Area-Level Time Models for Small Area Estimation of Poverty Indicators

M.D. Esteban, D. Morales, A. Pérez, and L. Santamaría

Abstract. Small area parameters usually take the form $h(y)$, where y is the vector containing the values of all units in the domain and h is a linear or nonlinear function. If h is not linear or the target variable is not normally distributed, then the unit-level approach has no standard procedure and each case should be treated with a specific methodology. Area-level linear mixed models can be generally applied to produce new estimates of linear and non linear parameters because direct estimates are weighted sums, so that the assumption of normality may be acceptable. In this communication we treat the problem of estimating small area non linear parameters, with special emphasis on the estimation of poverty indicators. For this sake, we borrow strength from time by using area-level linear time models. We consider two time-dependent area-level models, empirically investigate their behavior and apply them to estimate poverty indicators in the Spanish Living Conditions Survey.

1 Area-Level Linear Time Model

In small area estimation samples are drawn from a finite population, but estimations are required for subsets (called small areas or domains) where the effective sample sizes are too small to produce reliable direct estimates. An estimator of a small area parameter is called direct if it is calculated just with the sample data coming from the corresponding small area. Thus, the lack of sample data from the target small area affects seriously the accuracy of the direct estimators, and this fact has given rise to the development of new tools for obtaining more precise estimates. See a description of this theory in the monograph of Rao ([4]).

Area-level models relate direct estimates of small area means to area-level auxiliary variables. The idea is to borrow strength from other domains, related

M.D. Esteban, D. Morales, A. Pérez, and L. Santamaría
Universidad Miguel Hernández de Elche, Spain
e-mail: d.morales@umh.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 233–237.
springerlink.com

variables, past time instants and correlations, in order to produce new model-based estimates. In this work we consider the model

$$y_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{dt} + e_{dt}, \quad d = 1,\ldots,D, \quad t = 1,\ldots,m_d, \tag{1}$$

where y_{dt} is a direct estimator of the indicator of interest for area d and time instant t, $\mathbf{x}_{dt}$ is a vector containing the aggregated (population) values of p auxiliary variables, the random vectors $(u_{d1},\ldots,u_{dm_d})$, $d = 1,\ldots,D$, are i.i.d. AR(1), with variance and auto-correlation parameters σ_u^2 and ρ respectively, the errors e_{dt}'s are independent $N(0,\sigma_{dt}^2)$ with known σ_{dt}^2's, and the u_{dt}'s and the e_{dt}'s are independent. In the applications to real data we may also consider a simpler model obtained by restricting model (1) to $\rho = 0$. Model (1) is related to the model of Rao and Yu [3] in the sense that u_d is substituted by u_{dt} to take into account the area-by-time variability through specific random effects.

In matrix notation, model (1) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{2}$$

where

$$\mathbf{y} = \underset{1\le d\le D}{\text{col}}(\mathbf{y}_d), \quad \mathbf{y}_d = \underset{1\le t\le m_d}{\text{col}}(y_{dt}), \quad \mathbf{u} = \underset{1\le d\le D}{\text{col}}(\mathbf{u}_d), \quad \mathbf{u}_d = \underset{1\le t\le m_d}{\text{col}}(u_{dt}),$$

$$\mathbf{e} = \underset{1\le d\le D}{\text{col}}(\mathbf{e}_d), \quad \mathbf{e}_d = \underset{1\le t\le m_d}{\text{col}}(e_{dt}), \quad \mathbf{X} = \underset{1\le d\le D}{\text{col}}(\mathbf{X}_d), \quad \mathbf{X}_d = \underset{1\le t\le m_d}{\text{col}}(\mathbf{x}_{dt}),$$

$$\mathbf{x}_{dt} = \underset{1\le k\le p}{\text{col}'}(x_{dtk}), \quad \boldsymbol{\beta} = \underset{1\le k\le p}{\text{col}}(\beta_k), \quad \mathbf{Z} = \mathbf{I}_{M\times M}, \quad M = \sum_{d=1}^{D} m_d.$$

We assume that $\mathbf{u} \sim N(\mathbf{0},\mathbf{V}_u)$ and $\mathbf{e} \sim N(\mathbf{0},\mathbf{V}_e)$ are independent with covariance matrices

$$\mathbf{V}_u = \sigma_u^2\Omega(\rho), \quad \Omega(\rho) = \underset{1\le d\le D}{\text{diag}}(\Omega_d(\rho)), \quad \mathbf{V}_e = \underset{1\le d\le D}{\text{diag}}(\mathbf{V}_{ed}), \quad \mathbf{V}_{ed} = \underset{1\le t\le m_d}{\text{diag}}(\sigma_{dt}^2),$$

where the variances σ_{dt}^2 are known and

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

The BLU estimators and predictors of $\boldsymbol{\beta}$ and $\mathbf{u}$ are

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad \text{and} \qquad \widehat{\mathbf{u}} = \mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}),$$

where $\text{var}(\mathbf{y}) = \mathbf{V} = \sigma_u^2 \underset{1\le d\le D}{\text{diag}}(\Omega_d(\rho)) + \mathbf{V}_e = \underset{1\le d\le D}{\text{diag}}(\sigma_u^2\Omega_d(\rho) + \mathbf{V}_{ed}) = \underset{1\le d\le D}{\text{diag}}(\mathbf{V}_d)$. The model is fitted by using the residual maximum likelihood method and $\mu_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{dt}$ is predicted with the empirical bet linear unbiased predictor (EBLUP) $\widehat{\mu}_{dt} = \mathbf{x}_{dt}\widehat{\boldsymbol{\beta}} + \widehat{u}_{dt}$. If we do not take into account the error, e_{dt}, this is equivalent to predict $y_{dt} = \mathbf{a}'\mathbf{y}$, where $\mathbf{a} = \underset{1\le\ell\le D}{\text{col}}(\delta_{d\ell}\mathbf{a}_\ell)$ and $\mathbf{a}_\ell = \underset{1\le k\le m_\ell}{\text{col}}(\delta_{tk})$. The population mean $\overline{Y}_{dt}$ is estimated by means of $\widehat{\overline{Y}}_{dt}^{eblup} = \widehat{\mu}_{dt}$. Following Prasad and Rao [2], see also Rao [4] or Jiang and Lahiri [1], the mean squared error (MSE) of $\widehat{\overline{Y}}_{dt}^{eblup}$ takes the form

$$MSE(\widehat{\overline{Y}}_{dt}^{eblup}) = g_1(\theta) + g_2(\theta) + g_3(\theta),$$

where $\theta = (\sigma_u^2, \rho)$,

$$\begin{aligned} g_1(\theta) &= \mathbf{a}'\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{a}, \\ g_2(\theta) &= [\mathbf{a}'\mathbf{X} - \mathbf{a}'\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{V}_e^{-1}\mathbf{X}]\mathbf{Q}[\mathbf{X}'\mathbf{a} - \mathbf{X}'\mathbf{V}_e^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{a}], \\ g_3(\theta) &\approx \text{tr}\left\{(\nabla\mathbf{b}')\mathbf{V}(\nabla\mathbf{b}')'E\left[(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)'\right]\right\} \end{aligned}$$

and $\mathbf{Q} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, $\mathbf{T} = \mathbf{V}_u - \mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{V}_u$, $\mathbf{b}' = \mathbf{a}'\mathbf{Z}\mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}$. The estimator of $MSE(\widehat{\overline{Y}}_{dt}^{eblup})$ is

$$mse(\widehat{\overline{Y}}_{dt}^{eblup}) = g_1(\hat{\theta}) + g_2(\hat{\theta}) + 2g_3(\hat{\theta}). \tag{3}$$

2 Estimation of Poverty Indicators

Let us consider a finite population P_t partitioned into D domains P_{dt} at time period t, and denote their sizes by N_t and N_{dt}, $d = 1,\ldots,D$. Let z_{dtj} be an income variable measured in all the units of the population and let z_t be the poverty line, so that units with $z_{dtj} < z_t$ are considered as poor at time period t. The main goal of this section is to estimate the poverty incidence (proportion of individuals under poverty) and the poverty gap in Spanish domains. These two measures belongs to the family

$$Y_{\alpha;dt} = \frac{1}{N_{dt}}\sum_{j=1}^{N_{dt}} y_{\alpha;dtj}, \quad \text{where } y_{\alpha;dtj} = \left(\frac{z_t - z_{dtj}}{z_t}\right)^\alpha I(z_{dtj} < z_t), \tag{4}$$

$I(z_{dtj} < z_t) = 1$ 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 $Y_{0;dt}$ and the poverty gap is $Y_{1;dt}$.

We use data from the Spanish Living Conditions Survey (SLCS) corresponding to years 2004-2006 with sample sizes 44648, 37491, 34694 respectively. The SLCS 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 SLCS does not produce official estimates at the domain level (provinces $\times$ sex), but the analogous 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. In the particular case $y_{dtj} = 1$, for all $j \in P_{dt}$, we get the estimated domain size

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

Using this quantity, a direct estimator of the domain mean $\bar{Y}_{dt}$ is $\bar{y}_{dt} = \hat{Y}_{dt}^{dir}/\hat{N}_{dt}^{dir}$. The direct estimates of the domain means are used as responses in the area-level time model. The design-based variances of these estimators can be approximated by

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

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: First auxiliary variable is equal to one.
- AGE: Age groups for the intervals ≤ 15, $16-24$, $25-49$, $50-64$ and ≥ 65.
- EDUCATION: Highest level of education completed, with 4 categories for Less than primary education level, Primary education level, Secondary education level and University level.
- CITIZENSHIP: with 2 categories for Spanish and Not Spanish.
- LABOR: Labor situation with 4 categories for Below 16 years, Employed, Unemployed and Inactive.

The Poverty Threshold is fixed as the 60% of the median of the normalized incomes in Spanish households. The total number of normalized household members is

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

where $H_{dtj\geq 14}$ is the number of people aged 14 and over and $H_{dtj<14}$ 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 consider the linear model $\overline{y}_{dt} = \bar{\mathbf{X}}_{dt}\boldsymbol{\beta} + u_{dt} + e_{dt},\ d = 1,\ldots,D,$ where $\overline{y}_{dt} = \hat{Y}_{dt}^{dir}/\hat{N}_{dt}^{dir}$, $\sigma_{dt}^2 = \hat{V}_{\pi}(\bar{y}_{dt})$ and $\bar{\mathbf{X}}_d$ is the $1 \times p$ vector containing the population (aggregated) mean values of all the categories (except the last one) of the explanatory variables. Random effects errors are assumed to follow the distributional assumptions of model (1). Obtained EBLUP estimates of % poverty proportions $p_d = 100 \cdot \hat{Y}_{0;d,2006}^{eblup1}$ and poverty gaps $g_d = 100 \cdot \hat{Y}_{1;d,2006}^{eblup1}$ are presented in the Figure 1.

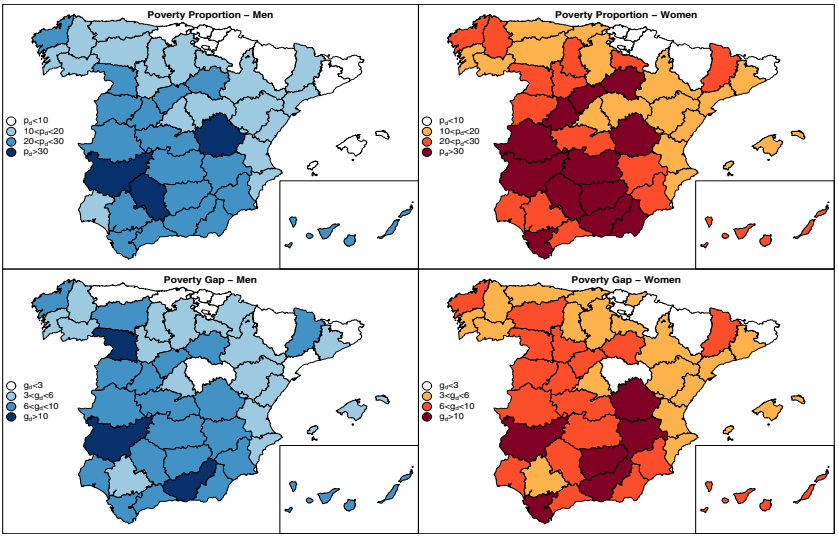

Fig. 1 Estimates of Spanish poverty proportions (top) and gaps (bottom) for men (left) and women (right) in 2006.

References

1. Jiang, J., Lahiri, P.: Mixed model prediction and small area estimation. Test 15, 1–96 (2006)
2. Prasad, N.G.N., Rao, J.N.K.: The estimation of the mean squared error of small-area estimators. J. Amer. Statist. Assoc. 85, 163–171 (1990)
3. Rao, J.N.K., Yu, M.: Small area estimation by combining time series and cross sectional data. Canad. J. Statist. 22, 511–528 (1994)
4. Rao, J.N.K.: Small Area Estimation. Wiley, New York (2003)