DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 61–77, http://dx.doi.org/10.1515/JOS-2017-0004

Space-Time Unit-Level EBLUP for Large Data Sets

Michele D'Aló[1], *Stefano Falorsi*[1], *and Fabrizio Solari*[1]

Most important large-scale surveys carried out by national statistical institutes are the repeated survey type, typically intended to produce estimates for several parameters of the whole population, as well as parameters related to some subpopulations. Small area estimation techniques are becoming more and more important for the production of official statistics where direct estimators are not able to produce reliable estimates. In order to exploit data from different survey cycles, unit-level linear mixed models with area and time random effects can be considered. However, the large amount of data to be processed may cause computational problems. To overcome the computational issues, a reformulation of predictors and the correspondent mean cross product estimator is given. The R code based on the new formulation enables the elaboration of about 7.2 millions of data records in a matter of minutes.

Key words: Small area estimation; time series; linear mixed model; small area estimation software.

1. Introduction

Large-scale surveys are usually aimed at providing estimates of target parameters for the whole population, as well as for relevant subpopulations defined at the sampling stage. Design-consistent and design-unbiased direct estimates are produced for the parameters of interest. However, in most surveys, the sample size is not large enough to guarantee reliable estimates for all the target subpopulations. When direct estimates cannot be provided, small area estimation (SAE) methods should be used to overcome the problem (see Rao 2003; Pfeffermann 2002, 2013). SAE methods, usually referred to as indirect estimators, cope with the lack of information from each domain by borrowing strength from samples that belong to other domains, with the result that it increases the effective sample size for each small area.

The most important surveys carried out by national statistical institutes are repeated surveys (see Duncan and Kalton 1987, and Kish 1987). The repeated nature of these surveys allows them to borrow strength not only from other areas but also from other survey cycles.

[1] Italian National Statistical Institute, via Cesare Balbo 16, 00184 Rome, Italy. Emails: dalo@istat.it, stfalors@istat.it, and solari@istat.it.

Acknowledgments: We would like to express our deepest appreciation to the reviewers and the associated editor for providing insightful comments and providing directions for improving the quality of this article. A special thank is also extended to Professor Maria Giovanna Ranalli for her support and for carefully aimed comments and suggestions.

In this context, Saei and Chambers (2003) proposed the use of unit-level linear mixed models (LMMs) with area and time random effects. However, this presents a computational challenge, since large amounts of data from different survey cycles have to be processed. The aim of this article is propose a method to overcome computational problems that may arise from using the predictors and correspondent errors given by Saei and Chambers (2003). For this reason, a reformulation of these expressions will be presented. Furthermore, these more efficient expressions will be applied to the estimation of the unemployment rate at Labour Market Area (LMA) level, using data from the Italian Labour Force Survey (LFS). The case study aims to show the potential gains in efficiency as a result of SAE methods borrowing strength from space and time. It does not aim to suggest a ready solution for official LFS statistics, which necessarily involves many other issues and considerations that are outside the scope of this article.

The LFS is a quarterly survey based on a two-stage stratified cluster design. Municipalities are the primary sampling units, and households are the secondary sampling units. The survey follows a rotating panel sample design, according to the rotation design 2-(2)-2. Households are interviewed in two consecutive quarters. After a two-quarter break, they are interviewed for an additional two consecutive quarters. The sample is uniformly spread across all the weeks, such that all territorial domains are represented in each month and in each of the four waves. The LFS is the main source of information on the Italian labour market and aims to produce monthly, quarterly, and yearly estimates of employment, unemployment, and inactivity rates for different planned territorial domains. Each sample contains information about approximately 170,000 respondents. LMAs, on the other hand, are unplanned areas that are defined every ten years based on daily commuting flows detected by the Population Census. At present, there are 611 LMAs, of which about 450 are included in at least one of the LFS samples in the years 2004–2014. The most unstable estimates refer to the estimation of the unemployment rate. In this case, the Coefficient of Variations (CVs) of the direct estimates are very large, and about three out of four CVs are larger than 30%. Therefore, SAE methods are needed in order to obtain more precise estimates of the unemployment rate that are suitable for dissemination. However, the areas are sampled with unequal selection probabilities in relation to the values of the target variable values. In such situations, standard SAE methods are biased; the magnitude of the bias depends on the sampling fraction and the covariance between the sampling weights and the target variable. However, in the LFS, bias resulting from informative sampling is considered to be small. Treatment of informative sampling in SAE is not considered in this article.

As mentioned above, when LMMs with area and time random effects are assumed, computational problems may result from the large amounts of data to be used in the estimation process. For instance, the data used in this article comes from the 44 LFS quarterly samples in 2004 to 2014, and the overall data size processed comprises about 7,200,000 records.

Usually, in order to overcome the computational problems deriving from large data sets, area-level models are applied. For instance, Rao and Yu (1994) proposed an extension of the basic Fay-Herriot (Fay and Herriot 1979) model to handle time series and cross-sectional data by means of an AR(1) model specification. Datta et al. (2002) and You (1999) used the Rao-Yu model but replace the AR(1) model specification with a random walk model. Pfeffermann and Burck (1990) proposed a general model involving

area-by-time specific random effects. Hidiroglou and You (2016) compared the performances of unit- and area-level models, showing that the former outperforms the latter in terms of bias and mean squared error. Furthermore, Gershunskaya (2015) showed that due to errors associated with the variance of direct estimates, in terms of mean squared error, there is no benefit to introducing temporal correlations between small areas over using the regular Fay-Herriot model. The benefits only become apparent when theoretical variances of direct estimates are used in Rao-Yu model specification.

To avoid the computational issues related to unit-level LMMs, formulas given in Saei and Chambers (2003) have been rewritten in order to involve only small dimensional matrices. The revised expressions, implemented in the ad hoc R function, enable the processing of millions of survey records from different survey cycles in a matter of minutes.

The two-way unit-level linear mixed model with area and time random effects is described in Section 2, while Section 3 is devoted to the reformulation of the expressions needed to compute small area estimates and errors. Section 4 describes some particular SAE methods obtained from the general model. Section 5 includes a case study based on LFS data that aimed to compare the empirical performances of alternative model specifications. Section 6 compares the computational performances of the available SAE software tools with the R function implementing the new expression presented in Section 3. In conclusion, Section 7 presents the most important conclusions of the work.

2. Two-Way Linear Mixed Model

Let d ($d = 1, \ldots, D$) and t ($t = 1, \ldots, T$) denote the generic domain and time indices respectively. For domain d and time t, let N_{dt} and n_{dt} denote population and sample sizes, respectively, and let y_{dti} be the observed value of the target variable for the generic unit i. The parameter of interest is the vector $\boldsymbol{\theta}$ including the population means $\bar{y}_{dt} = (1/N_{dt}) \sum_i y_{dti}$, for all domains and times ($d = 1, \ldots, D$, $t = 1, \ldots, T$). Other relevant parameters for large-scale repeated surveys, such as totals, or net changes between two survey cycles, can be expressed as a linear combination of $\boldsymbol{\theta}$. For this reason, the results in this article can be easily extended to the other types of parameters.

Let us suppose that the data follows the two-way unit-level additive LMM (see Searle et al. 1992; Saei and Chambers 2003)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}, \tag{1}$$

where $\mathbf{X}$, $\mathbf{Z}_1$, $\mathbf{Z}_2$ are known full rank matrices, and $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{e}$ are random vectors, independently distributed from each other. The random effect vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$, modeling between area and time variations not explained by fixed effects, include D and T levels respectively. Furthermore, we assume for $\alpha = 1, 2$, $\mathbf{u}_\alpha \sim N(\mathbf{0}, \mathbf{G}_\alpha)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, where the covariance matrices $\mathbf{G}_\alpha = \sigma_\alpha^2 \boldsymbol{\Omega}_\alpha(\rho_\alpha)$ and $\mathbf{R} = \sigma^2 \mathbf{W}^{-1}$, with $\mathbf{W}$ as a known diagonal matrix. In particular, for $\alpha = 1, 2$, σ_α^2 and ρ_α denote, respectively, the variance and a measure of correlation for the elements of $\mathbf{u}_\alpha$, while σ^2 is the variance of the generic element of $\mathbf{e}$. For notational simplicity, it will be useful to introduce the parametrisation $\phi_\alpha = \sigma_\alpha^2/\sigma^2$. Hence, $\mathbf{y}$ is $N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\omega})$ given by

$$\boldsymbol{\Sigma}(\boldsymbol{\omega}) = \sigma^2(\mathbf{W}^{-1} + \mathbf{Z}\boldsymbol{\Omega}\boldsymbol{\Omega}'),$$

where $\boldsymbol{\omega} = (\sigma^2, \phi_1, \rho_1, \phi_2, \rho_2)$ is the overall variance component vector, $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$, and $\boldsymbol{\Omega} = diag_\alpha\{\phi_\alpha \boldsymbol{\Omega}_\alpha(\rho_\alpha)\}$. The uncorrelated random effect case is obtained by setting $\boldsymbol{\Omega}_1(0) = \mathbf{I}_D$ and $\boldsymbol{\Omega}_2(0) = \mathbf{I}_T$. For models with correlated area effects and correlated time effects, different structures of covariance matrices of random effects can be assumed. For example, $\boldsymbol{\Omega}_1(\rho_1)$ may depend on the distances among the areas, while $\boldsymbol{\Omega}_2(\rho_2)$ may follow an auto-regressive model.

Once the sample is collected, it is useful to partition Model (1) into two parts, depending on whether or not units are observed. In the following, we use the subscripts s and r to refer to sampled and nonsampled population units, respectively. The predicted values for nonsampled population units of $\boldsymbol{\eta}_r = E[\mathbf{y}_r|\mathbf{X}_r, \boldsymbol{\beta}, \mathbf{u}] = \mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u}$ are (see Royall 1976)

$$\tilde{\boldsymbol{\eta}}_r(\boldsymbol{\omega}) = \mathbf{X}_r\tilde{\boldsymbol{\beta}} + \mathbf{Z}_r\tilde{\mathbf{u}}, \tag{2}$$

where $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})$, the Best Linear Unbiased Estimator (BLUE) of β, is given by

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X}_s'\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{X}_s\right]^{-1}\mathbf{X}_s'\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{y}_s,$$

and $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(\omega)$, the Best Linear Unbiased Predictor (BLUP) of $\mathbf{u} = \left[\mathbf{u}'_1\mathbf{u}'_2\right]'$, is

$$\tilde{\mathbf{u}} = \boldsymbol{\Omega}\mathbf{Z}_s'\boldsymbol{\Sigma}_{ss}^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\tilde{\boldsymbol{\beta}}\right).$$

Then, the BLUP of the target parameter $\boldsymbol{\theta}$ is

$$\tilde{\boldsymbol{\theta}}(\boldsymbol{\omega}) = \mathbf{L}_s\mathbf{y}_s + \mathbf{L}_r\tilde{\boldsymbol{\eta}}_r(\boldsymbol{\omega}), \tag{3}$$

where matrices $\mathbf{L}_s$ and $\mathbf{L}_r$ have the block-wise structure $diag_d\left\{diag_t\left\{\mathbf{l}'_{dt}\right\}\right\}$, being $\mathbf{l}'_{dt} = N_{dt}^{-1}\mathbf{1}'_{n_{dt}}$ and $\mathbf{l}'_{dt} = N_{dt}^{-1}\mathbf{1}'_{N_{r,dt}}$ for $\mathbf{L}_s$ and $\mathbf{L}_r$, respectively, and $N_{r,dt} = N_{dt} - n_{dt}$ is the number of nonsampled units in area d at time t.

The BLUP estimator $\tilde{\boldsymbol{\theta}}(\boldsymbol{\omega})$, given in (3), depends on the variance component vector $\boldsymbol{\omega}$, which is unknown in practical applications. By replacing $\boldsymbol{\omega}$ by an estimator, $\hat{\boldsymbol{\omega}}$, a two stage estimator called the Empirical Best Linear Unbiased Predictor (EBLUP) is obtained. Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) and the method of fitting constants can be applied to the estimation of fixed effects and variance components (for details see Harville 1977; Searle et al. 1992; Cressie 1992; Rao 2003; Saei and Chambers 2003). Then, the EBLUP of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\omega}}) = \mathbf{L}_s\mathbf{y}_s + \mathbf{L}_r\hat{\boldsymbol{\eta}}_r(\hat{\boldsymbol{\omega}}),$$

where $\hat{\boldsymbol{\eta}}_r(\hat{\boldsymbol{\omega}})$ is the EBLUP correspondent to (2).

3. Reformulation

In this section, computationally more efficient expressions for predicted area means and the mean cross product error are derived. Results 1 to 5 consist in rewriting the expressions given in Saei and Chambers (2003) as a function of terms dependent on area and time level matrices instead of unit-level matrices. In particular, Result 1 gives the expression of the predicted value for $\mathbf{u}_1$ and $\mathbf{u}_2$, while Result 2 provides the estimate of the regression coefficient $\boldsymbol{\beta}$. Result 3 gives the mean cross-product error (MCPE) for the BLUP of $\boldsymbol{\theta}$. Result 4 computes the expression for updating the variance component estimates when

EBLUP is performed, and finally Result 5 provides the MCPE of the EBLUP of $\boldsymbol{\theta}$. For the sake of simplicity and with obvious notation, we will use matrix operators col$\{\cdot\}$, row$\{\cdot\}$, diag$\{\cdot\}$, and matr$\{\cdot\}$. Different types of correlation matrices $\boldsymbol{\Omega}(\rho)$ can be used for both area and time effects, provided that they depend on a one-dimensional correlation parameter ρ. For instance, the spatial correlation can be specified either as a SAR model based on an adjacency matrix (Cressie 1993), or as exponential or gaussian correlation structures, while the time correlation can follow an AR(1) process.

Two alternative cases for fixed effects are considered. In the first case (Case A), a different regression coefficient vector $\boldsymbol{\beta}_t$, of dimension K, is defined for each time t, determining $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{'}, \ldots, \boldsymbol{\beta}_T^{'}\right)$ to be a $(T \times K)$-dimensional vector. In the second case (Case B), a common regression coefficient vector $\boldsymbol{\beta}$, of dimension K, is considered for all times t. The block-wise structure of matrix $\mathbf{X}$ under the two cases is given by

$$\mathbf{X} = \begin{cases} \text{col}_d\{\text{diag}_t\{\mathbf{X}_{dt}\}\}, & \text{for case } A \\ \text{col}_d\{\text{col}_t\{\{\mathbf{X}_{dt}\}\}, & \text{for case } B \end{cases},$$

where $\mathbf{X}_{dt}$ is the $N_{dt} \times K$ design matrix for area d and time t. The ith row of $\mathbf{X}_{dt}$ is $\mathbf{x}_{dti} = (x_{dti,1}, \ldots, x_{dti,K})^{'}$.

For the random effect part of the model, $\mathbf{u} = col_\alpha\{\mathbf{u}_\alpha\}$, and $\mathbf{Z} = row_\alpha\{\mathbf{Z}_\alpha\}$, where

$$\mathbf{Z}_\alpha = \begin{cases} \text{diag}_d\{\text{col}_t\{\mathbf{1}_{N_{dt}}\}\}, & \text{for } \alpha = 1 \\ \text{col}_d\{\text{diag}_t\{\mathbf{1}_{N_{dt}}\}\}, & \text{for } \alpha = 2 \end{cases}.$$

Finally, $\mathbf{W} = \text{diag}_d\{\text{diag}_t\{\mathbf{W}_{dt}\}\}$ in which $\mathbf{W}_{dt}$ is a diagonal $N_{dt}-$ dimensional matrix, whose generic element, $w_{dti} (i = 1, \ldots, N_{dt})$, is a known constant expressing the heteroscedasticity weight for the unit i in area d at time t.

It is worthwhile to note that matrices and vectors partitioned into sampled and nonsampled units have the same block-wise matrix structure of the corresponding nonpartitioned matrices and vectors, but matrices or vectors referred to area d and time t are, respectively, of size N_{dt} and $N_{r,dt}$ instead of N_{dt}.

Let us define the following quantities referred to as area d and time t:

$$f_{dt} = n_{dt}/N_{dt},$$

$$\bar{y}_{s,dt} = n_{dt}^{-1} \sum_i y_{s,dti},$$

$$\bar{y}_{w,dt} = w_{dt}^{-1} \sum_i w_{dti} y_{dti},$$

$$\bar{\mathbf{x}}_{w,dt} = w_{dt}^{-1} \sum_i w_{dti} \mathbf{x}_{dti},$$

$$\bar{\mathbf{x}}_{r,dt} = N_{r,dt}^{-1} \sum_{\mathbf{i}} \mathbf{x}_{r,dti}.$$

Then, the general aggregated expression of $\tilde{\boldsymbol{\theta}}(\boldsymbol{\omega})$ is

$$\tilde{\boldsymbol{\theta}} = \mathrm{col}_d\{\mathrm{col}_t\{\tilde{\bar{y}}_{dt}\}\},$$

where $\tilde{\bar{y}}_{dt} = \tilde{\bar{y}}_{dt}(\boldsymbol{\omega})$ is

$$\tilde{\bar{y}}_{dt} = f_{dt}\bar{y}_{s,dt} + (1 - f_{dt})\left(\bar{\mathbf{x}}'_{r,dt}\tilde{\boldsymbol{\beta}} + \tilde{u}_{1,d} + \tilde{u}_{2,t}\right), \tag{4}$$

in which $\tilde{u}_{1,d}$ and $\tilde{u}_{2,t}$ are the dth and tth element of $\tilde{\mathbf{u}}_\alpha$, $\alpha = 1, 2$. Let us define $\mathbf{T}^* = \mathbf{T}^*(\boldsymbol{\omega})$ as

$$\begin{aligned}\mathbf{T}^* &= \left[\mathbf{Z}'_s\mathbf{W}_s\mathbf{Z}_s + \boldsymbol{\Omega}^{-1}\right]^{-1} \\ &= \begin{bmatrix} \mathrm{diag}_d\{w_d\} + \phi_1^{-1}\boldsymbol{\Omega}_1^{-1}(\rho_1) & \mathrm{matr}_{dt}\{w_{dt}\} \\ \mathrm{matr}_{td}\{w_{dt}\} & \mathrm{diag}_t\{w_t\} + \phi_2^{-1}\boldsymbol{\Omega}_2^{-1}(\rho_2) \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{T}^*_{11} & \mathbf{T}^*_{12} \\ \mathbf{T}^*_{21} & \mathbf{T}^*_{22} \end{bmatrix},\end{aligned}$$

being $w_d = \sum_t w_{dt}$ and $w_t = \sum_d w_{dt}$, in which $w_{dt} = \sum_i w_{dti}$. Note that $\mathrm{matr}_{td}\{w_{dt}\}$ is the transpose of $\mathrm{matr}_{dt}\{w_{dt}\}$.

Result 1. The predicted values $\tilde{\mathbf{u}}_\alpha = \tilde{\mathbf{u}}_\alpha(\boldsymbol{\omega})$, $\alpha = 1,2$, are obtained as

$$\tilde{\mathbf{u}}_\alpha = \mathbf{T}^*_{\alpha 1} \cdot \mathrm{col}_d\left\{w_d\tilde{\bar{e}}_{w,d}\right\} + \mathbf{T}^*_{\alpha 2} \cdot \mathrm{col}_t\left\{w_t\tilde{\bar{e}}_{w,t}\right\}, \tag{5}$$

for $w_d\tilde{\bar{e}}_{w,d} = \sum_t w_{dt}\tilde{\bar{e}}_{w,dt}$ and $w_t\tilde{\bar{e}}_{w,t} = \sum_d w_{dt}\tilde{\bar{e}}_{w,dt}$, being $\tilde{\bar{e}}_{w,dt} = \tilde{\bar{e}}_{w,dt}(\boldsymbol{\omega})$ given by $\tilde{\bar{e}}_{w,dt} = \bar{y}_{w,dt} - \bar{\boldsymbol{x}}'_{w,dt}\tilde{\boldsymbol{\beta}}$.

Result 2. When case A is considered, the aggregated expression of $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})$ is

$$\tilde{\boldsymbol{\beta}} = [\mathbf{B}_{s,11} - \tilde{\mathbf{B}}_{s,12}]^{-1}[\mathbf{b}_{s,21} - \tilde{\mathbf{b}}_{s,22}], \tag{6}$$

being

$$\mathbf{B}_{s,11} = \mathrm{diag}_t\left\{\sum_d\sum_i w_{dti}\mathbf{x}_{dti}\mathbf{x}'_{dti}\right\}, \tag{7}$$

$$\mathbf{b}_{s,21} = \mathrm{col}_t\left\{\sum_d\sum_i w_{dti}\mathbf{x}'_{dti}\mathbf{y}_{dti}\right\}, \tag{8}$$

$$\tilde{\mathbf{B}}_{s,12} = \mathbf{B}_{\bar{x}_w}\mathbf{T}^*\mathbf{B}'_{\bar{x}_w},$$

$$\tilde{\mathbf{b}}_{s,22} = \mathbf{B}_{\bar{x}_w}\mathbf{T}^*\mathbf{b}_{\bar{y}_w},$$

where

$$\mathbf{B}_{\bar{x}_w} = [\mathrm{matr}_{td}\{w_{dt}\bar{\mathbf{x}}_{w,dt}\}, \mathrm{diag}_t\{w_t\bar{\mathbf{x}}_{w,t}\}],$$

$$\mathbf{b}_{\bar{y}_w} = [\mathrm{row}_d\{w_d\bar{y}_{w,d}\}, \mathrm{row}_t\{w_t\bar{y}_{w,t}\}]',$$

Under Case B, the external block-wise matrix operators in (7) and (8), $\text{diag}_t\{\cdot\}$ and $\text{col}_t\{\cdot\}$, are substituted by $\sum_t\{\cdot\}$, $\mathbf{B}_{\bar{x}_w} = [\text{row}_d\{w_d\bar{\mathbf{x}}_{w,d}\}, \text{row}_t\{w_t\bar{\mathbf{x}}_{w,t}\}]$, and $\mathbf{b}_{\bar{y}_w}$ does not change.

Result 3. Following Saei and Chambers (2003), the MCPE matrix of the BLUP $\tilde{\boldsymbol{\theta}}$ is given by

$$\text{MCPE}(\tilde{\theta}) = \text{E}[(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})'] = \mathbf{G}_1(\boldsymbol{\omega}) + \mathbf{G}_2(\boldsymbol{\omega}) + \mathbf{G}_4(\boldsymbol{\omega}), \tag{9}$$

where the aggregated expressions of $\mathbf{G}_1(\boldsymbol{\omega})$, $\mathbf{G}_2(\boldsymbol{\omega})$ and $\mathbf{G}_4(\boldsymbol{\omega})$ are:

$$\mathbf{G}_1(\boldsymbol{\omega}) = \sigma^2\mathbf{Z}_r^*\mathbf{T}^*\mathbf{Z}_r^{*'} = \sum_{\alpha}\sum_{\alpha'}\mathbf{a}_\alpha\mathbf{T}^*_{\alpha,\alpha'}\mathbf{a}_{\alpha'}, \tag{10}$$

$$\begin{aligned}\mathbf{G}_2(\boldsymbol{\omega}) =& \sigma^2\left(\mathbf{X}_r^* - \mathbf{Z}_r^*\mathbf{T}^*\mathbf{Z}_s'\mathbf{W}_s^{-1}\mathbf{X}_s\right)\left(\mathbf{B}_{s,11} - \tilde{\mathbf{B}}_{s,12}\right)^{-1} \\ &\times\left(\mathbf{X}_r^{*'} - \mathbf{X}_s'\mathbf{W}_s^{-1}\mathbf{Z}_s\mathbf{T}^*\mathbf{Z}_r^{*'}\right),\end{aligned} \tag{11}$$

$$\mathbf{G}_4(\boldsymbol{\omega}) = \sigma^2\mathbf{L}_r\mathbf{W}_r^{-1}\mathbf{L}_r' = \sigma^2(\text{diag}_d\{\text{diag}_t\{\mathbf{W}_{r,dt}\}\}), \tag{12}$$

being

$$\mathbf{X}_r^* = \mathbf{L}_r\mathbf{X}_r = \text{col}_d\left\{\text{diag}_t\left\{N_{r,dt}\bar{\mathbf{x}}_{r,dt}'\right\}\right\},$$

when case A is considered, while the internal operator $\text{diag}_t\{\cdot\}$ is substituted by $\text{col}_t\{\cdot\}$ under case B. In addition,

$$\mathbf{Z}_r^* = \mathbf{L}_r\mathbf{Z}_r = [\text{row}_d\{\text{diag}_t\{N_{r,dt}\}\}, \text{diag}_d\{\text{row}_t\{N_{r,dt}\}\}], \tag{13}$$

in which $\mathbf{a}_1 = \text{col}_d\{\text{diag}_t\{N_{r,dt}\}\}$, $\mathbf{a}_2 = \text{diag}_d\{\text{col}_t\{N_{r,dt}\}\}$, $\mathbf{a}_3 = \mathbf{a}_1' = \text{row}_d\{\text{diag}_t\{N_{r,dt}\}$, $\mathbf{a}_4 = \mathbf{a}_2' = \text{diag}_d\{\text{row}_t\{N_{r,dt}\}\}$.

Hence, the BLUP estimator, $\tilde{\boldsymbol{\theta}}$, given in Results 1 and 2, depends on the variance components vector $\boldsymbol{\omega}$, which is unknown in practical applications. Replacing $\boldsymbol{\omega}$ by an estimator, $\hat{\boldsymbol{\omega}}$, the correspondent EBLUP is obtained.

Result 4. The EBLUP $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\omega}})$ of $\boldsymbol{\theta}$ corresponding to (4) is given by

$$\hat{\boldsymbol{\theta}} = \text{col}_d\left\{\text{col}_t\left\{\hat{\bar{y}}_{dt}\right\}\right\}, \tag{14}$$

where $\hat{\bar{y}}_{dt}$ is the EBLUP of $\bar{y}_{dt}$. The explicit expression of $\hat{\bar{y}}_{dt} = \hat{\bar{y}}_{dt}(\hat{\boldsymbol{\omega}})$ is obtained by substituting the estimate $\hat{\boldsymbol{\omega}}$ of the variance component vector $\boldsymbol{\omega}$ into (6) and (5), namely $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$, $\hat{\mathbf{u}}_1 = \hat{\mathbf{u}}_1(\hat{\boldsymbol{\omega}})$ and $\hat{\mathbf{u}}_2 = \hat{\mathbf{u}}_2(\hat{\boldsymbol{\omega}})$.

REML estimates of model parameters are obtained following the iterative algorithm given in Saei and Chambers (2003). Compact expressions for updating the variance components from iteration k to iteration $k+1$ are:

$$\hat{\sigma}^2 = (n-Q)^{-1}\left(\sum_d\sum_t\sum_i w_{dti}y_{dti}\left(y_{dti} - \mathbf{x}_{dti}^{'}\hat{\boldsymbol{\beta}}\right) + \hat{\mathbf{u}}_1\mathbf{1}_D + \hat{\mathbf{u}}_2\mathbf{1}_T\right),$$

$$\hat{\varphi}_1 = \frac{1}{T}\left(\mathrm{tr}\left\{\hat{\mathbf{T}}_{s,11} + \hat{\mathbf{P}}_1\left(\hat{\mathbf{B}}_{11} - \hat{\mathbf{B}}_{21}\right)^{-1}\hat{\mathbf{P}}_1^{'}\right\} + \hat{\sigma}^{-2}\hat{\mathbf{u}}_1^{'}\boldsymbol{\Omega}_1^{-1}\hat{\mathbf{u}}_1\right),$$

$$\hat{\varphi}_2 = \frac{1}{D}\left(\mathrm{tr}\left\{\hat{\mathbf{T}}_{s,22} + \hat{\mathbf{P}}_2\left(\hat{\mathbf{B}}_{11} - \hat{\mathbf{B}}_{21}\right)^{-1}\hat{\mathbf{P}}_2^{'}\right\} + \hat{\sigma}^{-2}\hat{\mathbf{u}}_2^{'}\boldsymbol{\Omega}_2^{-1}\hat{\mathbf{u}}_2\right),$$

where Q denotes the number of columns of $\mathbf{X}$, $\hat{\mathbf{T}}_s = \hat{\mathbf{T}}^* + \hat{\mathbf{P}}(\hat{\mathbf{B}}_{11} - \hat{\mathbf{B}}_{21})^{-1}\hat{\mathbf{P}}'$, with $\hat{\mathbf{T}}^* = \mathbf{T}^*(\hat{\omega})$ and

$$\hat{\mathbf{P}} = \begin{bmatrix}\hat{\mathbf{P}}_1\\ \hat{\mathbf{P}}_2\end{bmatrix} = \begin{bmatrix}\mathrm{matr}_{dt}\left\{N_{dt}\bar{\mathbf{x}}_{dt}^{'}\right\}\hat{T}_{11}^* + \mathrm{diag}_t\left\{N_t\bar{\mathbf{x}}_t^{'}\right\}\hat{T}_{12}^*\\ \mathrm{matr}_{dt}\left\{N_{dt}\bar{\mathbf{x}}_{dt}^{'}\right\}\hat{T}_{21}^* + \mathrm{diag}_t\left\{N_t\bar{\mathbf{x}}_t^{'}\right\}\hat{T}_{22}^*\end{bmatrix},$$

$$\hat{\rho}_1(k+1) = \hat{\rho}_1(k) + I(\hat{\rho}_1) + \Delta(l_{\mathrm{REML}}(\hat{\rho}_1)), \tag{15}$$

$$\hat{\rho}_2(k+1) = \hat{\rho}_2(k) + I(\hat{\rho}_2) + \Delta(l_{\mathrm{REML}})(\hat{\rho}_2)), \tag{16}$$

where $\Delta(l_{\mathrm{REML}})(\cdot)$ is the derivative of the likelihood function with respect to the parameter of interest, whereas $I(\cdot)$ is the relevant element of the inverse of the information matrix. The expressions given above are updated iteratively together with the expression (6) for $\tilde{\boldsymbol{\beta}}$ given in Result 2 until convergence is attained.

Result 5. The MCPE of the EBLUP $\hat{\boldsymbol{\theta}}$ is given by the diagonal elements of the following matrix

$$\mathrm{MCPE}(\hat{\boldsymbol{\theta}}) = \mathrm{MCPE}(\tilde{\boldsymbol{\theta}}) + 2\mathbf{G}_3(\hat{\boldsymbol{\omega}}) = \mathbf{G}_1(\hat{\boldsymbol{\omega}}) + \mathbf{G}_2(\hat{\boldsymbol{\omega}}) + 2\mathbf{G}_3(\hat{\boldsymbol{\omega}}) + \mathbf{G}_4(\hat{\boldsymbol{\omega}}),$$

where $\mathbf{G}_1(\hat{\boldsymbol{\omega}})$, $\mathbf{G}_2(\hat{\boldsymbol{\omega}})$, $\mathbf{G}_4(\hat{\boldsymbol{\omega}})$ are computed, respectively, plugging into (9), (10), (11), and (12) the estimated values of the variance components. Matrix $\mathbf{G}_3(\hat{\boldsymbol{\omega}})$ takes into account the uncertainty of the estimation of the variance components. The explicit expression of $\mathbf{G}_3(\hat{\boldsymbol{\omega}})$ is

$$\mathbf{G}_3(\hat{\boldsymbol{\omega}}) = \hat{\sigma}^2\left[tr\left(\nabla_\alpha\hat{\boldsymbol{\Sigma}}_s^*\nabla_{\alpha'}^{'}\hat{\mathbf{B}}\right)\right],$$

where $\hat{\mathbf{B}}$ is the asymptotic covariance matrix of the REML estimates of the variance component vector $\boldsymbol{\omega}$. It depends on the diagonal elements of the inverse of the Fisher information matrix of REML estimators $\hat{\boldsymbol{\omega}}$. For more details, see Saei and Chambers (2003). Furthermore, ∇_α and $\boldsymbol{\Sigma}_s^*$ have the following expression

$$\nabla_\alpha = -\left(\mathbf{Z}_\alpha^*\hat{\mathbf{T}}^*\otimes\mathbf{I}_H\right)\left(\frac{\delta\boldsymbol{\Omega}^{-1}}{\delta\boldsymbol{\omega}}\right)_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}}\hat{\mathbf{T}}^*,$$

$$\boldsymbol{\Sigma}_s^* = \mathbf{A} + \mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega},$$

where, denoting with $\otimes$ the Kronecker product, $\mathbf{Z}^*_\alpha$ is the αth row of matrix $\mathbf{Z}^*_r$ given in (13), $\mathbf{I}_H$ is the identity matrix of dimension $H_{=4}$, and $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} \text{diag}_d\{w_{s,d}\} & \text{matr}_{dt}\{w_{s,dt}\} \\ \text{matr}_{td}\{w_{s,dt}\} & \text{diag}_t\{w_{s,t}\} \end{bmatrix},$$

for $w_{s,d} = \sum_t w_{s,dt}$, $w_{s,t} = \sum_d w_{s,dt}$ and $w_{s,dt} = \sum_i^{n_{dt}} w_{dti}$.

4. Particular Cases

Starting from the general LMM specification in Saei and Chambers (2003), we describe the more relevant model, and two random effect model specifications presented in the literature. The general model (1) will be denoted by $\text{M}^{\text{ST}}_{\text{CC}}$, where the superscript ST stands for model with spatial and temporal random effects, and subscript CC stands for using a correlation structure for both the random effects.

When N_d is large, $f_{dt} \cong 0$ and $\bar{\mathbf{x}}_{r,dt} \cong \bar{\mathbf{x}}_{dt}$, and the general formula (4) of the unit-level EBLUP with space and time correlation, $\hat{\bar{y}}_{dt}$, can be approximated by

$$\text{M}^{ST}_{CC} : \hat{\bar{y}}_{dt} = \bar{\mathbf{x}}'_{dt}\hat{\boldsymbol{\beta}} + \sum_{d'}\sum_{t'} \hat{\gamma}_{d't'}\hat{\bar{e}}_{w,d't'}, \tag{17}$$

where $\hat{\gamma}_{d't'} = w_{d't'}\hat{\Gamma}_{dt}(d',t')$, being $\hat{\Gamma}_{dt}(d',t') = \hat{T}^*_{11,dd'} + \hat{T}^*_{12,dt'} + \hat{T}^*_{21,td'} + \hat{T}^*_{22,tt'}$. The corresponding estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained by means of (14).

Special cases of (17) are obtained through particular settings for $\hat{\Gamma}_{dt}(d',t')$. Using analogous notation, $\text{M}^{\text{ST}}_{\text{II}}$ is the two-way model with independent and identically distributed area and time effects, while $\text{M}^{\text{ST}}_{\text{IC}}$ and $\text{M}^{\text{ST}}_{\text{CI}}$ denote, respectively, the two-way linear mixed model with independent area effects and correlated time effects, and spatially correlated area effects and independent time effects.

The case of two independent random effects, $\text{M}^{\text{ST}}_{\text{II}}$, is obtained when $\hat{\Gamma}_{dt}(d',t') = \hat{T}^*_{11,dd} + \hat{T}^*_{22,tt}$. Therefore, the expression for the estimator is given by

$$\text{M}^{\text{ST}}_{\text{II}} : \hat{\bar{y}}_{dt} = \bar{\mathbf{x}}'_{dt}\hat{\boldsymbol{\beta}} + \hat{\gamma}_d\hat{\bar{e}}_{w,d} + \hat{\gamma}_t\hat{\bar{e}}_{w,t},$$

where $\hat{\gamma}_d = w_d\hat{T}^*_{11,dd} = \hat{\sigma}^2_1/\left(\hat{\sigma}^2_1 + \hat{\sigma}^2/w_d\right)$ and $\hat{\gamma}_t = w_t\hat{T}^*_{22,tt} = \hat{\sigma}^2_2/\left(\hat{\sigma}^2_2 + \hat{\sigma}^2/w_t\right)$. This estimator may be applied in many real situations, for example, when the spatial and temporal correlation between area and time effects is lower than a given threshold. Furthermore, it may be useful for cross-sectional surveys in which index t, instead of representing time, represents a set of T domains which form a different partition of the population than the D areas.

In many practical situations, it may be useful to consider the two estimators $\text{M}^{\text{ST}}_{\text{CI}}$ and $\text{M}^{\text{ST}}_{\text{IC}}$. Model $\text{M}^{\text{ST}}_{\text{IC}}$ can be used for repeated business surveys, in which the small areas of interest are small domains different from territorial subpopulations (e.g., industry segments) and it is not possible, or straightforward, to define spatial correlation among domains.

One-way models $\text{M}^{\text{S}}_{\text{C}}$ and $\text{M}^{\text{S}}_{\text{I}}$, respectively, with spatially correlated area effects and independent and identically distributed area effects, allow traditional cross-sectional small

area estimation to borrow strength from other domains, but not from other survey cycles. Specifically, M_I^S corresponds to the standard model defined by Battese et al. (1988), while examples for M_C^S are given in Saei and Chambers (2003), and Petrucci and Salvati (2004). To borrow strength from other survey cycles but not from other domains, alternative modelisations for usual time series models are M_C^T and M_I^T, that is, linear mixed models with correlated time effects and independent and identically distributed time effects.

5. Application to Real Data

In this section we present a case study aimed at comparing several alternative SAE models and at testing different SAE software estimation tools. To this end, LFS data from 2004 to 2014 has been used to produce estimates of the unemployment rate at LMA level. The overall amount of data is about 7,200,000 records and about 25% of LMAs are not covered by the samples.

LMMs with both area and time random effects are considered, and their estimation is made possible by means of the expressions described in Section 3. The corresponding estimator has been applied to compute quarterly LMA unemployment rates and compared with other standard SAE methods.

The binary nature of the target variable should suggest the use of non-normal mixed models, for instance a binomial with a logistic link function. However, D'Aló et al. (2012) showed that the use of logistic models does not improve substantially the quality of the estimates with respect to normal model. Furthermore, Boonstra et al. (2007) do not find evidence for the superiority of logistic mixed models over their normal counterparts in the estimation of unemployment counts in Dutch municipalities. In addition, we are not usually interested in individual predictions, but rather in predicting area and time aggregates. Besides, for non-normal mixed models, easy interpretable closed-form expressions for predictors are not available. Linear mixed models only need area and time population totals for prediction, while non-normal models require cross-classified population totals for the fixed effects, even though only marginal effects are included in the model specification.

The LMMs and the correspondent estimators considered in the experimental study are reported in Table 1.

In addition to the direct estimator, EBLUPs arising from one-way and two-way unit-level LMMs are considered. Therefore, the estimator with area- and time-correlated

Table 1. List of models and estimators considered.

Model	Estimator
–	Direct
M_I^S	$EBLUP_I^S$
M_C^S	$EBLUP_C^S$
M_{CC}^{ST}	$EBLUP_{CC}^{ST}$
$M_I^{S(**)}$	$EBLUP\_ALL_I^S$
$M_C^{S(**)}$	$EBLUP\_ALL_C^S$

$^{(**)}$Model parameters are estimated using all LFS data from 2004 to 2014.

random effects, $\text{EBLUP}_{\text{CC}}^{\text{ST}}$, is compared with two SAE cross-sectional methods, specifically with the EBLUP with uncorrelated area random effects, $\text{EBLUP}_{\text{I}}^{\text{S}}$, and with spatially correlated area random effects, $\text{EBLUP}_{\text{C}}^{\text{S}}$. Furthermore, $\text{EBLUP}_{\text{CC}}^{\text{ST}}$ is computed using the whole set of available time series data, while the cross-sectional methods exploit only the last quarter data set. Then, in order to be able to set aside the effect of the amount of data, when comparing $\text{EBLUP}_{\text{CC}}^{\text{ST}}$ with its competitors, the one-way model parameters have also been estimated using the overall set of data. These last two estimators are denoted by $\text{EBLUP\_ALL}_{\text{I}}^{\text{S}}$ and $\text{EBLUP\_ALL}_{\text{C}}^{\text{S}}$, respectively.

In particular, for $\text{EBLUP}_{\text{CC}}^{\text{ST}}$, the between-area correlation matrix proposed by Saei and Chambers (2003) has been considered. This matrix is dependent on the distances among the areas and on a scale parameter ρ_1 connected to the spatial structure of the areas, and is given by

$$\boldsymbol{\Omega}_1(\rho_1) = \left[1 + \delta_{d,d'} \exp\left(\frac{\text{dist}(d,d')}{\rho_1}\right)\right]^{-1},$$

with $\delta_{d,d'} = 0$ if $d = d'$ and $\delta_{d,d'} = 1$ otherwise and $dist(d,d')$ denoting the Euclidean distance between area d and d'. Instead, the between-time correlation matrix arises from an autoregressive AR(1) process whose expression is

$$\boldsymbol{\Omega}_2(\rho_2) = \frac{1}{1-\rho_2^2}\begin{bmatrix} 1 & \rho_2 & \cdots & \rho_2^{T-1}\rho_2 \\ 1 & \cdots & \rho_2^{T-2} & \vdots \\ \ddots & \vdots & \rho_2^{T-1} & \rho_2^{T-2} \quad \cdots \\ 1 & & & \end{bmatrix}.$$

The scope of the empirical study is to assess the statistical properties of the estimators. To this aim, the estimates computed for the last quarter of 2011 (October 2011–December 2011) are compared with the correspondent 2011 Census values, which are referred to on 9 October 2011.

The auxiliary information used in the experimental study, that is, the cross-classification of 14 age groups by sex, is similar to what is used in the LFS calibration process. A common regression coefficient vector is defined for all the quarters. This is the hypothesis defined in Section 2 as Case B. We note that the assumption of fixed effects over time is not very realistic, but the correlated random effects are expected to smooth the estimates. A first comparison among the estimators has been carried out by means of Average Absolute Relative Error (AARE) and Average Squared Error (ASE), defined as

$$\text{AARE}(\hat{\boldsymbol{\theta}}) = \frac{1}{D}\sum_{d=1}^{D} \text{ARE}_d = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{\hat{\theta}_d}{\theta_d} - 1\right|,$$

$$\text{ASE}(\hat{\boldsymbol{\theta}}) = \frac{1}{D}\sum_{d=1}^{D} \text{SE}_d = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_d - \theta_d\right)^2,$$

where for domain d, $d = 1, \ldots, D$, $\hat{\theta}_d$ and θ_d are, respectively, the estimate computed with a given estimator and the true parameter of interest.

Table 2. AARE and ASE with respect to 2011 Census data.

Estimator	AARE	ASE$^{(*)}$
Direct	0.65	18.86
$EBLUP_I^S$	0.34	2.67
$EBLUP_C^S$	0.33	2.59
$EBLUP_{CC}^{ST}$	0.26	2.07
$EBLUP\_ALL_I^{S(**)}$	0.36	3.56
$EBLUP\_ALL_C^{S(**)}$	0.36	3.56

$^{(*)}$ASE is multiplied by 1,000.
$^{(**)}$Model parameters are estimated using all LFS data from 2004 to 2014.

Table 2 displays the values of AARE and ASE evaluated over the 611 LMAs. The $EBLUP_{CC}^{ST}$ outperforms the others estimators both in terms of AARE and ASE. It shows better performances than $EBLUP_I^S$ and $EBLUP_C^S$. $EBLUP_I^S$ and $EBLUP_C^S$ performed similarly, with a slight preference for the $EBLUP_C^S$. This implies there is no strong evidence for a significant spatial correlation. Therefore, the introduction of the time random effect substantially increases the efficiency of the estimates. In fact, the estimated value of the time correlation coefficient ρ_2, computed with (16), is equal to 0.73, while the estimate of the spatial parameter, obtained by means of (15), is 0.29. The spatial correlation defined for the area random effects allows us to obtain more accurate estimates for out-of-sample areas than the corresponding estimates computed only by synthetic prediction. Furthermore, the better performance of $EBLUP_{CC}^{ST}$ is not only due to the larger set of data involved in the estimation process. In fact, $EBLUP\_ALL_I^S$ and $EBLUP_C^S$, which use the same data as $EBLUP_{CC}^{ST}$, perform poorly because they do not capture the true time pattern of data.

Table 3 reports the value of the coefficients of variation for all the estimates, with the exception of $EBLUP\_ALL_I^S$ and $EBLUP\_ALL_C^S$. It shows that $EBLUP_{CC}^{ST}$ outperforms the other methods, aside from minimum and maximum values. The direct estimator shows a better coefficient of variation value only for the minimum value.

Figures 1a and 1b show the distribution of ARE and SE, respectively. The error distribution of the direct estimator is not included due to its poor performance. In accordance with Table 2, in both cases the distribution of the errors for $EBLUP_{CC}^{ST}$ is more concentrated around zero than the other distributions, with the exception of $EBLUP\_ALL_I^S$ and $EBLUP\_ALL_C^S$ for the ARE.

Figure 2 displays the spatial distribution of the estimates for direct estimator (a), $EBLUP_I^S$ (b), $EBLUP_C^S$ (c) and $EBLUP_{CC}^{ST}$ (d). The direct estimates are plotted for

Table 3. CV% distribution.

Estimator	Min.	1st Q	Median	Mean	3rd Q	Max.
Direct$^{(*)}$	0.72	31.57	52.01	54.56	77.34	119.80
$EBLUP_I^S$	4.83	19.83	26.99	26.42	33.95	45.46
$EBLUP_C^S$	4.83	19.71	27.18	26.35	33.79	45.60
$EBLUP_{CC}^{ST}$	1.19	4.23	6.16	7.80	9.44	27.97

$^{(*)}$There are 158 empty LMAs.

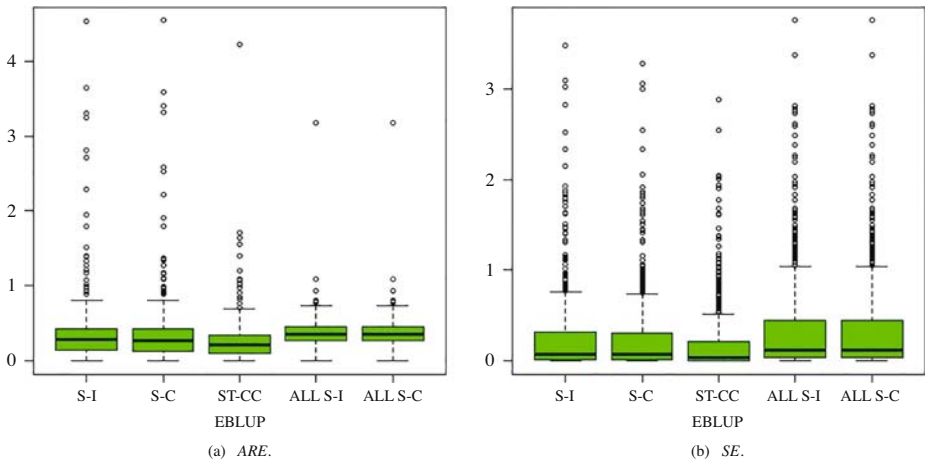


Fig. 1. ARE and SE distributions. SE is multiplied by 1,000.

provinces, while the estimates for the EBLUPs are plotted on LMAs. This is because the 110 provinces are planned domains for which the direct estimator produce reliable estimates. The spatial distribution of the direct estimates can be considered as a good picture of the spatial distribution of the true unemployment rates, and for that it can be used to benchmark SAE estimates. As showed in Figure 1, all the EBLUPs have analogous spatial patterns to the distribution of the direct estimates.

6. SAE Software for Unit-Level Linear Mixed Models

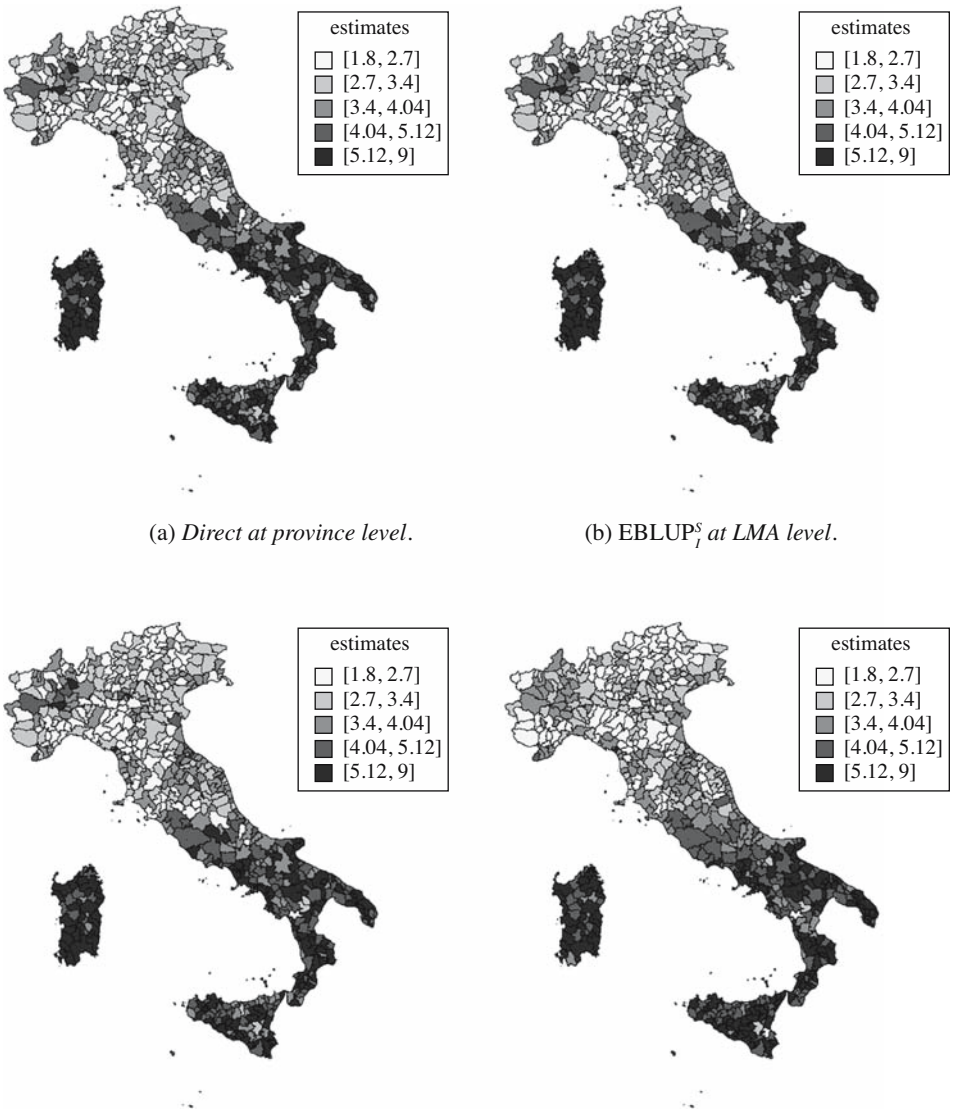
We implemented the new formulation given in Section 3 in an R function named space.time.eblup, which allows the computation of (a) estimates of the model parameters; (b) SAE estimates and their MSEs for sampled areas; (c) SAE estimates and their MSEs also for out-of-sample areas. In this section, the performance of this function is compared with the most used software tools, available for SAE or for LMMs fitting. An exhaustive review of available SAE software tools is provided by the Essnet SAE project.

The available SAE software packages carry out a complete estimation process with the computation of (a) and (b), but, usually, do not allow (c). LMMs can be estimated using general software tools for model fitting. In this case, they allow only (a), and extra work is needed to complete the estimation process, that is, (b) and (c).

The result of the comparative analysis of space.time.eblup compared with the other available functions and SAE packages shows evidence that space.time.eblup, in addition to performing a more complete estimation process, is more efficient in terms of runtime.

Table 4 reports SAE software tools developed recently by national or international projects dealing with small area estimation. All SAE software provides a complete tool for treating SAE problems, but only the R functions produced by SAMPLE are able to deal with LMMs that include area and time random effects. Specifically, time random effects are nested within area random effects instead of including additive random effects as in (1). Furthermore, no correlation structure can be specified for the area random effects.

Besides the software tools described in Table 4, R packages specifically dedicated to SAE are available for download at the CRAN, https://cran.r-project.org/. The SAE



(a) *Direct at province level*. (b) EBLUP$^{S}_{I}$ *at LMA level*.

(c) EBLUP$^{S}_{C}$ *at LMA level*. (d) EBLUP$^{ST}_{CC}$ *at LMA level*.

Fig. 2. Unemployment rate estimates for direct (a), EBLUP$^{S}_{I}$ (b), EBLUP$^{S}_{C}$ (c), EBLUP$^{ST}_{CC}$ (d). Legends display the estimated unemployment rate classes.

packages implementing unit-level LMMs are hbsae, JoSAE, rsae, sae, but do not include time random effects in the model. Furthermore, as far as the software toos describe in Table 4 are concerned, it is worthwhile to underline that they can only handle sets of data much smaller than the 7,200,000 records processed for the case study.

Besides SAE packages, there are many R packages that provide functions for fitting LMMs. A general package for LMMs is lme4. It can fit linear mixed models by means of the function lmer. These models can also be fitted using the function lme from the package

Table 4. Description of SAE software based on unit-level LMMs produced by projects on small area estimation.

Project	Enviroment	Area random effects	Time random effects
EURAREA	SAS	Correlated	No
BIAS	R	Uncorrelated	No
SAMPLE	R	Uncorrelated	Nested, Correlated
AMELI	R	Uncorrelated	No
ESSnet SAE	R	Correlated	No

nlme. This package supports various correlation and heteroscedasticity structures for the variance within.

Concerning the statistical software SAS, (see https://www.sas.com/), apart from the macro program codes developed by the EURAREA project, no ad hoc SAE software is available. The SAS procedure MIXED can fit a variety of LMMs. It performs model estimation that provides both fixed and random effects estimates, and variance components estimates.

Table 5 compares, in terms of computation times, the performances of the most used SAS and R functions to fit LMMs with the space.time.eblup function. Only the procedure MIXED in SAS allows us to handle the whole set of data used in the case study of the Italian LFS. However, when spatial and temporal correlation is introduced, it can only process much smaller data sets. The R functions tested to fit LMMs were not able to process the whole set of data, but only a subset including about 3,000,000 records related to the first 18 survey occasions. Furthermore, similarly to the SAS procedure MIXED, lme and lmer can fit models with correlated random effects only for very small sets of data. For this reason, the only comparison framework that can be set up is restricted to the 18 survey occasions sets of data, and without taking into account any type of correlation structure. Moreover, the space.time.eblup function is a complete SAE tool providing computation of (a), (b), and (c).

All the performances of R and SAS codes were run on an Intel CoreTM i7-3770K 3.50 GHz processor with 8 GB RAM on a 64 bit Windows 7 personal computer.

7. Conclusions

Since the most important surveys carried by national statistical institutes are repeated surveys, it is important to carefully consider SAE problems within this broad and relevant survey framework. Standard small area models usually take into account cross-sectional

Table 5. Comparison of performances, in terms of computer time, for R and SAS functions fitting unit-level LMMs and space.time.eblup R function for Italian LFS data, complete and reduced.

Package	Complete data set	Restricted data set
PROC MIXED(*)	30 sec	12 sec
lme(*)	–	3 min 00 sec
lmer(*)	–	2 min 21 sec
space.time.eblup	4 min 54 sec	2 min 18 sec

(*)Elaboration times are related to independent area and time random effects.

estimation. Nonetheless, in the context of repeated surveys, more realistic and efficient models can be considered by adding a temporal random effect for exploiting previous survey occasions data. It potentially allows us to increase the efficiency of results by using more realistic SAE models that can better capture the real variability of the phenomena under study. Furthermore, unit-level models have potentially more predictive power than area-level models, and they are able to exploit the individual correlations between target variable and fixed effects covariates.

As a consequence, large amount of data have to be processed and computational problems may occur. The empirical test, conducted on Italian LFS quarterly data, displayed good statistical performance, outperforming the other estimators. Furthermore, the new formulation was shown to be effective when dealing with extremely large amounts of data. As a matter of fact, the function space.time.eblup, implementing the new expressions was able to process 7,200,000 survey records from the 44 LFS quarterly samples from 2004 to 2014 in about five minutes. Therefore, the new formulation allows us to manage very large amounts of data, overcoming the computational limits underlying the software currently available. Moreover, it can provide a valuable starting point for building more sophisticated models.

Currently, only the R function is available for use. However, an R package will be produced and made available as soon as possible.

8. References

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. "An Error Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of American Statistical Association* 83: 28–36. Doi: http://dx.doi.org/10.1080/01621459.1988.10478561.

Boonstra, H., B. Buelens, and M. Smeets. 2007. "Estimation of Municipal Unemployment Fractions - A Simulation Study Comparing Different Small Area Estimators." Internal report, BPA-no. DMK-DMH-2007-04-20-HBTA, Herleen: Statistics Netherlands.

Cressie, N. 1992. "REML Estimation in Empirical Bayes Smoothing of Census Undercount." *Survey Methodology* 18: 75–94.

Cressie, N. 1993. *Statistics for Spatial Data*. New York: Wiley.

D'Aló, M., L. Di Consiglio, S. Falorsi, M.G. Ranalli, and F. Solari. 2012. "Use of Spatial Information in Small Area Models for Unemployment Rate Estimation at Sub-Provincial Areas in Italy." *Journal of the Indian Society of Agricultural Statistics* 66: 43–54.

Datta, G.S. P. Lahiri, and T. Maiti. 2002. "Empirical Bayes Estimation of Median Income of Four-Person Families by State Using Time Series and Cross-sectional Data." *Journal of Statistical Planning and Inference* 102: 83–97. Doi: http://dx.doi.org/10.1016/S0378-3758(01)00173-2.

Duncan, G.J. and G. Kalton. 1987. "Issues of Design and Analysis of Surveys across Time." *International Statistical Review* 55: 97–117. Doi: http://dx.doi.org/10.2307/1403273.

Fay, R. and R. Herriot. 1979. "Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74: 269–277. Doi: http://dx.doi.org/10.1080/01621459.1979.10482505.

Gershunskaya, J. 2015. "Combining Time Series and Cross-Sectional Data for the Current Employment Statistics Estimates." In Proceedings of the Section on Statistical Computing: American Statistical Association, August 9, 2015. 1085–1096. Alexandria, VAL: American Statistical Association. Available at: http://www.amstat.org/sections/srms/proceedings/y2015/files/233962.pdf (January 16, 2017).

Harville, D.A. 1977. "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." *Journal of the American Statistical Association* 72: 320–338. Doi: http://dx.doi.org/10.2307/2286797.

Hidiroglou, M.A. and Y. You. 2016. "Comparison of Unit Level and Area Level Small Area Estimators." *Survey Methodology* 42: 41–61. Available at: http://www.statcan.gc.ca/pub/12-001-x/2016001/article/14540-eng.pdf (January 16, 2017).

Kish, L. 1987. *Statistical Designs for Research*. New York: Wiley.

Petrucci, A. and N. Salvati. 2004. "Small Area Estimation Considering Spatially Correlated Errors: the Unit Level Rrandom Effects Model." Working Paper 2004/10, Department of Statistics, Florence University.

Pfeffermann, D. 2002. "Small Area Estimation: New Developments and Directions." *International Statistical Review* 70: 125–143. Doi: http://dx.doi.org/10.2307/1403729.

Pfeffermann, D. 2013. "New Important Developments in Small Area Estimation." *Statistical Science* 28: 40–68. Doi: http://dx.doi.org/10.1214/12-sts395.

Pfeffermann, D. and L. Burck. 1990. "Robust Small Area Estimation Combining Time Series and Cross-Sectional Data." *Survey Methodology* 16: 217–237.

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

Rao, J.N.K. and M. Yu. 1994. "Small Area Estimation by Combining Time Series and Cross-Sectional Data." *Canadian Journal of Statistics* 22: 511–528. Doi: http://dx.doi.org/10.2307/3315407.

Royall, R.M. 1976. "The Linear Least-Squares Prediction Approach to Two-Stage Sampling." *Journal of the American Statistical Association* 71: 657–664. Doi: http://dx.doi.org/10.1080/01621459.1976.10481542.

Saei, A. and R. Chambers. 2003. "Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects". *Methodology Working Paper M03/15*. University of Southampton: Southampton Statistical Sciences Research Institute. Available at: http://eprints.soton.ac.uk/8165/1/8165-01.pdf (January 16, 2017).

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. *Variance Components*. New York: Wiley.

You, Y. 1999. "Hierarchical Bayes and Related Methods for Model Based Small Area Estimation." Ph.D. Thesis, School of Mathematics and Statistics, Carleton University.

Received September 2015
Revised September 2016
Accepted October 2016