SORT 39 (1) January-June 2015, 19-34

Small area estimation of poverty indicators under partitioned area-level time models*

Domingo Morales[1], Maria Chiara Pagliarella[2] and Renato Salvatore[2]

Abstract

This paper deals with small area estimation of poverty indicators. Small area estimators of these quantities are derived from partitioned time-dependent area-level linear mixed models. The introduced models are useful for modelling the different behaviour of the target variable by sex or any other dichotomic characteristic. The mean squared errors are estimated by explicit formulas. An application to data from the Spanish Living Conditions Survey is given.

MSC: 62D05, 62J05.

Keywords: Area-level models, small area estimation, time correlation, poverty indicators.

1. Introduction

In most European countries, the estimation of poverty is done by using the Living Conditions Survey (LCS) data. The Spanish LCS (SLCS) uses a stratified two-stage design within each Autonomous Community. As most provinces have a very small sample size, the direct estimates at that level have a low accuracy. The problem is thus that domain sample sizes are too small to carry out direct estimations. This situation may be treated by using small area estimation techniques. Small Area Estimation (SAE) is a part of the statistical science that combines survey sampling and finite population inference with statistical models. See a description of this theory in the monograph of Rao (2003), or in the reviews of Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002, 2012) and more recently Jiang and Lahiri (2006).

* Supported by the grants EU-FP7-SSH-2007-1, MTM2012-37077-C02-01.

[1] Centro de Investigación Operativa, Universidad Miguel Hernández de Elche, Spain. d.morales@umh.es

[2] Dipartimento di Economia Politica e Statistica, Università degli Studi di Siena, Italy. mariachiara.pagliarella@unisi.it

[3] Dipartimento di Economia e Giurisprudenza, Università di Cassino e del Lazio Meridionale, Italy. rsalvatore@unicas.it

Received: April 2014

Accepted: September 2014

This paper deals with the estimation of poverty indicators by using area-level models. For this sake, Esteban et al. (2012a,b) proposed several area-level time models. They argue that employing data from past periods produce a significant improvement of the estimation process. Marhuenda et al. (2013) introduced some more complex area-level linear mixed models that take into account for temporal and spatial correlation. The first two papers gave empirical best linear unbiased prediction (EBLUP) estimates of poverty estimators for Spanish provinces crossed by sex. The third one did not give estimates by sex. Many socio-economic indicators, such as those related with poverty and labour, behave differently in the subpopulations of men and women. This is why, we adapt some of the temporal models appearing in Esteban et al. (20121,b) and Marhuenda et al. (2013) to this situation.

In this paper we use four time-dependent area-level linear mixed models to obtain small area estimates of poverty indicators. Two of them are specified with a partition of the population in two groups. This fact allows modelling, for example, a different behaviour of the target variable by sex, as it was done by Herrador et al. (2011). This is an important modelling tool as many socioeconomic indicators behave differently for men and women. Following Esteban et al. (2012b), the first partitioned model assumes that time dependency is explained by the auxiliary variables and the second one contains a correlation parameter in the distribution of the random intercept. The estimates of the model parameters are obtained by using the residual maximum likelihood (REML) estimation method. These estimates are then used to construct empirical best linear unbiased predictors of poverty indicators by sex of the Spanish provinces. Estimation of the mean squared error (MSE) of model-based estimators is an important issue that has no easy solution. In this paper we follow Prasad and Rao (1990) and Das, Jiang and Rao (2004) to introduce an approximation of the MSE and the corresponding MSE estimator.

The rest of the paper is organized as follows. Section 2 introduces the considered area-level time models and the corresponding model-based estimators of poverty indicators. Section 3 describes the estimation problem of interest and presents an application to data from the SLCS. The target is to estimate poverty indicators by sex in the Spanish provinces. Finally, Section 4 gives a discussion on the findings of this paper.

2. The area-level partitioned time models

2.1. The models

Let us consider a population partitioned in D domains. Assume that domains are classified in two groups of sizes D_A and D_B ($D_A + D_B = D$) that behave differently with respect to some socioeconomic characteristic. For example, let us consider a country divided in provinces. Assume that a statistical agency is interested in estimating some poverty indicators of regions by sex. In that situation, they can define the domains as regions crossed by sex, so that they have $D_A = D_B$ and $D = 2D_A = 2D_B$. Another example is a state partitioned in D_A urban-type counties and D_B rural-type counties, where the interest is the estimation of some labour indicators at the county level. In what follows, we will introduce some models adapted to these kind of situations.

Let us consider the model (model 3)

$$y_{dt} = \mathbf{x}_{dt}^{\mathsf{T}} \boldsymbol{\beta} + u_{dt} + e_{dt}, \quad d = 1, \dots, D = D_A + D_B, \quad t = 1, \dots, m_d, \tag{1}$$

where y_{dt} is a direct estimator of the indicator of interest for area d and time instant t, and $\mathbf{x}_{dt}^{\mathsf{T}}$ is a row vector containing the aggregated (population) values of p auxiliary variables. The index d is used for domains and the index t for time instants. We assume that the random vectors $(u_{d1}, \dots, u_{dm_d})$, $d \leq D_A$, follow independent and identically distributed (i.i.d.) first order auto-regressive processes with variance and auto-correlation parameters σ_A^2 and ρ_A respectively; in short, $(u_{d1}, \dots, u_{dm_d}) \sim_{iid} AR1(\sigma_A^2, \rho_A)$, $d \leq D_A$. We further assume that $(u_{d1}, \dots, u_{dm_d}) \sim_{iid} AR1(\sigma_B^2, \rho_B)$, $d > D_A$, and that the errors e_{dt}'s are independent $N(0, \sigma_{dt}^2)$ with known variances σ_{dt}^2's. Finally we assume that the $(u_{d1}, \dots, u_{dm_d})$'s and the e_{dt}'s are mutually independent.

The introduction of the partitioned model (1) is motivated by the observed different behaviour by sex of poverty indicators in Spanish data. Further, we also consider the models restricted to $\rho_A = \rho_B$ (model 2), restricted to $\rho_A = \rho_B = 0$ (model 1) and restricted to $\rho_A = \rho_B = 0$ and $\sigma_A^2 = \sigma_B^2$ (model 0). For the sake of brevity, we only present the theoretical developments for the partitioned model 3.

In matrix notation the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where $\mathbf{y}$ can be decomposed in the form $\mathbf{y} = (\mathbf{y}_A^{\mathsf{T}}, \mathbf{y}_B^{\mathsf{T}})^{\mathsf{T}}$, with $\mathbf{y}_A = \underset{d \leq D_A}{\mathrm{col}}(\mathbf{y}_d)$, $\mathbf{y}_B = \underset{d > D_A}{\mathrm{col}}(\mathbf{y}_d)$ and $\mathbf{y}_d = \underset{1 \leq t \leq m_d}{\mathrm{col}}(y_{dt})$, and similarly for $\mathbf{u}$ and $\mathbf{e}$, $\mathbf{X}$ can be decomposed in the form $\mathbf{X} = (\mathbf{X}_A^{\mathsf{T}}, \mathbf{X}_B^{\mathsf{T}})^{\mathsf{T}}$, with $\mathbf{X}_A = \underset{d \leq D_A}{\mathrm{col}}(\mathbf{X}_d)$, $\mathbf{X}_B = \underset{d > D_A}{\mathrm{col}}(\mathbf{X}_d)$ and $\mathbf{X}_d = \underset{1 \leq t \leq m_d}{\mathrm{col}}(\mathbf{x}_{dt}^{\mathsf{T}})$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{p \times 1}$, $\mathbf{Z} = \mathbf{I}_M$ and $M = \sum_{d=1}^{D} m_d$. We use the notation col$(\cdots)$ to denote a column vector, or set of column vectors, composed of the elements of the argument, which can be scalars or vectors. In this notation, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{V}_u)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V}_e)$ are independent with covariance matrices

$$\mathbf{V}_u = \mathrm{var}(\mathbf{u}) = \mathrm{diag}(\sigma_A^2 \Omega_A, \sigma_B^2 \Omega_B), \quad \mathbf{V}_e = \mathrm{var}(\mathbf{e}) = \underset{1 \leq d \leq D}{\mathrm{diag}}(\mathbf{V}_{ed}),$$

where $\Omega_A = \underset{d \leq D_A}{\mathrm{diag}}(\Omega_d)$, $\Omega_B = \underset{d > D_A}{\mathrm{diag}}(\Omega_d)$, $\mathbf{V}_{ed} = \underset{1 \leq t \leq m_d}{\mathrm{diag}}(\sigma_{dt}^2)$ and

$$\Omega_d = \Omega_d(\rho) = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \cdots & \rho^{m_d-2} & \rho^{m_d-1} \\ \rho & 1 & \ddots & & \rho^{m_d-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{m_d-2} & & \ddots & 1 & \rho \\ \rho^{m_d-1} & \rho^{m_d-2} & \cdots & \rho & 1 \end{pmatrix}_{m_d \times m_d}, \quad \rho = \rho_A, \rho_B.$$

The covariance matrix of vector $\mathbf{y}$ is $\mathbf{V}=\mathbf{V}(\boldsymbol{\theta})=\text{var}(\mathbf{y})=\text{diag}(\mathbf{V}_A,\mathbf{V}_B)$, where $\mathbf{V}_A=\underset{d\leq D_A}{\text{diag}}(\mathbf{V}_d)$, $\mathbf{V}_B=\underset{d>D_A}{\text{diag}}(\mathbf{V}_d)$, $\mathbf{V}_d=\sigma_A^2\Omega_d+\mathbf{V}_{ed}$ if $d\leq D_A$, $\mathbf{V}_d=\sigma_B^2\Omega_d+\mathbf{V}_{ed}$ if $d>D_A$ and $\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3,\theta_4)=(\sigma_A^2,\rho_A,\sigma_B^2,\rho_B)$. The residual loglikelihood is

$$\begin{aligned}l_{reml}=l_{reml}(\boldsymbol{\theta})=&-\frac{M-p}{2}\log 2\pi+\frac{1}{2}\log|\mathbf{X}^{\mathsf{T}}\mathbf{X}|-\frac{1}{2}\log|\mathbf{V}_A|-\frac{1}{2}\log|\mathbf{V}_B|\\&-\frac{1}{2}\log|\mathbf{X}_A^{\mathsf{T}}\mathbf{V}_A^{-1}\mathbf{X}_A+\mathbf{X}_B^{\mathsf{T}}\mathbf{V}_B^{-1}\mathbf{X}_B|-\frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{P}\mathbf{y},\end{aligned}$$

where $\mathbf{P}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}$. The scores and the Fisher information matrix components are

$$S_a=\frac{\partial l_{reml}}{\partial\theta_a},\quad F_{ab}=-E\left[\frac{\partial l_{reml}^2}{\partial\theta_a\partial\theta_b}\right],\quad a,b=1,2,3,4.$$

To calculate the residual maximum likelihood (REML) estimate, $\hat{\boldsymbol{\theta}}$, we apply the Fisher-scoring algorithm with the updating formula

$$\boldsymbol{\theta}^{k+1}=\boldsymbol{\theta}^{k}+\mathbf{F}^{-1}(\boldsymbol{\theta}^{k})\mathbf{s}(\boldsymbol{\theta}^{k}),$$

where $\mathbf{s}$ and $\mathbf{F}$ are the column vector of scores and the Fisher information matrix respectively. As seeds we use $\rho_A^{(0)}=\rho_B^{(0)}=0$, and $\sigma_A^{2(0)}=\sigma_B^{2(0)}=\hat{\sigma}_{uH}^2$, where $\hat{\sigma}_{uH}^2$ is the Henderson 3 estimator under model with $\rho_A=\rho_B=0$ and $\sigma_A^2=\sigma_B^2$. The REML estimator of $\boldsymbol{\beta}$ and the REML empirical best linear unbiased predictor (EBLUP) of $\mathbf{u}$ are

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^{\mathsf{T}}\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\hat{\mathbf{V}}^{-1}\mathbf{y},\quad\hat{\mathbf{u}}=\hat{\mathbf{V}}_u\mathbf{Z}^{\mathsf{T}}\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\hat{\mathbf{V}}=\mathbf{V}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{V}}_u=\mathbf{V}_u(\hat{\boldsymbol{\theta}})$.

2.2. Statistical inference on the model parameters

The asymptotic distributions of of the REML estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are

$$\hat{\boldsymbol{\theta}}\sim N_4(\boldsymbol{\theta},\mathbf{F}^{-1}(\boldsymbol{\theta})),\quad\hat{\boldsymbol{\beta}}\sim N_p(\boldsymbol{\beta},(\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}).$$

Asymptotic confidence intervals at the level $1-\alpha$ for θ_a and β_j are

$$\hat{\theta}_a\pm z_{\alpha/2}\,\nu_{aa}^{1/2},\ a=1,2,3,4,\quad\hat{\beta}_j\pm z_{\alpha/2}q_{jj}^{1/2},\ j=1,\ldots,p,$$

where $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{\kappa}$, $\mathbf{F}^{-1}(\boldsymbol{\theta}^{\kappa}) = (\nu_{ab})_{a,b=1,2,3,4}$, $(\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}(\boldsymbol{\theta}^{\kappa})\mathbf{X})^{-1} = (q_{ij})_{i,j=1,\dots,p}$, κ is the final iteration of the Fisher-scoring algorithm and z_α is the α-quantile of the standard normal distribution $N(0,1)$. Observed $\hat{\beta}_j = \beta_0$, the p-value for testing the hypothesis $H_0: \beta_j = 0$ is

$$p = 2P_{H_0}(\hat{\beta}_j > |\beta_0|) = 2P(N(0,1) > \beta_0/\sqrt{q_{jj}}).$$

Let $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, $\hat{\rho}_A$ and $\hat{\rho}_B$ be the unrestricted REML estimators of σ_A^2 and σ_B^2, ρ_A and ρ_B respectively. Let $\tilde{\sigma}_A^2$, $\tilde{\sigma}_B^2$ and $\tilde{\rho}$ be the REML estimator of σ_A^2, σ_B^2 and of the common value $\rho_A = \rho_B$ under H_0 (model 2). Under model 3, the REML likelihood ratio statistic (LRS) for testing $H_0: \rho_A = \rho_B$ is

$$\lambda = -2[l_{REML}(\tilde{\sigma}_A^2, \tilde{\sigma}_B^2, \tilde{\rho}) - l_{REML}(\hat{\sigma}_A^2, \hat{\sigma}_B^2, \hat{\rho}_A, \hat{\rho}_B)].$$

The asymptotic distribution of λ under H_0 is χ_1^2. The null hypothesis is rejected at the level α if $\lambda > \chi_{1,\alpha}^2$.

Under model 2, the REML LRS for testing $H_0: \rho = 0$ is

$$\lambda = -2[l_{REML}(\tilde{\sigma}_A^2, \tilde{\sigma}_B^2) - l_{REML}(\hat{\sigma}_A^2, \hat{\sigma}_B^2, \hat{\rho})],$$

where $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ and $\hat{\rho}$ are the unrestricted REML estimators of σ_A^2, σ_B^2 and ρ respectively, $\tilde{\sigma}_A^2$ and $\tilde{\sigma}_B^2$ are the REML estimator of σ_A^2 and σ_B^2 under H_0 (model 1). The asymptotic distribution of λ under H_0 is χ_1^2, so the null hypothesis is rejected at the level α if $\lambda > \chi_{1,\alpha}^2$.

2.3. The EBLUP and its mean squared error

We are interested in predicting the value of $\mu_{dt} = \mathbf{x}_{dt}^{\mathsf{T}}\boldsymbol{\beta} + u_{dt}$ by using the EBLUP $\hat{\mu}_{dt} = \mathbf{x}_{dt}^{\mathsf{T}}\hat{\boldsymbol{\beta}} + \hat{u}_{dt}$. If we do not take into account the error, e_{dt}, this is equivalent to predict $y_{dt} = \mathbf{a}^{\mathsf{T}}\mathbf{y}$, where $\mathbf{a} = \underset{1\le\ell\le D}{\text{col}}(\underset{1\le k\le m_\ell}{\text{col}}(\delta_{d\ell}\delta_{tk}))$ is a vector having one 1 in the position $t + \sum_{\ell=1}^{d-1} m_\ell$ and 0's in the remaining cells. To estimate $\overline{Y}_{dt}$ we use $\widehat{\overline{Y}}_{dt}^{eblup} = \hat{\mu}_{dt}$. The mean squared error of $\widehat{\overline{Y}}_{dt}^{eblup}$ can be approximated by considering the formula established by Prasad and Rao (1980) for moment-based estimators of model parameters in the Fay-Herriot model. This formula was later extended by Datta and Lahiri (2000) and Das, Jiang and Rao (2004) to a wide variety of linear mixed models when the model parameters are estimated by the ML and REML method. By adapting the mean squared error formula to model 3, we get

$$MSE(\widehat{\overline{Y}}_{dt}^{eblup}) = g_{1dt}(\boldsymbol{\theta}) + g_{2dt}(\boldsymbol{\theta}) + g_{3dt}(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\sigma_A^2, \rho_A, \sigma_B^2, \rho_B)$,

$$\begin{aligned}
g_{1dt}(\boldsymbol{\theta}) &= \mathbf{a}^\mathsf{T}\mathbf{Z}\mathbf{T}\mathbf{Z}^\mathsf{T}\mathbf{a}, \\
g_{2dt}(\boldsymbol{\theta}) &= [\mathbf{a}^\mathsf{T}\mathbf{X} - \mathbf{a}^\mathsf{T}\mathbf{Z}\mathbf{T}\mathbf{Z}^\mathsf{T}\mathbf{V}_e^{-1}\mathbf{X}]\mathbf{Q}[\mathbf{X}^\mathsf{T}\mathbf{a} - \mathbf{X}^\mathsf{T}\mathbf{V}_e^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^\mathsf{T}\mathbf{a}], \\
g_{3dt}(\boldsymbol{\theta}) &\approx \operatorname{tr}\left\{\nabla\mathbf{b}^\mathsf{T}\mathbf{V}\nabla\mathbf{b}\,E\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\mathsf{T}\right]\right\}.
\end{aligned}$$

$\mathbf{T} = \mathbf{V}_u - \mathbf{V}_u\mathbf{Z}^\mathsf{T}\mathbf{V}^{-1}\mathbf{Z}\mathbf{V}_u$, $\mathbf{Q} = (\mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X})^{-1}$, $\mathbf{b}^\mathsf{T} = \mathbf{a}^\mathsf{T}\mathbf{Z}\mathbf{V}_u\mathbf{Z}^\mathsf{T}\mathbf{V}^{-1}$, $\nabla\mathbf{b}^\mathsf{T} = (\frac{\partial\mathbf{b}^\mathsf{T}}{\partial\sigma_A^2}, \frac{\partial\mathbf{b}^\mathsf{T}}{\partial\sigma_B^2}, \frac{\partial\mathbf{b}^\mathsf{T}}{\partial\rho_A}, \frac{\partial\mathbf{b}^\mathsf{T}}{\partial\rho_B})$.

The estimator of $MSE(\widehat{\overline{Y}}_{dt}^{eblup})$ is

$$mse_{dt}(\widehat{\overline{Y}}_{dt}^{eblup}) = g_{1dt}(\hat{\boldsymbol{\theta}}) + g_{2dt}(\hat{\boldsymbol{\theta}}) + 2g_{3dt}(\hat{\boldsymbol{\theta}}).$$

3. Estimation of poverty indicators

3.1. The indicators and the data

Let z_{dtj} be an income variable measured in all the units j of the population and let z_t be the poverty line, so that units from domain d with $z_{dtj} < z_t$ are considered as poor at time period t. Let N_t and N_{dt}, $d = 1, \ldots, D$, be the population size at time t and the population size of each domain d at time t respectively. Foster et al. (1984) introduced the family of poverty indicators

$$\overline{Y}_{\alpha,dt} = \frac{1}{N_{dt}}\sum_{j=1}^{N_{dt}} y_{\alpha,dtj}, \quad \text{where } y_{\alpha,dtj} = \left(\frac{z_t - z_{dtj}}{z_t}\right)^\alpha I(z_{dtj} < z_t), \tag{2}$$

$I(z_{dtj} < z_t) = 1$ if $z_{dtj} < z_t$ and $I(z_{dtj} < z_t) = 0$ otherwise. The proportion of units under poverty in the domain d and period t is thus $\overline{Y}_{0,dt}$ and the poverty gap is $\overline{Y}_{1,dt}$.

The Spanish Statistical Office fixes the Poverty Threshold z_t at the 60% of the median of the normalized incomes in Spanish households. The aim of normalizing the household income is to adjust for the varying size and composition of households. The definition of the total number of normalized household members uses a scale giving a weight 1.0 to the first adult, 0.5 to the second and each subsequent person aged 14 and over and 0.3 to each child aged under 14 in the household. The *normalized size* of a household is the sum of the weights assigned to each person. So for each household h in domain d and time t, the total number of normalized members is

$$H_{dth} = 1 + 0.5(H_{dth\geq 14} - 1) + 0.3H_{dth<14},$$

where $H_{dth\geq 14}$ is the number of people aged 14 and over and $H_{dth<14}$ is the number of children aged under 14. The normalized net annual income of a household is obtained by dividing its net annual income by its normalized size. The Spanish poverty thresholds (in euros) in 2004-06 are $z_{2004} = 6098.57$, $z_{2005} = 6160.00$ and $z_{2006} = 6556.60$ respectively. These are the z_t-values used in the calculation of the direct estimates of the poverty incidence and gap.

We use data from the Spanish Living Conditions Survey (SLCS) corresponding to years 2004-2006. The SLCS started in 2004 with an annual periodicity and is the Spanish version of the European Statistics on Income and Living Conditions (EU-SILC), which is one of the statistical operations that have been harmonized for EU countries. We consider $D = 104$ domains obtained by crossing 52 provinces with 2 sexes.

The direct estimator of the total, $Y_{dt} = \sum_{j=1}^{N_{dt}} y_{dtj}$, is

$$\hat{Y}_{dt}^{dir} = \sum_{j\in S_{dt}} w_{dtj} y_{dtj}.$$

where S_{dt} is the domain sample at time period t and the w_{dtj}'s are the official calibrated sampling weights which take into account for non response. The estimated domain size

$$\hat{N}_{dt}^{dir} = \sum_{j\in S_{dt}} w_{dtj}.$$

Using these quantities, a direct estimator of the domain mean, $\bar{Y}_{dt}$, is $\bar{y}_{dt} = \hat{Y}_{dt}^{dir}/\hat{N}_{dt}^{dir}$. The design-based variances of these estimators can be approximated by

$$\hat{V}_\pi(\hat{Y}_{dt}^{dir}) = \sum_{j\in S_{dt}} w_{dtj}(w_{dtj}-1)\left(y_{dtj}-\bar{y}_{dt}\right)^2 \text{ and } \hat{V}_\pi(\bar{y}_{dt}) = \hat{V}\left(\hat{Y}_{dt}^{dir}\right)/\hat{N}_{dt}^2. \qquad (3)$$

The last formulas are obtained from Särndal et al. (1992), pp. 43, 185 and 391, with the simplifications $w_{dtj} = 1/\pi_{dtj}$, $\pi_{dtj,dtj} = \pi_{dtj}$ and $\pi_{dti,dtj} = \pi_{dti}\pi_{dtj}$, $i \neq j$ in the second order inclusion probabilities.

As we are interested in the cases $y_{dtj} = y_{\alpha,dtj}$, $\alpha = 0,1$, we select the direct estimates of the poverty incidence and poverty gap at domain d and time period t (i.e. $\bar{y}_{0,dt}$ and $\bar{y}_{1,dt}$ respectively) as target variables for the time dependent area-level models.

The considered auxiliary variables are the known domain means of the category indicators of the following variables. INTERCEPT: *constant* equal to 1. AGE: Age groups are *age1-age5* for the intervals ≤ 15, $16-24$, $25-49$, $50-64$ and ≥ 65. EDUCATION: Highest level of education completed, with 4 categories denoted by *edu0* for Less than primary education level, *edu1* for Primary education level, *edu2* for Secondary education level and *edu3* for University level. LABOUR: Labour situation with 4 categories taking the values *lab0* for Below 16 years, *lab1* for Employed, *lab2* for Unemployed and *lab3* for Inactive.

3.2. The application

In this section we present an application to real data of model 3 defined in (1). We compare the obtained results with the corresponding ones under the same model restricted to $H_0 : \rho_A = \rho_B$ (model 2), $H_0 : \rho_A = \rho_B = 0$ (model 1) and $H_0 : \rho_A = \rho_B = 0, \sigma_A^2 = \sigma_B^2$ (model 0). Finally the main goal is to estimate the poverty incidence (proportion of individuals under poverty) and the poverty gap in Spanish domains for the three models.

The final selected models include only the auxiliary variables appearing in Table 1. We have included three statistically significant variables that have a relevant meaning in the socio-economic sense. We have selected the variables *age4* (age group 50-65), *edu2* (secondary education completed) and *lab2* (unemployed). Regression parameters and their corresponding p-values are also presented in Table 1 for $\alpha = 0$ and $\alpha = 1$.

By observing the signs of the regression parameters for $\alpha = 0$ (poverty proportion), we interpret that there is an inverse relation between poverty proportion and the categories *age4* and *edu2* of explanatory variables. That is, poverty incidence tends to be smaller in those domains with larger proportion of population in the subset defined by age between 50 and 64, and by secondary education level completed. On the other hand, poverty incidence tends to be larger in those domains with larger proportion of population in the subset defined by *lab2*, i.e. in the category of unemployed people. All the p-values are lower than 0.05 for all the considered auxiliary variables, except for *lab2* in model 3. By doing the same exercise with the signs of the regression parameters in the case $\alpha = 1$ (poverty gap), we can give the same interpretations as before. Again all the p-values are lower than 0.05.

The asymptotic confidence intervals (CIs) for the β's at the 90% confidence level are presented in Table 2 (top) for $\alpha = 0$ and in Table 2 (bottom) for $\alpha = 1$. The columns with labels INF and SUP contains the low and upper limits respectively. By

Table 1: *β-parameters and p-values for $\alpha = 0$ (left) and $\alpha = 1$ (right).*

	$\alpha = 0$				$\alpha = 1$			
model 3	*constant*	*age4*	*edu2*	*lab2*	*constant*	*age4*	*edu2*	*lab2*
β	0.622	−1.881	−0.272	0.260	0.215	−0.741	−0.100	0.320
p-value	0.000	0.000	0.000	0.284	0.000	0.000	0.002	0.004
model 2	*constant*	*age4*	*edu2*	*lab2*	*constant*	*age4*	*edu2*	*lab2*
β	0.778	−2.603	−0.425	0.772	0.237	−0.874	−0.115	0.413
p-value	0.000	0.000	0.000	0.026	0.000	0.000	0.002	0.002
model 1	*constant*	*age4*	*edu2*	*lab2*	*constant*	*age4*	*edu2*	*lab2*
β	0.713	−2.284	−0.445	1.264	0.232	−0.827	−0.123	0.472
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
model 0	*constant*	*age4*	*edu2*	*lab2*	*constant*	*age4*	*edu2*	*lab2*
β	0.730	−2.632	−0.411	1.829	0.198	−0.719	−0.107	0.667
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000

Table 2: *90% confidence intervals for* $\alpha = 0$ *(top) and for* $\alpha = 1$ *(bottom).*

	$model\,3$		$model\,2$		$model\,1$		$model\,0$	
ITEMS	INF	SUP	INF	SUP	INF	SUP	INF	SUP
constant	0.527	0.717	0.646	0.911	0.632	0.794	0.618	0.842
age4	−2.344	−1.418	−3.224	−1.982	−2.657	−1.912	−3.219	−2.045
edu2	−0.385	−0.159	−0.589	−0.262	−0.547	−0.342	−0.562	−0.260
lab2	−0.140	0.661	0.200	1.344	0.879	1.649	1.309	2.349
constant	0.173	0.257	0.188	0.286	0.199	0.264	0.154	0.242
age4	−0.941	−0.542	−1.102	−0.646	−0.978	−0.676	−0.952	−0.486
edu2	−0.152	−0.048	−0.177	−0.054	−0.166	−0.081	−0.169	−0.046
lab2	0.136	0.505	0.198	0.628	0.316	0.629	0.459	0.874

observing these confidence intervals, we conclude that all the regression parameters are significantly different from zero in both cases. The only exception is *lab2* in model 3 for $\alpha = 0$.

Table 3 presents the CIs for the variance components at the 90% confidence level, under models 3-0, for $\alpha = 0$ and $\alpha = 1$. The columns with labels INF and SUP contains the low and upper limits respectively. The column with label 0 ∈CI contains T (true) if 0 belongs to the CI and F (false) otherwise. Concerning model 3, we observe that the CIs for $\rho_A - \rho_B$ and $\sigma_A^2 - \sigma_B^2$ contain the 0. In the case of $\alpha = 0$, the observed value of the likelihood ratio statistics for testing $H_0 : \rho_A = \rho_B$ is $\lambda = 0.5738$ and the corresponding p-value is 0.4487. In the case of $\alpha = 1$, the observed value of the likelihood ratio statistics for testing $H_0 : \rho_A = \rho_B$ is $\lambda = 3.8195$ and the corresponding p-value is 0.0506. These facts suggest that model 3 is not the model fitting best to data.

Table 3: *90% confidence intervals for variances.*

		$\alpha = 0$			$\alpha = 1$		
Model	Parameter	INF	SUP	0 ∈CI	INF	SUP	0 ∈CI
3	σ_A^2	0.0002	0.0008	F	0.0003	0.0005	F
	σ_B^2	0.0005	0.0014	F	0.0002	0.0004	F
	$\sigma_A^2 - \sigma_B^2$	−0.0010	0.0001	T	−0.0000	0.0003	T
	ρ_A	0.8662	0.9957	F	0.5416	0.7484	F
	ρ_B	0.8598	0.9344	F	0.6017	0.8843	F
	$\rho_A - \rho_B$	−0.0409	0.1087	T	−0.2734	0.0774	T
2	σ_A^2	0.0101	0.0154	F	0.0014	0.0019	F
	σ_B^2	0.0023	0.0038	F	0.0004	0.0005	F
	$\sigma_A^2 - \sigma_B^2$	0.0070	0.0124	F	0.0009	0.0015	F
	ρ	0.4050	0.6108	F	0.3528	0.5756	F
1	σ_A^2	0.0025	0.0040	F	0.0004	0.0004	F
	σ_B^2	0.0028	0.0045	F	0.0006	0.0007	F
0	σ_u^2	0.0025	0.0040	F	0.0004	0.0006	F

Table 4: *Normalized Euclidean distances for $\alpha = 0,1$.*

	Model 3		Model 2		Model 1		Model 0	
α	Men	Women	Men	Women	Men	Women	Men	Women
0	0.0194	0.0255	0.0083	0.0421	0.0285	0.0486	0.0648	0.0673
1	0.0115	0.0116	0.0121	0.0221	0.0188	0.0229	0.0290	0.0303

For models 2-0 Table 3 shows that the CIs for σ_A^2, σ_B^2 and σ_u^2 do not contain the origin 0 in any case, so the variances are significatively positive. Table 3 also presents the CIs for the difference of variances $\sigma_A^2 - \sigma_B^2$ and the CIs for ρ under model 2. The variances σ_A^2 and σ_B^2 can be considered as different at the 90% confidence level and the correlation parameter ρ is significantly greater than zero in both cases ($\alpha = 0$ and $\alpha = 1$). In the case $\alpha = 0$ the REML likelihood ratio statistic (LRS) for testing $H_0 : \sigma_A^2 = \sigma_B^2$ takes the value 1210.06 and its corresponding p-value is 0.00. In the case of $\alpha = 1$ the value of the REML LRS for testing $H_0 : \sigma_A^2 = \sigma_B^2$ is 1599.96 and the corresponding p-value is 0.00. In both cases we reject the null hypothesis of equality of variances. Therefore we can recommend model 2 for both poverty indicators.

Table 4 presents the normalized Euclidean distances between the direct and the EBLUPs estimates in both cases $\alpha = 0$ and $\alpha = 1$. We use the formula

$$D(\mathbf{y}_1, \mathbf{y}_2) = \left(\frac{1}{M} \sum_{d=1}^{D} \sum_{t=1}^{m_d} (y_{1dt} - y_{2dt})^2 \right)^{1/2}.$$

The obtained results are somehow expected. The models with more parameters present the lower normalized Euclidean distances. The extreme case would be a saturated model with as many parameters as observations, which has a perfect fit to data. As our target is explaining the data relationships, instead of looking for the best way of predicting the observed y-values, we do not modify our decision about model 3.

For being more confident about our decision of selecting model 2 as true generating model, we still give some diagnostics for models 0-2. At this stage, we drop out Model 3 from the selection procedure because of the hypotheses tested in Table 3.

Residuals $\hat{e}_{dt} = \bar{y}_{dt} - \bar{\mathbf{x}}_{dt}^{\intercal} \hat{\boldsymbol{\beta}} - \hat{u}_{dt}$ of fitted models 2, 1 and 0 are plotted against the observed values $\bar{y}_{dt}$ in the Figure 1 for $\alpha = 0$ (left) and $\alpha = 1$ (right). The dispersion graph shows a great difference in the pattern of the plots, passing from the basic model 0 to the more complex model 2. In particular, residuals of model 2 present a more flattened shape than the ones of the other two models. Figure 2 presents the boxplots of residuals of models 0-2 and also shows that partitioned models 1 and 2 fit much better to the data than model 0. This conclusion coincides with the results appearing in Table 4, where Euclidean distances decrease as moving from model 0 to model 2. So we conclude that model 2 fits better to the direct estimates and therefore we can recommend it.

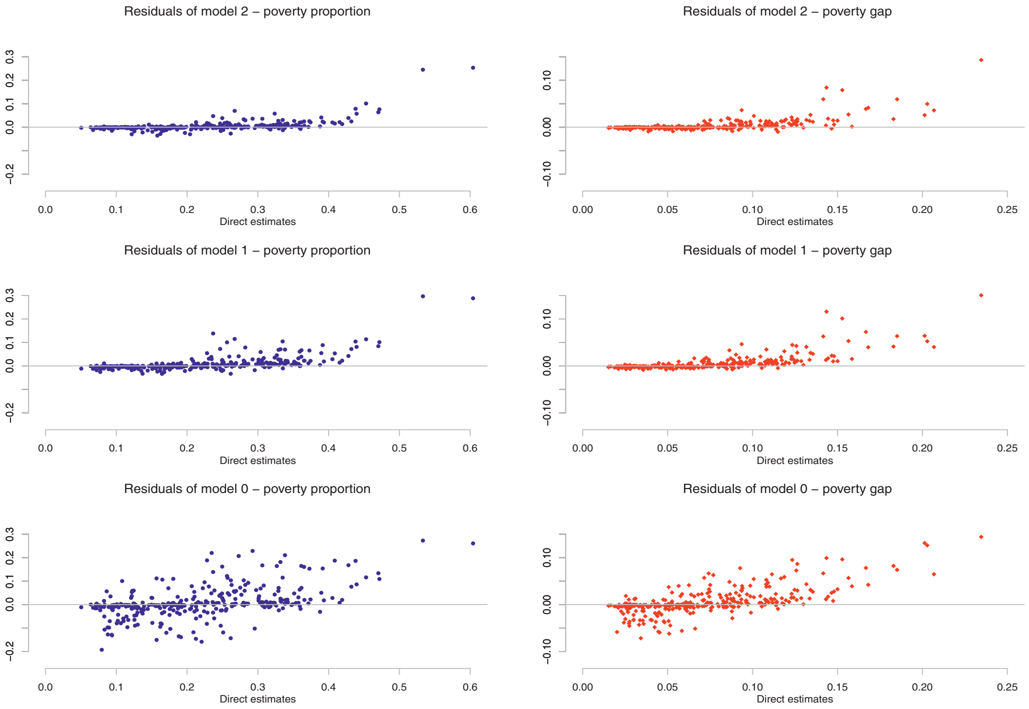

Figure 1: *Residuals versus direct estimates.*

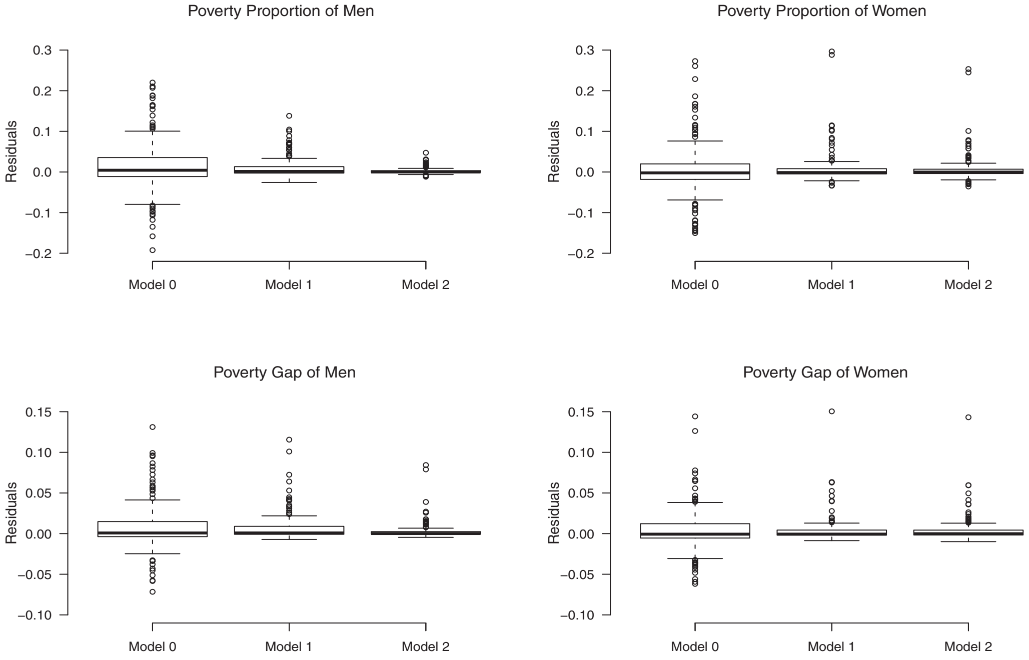

Figure 2: *Boxplots of residuals of models 0-2.*

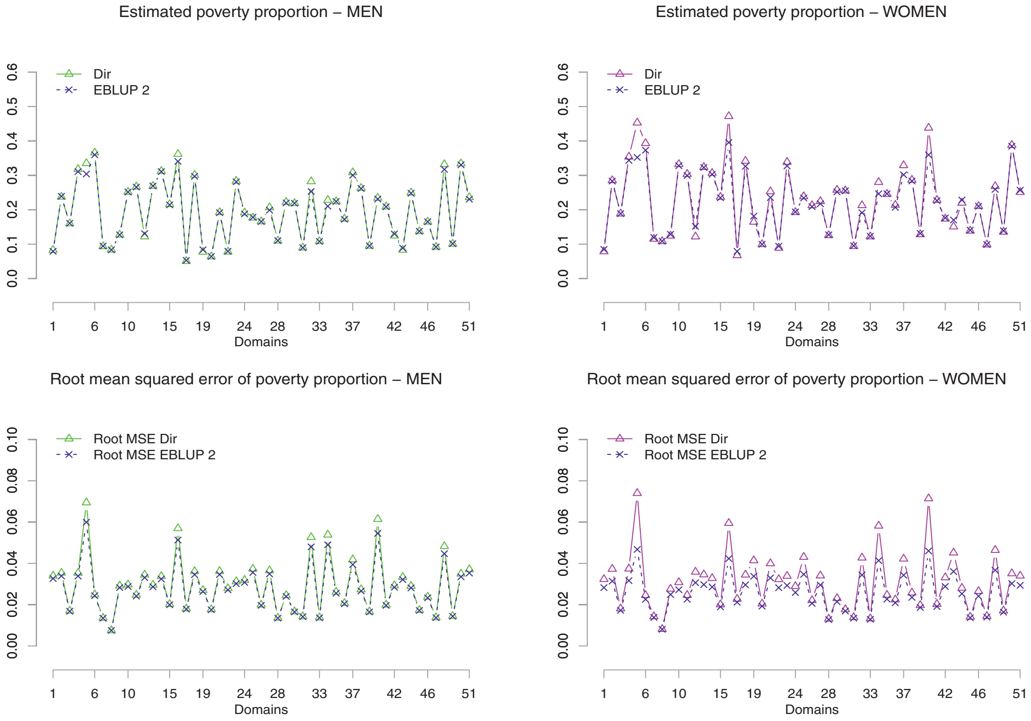

Figure 3: *Estimates of poverty proportion (top) and squared root of their estimated MSEs (bottom) respectively for men (on the left) and women (on the right) in 2006.*

The poverty proportion estimates, direct and EBLUP under model 2, are plotted in the Figure 3 with respect to the partition of the domains in men (left) and women (right). Figure 4 presents the same plots for the poverty gap. Concerning the root MSEs, these figures show that the EBLUPs under model 2 have lower MSE than the direct estimator. Therefore it is worthwhile using model-based estimators instead of the direct ones. As the estimated root MSE of the direct estimate of domain 42 is too large, Figure 3 does not plot the estimates of this domain and renumbers domains 43 to 52 as 42 to 51.

In the Figure 5 the Spanish provinces are plotted in 4 colored categories depending on the values of the EBLUP2 estimates in % of the poverty proportions and the gaps, i.e. $p_d = 100\cdot\hat{\overline{Y}}_{0;d,2006}^{eblup2}$ and $g_d = 100\cdot\hat{\overline{Y}}_{1;d,2006}^{eblup2}$. We observe that the Spanish regions where the proportion of the population under the poverty line is smallest are those situated in the north and east, like Cataluña, Aragón, Navarra, País Vasco, Cantabria and Baleares. On the other hand the Spanish regions with higher poverty proportion are those situated in the centre-south, like Andalucía, Extremadura, Murcia, Castilla La Mancha, Canarias, Ceuta and Melilla. In an intermediate position we can find regions that are in the centre-north of Spain, like Galicia, La Rioja, Castilla León, Asturias, Comunidad Valenciana and Madrid. If we investigate how far the annual net incomes of population under the

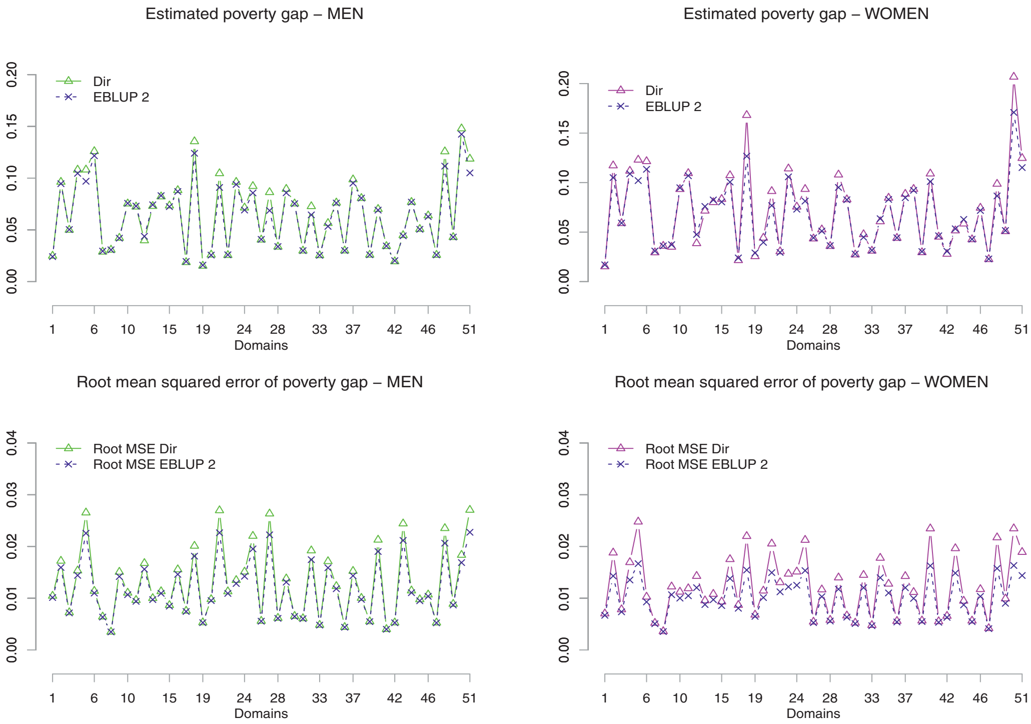

Figure 4: *Estimates of poverty gap (top) and squared root of their estimated MSEs (bottom) respectively for men (on the left) and women (on the right) in 2006.*

Table 5: *Estimated poverty proportions ($\alpha = 0$) and RMSE's in 2006.*

	Men					Women				
Province	n_d	DIR	EB$_2$	RMSE$_\star$	RMSE$_2$	n_d	DIR	EB$_2$	RMSE$_\star$	RMSE$_2$
Soria	24	0.247	0.231	0.107	0.080	18	0.604	0.351	0.126	0.057
Segovia	60	0.234	0.231	0.061	0.055	60	0.438	0.360	0.071	0.046
Palencia	73	0.228	0.210	0.054	0.049	72	0.280	0.246	0.058	0.041
Álava	98	0.083	0.079	0.034	0.033	100	0.079	0.085	0.032	0.028
Zamora	109	0.332	0.317	0.048	0.045	100	0.268	0.259	0.046	0.037
Huelva	124	0.192	0.191	0.036	0.035	124	0.253	0.235	0.040	0.033
Burgos	169	0.127	0.127	0.029	0.028	167	0.124	0.129	0.028	0.025
Albacete	173	0.237	0.239	0.035	0.034	193	0.285	0.283	0.037	0.031
Granada	189	0.301	0.297	0.036	0.035	229	0.342	0.326	0.034	0.030
Crdoba	221	0.312	0.311	0.034	0.032	233	0.307	0.303	0.033	0.029
Cáceres	261	0.252	0.252	0.030	0.029	303	0.332	0.328	0.031	0.027
Tenerife	373	0.263	0.262	0.027	0.027	397	0.286	0.283	0.026	0.024
Sevilla	473	0.209	0.209	0.020	0.020	492	0.228	0.227	0.020	0.019
Zaragoza	556	0.101	0.101	0.014	0.014	577	0.136	0.139	0.017	0.016
Barcelona	1367	0.083	0.084	0.008	0.008	1494	0.108	0.109	0.008	0.008

Table 6: *Estimated poverty gapss ($\alpha = 1$) and RMSE's for men in 2006.*

	Men					Women				
Province	n_d	DIR	EB$_2$	RMSE$_\star$	RMSE$_2$	n_d	DIR	EB$_2$	RMSE$_\star$	RMSE$_2$
Soria	24	0.153	0.074	0.088	0.038	18	0.235	0.091	0.111	0.023
Segovia	60	0.070	0.069	0.021	0.019	61	0.123	0.102	0.025	0.017
Palencia	73	0.056	0.053	0.017	0.016	78	0.052	0.053	0.020	0.015
lava	98	0.025	0.024	0.010	0.010	87	0.107	0.101	0.018	0.014
Zamora	109	0.126	0.112	0.024	0.021	100	0.099	0.087	0.022	0.016
Huelva	124	0.105	0.091	0.027	0.023	124	0.091	0.077	0.021	0.015
Burgos	169	0.042	0.042	0.015	0.014	165	0.089	0.085	0.014	0.012
Albacete	173	0.096	0.095	0.017	0.016	181	0.053	0.051	0.012	0.010
Granada	189	0.135	0.124	0.020	0.018	194	0.112	0.109	0.017	0.014
Crdoba	221	0.082	0.083	0.011	0.011	230	0.114	0.106	0.015	0.012
Cceres	261	0.075	0.076	0.011	0.011	247	0.207	0.171	0.023	0.016
Tenerife	373	0.081	0.081	0.010	0.010	397	0.093	0.092	0.011	0.010
Sevilla	473	0.034	0.034	0.004	0.004	501	0.043	0.044	0.005	0.005
Zaragoza	556	0.043	0.043	0.009	0.009	605	0.027	0.028	0.005	0.005
Barcelona	1367	0.031	0.031	0.003	0.003	1494	0.036	0.036	0.004	0.004

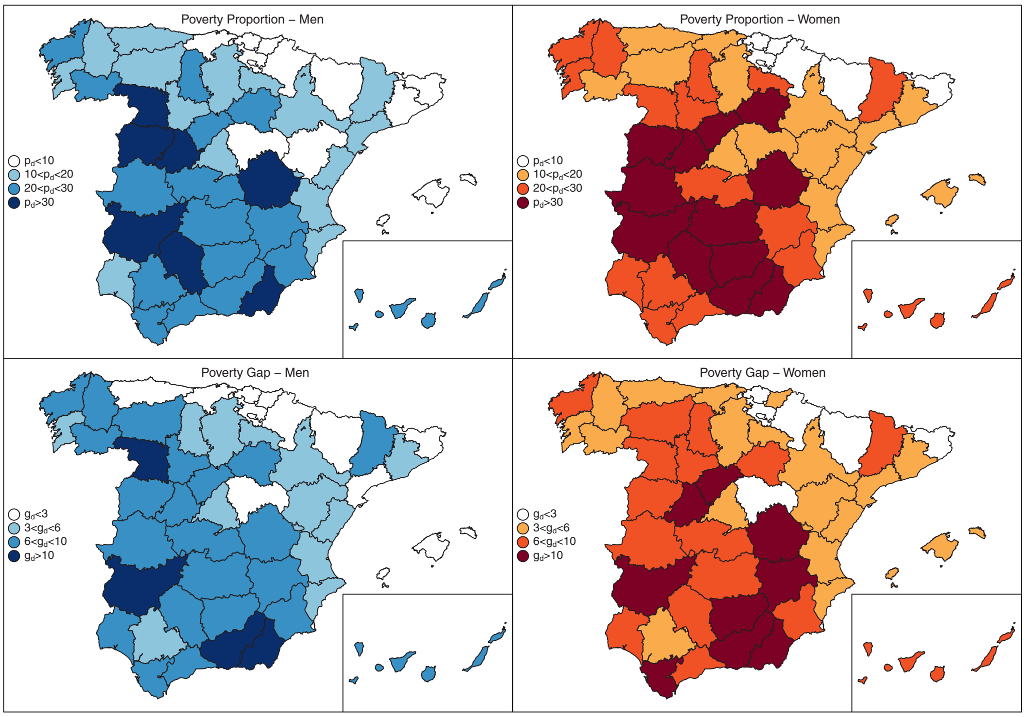

Figure 5: *EBLUP2 estimates of poverty proportions (top) and gaps (bottom) for men (left) and women (right) in 2006.*

poverty line z_{2006} are from z_{2006}, we observe that in the Spanish regions situated in the centre-north there exist a distance that is generally lower than the 6% of z_{2006}. However, the cited distance is in general greater than 6% of z_{2006} in the centre-south.

Tables 5-6 present the direct and EBLUP estimates under model 2 of poverty proportions ($\alpha = 0$) and poverty gaps ($\alpha = 1$) for some Spanish provinces. The provinces were selected accordingly with the quantiles of the set of domain sample sizes n_d. The EBLUP estimates under the model 2 are labelled by EB_2 and the direct estimates by DIR. The squared root of MSEs are labelled by $\text{RMSE}_\star$ for the direct estimator and by RMSE_2 for the EBLUP under the model 2 respectively. Numerical results are sorted by sex. Regarding the reduction of the MSE when passing from direct to EBLUP estimates, we observe that model 2 performs better in domains with small sample size.

4. Discussion

As poverty indicators are nonlinear, unit-level model-based estimation approaches cannot always be used. However, their direct estimators are weighted sums that can be modelled by area-level models. Area-level models thus provide an easy-to-apply solution. These idea motivates the introduction of partitioned temporal models that borrow strength from time. The use of information from past time instants, the greater availability of auxiliary variables at the domain level and the possibility of introducing modelling differences by sex might compensate the loss of information when passing from unit-level models to area-level models. We thus considered four area-level linear mixed models and we applied the methodology to Spanish EU-SILC data.

We would also like to point out that model (1) and its particularizations have some features of interest, from a methodological point of view. It is somewhat different from the Rao-Yu model (Rao and Yu, 1994, and Rao, 2003), viewed as an extension of the Fay-Herriot area-level model in the case of time-correlated data. As we can note, the covariance matrix of the model does not contain the variance component connected with the random-effect at the domains, as clusters of time-correlated data. This fact permits to the random time-area effect to absorb completely the variation of the EBLUP due to the correlated observation, without considering any cluster-oriented random-effect components.

Another characteristic of main interest of the model (1), is that is a “partitioned” model. This means that different variance components in the covariance matrix of the random-area effects can accommodate different inputs of information, due to some relevant issues related to the specific levels of auxiliary variables. In the case of the application on the poverty indicators in Spain, the partitioning of the variance of the random-effect is significative for the gender-based class of survey domains. In fact, relevant differences in terms of the data in these classes of domains, as inputs in the fixed-effects regression, seems to drive at the same time to different variations in the related class of random-area effects.

The R programming language has been employed for doing all the computations in this paper. The deliverable D22 on software for small area estimation of the European SAMPLE project (http://www.sample-project.eu/) gives a primary version of the employed R codes.

References

Das, K., Jiang, J. and Rao, J. N. K. (2004). Mean squared error of empirical predictor. Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818–840.

Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613–627.

Esteban, M. D., Morales, D., Pérez, A. and Santamaría, L. (2012a). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, 56, 2840–2855.

Esteban, M. D., Morales, D., Pérez, A. and Santamaría, L. (2012b). Two area-level time models for estimating small area poverty indicators. *Journal of the Indian Society of Agricultural Statistics*, 66, 75–89.

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

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55–93.

Herrador, M., Esteban, M. D., Hobza, T. and Morales, D. (2011). A Fay-Herriot model with different random effect variances. *Communications in Statistics (Theory and Methods)*, 10, 785–797.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15, 1–96.

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308–325.

Pfeffermann, D. (2002). Small area estimation-new developments and directions. *International Statistical Review*, 70, 125–143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28, 1–134.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163–171.

Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175–186.

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

Rao, J. N. K. and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511–528.

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