLE 1: Population size, sample size, direct and EB estimates of poverty incidences (×100), estimated MSEs of direct and EB estimators (×10^4) and CVs of direct and EB estimators (×100) for the Spanish domains with sample size closest to minimum, first quartile, median, third quartile, and maximum.

Province	Gen	N_d	n_d	$\hat{F}^{w}_{0d}$	$\hat{F}^{\mathrm{EB}}_{0d}$	$\mathrm{var}(\hat{F}^{w}_{0d})$	$\mathrm{mse}_*(\hat{F}^{\mathrm{EB}}_{0d})$	$\mathrm{cv}(\hat{F}^{w}_{0d})$	$\mathrm{cv}(\hat{F}^{\mathrm{EB}}_{0d})$
Soria	F	17,211	17	60.41	31.48	158.6708	27.0518	20.85	16.52
Tarragona	M	264,627	129	12.46	14.86	8.5695	5.7605	23.50	16.15
Córdoba	F	364,583	230	30.66	33.32	10.7598	5.0252	10.70	6.73
Badajoz	M	351,985	472	36.58	36.56	6.1853	1.7031	6.80	3.57
Barcelona	F	2,752,431	1,483	10.82	13.10	0.6605	0.4944	7.51	5.37

from the year 2006 has been used. The welfare variable for the individuals is the equivalized annual net 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 variable in the nested-error regression model. As auxiliary variables, we have considered the indicators of the five quinquennial groupings of the variable age, the indicator of having Spanish nationality, the indicators of the three levels of the variable education level, and the indicators of the three categories of the variable employment, with categories "unemployed," "employed," and "inactive." For each auxiliary 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 using the parametric bootstrap estimator $\mathrm{mse}_*(\hat{F}^{\mathrm{EB}}_{\alpha d})$ given by (25) with $B = 500$ replicates. Values of EB estimators, $\hat{F}^{\mathrm{EB}}_{\alpha d}$, and associated coefficients of variation (CVs) for the poverty incidence ($\alpha = 0$) and the poverty gap ($\alpha = 1$) are listed respectively in Tables 1 and 2 for a few representative domains (provinces × gender), where $\mathrm{cv}(\hat{F}^{\mathrm{EB}}_{\alpha d}) = \{\mathrm{mse}_*(\hat{F}^{\mathrm{EB}}_{\alpha d})\}^{1/2}/\hat{F}^{\mathrm{EB}}_{\alpha d}$. Tables 1 and 2 also show the values of the direct estimator (4) and the 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. Tables with full results for all domains can be found in Molina and Rao (2009). The CVs of EB estimators are much smaller than those of the direct estimators for practically all domains. Moreover, the reduction in CV tends to be greater for domains with smaller sample sizes. National statistical

TABLE 2: Population size, sample size, direct and EB estimates of poverty gaps (×100), estimated MSEs of direct and EB estimators (×10^4) and CVs of direct and EB estimators (×100) for the Spanish domains with sample size closest to minimum, first quartile, median, third quartile, and maximum.

Province	Gen	N_d	n_d	$\hat{F}^{w}_{1d}$	$\hat{F}^{\mathrm{EB}}_{1d}$	$\mathrm{var}(\hat{F}^{w}_{1d})$	$\mathrm{mse}_*(\hat{F}^{\mathrm{EB}}_{1d})$	$\mathrm{cv}(\hat{F}^{w}_{1d})$	$\mathrm{cv}(\hat{F}^{\mathrm{EB}}_{1d})$
Soria	F	17,211	17	23.46	11.84	122.9756	5.5980	47.27	19.99
Tarragona	M	264,627	129	1.95	4.53	0.2800	1.0997	27.15	23.14
Córdoba	F	364,583	230	8.01	12.26	1.1694	1.1819	13.50	8.87
Badajoz	M	351,985	472	12.59	14.11	1.2979	0.3086	9.05	3.94
Barcelona	F	2,752,431	1,483	3.60	3.92	0.1297	0.1027	10.00	8.17

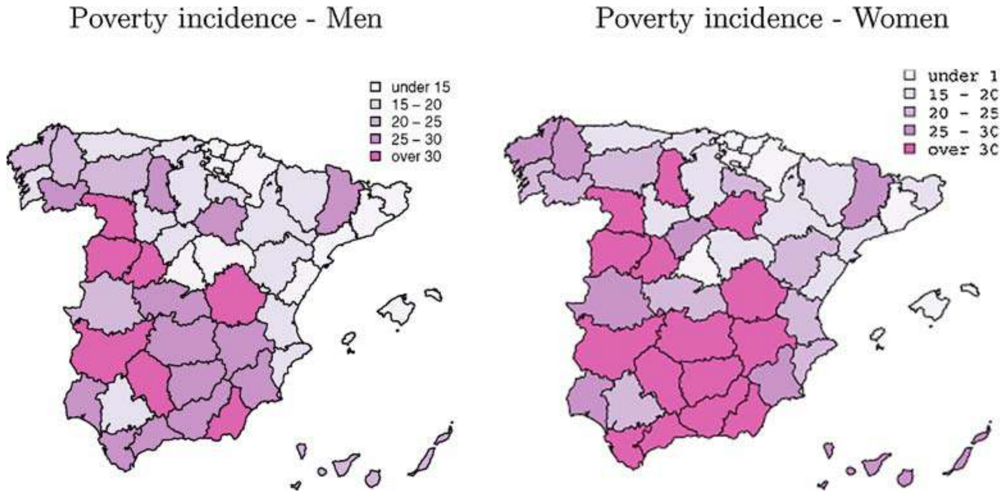

FIGURE 4: Cartograms of estimated percent poverty incidences in Spanish provinces for Men and Women. Canary islands have been moved from their original position in the map to the bottom-right corner. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

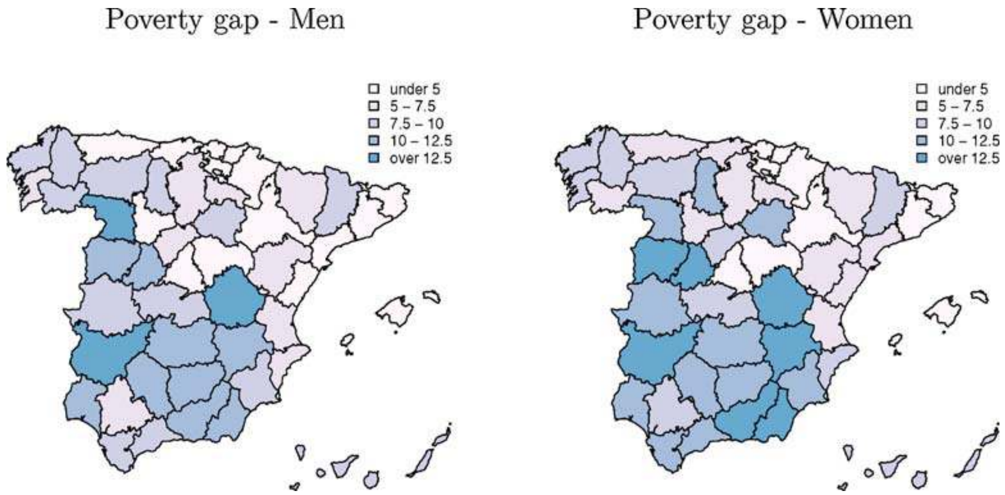

FIGURE 5: Cartograms of estimated percent poverty gaps in Spanish provinces for Men and Women. Canary islands have been moved from their original position in the map to the bottom-right corner. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

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 87 (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 25 domains but the CVs of EB estimators exceeded 20% only for one domain.

Cartograms of the estimated poverty incidences and the poverty gaps in Spanish provinces for males and females have been constructed using the EB estimates, see Figures 4 and 5. In these maps we can see that the poorer provinces concentrate mainly in the southern and western parts of

Spain. Provinces with critical poverty incidences (over 30%) for men are, in the south: Almería and Córdoba; west: Badajoz, Ávila, Salamanca 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 Lleida (north-east) seems unexpected considering that this province belongs to the region of Catalonia, which is commonly considered as a "richer" region.

The poverty gap measures the degree of poverty instead of the number 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.

9. CONCLUSIONS

In this paper Empirical Best (EB) methodology to estimate poverty measures for small areas is proposed. A parametric bootstrap method is used for mean squared error (MSE) estimation. Simulation results show good performance of EB estimators in comparison with direct and ELL estimators.

Model (20) illustrates a parallelism between ELL and EB methods. When the clusters in ELL method are taken to be equal to the small areas, ELL method generates a full population or census file of responses Y_{dj} from the bootstrap model (24). Then the poverty measure is calculated from this census file. The bootstrap procedure is replicated several times and the computed poverty measures are averaged over bootstrap replications. The EB method also creates new census files, but first plugging in the observed sample elements Y_{dj} in their corresponding place, and then generating only the out-of-sample values from the conditional model (20). The main difference between model (20) and bootstrap model (24) 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 (18). The rest of the procedure is the same as in the ELL method. Thus, this term provides an improvement for sampled areas that are not fully explained by auxiliary variables and therefore reduces the MSE of estimators significantly.

Note that the ELL method uses the census values $\mathbf{x}_{dj}$, $j \in \Omega_d$ without requiring the knowledge of the partition into s_d and r_d, unlike our EB method. This feature of ELL method may be attractive to users since the users are not required to identify the non-sampled $\mathbf{x}_{dj}$ in the census x-file. But, as shown in Section 7, the ELL method pays a heavy price for sampled areas in terms of efficiency and it can be even less efficient than a direct area-specific estimator.

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.

In simulations, EB method has also been applied to estimate various other small area measures, including non-separable functions such as quantiles and the Gini coefficient. The conclusions drawn here with respect to bias and MSE of EB estimators in comparison to direct and ELL estimators seem to hold also for other measures.

We have also developed hierarchical Bayes (HB) methods, jointly with B. Nandram, for making inferences on poverty indicators and other measures for small areas, using an efficient Bayesian sampling-based approach for nested error linear regression models. HB methods provide "exact" inferences via the posterior distribution of the desired population measures, including posterior credible intervals on the parameters of interest. These results will be reported in a separate paper. We are also studying the use of survey weights by developing pseudo-EB estimators, similar to those of You and Rao (2002) and Jiang and Lahiri (2006), for the area means.

ACKNOWLEDGEMENTS

The authors wish to thank the referees, associate editor, Domingo Morales and P.L.N. Silva for constructive suggestions. This research was supported by Spanish grants SEJ2007-64500, MTM2009-09473, the European project SAMPLE, FP7-SSH-2007-1 and a grant to J.N.K. Rao from the Natural Sciences and Engineering Research Council of Canada.

BIBLIOGRAPHY

F. Ballini, G. Betti, S. Carrette & L. Neri (2006). Poverty and inequality mapping in the Commonwealth of Dominica. *Estudios Economicos*, 2, 123–162.

G. E. Battese, R. M. Harter & W. A. Fuller (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.

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

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

R. E. Fay & R. A. Herriot (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.

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

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

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

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

J. Jiang & P. Lahiri (2006). Estimation of finite population domain means: a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301–311.

I. Molina & J. N. K. Rao (2009). Small area estimation of poverty indicators. Working Paper 09-15, Statistics and Econometric Series 05, Universidad Carlos III de Madrid, http://hdl.handle.net/10016/5644.

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

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

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

A. Tarozzi & A. Deaton (2009). Using census and survey data to estimate poverty and inequality for small areas. *Review of Economics and Statistics*, 91, 773–792.

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

Received 6 April 2009
Accepted 27 November 2009