An Application of Linear Model with Both Fixed and Random Effects in Small Area Estimation

TOMÁŠ HOBZA

Institute of Information Theory and Automation, Prague

hobza@fjfi.cvut.cz

DOMINGO MORALES

Operations Research Center, Miguel Hernández University of Elche, Spain

Contents of the talk:

- **Motivation**
- **The proposed model**
- **BLU predictor of mean of a small area**
- **Simulation**
- **Application to the Labour Force Survey**

What is small area estimation?

Sample survey: 1500 - 2000 respondents in Czech Republic ČR

How to get precise estimates eg. for region of Beroun?

Small area: if the domain specific sample is not large enough to support direct estimates of adequate precision

a) geographic area - state, province, county, municipality ...

b) socio-demographic group - specific age-sex-race group, or e.g. unemployed people between 20-30 years etc.

How to increase precision of estimates in small areas?

- to increase the number of respondents - expensive, impossible
- to use SAE - employs a statistical model that "borrows strength" from data collected in other small areas or at other time periods (also use auxiliary data such as administrative data or data from census)

Types of indirect estimators:

- data from different domain but not from another time period - domain indirect
- data from another time period but not from other domain - time indirect
- data from different domain as well as another time period - domain and time indirect

Auxiliary data available:

a) only at the aggregated level for each small area - area level model
b) for the individual units in the population - unit level model

Let's suppose 2 data sets elaborated by INE

- Spanish Labour Force Survey (SLFS) 2003 in the Canary Islands
  $n = 7728$ records
  2 provinces, 34 NUT4 areas
  $D = 46$ domains (areas crossed with sex)
- aggregated data at domain level obtained from administrative registers

Variable	Description
AREA	NUT4 areas: 1-23
PROVINCE	NUT3 areas: 1 for Las Palmas, 2 for Tenerife
RURAL	degree of rurality: 1 if low, 2 if high
SEX	sex categories: 1 if man, 2 if woman
AGE	age categories: 1 for 16-24, 2 for 25-54, 3 for ≥ 55
CLAIM	unemployment claimant: 1 if yes, 2 if no
DOMAIN (d)	sex-area categories: 1-46 for (1,1),...,(1,23),(2,1),...,(2,23)
UNEMPLOYED (y)	unemployment status: 1 if yes, 0 if no
SEXAGECLAIM ($\boldsymbol{x}_1$)	SEX$*$AGE$*$CLAIM categories: 1-12, for (1,1,1), (1,1,2),(1,2,1),...,(2,3,2)
CLUSTER ($\boldsymbol{x}_2$)	PROVINCE$*$RURAL categories: 1-4 for (1,1),(1,2),(2,1),(2,2)
WEIGHT (w)	scaled and calibrated inverses of inclusion probabilities

Table 1. Description of the variables in the Labour Force data file.

If we denote

$$P_d \text{ – domain population}, \qquad s_d \text{ – domain sample}$$

totals of variables y, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in domain d are

$$Y_d = \sum_{j \in P_d} y_{dj}, \quad \boldsymbol{x}_{kd} = \sum_{j \in P_d} \boldsymbol{x}_{kdj}, \quad k = 1, 2,$$

and direct estimate of Y_d and its variance estimator are

$$y_d = \sum_{j \in s_d} w_{dj} y_{dj}, \quad \sigma_d^2 = \sum_{j \in s_d} w_{dj}(w_{dj} - 1) y_{dj}^2.$$

By taking $\boldsymbol{x}_d^t = (\boldsymbol{x}_{1,d}^t, \boldsymbol{x}_{2,d}^t)^t$ we can formulate the area level linear mixed model

$$y_d = \boldsymbol{x}_d^t \boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D$$

where $u_d \sim N(0, \sigma_u^2)$ and $e_d \sim N(0, \sigma_d^2)$ are independent.

By taking $\boldsymbol{x}_d^t = (\boldsymbol{x}_{1,d}^t, \boldsymbol{x}_{2,d}^t)^t$ we can formulate the area level linear mixed model

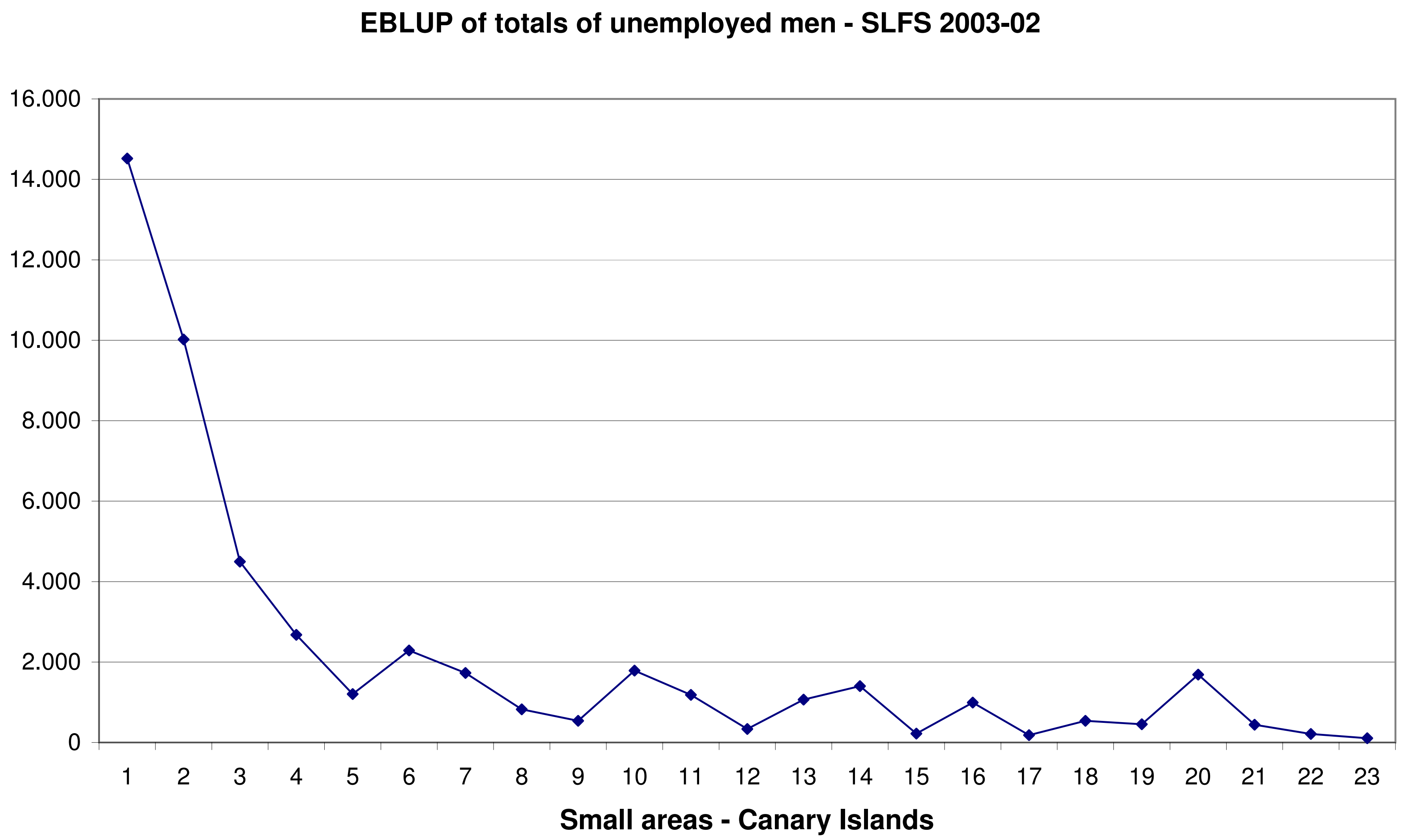
$$y_d = \boldsymbol{x}_d^t \boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D$$

where $u_d \sim N(0, \sigma_u^2)$ and $e_d \sim N(0, \sigma_d^2)$ are independent.

Battesse et al. (1988) – proposed for the first time a nested-error regression model in the setup of SAE

Searle et al. (1982) – provide a detailed description of these models

Ghosh and Rao (1994), Rao (2003) and Jiang and Lahiri (2006) – discuss their applications to SAE


EBLUP of totals of unemployed men - SLFS 2003-02
16.000
14.000
12.000
10.000
8.000
6.000
4.000
2.000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Small areas - Canary Islands

Njuho and Milliken (2005) developed theory for a case where a factor has both fixed and random effect level under a one-way ANOVA model

In this contribution their model is extended to a linear regression model with an intercept being fixed in a part of the domains and being random in the rest of the domains.

Njuho and Milliken (2005) developed theory for a case where a factor has both fixed and random effect level under a one-way ANOVA model

In this contribution their model is extended to a linear regression model with an intercept being fixed in a part of the domains and being random in the rest of the domains.

The supposed model can be written in terms of fixed effect (F) part and random effect (R) part in the following way:

$$(F) \qquad y_{dj} = x_{dj}^t \boldsymbol{\gamma} + \mu_d + e_{dj}, \quad d = 1, \dots, D_F, \ j = 1, \dots, N_d,$$

$$(R) \qquad y_{dj} = x_{dj}^t \boldsymbol{\gamma} + u_d + e_{dj}, \quad d = D_F + 1, \dots, D, \ j = 1, \dots, N_d,$$

Using matrix notation parts (F) and (R) of the model can be written in the form

$$\boldsymbol{y}_F = X_F\boldsymbol{\gamma} + \underset{1\leq d\leq D_F}{\mathrm{diag}}\left(\mathbf{1}_{N_d}\right)\boldsymbol{\mu} + \boldsymbol{e}_F = \left[X_F \;\; \underset{1\leq d\leq D_F}{\mathrm{diag}}\left(\mathbf{1}_{N_d}\right)\right]\begin{pmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\mu}\end{pmatrix} + \boldsymbol{e}_F\,,$$

Using matrix notation parts (F) and (R) of the model can be written in the form

$$\boldsymbol{y}_F = X_F\boldsymbol{\gamma} + \underset{1\leq d\leq D_F}{\operatorname{diag}}(\mathbf{1}_{N_d})\,\boldsymbol{\mu} + \boldsymbol{e}_F = \left[X_F \;\; \underset{1\leq d\leq D_F}{\operatorname{diag}}(\mathbf{1}_{N_d})\right]\begin{pmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\mu}\end{pmatrix} + \boldsymbol{e}_F\,,$$

$$\boldsymbol{y}_R = X_R\boldsymbol{\gamma} + \underset{D_F+1\leq d\leq D}{\operatorname{diag}}(\mathbf{1}_{N_d})\,\boldsymbol{u} + \boldsymbol{e}_R = \left[X_R \;\; \underset{D_F+1\leq d\leq D}{\operatorname{diag}}(\mathbf{1}_{N_d})\right]\begin{bmatrix}\boldsymbol{\gamma}\\ \boldsymbol{u}\end{bmatrix} + \boldsymbol{e}_R\,.$$

So that we can express the complete model in the form

$$\begin{pmatrix} \boldsymbol{y}_F \\ \boldsymbol{y}_R \end{pmatrix} = \begin{bmatrix} X_F & \operatorname*{diag}_{1\leq d\leq D_F}(\mathbf{1}_{N_d}) \\ X_R & \mathbf{0}_{N_R\times D_F} \end{bmatrix} \begin{pmatrix} \gamma \\ \boldsymbol{\mu} \end{pmatrix} + \begin{bmatrix} \mathbf{0}_{N_F\times D_R} \\ \operatorname*{diag}_{D_F+1\leq d\leq D}(\mathbf{1}_{N_d}) \end{bmatrix} \boldsymbol{u} + \begin{pmatrix} \boldsymbol{e}_F \\ \boldsymbol{e}_R \end{pmatrix}$$

So that we can express the complete model in the form

$$\begin{pmatrix} \boldsymbol{y}_F \\ \boldsymbol{y}_R \end{pmatrix} = \begin{bmatrix} X_F & \operatorname*{diag}_{1\le d\le D_F} (\mathbf{1}_{N_d}) \\ X_R & \mathbf{0}_{N_R\times D_F} \end{bmatrix} \begin{pmatrix} \gamma \\ \boldsymbol{\mu} \end{pmatrix} + \begin{bmatrix} \mathbf{0}_{N_F\times D_R} \\ \operatorname*{diag}_{D_F+1\le d\le D} (\mathbf{1}_{N_d}) \end{bmatrix} \boldsymbol{u} + \begin{pmatrix} \boldsymbol{e}_F \\ \boldsymbol{e}_R \end{pmatrix}$$

or more simply

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

where $\boldsymbol{y} = \boldsymbol{y}_{N\times 1}$, $\boldsymbol{X} = \boldsymbol{X}_{N\times(p+D_F)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{(p+D_F)\times 1}$,
$\boldsymbol{Z} = \boldsymbol{Z}_{N\times D_R}$, $\boldsymbol{u} = \boldsymbol{u}_{D_R\times 1}$ and $\boldsymbol{e} = \boldsymbol{e}_{N\times 1}$ with $N = \sum_{d=1}^{D} N_d$.

Assumptions:

- $\text{rank}(\boldsymbol{X}) = p + D_F$
- $\boldsymbol{u} = \boldsymbol{u}_{D_R \times 1} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_u)$ and $\boldsymbol{e} = \boldsymbol{e}_{N \times 1} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_e)$ are independent
- $\boldsymbol{\Sigma}_u = \sigma_u^2 \boldsymbol{I}_{D_R}$
- $$\boldsymbol{\Sigma}_e = \text{diag}\left[\sigma_{e_F}^2 \operatorname*{diag}_{1 \le d \le D_F}(\boldsymbol{W}_d^{-1}),\ \sigma_{e_R}^2 \operatorname*{diag}_{D_F+1 \le d \le D}(\boldsymbol{W}_d^{-1})\right]$$

  and $\boldsymbol{W}_d = \text{diag}(w_{d1}, \ldots, w_{dN_d})_{N_d \times N_d}, d = 1, \ldots, D$, is the corresponding part of the matrix

  $$\boldsymbol{W}_N = \text{diag}(w_{11}, \ldots, w_{D,N_D})_{N \times N},$$

  $w_{11} > 0, \ldots, w_{D,N_D} > 0$ known.

Assumptions:

- $\text{rank}(\boldsymbol{X}) = p + D_F$
- $\boldsymbol{u} = \boldsymbol{u}_{D_R \times 1} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_u)$ and $\boldsymbol{e} = \boldsymbol{e}_{N \times 1} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_e)$ are independent
- $\boldsymbol{\Sigma}_u = \sigma_u^2 \boldsymbol{I}_{D_R}$
- $$\boldsymbol{\Sigma}_e = \text{diag}\left[\sigma_{e_F}^2 \operatorname*{diag}_{1 \leq d \leq D_F}(\boldsymbol{W}_d^{-1}), \sigma_{e_R}^2 \operatorname*{diag}_{D_F+1 \leq d \leq D}(\boldsymbol{W}_d^{-1})\right]$$

  and $\boldsymbol{W}_d = \text{diag}(w_{d1}, \ldots, w_{dN_d})_{N_d \times N_d}, d = 1, \ldots, D$, is the corresponding part of the matrix

  $$\boldsymbol{W}_N = \text{diag}(w_{11}, \ldots, w_{D,N_D})_{N \times N},$$

  $w_{11} > 0, \ldots, w_{D,N_D} > 0$ known.

Thus

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}) \quad \text{with} \quad \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\Sigma}_u\boldsymbol{Z}^t + \boldsymbol{\Sigma}_e = diag(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_D).$$

When $\sigma_{e_F}^2 > 0$, $\sigma_{e_R}^2 > 0$ and $\sigma_u^2 > 0$ are known,

the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p+D_F})^t$ is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{y}$$

and the best linear unbiased predictor (BLUP) of $\boldsymbol{u} = (u_1, \ldots, u_{D_R})^t$ is

$$\widehat{\boldsymbol{u}} = \boldsymbol{\Sigma}_u \boldsymbol{Z}^t \boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{X} \widehat{\boldsymbol{\beta}} \right).$$

When $\sigma^2_{e_F} > 0$, $\sigma^2_{e_R} > 0$ and $\sigma^2_u > 0$ are known,

the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p+D_F})^t$ is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{y}$$

and the best linear unbiased predictor (BLUP) of $\boldsymbol{u} = (u_1, \ldots, u_{D_R})^t$ is

$$\widehat{\boldsymbol{u}} = \boldsymbol{\Sigma}_u \boldsymbol{Z}^t \boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{X} \widehat{\boldsymbol{\beta}} \right).$$

The parametric space of the model is

$$\Theta = \{\boldsymbol{\theta}^t = (\boldsymbol{\beta}^t, \sigma^2_u, \sigma^2_{e_F}, \sigma^2_{e_R}); \boldsymbol{\beta} \in R^{p+D_F}, \sigma^2_u \geq 0, \sigma^2_{e_F} > 0, \sigma^2_{e_R} > 0\}$$

and MLE of the unknown parameters can be found e.g. by the Fisher-Scoring algorithm.

Now let's consider a finite population of $N = N_F + N_R$ elements following the introduced model.

From the population a sample of size n with n_d elements in area d, $n = \sum_{d=1}^{D} n_d$, is selected.

We can reorder the population so that

$$\boldsymbol{y} = (\boldsymbol{y}_s^t, \boldsymbol{y}_r^t)^t,$$

where

$\boldsymbol{y}_s$ – vector of n observed elements

and

$\boldsymbol{y}_r$ – vector of $N - n$ unobserved elements.

In this notation we can write

$$E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{V} = V[\boldsymbol{y}] = \begin{pmatrix} \boldsymbol{V}_{ss} & \boldsymbol{V}_{sr} \\ \boldsymbol{V}_{rs} & \boldsymbol{V}_{rr} \end{pmatrix}.$$

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

$$\overline{Y}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj} = \boldsymbol{a}^t\boldsymbol{y},$$

where $\boldsymbol{a}^t = \frac{1}{N_d}\left(\mathbf{0}^t_{N_1},\ldots,\mathbf{0}^t_{N_{d-1}},\mathbf{1}^t_{N_d},\mathbf{0}^t_{N_{d+1}},\ldots,\mathbf{0}^t_{N_D}\right)$ and $\mathbf{0}^t_m = (0,\ldots,0)_{1\times m}$.

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

$$\overline{Y}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj} = \boldsymbol{a}^t\boldsymbol{y},$$

where $\boldsymbol{a}^t = \frac{1}{N_d}\left(\mathbf{0}^t_{N_1},\ldots,\mathbf{0}^t_{N_{d-1}},\mathbf{1}^t_{N_d},\mathbf{0}^t_{N_d+1},\ldots,\mathbf{0}^t_{N_D}\right)$ and $\mathbf{0}^t_m = (0,\ldots,0)_{1\times m}$.

From the general theorem of prediction it follows

$$\widehat{\overline{Y}}_d^{blup} = \boldsymbol{a}^t_s\boldsymbol{y}_s + \boldsymbol{a}^t_r\left[\boldsymbol{X}_r\widehat{\boldsymbol{\beta}} + \boldsymbol{V}_{rs}\boldsymbol{V}^{-1}_{ss}(\boldsymbol{y}_s - \boldsymbol{X}_s\widehat{\boldsymbol{\beta}})\right],$$

where

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t_s\boldsymbol{V}^{-1}_{ss}\boldsymbol{X}_s)^{-1}\boldsymbol{X}^t_s\boldsymbol{V}^{-1}_{ss}\boldsymbol{y}_s$$

.

In our case

$$\widehat{\overline{Y}}_d^{blup} = \overline{\boldsymbol{X}}_d \widehat{\boldsymbol{\beta}} + f_d \left(\widehat{\overline{Y}}_d - \widehat{\overline{\boldsymbol{X}}}_d \widehat{\boldsymbol{\beta}} \right)$$

for $1 \leq d \leq D_F$

In our case

$$\widehat{\overline{Y}}_d^{blup} = \overline{\boldsymbol{X}}_d \widehat{\boldsymbol{\beta}} + f_d \left(\widehat{\overline{Y}}_d - \widehat{\overline{\boldsymbol{X}}}_d \widehat{\boldsymbol{\beta}} \right)$$

for $1 \leq d \leq D_F$ and

$$\widehat{\overline{Y}}_d^{blup} = (1 - f_d) \left[\overline{\boldsymbol{X}}_d \widehat{\boldsymbol{\beta}} + \gamma_d^w \left(\widehat{\overline{Y}}_d^{direct} - \widehat{\overline{\boldsymbol{X}}}_d^{direct} \widehat{\boldsymbol{\beta}} \right) \right] + f_d \left[\widehat{\overline{Y}}_d + (\overline{\boldsymbol{X}}_d - \widehat{\overline{\boldsymbol{X}}}_d) \widehat{\boldsymbol{\beta}} \right]$$

for $D_F + 1 \leq d \leq D$,

BLU predictor of mean of a small area

In our case

$$\widehat{\overline{Y}}_d^{blup} = \overline{\boldsymbol{X}}_d \widehat{\boldsymbol{\beta}} + f_d \left(\widehat{\overline{Y}}_d - \widehat{\overline{\boldsymbol{X}}}_d \widehat{\boldsymbol{\beta}} \right)$$

for $1 \leq d \leq D_F$ and

$$\widehat{\overline{Y}}_d^{blup} = (1-f_d) \left[\overline{\boldsymbol{X}}_d \widehat{\boldsymbol{\beta}} + \gamma_d^w \left(\widehat{\overline{Y}}_d^{direct} - \widehat{\overline{\boldsymbol{X}}}_d^{direct} \widehat{\boldsymbol{\beta}} \right) \right] + f_d \left[\widehat{\overline{Y}}_d + (\overline{\boldsymbol{X}}_d - \widehat{\overline{\boldsymbol{X}}}_d) \widehat{\boldsymbol{\beta}} \right]$$

for $D_F + 1 \leq d \leq D$,

where $\widehat{\overline{Y}}_d^{direct} = \frac{1}{w_d} \sum_{j=1}^{n_d} w_{dj} y_{dj}$, $\quad \widehat{\overline{\boldsymbol{X}}}_d^{direct} = \frac{1}{w_d} \sum_{j=1}^{n_d} w_{dj} \boldsymbol{x}_{dj}^t$,

$\gamma_d^w = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_{e_R}^2}{w_d}}$, $\quad \overline{\boldsymbol{X}}_d = 1/N_d \sum_{j=1}^{N_d} \boldsymbol{x}_{dj}^t$, $\quad \widehat{\overline{\boldsymbol{X}}}_d = 1/n_d \sum_{j=1}^{n_d} \boldsymbol{x}_{dj}^t$,

$\widehat{\overline{Y}}_d = 1/n_d \sum_{j=1}^{n_d} y_{dj}$ and $f_d = n_d/N_d$.

BLU predictor of mean of a small area

In our case

$$\widehat{\overline{Y}}_d^{blup} = \overline{\boldsymbol{X}}_d\widehat{\boldsymbol{\beta}} + f_d\left(\widehat{\overline{Y}}_d - \widehat{\overline{\boldsymbol{X}}}_d\widehat{\boldsymbol{\beta}}\right)$$

for $1 \leq d \leq D_F$ and

$$\widehat{\overline{Y}}_d^{blup} = (1-f_d)\left[\overline{\boldsymbol{X}}_d\widehat{\boldsymbol{\beta}} + \gamma_d^w\left(\widehat{\overline{Y}}_d^{direct} - \widehat{\overline{\boldsymbol{X}}}_d^{direct}\widehat{\boldsymbol{\beta}}\right)\right] + f_d\left[\widehat{\overline{Y}}_d + (\overline{\boldsymbol{X}}_d - \widehat{\overline{\boldsymbol{X}}}_d)\widehat{\boldsymbol{\beta}}\right]$$

for $D_F + 1 \leq d \leq D$,

where $\widehat{\overline{Y}}_d^{direct} = \frac{1}{w_d}\sum_{j=1}^{n_d} w_{dj}y_{dj}$, $\quad \widehat{\overline{\boldsymbol{X}}}_d^{direct} = \frac{1}{w_d}\sum_{j=1}^{n_d} w_{dj}\boldsymbol{x}_{dj}^t$,

$\gamma_d^w = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_{eR}^2}{w_d}}$, $\quad \overline{\boldsymbol{X}}_d = 1/N_d\sum_{j=1}^{N_d}\boldsymbol{x}_{dj}^t$, $\quad \widehat{\overline{\boldsymbol{X}}}_d = 1/n_d\sum_{j=1}^{n_d}\boldsymbol{x}_{dj}^t$,

$\widehat{\overline{Y}}_d = 1/n_d\sum_{j=1}^{n_d} y_{dj}$ and $f_d = n_d/N_d$.

Estimator $\widehat{\overline{Y}}_d^{eblup}$ of $\overline{Y}_d$ is obtained by substituting variance components by their MLE's

The mean squared error of $\widehat{\overline{Y}}_d^{eblup}$ is estimated by using the following formula

$$mse(\widehat{\overline{Y}}_d^{eblup}) = g_{1d}(\widehat{\boldsymbol{\sigma}}) + g_{2d}(\widehat{\boldsymbol{\sigma}}) + 2g_{3d}(\widehat{\boldsymbol{\sigma}}) + g_{4d}(\widehat{\boldsymbol{\sigma}}) - g_{d5}(\widehat{\boldsymbol{\sigma}}).$$

Prasad and Rao (1990), Das, Jiang and Rao (2001)

$$g_{1d}(\boldsymbol{\sigma}) = \begin{cases} 0 & \text{if } 1 \le d \le D_F, \\ (1-f_d)^2(1-\gamma_d^w)\sigma_u^2 & \text{if } D_F+1 \le d \le D, \end{cases}$$

$$g_{2d}(\boldsymbol{\sigma}) = \begin{cases} (1-f_d)^2 \overline{\boldsymbol{X}}_d^* \boldsymbol{P}_s \overline{\boldsymbol{X}}_d^{*t} & \text{if } 1 \le d \le D_F, \\ (1-f_d)^2 \left(\overline{\boldsymbol{X}}_d^* - \gamma_d^w \widehat{\overline{\boldsymbol{X}}}_d^{direct}\right) \boldsymbol{P}_s \left(\overline{\boldsymbol{X}}_d^* - \gamma_d^w \widehat{\overline{\boldsymbol{X}}}_d^{direct}\right)^t & \text{if } D_F+1 \le d \le D \end{cases}$$

$$g_{3d}(\boldsymbol{\sigma}) = 0 \quad \text{if } 1 \le d \le D_F; \text{ otherwise}$$

$$g_{3d}(\boldsymbol{\sigma}) = (1-f_d)^2 \left(\sigma_u^2 + \frac{\sigma_{eR}^2}{w_d}\right)^{-3} \frac{1}{w_d^2} \left\{\sigma_{eR}^4 \mathrm{V}(\hat{\sigma}_u^2)\right\} - 2\sigma_u^2\sigma_{eR}^2 \mathrm{cov}(\hat{\sigma}_u^2, \hat{\sigma}_{eR}^2) + \sigma_u^4 \mathrm{V}(\hat{\sigma}_{eR}^2),$$

$$g_{4d}(\boldsymbol{\sigma}) = \begin{cases} \frac{\sigma_{eF}^2(\mathcal{V}_d - \nu_d)}{N_d^2} & \text{if } 1 \le d \le D_F, \\ \frac{\sigma_{eR}^2(\mathcal{V}_d - \nu_d)}{N_d^2} & \text{if } D_F+1 \le d \le D, \end{cases}$$

where $\mathcal{V}_d = \sum_{j=1}^{N_d} w_{dj}^{-1}$, $\nu_d = \sum_{j=1}^{n_d} w_{dj}^{-1}$.

The true model is a model with fixed effects

We consider the proposed model with

$$D = 30 \quad \text{small areas,} \qquad D_F = 3 \quad \text{small areas with fixed effect,}$$

$$N_d = 100, \quad 1 \leq d \leq D, \text{ totals of units in each area .}$$

The true model is a model with fixed effects

We consider the proposed model with

$$D = 30 \quad \text{small areas,} \qquad D_F = 3 \quad \text{small areas with fixed effect,}$$

$$N_d = 100, \quad 1 \leq d \leq D, \text{ totals of units in each area .}$$

Algorithm:

1. Population generation

- Matrices $\boldsymbol{X}_F, \boldsymbol{X}_R$. Set

$$a_d = 1, b_d = d + D_R + 1 \quad \text{for} \quad d = 1, \ldots, D_F$$

$$a_d = 1, b_d = d - D_F + 1 \quad \text{for} \quad d = D_F + 1, \ldots, D$$

  and for $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, do

$$x_{dj} = (b_d - a_d)\frac{j}{1 + n_d} + a_d.$$

- Weights. Do $w_{dj} = x_{dj}^{-\ell}$ with $\ell = 1/2$ for all d, j.

- Weights. Do $w_{dj} = x_{dj}^{-\ell}$ with $\ell = 1/2$ for all d, j.
- Target variable $\boldsymbol{y}$. For $d = 1, \dots, D_F$, $j = 1, \dots, n_d$, take

$$\gamma = 1, \quad \mu_d = 12 + d \quad \text{and} \quad \sigma_{e_F}^2 = 2$$

and generate

$$y_{dj} = x_{dj}\gamma + \mu_d + w_{dj}^{-1/2} e_{dj}, \quad \text{where } e_{dj} \sim \mathcal{N}(0, \sigma_{e_F}^2).$$

- Weights. Do $w_{dj} = x_{dj}^{-\ell}$ with $\ell = 1/2$ for all d, j.
- Target variable $\boldsymbol{y}$. For $d = 1, \ldots, D_F$, $j = 1, \ldots, n_d$, take

$$\gamma = 1, \quad \mu_d = 12 + d \quad \text{and} \quad \sigma^2_{e_F} = 2$$

and generate

$$y_{dj} = x_{dj}\gamma + \mu_d + w_{dj}^{-1/2} e_{dj}, \quad \text{where } e_{dj} \sim \mathcal{N}(0, \sigma^2_{e_F}).$$

For $d = D_F + 1, \ldots, D$, $j = 1, \ldots, n_d$, take

$$\gamma = 1, \quad \sigma^2_u = 1 \quad \text{and} \quad \sigma^2_{e_R} = 1$$

and generate

$$y_{dj} = x_{dj}\gamma + u_d + w_{dj}^{-1/2} e_{dj}, \quad \text{where } u_d \sim \mathcal{N}(0, \sigma^2_u), \;\; e_{dj} \sim \mathcal{N}(0, \sigma^2_{e_R}).$$

2. Sample extraction

From each small area we extract a sample of size n_d, where

$$n_d = \begin{cases} c \cdot q & \text{for } 1 \leq d \leq D_F, \\ q & \text{for } D_F + 1 \leq d \leq D \end{cases} \quad \text{and} \quad c = 2, q = 5\,.$$

2. Sample extraction

From each small area we extract a sample of size n_d, where

$$n_d = \begin{cases} c \cdot q & \text{for } 1 \leq d \leq D_F, \\ q & \text{for } D_F + 1 \leq d \leq D \end{cases} \quad \text{and} \quad c = 2, q = 5\,.$$

3. Parameter estimation and prediction

From the simulated population we calculate

- the population mean of each area d:

$$\overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}$$

.

From the extracted sample we calculate

- the MLEs $\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_u^2, \widehat{\sigma}_e^2$
- the EBLUP $\widehat{\overline{Y}}_d^{eblup}$ of the mean of each area d
- The MSE estimator $mse_d(\widehat{\overline{Y}}_d^{eblup})$

From the extracted sample we calculate

- the MLEs $\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_u^2, \widehat{\sigma}_e^2$
- the EBLUP $\widehat{\overline{Y}}_d^{eblup}$ of the mean of each area d
- The MSE estimator $mse_d(\widehat{\overline{Y}}_d^{eblup})$

Under the assumption $D_F = 0$

- the MLEs $\widehat{\boldsymbol{\beta}}^*, \widehat{\sigma}_u^{2*}, \widehat{\sigma}_e^{2*}$
- the EBLUP $\widehat{\overline{Y}}_d^{eblup*}$
- The MSE estimator $mse(\widehat{\overline{Y}}_d^{eblup*})$

4. Repetition and performance measures

Steps 1-3 are repeated $K = 10000$ times obtaining thus in each iteration

$$\overline{Y}_d^{(k)}, \quad \widehat{\overline{Y}}_d^{eblup(k)} \quad \text{and} \quad \widehat{\overline{Y}}_d^{eblup*(k)}.$$

4. Repetition and performance measures

Steps 1-3 are repeated $K = 10000$ times obtaining thus in each iteration

$$\overline{Y}_d^{(k)}, \quad \widehat{\overline{Y}}_d^{eblup(k)} \quad \text{and} \quad \widehat{\overline{Y}}_d^{eblup*(k)}.$$

Calculated performance measures:

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

4. Repetition and performance measures

Steps 1-3 are repeated $K = 10000$ times obtaining thus in each iteration

$$\overline{Y}_d^{(k)}, \quad \widehat{\overline{Y}}_d^{eblup(k)} \quad \text{and} \quad \widehat{\overline{Y}}_d^{eblup*(k)}.$$

Calculated performance measures:

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

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

4. Repetition and performance measures

Steps 1-3 are repeated $K = 10000$ times obtaining thus in each iteration

$$\overline{Y}_d^{(k)}, \quad \widehat{\overline{Y}}_d^{eblup(k)} \quad \text{and} \quad \widehat{\overline{Y}}_d^{eblup*(k)}.$$

Calculated performance measures:

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

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

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

4. Repetition and performance measures

Steps 1-3 are repeated $K = 10000$ times obtaining thus in each iteration

$$\overline{Y}_d^{(k)}, \quad \widehat{\overline{Y}}_d^{eblup(k)} \quad \text{and} \quad \widehat{\overline{Y}}_d^{eblup*(k)}.$$

Calculated performance measures:

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

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

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

$$mse_d = \frac{1}{K}\sum_{k=1}^{K} mse(\widehat{\overline{Y}}_d^{eblup(k)}), \quad mse_d^* = \frac{1}{K}\sum_{k=1}^{K} mse(\widehat{\overline{Y}}_d^{eblup*(k)}).$$

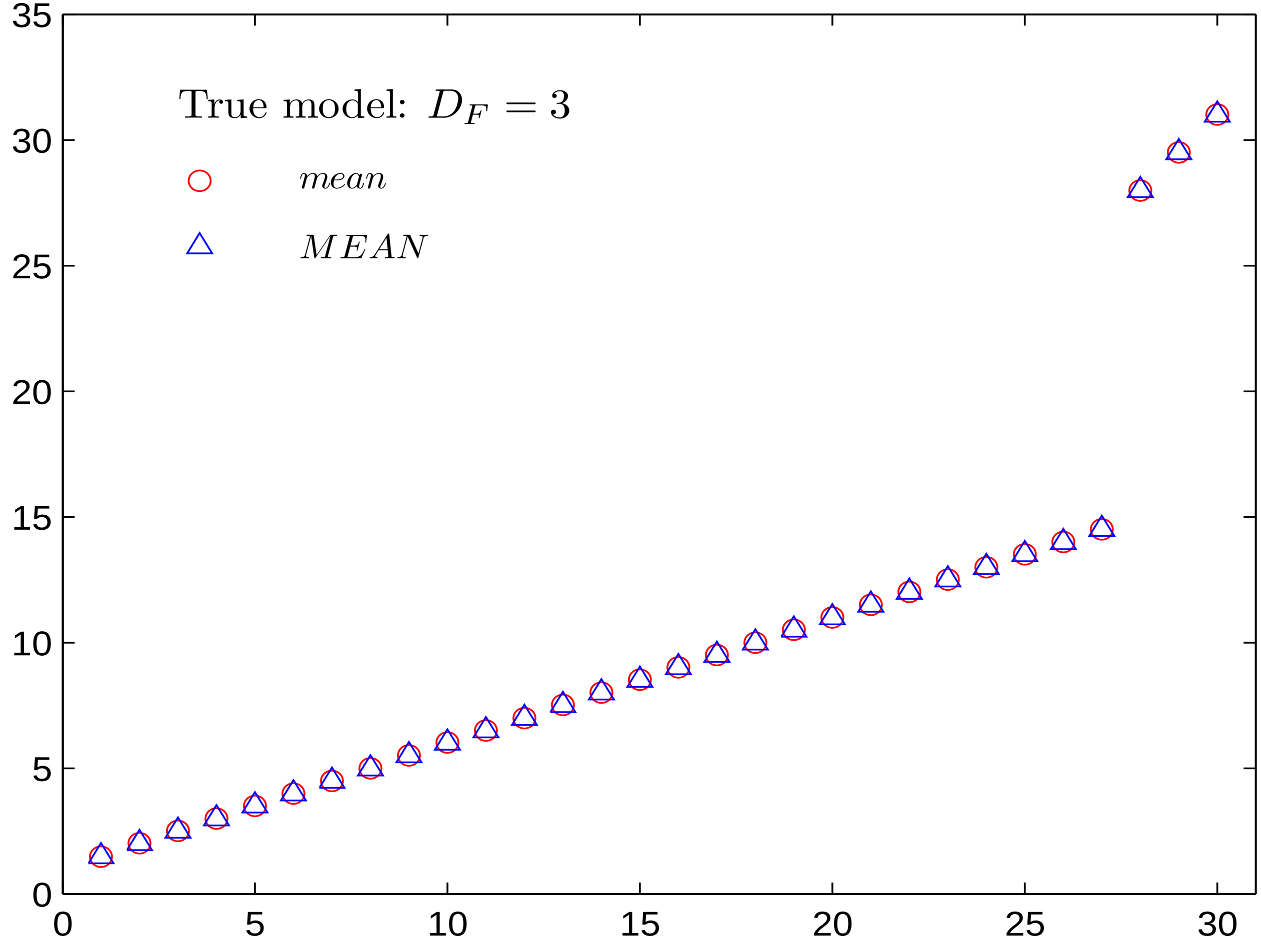


Figure 2. $MEAN_d$ and $mean_d$ values for $\boldsymbol{\mu} = (13, 14, 15)$ and $D_F = 3$.

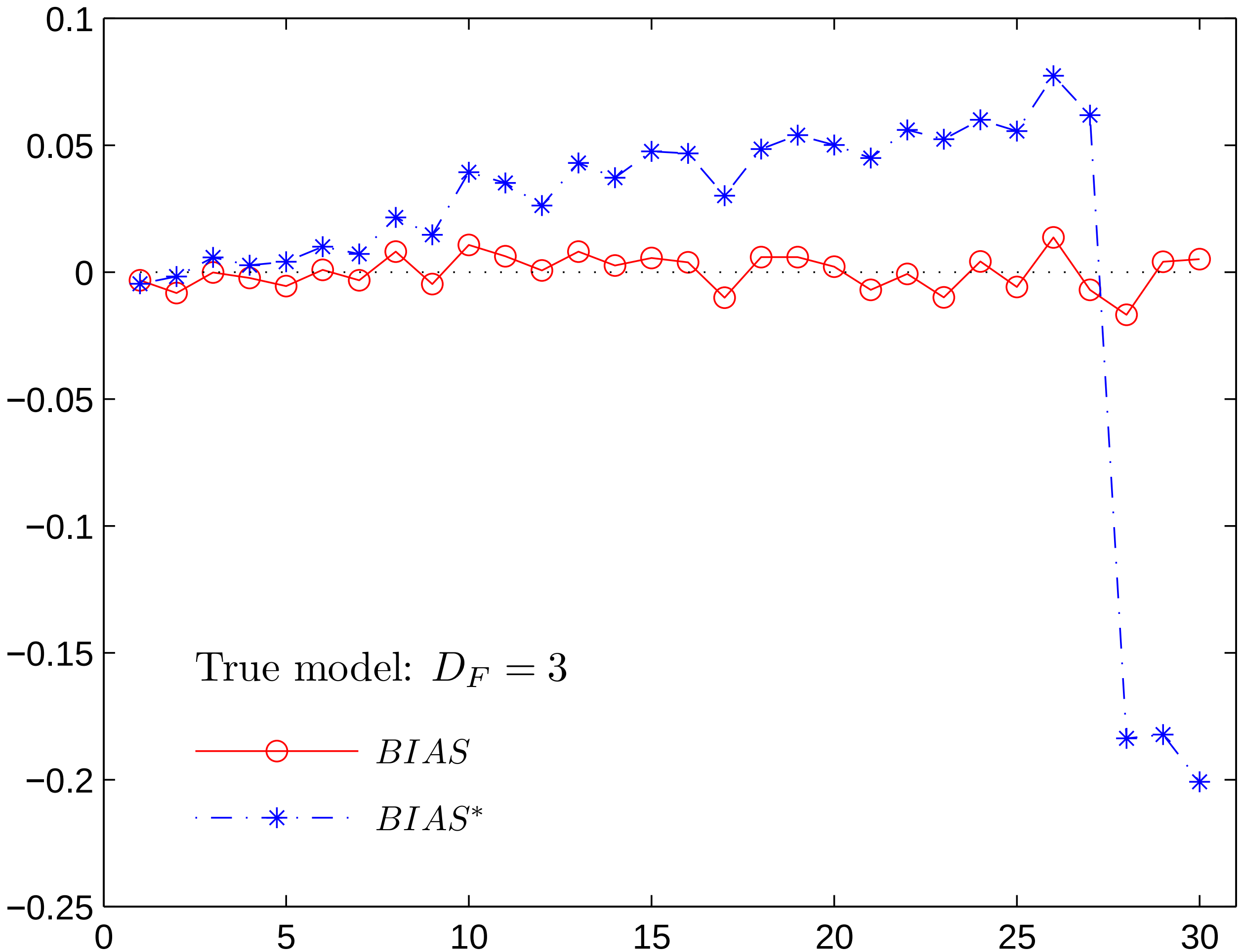


Figure 3. $BIAS_d$ and $BIAS_d^*$ values for $\boldsymbol{\mu} = (13, 14, 15)$ and $D_F = 3$.

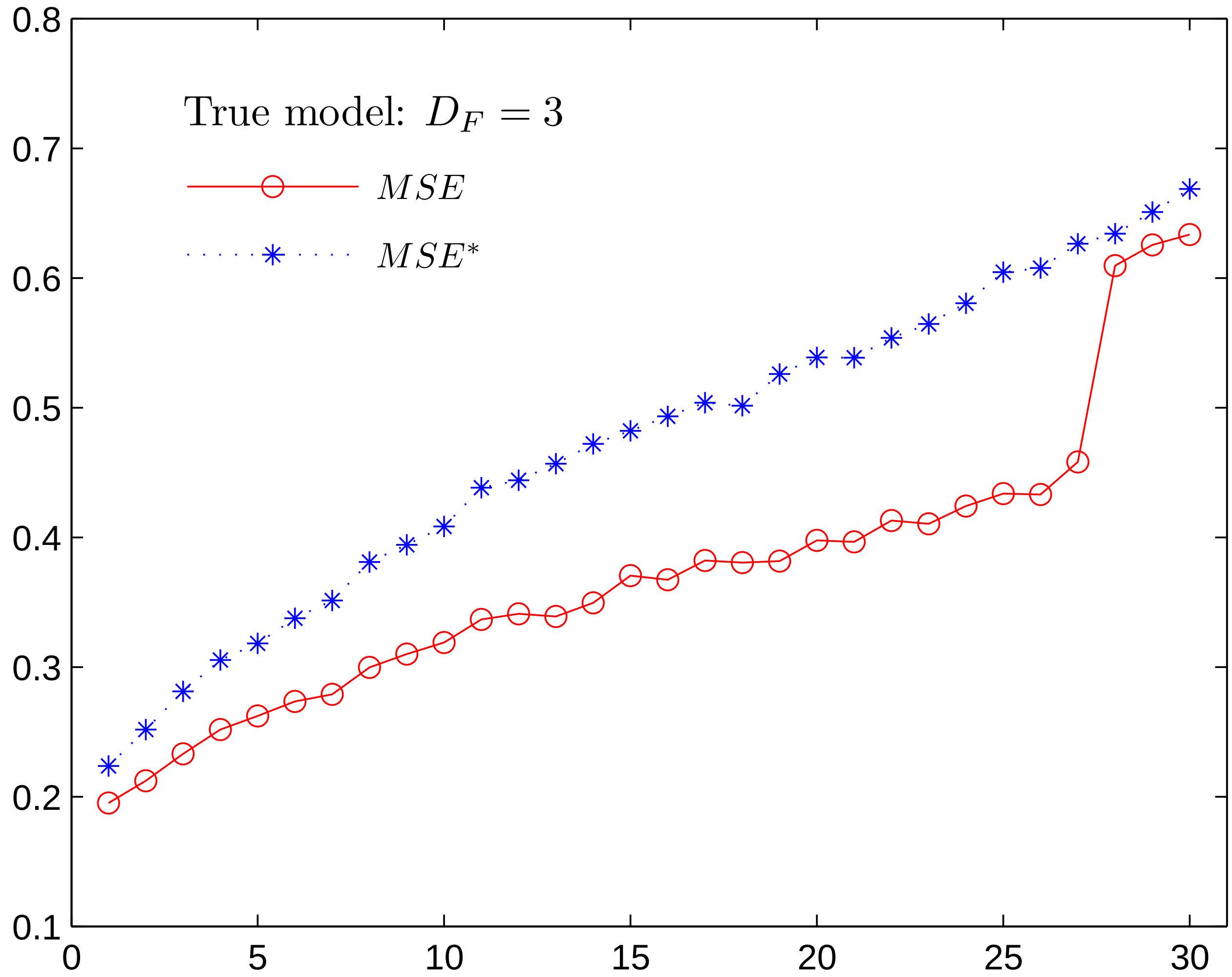


Figure 4. MSE_d and MSE_d^* (right) values for $\boldsymbol{\mu} = (13, 14, 15)$ and $D_F = 3$.

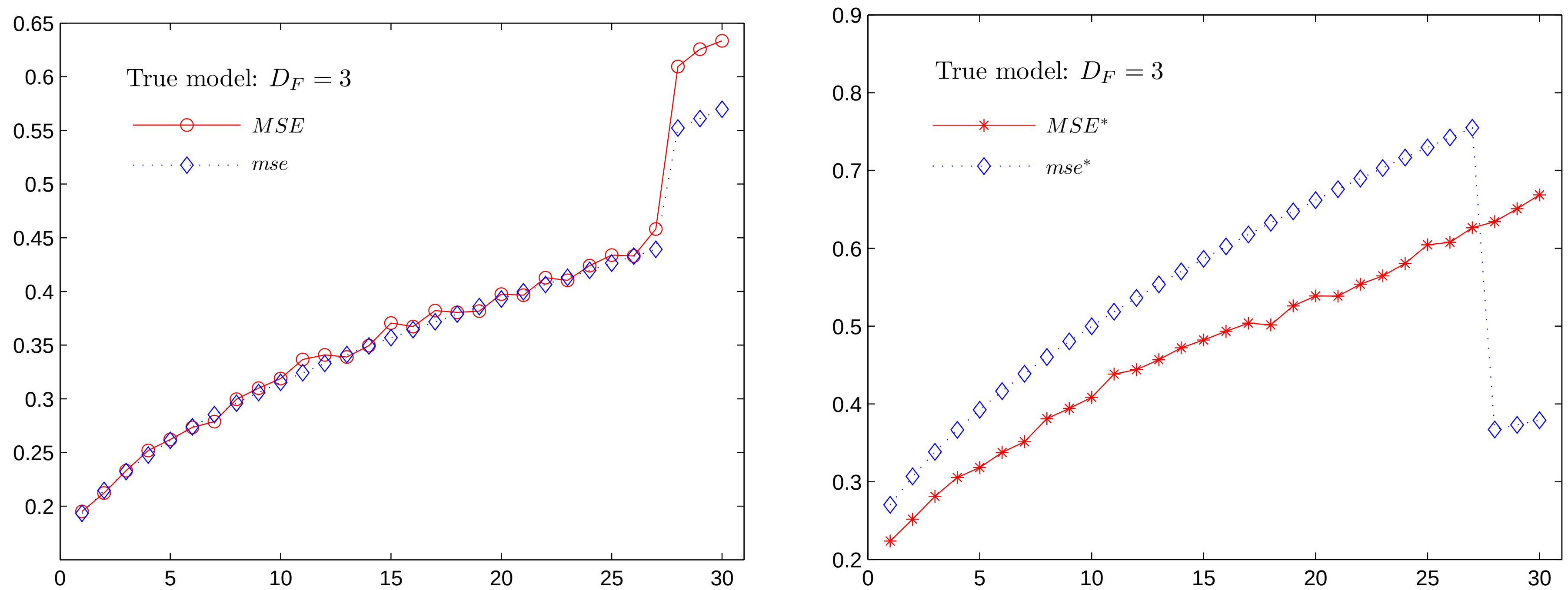


Figure 5. MSE_d, mse_d (left) and MSE^*_d, mse^*_d (right) values for $\boldsymbol{\mu} = (13, 14, 15)$ and $D_F = 3$.

The true model is a model without fixed effects

We repeat the simulation for the case that the population is generated from the model without fixed effects, i.e. with $D_F = 0$.

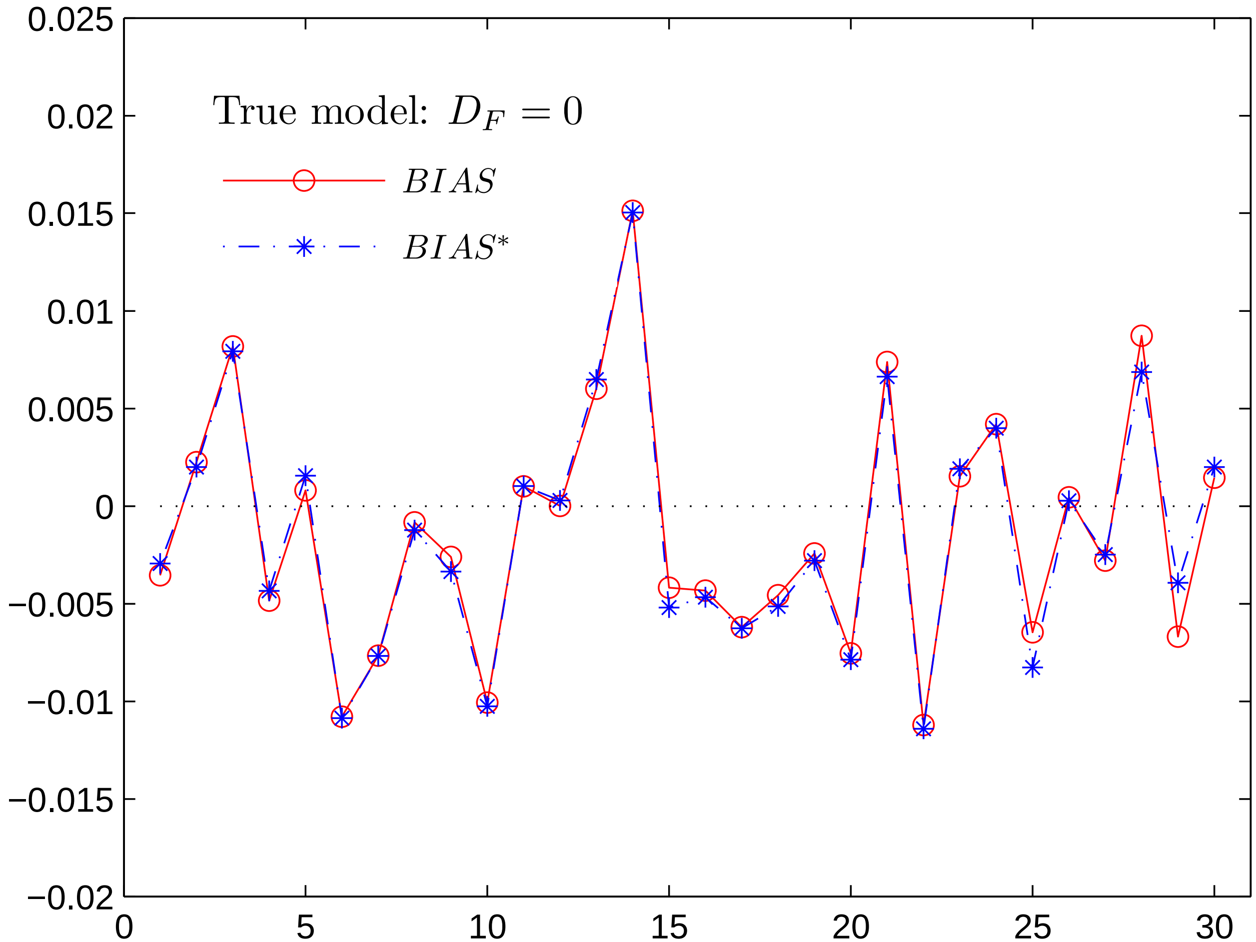


Figure 6. $BIAS_d$ and $BIAS_d^*$ values for $\boldsymbol{\mu} = (3, 4, 5)$ and $D_F = 0$.

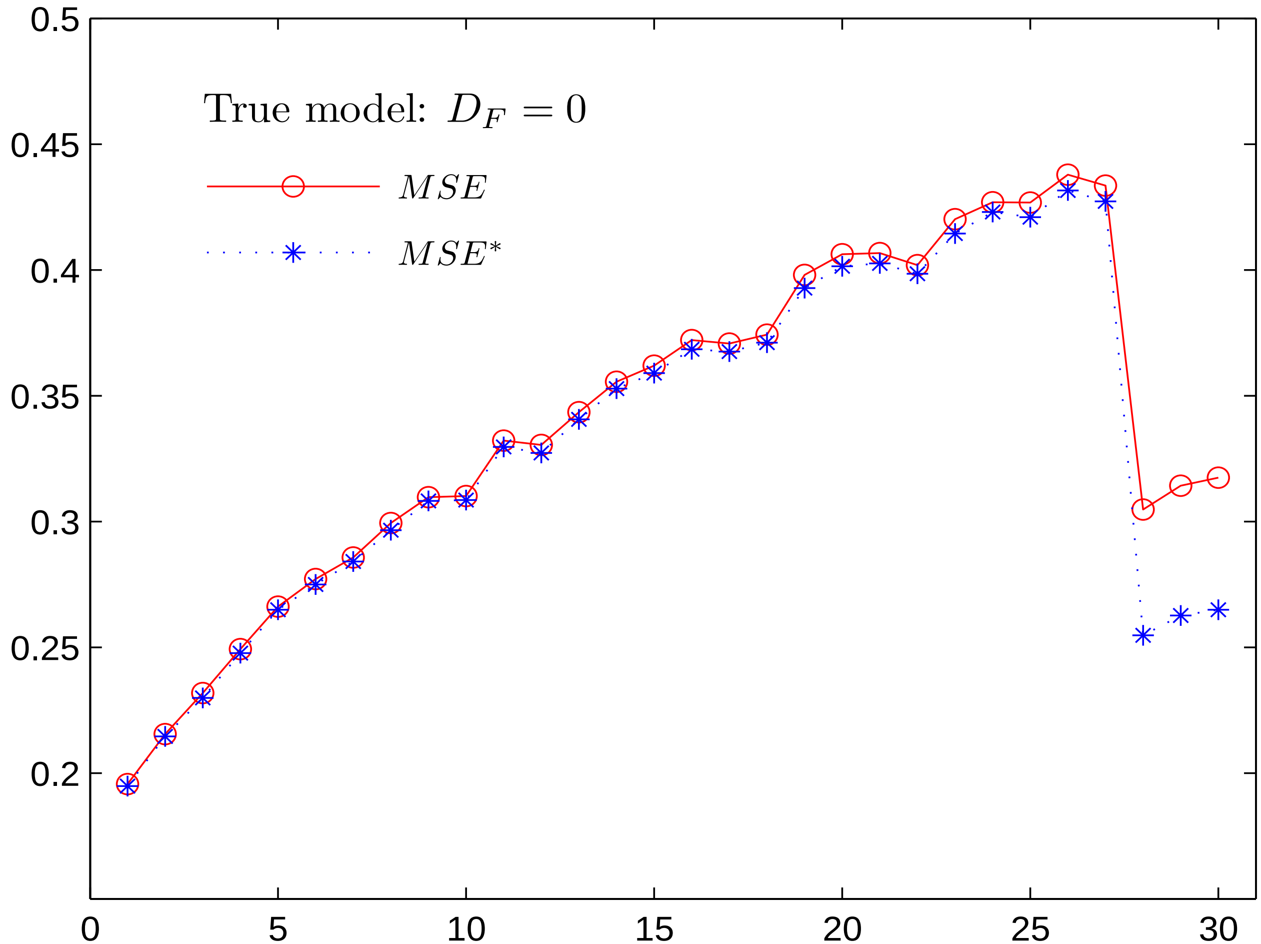
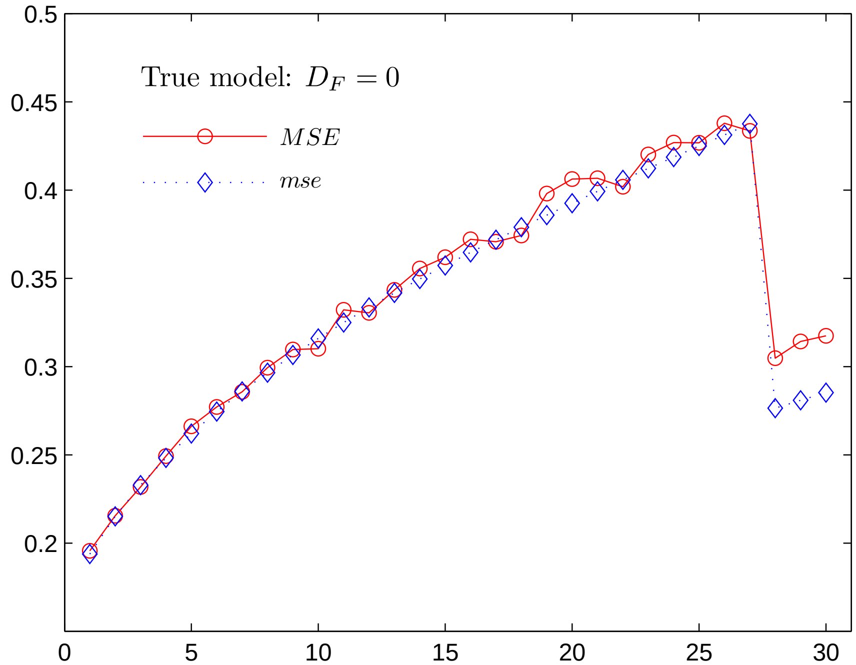


Figure 7. MSE_d and MSE_d^* (right) values for $\boldsymbol{\mu} = (3, 4, 5)$ and $D_F = 0$.



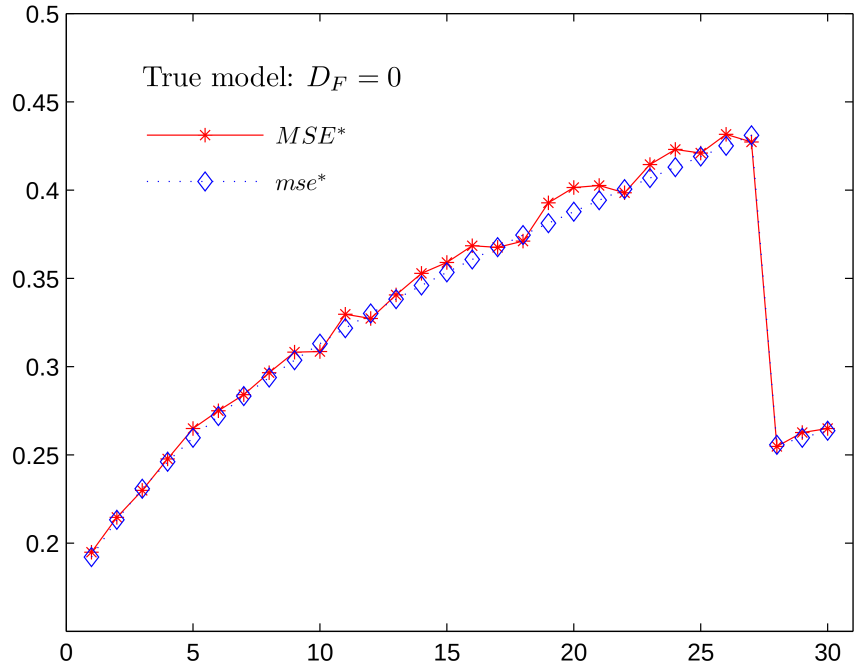


Figure 8. MSE_d, mse_d (left) and MSE^*_d, mse^*_d (right) values for $\boldsymbol{\mu} = (3, 4, 5)$ and $D_F = 0$.

We apply the introduced methodology to the sample of SLFS introduced in the motivation.

Target: to estimate domain totals of unemployed people with EBLUP estimators

We consider 5 cases:

Case 1 - $D_F = 0$

Case 2 - $D_F = 2$

Case 3 - $D_F = 8$

Case 4 - $D_F = 17$

Case 5 - $D_F = 23$

Application to the Labour Force Survey

d	EB 1	CV 1	EB 2	CV 2	EB 3	CV 3	EB 4	CV 4	EB 5	CV 5
1	14520	9,37	14585	9,35	14387	9,87	14333	9,52	14232	9,54
2	10019	11,20	10035	11,20	9990	11,72	10039	11,22	9773	11,44
3	4496	12,59	4492	12,58	4450	13,29	4476	12,68	4465	12,64
4	2676	21,32	2675	21,28	2716	21,94	2721	21,03	2719	20,92
5	1204	43,88	1204	43,76	1266	43,60	1264	41,91	1268	41,56
6	2288	18,54	2289	18,49	2334	18,99	2344	18,16	2347	18,03
7	1728	22,00	1726	21,98	1712	23,21	1720	22,18	1714	22,13
8	824	46,63	824	46,52	850	47,30	875	44,13	877	43,75
9	539	62,24	540	62,03	563	50,33	554	60,77	554	60,40
10	1788	47,38	1789	47,23	1824	39,22	1830	46,44	1835	46,07
11	1184	21,86	1184	21,81	1177	18,58	1193	21,79	1193	21,67
12	336	87,54	335	87,62	340	73,01	371	79,67	368	79,75
13	1065	34,40	1064	34,36	1070	28,93	1114	33,06	1111	32,96
14	1402	21,07	1401	21,04	1411	17,70	1402	21,16	1397	21,12
15	219	187,24	220	186,34	228	152,41	289	142,99	293	140,26
16	993	28,84	994	28,76	1003	24,12	1011	28,45	1013	28,24
17	182	122,14	181	122,22	193	97,14	217	102,82	216	102,68
18	537	42,51	536	42,51	533	36,21	529	39,69	535	42,70
19	453	44,07	453	43,99	467	36,15	461	39,85	501	39,97
20	1686	28,43	1686	28,38	1680	24,12	1688	26,12	1715	27,99
21	441	42,86	441	42,73	459	34,83	452	38,44	493	38,45
22	211	106,92	211	106,57	215	88,70	217	96,02	230	98,58
23	105	94,15	104	94,21	103	81,42	103	88,10	107	92,59
Total	48896	1157	48969	1155	48971	**993**	49203	1046	48957	1053

Table 2. EBLUP and CV estimates of totals of unemployed men in the SLFS 2003-02 of Canary Islands for cases 1-5.

- In the simulation experiment it is shown that if the proposed model ($D_F = 3$) is true and the standard linear mixed model ($D_F = 0$) is used, then a severe lack of precision is achieved.
- However if the true model is the standard linear mixed model ($D_F = 0$), then the reduction of precision because of using the proposed model ($D_F = 3$) is quite moderate.
- An application to real data shows that the best model is found by using a model with both fixed and random effects.

References

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

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

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

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

Njuho, P.M. and Milliken, G.A. (2005). Analysis of linear models with one factor having both fixed and random effects. *Communications in Statistics - Theory and Methods*, **34**, 1979-1989.

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. (2003). *Small area estimation*. John Wiley and Sons, Inc., New-York.

Searle, S.R., Casella, G. and McCullogh, C.E. (1982). *Variance components*. John Wiley and Sons, Inc., New-York.