Small area estimation of poverty proportions under random regression coefficient models *

Tomáš Hobza[1] and Domingo Morales[2]

[1] Department of Mathematics, Czech Technical University in Prague, Czech Republic
hobza@fjfi.cvut.cz

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

Summary. In this paper a random regression coefficient model is used to provide estimates of small area poverty proportions. As poverty variable is dichotomic at the individual level, the sample data from Spanish Living Conditions Survey is previously aggregated to the level of census sections. EBLUP estimates based on the proposed model are obtained. A closed-formula procedure to estimate the mean squared error of the EBLUP estimators is given and empirically studied. Results of several simulations studies are reported as well as an application to real data.

1 Introduction

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 [8], or in the reviews of Ghosh and Rao (1994), Rao (1999), Pfeffermann [5] and more recently Jiang and Lahiri [3]. Mixed models increase the effective information used in the estimation process by linking all observations of the sample, and at the same time they can allow for between-area variation. Further flexibility is obtained by using random coefficient regression models, which allows the coefficient of auxiliary variables to vary across sampling units or domains. Moura and Holt (1999) suggested the application of random coefficient models in small area estimation. This paper follows their recommendation and presents and application to the estimation of poverty proportions by using data from the Spanish Living Conditions Survey.

*

The paper is organized as follows. Section 2 introduces the considered random coefficient model and Section 3 derives the corresponding EBLUP estimates. Section 4 deals with the problem of estimating mean squared errors. Section 5 presents several simulation experiments designed to investigate some practical issues. Section 6 is devoted to the application to real data. Finally, Section 7 gives some conclusions.

2 A random regression coefficient model

We consider two models. The first one, which will be called *Model B* in the sequel, is the random regression coefficient model

$$y_{dj}=\sum_{k=0}^{p}\beta_k x_{kdj}+\sum_{k=0}^{p}u_{kd}x_{kdj}+e_{dj},\quad d=1,\dots,D,\ j=1,\dots,n_d, \tag{1}$$

where y_{dj} is the jth observation from area d, x_{kdj} are auxiliary variables and β_k are unknown regression parameters. Further, random regression coefficients $u_{kd}\overset{iid}{\sim}N(0,\sigma_k^2)$ and random errors $e_{dj}{\sim}N(0,w_{dj}^{-1}\sigma_e^2)$ are independent, $d=1,\dots,D$, $j=1\dots,n_d$, $k=0,\dots,p$. If $x_{0dj}=1$ for any d and j then model (1) contains a random intercept of the form β_0+u_{0d} for area d. The model variance and covariance parameters are σ_e^2, σ_k^2, $k=0,\dots,p$, $(2+p$ parameters).

In this paper we will compare model (1) with the standard nested regression model (denoted as *Model A*)

$$y_{dj}=\sum_{k=0}^{p}\beta_k x_{kdj}+u_{0d}+e_{dj},\quad d=1,\dots,D,\ j=1,\dots,n_d, \tag{2}$$

where $u_{0d}\overset{iid}{\sim}N(0,\sigma_0^2)$ and $e_{dj}\overset{iid}{\sim}N(0,w_{dj}^{-1}\sigma_e^2)$ are independent, $d=1,\dots,D$, $j=1\dots,n_d$. In this section we briefly describe some basic facts for the application of Model B to small area estimation. The corresponding derivations for Model A are straightforward.

In matrix notation model (1) can be written in the form

$$\boldsymbol{y}=\boldsymbol{X\beta}+\sum_{\boldsymbol{k=0}}^{\boldsymbol{p}}\boldsymbol{Z_k u_k}+\boldsymbol{e}, \tag{3}$$

where $n=\sum_{d=1}^{D}n_d$, $\beta=\beta_{(p+1)\times 1}$, $\boldsymbol{y}=\underset{\boldsymbol{1\le d\le D}}{\mathrm{col}}(\boldsymbol{y_d})$, $\boldsymbol{y_d}=\underset{\boldsymbol{1\le j\le n_d}}{\mathrm{col}}(\boldsymbol{y_{dj}})$, $\boldsymbol{e}=\underset{\boldsymbol{1\le d\le D}}{\mathrm{col}}(\boldsymbol{e_d})$, $\boldsymbol{e_d}=\underset{\boldsymbol{1\le j\le n_d}}{\mathrm{col}}(\boldsymbol{e_{dj}})$, $\boldsymbol{u_k}=\underset{\boldsymbol{1\le d\le D}}{\mathrm{col}}(\boldsymbol{u_{kd}})$, $\boldsymbol{X}=\underset{\boldsymbol{1\le d\le D}}{\mathrm{col}}(\boldsymbol{X_d})$, $\boldsymbol{X_d}=\underset{\boldsymbol{0\le k\le p}}{\mathrm{col}^{\boldsymbol{t}}}(\boldsymbol{x_{k,n_d}})$, $\boldsymbol{x_{k,n_d}}=\underset{\boldsymbol{1\le j\le n_d}}{\mathrm{col}}(\boldsymbol{x_{kdj}})$, $\boldsymbol{Z_k}=\underset{\boldsymbol{1\le d\le D}}{\mathrm{diag}}(\boldsymbol{x_{k,n_d}})$, $\boldsymbol{I_a}=\underset{\boldsymbol{1\le j\le a}}{\mathrm{diag}}(\boldsymbol{1})$, $W=\underset{1\le d\le D}{\mathrm{diag}}(W_d)$, $\boldsymbol{W_d}=\underset{\boldsymbol{1\le j\le n_d}}{\mathrm{diag}}(\boldsymbol{w_{dj}})$, with $w_{dj}>0$ known, $d=1,\dots,D$, $j=1,\dots,n_d$. Note that model (1) is a multilevel model that can alternatively be written as in Moura and Holt [4], i.e.

$$\boldsymbol{y_d}=\boldsymbol{X_d\gamma_d}+\boldsymbol{e_d},\quad \boldsymbol{\gamma_d}=\boldsymbol{\beta}+\boldsymbol{u_{.d}},\quad \boldsymbol{d=1,\dots,D}, \tag{4}$$

where $\boldsymbol{u_{.d}}=\underset{\boldsymbol{0\le k\le p}}{\mathrm{col}}(\boldsymbol{u_{kd}})$.

Variance matrices of Model B are $V_e = \mathrm{var}(e) = \sigma_e^2 W^{-1}$, $V_{u_k} = \mathrm{var}(u_k) = \sigma_k^2 I_D$, $k = 0, 1, \ldots, p$, and

$$\boldsymbol{V} = \mathrm{var}(\boldsymbol{y}) = \boldsymbol{V}_e + \sum_{\boldsymbol{k=0}}^{\boldsymbol{p}} \boldsymbol{Z}_{\boldsymbol{k}} \boldsymbol{V}_{\boldsymbol{u_k}} \boldsymbol{Z}_{\boldsymbol{k}}^{\boldsymbol{t}} = \underset{\boldsymbol{1 \le d \le D}}{\mathrm{diag}} (\boldsymbol{V}_{\boldsymbol{d}}),$$

where

$$V_d = \sigma_e^2 W_d^{-1} + \sum_{k=0}^{p} \sigma_k^2 x_{k,n_d} x_{k,n_d}^t, \quad d = 1, \ldots, D.$$

For model fitting it is worthwhile to consider the alternative parameters $\sigma^2 = \sigma_e^2$, $\varphi_k = \sigma_k^2/\sigma_e^2$, $k = 0, 1, \ldots, p$, in such a way that $V = \sigma^2 \Sigma$ and $V_d = \sigma^2 \Sigma_d$, where $\boldsymbol{\Sigma} = \underset{\boldsymbol{1 \le d \le D}}{\mathrm{diag}} (\boldsymbol{\Sigma}_{\boldsymbol{d}})$ and

$$\boldsymbol{\Sigma}_{\boldsymbol{d}} = \boldsymbol{W}_{\boldsymbol{d}}^{\boldsymbol{-1}} + \sum_{\boldsymbol{k=0}}^{\boldsymbol{p}} \boldsymbol{\varphi}_{\boldsymbol{k}} \boldsymbol{x}_{\boldsymbol{k,n_d}} \boldsymbol{x}_{\boldsymbol{k,n_d}}^{\boldsymbol{t}}, \quad \boldsymbol{d = 1, \ldots, D.} \tag{5}$$

Let $\varphi = (\sigma^2, \varphi_0, \varphi_1, \ldots, \varphi_p)$ be the vector of variance components, with $\sigma^2 > 0$, $\varphi_0 > 0$, $\varphi_1 > 0, \ldots, \varphi_p > 0$. Let $\boldsymbol{u} = \underset{\boldsymbol{0 \le k \le p}}{\mathrm{col}} (\boldsymbol{u}_{\boldsymbol{k}})$ with variance $\boldsymbol{V}_{\boldsymbol{u}} = \mathrm{var}(\boldsymbol{u}) = \underset{\boldsymbol{0 \le k \le p}}{\mathrm{diag}} (\boldsymbol{V}_{\boldsymbol{u_k}})$ and $\boldsymbol{Z} = \underset{\boldsymbol{0 \le k \le p}}{\mathrm{col}^{\boldsymbol{t}}} (\boldsymbol{Z}_{\boldsymbol{k}})$. Using this notation the model (3) can be written in the general form

$$y = X\beta + Zu + e.$$

If φ is known, then the BLUE of $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^t$ is

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y = \left(\sum_{d=1}^{D} X_d^t \Sigma_d^{-1} X_d \right)^{-1} \left(\sum_{d=1}^{D} X_d^t \Sigma_d^{-1} y_d \right)$$

and the BLUP of u is $\hat{\boldsymbol{u}} = \boldsymbol{V}_{\boldsymbol{u}} \boldsymbol{Z}^{\boldsymbol{t}} \boldsymbol{V}^{\boldsymbol{-1}} (\boldsymbol{y} - \boldsymbol{X} \hat{\boldsymbol{\beta}})$, i.e.

$$\hat{u} = \underset{0 \le k \le p}{\mathrm{diag}} (V_{u_k}) \underset{0 \le k \le p}{\mathrm{col}} (Z_k^t) \underset{1 \le d \le D}{\mathrm{diag}} (V_d^{-1}) \underset{1 \le d \le D}{\mathrm{col}} (y_d - X_d \hat{\beta}).$$

The empirical BLUE and BLUP (EBLUE and EBLUP) are obtained by substituting the variance parameters by convenient estimates. We will now describe the Fisher-scoring algorithm to calculate the residual maximum likelihood estimates of the variance components.

The REML log-likelihood is

$$l_{REML}(\sigma) = -\frac{1}{2}(n-p)\log 2\pi - \frac{1}{2}(n-p)\log \sigma^2 - \frac{1}{2}\log |K^t \Sigma K| - \frac{1}{2\sigma^2} y^t P y,$$

where

$$P = K(K^t \Sigma K)^{-1} K^t = \Sigma^{-1} - \Sigma^{-1} X (X^t \Sigma^{-1} X)^{-1} X^t \Sigma^{-1},$$
$$K = W - WX(X^t W X)^{-1} X^t W$$

are such that $PX = 0$ and $P\Sigma P = P$. From (5) it follows that Σ can be written in the form

$$\Sigma = W^{-1} + \sum_{k=0}^{p} \varphi_k A_k,$$

where $\boldsymbol{A_k = Z_k Z_k^t =} \underset{\boldsymbol{1 \le d \le D}}{\text{diag}} \boldsymbol{(x_{k,n_d} x_{k,n_d}^t)}$, $k = 0, 1, \ldots, p$. As $\frac{\partial P}{\partial \varphi_k} = -P A_k P$, by taking partial derivatives with respect to σ^2 and φ_k, $k = 0, 1, \ldots, p$, one gets

$$S_{\sigma^2} = -\frac{n-p}{2\sigma^2} + \frac{1}{2\sigma^4}\, y^t P y, \quad S_{\varphi_k} = -\frac{1}{2}\,\text{tr}\{PA_k\} + \frac{1}{2\sigma^2}\, y^t P A_k P y,$$

$k = 0, 1, \ldots, p$. The second partial derivatives are

$$\begin{aligned} H_{\sigma^2\sigma^2} &= \frac{n-p}{2\sigma^4} - \frac{1}{\sigma^6} y^t P y, \quad H_{\sigma^2\varphi_i} = -\frac{1}{2\sigma^4} y^t P A_i P y, \\ H_{\varphi_i\varphi_j} &= \frac{1}{2}\,\text{tr}\{PA_i P A_j\} - \frac{1}{\sigma^2} y^t P A_i P A_j P y, \quad i, j = 0, 1, \ldots, p. \end{aligned}$$

By taking expectations and multiplying by -1, we obtain the components of the Fisher information matrix $(i, j = 0, 1, \ldots, p)$

$$F_{\sigma^2\sigma^2} = \frac{n-p}{2\sigma^4}, \quad F_{\sigma^2\varphi_j} = \frac{1}{2\sigma^2}\,\text{tr}\{PA_j\}, \quad F_{\theta_i\varphi_j} = \frac{1}{2}\,\text{tr}\{PA_i P A_j\}.$$

To calculate the REML estimates, the Fisher-scoring updating formula is

$$\varphi^{k+1} = \varphi^k + F^{-1}(\varphi^k) S(\varphi^k).$$

The following seeds can be used as starting values in the Fisher-scoring algorithm

$$\sigma^{2(0)} = \theta_0^{(0)} = \varphi_1^{(0)} = \ldots = \varphi_p^{(0)} = S^2/(p+2),$$

where $\boldsymbol{S^2 = \frac{1}{n-p}(y - X\tilde{\beta})^t W (y - X\tilde{\beta})}$ and $\tilde{\beta} = (X^t W X)^{-1} X^t W y$.

The asymptotic distributions of the REML estimators of φ and β are

$$\boldsymbol{\hat{\varphi} \sim N_{p+2}(\varphi, F^{-1}(\varphi)), \quad \hat{\beta} \sim N_{p+1}(\beta, (X'V^{-1}X)^{-1}),}$$

so that the $1 - \alpha$ asymptotic confidence intervals for φ_k and β_k are

$$\hat{\varphi}_k \pm z_{\alpha/2}\, \nu_{kk}^{1/2}, \quad \text{and} \quad \hat{\beta}_k \pm z_{\alpha/2}\, q_{kk}^{1/2}, \; k = 0, 1, \ldots, p,$$

where $\hat{\varphi} = \varphi^\kappa$, κ is the last iteration in the Fisher-scoring algorithm, $F^{-1}(\varphi^\kappa) = (\nu_{k\ell})_{k,\ell=-1,0,\ldots,p}$, $(X'V^{-1}(\varphi^\kappa)X)^{-1} = (q_{k\ell})_{k,\ell=0,1,\ldots,p}$ and z_α is the α-quantile of the $N(0,1))$ distribution. The confidence interval for σ^2 is obtained in the same way by using the corresponding diagonal element of the matrix F^{-1}.

3 EBLUP of the domain mean

In this section we consider a finite population of N elements following the model introduced in (1) with population sizes N_d in the place of sample sizes n_d. From the population a sample of size n with n_d elements in area d, $n = \sum_{d=1}^{D} n_d$, is selected. Without loss of generality we can reorder the population so that $y = (y_s^t, y_r^t)^t$, where y_s is the vector of n observed elements and y_r is the vector of $N - n$ unobserved elements. In the following, the index s for the sample and the index r for the rest of the population will be used when appropriate. In this notation and taking into account the reordering we can write

$$\boldsymbol{V} = \mathrm{var}[\boldsymbol{y}] = \begin{pmatrix} \boldsymbol{V_{ss}} & \boldsymbol{V_{sr}} \\ \boldsymbol{V_{rs}} & \boldsymbol{V_{rr}} \end{pmatrix} .$$

We are interested in the estimation of the mean of the small area d, i.e.

$$\overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj} = a^t y = a_s^t y_s + a_r^t y_r,$$

where $a^t = \frac{1}{N_d} \left(\mathbf{0}_{N_1}^t, \dots, \mathbf{0}_{N_{d-1}}^t, \mathbf{1}_{N_d}^t, \mathbf{0}_{N_d+1}^t, \dots, \mathbf{0}_{N_D}^t \right)$ and $\mathbf{0}_m^t = (0, \dots, 0)_{1\times m}$. From the general theorem of prediction it follows that the BLU predictor of $\overline{Y}_d$, under Model B, is

$$\boldsymbol{\widehat{\overline{Y}}_d^{blupB}} = \boldsymbol{a_s^t y_s + a_r^t \left[X_r \widehat{\beta} + \widehat{V}_{rs} \widehat{V}_{ss}^{-1} (y_s - X_s \widehat{\beta}) \right] .} \tag{6}$$

In our case it holds $V_{e,rs} = 0$, $V_{rs} = Z_r V_u Z_s^t + V_{e,rs} = Z_r V_u Z_s^t$ and $\boldsymbol{\hat{u}} = \boldsymbol{\hat{V}_u Z_s^t \hat{V}_{ss}^{-1} (y_s - X_s \hat{\beta})}$, so

$$\begin{aligned} \widehat{\overline{Y}}_d^{blupB} &= a_s^t y_s + a_r^t \left[X_r \hat{\beta} + Z_r \hat{V}_u Z_s^t \hat{V}_{ss}^{-1} (y_s - X_s \hat{\beta}) \right] \\ &= a^t \left[X \hat{\beta} + \sum_{k=0}^{p} Z_k \hat{u}_k \right] + a_s^t \left[y_s - X_s \hat{\beta} - \sum_{k=0}^{p} Z_{k,s} \hat{u}_k \right] . \end{aligned}$$

Since a^t can be written in the form $\boldsymbol{a^t = \frac{1}{N_d} \underset{1\le\ell\le D}{\mathrm{col}^t} \{\delta_{d\ell} 1_{N_\ell}^t\}}$, where $\delta_{ab} = 1$ if $a = b$ and $\delta_{ab} = 0$ if $a \neq b$, it holds that $\boldsymbol{a^t X \hat{\beta} = \sum_{k=0}^{p} \overline{X}_{kd} \hat{\beta}_k}$ and

$$\boldsymbol{a^t Z_k \hat{u}_k = \frac{1}{N_d} \underset{1\le\ell\le D}{\mathrm{col}^t} \{\delta_{d\ell} 1_{N_\ell}^t\} \underset{1\le\ell\le D}{\mathrm{diag}} (x_{k,N_\ell}) \hat{u}_k = \overline{X}_{kd} \hat{u}_{kd},}$$

where $\overline{X}_{kd} = \frac{1}{N_d} \sum_{j=1}^{N_d} x_{kdj}$. Thus the EBLUP B of $\overline{Y}_d$ is

$$\widehat{\overline{Y}}_d^{eblupB} = \sum_{k=0}^{p} \overline{X}_{kd} \hat{\beta}_k + \sum_{k=0}^{p} \overline{X}_{kd} \hat{u}_{kd} + f_d \left[\overline{y}_{d,s} - \sum_{k=0}^{p} \overline{X}_{kd,s} \hat{\beta}_k - \sum_{k=0}^{p} \overline{X}_{kd,s} \hat{u}_{kd} \right] ,$$

where $\overline{y}_{d,s} = \frac{1}{n_d} \sum_{j=1}^{n_d} y_{dj}$, $\overline{X}_{kd,s} = \frac{1}{n_d} \sum_{j=1}^{n_d} x_{kdj}$ and $f_d = \frac{n_d}{N_d}$. EBLUP under Model A is similarly introduced and it is denoted by EBLUP A in the sequel. The mean squared error (MSE) of the EBLUP and its proposed estimator are given in the next section.

4 MSE of EBLUP

Following Prasad and Rao [6] and Das, Jiang and Rao [1], the mean squared error (MSE) of the EBLUP of $\overline{Y}_d$, under Model B, is

$$MSE(\widehat{\overline{Y}}_d^{eblupB}) = g_1(\varphi) + g_2(\varphi) + g_3(\varphi) + g_4(\varphi),$$

where

$$\begin{aligned}
g_1(\varphi) &= a_r^t Z_r T_s Z_r^t a_r,\\
g_2(\varphi) &= [a_r^t X_r - a_r^t Z_r T_s Z_s^t V_{e,s}^{-1} X_s] Q_s [X_r^t a_r - X_s^t V_{e,s}^{-1} Z_s T_s Z_r^t a_r],\\
g_3(\varphi) &\approx \operatorname{tr}\left\{(\nabla b^t) V_s (\nabla b^t)^t E\left[(\hat{\varphi} - \varphi)(\hat{\varphi} - \varphi)^t\right]\right\},\\
g_4(\varphi) &= a_r^t V_{e,r} a_r,
\end{aligned}$$

and $T_s = V_u - V_u Z_s^t V_s^{-1} Z_s V_u$, $Q_s = (X_s^t V^{-1} X_s)^{-1}$, $b^t = a_r^t Z_r V_u Z_s^t V_s^{-1}$. The Prasad-Rao (PR) estimator of $MSE(\hat{\overline{Y}}_d^{eblupB})$ is

$$mse_d^B = mse(\hat{\overline{Y}}_d^{eblup\,B}) = g_1(\hat{\varphi}) + g_2(\hat{\varphi}) + 2g_3(\hat{\varphi}) + g_4(\hat{\varphi}),$$

where $\hat{\varphi}$ is REML estimator of φ. In what follows we present the calculation of $g_1 - g_4$ for Model B. The derivations under Model A are straightforward. We employ the notation $mse_d^\ell = mse(\hat{\overline{Y}}_d^{eblup\ell})$, $\ell = A, B$, under Models A and B.

4.1 Calculation of $g_1(\varphi)$ under Model B

To calculate $g_1(\varphi) = a_r^t Z_r T_s Z_r^t a_r$, basic elements are

$$\boldsymbol{a_r^t = \frac{1}{N_d} \underset{1\le\ell\le D}{\mathrm{col}^t}(\delta_{d\ell} 1_{N_\ell - n_\ell}^t)}, \quad \boldsymbol{Z_r = \underset{0\le k\le p}{\mathrm{col}^t}(Z_{k,r})}, \quad \boldsymbol{V_u = \sigma^2 \underset{0\le k\le p}{\mathrm{diag}}(\varphi_k I_D)}$$

and

$$\begin{aligned}
T_s &= V_u - V_u Z_s^t V_s^{-1} Z_s V_u = \sigma^2 \underset{0\le k\le p}{\mathrm{diag}}(\varphi_k I_D)\\
&\quad - \sigma^2 \underset{0\le k\le p}{\mathrm{col}}(\varphi_k Z_{k,s}^t) \underset{1\le\ell\le D}{\mathrm{diag}}(\Sigma_{\ell,s}^{-1}) \underset{0\le k\le p}{\mathrm{col}^t}(\varphi_k Z_{k,s}) = (T_{k_1 k_2})_{k_1,k_2=0,1,\dots,p}\,.
\end{aligned}$$

where $\delta_{k_1 k_2} = 0$ if $k_1 \neq k_2$, $\delta_{k_1 k_2} = 1$ if $k_1 = k_2$ and

$$T_{k_1 k_2} = \sigma^2 \varphi_{k_1} \delta_{k_1 k_2} I_D - \sigma^2 \varphi_{k_1} \varphi_{k_2} Z_{k_1,s}^t \underset{1\le\ell\le D}{\mathrm{diag}}(\Sigma_{\ell,s}^{-1}) Z_{k_2,s}.$$

Therefore

$$\begin{aligned}
g_1(\theta) &= \frac{1}{N_d^2} \underset{1\le\ell\le D}{\mathrm{col}^t}(\delta_{d\ell} 1_{N_\ell - n_\ell}^t) \underset{0\le k\le p}{\mathrm{col}^t}(Z_{k,r}) T_s \underset{0\le k\le p}{\mathrm{col}}(Z_{k,r}^t) \underset{1\le\ell\le D}{\mathrm{col}}(\delta_{d\ell} 1_{N_\ell - n_\ell})\\
&= (1 - f_d)^2 \sigma^2 \left\{ \sum_{k=0}^{p} \varphi_k \overline{X}_{kd}^{*2} - \sum_{k_1=0}^{p} \sum_{k_2=0}^{p} \varphi_{k_1} \varphi_{k_2} \overline{X}_{k_1 d}^{*} x_{k_1,n_d}^t \Sigma_{d,s}^{-1} x_{k_2,n_d} \overline{X}_{k_2 d}^{*} \right\},
\end{aligned}$$

where $f_d = n_d / N_d$ and $\overline{X}_{kd}^* = \frac{1}{N_d - n_d} \sum_{j\in r} x_{kdj} = (1 - f_d)^{-1} (\overline{X}_{kd} - f_d \overline{X}_{kd,s})$.

4.2 Calculation of $g_2(\varphi)$ under Model B

From the definition of $g_2(\varphi)$ it follows that it can be written in the form

$$g_2(\varphi) = [a_1^t - a_2^t] Q_s [a_1 - a_2],$$

where Q_s is defined on page 271. The first vector from the difference $[a_1^t - a_2^t]$ is

$$\boldsymbol{a_1^t = a_r^t X_r = \frac{1}{N_d} 1^t_{N_d - n_d} X_{rd} = (1 - f_d) \overline{X}^*_d,}$$

where $\overline{X}^*_d = (\overline{X}^*_{0d}, \overline{X}^*_{1d}, \ldots, \overline{X}^*_{pd})$. The second vector can be written as

$$a_2^t = a_r^t \underset{0 \le k \le p}{\mathrm{col}^t} (Z_{k,r}) T_s \underset{0 \le k \le p}{\mathrm{col}} (Z^t_{k,s}) \sigma^{-2} W_s X_s$$

and after some straightforward algebra it takes the form

$$\begin{aligned} a_2^t = (1 - f_d) \Bigg\{ & \sum_{k=0}^{p} \varphi_k \overline{X}^*_{kd} x^t_{k,n_d} \\ & - \sum_{k_1=0}^{p} \sum_{k_2=0}^{p} \varphi_{k_1} \varphi_{k_2} \overline{X}^*_{k_1 d} x^t_{k_1,n_d} \Sigma^{-1}_{d,s} x_{k_2,n_d} x^t_{k_2,n_d} \Bigg\} W_{d,s} X_{d,s}. \end{aligned}$$

4.3 Calculation of $g_3(\varphi)$ under Model B

We recall that $\boldsymbol{g_3(\varphi) \approx}$ tr $\boldsymbol{\left\{(\nabla b^t) V_s (\nabla b^t)^t E\left[(\hat{\varphi} - \varphi)(\hat{\varphi} - \varphi)^t\right]\right\}}$, where

$$b^t = a_r^t Z_r V_u Z_s^t V_s^{-1} = a_r^t \sum_{k=0}^{p} \varphi_k Z_{k,r} Z^t_{k,s} \underset{1 \le \ell \le D}{\mathrm{diag}} (\Sigma^{-1}_{\ell,s}).$$

As $\frac{\partial \Sigma_{\ell,s}}{\partial \sigma^2} = 0$ and $\frac{\partial \Sigma_{\ell,s}}{\partial \varphi_k} = x_{k,n_\ell} x^t_{k,n_\ell}$ $(k = 0, \ldots, p)$, the derivative with respect to σ^2 is $\frac{\partial b^t}{\partial \sigma^2} = 0$ and the remaining derivatives are

$$\begin{aligned} \frac{\partial b^t}{\partial \varphi_k} &= a_r^t Z_{k,r} Z^t_{k,s} \underset{1 \le \ell \le D}{\mathrm{diag}} (\Sigma^{-1}_{\ell,s}) \\ &\quad - a_r^t \left(\sum_{i=0}^{p} \varphi_i Z_{i,r} Z^t_{i,s} \right) \underset{1 \le \ell \le D}{\mathrm{diag}} (\Sigma^{-1}_{\ell,s} x_{k,n_\ell} x^t_{k,n_\ell} \Sigma^{-1}_{\ell,s}), \quad k = 0, 1, \ldots, p. \end{aligned}$$

As $\boldsymbol{Z_{k,r} = \underset{1 \le \ell \le D}{\mathrm{diag}} (x_{k,N_\ell - n_\ell})}$, we obtain for $k = 0, 1, \ldots, p$

$$\begin{aligned} \frac{\partial b^t}{\partial \varphi_k} = (1 - f_d) \Bigg[& \underset{1 \le \ell \le D}{\mathrm{col}^t} (\delta_{d\ell} \overline{X}^*_{k\ell}\, x^t_{k,n_\ell} \Sigma^{-1}_{\ell,s}) \\ & - \underset{1 \le \ell \le D}{\mathrm{col}^t} \left(\delta_{d\ell} \left(\sum_{i=0}^{p} \varphi_i \overline{X}^*_{i\ell}\, x^t_{i,n_\ell} \right) \Sigma^{-1}_{\ell,s} x_{k,n_\ell} x^t_{k,n_\ell} \Sigma^{-1}_{\ell,s} \right) \Bigg]. \end{aligned}$$

Let us define $H(\varphi) = (h_{k_1,k_2})_{k_1,k_2 = -1,0,1,\ldots,p}$, where $h_{-1,k} = h_{k,-1} = 0$, $k = -1, 0, 1, \ldots, p$ and

$$h_{k_1,k_2} = \frac{\partial b^t}{\partial \varphi_{k_1}} V_s \left(\frac{\partial b^t}{\partial \varphi_{k_2}} \right)^t = \sigma^2 (1-f_d)^2 \Bigg\{ \overline{X}^*_{k_1 d} x^t_{k_1,n_d} \Sigma^{-1}_{d,s} x_{k_2,n_d} \overline{X}^*_{k_2 d}$$
$$- \overline{X}^*_{k_1 d} x^t_{k_1,n_d} \Sigma^{-1}_{d,s} x_{k_2,n_d} x^t_{k_2,n_d} \Sigma^{-1}_{d,s} \sum_{i=0}^{p} \varphi_i x_{i,n_d} \overline{X}^*_{id}$$
$$- \left(\sum_{i=0}^{p} \varphi_i \overline{X}^*_{id} x^t_{i,n_d} \right) \Sigma^{-1}_{d,s} x_{k_1,n_d} x^t_{k_1,n_d} \Sigma^{-1}_{d,s} x_{k_2,n_d}$$
$$\cdot \left[\overline{X}^*_{k_2 d} - x^t_{k_2,n_d} \Sigma^{-1}_{d,s} \sum_{i=0}^{p} \varphi_i x_{i,n_d} \overline{X}^*_{id} \right] \Bigg\}$$

for any $k_1, k_2 = 0, 1, \ldots, p$. Then

$$\boldsymbol{g_3(\varphi)} \approx \operatorname{tr} \left\{ \boldsymbol{H(\varphi) F^{-1}(\varphi)} \right\},$$

where $F(\varphi)$ is the REML Fisher information matrix which approximates the covariance matrix $E\left[(\hat{\varphi}-\varphi)(\hat{\varphi}-\varphi)^t\right]$.

4.4 Calculation of $g_4(\varphi)$ under Model B

We recall that $g_4(\varphi) = a_r^t V_{e,r} a_r$, where

$$\boldsymbol{a_r^t = \frac{1}{N_d} \underset{1 \le \ell \le D}{\mathrm{col}^t} (\delta_{d\ell} 1^t_{N_\ell - n_\ell}), \quad V_{e,r}^{-1} = \sigma^{-2} W_r = \sigma^{-2} \underset{1 \le d \le D}{\mathrm{diag}} \{W_{d,r}\}.}$$

Therefore

$$g_4(\varphi) = \frac{\sigma^2}{N_d^2} 1^t_{N_d - n_d} \underset{j \in r}{\mathrm{diag}} \{w_{dj}^{-1}\} 1_{N_d - n_d} = \frac{\sigma^2}{N_d^2} \sum_{j \in r_d} \frac{1}{w_{dj}}.$$

5 Simulation experiments

In this section we present several simulation experiments. The first one is designed to check the behavior of the REML estimates under Model B. The second simulation experiment is planned to study the behavior of EBLUP a, $a = A, B$, under Models A and B. Finally, the fourth simulation experiment is carried out to analyze the behavior of the MSE estimates.

In all the simulations, samples are generated as follows.

- Explanatory variable: Take $a_d = 1$, $b_d = 2 + \frac{8d}{D}$, $d = 1, \ldots, D$. For $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, generate
$$x_{1dj} = (b_d - a_d) U_{dj} + a_d \ \text{ with } U_{dj} = \frac{j}{n_d + 1}, \quad j = 1, \ldots, n_d.$$
- Random effects and errors: For $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, generate
$$u_{0d} \sim N(0, \sigma^2 \varphi_0), \quad u_{1d} \sim N\left(0, \sigma^2 \varphi_1\right), \quad e_{dj} \sim N(0, \sigma^2),$$
with $\sigma^2 = \varphi_0 = 1$ and $\varphi_1 = 2$.
- Target variable: For $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, generate
$$y_{dj} = \beta_0 + \beta_1 x_{dj} + u_{1d} x_{dj} + u_{0d} + w_{dij}^{-1/2} e_{dj}, \quad \text{with } \beta_0 = 2,\ \beta_1 = 1.$$
(Just skipping the term $u_{1d} x_{dj}$ in the case of Model A.)

5.1 Simulation 1

The steps of the simulation experiment are:

1. Repeat $K = 10^4$ times $(k = 1, \dots, K)$
   1.1. Generate a sample of size $n = \sum_{d=1}^{D} n_d$ and calculate the REML estimates $\gamma_{(k)} \in \{\hat{\beta}_{0(k)}, \hat{\beta}_{(1k)}, \hat{\sigma}^2_{(k)}, \hat{\varphi}_{0(k)}, \hat{\varphi}_{1(k)}\}$.
2. Output:

$$EMSE(\hat{\gamma}) = \frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_{(k)} - \gamma)^2, \quad BIAS(\hat{\gamma}) = \frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_{(k)} - \gamma).$$

Table 1. BIAS and EMSE for $K = 10^4$ under Model B

n	300		600		1200		2400	
n_d	5		10		20		40	
$D = 60$	BIAS	EMSE	BIAS	EMSE	BIAS	EMSE	BIAS	EMSE
$\beta_0 = 2$	-0.001	0.052	-0.001	0.032	-0.002	0.024	0.000	0.020
$\beta_1 = 1$	-0.001	0.020	0.000	0.018	-0.001	0.018	0.000	0.017
$\sigma^2 = 1$	0.006	0.010	0.002	0.004	0.001	0.002	0.001	0.001
$\varphi_0 = 1$	-0.050	0.335	-0.007	0.129	-0.001	0.070	0.002	0.050
$\varphi_1 = 1$	-0.020	0.055	-0.005	0.043	-0.002	0.038	-0.002	0.037

Table 1 presents the obtained performance measures. In all the presented cases we observe that EMSE decreases as sample size increases. The conclusion is that the implemented Fisher-scoring algorithm is running properly and thus the obtained REML parameter estimates are reliable.

5.2 Simulation 2

The second simulation experiment is designed to investigate the behavior of EBLUPa, $a = A, B$, under Models A and B. The steps of simulation experiment are:

1. Generate deterministically $N = \sum_{d=1}^{D} N_d$ x-values with $N_d = 100$, $D = 60$ as described at the beginning of this section and calculate $\overline{X}_d$, $d = 1, \dots, D$.
2. Repeat $K = 10^4$ times $(k = 1, \dots, K)$
   2.1. Generate a population of size N and extract a sample of size $n = \sum_{d=1}^{D} n_d$ $(n_d = 10)$ under Model B (Model A).
   2.2 Calculate the REML estimates under Models A and B.
   2.3 Calculate the true value $\overline{Y}_d^{(k)}$ and its estimates $\widehat{\overline{Y}}_d^{eblup\,a(k)}$ for $a = A, B$.
3. For any $a = A, B$ the output is:

$$mean_d^a = \frac{1}{K}\sum_{k=1}^{K}\widehat{\overline{Y}}_d^{eblup\,a(k)}, \quad MEAN_d = \frac{1}{K}\sum_{k=1}^{K}\overline{Y}_d^{(k)},$$

$$EMSE_d^a = \frac{1}{K}\sum_{k=1}^{K}(\widehat{\overline{Y}}_d^{eblup\,a(k)} - \overline{Y}_d^{(k)})^2, \quad EMSE^a = \frac{1}{D}\sum_{d=1}^{D} EMSE_d^a,$$

and

$$BIAS_d^a = \frac{1}{K}\sum_{k=1}^{K}(\widehat{\overline{Y}}_d^{eblup\,a(k)} - \overline{Y}_d^{(k)}), \quad BIAS^a = \frac{1}{D}\sum_{d=1}^{D} BIAS_d^a.$$

Table 2 presents the basic performance measures of simulations 2. $DIFr$, $r = A, B$, is used to denote the differences $EMSE^a - EMSE^r$, $a = A, B$, $a \neq r$. Figure 1 plot $EMSE_d^a$ of estimators $\widehat{\overline{Y}}_d^{eblup\,a}$, $a = A, B$ under Model A and B, respectively.

We observe that if Model B is true, EBLUP estimate may lose a significative amount of precision by assuming the wrong Model A. However, the loss of efficiency negligible in the reciprocal case.

Table 2. BIAS and EMSE for $D = 60$ and $K = 10^4$

	Model B		Model A	
$N_d = 100$, $n_d = 10$	eblupA	eblupB	eblupA	eblupB
$10^2 BIAS$	0.0046	0.0065	0.0992	0.0993
$10^2 EMSE$	10.7272	8.513	8.2237	8.2253
$10^2 DIFr$	2.2142			0.0016

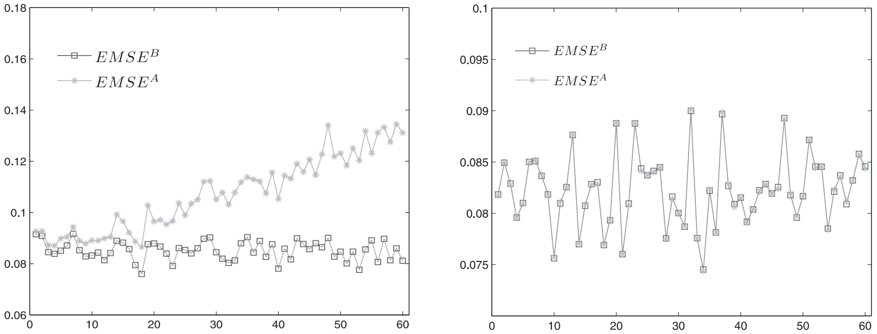

Fig. 1. $EMSE_d$ values under the true Model B (left) and Model A (right)

5.3 Simulation 3

The third simulation experiment is designed to analyze the behavior of the MSE estimates. The steps of the simulation experiment under Model B (Model A) are:

1-2. Do steps 1-2.3 as in Simulation 2. Do new step 2.4 as follows.
  2.4. Calculate the MSE estimates $mse_d^{A(k)}$ and $mse_d^{B(k)}$.

3. For $a = A, B$ the output is:

$$E_d^a = \frac{1}{K}\sum_{k=1}^{K}(mse_d^{a(k)} - EMSE_d^a)^2, \quad B_d^a = \frac{1}{K}\sum_{k=1}^{K}(mse_d^{a(k)} - EMSE_d^a).$$

$$E^a = \frac{1}{D}\sum_{d=1}^{D} E_d^a, \quad B^a = \frac{1}{D}\sum_{d=1}^{D} B_d^a,$$

where the values $EMSE_d^a$ are taken from the results of Simulation 2.

Table 3 presents basic performance measures of Simulation 3. From the table it can be seen that the two estimators mse_d^B and mse_d^A have basically the same behavior under the true Model A. However, under the true Model B mse_d^A has a very poor behavior when it is used to estimate $MSE(\widehat{\overline{Y}}_d^{eblupA})$.

Table 3. B^a and E^a values for $K = 10^4$

$N_d = 100$	Model B		Model A	
$n_d = 10$	mse_d^A	mse_d^B	mse_d^A	mse_d^B
$10^2 B$	46.5629	0.0098	-0.0146	0.0054
$10^2 E$	23.0612	0.0029	0.0025	0.0026

6 Estimation of poverty proportions

In this section we use data from the 2006 Spanish Living Conditions Survey (SLCS) with global sample size 34694. 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. Its main goal is to provide a reference source on comparative statistics on the distribution of income and social exclusion in the European environment. The sample includes 16000 dwellings distributed in 2000 census sections.

We consider $D = 52$ domains (provinces) and we are interested in estimating the domain averages of the household normalized net annual incomes. 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 is the modified OECD scale used by EUROSTAT, where OECD is the acronym for the Organization for Economic Cooperation and Development. This scale gives a weight of 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 the total number of normalized household members is

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

where $H_{di\geq 14}$ is the number of people aged 14 and over and $H_{di<14}$ is the number of children aged under 14. The normalized net annual income of a household (z)

is obtained by dividing its net annual income by its normalized size. Following the standards of the Spanish Statistical Office, the Poverty Threshold is fixed as the 60% of the median of the normalized incomes in Spanish households. The Spanish poverty thresholds (in euros) in 06 is $z_{2006} = 6556.60$. This is z_0-value used in the calculation of the direct estimates of the poverty proportion

$$\overline{Y}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj}, \quad y_{dj} = I(z_{dj} < z_0),$$

where $I(z_{dj} < z_0) = 1$ if $z_{dj} < z_0$ and $I(z_{dj} < z_0) = 0$ otherwise.

The considered auxiliary variables are *nationality* (x_0) and *employed* (x_1), both with values 0-1 at the individual level (1 for Spanish citizenship and employed). In the SLCS the target variable y is measured at the household level and the auxiliary variables x_1 and x_2 at the individual level. For this reason a data file has been built containing the survey data aggregated at the level of census sections (territories with around 2000 people). In the census section file the y variable remains unchanged and the x-variables are calculated by taking weighted averages on the territory.

Table 4 presents the REML estimates of model parameters and the corresponding 90% confidence intervals. We observe that confidence intervals for parameters φ_0 and φ_1 are strictly positive, suggesting that Model B fits better to data than Model A.

Figure 2 presents the domain mean estimates and their estimated mean squared error. It shows that EBLUP B has slightly different behavior from EBLUP A estimates. Figure 2 also shows that the EBLUP estimates behave more smoothly than the direct ones, which are calculated by means of the formula

$$\hat{\overline{Y}}_d^{dir} = \frac{1}{\hat{N}_d}\sum_{j=1}^{n_d} \omega_{dj} y_{dj}, \quad \hat{N}_d = \sum_{j=1}^{n_d} \omega_{dj},$$

where the ω_{dj}'s are SLCS calibrated sampling weights.

Concerning mean squared errors, EBLUP B is the estimators giving the best results. EBLUP estimators produce some gain of efficiency with respect to the direct ones. For comparison purposes, design-based mean squared errors of direct estimators where approximated by

$$mse(\hat{\overline{Y}}_d^{dir}) = \frac{1}{\hat{N}_d^2}\sum_{j=1}^{n_d} \omega_{dj}(\omega_{dj}-1)\big(y_{dj} - \hat{\overline{Y}}_d^{dir}\big)^2. \tag{7}$$

The last formula is taken from Särndal *et al.* [9], pp. 43, 185 and 391, with the simplifications $\omega_{dj} = 1/\pi_{dj}$, $\pi_{dj,dj} = \pi_{dj}$ and $\pi_{di,dj} = \pi_{di}\pi_{dj}$, $i \neq j$ in the second order inclusion probabilities.

By observing the signs of the regression parameters, we interpret that poverty proportion tends to be smaller in those domains with larger proportion of people with non Spanish citizenship (may be because immigrants tends to go to regions with greater richness where it is easier to find job) and larger proportion of employed people.

Table 4. Parameter estimates and 90% confidence intervals for models B and A

	Model A		**Model B**	
	Estim.	90% CI	Estim.	90% CI
β_0	**0.2942**	(0.1722 , 0.4162)	**0.3336**	(0.2197 , 0.4475)
β_1	**-0.2900**	(-0.4946 , -0.0854)	**-0.2958**	(-0.5294 , -0.0621)
σ^2	**0.0453**	(0.0430 , 0.0477)	**0.0457**	(0.0433 , 0.0480)
φ_0	**0.1481**	(0.0865 , 0.2098)	**0.0689**	(0.0061 , 0.1317)
φ_1			**0.1382**	(0.0478 , 0.2287)

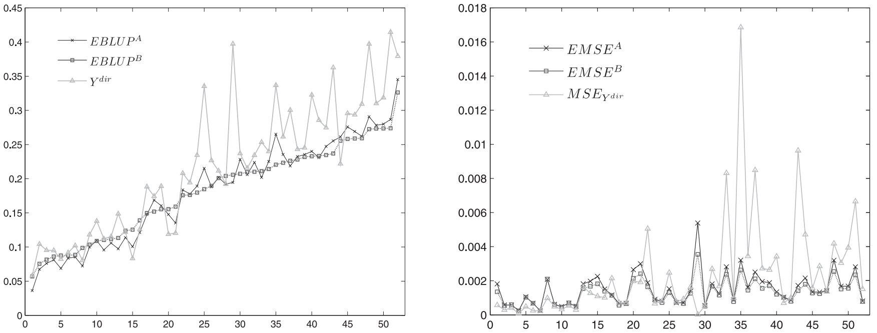

Fig. 2. Direct estimates and EBLUP estimates (left) and its estimated mean square error (right)

7 Conclusions

This paper investigate the use of EBLUPs, based on random regression coefficient models, in small area estimation. By looking at the presented simulations and application to real data, we may conclude that fixed regression coefficients are sometimes too rigid for modeling real data. Some extra variability, and better performance of EBLUP estimates, might be obtained by allowing some variability on the regression (beta) parameters.

Acknowledgement. The authors are grateful to the Czech and the Spanish Government for their economical support under the grants MSMTV 1M0572, MSM6840770039 and MTM2009-09473. The authors also thanks the Instituto Nacional de Estadística for providing the Spanish EU-SILC data.

References

1. Das K, Jiang J, Rao JNK (2004) Mean squared error of empirical predictor. Ann Statist 32:818–840

2. Ghosh M, Rao JNK (1994) Small area estimation: An appraisal. Statist Sci 9:55–93
3. Jiang J, Lahiri P (2006) Mixed model prediction and small area estimation. TEST 15:1–96
4. Moura FAS, Holt D (1999). Small area estimation using multilevel models. Surv Meth 25(N.1):73–80
5. Pfeffermann D (2002) Small Area Estimation. New Developments and Directions. Int Statist Rev 70:125–143
6. Prasad NGN, Rao JNK (1990) The estimation of the mean squared error of small-area estimators. J Amer Statist Assoc 85:163–171
7. Rao JNK (1999) Some recent advances in model-based small area estimation. Surv Meth 25:175–186
8. Rao JNK (2003) Small Area Estimation. John Wiley, New York
9. Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, Berlin