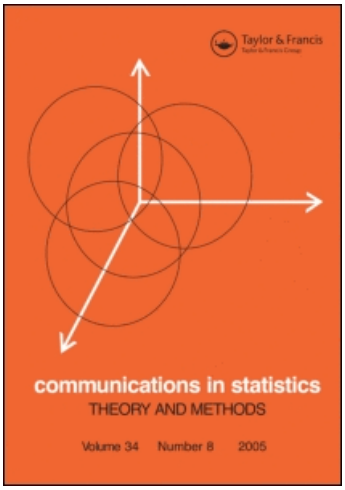

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

A Fay-Herriot Model with Different Random Effect Variances

M. Herrador[a]; M. D. Esteban[b]; T. Hobza[cd]; D. Morales[b]
[a] Instituto Nacional de Estadística, Madrid, Spain [b] Operations Research Center, Miguel Hernández University of Elche, Elche, Spain [c] Department of Mathematics, Czech Technical University in Prague, Prague, Czech Republic [d] Institute of Information Theory and Automation of the ASCR, Prague, Czech Republic

Online publication date: 02 December 2010

To cite this Article 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, 40: 5, 785 — 797
To link to this Article: DOI: 10.1080/03610920903480858
URL: http://dx.doi.org/10.1080/03610920903480858

Communications in Statistics—Theory and Methods, 40: 785–797, 2011

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920903480858

A Fay–Herriot Model with Different Random Effect Variances

M. HERRADOR[1], M. D. ESTEBAN[2], T. HOBZA[3,4], AND D. MORALES[2]

[1]Instituto Nacional de Estadística, Madrid, Spain
[2]Operations Research Center, Miguel Hernández University of Elche, Elche, Spain
[3]Department of Mathematics, Czech Technical University in Prague, Prague, Czech Republic
[4]Institute of Information Theory and Automation of the ASCR, Prague, Czech Republic

A modification of the Fay–Herriot model is introduced to treat situations where small areas are divided in two groups and domain random effects have different variances across the groups. The model is applicable to data having a large subset of domains where direct estimates of the variable of interest cannot be described in the same way as in its complementary subset of domains. This is generally the case when domains are constructed by crossing geographical characteristics with sex. Algorithms and formulas to fit the model, to calculate EBLUPs and to estimate mean squared errors are given. Monte Carlo simulation experiments are presented to illustrate the gain of precision obtained by using the proposed model and to get some practical conclusions. A motivating application to Spanish Labour Force Survey data is also given.

Keywords EBLUP; Fay–Herriot model; Labour Force Survey; Linear mixed models; Small area estimation.

Mathematics Subject Classification Primary 62J05; Secondary 62D06.

1. Introduction

Linear mixed models are widely used in applied statistics. Searle et al. (1982) provided a detailed description of linear mixed models and Ghosh and Rao (1994), and more recently Rao (2003) and Jiang and Lahiri (2006), discussed their applications to small area estimation. In this last setup, the basic area level linear mixed model was introduced by Fay and Herriot (1979). This model typically assumes that the domain random effect have a common constant variance.

Received November 12, 2008; Accepted November 11, 2009
Address correspondence to D. Morales, Operations Research Center, Miguel Hernández University of Elche, Elche, Spain; E-mail: d.morales@umh.es

However, when estimating totals or means we may often find that domains can be divided in two groups where direct estimates behave in a different manner within them; for example they might have different variability. This situation may happen if we are interested in producing estimates by sex. In those cases, traditional random intercept models do not fit well to data and some extra parameters are needed in the model.

In this article, we extend the Fay–Herriot model to an area level linear regression model with random intercepts having one of two possible variances. Estimation procedures for the variance components and regression parameters are considered and EBLUP estimators of domain parameters are derived. The approximation given by Prasad and Rao (1990) and extended to a general class of linear mixed models by Das et al. (2004) is applied to obtain estimators of the mean squared errors of the EBLUP estimates.

This article is organized as follows. In Sec. 2, we introduce the proposed model, we give a Fisher-scoring algorithm to calculate the maximum likelihood estimators of model parameters, we derive the expression of the EBLUP estimator of a domain linear parameter, and we give an estimator of its mean squared error (MSE). In Secs. 3–4, we carry out simulation experiments to investigate the behavior of the EBLUP estimates under some proposed setups. In Sec. 5, we illustrate the use of the proposed model with data from the Spanish Labour Force Survey (SLFS) and from some administrative registers. Finally, in Sec. 6 we give some conclusions.

2. The Model

We suppose a model where domains are divided in two groups, denoted by A and B, and variances of random intercepts varies across the groups. The model is

$$y_d = \boldsymbol{x}_d\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D = D_A + D_B, \tag{2.1}$$

where $u_1, \ldots, u_{D_A} \sim N(0, \sigma_A^2)$, $u_{D_A+1}, \ldots, u_D \sim N(0, \sigma_B^2)$, $e_1 \sim N(0, \sigma_1^2), \ldots, e_D \sim N(0, \sigma_D^2)$; they are all mutually independent and the variances $\sigma_1^2, \ldots, \sigma_D^2$ are known. In matrix notation the model is

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

where $\boldsymbol{y} = \boldsymbol{y}_{D\times 1} = (\boldsymbol{y}_A^t, \boldsymbol{y}_B^t)^t$, $\boldsymbol{X} = \boldsymbol{X}_{D\times p} = (\boldsymbol{X}_A^t, \boldsymbol{X}_B^t)^t$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{p\times 1}$, $\boldsymbol{Z} = \boldsymbol{I}_{D\times D}$, $\boldsymbol{u} = \boldsymbol{u}_{D\times 1} = (\boldsymbol{u}_A^t, \boldsymbol{u}_B^t)^t$ and $\boldsymbol{e}_{D\times 1} = (\boldsymbol{e}_A^t, \boldsymbol{e}_B^t)^t$. In this case, $\boldsymbol{V}_u = \mathrm{var}(\boldsymbol{u}) = \mathrm{diag}(\sigma_A^2\boldsymbol{I}_{D_A}, \sigma_B^2\boldsymbol{I}_{D_B})$ and $\boldsymbol{V} = \mathrm{var}(\boldsymbol{y}) = \mathrm{diag}(\boldsymbol{V}_A, \boldsymbol{V}_B)$ with $\boldsymbol{V}_A = \mathrm{diag}(v_1^2, \ldots, v_{D_A}^2)$, $\boldsymbol{V}_B = \mathrm{diag}(v_{D_A+1}^2, \ldots, v_D^2)$, $v_d^2 = \sigma_d^2 + \sigma_A^2$ if $d = 1, \ldots, D_A$ and $v_d^2 = \sigma_d^2 + \sigma_B^2$ if $d = D_A + 1, \ldots, D$.

If $\sigma_A^2 > 0$ and $\sigma_B^2 > 0$ are known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) of $\boldsymbol{u}$ are

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y} \quad \text{and} \quad \hat{\boldsymbol{u}} = \boldsymbol{V}_u\boldsymbol{Z}^t\boldsymbol{V}^{-1}\big(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\big).$$

Components of $\hat{\boldsymbol{u}}$ are

$$\hat{u}_d = \left(\frac{\sigma_A^2}{\sigma_A^2 + \sigma_d^2} I_{\{d\le D_A\}}(d) + \frac{\sigma_B^2}{\sigma_B^2 + \sigma_d^2} I_{\{d>D_A\}}(d)\right)(y_d - \boldsymbol{x}_d\hat{\boldsymbol{\beta}}), \quad d = 1, \ldots, D,$$

and BLUP of the components of the linear parameter $\boldsymbol{\tau} = \boldsymbol{X\beta} + \boldsymbol{Zu}$ are

$$\hat{\tau}_d^{blup} = \boldsymbol{x}_d\hat{\boldsymbol{\beta}} + \boldsymbol{z}_d\hat{\boldsymbol{u}} = \boldsymbol{x}_d\hat{\boldsymbol{\beta}} + \hat{u}_d, \quad d = 1, \ldots, D, \tag{2.2}$$

where $\boldsymbol{x}_d$ ($\boldsymbol{z}_d$) is the row d of matrix $\boldsymbol{X}$ ($\boldsymbol{Z}$). EBLUP of the components of $\boldsymbol{\tau}$ are obtained by substituting σ_A^2 and σ_B^2 by estimators $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$, respectively, in (2.2).

2.1. *Maximum Likelihood Estimates*

The parameter space of the supposed model is

$$\Theta = \left\{\boldsymbol{\theta}^t = (\boldsymbol{\beta}^t, \sigma_A^2, \sigma_B^2) : \boldsymbol{\beta} \in R^p, \sigma_A^2 \geq 0, \sigma_B^2 \geq 0\right\} \tag{2.3}$$

and the corresponding log-likelihood functions is

$$\ell(\boldsymbol{\beta}, \sigma_A^2, \sigma_B^2; \boldsymbol{y}) = -\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^t\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}).$$

The derivatives of the log-likelihood function with respect to parameters are

$$\boldsymbol{S}_\beta = \sum_{d=1}^{D} \boldsymbol{x}_d^t \frac{y_d - \boldsymbol{x}_d\boldsymbol{\beta}}{v_d^2},$$

$$S_{\sigma_A^2} = -\frac{1}{2}\sum_{d=1}^{D_A}\frac{1}{v_d^2} + \frac{1}{2}\sum_{d=1}^{D_A}\frac{(y_d - \boldsymbol{x}_d\boldsymbol{\beta})^2}{v_d^4},$$

$$S_{\sigma_B^2} = -\frac{1}{2}\sum_{d=D_A+1}^{D}\frac{1}{v_d^2} + \frac{1}{2}\sum_{d=D_A+1}^{D}\frac{(y_d - \boldsymbol{x}_d\boldsymbol{\beta})^2}{v_d^4}.$$

The components of the Fisher information matrix are $\boldsymbol{F}_{\beta\sigma_A^2} = \boldsymbol{F}_{\beta\sigma_B^2} = \boldsymbol{F}_{\sigma_A^2\sigma_B^2} = \boldsymbol{0}$ and

$$\boldsymbol{F}_{\beta\beta} = \sum_{d=1}^{D} v_d^{-2}\boldsymbol{x}_d^t\boldsymbol{x}_d, \quad F_{\sigma_A^2\sigma_A^2} = \frac{1}{2}\sum_{d=1}^{D_A} v_d^{-4}, \quad F_{\sigma_B^2\sigma_B^2} = \frac{1}{2}\sum_{d=D_A+1}^{D} v_d^{-4}.$$

Updating equations of the Fisher-scoring algorithm are

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + F_{\beta^{(k)}\beta^{(k)}}^{-1} S_{\beta^{(k)}}, \quad \sigma_C^{2(k+1)} = \sigma_C^{2(k)} + F_{\sigma_C^{2(k)}\sigma_C^{2(k)}}^{-1} S_{\sigma_C^{2(k)}}, \quad C = A, B. \tag{2.4}$$

2.2. *MSE of EBLUP*

Prasad and Rao (1990) gave an approximation to the mean squared error of the EBLUP in Fay–Herriot models. In our case, the approximation is

$$MSE\left(\hat{\tau}_d^{eblup}\right) \approx g_1\left(\sigma_A^2, \sigma_B^2\right) + g_2\left(\sigma_A^2, \sigma_B^2\right) + g_3\left(\sigma_A^2, \sigma_B^2\right),$$

where

$$g_1\left(\sigma_A^2, \sigma_B^2\right) = \frac{\sigma_A^2\sigma_d^2}{\sigma_A^2 + \sigma_d^2} I_{\{d \leq D_A\}}(d) + \frac{\sigma_B^2\sigma_d^2}{\sigma_B^2 + \sigma_d^2} I_{\{d > D_A\}}(d),$$

$$g_2(\sigma_A^2, \sigma_B^2) = \left(\frac{\sigma_d^4}{(\sigma_A^2 + \sigma_d^2)^2} I_{\{d \le D_A\}}(d) + \frac{\sigma_d^4}{(\sigma_B^2 + \sigma_d^2)^2} I_{\{d > D_A\}}(d) \right) \boldsymbol{x}_d \boldsymbol{F}_{\beta\beta}^{-1} \boldsymbol{x}_d^t$$

$$g_3(\sigma_A^2, \sigma_B^2) = \frac{\sigma_d^4}{(\sigma_A^2 + \sigma_d^2)^3} var(\hat{\sigma}_A^2) I_{\{d \le D_A\}}(d) + \frac{\sigma_d^4}{(\sigma_B^2 + \sigma_d^2)^3} var(\hat{\sigma}_B^2) I_{\{d \le D_A\}}(d),$$

where $var(\hat{\sigma}_C^2) \approx F^{-1}_{\sigma_C^2, \sigma_C^2}$, $C = A, B$. Mean squared error is estimated by

$$mse(\hat{\tau}_d^{eblup}) = g_1(\hat{\sigma}_A^2, \hat{\sigma}_B^2) + g_2(\hat{\sigma}_A^2, \hat{\sigma}_B^2) + 2g_3(\hat{\sigma}_A^2, \hat{\sigma}_B^2). \tag{2.5}$$

3. Simulation Experiment at the Area Level

The scope of this simulation experiment is to investigate the loss of precision of the EBLUP based on the standard Fay–Herriot model when the true model is (2.1). For this sake, we consider the model (2.1) with D ($D = 60$) and $D_A = D/5$ ($D_A = 12$). The algorithm of the simulation experiment is described by the following steps.

1. Sample generation

Model parameters are $\sigma_A^2 = 1$, $\sigma_B^2 = 3$, $\beta_1 = \beta_2 = 1$ and $\sigma_d^2 = 1$, $d = 1, \ldots, D$. The $p = 2$ auxiliary variables are

$$x_{1d} = 1, \quad d = 1, \ldots, D_A; \quad x_{1d} = 4 + \frac{d}{D - D_A}, \quad d = D_A + 1, \ldots, D;$$

$$x_{2d} = \frac{d}{D}, \quad d = 1, \ldots, D.$$

Target variable is

$$y_d = \beta_1 x_{1d} + \beta_2 x_{2d} + u_d + e_d, \quad d = 1, \ldots, D,$$

where $u_d \sim \mathcal{N}(0, \sigma_A^2)$ if $d = 1, \ldots, D_A$, $u_d \sim \mathcal{N}(0, \sigma_B^2)$ if $d = D_A + 1, \ldots, D$ and $e_d \sim \mathcal{N}(0, \sigma_d^2)$ are independent.

2. Parameter estimation and prediction

For each area d, the parameter of interest is

$$\tau_d = \beta_1 x_{1d} + \beta_2 x_{2d} + u_d, \quad d = 1, \ldots, D.$$

We calculate: (1) the maximum likelihood estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ of the model parameters, using the Fisher-Scoring algorithm (2.4) with the corresponding formulas for the Fisher information matrix $\boldsymbol{F}$ and for the vector of scores $\boldsymbol{S}$ from model (2.1); (2) the EBLUP $\hat{\tau}_d^{eblup}$ of τ_d using the formula (2.2); (3) the MSE estimator $mse_d(\hat{\tau}_d^{eblup})$ using the formula (2.5); (4) the maximum likelihood estimates $\hat{\beta}_1^*$, $\hat{\beta}_2^*$, $\hat{\sigma}_A^{2*}$, $\hat{\sigma}_B^{2*}$ using the Fisher-Scoring algorithm (2.4) under the assumption that $D_A = 0$, i.e., under the standard Fay–Herriot model; (5) the corresponding EBLUP $\hat{\tau}_d^{eblup*}$ of τ_d^* using the formulas (2.2) under $D_A = 0$; and (6) the MSE estimator $mse(\hat{\tau}_d^{eblup*})$ using the formula (2.5) under $D_A = 0$.

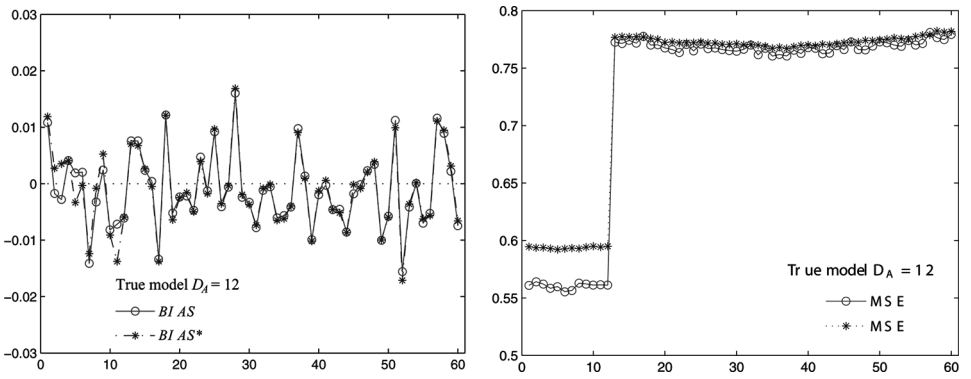

Figure 1. $BIAS_d$, $BIAS_d^*$ (left) and MSE_d, MSE_d^* (right) values for $D_A = 12$.

3. Repetition and performance measures

Steps 1–2 are repeated $K = 10^4$ times obtaining thus in each iteration $\tau_d^{(k)}$, $\hat{\tau}_d^{eblup(k)}$, and $mse\left(\hat{\tau}_d^{eblup(k)}\right)$. The following performance measures are calculated:

$$MEAN_d = \frac{1}{K}\sum_{k=1}^{K}\tau_d^{(k)}, \quad mean_d = \frac{1}{K}\sum_{k=1}^{K}\hat{\tau}_d^{eblup(k)}, \quad BIAS_d = mean_d - MEAN_d,$$

$$MSE_d = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\tau}_d^{eblup(k)} - \tau_d^{(k)}\right)^2, \quad mse_d = \frac{1}{K}\sum_{k=1}^{K} mse\left(\hat{\tau}_d^{eblup(k)}\right),$$

and also, in the same way, $mean_d^*$, $BIAS_d^*$, MSE_d^*, and mse_d^*.

Concerning the estimation of τ_d, performance measures are plotted in Fig. 1. Concerning the estimation of the mean squared error of estimators of τ_d, performance measures are plotted in Figs. 2 and 3. In Fig. 1 (left), we observe that both EBLUPs are basically unbiased. We do not notice any significant increment of bias because of not calculating the EBLUP under the true model. In Fig. 1 (right), we observe that MSEs of EBLUPs derived under the incorrect model with $D_A = 0$

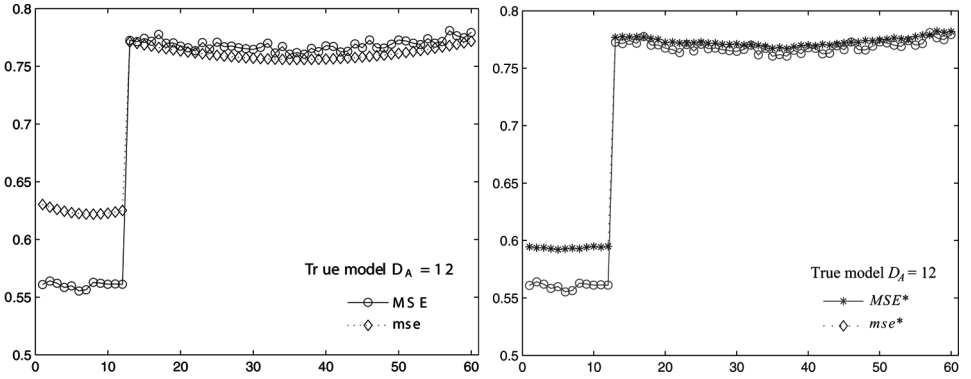

Figure 2. MSE_d, mse_d (left) and MSE_d^*, mse_d^* (right) values for $D_A = 12$.

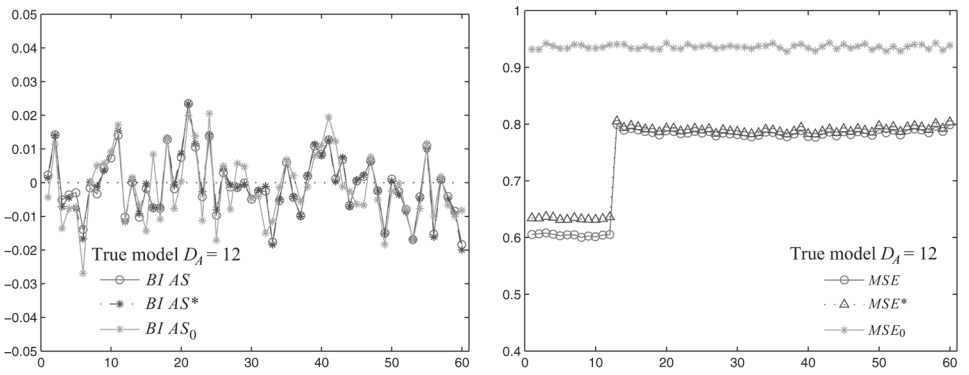

Figure 3. $BIAS_d$, $BIAS_d^*$, $BIAS_d^0$ (left) and MSE_d, MSE_d^* MSE_d^0 (right) values for $D_A = 12$.

(MSE^*) are greater than the ones of the EBLUPs derived under the true model (MSE) in the domains of the part A and slightly greater in the remaining domains.

In Fig. 2, we observe a similar pattern. If the EBLUP and its MSE estimator are derived under the true model with $D_A = 12$, then the MSE estimator is basically unbiased having a small positive bias in the part A. However, if they are derived under the incorrect model with $D_A = 0$, then a high positive bias appears in the domains of the part A and a small negative bias in the domains of the part B.

4. Simulation Experiment at the Unit Level

In this section, we implement a simulation experiment based on a unit-level model producing basically the same area-level model as the one considered in Sec. 3. The target is to investigate the gain of precision with respect to direct estimates when the area-level model holds. We consider a population generated by a unit-level model with D ($D = 60$) small areas and $D_A = 12$ small areas in group A. We extract deterministic samples within each domain. The algorithm of the simulation experiment is described by the following steps.

1. Population generation

Take $p = 2$, $\beta_1 = \beta_2 = 1$, $\sigma_A^2 = 1$, $\sigma_B^2 = 3$, $\sigma_{ed}^2 = n_d$, $N_d = 100$, and $n_d = 5$. For $j = 1, \ldots, N_d$, calculate

$$x_{1dj} = 1, d = 1, \ldots, D_A, \quad x_{1dj} = 4 + \frac{d}{D - D_A}, \quad d = D_A + 1, \ldots, D.$$

For $d = 1, \ldots, D$, $j = 1, \ldots, N_d$, calculate

$$a_{dj} = \frac{j}{1 + N_d}, \quad x_{2dj} = \frac{d + a_{dj}}{D}, \quad \overline{X}_{kd} = \frac{1}{N_d} \sum_{j=1}^{N_d} x_{kdj}, \quad k = 1, 2.$$

Generate the y-values

$$y_{dj} = x_{1dj}\beta_1 + x_{2dj}\beta_2 + u_d + e_{dj}, \quad d = 1, \ldots, D, \quad j = 1, \ldots, N_d,$$

where $u_d \sim \mathcal{N}(0, \sigma_A^2)$ if $d = 1, \ldots, D_A$, $u_d \sim \mathcal{N}(0, \sigma_B^2)$ if $d = D_A + 1, \ldots, D$ and $e_{dj} \sim N(0, \sigma_{ed}^2)$ are independent, and calculate $\overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}$.

2. Sample extraction

From each domain d we extract a sample s_d of size n_d by selecting the elements with indexes

$$(d, j(\ell)) = \left(d, \left[\frac{N_d}{n_d + 1}\right] \cdot \ell\right), \quad d = 1, \ldots, D, \ \ \ell = 1, \ldots, n_d.$$

We calculate the direct estimator of $\overline{Y}_d$ and the estimator of its design-based variance under simple random sampling (without replacement), i.e.,

$$\overline{y}_d = \frac{1}{n_d} \sum_{j \in s_d} y_{dj} \quad \text{and} \quad \sigma_d^2 = \widehat{V}_\pi(\overline{y}_d) = \frac{N_d - n_d}{N_d n_d (n_d - 1)} \sum_{j \in s_d} (y_{dj} - \overline{y}_d)^2.$$

3. Parameter estimation and prediction

For each area d the parameter of interest is

$$\overline{Y}_d = \tau_d + \overline{e}_d = \overline{X}_{1d}\beta_1 + \overline{X}_{2d}\beta_2 + u_d + \overline{e}_d, \quad d = 1, \ldots, D,$$

where $\overline{e}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} e_{dj}$. To obtain EBLUP estimates of $\overline{Y}_d$ we assume that the area-level model derived from the unit-level model holds, i.e., we assume that

$$\overline{y}_d = \overline{X}_{1d}\beta_1 + \overline{X}_{2d}\beta_2 + u_d + \varepsilon_d, \quad d = 1, \ldots, D,$$

where $u_d \sim \mathcal{N}(0, \sigma_A^2)$ if $d = 1, \ldots, D_A$, $u_d \sim \mathcal{N}(0, \sigma_B^2)$ if $d = D_A + 1, \ldots, D$ and $\varepsilon_d \sim N(0, \sigma_d^2)$ are independent. An important difference of this simulation experiment is that we use $\hat{\tau}_d^{eblup}$ to estimate $\overline{Y}_d$ (instead of τ_d). Note that $\overline{Y}_d - \tau_d = \overline{e}_d$ has a 3-sigma range $\pm 3(5/100)^{1/2} = \pm 0.671$, which is not negligible. We calculate: (1) the maximum likelihood estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}_A^2$, and $\hat{\sigma}_B^2$ of the model parameters; (2) the EBLUP $\hat{\tau}_d^{eblup}$ of the mean of each area d; (3) the MSE estimator $mse_d\big(\hat{\tau}_d^{eblup}\big)$; (4) the maximum likelihood estimates $\hat{\beta}_1^*$, $\hat{\beta}_2^*$, $\hat{\sigma}_A^{2*}$ and $\hat{\sigma}_B^{2*}$ under the assumption $D_A = 0$, i.e., under the standard Fay–Herriot model; (5) the EBLUP $\hat{\tau}_d^{eblup*}$ of the mean of each area d under $D_A = 0$; and (6) the MSE estimator $mse(\hat{\tau}_d^{eblup*})$ under $D_A = 0$.

4. Repetition and performance measures

Steps 1–3 are repeated $K = 10^4$ times obtaining in each iteration $\overline{Y}_d^{(k)}$, $\overline{y}_d^{(k)}$, $\hat{\tau}_d^{eblup(k)}$, $\widehat{V}_\pi\big(\overline{y}_d^{(k)}\big)$, and $mse\big(\hat{\tau}_d^{eblup(k)}\big)$. The following performance measures are calculated:

$$MEAN_d = \frac{1}{K} \sum_{k=1}^{K} \overline{Y}_d^{(k)}, \quad mean_d^0 = \frac{1}{K} \sum_{k=1}^{K} \overline{y}_d^{(k)}, \quad mean_d = \frac{1}{K} \sum_{k=1}^{K} \hat{\tau}_d^{eblup(k)},$$

$$BIAS_d^0 = mean_d^0 - MEAN_d, \quad BIAS_d = mean_d - MEAN_d,$$

$$MSE_d^0 = \frac{1}{K} \sum_{k=1}^{K} \left(\overline{y}_d^{(k)} - \overline{Y}_d^{(k)}\right)^2, \quad MSE_d = \frac{1}{K} \sum_{k=1}^{K} \left(\hat{\tau}_d^{eblup(k)} - \overline{Y}_d^{(k)}\right)^2,$$

$$mse_d^0 = \frac{1}{K} \sum_{k=1}^{K} \widehat{V}_\pi\big(\overline{y}_d^{(k)}\big), \quad mse_d = \frac{1}{K} \sum_{k=1}^{K} mse\big(\hat{\tau}_d^{eblup(k)}\big),$$

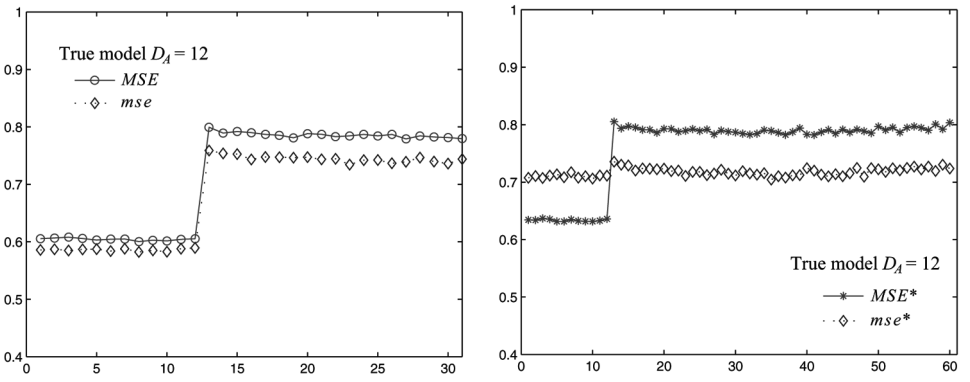

Figure 4. MSE_d, mse_d (left) and MSE_d^*, mse_d^* (right) values for $D_A = 12$.

$$E_d^0 = \left[\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{V}_\pi\left(\overline{y}_d^{(k)}\right) - MSE_d^0\right)^2\right]^{1/2},$$

$$E_d = \left[\frac{1}{K}\sum_{k=1}^{K}\left(mse\left(\hat{\tau}_d^{eblup(k)}\right) - MSE_d\right)^2\right]^{1/2},$$

and also, in the same way, $mean_d^*$, $BIAS_d^*$, MSE_d^*, mse_d^*, and E_d^*.

Figure 3 plot the performance measures for the estimators of $\overline{Y}_d$ and Figs. 4 and 5 plot the performance measures for the estimation of the corresponding mean squared errors.

In Fig. 3 (left), we observe that if the EBLUP estimator is derived under the true model with $D_A = 12$, then its bias is basically negligible. The same happens for the direct estimate and for the EBLUP derived under incorrect model with $D_A = 0$. So, bias of population mean estimates does not seem to play a relevant role in this study. In Fig. 3 (right), we observe that the MSEs of EBLUPs derived under the incorrect model are of similar size of the ones of EBLUPs derived under the true

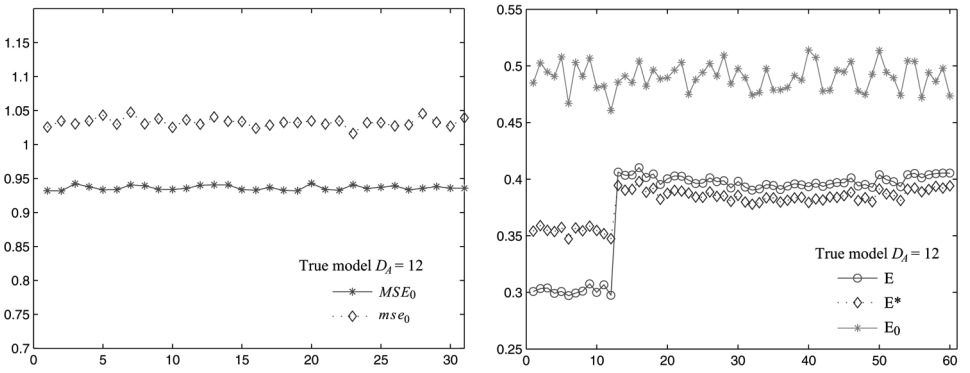

Figure 5. MSE_d^0, mse_d^0 (left) and E_d, E_d^*, E_d^0 (right) values for $D_A = 12$.

model in the domains of the part B and they are greater in the remaining domains. From Figs. 3 (right), 4 (left), and 5 (left) we can state that if the assumed model with $D_A = 12$ is correct, then the MSE of the EBLUP is lower than the one of the direct estimator.

In Fig. 4, we observe that if the EBLUP and its MSE estimator are derived under the true model with $D_A = 12$, then the MSE estimator has a small negative bias as it has not taken into account the variability of $\overline{e}_d$. However, if they are derived under the incorrect model with $D_A = 0$, then a positive bias appears in the domains of the part A and a negative bias in the domains of the part B. The estimation of the mean squared error of the direct estimator presents a positive bias, as it can be seen in Fig. 5 (left). This happens because the randomness of $\overline{y}_d$ comes from the underlying model and not from a simple random sampling extraction mechanism on a fixed population. In Fig. 5 (right) we observe that the estimation of the MSE of the EBLUP under the correct model is the most precise in the domains of the part A. However, in the rest of the domains the estimation of the MSE of the EBLUP derived under the incorrect model with $D_A = 0$ is slightly better. This last fact is simply because mse* has in general lower variance than mse, as it is derived from a model with less parameters.

5. An Application to the Spanish Labour Force Survey

In order to illustrate the use of the introduced model and methodology on small area estimation, we present an application to the estimation of proportion of unemployed people by sex in the Canary Islands. Data sets were elaborated by the Spanish National Statistics Institute (INE) and contain aggregated data from the Canary Islands in the the second trimester of 2003. Statistical sources were the Spanish Labour Force Survey (SLFS) and the Spanish administrative register of unemployment. In the data set there are $D = 50$ records corresponding to 25 areas and $D = 50$ domains (areas crossed with sex). At the unit level, the population of interest contains all the individuals aged 16 or more with legal residence in the Canary Islands during the studied period. At the aggregated level, target variable is the direct estimate of the domain mean of ILO (International Labour Office) unemployed people. Auxiliary variables are the population means ($\overline{\boldsymbol{X}}_d$) of the 6 AGE $*$ WORK categories described in Table 1.

Table 1
Description of the variables in the data file

Variable	Description
AREA	Small territories of Canary Islands: 1–25
SEX	Sex categories: 1 if man, 2 if woman
AGE	Age categories: 1 for 16–24, 2 for 25–54, 3 for $\geq$55
WORK	Registered in the unemployment public office: 1 if yes, 2 if no
DOMAIN (d)	Sex-area categories: 1–50 for $(1,1),\ldots,(1,25),(2,1),\ldots,(2,25)$
UNEMPLOYED (y)	ILO unemployment status: 1 if yes, 0 if no
AGEWORK ($\boldsymbol{x}$)	AGE $*$ WORK categories: 1–6, for $(1,1),(1,2),(2,1),(2,2),(3,1),(3,2)$

Let P_d and s_d denote the domain population and sample, respectively, and let N_d be the population size. Means of variables y and $\boldsymbol{x}$ in domain d are

$$\overline{Y}_d = \frac{1}{N_d} \sum_{j \in P_d} y_{dj}, \quad \overline{\boldsymbol{X}}_d = \frac{1}{N_d} \sum_{j \in P_d} \boldsymbol{x}_{dj},$$

and direct estimates of N_d, $\overline{Y}_d$ and of the design-based variance of $\overline{Y}_d$ are

$$\widehat{N}_d = \sum_{j \in s_d} w_{dj}, \quad \widehat{\overline{Y}}_d^{dir} = \frac{1}{\widehat{N}_d} \sum_{j \in s_d} w_{dj} y_{dj}, \quad \widehat{V}_\pi\left(\widehat{\overline{Y}}_d^{dir}\right) = \frac{1}{\widehat{N}_d^2} \sum_{j \in s_d} w_{dj}(w_{dj} - 1)\left(y_{dj} - \widehat{\overline{Y}}_d^{dir}\right)^2,$$

where the w_{dj} are the calibrated sampling weights of the SLFS. Formula of $\widehat{V}_\pi(\widehat{\overline{Y}}_d^{dir})$ is obtained from Särndal et al. (1992) under the assumptions that sampling weights are the inverses of the first order inclusion probabilities, $w_{dj} = 1/\pi_{dj}$, and that equalities $\pi_{dii} = \pi_{di}$ and $\pi_{dij} = \pi_{di}\pi_{dj}$, if $i \neq j$, hold for the second order inclusion probabilities.

Table 2
Estimated domain means and CV's ($\times 100$) for women

Area	n	dir	eb2	eb00	eb0	cv(dir)	cv(eb2)	cv(eb00)	cv(eb0)
1	1247	0.0777	0.0706	0.0773	0.0774	10.14	13.05	10.06	9.68
2	859	0.0742	0.0677	0.0745	0.0740	12.54	14.49	12.23	11.73
3	152	0.0963	0.0667	0.0848	0.0878	30.03	16.66	28.31	22.45
4	61	0.0175	0.0155	0.0172	0.0153	98.83	69.65	96.40	94.90
5	41	0.0190	0.0399	0.0243	0.0210	98.95	30.32	72.05	73.12
6	18	0.0936	0.0544	0.0688	0.0799	69.00	20.79	49.60	32.48
7	74	0.0362	0.0531	0.0398	0.0405	69.71	21.24	54.60	44.49
8	78	0.0478	0.0541	0.0490	0.0485	56.06	21.95	46.69	38.63
9	76	0.0231	0.0309	0.0237	0.0197	69.83	35.59	64.74	69.68
10	143	0.0530	0.0658	0.0526	0.0546	34.91	20.59	33.43	27.92
11	34	0.0584	0.0722	0.0560	0.0679	68.53	25.98	54.20	37.42
12	156	0.0332	0.0455	0.0366	0.0381	41.86	22.76	36.13	31.62
13	160	0.0570	0.0724	0.0576	0.0565	35.34	18.18	33.02	28.93
14	77	0.0469	0.0391	0.0468	0.0485	57.50	25.88	48.72	38.02
15	132	0.0580	0.0572	0.0560	0.0546	37.83	19.17	34.51	30.09
16	41	0.0738	0.0795	0.0620	0.0910	55.83	26.83	46.35	32.07
17	95	0.1207	0.0879	0.0992	0.0885	27.44	12.78	26.90	23.53
18	111	0.0487	0.0327	0.0487	0.0478	40.14	31.78	37.08	32.63
19	43	0.0430	0.0417	0.0492	0.0471	69.34	23.32	47.82	40.11
20	214	0.1074	0.0715	0.0980	0.0936	20.53	15.78	19.99	17.68
21	20	0.1027	0.0237	0.0490	0.0308	66.90	46.06	67.40	79.37
22	71	0.0993	0.0622	0.0882	0.0828	39.37	15.69	31.07	25.26
23	19	0.0478	0.0610	0.0960	0.0859	97.73	24.78	36.48	32.39
24	79	0.0244	0.0539	0.0333	0.0358	70.32	20.48	47.84	39.48
25	15	0.1684	0.0369	0.0807	0.0761	63.59	34.30	50.09	36.82

By taking $\sigma_d^2 = \widehat{V}_\pi(\widehat{\overline{Y}}_d^{dir})$ we formulate the area-level linear mixed model

$$\widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D = D_A + D_B, \tag{5.1}$$

where $u_1, \ldots, u_{D_A} \sim N(0, \sigma_A^2)$, $u_{D_A+1}, \ldots, u_D \sim N(0, \sigma_B^2)$, $e_1 \sim N(0, \sigma_1^2), \ldots, e_D \sim N(0, \sigma_D^2)$ and they are all mutually independent. We consider the cases $D_A = D/2 = 25$ and $D_A = 0$ to obtain EBLUP estimates labeled eb2 and eb0, respectively. The domains corresponding to women are in part A and the ones corresponding to men are in part B. We also consider the case where data is completely separated by sex in two parts and different standard Fay Herriot models are fitted to each part. The obtained estimators under this last approach are labeled by eb00. Coefficients of variation (CV) are calculated with the formulas

$$\text{cv}(dir) = \frac{\left[\widehat{V}_\pi\left(\widehat{\overline{Y}}_d^{dir}\right)\right]^{1/2}}{\widehat{\overline{Y}}_d^{dir}}, \text{cv}(eb\ell) = \frac{\left[mse\left(\widehat{\overline{Y}}_d^{eb\ell}\right)\right]^{1/2}}{\widehat{\overline{Y}}_d^{eb\ell}}, \quad \ell = 0, 2, 00.$$

Table 3
Estimated domain means and CV's (×100) for men

Area	n	dir	eb2	eb00	eb0	cv(dir)	cv(eb2)	cv(eb00)	cv(eb0)
1	1149	0.0682	0.0696	0.0700	0.0688	11.55	10.82	10.31	10.88
2	726	0.0644	0.0674	0.0636	0.0635	14.78	13.19	13.13	13.95
3	144	0.0164	0.0280	0.0287	0.0202	72.14	36.90	32.99	52.39
4	60	0.0134	0.0065	0.0058	0.0095	99.17	177.46	182.56	123.11
5	41	0.0746	0.0156	0.0117	0.0179	55.50	101.74	132.41	120.84
6	12	0.0000	0.0000	0.0000	0.0039				
7	80	0.0772	0.0488	0.0361	0.0420	40.18	31.79	40.70	46.07
8	85	0.0304	0.0194	0.0176	0.0220	57.63	67.81	64.28	64.23
9	73	0.0250	0.0281	0.0279	0.0257	70.05	45.24	40.08	55.11
10	115	0.0618	0.0659	0.0688	0.0673	37.12	22.10	21.92	25.86
11	35	0.0640	0.0470	0.0482	0.0507	68.17	32.56	33.62	44.06
12	143	0.0800	0.0508	0.0461	0.0568	31.80	28.33	26.70	30.90
13	167	0.0630	0.0675	0.0734	0.0695	29.59	20.49	16.71	21.35
14	99	0.0148	0.0175	0.0179	0.0158	71.07	55.26	53.73	61.70
15	126	0.0819	0.0662	0.0559	0.0638	29.38	22.00	21.65	26.66
16	44	0.0497	0.0727	0.0563	0.0541	68.72	26.50	48.73	49.07
17	86	0.0930	0.0963	0.0855	0.0870	34.45	19.00	20.07	23.72
18	97	0.0204	0.0146	0.0120	0.0156	70.27	80.33	84.64	78.63
19	37	0.1099	0.0569	0.0477	0.0541	47.42	27.98	27.43	40.38
20	193	0.1217	0.0821	0.0648	0.0820	21.13	18.14	19.37	21.50
21	20	0.0434	0.0174	0.0069	0.0096	98.05	95.84	241.65	234.36
22	73	0.0489	0.0644	0.0670	0.0627	56.44	23.33	19.69	29.21
23	19	0.0458	0.0496	0.0484	0.0459	97.93	33.72	46.73	55.04
24	75	0.0877	0.0676	0.0578	0.0636	36.58	22.59	21.95	30.14
25	13	0.0000	0.0000	0.0000	0.0033				

Table 4
Quartiles of CV's ($\times 100$) for women (left) and men (right)

Quartile	cv(dir)	cv(eb2)	cv(eb00)	cv(eb0)	cv(dir)	cv(eb2)	cv(eb00)	cv(eb0)
1	35.34	**18.18**	33.02	27.92	33.12	22.05	**20.86**	26.26
2	56.06	**21.95**	46.35	32.48	55.50	**28.33**	32.99	44.06
3	69.34	**26.83**	50.09	39.48	70.16	**50.25**	51.23	58.41.

Tables 2 and 3 present the estimates of the means (proportions) of unemployed women and men and the corresponding estimates of their CV's (multiplied by 100), in the Canary Islands during the second trimester of 2003. The three EBLUP estimators have in general lower coefficients of variation then the direct estimator, as can be particularly seen in the Table 2 for women. In areas 6 and 25 of Table 3, with sample sizes 12 and 13, respectively, none of the interviewed people was unemployed. This is why their direct estimates are equal to 0. In the Table 4 we present the first, second, and third quartiles of the sets of coefficients of variation for the groups of women and men. We observe that the eb2 estimator is the one presenting the smallest quartiles and thus is the one that we recommend.

6. Conclusions

In this article, an area-level mixed model having random intercepts with variances varying across two groups has been introduced. The scope is to estimate linear parameters of small areas when the set of the areas can be divided in two parts with different variabilities. Algorithms and formulas to fit the model, to calculate EBLUP and to estimate mean squared errors are given. An appealing property of the EBLUP based on the proposed model is that it can be useful to model the behavior of the direct estimator by sex. In the presented simulation experiments, it is shown that if the proposed model is true and the standard linear mixed model is used, then a lack of precision is achieved. An application to real data from the Spanish Labour Force survey shows that the introduced new EBLUP give better results than applying the standard Fay–Herriot model to the whole set of direct estimates or than applying by sex two independent standard Fay–Herriot models.

Acknowledgments

The authors thank one of the referees for their valuable suggestions, including the proposal of introducing model (2.1). This work was supported by the grants MTM2009-09473 and MSMTV 1M0572.

References

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

Fay, R. E., Herriot, R. A. (1979). Estimates of income for small places: an application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* 74(366):269–277.

Ghosh, M., Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statist. Sci.* 9:55–93.

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

Prasad, N. G. N., Rao, J. N. K. (1990). The estimation of the men squared error of small-area estimators. *J. Amer. Statist. Assoc.* 85:163–171.

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

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

Searle, S. R., Casella, G., McCullogh, C. E. (1982). *Variance Components*. New York: John Wiley.