REVSTAT – Statistical Journal
Volume 9, Number 2, June 2011, 155–180

A SPATIAL UNIT LEVEL MODEL FOR SMALL AREA ESTIMATION

Authors: Pedro S. Coelho
– ISEGI – Universidade Nova de Lisboa, Portugal
Faculty of Economics, Ljubljana University, Slovenia
psc@isegi.unl.pt

Luís N. Pereira
– Escola Superior de Gestão, Hotelaria e Turismo,
Centro de Investigação sobre o Espaço e as Organizações,
Universidade do Algarve, Portugal
lmper@ualg.pt

Received: May 2010 Revised: April 2011 Accepted: May 2011

Abstract:

- This paper approaches the problem of small area estimation in the framework of spatially correlated data. We propose a class of estimators allowing the integration of sample information of a spatial nature. Those estimators are based on linear models with spatially correlated small area effects where the neighbourhood structure is a function of the distance between small areas. Within a Monte Carlo simulation study we analyze the merits of the proposed estimators in comparison to several traditional estimators. We conclude that the proposed estimators can compete in precision with competitive estimators, while allowing significant reductions in bias. Their merits are particularly conspicuous when analyzing their conditional properties.

Key-Words:

- *combined estimator; empirical best linear unbiased prediction; small area estimation; spatial models; unit level models.*

AMS Subject Classification:

- 62D05, 62F40, 62J05.

1. INTRODUCTION

Sample survey data are extensively used to provide reliable direct estimates of parameters of interest for the whole population and for domains of different kinds and sizes. When the domains were not originally planned, they usually are poorly represented in the sample or even not represent at all. These domains are called small areas and they usually correspond to small geographical areas, such as a municipality or a census division, or a small subpopulation like a particular economic activity or a subgroup of people obtained by cross-classification of demographic characteristics. Traditionally, sample sizes are chosen to provide reliable estimates for large domains and the lack of sample data from the target small area seriously affects the precision of estimates obtained from area-specific direct estimators. This fact has given rise to the development of various types of estimators that combine both the survey data for the target small areas and auxiliary information from sources outside the survey, often related to recent censuses and current administrative data, in order to increase precision. Under this context, the use of indirect estimators has been extensively applied. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through auxiliary data.

Although traditional indirect estimators based on implicit models, which include synthetic and composite estimators, are easy to apply, they usually present undesirable properties. For that reason, other model based methods of small area estimation have been suggested in the literature. These methods can make specific allowance for local variation through complex error structures in the models that link the small areas, can be validated from the sample data and can handle complex cases such as cross-sectional, time series and spatial data. Such methods are often based on explicit Linear Mixed Models. The Best Linear Unbiased Prediction (BLUP) approach, using Henderson's method ([13]), is the most popular technique for estimating small area parameters of interest (usually the mean or the total). Under this approach and from the model point of view the small area parameters of interest are functions of fixed ($\boldsymbol{\beta}$) and random ($\mathbf{u}$) effects. Consequently, the prediction of small area parameters of interest is based on the estimation/prediction of these model effects. In practice this type of models always involves unknown variance components in the variance-covariance structure of random effects. When these unknown components are substituted by consistent estimates the resulting estimator is usually named as Empirical Best Linear Unbiased Predictor (EBLUP).

In the context of unit level spatial data, little work has been done on model-based methods of small area estimation. [25], [26] and [5] proposed a spatial unit level random effects model with spatial dependence incorporated in the error structure through a simultaneous autoregressive (SAR) error process.

Other findings are due to [4], [36], [30], [31], [32] and [40]. All these approaches consider a contiguity matrix to describe the neighbourhood structure between small areas. Nevertheless, there has been a lack of work regarding the explicit modeling of spatial correlation as a function of the distance between observations or small areas.

The main aim of this paper is to propose an approach to the problem of small area estimation in circumstances in which the sample data are of a spatial nature (or in other contexts in which it is possible to establish some kind of proximity between the domains of study), using an estimator that explicitly consider spatial correlation as a function of distance between small areas of study. This estimator, applicable to unit level data, exploits both auxiliary information relating to other known variables on the population and structures of spatial correlation between the sample data through the specification of an adequate non-diagonal structure for the variance-covariance matrices of random effects. It is based on a general class of models that includes some of the existing models as special cases and can be understood as an EBLUP of the small area totals. Consequently, it does not require the specification of a specific prior distribution for model random effects. We also aim to evaluate this estimator in comparison with traditional synthetic and composite estimators that do not explicitly consider spatial variability. The paper is organized under five sections. Section 1 introduces the context of the small area estimation and the goals of the paper. Section 2 reviews some traditional indirect estimators. Section 3 proposes an EBLUP estimator for small area totals based on spatial unit level data. The estimator is assisted by a class of models that fits into the general linear mixed theory. Section 4 describes the design of the Monte Carlo simulation study and presents empirical results. This study analyzes the performance of the proposed estimator over the direct and indirect estimators using a real data set from an agricultural survey conducted by the Portuguese Statistical Office. Discussion of the main findings of this study, along with some of its limitations and possible future developments are the subject of Section 5.

2. INDIRECT ESTIMATORS

One possible approach for "borrow information" in the context of small area estimation based on implicit models is to use direct modified estimators. These estimators maintain certain design-based properties such as approximately unbiased. This is the case of the regression estimator ([37])

$$\hat{\tau}_{d,\text{reg}} = \hat{\tau}_d + (\boldsymbol{\tau}_{\mathbf{x}d} - \hat{\boldsymbol{\tau}}_{\mathbf{x}d})' \hat{\boldsymbol{\beta}} , \qquad d = 1, ..., D , \tag{2.1}$$

where $\hat{\tau}_d$ is an estimator of the d^{th} domain total of the interest variable, usually the Horvitz–Thompson or a post-stratified estimator and $\hat{\boldsymbol{\tau}}_{\mathbf{x}d}$ have the same

meaning in relation to the vector of auxiliary variables $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$, $\hat{\boldsymbol{\beta}} = \left[\sum_{d\in U}\sum_{i\in s_d} \upsilon_i^{-2}\pi_i^{-1}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\sum_{d\in U}\sum_{i\in s_d}\upsilon_i^{-2}\pi_i^{-1}\mathbf{x}_i\, y_i$ are estimators of the regression coefficients obtained using data from the whole sample, υ_i^2 are regression weights and π_i the inclusion probabilities resulting from the sampling design. Estimator (2.1) is approximately unbiased, since $E(\hat{\tau}_{d,\mathrm{reg}}) = \tau_d + \boldsymbol{\tau}_{\mathbf{x}d}' E(\hat{\boldsymbol{\beta}}) - E(\hat{\boldsymbol{\tau}}_{\mathbf{x}d}'\boldsymbol{\beta}) \approx \tau_d$ (supposing $E\left[(\hat{\boldsymbol{\tau}}_{\mathbf{x}d} - \boldsymbol{\tau}_{\mathbf{x}d})'(\hat{\boldsymbol{\beta}} - \mathbf{b})\right] \ll \boldsymbol{\tau}_{\mathbf{x}d}\,\mathbf{b}$, where $\mathbf{b} = E(\hat{\boldsymbol{\beta}})$ from the design-based perspective[1]). Although using information from outside the domain for estimating the regression coefficients, usually these estimators still show low precision.

An alternative is the synthetic estimation (whose properties depend on the assumptions of a postulated model). From the design-based point of view these estimators can be biased and inconsistent. A synthetic regression estimator can be presented as:

$$\hat{\tau}_{d,\mathrm{sreg}} = \boldsymbol{\tau}_{\mathbf{x}d}'\,\hat{\boldsymbol{\beta}}\ , \qquad d = 1, ..., D\ , \tag{2.2}$$

where $\hat{\boldsymbol{\beta}}$ is obtained as before. A more extreme attitude under a pure model-based approach would ignore the inclusion probabilities in estimating the regression parameters. The design-based bias of estimator (2.2) is $B(\hat{\tau}_{d,\mathrm{sreg}}) \approx \boldsymbol{\tau}_{\mathbf{x}d}'(\mathbf{b} - \mathbf{b}_d)$, assuming the regression weights are such that $\upsilon_i^2 \propto \sum_{j=1}^{p} a_j\, x_{ij},\ i \in U_d$, where $a_j,\ j = 1, ..., p$, are arbitrary constants. This condition is always assured in the most typical situation of a non-weighted regression where the parameters υ_i^2 are assumed constant. Further, typically estimator (2.2) has smaller variance than the direct modified regression estimator but it is biased from the design-based point of view. When dealing with small areas the reduction in variance associated with the synthetic estimator can be such that it will assure a mean square error (MSE) lower than the one obtained through the use of the direct modified estimator. There will always be a risk of a high bias and consequently the invalidity of any confidence intervals obtained under repeated sampling. A significant advantage of synthetic estimation lies in the fact that it is always possible to obtain domain estimates, even in situations where the sample is very small or even zero. In order to prevent the quality of the estimator being totally dependent on the postulated model, some combined or composite estimators have been proposed. A combined estimator typically presents the form of the weighted average of a design-based estimator (approximately unbiased but with high variance) and a synthetic estimator (biased but with low variance):

$$\hat{\tau}_{d,\mathrm{com}} \ = \ \lambda_d\,\hat{\tau}_{d,\mathrm{des}} + (1 - \lambda_d)\,\hat{\tau}_{d,\mathrm{syn}}\ , \qquad d = 1, ..., D\ , \tag{2.3}$$

with $0 \le \lambda_d \le 1$. These estimators can be classified in two main types (according to the way the weights λ_d are chosen): sample-size dependent weights and data dependent weights. It is also possible to assume that the weights are chosen in

[1] This condition supposes there is a sufficiently week correlation between $\hat{\boldsymbol{\tau}}_{\mathbf{x}d}$ and $\hat{\boldsymbol{\beta}}$ what is usually easily achieved as $\hat{\boldsymbol{\tau}}_{\mathbf{x}d}$ and $\hat{\boldsymbol{\beta}}$ are estimated at different aggregation levels.

a deterministic way using, for example, some previous knowledge or an informed guess. This would result in what [39] call weights fixed in advance. A good example of a combined regression estimator where the weights depend on sample size is the dampened regression estimator ([38]):

$$\hat{\tau}_{d,\text{dreg}} = \lambda_d \hat{\tau}_{d,\text{reg}} + (1-\lambda_d)\hat{\tau}_{d,\text{sreg}} , \qquad d = 1, ..., D , \tag{2.4}$$

where $\lambda_d = 1$ if $\hat{N}_d \geq N_d$ and $\lambda_d = 0$ otherwise, where N_d is the d^{th} domain population size and h is a positive constant. The authors suggested to use $h = 2$. The basic idea for choosing h is to assure that the bias contribution from the synthetic component of the estimator is kept within acceptable limits. Another possible approach for "borrow information" in the context of small area estimation is to use a data-dependent combined estimator, through the modeling of the bias of the synthetic part of the estimator, thus producing indirect estimates for the weights. Many of the models that have been proposed include random area effects and can be seen as particular cases of linear mixed models. One of the best known models applicable at unit level is the nested error regression model ([8], [3]). All these approaches implicitly consider some kind of sectional correlation and the domain estimators are obtained through EBLUP, empirical Bayes or hierarchical Bayes approaches. The well-known nested error regression model ([3]) has the form $y_{di} = \mathbf{x}'_{di}\boldsymbol{\beta} + u_d + \epsilon_{di},\ d = 1, ..., D,\ i = 1, ..., N_d,$ where u_d and ϵ_d are assumed to be iid with zero means. It is also assumed that u_d and ϵ_{di} are mutually independent, $V_m(u_d) = \sigma_u^2$ and $V_m(\epsilon_{di}) = \sigma^2 \upsilon_{di}^2$ where υ_{di} are known constants. Here a common covariance between any two observations in the same small area is assumed, $\text{Cov}_m(y_{di}, y_{dj}) = \sigma_u^2\ (i \neq j)$. In this kind of models it is assumed that there is no sample selection bias, resulting that they are assumed to hold both for the population and for the sample. This may be a very limiting assumption since small domain estimation is frequently needed in the context of informative sampling designs.

An alternative to the EBLUP in the context of informative sampling designs is the pseudo-EBLUP estimator ([29]). This estimator, based on the nested error regression model, depends on survey weights and it is design-consistent.

3. A COMBINED ESTIMATOR FOR SPATIAL DATA

3.1. A class of models

Let y_{di} be the value of the interest variable for unit i $(i = 1, ..., n_d)$ in small area d $(d = 1, ..., D)$ and let $\mathbf{x}'_{di} = (x_{di1}, ..., x_{dip})$ a vector of p unit level explana-

tory variables referring to the same unit. Consider the following class of models:

$$(3.1) \quad y_{di} = \mathbf{x}'_{di}\boldsymbol{\beta} + \sum_{h=1}^{H} \mathbf{x}'_{(1)di}\, q_{h,di}\, \mathbf{u}_h^{(1)} + \mathbf{x}'_{(2)di}\, \mathbf{u}_d^{(2)} + \epsilon_{di}\,, \quad d = 1, ..., D, \;\; i = 1, ..., n_d\,,$$

where $\boldsymbol{\beta}$ is a vector of p fixed effects; $\mathbf{x}'_{(j)di}$ is a vector of p_j explanatory variables (typically a subvector of $\mathbf{x}'_{di}$) for the i^{th} unit in small area d; $q_{h,di}$ are design variables used to take into account the sampling design and indicate that unit di belongs to a stratum or a sampling unit h ($h = 1, ..., H$); $\mathbf{u}_h^{(1)} = \text{col}_{1 \le j \le p_1}(u_{hj})$ is a vector of p_1 random (or fixed) design effects associated with stratum (or sampling unit) h; $\mathbf{u}_d^{(2)} = \text{col}_{1 \le j \le p_2}(u_{dj})$ is a vector of p_2 random effects associated with domain d; l_d represents the geographical location associated to the centroid of domain d; $f(l_d - l_e)$ is a function of the vector $l_d - l_e$ and ϵ_{di} is the residual term associated with unit di. We assume that $E_m\big(\mathbf{u}_h^{(1)}\big) = \mathbf{0}$, $E_m\big(\mathbf{u}_d^{(2)}\big) = \mathbf{0}$, $E_m(\epsilon_{di}) = 0$, $E_m\big(\mathbf{u}_h^{(1)}\mathbf{u}_g^{\prime(1)}\big) = \begin{cases} \boldsymbol{\Sigma}^{(1)}, & h = g \\ 0, & \text{otherwise} \end{cases}$, $E_m(\epsilon_{di}\,\epsilon_{ej}) = \begin{cases} \sigma^2_{di}, & d = e,\; i = j \\ 0, & \text{otherwise} \end{cases}$, $E_m\big(\mathbf{u}_d^{(2)}\mathbf{u}_e^{\prime(2)}\big) = \boldsymbol{\Sigma}^{(2)} f(l_d - l_e)$, with $\boldsymbol{\Sigma}^{(1)} = \left\{\sigma^2_{U_{jk}^{(1)}}\right\}$ $(j, k = 1, ..., p_1)$, $\sigma^2_{U_{jk}^{(1)}} = E(u_{hj}\, u_{hk})$, $\boldsymbol{\Sigma}^{(2)} = \left\{\sigma^2_{U_{jk}^{(2)}}\right\}$ $(j, k = 1, ..., p_2)$, and $\sigma^2_{U_{jk}^{(2)}} = E(u_{dj}\, u_{dk})$. The model is applied to data from a sample of total size $n = \sum_{d=1}^{D} n_d$, where n_d is the number of sampling units in area d. It is also assumed that random effects associated with different aggregation levels are not correlated, $E\big(\mathbf{u}^{(j)}\,\mathbf{u}'^{(k)}\big) = \mathbf{0}$ for $j \neq k$, and that the errors are non-correlated with random effects, $E\big(\mathbf{u}^{(j)}\boldsymbol{\epsilon}'\big) = \mathbf{0}$, $\forall j$.

In the proposed model, domain effects show a structure of spatial variability. The covariance between the random effects associated with domains d and e depends on the vector defined by their geographical coordinates $l_d - l_e$. Some functions that can be applied to this context are presented in [28]. When the spatial covariance only depends on the distance $|l_d - l_e|$, then the function $f(|l_d - l_e|)$ is said to be isotropic ([6]) and is typically such that $\lim_{|l_d - l_e| \to 0} f = 1$. As the domains are not points in space, but areas, these coordinates are defined by their centroids. The assumption is that the lowest level of aggregation for which the georeferencing is available is the domain. Situations where the level of aggregation for which the referencing is available does not coincide with the domains of study can generate special cases of model (3.1). In particular, when georeferencing is possible at unit level, spatial variation can be modeled through variances-covariances of the errors vector $\boldsymbol{\epsilon}$.

Domain random effects represent the characteristics specific to the domain of study that affect the values of the interest variable and are not represented by the fixed effects at a higher level of aggregation. They can be thought of as modeling the bias of the synthetic part of the model. Moreover, these domain random effects will now have the additional role of bringing information from other domains, to explain the values of the interest variable in each domain.

Design effects are used to take into account the sampling design. The goal is to allow the model to be applied to contexts with informative sampling designs, overcoming the limitations ([27], [20]) of other data-dependent combined estimators that implicitly assume that the sampling design is ignorable.

The methodology proposed can therefore be seen as model assisted. The sample s is the result of a two-step procedure. First it is supposed that the finite population can be approximately described by a superpopulation model. In the second step it is assumed that a sample is drawn from the finite population through a specific sampling design. It is assumed that the sample can be approximately described by model (3.1), which has taken into account the existence of these two steps.

3.2. Estimation of model parameters

The model 3.1 can be presented as a special case of the general linear mixed model, grouping the unit-specific models over the population:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \ , \tag{3.2}$$

where $\mathbf{y}$ is a vector of the target variable, $\mathbf{X}$ is a design matrix of explanatory variables with rows given by $\mathbf{x}'_{di}$, $\mathbf{Z} = \left[\mathbf{Z}_{(1)}\, \mathbf{Z}_{(2)}\right]$ is a design matrix, $\mathbf{u} = \mathrm{col}_{1\leq j\leq 2}\left(\mathbf{u}^{(j)}\right)$ is a vector of random effects and $\boldsymbol{\epsilon}$ is a vector of errors. The covariance matrix of $\mathbf{u}$ is given by $\mathbf{G} = V_m(\mathbf{u}) = \mathrm{blockdiag}_{1\leq j\leq 2}\left[\mathbf{G}^{(j)}\right]$, where $\mathbf{G}^{(1)} = \mathrm{blockdiag}_{1\leq h\leq H}\left\{\boldsymbol{\Sigma}^{(1)}\right\}$ and $\mathbf{G}^{(2)} = \mathbf{F}\otimes\boldsymbol{\Sigma}^{(2)}$ with $\mathbf{F} = \left\{f(l_d - l_e)\right\},\ d,e = 1,...,D$. Further, $\mathbf{R} = V_m(\boldsymbol{\epsilon}) = \mathrm{diag}_{\substack{1\leq i\leq n_d \\ 1\leq d\leq D}}\{\sigma^2_{di}\}$, $E_m\left(\mathbf{u}^{(1)}\right) = \mathbf{0}$, $E_m\left(\mathbf{u}^{(2)}\right) = \mathbf{0}$ and $E_m(\boldsymbol{\epsilon}) = \mathbf{0}$. Both covariance matrices $\mathbf{G}$ and $\mathbf{R}$ involve unknown variance components, represented by $\boldsymbol{\theta}$. Also the flexibility of the proposed class of models recommends proceeding in each application to the selection of a specific model, i.e. to the choice of the explanatory variables and appropriate variance-covariance structures to $\mathbf{u}$ and $\boldsymbol{\epsilon}$. This step in model selection and diagnosis is crucially important to obtaining a model that can adequately describe the behavior of the target population and can be performed as a systematic procedure like those proposed by [7] or [43].

Once a specific model has been selected, the variance components $\boldsymbol{\theta}$ need to be estimated in order to assess the variability of estimators or to predict the fixed and random effects. Several methods are available for estimating variance components, such as the analysis of variance (ANOVA) method ([12]), the minimum norm quadratic unbiased estimation (MINQUE) method ([33], [34], [35]) and the likelihood methods. Some references about the maximum likelihood estimation (MLE) method due to Fisher may be found in [9], [1], [23], [21], [15] and [18]. On the other hand, references about the residual maximum likelihood estimation

(RMLE) method proposed by [41] and its extensions can be found in [24], [10], [11], [2], [42], [14], [16], [17], among others. For details about estimation general linear mixed models see [19].

Now consider the decomposition of all matrices into sample and non-sample components, where the subscript s is associated with the n sample units and r is associated with the $(N-n)$ non-sample units. The omission of the subscript indicates that the respective matrices allude to the whole population $U \equiv s \cup r$. Assuming model 3.2 holds and variance components are known, the best linear unbiased estimator of $\boldsymbol{\beta}$ and the best linear unbiased predictor of $\boldsymbol{\theta}$ are given by

$$\widetilde{\boldsymbol{\beta}} = \left(\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{y}_s \,, \tag{3.3}$$

$$\widetilde{\mathbf{u}} = \mathbf{G}\, \mathbf{Z}_s' \mathbf{V}_s^{-1} \left(\mathbf{y}_s - \mathbf{X}_s \widetilde{\boldsymbol{\beta}}\right) , \tag{3.4}$$

where $\mathbf{V}_s = E\left[(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})\,(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\right] = \mathbf{Z}_s\,\mathbf{G}\,\mathbf{Z}_s' + R_{ss}$. The vectors $\mathbf{u}^{(j)}$ may be predicted using $\widetilde{\mathbf{u}}^{(j)} = \mathbf{G}^{(j)}\,\mathbf{Z}_{(j)s}'\,\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}})$, while the predictors of the errors $\boldsymbol{\epsilon}$, can be obtained as $\boldsymbol{\epsilon} = \mathbf{R}_{.s}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}})$, where $\mathbf{R}_{.s} = [\mathbf{R}_{ss}'\,\mathbf{R}_{rs}']'$.

When the covariance matrix $\mathbf{R} = \begin{bmatrix} \mathbf{R}_{ss} & \mathbf{R}_{sr} \\ \mathbf{R}_{rs} & \mathbf{R}_{rr} \end{bmatrix}$ is block-diagonal, i.e. when there is no correlation between errors associated with the observations inside and outside the sample, then $\mathbf{R}_{rs} = 0$ and $\widetilde{\boldsymbol{\epsilon}}_r = \mathbf{0}$. This is the case for model (3.1). Nevertheless, it should be noted that some situations can be devised, particularly when the spatial correlation can be established at unit level, where there is a correlation between model errors that can be used in the prediction of $\boldsymbol{\epsilon}_r$.

3.3. Estimation of domain totals

The objective of the inference can be seen as to predict the total of an interest variable, τ_d, that under the model corresponds to the summation of the realizations of the variable of interest over all the elements in the small area d:

$$\tau_d = \sum_{i \in U_d} y_{di} = \boldsymbol{\tau}_{\mathbf{x},d}'\boldsymbol{\beta} + \sum_{h=1}^{H} \boldsymbol{\tau}_{\mathbf{x}(1),hd}'\,\mathbf{u}_h^{(1)} + \tau_{\mathbf{x}(2),d}'\,\mathbf{u}_{ad}^{(2)} + \tau_{\epsilon,d} \,, \tag{3.5}$$

where $\tau_{\epsilon,d} = \sum_{i \in U_d} \epsilon_{di}$. It should be noted that, from the model-based point of view, (3.5) is a predictable function producing inference in the narrow inference space [22]. An estimator for the small area total, τ_d, can be obtained as

$$\widetilde{\tau}_d = \mathbf{1}_{N_d}'\widetilde{\mathbf{y}}_d = \sum_{i \in U_d} \widetilde{y}_{di} = \boldsymbol{\tau}_{\mathbf{x}d}'\widetilde{\boldsymbol{\beta}} + \mathbf{v}_{\boldsymbol{\tau}s}'\mathbf{V}_s^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\right) , \tag{3.6}$$

where $\widetilde{\mathbf{y}}_d$ is the EBLUP of the vector $\mathbf{y}_d$ and $\mathbf{v}_{\boldsymbol{\tau}s}' = \boldsymbol{\tau}_{\mathbf{z}d}'\,\mathbf{G}\,\mathbf{Z}_s' + \mathbf{1}_{N_d}'\mathbf{R}_{d\cdot,s}$ is the line vector of the model-based covariances between the small area total τ_d and

the observable vector $\mathbf{y}_s$, $\mathbf{R}_{d\cdot,s} = E(\boldsymbol{\epsilon}_{ad}\boldsymbol{\epsilon}_s)$, and $\mathbf{1}'_{N_d}$ is a unit vector of size N_d. It should be noted that the estimator $\widetilde{\tau}_d$ is the EBLUP of τ_d, given the observable random vector $\mathbf{y}_s$ (cf. Appendix 1).

When $\mathbf{R}_{d,rs}$ is a null matrix, then the EBLUP of the total is τ_d is given by a simplified expression:

$$\widetilde{\tau}_d \;=\; \widetilde{E}(\tau_{d,r}|\mathbf{u}) \;=\; \tau_{\mathbf{y},d,s} + \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\tau}'_{\mathbf{z},d,r}\,\mathbf{G}\,\mathbf{Z}'_s\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\;, \tag{3.7}$$

where $\tau_{\mathbf{y},d,s}$ is the observed sample total in small area d (cf. Appendix 2). It should also be noted that many of the regression estimators that have been proposed for small area estimation may be viewed as EBLUP of domain totals for particular cases of the class of models (3.1). For instance, the form of the nested error regression model and the random coefficient model presented in Section 2 accord with class (3.1), with $\mathbf{u}_d^{(2)}$ scalar, $\mathbf{G}^{(1)} = \mathbf{0}$, $\mathbf{G}^{(2)} = \sigma_u^2\,\mathbf{I}_D$ and $\mathbf{R} = \sigma^2\,\mathbf{I}_n$. Also, the model underlying the synthetic regression estimator (2.2) is equivalent to considering (3.1) with $\mathbf{u}_h^{(1)}$ scalar and taken as a fixed effect, $\mathbf{G}^{(2)} = \mathbf{0}$ and $\mathbf{R} = \sigma^2\,\mathbf{I}_n$. Moreover, the direct modified regression estimator (2.1), can be obtained considering $\mathbf{u}_h^{(1)}$ and $\mathbf{u}_d^{(2)}$ scalars and taken as a fixed effects and $\mathbf{R} = \sigma^2\,\mathbf{I}_n$.

3.4. Domains not represented in the sample

Situations may arise where some domains are not represented in the sample. If no sample falls into small area d, then the respective random effects $\mathbf{u}_d^{(2)}$ may still be predicted if there is covariance between $\mathbf{u}_d^{(2)}$ and at least one of the small area random effects represented in the sample $\mathbf{u}_e^{(2)}$ $(e = 1, ..., D;\; e \neq d)$. We have then

$$\widetilde{\mathbf{u}}_d^{(2)} \;=\; \mathbf{G}_{d,\cdot}^{(2)}\,\mathbf{Z}'_{(2)s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\;, \tag{3.8}$$

where $\mathbf{G}_{d,\cdot}^{\prime(2)} = \underset{\substack{n_e\neq 0\\ 1\le e\le D}}{\mathrm{col}}\big[\mathbf{G}_{d,e}^{\prime(2)}\big] = E\big[\mathbf{u}^{(2)}\mathbf{u}_d^{\prime(2)}\big]$ and $\mathbf{G}_{d,e}^{(2)} = \boldsymbol{\Sigma}^{(2)} f(l_d - l_e)$. In an extreme situation where the small area effects $\mathbf{u}_d^{(2)}$ are not correlated with any other small area effect for a domain represented in the sample, i.e. when $\mathbf{G}_{d,e}^{(2)} = 0$, $\forall\, e \neq d\colon n_e \neq 0$, then $\widetilde{\mathbf{u}}_d^{(2)} = \mathbf{0}$. The estimator $\widetilde{\tau}_d$ is then reduced to a form similar to a following synthetic estimator:

$$\widetilde{\tau}_d \;=\; \tau_{y,d,s} + \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \sum_{h=1}^{H}\boldsymbol{\tau}'_{\mathbf{x}(1),dh,r}\,\widetilde{\mathbf{u}}_h^{(1)}\;. \tag{3.9}$$

It may be noted that this estimator may be written in the same generic form

$$\hat{\tau}_d \;=\; \boldsymbol{\tau}'_{\mathbf{x},d}\hat{\boldsymbol{\beta}} + \mathbf{f}'\big(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}\big)\;, \tag{3.10}$$

where $\mathbf{f}$ is used to weight the regression residuals. This form puts in evidence that estimator (3.9) can be seen as a combined estimator where the weights in $\mathbf{f}'$ allow a correction of the synthetic part of the estimator $\boldsymbol{\tau}'_{\mathbf{x},d}\widetilde{\boldsymbol{\beta}}$ through the prediction errors in the domain that is the target of inference, but also in other domains spatially correlated. When no correlation between domains is specified, the correction factor depends only on the prediction errors in the target small area and the estimator is reduced to a similar form to the data-dependent combined estimators presented in Section 2.

These characteristics seem to be particularly interesting when estimating in small domains, where the available sample size is small, since it borrows information from outside the domain of study in order to assist the estimation. Moreover, taking advantage of the potential spatial correlation of data it is possible to avoid the reduction of the proposed estimators to pure synthetic estimators even when the sample size in the domain is null.

4. MONTE CARLO SIMULATION STUDY

4.1. Generation of the pseudo-population

For the simulation a pseudo-population is used. This population is obtained from a real data set containing the responses to the 1993 wave of the Agricultural Structure Survey. It is an agricultural survey conducted by the Portuguese Statistical Office in the period between agricultural censuses. The responses for the variable total production of cereals were extracted and circumscribed to the NUTSII of *Alentejo*. The total sample size in this region is 7,060 and the population size 47,049. The design for the Agricultural Structure Survey is based on stratified sampling. The sample is first stratified using the *região agrária* as the level for geographic stratification. A *região agrária* is an administrative division used for agricultural purposes. In each *região agrária* a new stratification is established based on Used Agricultural Surface (UAS) classes. In the same *região agrária* some other strata are defined based on the value of other variables considered weakly correlated with UAS. In *Alentejo* there are 19 strata.

For simulation purposes a pseudo-population is generated by replicating the agricultural establishments in the sample proportionally to the inverse of their inclusion probabilities. The sampling frame resulting from this replication includes the value of production of cereals for each establishment in 1993, and the same value reported for 1989 (year of the agricultural census). The production in 1989 is used as an auxiliary variable in the models used in the simulation. Also, geographical coordinates associated with the centroids of *freguesias* were recorded.

This was the lowest level of aggregation for which geographical referencing was available. This means that geographical differentiation between establishments included in the same *freguesia* is not available. A *freguesia* is an administrative division that segments the *Alentejo* into 284 sub-regions.

4.2. Description of the simulations

Using the sampling frame corresponding to the pseudo-population of agricultural establishments we have run a Monte Carlo simulation. The goal is to evaluate the design-based properties of a set of alternative estimators. Note that the approach followed in this paper is to evaluate the properties and relative merits of the proposed estimators through simulation. In fact, due to the complexity of these estimators their design-based properties (e.g. bias, variance) are impossible to obtain through analytical methods. Also, their model-based properties would be of limited interest from the point of view of a benchmark with alternative direct estimators used in this simulation whose properties only make sense to evaluate from the design-based perspective. The target parameter is the total of the variable production of cereals at *freguesia* level. The number of simulations performed is 560. In each simulation a sample is drawn from the pseudo-population U^*, using a stratified design similar to the one used in the Agricultural Structures Survey. The only difference in relation to that survey design is that the sample size by stratum was reduced to 30% of the original size (2,118 establishments). The goal is to simulate a framework similar to that survey, but with a smaller sample size, enabling the evaluation of the estimators' behavior in "critical" situations where the domain sample size is very small (sometimes only a few units or even none). This sampling design leads to a relative precision of 7.5% (for a 95% confidence-level) in the estimation of the total of the interest variable at the population level, using the Horvitz–Thompson estimator. The expected sample sizes for the 284 domains of interest vary from 0.3 to 45.8 units.

4.3. Estimators

The estimators analyzed in the simulation are presented in this section. They are mainly implementations of the direct, synthetic and combined regression estimators presented in Sections 2 and 3. It should be noted that all the regression estimators include the same auxiliary variables (associated with the fixed effects), allowing a fair comparison of their relative merits. In what follows, the notation ad is used to represent the small area d of region a, where the regions correspond to the level of aggregation of NUTSIII and the small area of interest to *freguesia*. Table 1 summarizes the estimators used in the simulation.

Table 1: Estimators used in the simulation study.

Estimator	Description
$\hat{\tau}_{ad1}$	Horvitz–Thompson estimator
$\hat{\tau}_{ad2}$	Direct modified regression estimator (2.1)
$\hat{\tau}_{ad3}$	Dampened regression estimator (2.4)
$\hat{\tau}_{ad4}$	Pure synthetic regression estimator (2.2) with fixed effects estimated ignoring the sampling design
$\hat{\tau}_{ad5}$	Synthetic regression estimator where the sampling design is explicitly considered through the inclusion of a vector of design variables E_{hi} indicating the belonging of each establishment i to the strata $h=1,...,H$
$\widetilde{\tau}_{ad6}$	Data-dependent combined regression estimator based on the nested error unit level regression model
$\widetilde{\tau}_{ad7}$	Data-dependent combined regression estimator, similar to $\widetilde{\tau}_{ad6}$ but including fixed strata effects β_{h0}, $h=1,...,H$
$\widetilde{\tau}_{ad8}$	Data-dependent combined regression estimator, based on a model included in the proposed class of models (3.1), with random small area effects presenting a spatial covariance structure following an isotropic exponential model.

The isotropic exponential model used to represent spatial variability in $\hat{\tau}_{ad8}$ was suggested in the model diagnosis phase. We have tested several structures (exponential, spherical, linear, log-linear and Gaussian), through the evaluation the significance of covariance parameters (using Wald tests) and information criteria (such as AIC and BIC). Among the structures that showed statistical significance we retained the one that minimized the several information criteria. Although we have chosen the exponential model, some of the other structures resulted in very similar adjustments. Also note that for the data-dependent regression estimators the variance components are estimated through REML method. The only exception regards estimator $\widetilde{\tau}_{ad8}$, where the parameter c_e was estimated *a priori* through the adjustment of an exponential semivariogram to an empirical semivariogram.

Note that the estimators included in the simulation vary in nature: $\hat{\tau}_{ad1}$ and $\hat{\tau}_{ad2}$ are design-based estimators, $\hat{\tau}_{ad4}$ and $\hat{\tau}_{ad5}$ are synthetic estimators, while the others can be classified as combined estimators as described in previous sections. The included estimators also differentiate in the way the sampling design is (or not) taking into account: in $\hat{\tau}_{ad1}$, $\hat{\tau}_{ad2}$ and $\hat{\tau}_{ad3}$ the sampling design information is taking into account using sampling weights, $\hat{\tau}_{ad5}$, $\hat{\tau}_{ad7}$ and $\hat{\tau}_{ad8}$ include fixed strata effects, while in $\hat{\tau}_{ad4}$ and $\hat{\tau}_{ad6}$ the sampling design information is ignored.

4.4. Precision and bias measures

The estimators under consideration are evaluated using a set of precision and bias measures. In what follows K represents the number of simulations, and $\hat{\tau}_{kd}$ the d^{th} small area estimate of the total obtained from the simulation k $(k = 1, ..., K)$.

4.4.1. Unconditional analysis

Taking into account the high number of small areas in the population (284) and in order to facilitate the presentation of the simulation results, the small areas are divided into six groups. Each group g contains D_g small areas. Table 2 presents the definition of each group and the number of small areas involved.

Table 2: Small area groups in the simulation study.

Group	Expected sample size	Number of small areas
0	—	20
1	$[0; 2]$	20
2	$[2; 3.5]$	43
3	$[3.5; 5]$	49
4	$[5; 10]$	87
5	$[10; +\infty]$	65

Groups 1 to 5 were defined according to the expected sample size of the small areas. Group 0 includes small areas for which the total of the interest variable is zero (*freguesias* where there is no cereal production) regardless their size. The goal is to separate these small areas from the other groups to prevent them from changing the conclusions regarding the relative merits of the estimators. The Monte Carlo relative error for the estimators' expected value is on average 8.0% in group 1 and varies between 3.4% and 4.1% in groups 2 to 5.

For the unconditional analysis the following measures were considered for each group g:

Average absolute bias: $AAB_g = D_g^{-1} \sum_{d=1}^{D_g} AB_d$, where $AB_d = K^{-1} \sum_{j=1}^{K} |\hat{\tau}_{jd} - \tau_d|$;

Average MSE: $AMSE_g = D_g^{-1} \sum_{d=1}^{D_g} MSE_d$, where $MSE_d = K^{-1} \sum_{j=1}^{K} (\hat{\tau}_{jd} - \tau_d)^2$;

Average variance: $AV_g = D_g^{-1} \sum_{d=1}^{D_g} V_d$, where $V_d = K^{-1} \sum_{j=1}^{K} (\hat{\tau}_{jd} - \bar{\hat{\tau}}_d)^2$;

Average absolute bias ratio: $AABR_g = D_g^{-1} \sum_{d=1}^{D_g} ABR_d$, where $ABR_d = AB_d / \sqrt{V_d}$;

Average coverage rate for a design-based $100(1-\alpha)$ confidence interval: $ACR_g = D_g^{-1} \sum_{d=1}^{D_g} TC_d$,

where $TC_d = 100 \times R_d / K$ and R_d represents the number of simulations for which the confidence interval $\hat{\tau}_{jd} \pm t_{\alpha/2} \sqrt{V_d}$ contains the true parameter τ_d .

4.4.2. Conditional analysis

A conditional analysis was also conducted using a set of precision and bias measures for each small area d. The superscript (n_d) indicates that the respective measure is conditioned to the realized sample size, n_d, in small area d. They are:

Conditional relative bias: $CRB_d^{(n_d)} = K_{n_d}^{-1} \sum_{j=1}^{K_{n_d}} (\hat{\tau}_{jd} - \tau_d) / \tau_d$;

Conditional relative standard error: $CRSE_d^{(n_d)} = \sqrt{K_{n_d}^{-1} \sum_{j=1}^{K_{n_d}} (\hat{\tau}_{jd} - \tau_d)^2} / \tau_d$;

Conditional variation coefficient: $CVC_d^{(n_d)} = \sqrt{V_d^{(n_d)}} / \tau_d$,

where $V_d^{(n_d)} = K_{n_d}^{-1} \sum_{j=1}^{K_{n_d}} (\hat{\tau}_{jd} - \bar{\hat{\tau}}_d)^2$ is the conditional variance ;

Conditional bias ratio: $CBR_d^{(n_d)} = B_d^{(n_d)} / \sqrt{V_d^{n_d}}$,

where $B_d^{(n_d)} = K_{n_d}^{-1} \sum_{j=1}^{K_{n_d}} (\hat{\tau}_{jd} - \tau_d)$ is the conditional bias ;

Coverage rate of the conditional design-based confidence interval: $CR_d^{(n_d)} = 100 \times R_d^{(n_d)} / K_{n_d}$,

where $R_d^{(n_d)}$ represents the number of simulations for which the confidence interval $\hat{\tau}_{jd} \pm t_{\alpha/2} \sqrt{V_d^{(n_d)}}$ contains the true parameter τ_d .

4.5. Results

4.5.1. Unconditional analysis

Table 3 summarizes the unconditional results of the simulation study. The values for absolute bias, variance and MSE are presented relatively to the respective value associated with $\hat{\tau}_{ad2}$.

Table 3: Unconditional results.

Group	$\hat{\tau}_{ad1}$	$\hat{\tau}_{ad2}$	$\hat{\tau}_{ad3}$	$\hat{\tau}_{ad4}$	$\hat{\tau}_{ad5}$	$\widetilde{\tau}_{ad6}$	$\widetilde{\tau}_{ad7}$	$\widetilde{\tau}_{ad8}$
Absolute bias								
0	0.00	1.00	2.39	96.54	19.67	91.80	15.50	11.07
1	0.60	1.00	3.44	23.64	10.32	22.57	9.63	9.46
2	1.13	1.00	4.15	22.72	15.14	24.30	12.07	12.59
3	1.33	1.00	4.65	45.75	18.54	44.16	13.72	13.55
4	1.46	1.00	4.40	51.13	20.76	49.14	13.60	14.47
5	1.09	1.00	3.36	77.28	22.51	68.04	14.14	18.14
Variance								
0	0.00	1.00	1.07	0.67	0.03	0.60	0.21	0.29
1	1.10	1.00	0.79	0.03	0.01	0.09	0.06	0.09
2	2.06	1.00	0.76	0.06	0.02	0.26	0.18	0.25
3	1.77	1.00	0.80	0.10	0.03	0.45	0.34	0.40
4	2.01	1.00	0.79	0.19	0.05	1.18	0.94	1.10
5	1.84	1.00	0.85	0.45	0.10	2.29	1.91	2.13
MSE								
0	0.00	1.00	1.08	14.74	0.68	13.18	0.63	0.61
1	1.09	1.00	0.85	0.93	0.40	0.91	0.37	0.39
2	2.06	1.00	0.79	0.92	0.45	1.21	0.47	0.57
3	1.77	1.00	0.82	1.67	0.45	1.89	0.58	0.63
4	2.01	1.00	0.82	3.36	0.63	3.99	1.20	1.46
5	1.84	1.00	0.87	7.88	0.76	7.39	2.27	2.76
Bias ratio								
0	0.01	0.04	0.09	5.39	4.55	5.15	2.29	1.68
1	0.03	0.04	0.19	5.36	6.40	3.86	2.58	1.63
2	0.03	0.03	0.18	3.64	3.70	2.18	1.44	1.26
3	0.03	0.03	0.14	3.45	3.05	2.13	0.85	0.71
4	0.04	0.03	0.15	3.74	3.58	2.11	0.75	0.62
5	0.03	0.03	0.12	3.70	2.60	1.76	0.40	0.42
Coverage Rate								
0	1.00	0.96	0.96	0.01	0.16	0.00	0.59	0.71
1	0.97	0.97	0.95	0.07	0.32	0.21	0.51	0.69
2	0.96	0.96	0.94	0.28	0.37	0.48	0.65	0.70
3	0.96	0.95	0.94	0.23	0.46	0.50	0.82	0.84
4	0.96	0.95	0.95	0.23	0.35	0.48	0.84	0.86
5	0.96	0.95	0.95	0.18	0.44	0.60	0.91	0.91

As expected the two design-based estimators $\hat{\tau}_{ad1}$ and $\hat{\tau}_{ad2}$ are approximately unbiased, even for very small areas. For these estimators the average coverage rate is very near the nominal confidence level (95%). Nevertheless, especially in the smallest domains, they show variance and MSE substantially higher than those observed for the synthetic and combined estimators. It is also to be noted that $\hat{\tau}_{ad2}$ brings significant precision gains when compared with the Horvitz–Thompson estimator.

On the other hand, the two synthetic estimators show very different behavior. $\hat{\tau}_{ad4}$, which can be viewed as a pure synthetic estimator shows disastrous behavior both in terms of bias and precision. These results are clear evidence of the effects of ignoring an informative sampling design. On the other hand, the synthetic estimator with fixed strata effects, $\hat{\tau}_{ad5}$, shows significant precision gains when compared to the direct regression estimator. These precision gains show a tendency to decrease as the expected sample size in the small areas increases. Its major drawback is related to the high bias, which originates average bias ratios that are always above 2.6, compromising the construction of design-based confidence intervals. In fact the average coverage rate for this estimator is always below 0.46, and in many cases near 0.30.

The combined regression estimator with sample-size dependent weights, $\hat{\tau}_{ad3}$, shows a systematic precision gain when compared to the direct regression estimator (with the exception of group 0, the ratio between the average MSE of the two estimators is between 0.79 and 0.87). Nevertheless, these gains are always moderate and substantially lower than the ones observed for the synthetic regression estimator. This estimator also shows a very good behavior in what regards bias. In fact, although having an absolute bias higher than those observed for the direct estimators, the average bias ratio is always lower than 0.2, which originates an average coverage rate very near the nominal confidence level. With regard to the combined estimators with data-dependent weights it can once again be observed that disregard of the sampling design, as in estimator $\widetilde{\tau}_{ad6}$, produces undesirable properties both in terms of bias and precision. In fact, $\widetilde{\tau}_{ad6}$ systematically shows higher MSE than the direct regression estimator (with the exception of group 1). It also exhibits dramatic biases (of the same magnitude as the synthetic estimator $\hat{\tau}_{ad4}$) and bias ratios that on average are situated between 1.76 and 5.15.

The combined estimators $\widetilde{\tau}_{ad7}$ and $\widetilde{\tau}_{ad8}$, which explicitly consider strata effects, show very different behavior. These estimators show average MSE that for the smallest small areas (groups 0 to 3) is near those observed for the best synthetic estimator, while allowing significant reduction in bias and bias ratio.

In particular, $\widetilde{\tau}_{ad7}$ based on a nested error model with fixed strata effects reveals important precision gains when compared to the direct regression estimator, $\hat{\tau}_{ad2}$, or the sample-size dependent regression estimator, $\hat{\tau}_{ad3}$, in groups 0 to 3.

It is to be noted that for groups 4 and 5 (corresponding to expected sample sizes higher than 5 units) $\widetilde{\tau}_{ad7}$ shows some precision loss regarding the direct regression estimator, which is particularly important in group 5. It is also significant that, in the groups 0 to 3, $\widetilde{\tau}_{ad7}$ exhibits a MSE that is similar or even smaller than that associated with the synthetic estimator $\hat{\tau}_{ad5}$. In these groups the increase in variance in $\widetilde{\tau}_{ad7}$ is more than compensated by the reduction in bias. In what regards bias measures, the estimator $\widetilde{\tau}_{ad7}$ shows a behavior that is situated between those recorded for the direct and the synthetic estimators. The average absolute bias ratios vary from 0.40 to 2.58 (increasing with the reduction of the expected sample size) and are strikingly lower than the ones associated with the synthetic estimators (varying between 15% and 50% of those obtained for the best synthetic estimator). This results in average coverage rates for a design-based confidence interval that are substantially higher than those observed for the synthetic estimators.

The estimator $\widetilde{\tau}_{ad8}$ considers a spatial covariance based on an isotropic exponential structure at the small area level. For all small area groups (with the exception of group 0) it shows a small loss of precision when compared to $\widetilde{\tau}_{ad7}$, but still allows important precision gains regarding the direct estimators and the sample-size dependent regression estimator, $\hat{\tau}_{ad3}$, in groups 0 to 3. It is worth noting that this decline in precision is mainly induced by an increase in variance, since $\widetilde{\tau}_{ad8}$ shows average absolute bias that is very near or even smaller than for $\widetilde{\tau}_{ad7}$ (mainly in the smaller areas). The bias ratios for $\widetilde{\tau}_{ad8}$ are, for groups 0 to 4, substantially smaller than those observed for $\widetilde{\tau}_{ad7}$, varying now from 0.42 to 1.68. The reduction in the bias ratio tends to diminish with the increase in the expected sample size, resulting that in group 5 the bias ratio of the two estimators is similar. In fact it is in smaller areas that $\widetilde{\tau}_{ad7}$ approximates more closely a synthetic estimator, allowing the additional sample information used in $\widetilde{\tau}_{ad8}$ (from other spatially correlated small areas) to contribute to bias reduction. Between the combined estimators that allow precision gains in small areas groups 0 to 3, $\widetilde{\tau}_{ad8}$ is the one that shows the best behavior in terms of average bias ratio, which varies from 16% to 37% of those obtained for the best synthetic estimator.

4.5.2. Conditional analysis

Figure 1 and Table 4 summarize the simulation's conditional results for one small area in the study. Considering the large number of small areas in the study (284), these data are only intended to illustrate typical results associated with one of the smallest areas. The results refer to a small area with expected sample size of 4.2 units, thus belonging to group 3.

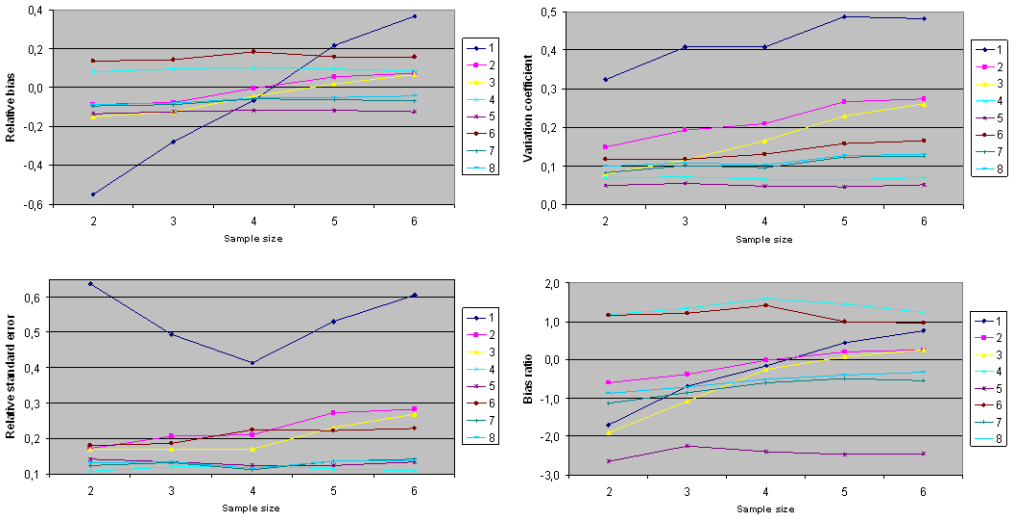

Figure 1: Conditional results.

Table 4: Unconditional coverage rates.

Sample size	$\hat{\tau}_{ad1}$	$\hat{\tau}_{ad2}$	$\hat{\tau}_{ad3}$	$\hat{\tau}_{ad4}$	$\hat{\tau}_{ad5}$	$\widetilde{\tau}_{ad6}$	$\widetilde{\tau}_{ad7}$	$\widetilde{\tau}_{ad8}$
2	0.52	0.90	0.51	0.80	0.31	0.83	0.81	0.84
3	0.88	0.89	0.81	0.74	0.32	0.77	0.90	0.88
4	0.98	0.95	0.95	0.61	0.29	0.72	0.91	0.90
5	0.91	0.96	0.96	0.66	0.35	0.89	0.94	0.94
6	0.88	0.97	0.96	0.78	0.27	0.89	0.92	0.92

The two design-based estimators $\hat{\tau}_{ad1}$ and $\hat{\tau}_{ad2}$ show bad conditional properties, namely in what regards bias. In fact, both estimators show important conditional biases and bias ratios when the effective sample size departs from the expected sample size. This phenomenon is particularly notable in the Horvitz–Thompson estimator. This bias tends to be negative for effective sample sizes that are smaller than expected and positive for expected sample sizes that are larger than expected. Also, when effective sample size is smaller than the expected sample size, the conditional variation coefficients tend to show a rising pattern with the increase in the effective sample size. The combined result of this bias and variance behavior is a conditional relative standard error that for both estimators tends to increase with the effective sample size (for sample sizes above the expected). When the effective sample size is significantly smaller or greater than the expected sample size, these estimators (and mainly $\hat{\tau}_{ad1}$) show a significant degradation in precision. These patterns are particularly notable in the small areas with a very small sample size.

On the other hand, although the synthetic estimators show very high bias and bias ratios (resulting in very low conditional coverage rates for a design-based confidence interval), they are seen to be approximately constant and therefore independent of the effective sample size for each small area. The conditional variation coefficient is clearly constant showing independence from the sample size in the small area. The combined result of these patterns is a relative standard error which is also approximately invariant with the effective sample size. From this conditional point of view, the synthetic regression estimator can still be considered one of the most precise estimators.

For effective sample sizes that are smaller than the expected sample size the combined regression estimator with sample-size dependent weights, $\widetilde{\tau}_{ad3}$, shows important conditional biases and bias ratios that for significant departures are similar to those observed for the synthetic estimators. For sample sizes that are greater than expected $\widetilde{\tau}_{ad3}$ tends to show a behavior similar to the direct regression estimator $\hat{\tau}_{ad2}$. Therefore, the resulting conditional coverage rates for a design-based confidence interval also show a behavior similar to a synthetic estimator for sample sizes that are smaller than expected and similar to the direct regression estimator when they are higher than expected. The relative standard error also tends to show the bad property observed for the direct estimator, characterized by an increase with the effective sample size, mainly for the smallest areas.

The combined estimators $\widetilde{\tau}_{ad7}$ and $\widetilde{\tau}_{ad8}$ show interesting conditional properties as they show a mixed behavior between the direct regression estimator $\hat{\tau}_{ad2}$ and the synthetic regression estimator $\hat{\tau}_{ad5}$. This behavior is characterized by a significant resistance of precision and bias to departures between the effective sample size and the expected sample size.

It can be observed that $\widetilde{\tau}_{ad7}$ shows a conditional bias and bias ratio that are approximately constant, although with a slight tendency to increase with the reduction of the effective sample size. As to bias, this estimator shows a clear advantage when compared to the synthetic estimators and even when compared to the direct estimators and the sample-size dependent combined estimator, particularly when the sample size departs from the expected sample size. This results in conditional coverage rates which in extreme situations are closer to the confidence level than those associated with some design-based estimators. The conditional variance only shows a very slight tendency to rise with an increase in the effective sample size, resulting in a conditional relative standard error that is approximately constant. From the precision point of view it can be seen that this estimator is still competitive with the best synthetic estimator and maintains important precision gains when compared to the direct estimators and the sample-size dependent combined estimator (especially when the effective and expected sample sizes are different).

The estimator $\widetilde{\tau}_{ad8}$ magnifies these bias and bias ratio reductions as it continues to show smaller conditional bias and bias ratios than $\widetilde{\tau}_{ad7}$. Although not seen in this illustrative small area, global results showed that these bias reductions are particularly notable for effective sample sizes smaller than the expected sample sizes. In fact, it is in the smaller areas and particularly when the effective sample sizes are smaller than expected that there is an opportunity to reduce bias by borrowing information from other domains through the use of spatial correlations. The conditional variation coefficient still shows significant resistance to departures from the expected sample size. For sample sizes below the expected the variation coefficient tends to be slightly higher than the one obtained for $\widetilde{\tau}_{ad7}$, and still shows a pattern of a slight increase with the growth of the effective sample size. This increase is now lessened since a part of the variance is due to data provided by small areas in the neighborhood of the target small area. The conditional relative standard error continues to be reasonably constant and not substantially higher than the one obtained for $\widetilde{\tau}_{ad7}$. Overall, it can be concluded that among the combined estimators analyzed $\widetilde{\tau}_{ad8}$ is the estimator that exhibits the best conditional properties for bias and coverage rates for a design-based confidence interval.

5. MAIN FINDINGS AND DISCUSSION

The results of the empirical study show that the combined estimators obtained from the model classes proposed can compete in precision with the best synthetic estimators analyzed, while also allowing large reductions at the level of bias and, particularly, the bias ratio. They manage to show better precision than synthetic estimators for very small domains, and thus provide an important alternative to such estimators. The results attained seem to confirm that the combination of a synthetic and a direct component manages to take into account a significant part of the bias in the purely synthetic estimator, trading it for an increase in variance.

It should be noted that for this population the proposed estimators only prove interesting for inference related to domains of a small expected sample size (up to 5–10 units for the population analyzed). For larger sample sizes they cease to show precision gains in comparison with the best direct estimators (particularly with some direct modified regression estimators).

When the adjusted data displays spatial variability, the estimators that take advantage of the spatial correlation between observations tend to present reductions in bias (and mainly bias ratio) when compared with estimators that ignore this variability. These reductions are usually accompanied by a modest

loss of precision, resulting in bias ratios that are generally substantially lower than those obtained for these other estimators. This fact is easy to understand if we take into account that the consideration of spatial information implies the use of observations that are exogenous to each small area when estimating its random effect. It is natural that the inclusion of such information will also introduce some additional variability in the resulting estimator. The spatial information permits a repositioning of the estimator, which will display behavior that is further away from that presented by a synthetic estimator and gain the characteristics of a direct estimator. It should be pointed out that when the sample size in the inference target domain is very small or even non-existent, the introduction of spatial information relating to other domains can prevent the estimators being reduced to 'pure' synthetic estimators and maintain mixed characteristics between a synthetic and a design-based estimator. This fact helps to explain the good behavior of these estimators in domains with a very small sample size.

The proposed estimators clearly show interesting conditional characteristics, as they tend to behave in a way that is typified by strong robustness, both in precision and bias, to differences in effective and expected sample size. Their remarkable conditional behavior is clearly demonstrated by the fact that their conditional bias ratios are in many cases lower than those registered for direct estimators, specially when there are significant discrepancies between the effective and expected sample size. In particular, estimators that exploit spatial correlation continued to show reductions in conditional bias and conditional bias ratios when compared with estimators that ignore this variability. In fact, we can conclude that while the proposed estimator shows interesting unconditional properties, it is within a conditional point of view that its advantage over competitive estimators strikes.

One of the main limitations of this study lies on the fact that only the specification of isotropic spatial covariance structures was considered. In fact, in a context where the differences between the coast and the hinterland are presumably very different from those between the north and south, resort to anisotropic spatial models can allow the reality to be more satisfactorily represented. However, the sheer complexity of calculation presented by these structures, arising from the need to process a considerable amount of data, rendered the estimation of these models unviable. Although the proposed estimator is thought to be applicable to the context where data of spatial nature is present, it would be interesting to test its application to other contexts, whenever is possible to establish some kind of proximity between the small areas of study.

It should be also stressed that the conclusions presented are depended of the used data set. Although we have used a realistic data set based on real data from a National Statistical Office, the use of different data sets, for example exhibiting different spatial correlation, can lead to different results and possibly different conclusions. Therefore, the proposed estimators should be tested with

other sets of real and artificial data before they are selected for application in other contexts. In fact, empirical studies have revealed to be a fundamental stage in the process of choosing an estimator. The results of such studies can, moreover, help to create greater confidence on the part of potential users of these kinds of estimators.

A. APPENDIX 1

The estimation of τ_d is performed thought the prediction of the realizations of the vector $\mathbf{y}_d$. Under the model (3.2) the EBLUP is:

$$\begin{aligned}\widetilde{\mathbf{y}}_d &= \mathbf{X}_d\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_d\widetilde{\mathbf{u}} + \widetilde{\boldsymbol{\epsilon}}_d \\ &= \mathbf{X}_d\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_d\mathbf{G}\,\mathbf{Z}_s'\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big) + \mathbf{R}_{d,s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big) \\ &= \mathbf{X}_{ad}\widetilde{\boldsymbol{\beta}} + \mathbf{V}_{ad,s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\,,\end{aligned}$$

where the subscript d indicates that the respective matrices only include observations from the small area d, $\mathbf{R}_{d,s} = E(\boldsymbol{\epsilon}_d\boldsymbol{\epsilon}_s)$ and $\mathbf{V}_{ad,s} = E\big[(\mathbf{y}_d - \mathbf{X}_d\boldsymbol{\beta})\,(\mathbf{y}_d - \mathbf{X}_d\boldsymbol{\beta})'\big]$ $= E(\mathbf{Z}_d\,\mathbf{u}\,\mathbf{u}'\,\mathbf{Z}_s') + E(\boldsymbol{\epsilon}_d\,\boldsymbol{\epsilon}_s) = \mathbf{Z}_d\mathbf{G}\,\mathbf{Z}_s' + \mathbf{R}_{d,s}$. With the EBLUP $\widetilde{\mathbf{y}}_d$, the estimator of τ_d may be obtained as:

$$\begin{aligned}\widetilde{\tau}_d = \sum_{i\in U_d}\widetilde{y}_{di} &= \boldsymbol{\tau}_{\mathbf{x}d}'\widetilde{\boldsymbol{\beta}} + \boldsymbol{\tau}_{\mathbf{z},d}'\mathbf{G}\,\mathbf{Z}_s'\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big) + \mathbf{1}_{N_d}'\mathbf{R}_{d\cdot,s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big) \\ &= \boldsymbol{\tau}_{\mathbf{x}d}'\widetilde{\boldsymbol{\beta}} + \mathbf{v}_{\boldsymbol{\tau}s}'\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\,,\end{aligned}$$

where $\mathbf{R}_{d\cdot,s} = E(\boldsymbol{\epsilon}_{ad}\,\boldsymbol{\epsilon}_s)$ and $\mathbf{v}_{\boldsymbol{\tau}s}' = E\big[(\tau_d - \boldsymbol{\tau}_{\mathbf{x},d}'\boldsymbol{\beta})\,(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\big] = \boldsymbol{\tau}_{\mathbf{z},d}'\mathbf{G}\,\mathbf{Z}_s' + \mathbf{1}_{N_d}'\mathbf{R}_{d\cdot,s}$.

B. APPENDIX 2

The vector $\widetilde{\mathbf{y}}_d$ may be decomposed into $\widetilde{\mathbf{y}}_d' = (\widetilde{\mathbf{y}}_{d,s}', \widetilde{\mathbf{y}}_{d,r}')'$. From mixed model theory it is straightforward that $\widetilde{\mathbf{y}}_{d,s} = \mathbf{y}_{d,s}$, with the unobservable part of $\mathbf{y}_d$ predicted by

$$\widetilde{\mathbf{y}}_{d,r} = \mathbf{X}_{d,r}\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_{d,r}\widetilde{\mathbf{u}} + \widetilde{\boldsymbol{\epsilon}}_{d,r} = \mathbf{X}_{d,r}\widetilde{\boldsymbol{\beta}} + \mathbf{V}_{dr,s}\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\,,$$

where $\mathbf{R}_{dr,s} = E(\boldsymbol{\epsilon}_{d,r}\,\boldsymbol{\epsilon}_s)$ and $\mathbf{V}_{dr,s} = E\big[(\mathbf{y}_{d,r} - \mathbf{X}_{d,r}\boldsymbol{\beta})\,(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\big] = E(\mathbf{Z}_{d,r}\,\mathbf{u}\,\mathbf{u}'\,\mathbf{Z}_s')$ $+ E(\boldsymbol{\epsilon}_{d,r}\,\boldsymbol{\epsilon}_s) = \mathbf{Z}_{d,r}\mathbf{G}\,\mathbf{Z}_s' + \mathbf{R}_{dr,s}$. When $\mathbf{R}_{dr,s}$ is a null matrix, then the covariances between the unobservable vector $\mathbf{y}_{d,r}$ and the observable vector $\mathbf{y}_s$ are uniquely determined by the random effects $\mathbf{u}$. Consequently the EBLUP of $\mathbf{y}_{d,r}$ coincides with the EBLUP of $E(\mathbf{y}_{d,r}|\mathbf{u})$:

$$\widetilde{\mathbf{y}}_{d,r} = \widetilde{E}(\mathbf{y}_{d,r}|\mathbf{u}) = \mathbf{X}_{d,r}\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_{d,r}\mathbf{G}\,\mathbf{Z}_s'\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\,.$$

The EBLUP total for the unobservable part of the small area $\tau_{d,r}$ is now equal to the EBLUP of $E(\tau_{d,r}|\mathbf{u})$, with

$$\begin{aligned}\widetilde{\tau}_{d,r} \;=\; \widetilde{E}(\tau_{d,r}|\mathbf{u}) \;&=\; \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \sum_{h=1}^{H} \boldsymbol{\tau}'_{\mathbf{x}(1),dh,r}\widetilde{\mathbf{u}}_h^{(1)} + \boldsymbol{\tau}'_{\mathbf{x}(2),d,r}\widetilde{\mathbf{u}}_d^{(2)} \\ &=\; \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\tau}'_{\mathbf{z},d,r}\,\mathbf{G}\,\mathbf{Z}'_s\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\;,\end{aligned}$$

and the estimator of τ_d is given in a simplified expression by

$$\widetilde{\tau}_d \;=\; \tau_{\mathbf{y},d,s} + \boldsymbol{\tau}'_{\mathbf{x},d,r}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\tau}'_{\mathbf{z},d,r}\,\mathbf{G}\,\mathbf{Z}'_s\mathbf{V}_s^{-1}\big(\mathbf{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}\big)\;,$$

where $\tau_{\mathbf{y},d,s}$ is the observed sample total in small area d.

REFERENCES

[1] Amemiya, T. (1971). The Estimation of the Variances in a Variance-Components Model, *International Economic Review*, **12**, 1–13.

[2] Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator, *Biometrika*, **70**(2), 343–365.

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

[4] Bernardinelli, L. and Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk, *Statistics in Medicine*, **11**, 983–1007.

[5] Chandra, H.; Salvati, N. and Chambers, R. (2007). Small area estimation for spatially correlated populations — a comparison of direct and indirect model-based methods, *Statistics in Transition – new series*, **8**(2), 331–350.

[6] Cressie, N. (1991). *Statistics for Spatial Data*, John Wiley & Sons, New York.

[7] Diggle, P. (1988). An approach to the analysis of repeated measurements, *Biometrics*, **44**, 959–971.

[8] Fuller, W.A. and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structures, *Journal of the American Statistical Association*, **68**, 625–632.

[9] Hartley, H. and Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model, *Biometrika*, **54**, 93–108.

[10] Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika*, **61**(2), 383–385.

[11] Harville, D.A. (1977). Maximum likelihood approaches to variance components estimation and to related problems, *Journal of the American Statistical Association*, **72**, 320–337.

[12] Henderson, C.R. (1953). Estimation of variance and covariance components, *Biometrics*, **9**, 226–252.

[13] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics*, **31**, 423–447.

[14] Heyde, C.C. (1997). *Quasi-Likelihood and Its Application*, Springer-Verlag, New York.

[15] Jennrich, R. and Schluchter, M. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics*, **42**, 805–820.

[16] Jiang, J. (1996). REML estimation: Asymptotic behaviour and related topics, *The Annals of Statistics*, **24**(1), 255–286.

[17] Jiang, J. (1997). A derivation of BLUP — best linear unbiased predictor, *Statistics and Probability Letters*, **25**, 321–324.

[18] Jiang, J. (1998). Consistent estimators in generalized linear mixed models, *Journal of the American Statistical Association*, **93**, 720–729.

[19] Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer-Verlag, New York.

[20] Kott, P. (1989). Robust small domain estimation using random effects modeling, *Survey Methodology*, **15**(1), 3–12.

[21] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, **38**, 963–974.

[22] McLean, R.; Sanders, W. and Stroup, W. (1991). A unified approach to mixed linear models, *The American Statistician*, **45**, 54–64.

[23] Miller, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance, *The Annals of Statistics*, **5**(4), 146–762.

[24] Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika*, **58**(3), 545–554.

[25] Petrucci, A. and Salvati, N. (2004). *Small Area Estimation considering spatially correlated errors: the unit level random effects model*, Working Paper No. 2004/10, Dipartimento di Statistica "G. Parenti", Firenze.

[26] Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment, *Journal of Agricultural, Biological and Environmental Statistics*, **11**(2), 169–182.

[27] Pfeffermann, D. (1993). The role of sampling weights when modeling survey data, *International Statistical Review*, **61**(2), 317–337.

[28] Pinheiro, J. and Coelho, P. (2004). Spatial variability in the general mixed model, *Revista de Estatística*, **2**, 35–80.

[29] Prasad, N.G.N. and Rao, J.N.K. (1999). On robust small area estimation using a simple random effects model, *Survey Methodology*, **25**(1), 67–72.

[30] Pratesi, M. and Salvati, N. (2004). *Spatial EBLUP in agricultural survey. An application based on census data*, Report No. 256, Department of Statistics and Mathematics, University of Pisa, Pisa.

[31] Pratesi, M. and Salvati, N. (2004). *Small area estimation: the EBLUP estimator with autoregressive random area effects*, Report No. 261, Department of Statistics and Mathematics, University of Pisa, Pisa.

[32] Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects, *Statistical Methods and Applications*, **17**(1), 113–141.

[33] Rao, C.R. (1970). Estimation of heteroscedastic variances in linear models, *Journal of the American Statistical Association*, **65**, 161–172.

[34] Rao, C.R. (1971). Estimation of variance and covariance components — MINQUE theory, *Journal of Multivariate Analysis*, **1**, 257–275.

[35] Rao, C.R. (1972). Estimation of variance and covariance components in linear models, *Journal of the American Statistical Association*, **67**, 112–115.

[36] Salvati, N. (2004). *Small area estimation by spatial models: the spatial empirical best linear unbiased prediction (spatial EBLUP)*, Working Paper No. 2004/3, Dipartimento di Statistica "G. Parenti", Firenze.

[37] Särndal, C.E. (1984). Design-consistent versus model-dependent estimators for small domains, *Journal of the American Statistical Association*, **79**, 624–631.

[38] Särndal, C.E. and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis, *Journal of the American Statistical Association*, **84**, 266–275.

[39] Singh, M.B.; Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data, *Survey Methodology*, **20**, 3–22.

[40] Singh, B.B.; Shukla, G.K. and Kundu, D. (2005). Spatio-temporal models in small area estimation, *Survey Methodology*, **31**(2), 183–195.

[41] Thompson, W.A. (1962). The problem of negative estimates of variance components, *The Annals of Mathematical Statistics*, **33**(1), 273–289.

[42] Verbyla, A.P. (1990). A conditional derivation of residual maximum likelihood, *Australian Journal of Statistics*, **32**(2), 227–230.

[43] Wolfinger, R. (1993). Covariance structures selection in general mixed models, *Communications in Statistics, Simulation and Computing*, **22**(4), 1079–1106.