UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 07-34
Statistic and Econometric Series 08
April 2007

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

BOOTSTRAP FOR ESTIMATING THE MEAN SQUARED ERROR OF THE SPATIAL EBLUP*

Isabel Molina[1], Nicola Salvati[2] and Monica Pratesi[3]

Abstract

This work assumes that the small area quantities of interest follow a Fay-Herriot model with spatially correlated random area effects. Under this model, parametric and nonparametric bootstrap procedures are proposed for estimating the mean squared error of the EBLUP (Empirical Best Linear Unbiased Predictor). A simulation study compares the bootstrap estimates with an asymptotic analytical approximation and studies the robustness to non-normality. Finally, two applications with real data are described.

Keywords: Spatial correlation; Simultaneously autoregressive process; Small area estimation; Parametric bootstrap; Nonparametric bootstrap.
JEL Classification: c13, c14, c15.

* This work has been developed under the support of the project PRIN "Metodologie di stima e problemi non campionari nelle indagini in campo agricolo-ambientale", awarded by the Italian Government to the Universities of Florence, Cassino, Pisa and Perugia. It has been also supported by the Spanish grants MTM2006-05693 and SEJ2004-03303.

[1] Depto. de Estadística, Univ. Carlos III de Madrid, E-mail: imolina@est-econ.uc3m.es
[2] Dipto. di Statistica e Matematica Applicata all'Economia, Univ. di Pisa, E-mail: salvati@ec.unipi.it
[3] Dipto. di Statistica e Matematica Applicata all'Economia, Univ. di Pisa, E-mail: m.pratesi@ec.unipi.it

Bootstrap for estimating the mean squared error of the Spatial EBLUP

Isabel Molina* Nicola Salvati† Monica Pratesi‡

Abstract

This work assumes that the small area quantities of interest follow a Fay-Herriot model with spatially correlated random area effects. Under this model, parametric and nonparametric bootstrap procedures are proposed for estimating the mean squared error of the EBLUP (Empirical Best Linear Unbiased Predictor). A simulation study compares the bootstrap estimates with an asymptotic analytical approximation and studies the robustness to non-normality. Finally, two applications with real data are described.

Keywords: Spatial correlation; Simultaneously autoregressive process; Small area estimation; Parametric bootstrap; Nonparametric bootstrap.

1 Introduction

Due to monetary limitations, surveys conducted by national statistical offices usually cannot provide direct estimates at small geographical areas, or for some domains or subgroups of the population, especially when the variable of interest has low frequency. The term "direct" refers to an estimate for an area/domain that is calculated using solely the data from that area/domain. For instance, in Spain, the Survey on Income and Living Conditions is planned to provide reliable direct estimates for Autonomous Communities, but not for Provinces or regions inside Provinces. Small Area Estimation (SAE) deals with estimating in such smaller regions or domains, called small areas, making use of the data from all areas that share common features. A broadly established tool in SAE are the regression models

*Depto. de Estadística, Univ. Carlos III de Madrid, imolina@est-econ.uc3m.es

†Dipto. di Statistica e Matematica Applicata all'Economia, Univ. di Pisa, salvati@ec.unipi.it

‡Dipto. di Statistica e Matematica Applicata all'Economia, Univ. di Pisa, m.pratesi@ec.unipi.it

with random area effects, since the random effects allow for between area variation apart from that explained by the auxiliary variables. Among these, Fay-Herriot (FH) models (Fay & Herriot, 1979) are used when the available auxiliary data are aggregated at the area level.

Spatial correlation among data from neighboring small areas is observed in many practical applications. If there are not covariates explaining sufficiently this between-area correlation, then it should be somehow represented in the covariance structure of the model. However, the introduction of a dependency structure among small areas entails a serious conceptual difference with respect to the traditional framework of independent small areas, where the overall covariance matrix is block-diagonal (Prasad & Rao, 1990).

A model with spatially correlated random effects in the context of SAE was firstly introduced by Cressie (1991). Recently, an extension of the FH model through the Simultaneously Autoregressive (SAR) process has been considered by Salvati (2004), Pratesi & Salvati (2005, 2006), Singh et al. (2005) and Petrucci & Salvati (2006). When all parameters involved in the covariance matrix are known, Pratesi & Salvati (2005) introduced the Spatial Best Linear Unbiased Predictor (SBLUP). In order to include the effect of the spatial correlation on the confidence interval width, they obtained an estimator of the mean squared error of the SBLUP. Recent results show that the coverage of the 95% confidence interval is appreciable and that the mean squared error (MSE) of the SBLUP does not exceed the MSE of the traditional BLUP (Pratesi & Salvati, 2006).

In practice there are unknown parameters in the model covariance matrix, called here variance components, that must be estimated from the sample data. Replacing the derived estimates for the parameters in the SBLUP leads to the so called Spatial Empirical Best Linear Unbiased Predictor (SEBLUP). Singh et al. (2005) proposed a second order approximation of the MSE of the SEBLUP. However, this approximation might produce too optimistic or conservative confidence intervals depending on the strength of the spatial correlation and on the values of the sampling variances (Pratesi & Salvati, 2006). Moreover, analytical approximations usually rely on strong model assumptions and require large number of small areas to approximate well the true values.

Resampling techniques are nowadays accepted as a good alternative to asymptotic analytical approximations. They are attractive for practitioners because of their conceptual simplicity and their easy application to complex statistical models. Furthermore, they usually require less assumptions and their performance relies less in the number or small areas.

Some resampling procedures have been already proposed in the small area framework. See for instance the jackknife method of Jiang & Lahiri (2002), the more recent parametric

bootstrap approaches of González-Manteiga et al. (2005, 2007) and Hall & Maiti (2006a), and the nonparametric bootstrap of Hall & Maiti (2006b).

To our knowledge, the bootstrap-based estimation of the MSE of the SEBLUP under the extended FH model with spatial correlation has not been intended yet. This work extends the parametric bootstrap of González-Manteiga et al. (2005) to the situation of this paper, and introduces a nonparametric approach that resamples both the random effects and the errors from the empirical distribution of their respective estimators.

A simulation study compares the efficiency of the analytical and the bootstrap MSE estimators introduced in the paper for different values of the spatial correlation, and analyzes the robustness of the bootstrap procedures to the absence of normality in the random effects and errors.

The paper is organized as follows. Section 2 presents the FH model with spatially correlated random area effects, describes how the Spatial EBLUP is obtained from the model and comments on the available model fitting methods. Section 3 discusses the estimation of the MSE of the Spatial EBLUP and describes an analytical approximation of this MSE. Section 4 introduces the mentioned parametric and nonparametric bootstrap methods for estimating the MSE. Then Section 5 describes the simulation study carried out for comparing the MSE estimators. Two real life applications are illustrated in Section 6, and finally, some conclusions are drawn in Section 7.

2 Model with Spatially Correlated Random Effects

The basic FH model relates linearly the small area quantities of inferential interest θ_i (for example, totals y_i or means $\bar{y}_i$) to some area level auxiliary covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$, and includes random effects v_i associated to the areas; that is,

$$\theta_i = \mathbf{x}_i\boldsymbol{\beta} + z_i v_i, \quad i = 1, \ldots, m. \tag{1}$$

Here z_i are known positive constants, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression parameters, v_i are independent and identically distributed random variables with mean 0 and variance σ_u^2. Moreover, it assumes that design-unbiased direct estimators $\hat{\theta}_i$ of θ_i are available for the m small areas and that they can be expressed as

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1, \ldots, m, \tag{2}$$

where e_i are independent sampling errors with mean 0 and known variances denoted by ψ_i, and independent of the random effects v_i (Ghosh & Rao, 1994). Combining (1) and (2),

the obtained model is

$$\hat{\theta}_i = \mathbf{x}_i\boldsymbol{\beta} + z_i v_i + e_i, \quad i = 1, \ldots, m. \tag{3}$$

Let us define the vectors $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)^T$, $\mathbf{v} = (v_1, \ldots, v_m)^T$ and $\mathbf{e} = (e_1, \ldots, e_m)^T$, and the matrices $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_m^T)^T$ and $\mathbf{Z} = \text{diag}(z_1, \ldots, z_m)$. In matrix notation, the model is

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{4}$$

which is a special case of the general linear mixed model with diagonal covariance structure.

Model (4) can be extended to allow for spatially correlated area effects as follows. Let $\mathbf{v}$ be the result of a SAR process with parameter ρ and proximity matrix $\mathbf{W}$ (Anselin, 1992; Cressie, 1993), i.e.,

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I}_m - \rho\mathbf{W})^{-1}\mathbf{u}, \tag{5}$$

where $\mathbf{u} = (u_1, \ldots, u_m)^T$ has mean $\mathbf{0}$ and covariance matrix $\sigma_u^2\mathbf{I}_m$, and $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. From (5), it can be easily seen that $\mathbf{v}$ has mean $\mathbf{0}$ and covariance matrix equal to

$$\mathbf{G} = \sigma_u^2[(\mathbf{I}_m - \rho\mathbf{W})(\mathbf{I}_m - \rho\mathbf{W}^T)]^{-1}. \tag{6}$$

Combining (4) and (5), since $\mathbf{e}$ is independent of $\mathbf{v}$, the model is

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I}_m - \rho\mathbf{W})^{-1}\mathbf{u} + \mathbf{e}. \tag{7}$$

The covariance matrix of $\hat{\boldsymbol{\theta}}$ is equal to

$$\mathbf{V} = \boldsymbol{\psi} + \mathbf{Z}\mathbf{G}\mathbf{Z},$$

where $\boldsymbol{\psi} = \text{diag}(\psi_1, \ldots, \psi_m)$ is the known $m \times m$ variance matrix of the vector of sampling errors $\mathbf{e}$. Under model (7), the Spatial BLUP of the quantity of interest $\theta_i = \mathbf{x}_i\boldsymbol{\beta} + z_i v_i$ is

$$\tilde{\theta}_i(\sigma_u^2, \rho) = \mathbf{x}_i\tilde{\boldsymbol{\beta}} + z_i\mathbf{b}_i^T\mathbf{G}\mathbf{Z}\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \tag{8}$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$ is an asymptotically consistent estimator of the regression parameter $\boldsymbol{\beta}$ and $\mathbf{b}_i^T$ is the $1 \times m$ vector $(0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the i-th position. We consider that the proximity matrix $\mathbf{W}$ is defined in row standardized form; that is, $\mathbf{W}$ is row stochastic. Then, $\rho \in (-1, 1)$ is called spatial autocorrelation parameter (Banerjee et al., 2004).

The estimator (8) depends on the unknown variance components σ_u^2 and ρ. The two stage estimator $\tilde{\theta}_i(\hat{\sigma}_u^2, \hat{\rho})$ obtained by replacing these parameters by asymptotically consistent estimators $\hat{\sigma}_u^2$ and $\hat{\rho}$ is called Spatial EBLUP (Salvati, 2004; Pratesi & Salvati, 2005; Singh et al., 2005; Petrucci & Salvati, 2006). Assuming normality of the random effects,

σ_u^2 and ρ can be estimated by ML or REML procedures. The ML or REML estimators can be obtained iteratively using the Nelder-Mead algorithm (Nelder and Mead 1965) and the scoring algorithm in sequence. The use of these procedures one after the other is necessary because the log-likelihood function has a global maximum and some local maxima (Pratesi & Salvati, 2005).

Avoiding distributional assumptions, Kelejian & Prucha (1999) proposed a generalized moments (GM) method for estimating the variance components σ_u^2 and ρ of the model. In Section 5 we compare the accuracy of the two methods, ML and GM, under several probability distributions of the random effects and errors.

3 Analytical approximation of the MSE

Under normality of random effects and errors, the MSE of the Spatial EBLUP can be decomposed as (Rao, 2003)

$$\begin{aligned} \mathrm{MSE}\left[\tilde{\theta}_i(\hat{\sigma}_u^2,\hat{\rho})\right] &= \mathrm{MSE}\left[\tilde{\theta}_i(\sigma_u^2,\rho)\right] + E\left\{\left[\tilde{\theta}_i(\hat{\sigma}_u^2,\hat{\rho}) - \tilde{\theta}_i(\sigma_u^2,\rho)\right]^2\right\} \\ &= g_{1i}(\sigma_u^2,\rho) + g_{2i}(\sigma_u^2,\rho) + g_{3i}(\sigma_u^2,\rho), \end{aligned} \tag{9}$$

where $g_{1i}(\sigma_u^2,\rho)$ represents the uncertainty due to the estimation of the random effects and is of order $O(1)$ for large m, $g_{2i}(\sigma_u^2,\rho)$ is due to the estimation of $\boldsymbol{\beta}$ and is of order $O(m^{-1})$, and the last term measures the uncertainty of the Spatial EBLUP that results from the estimation of the variance components σ_u^2 and ρ. While the exact analytical expression of the terms $g_{1i}(\sigma_u^2,\rho)$ and $g_{2i}(\sigma_u^2,\rho)$ can be expressed by a closed formula, the last quantity can not be calculated analytically, and therefore approximation is necessary (Pratesi & Salvati, 2005). Under normality, an approximation of g_{3i} can be obtained following the results of Kackar & Harville (1984), as

$$\begin{aligned} \tilde{g}_{3i}(\sigma_u^2,\rho) = \mathrm{tr}\Bigg\{ &\begin{bmatrix} \mathbf{b}_i^T\left(\mathbf{C}^{-1}\mathbf{Z}\mathbf{V}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}\mathbf{V}^{-1})\right) \\ \mathbf{b}_i^T\left(\mathbf{A}\mathbf{Z}\mathbf{V}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}\mathbf{V}^{-1})\right) \end{bmatrix} \mathbf{V}\times \\ &\times \begin{bmatrix} \mathbf{b}_i^T\left(\mathbf{C}^{-1}\mathbf{Z}\mathbf{V}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}\mathbf{V}^{-1})\right) \\ \mathbf{b}_i^T\left(\mathbf{A}\mathbf{Z}\mathbf{V}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}\mathbf{V}^{-1})\right) \end{bmatrix}^T \bar{\mathbf{V}}(\hat{\sigma}_u^2,\hat{\rho}) \Bigg\} \end{aligned} \tag{10}$$

where $\mathbf{C} = (\mathbf{I}_m - \rho\mathbf{W})(\mathbf{I}_m - \rho\mathbf{W}^T)$, $\mathbf{A} = \sigma_u^2[-\mathbf{C}^{-1}(2\rho\mathbf{W}\mathbf{W}^T - 2\mathbf{W})\mathbf{C}^{-1}]$ and $\bar{\mathbf{V}}(\hat{\sigma}_u^2,\hat{\rho})$ is the asymptotic covariance matrix of $\hat{\sigma}_u^2$ and $\hat{\rho}$. This leads to the approximation

$$\mathrm{MSE}[\tilde{\theta}_i(\hat{\sigma}_u^2,\hat{\rho})] \approx g_{1i}(\sigma_u^2,\rho) + g_{2i}(\sigma_u^2,\rho) + \tilde{g}_{3i}(\sigma_u^2,\rho). \tag{11}$$

In practical applications, the estimator $\tilde{\theta}_i(\hat{\sigma}_u^2, \hat{\rho})$ should be accompanied with an estimate of the MSE. For FH models with diagonal covariance matrix $\mathbf{V}$, that is, with independent small areas, an approximately unbiased estimator of (11) was obtained through Taylor linearization by Prasad & Rao (1990).

In the case of correlated random area effects like the SAR process, the small areas are not independent and then $\mathbf{V}$ is not diagonal. However, following the results of Harville & Jeske (1992), Zimmerman & Cressie (1992) have extended the Prasad-Rao estimator of the MSE to models with more general covariance structure. The authors refer to geostatistical models, in which the correlation matrix is directly specified, and they assume that the covariance function is linear in the parameters. This situation is likely to occur under geostatistical models where the covariance function depends on the distance between locations. Under SAR models, the covariance is assumed to depend on a proximity matrix that specifies the proximity between the areas. Even so, the SAR models lead to a covariance function that is similar to the Bessel variogram model (Griffith & Csillag, 1993). Then following the results of Zimmerman & Cressie (1992), when $\hat{\sigma}_u^2$ and $\hat{\rho}$ are REML estimators, an approximately unbiased estimator of the MSE is given by the expression

$$\text{mse}[\tilde{\theta}_i(\hat{\sigma}_u^2, \hat{\rho})] \approx g_{1i}(\hat{\sigma}_u^2, \hat{\rho}) + g_{2i}(\hat{\sigma}_u^2, \hat{\rho}) + 2\tilde{g}_{3i}(\hat{\sigma}_u^2, \hat{\rho}). \tag{12}$$

If $\hat{\sigma}_u^2$ and $\hat{\rho}$ are obtained by ML, then an approximately unbiased estimator of the MSE is

$$\text{mse}[\tilde{\theta}_i(\hat{\sigma}_u^2, \hat{\rho})] \approx g_{1i}(\hat{\sigma}_u^2, \hat{\rho}) - \mathbf{b}_{ML}^T(\hat{\sigma}_u^2, \hat{\rho}) \bigtriangledown g_{1i}(\hat{\sigma}_u^2, \hat{\rho}) + g_{2i}(\hat{\sigma}_u^2, \hat{\rho}) + 2\tilde{g}_{3i}(\hat{\sigma}_u^2, \hat{\rho}). \tag{13}$$

The extra term $\mathbf{b}_{ML}^T(\hat{\sigma}_u^2, \hat{\rho}) \bigtriangledown g_{1i}(\hat{\sigma}_u^2, \hat{\rho})$ accounts for the bias of $g_{1i}(\hat{\sigma}_u^2, \hat{\rho})$. Ignoring this term could lead to underestimation of the MSE (see e.g. Pratesi & Salvati, 2005, 2006; Petrucci & Salvati, 2006). Singh et al. (2005) derived a different estimator of the MSE for large m neglecting all $o(m^{-1})$ terms. Their estimator differs from (12) and (13) in the subtraction of an extra term called here $g_4(\hat{\sigma}_u^2, \hat{\rho})$. Up to terms of order $o(m^{-1})$, this term is equal to

$$g_4(\hat{\sigma}_u^2, \hat{\rho}) = \frac{1}{2} \operatorname{tr} \left\{ [\mathbf{I}_2 \otimes (\boldsymbol{\psi}\mathbf{V}^{-1})]\mathbf{H}[\mathcal{I}^{-1}(\sigma_u^2, \rho) \otimes (\mathbf{V}^{-1}\boldsymbol{\psi})] \right\} \tag{14}$$

where $\otimes$ denotes the Kronecker product, $\mathcal{I}(\sigma_u^2, \rho)$ is the Fisher information matrix and $\mathbf{H}$ is a partitioned matrix of order $2m \times 2m$ defined as

$$\mathbf{H} = \begin{pmatrix} \dfrac{\partial^2 \mathbf{V}}{\partial (\sigma_u^2)^2} & \dfrac{\partial^2 \mathbf{V}}{\partial \sigma_u^2 \partial \rho} \\ \dfrac{\partial^2 \mathbf{V}}{\partial \rho \partial \sigma_u^2} & \dfrac{\partial^2 \mathbf{V}}{\partial \rho^2} \end{pmatrix}$$

4 Bootstrap approximation of the MSE

This section describes two alternative bootstrap procedures designed for estimating the MSE of the Spatial EBLUP $\tilde{\theta}_i(\hat{\sigma}_u^2, \hat{\rho})$. Remind that in the expression of the MSE given in (9), there are exact closed formulas for $g_{1i}(\sigma_u^2, \rho)$ and $g_{2i}(\sigma_u^2, \rho)$. However, the term $g_{3i}(\sigma_u^2, \rho)$, that represents the additional uncertainty of the Spatial EBLUP due to estimating the variance components $\hat{\sigma}_u^2$ and $\hat{\rho}$, can not be calculated analytically and then requires approximation. Thus, the bootstrap approaches, as they appear below, are written only for obtaining an estimate of g_{3i}. This term is then used in (15) to calculate the final estimate of the MSE. The first procedure is a parametric bootstrap that extends the ideas of González-Manteiga et al. (2005) to the FH model with spatial correlation. The final estimate of the MSE obtained by this procedure is consistent if the model parameter estimates are consistent. This can be proved by the method of imitation as in González-Manteiga et al. (2005), using the asymptotic formula of the MSE obtained by Singh et al. (2005).

In the following, for a function $\mathbf{B}(\sigma_u^2, \rho)$ of σ_u^2 and ρ, we will write simply $\mathbf{B}$ when $\mathbf{B}$ is evaluated at the true values of σ_u^2 and ρ. The parametric bootstrap works as follows:

PARAMETRIC BOOTSTRAP

1) Fit model (7) to the initial data $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)^T$, obtaining estimates $\hat{\sigma}_u^2$, $\hat{\rho}$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_u^2, \hat{\rho})$.

2) Generate a vector $\boldsymbol{\omega}_1^*$ whose elements are m independent copies of a $N(0,1)$. Construct the bootstrap vectors $\mathbf{u}^* = \hat{\sigma}_u \boldsymbol{\omega}_1^*$ and $\mathbf{v}^* = (\mathbf{I}_m - \hat{\rho}\mathbf{W})^{-1}\mathbf{u}^*$, and calculate the bootstrap quantity of interest $\boldsymbol{\theta}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^*$.

3) Generate a vector $\boldsymbol{\omega}_2^*$ with m independent copies of a $N(0,1)$, independently of the generation of $\boldsymbol{\omega}_1^*$, and construct the random errors $\mathbf{e}^* = \boldsymbol{\psi}^{1/2}\boldsymbol{\omega}_2^*$.

4) Construct the bootstrap data $\hat{\boldsymbol{\theta}}^* = \boldsymbol{\theta}^* + \mathbf{e}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* + \mathbf{e}^*$.

5) Regarding $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\rho}$ as the real values of the parameters, fit model (7) to the bootstrap data $\hat{\boldsymbol{\theta}}^*$, obtaining new bootstrap estimates $\tilde{\boldsymbol{\beta}}^*(\hat{\sigma}_u^2, \hat{\rho})$, $\hat{\sigma}_u^{2*}$, $\hat{\rho}^*$ and $\tilde{\boldsymbol{\beta}}^*(\hat{\sigma}_u^{2*}, \hat{\rho}^*)$.

6) Calculate the bootstrap Spatial BLUP from the bootstrap data $\hat{\boldsymbol{\theta}}^*$ and assuming that the real values of the parameters σ_u^2 and ρ are respectively $\hat{\sigma}_u^2$ and $\hat{\rho}$, that is

$$\tilde{\theta}_i^{S*}(\hat{\sigma}_u^2, \hat{\rho}) = \mathbf{x}_i\tilde{\boldsymbol{\beta}}^*(\hat{\sigma}_u^2, \hat{\rho}) + z_i\mathbf{b}_i^T\mathbf{G}(\hat{\sigma}_u^2, \hat{\rho})\mathbf{Z}\mathbf{V}(\hat{\sigma}_u^2, \hat{\rho})^{-1}[\hat{\boldsymbol{\theta}}^* - \mathbf{X}\tilde{\boldsymbol{\beta}}^*(\hat{\sigma}_u^2, \hat{\rho})].$$

Calculate also the bootstrap Spatial EBLUP as

$$\tilde{\theta}_i^{S*}(\hat{\sigma}_u^{2*}, \hat{\rho}^*) = \mathbf{x}_i\tilde{\boldsymbol{\beta}}^*(\hat{\sigma}_u^{2*}, \hat{\rho}^*) + z_i\mathbf{b}_i^T\mathbf{G}(\hat{\sigma}_u^{2*}, \hat{\rho}^*)\mathbf{Z}\mathbf{V}(\hat{\sigma}_u^{2*}, \hat{\rho}^*)^{-1}[\hat{\boldsymbol{\theta}}^* - \mathbf{X}\tilde{\boldsymbol{\beta}}^*(\hat{\sigma}_u^{2*}, \hat{\rho}^*)].$$

7) Repeat steps 2)–6) B times. Let $\hat{\sigma}_u^{2*(b)}$ and $\hat{\rho}^{*(b)}$ be the bootstrap estimates obtained in b-th bootstrap replication. Additionally, let $\tilde{\theta}_i^{S*(b)}(\hat{\sigma}_u^2, \hat{\rho})$ be the bootstrap Spatial BLUP and $\tilde{\theta}_i^{S*(b)}(\hat{\sigma}_u^{2*(b)}, \hat{\rho}^{*(b)})$ the bootstrap Spatial EBLUP obtained in the b-th bootstrap replication.

8) A bootstrap estimator of g_{3i} is

$$g_{3i}^* = B^{-1} \sum_{b=1}^{B} \left[\tilde{\theta}_i^{S*(b)}(\hat{\sigma}_u^{2*(b)}, \hat{\rho}^{*(b)}) - \tilde{\theta}_i^{S*(b)}(\hat{\sigma}_u^2, \hat{\rho}) \right]^2.$$

The second procedure is a nonparametric bootstrap, where the bootstrap random effects $(u_1^*, \ldots, u_D^*)^T$ and the random errors $(e_1^*, \ldots, e_D^*)^T$ are obtained by resampling respectively from the empirical distribution of the predicted random effects $(\hat{u}_1, \ldots, \hat{u}_D)^T$ and the residuals $\hat{\mathbf{e}} = \hat{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{v}} = (\hat{e}_1, \ldots, \hat{e}_m)^T$, both previously standardized. This method should be robust to non-normality of any of the random components of the model. It works by replacing in the parametric bootstrap, steps 2) and 3) by the new steps 2') and 3') below:

NONPARAMETRIC BOOTSTRAP

2') With the estimates $\hat{\sigma}_u^2$, $\hat{\rho}$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_u^2, \hat{\rho})$ obtained in step 1), calculate predictors of $\mathbf{v}$ and $\mathbf{u}$ as

$$\begin{aligned} \hat{\mathbf{v}} &= \mathbf{G}(\hat{\sigma}_u^2, \hat{\rho})\mathbf{Z}\mathbf{V}(\hat{\sigma}_u^2, \hat{\rho})^{-1}[\hat{\boldsymbol{\theta}} - \mathbf{X}\tilde{\boldsymbol{\beta}}(\hat{\sigma}_u^2, \hat{\rho})], \\ \hat{\mathbf{u}} &= (\mathbf{I} - \hat{\rho}\mathbf{W})\hat{\mathbf{v}} = (\hat{u}_1, \ldots, \hat{u}_m)^T. \end{aligned}$$

Consider the theoretical predictor $\tilde{\mathbf{u}} = (\mathbf{I} - \rho\mathbf{W})\mathbf{G}\mathbf{Z}\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\tilde{\boldsymbol{\beta}})$. The covariance matrix of $\tilde{\mathbf{u}}$ is

$$\mathbf{V}_\mathbf{u} = (\mathbf{I} - \rho\mathbf{W})\mathbf{G}\mathbf{Z}\mathbf{P}\mathbf{Z}\mathbf{G}(\mathbf{I} - \rho\mathbf{W}^T),$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}.$$

Consider now the estimated matrix $\hat{\mathbf{V}}_\mathbf{u} = \mathbf{V}_\mathbf{u}(\hat{\sigma}_u^2, \hat{\rho})$. Let $\hat{\mathbf{V}}_\mathbf{u}^{1/2}$ be a root square of a generalized inverse of $\hat{\mathbf{V}}_\mathbf{u}$. A simple choice of root square can be obtained through the spectral decomposition of $\hat{\mathbf{V}}_\mathbf{u}$, in the form

$$\hat{\mathbf{V}}_\mathbf{u}^{-1/2} = \mathbf{M}_\mathbf{u}\boldsymbol{\Delta}_\mathbf{u}^{-1/2}\mathbf{M}_\mathbf{u}^T,$$

where $\boldsymbol{\Delta}_\mathbf{u}$ is a diagonal matrix with the $m - p$ non-zero eigenvalues of $\hat{\mathbf{V}}_\mathbf{u}$, and $\mathbf{M}_\mathbf{u}$ is the matrix with the corresponding eigenvectors in the columns. With the obtained

root square, standardize $\hat{\mathbf{u}}$ as $\hat{\mathbf{u}}^S = \hat{\mathbf{V}}_{\mathbf{u}}^{-1/2}\hat{\mathbf{u}} = (\hat{u}_1^S, \ldots, \hat{u}_m^S)^T$. Then, by the consistency of the estimators $\hat{\sigma}_u^2$ and $\hat{\rho}$, for large m the covariance matrix of $\hat{\mathbf{u}}^S$ is approximately equal to the identity matrix. It is convenient to re-standardize the elements $\hat{u}_i^S$ in the form

$$\hat{u}_i^{SS} = \frac{\hat{\sigma}_u(\hat{u}_i^S - m^{-1}\sum_{i=1}^m \hat{u}_i^S)}{\sqrt{m^{-1}\sum_{d=1}^m(\hat{u}_d^S - m^{-1}\sum_{i=1}^m \hat{u}_i^S)^2}}, \quad i = 1, \ldots, m.$$

Construct the vector $\mathbf{u}^* = (u_1^*, \ldots, u_m^*)^T$, whose elements are obtained by extracting a simple random sample with replacement of size m, from the set $\{\hat{u}_1^{SS}, \ldots, \hat{u}_m^{SS}\}$. Then obtain $\mathbf{v}^* = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{u}^*$ and calculate the bootstrap quantity of interest $\boldsymbol{\theta}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* = (\theta_1^*, \ldots, \theta_m^*)^T$

3') Compute the vector of residuals $\hat{\mathbf{e}} = \hat{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{v}} = (\hat{e}_1, \ldots, \hat{e}_m)^T$. Consider the theoretical vector $\tilde{\mathbf{e}} = \hat{\boldsymbol{\theta}} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{Z}\tilde{\mathbf{v}}$, where $\tilde{\mathbf{v}} = \mathbf{G}\mathbf{Z}\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\tilde{\boldsymbol{\beta}})$. The covariance matrix of $\tilde{\mathbf{e}}$ is given by $\mathbf{V}_{\mathbf{e}} = \boldsymbol{\psi}\mathbf{P}\boldsymbol{\psi}$. Standardize the residuals by $\mathbf{e}^S = \hat{\mathbf{V}}_{\mathbf{e}}^{-1/2}\mathbf{e} = (e_1^S, \ldots, e_m^S)^T$, where $\hat{\mathbf{V}}_{\mathbf{e}}^{-1/2}$ is a root square of a generalized inverse of $\hat{\mathbf{V}}_{\mathbf{e}} = \mathbf{V}_{\mathbf{e}}(\hat{\sigma}_u^2, \hat{\rho})$. Again, re-standardize this values

$$\hat{e}_i^{SS} = \frac{(\hat{e}_i^S - m^{-1}\sum_{i=1}^m \hat{e}_i^S)}{\sqrt{m^{-1}\sum_{d=1}^m(\hat{e}_d^S - m^{-1}\sum_{i=1}^m \hat{e}_i^S)^2}}, \quad i = 1, \ldots, m.$$

Construct $\mathbf{r}^* = (r_1^*, \ldots, r_m^*)^T$ by extracting a simple random sample with replacement of size m from the set $\{\hat{e}_1^{SS}, \ldots, \hat{e}_m^{SS}\}$. Then take $\mathbf{e}^* = (e_1^*, \ldots, e_m^*)^T$, where $e_i^* = \psi_i^{1/2} r_i^*$, $i = 1, \ldots, m$.

With the obtained bootstrap estimate g_{3i}^* of g_{3i}, an estimate of the MSE of the Spatial EBLUP can be obtained by adding the estimated values of g_{1i} and g_{2i}, and a bootstrap correction of the bias induced by the estimation of these two quantities, as

$$\text{mse}[\tilde{\theta}_i(\hat{\sigma}_u^2, \hat{\rho})] = 2\left[g_{1i}(\hat{\sigma}_u^2, \hat{\rho}) + g_{2i}(\hat{\sigma}_u^2, \hat{\rho})\right] - B^{-1}\sum_{b=1}^{B}\left[g_{1i}(\hat{\sigma}_u^{2*(b)}, \hat{\rho}^{*(b)}) + g_{2i}(\hat{\sigma}_u^{2*(b)}, \hat{\rho}^{*(b)})\right] + g_{3i}^*. \tag{15}$$

Remark 4.1. *When there are doubts of the normality assumption either for the random effects or for the errors, it is possible to combine step 2') with 3), or step 2) with 3') of the two bootstrap procedures. The result is a semiparametric bootstrap that avoids the normality assumption on the desired component of the model.*

5 Simulation study

In this section we describe the simulation experiments carried out with the following main objectives in mind: (a) to find empirical evidence on the reasonable conjecture that taking into account the spatial correlation among small areas improves the precision of small area estimators; (b) to study the small-sample behavior of the proposed bootstrap procedures for estimating the term g_{3i} involved in the mean squared error, for different values of the spatial correlation parameter ρ and for different patterns of sampling variances ψ_i; (c) to analyze the robustness of the bootstrap procedures to non-normality of the random effects and errors.

The experiments are based on a real population, the map of the $m = 287$ municipalities (small areas) of Tuscany. We considered a model with $p = 2$, that is, one explanatory variable and a constant, with $m \times 2$ design matrix $\mathbf{X} = [\mathbf{1}_m \ \mathbf{x}]$, where $\mathbf{1}_m$ is a column vector of ones of size m and $\mathbf{x} = (x_1, \ldots, x_m)^T$ contains the values of the explanatory variable. These values x_i were generated from a uniform distribution in the interval $(0, 1)$. The true model coefficients were $\boldsymbol{\beta} = (1, 2)^T$, the random effects variance $\sigma_u^2 = 1$ and the spatial correlation parameter $\rho \in \{-0.75, -0.5, -0.25, 0.25, 0.5, 0.75\}$. The matrix of sampling variances $\boldsymbol{\psi} = \text{diag}(\psi_1, \ldots, \psi_m)$ was taken as $\psi_i = 0.7$ for $1 \leq i \leq 60$; $\psi_i = 0.6$ for $61 \leq i \leq 120$; $\psi_i = 0.5$ for $121 \leq i \leq 180$; $\psi_i = 0.4$ for $181 \leq i \leq 240$ and finally $\psi_i = 0.3$ for $241 \leq i \leq 287$ (Datta et al., 2005). The $m \times m$ row-standardized proximity matrix $\mathbf{W}$ was obtained from the neighborhood structure of the municipalities in Tuscany. This matrix was kept constant for all simulations. We considered three possible probability distributions for the random area effects and errors, namely Normal, Gumbel and Student t distribution with 6 degrees of freedom, all standardized to have zero mean and unit variance. The last two distributions represent two different sources of discrepancy to normality, since the Gumbel distribution is asymmetric and the Student t has heavy tails.

A first experiment was carried out for comparing the performance of the ML and the GM (Kelejian & Prucha, 1999) methods for estimating σ_u^2 and ρ. For this, $L = 1000$ Monte Carlo data sets were generated as described above, and the model was fitted to each data set by the two methods, ML and GM. The comparison is based on the empirical relative bias and the empirical relative mean squared error of the estimators. For an estimator $\hat{\delta}$ of a parameter δ, these quantities are defined respectively as

$$\text{RB}(\hat{\delta}) = \frac{1}{L}\sum_{j=1}^{L}\frac{\hat{\delta}^{(j)}}{\delta} - 1, \quad \text{RMSE}(\hat{\delta}) = \frac{1}{L}\sum_{j=1}^{L}\frac{(\hat{\delta}^{(j)} - \delta)^2}{\delta},$$

where $\hat{\delta}^{(j)}$ is the estimate obtained for the j-th data set.

	RB($\hat{\sigma}_u^2$)		RB($\hat{\rho}$)		RMSE($\hat{\sigma}_u^2$)		RMSE($\hat{\rho}$)	
	ML	GM	ML	GM	ML	GM	ML	GM
Normal	0,062	0,101	-0,059	-0,183	0,026	0,027	0,007	0,030
Student t_6	0,054	0,100	-0,052	-0,180	0,036	0,036	0,007	0,030
Gumbel	0,149	0,163	-0,075	-0,190	0,058	0,056	0,009	0,032

Table 1: Relative bias and relative mean squared error of the estimators of $\sigma_u^2 = 1$ and $\rho = 0.75$, by ML and GM methods, for Normal, Gumbel and Student t_6 distributions.

Table 1 lists the previous indicators obtained by the two estimation methods, under the three considered probability distributions, taking the same distribution for the random effects u_i and the errors e_i, and for $\rho = 0.75$. This table shows that ML estimates have smaller relative bias and not greater relative mean squared error than the GM estimates, except for the estimator of σ_u^2 obtained under the Gumbel distribution. Moreover, for that parameter the differences between the two estimation methods are smaller. Observe that the ML method estimates ρ better than the GM method even under the two non-normal distributions. Nevertheless, an advantage of the GM method, apart from being distribution-free, is that it is rather faster than ML. Thus, the GM method is convenient under nonparametric settings and when applying some computationally intensive procedure like bootstrap.

Concerning target (a), $L = 1000$ Monte Carlo data sets were generated as described before, taking Normal distribution for the random effects and errors. Then two models were fitted to each data set: the spatial model (4)-(5), and the non-spatial model obtained by assuming that in model (4), the vector of random effects $\mathbf{v} = (v_1, \ldots, v_m)^T$ has independent and identically distributed elements v_i, with zero mean and variance σ_u^2. Figures 1 and 2 plot the empirical values of the mean squared errors of the Spatial EBLUP obtained from the former model, and the NonSpatial EBLUP resulting from the latter model, for the $m = 287$ small areas, for $\rho = 0.75$ and $\rho = 0.25$, respectively. The piecewise decreasing shape that we observe in the level of these two figures is due to the decreasing patterns of sampling variances ψ_i. Figure 1 shows that ignoring the spatial correlation structure of small areas leads to an increase in the MSE. However, this increase is smaller for areas with smaller sampling variances and in the case of weak spatial correlation, see Figure 2 for $\rho = 0.25$. This last figure also suggests that modelling the spatial correlation seems to be convenient even when this correlation is weak, since there is no loss in efficiency.

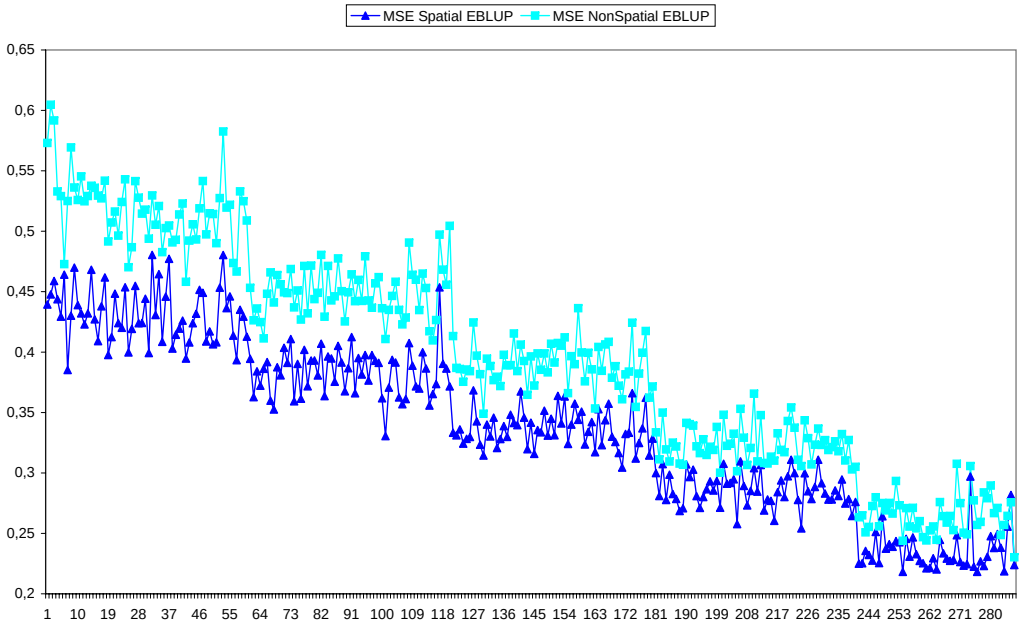

Figure 1: Empirical MSE of the Spatial EBLUP and the NonSpatial EBLUP for the $m = 287$ small areas, for $\rho = 0.75$.

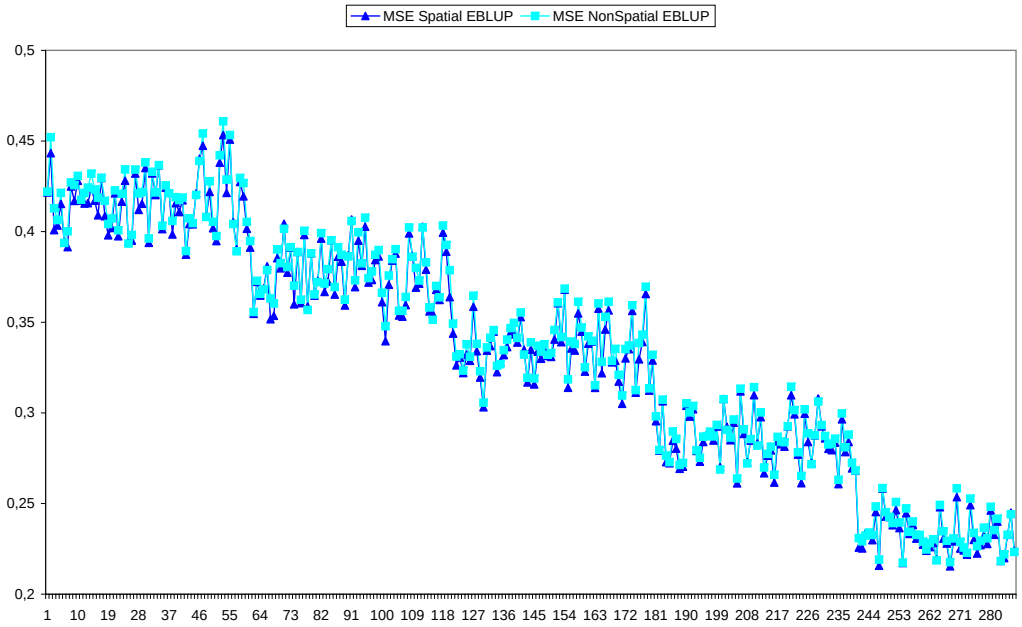

Figure 2: Empirical MSE of the Spatial EBLUP and the NonSpatial EBLUP for the $m = 287$ small areas, for $\rho = 0.25$.

Target (b) deals with comparing the analytical estimate of g_{3i} obtained by substituting $\hat{\sigma}_u^2$ and $\hat{\rho}$ in expression (10), with the parametric bootstrap estimate. For this, $L = 250$ Monte Carlo data sets were generated, and for each data set, the estimated values of $\tilde{g}_{3i}$, $i = 1, \ldots, m$, were calculated, and the parametric bootstrap procedure was applied with $B = 250$ bootstrap replicates, deriving bootstrap estimates g_{3i}^*, $i = 1, \ldots, m$. The empirical values of g_{3i}, which are the reference values for comparison, were computed previously with 1000 Monte Carlo replicates to ensure better accuracy.

Figures 3–6 plot the ratios of the analytical estimates $\tilde{g}_{3i}(\hat{\sigma}_u^2, \hat{\rho})$ and the parametric bootstrap estimates g_{3i}^* over the empirical values, under normality of random effects and errors and with ML estimation of the parameters σ_u^2 and ρ, for $\rho = 0.75, 0.5, 0.25, -0.5$ respectively. The straight lines in each plot correspond to the empirical values. First of all we want to point out that the term g_{3i} has very small range of variation: our reference empirical values range in the interval $(0.0007, 0.004)$. The result is tenable for small, medium and high correlation and it is confirmed for all considered patterns of sampling variances. For $\rho = 0.75$ (Figure 3), the ratio of the analytical estimates to the empirical ones highlight an underestimation of the true g_{3i} value for almost every area. However, Figures 4–6 indicate that this bias disappears as long as the spatial correlation decreases. The bootstrap estimates, although also slightly biased, are more stable, taking generally values closer to the empirical values, for all considered values of the spatial correlation parameter and for all patterns of error variances. The plots corresponding to $\rho \in \{-0.25, -0.75\}$ are omitted because of their similarity with Figure 6 for the average value $\rho = -0.5$.

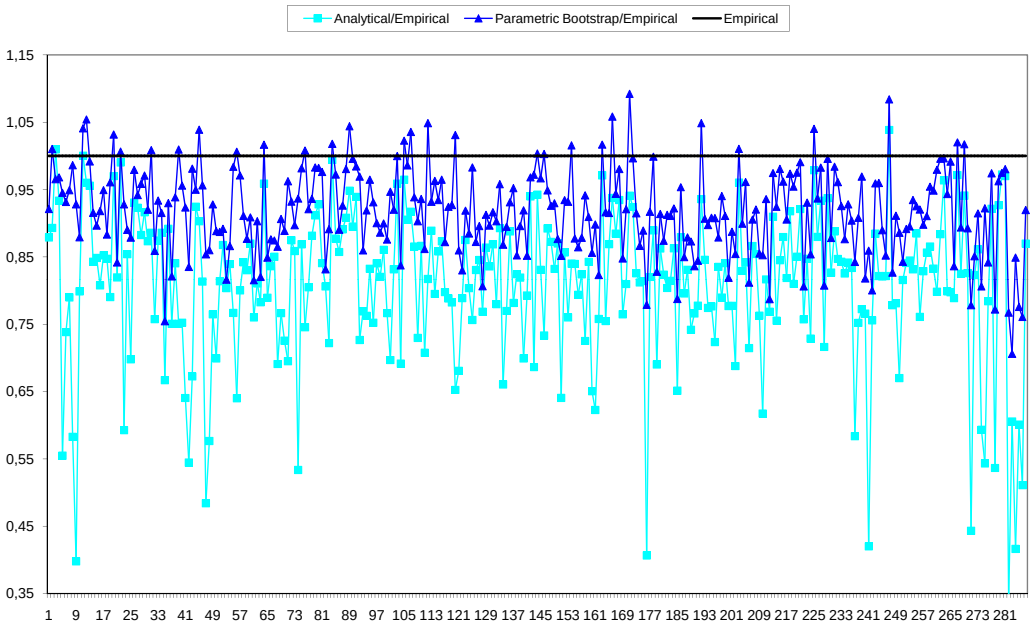

Figure 3: Ratios of the empirical g_{3i}, estimated $\tilde{g}_{3i}$ and parametric bootstrap estimates g^*_{3i} over the empirical values, for the $m = 287$ small areas, with $\rho = 0.75$.

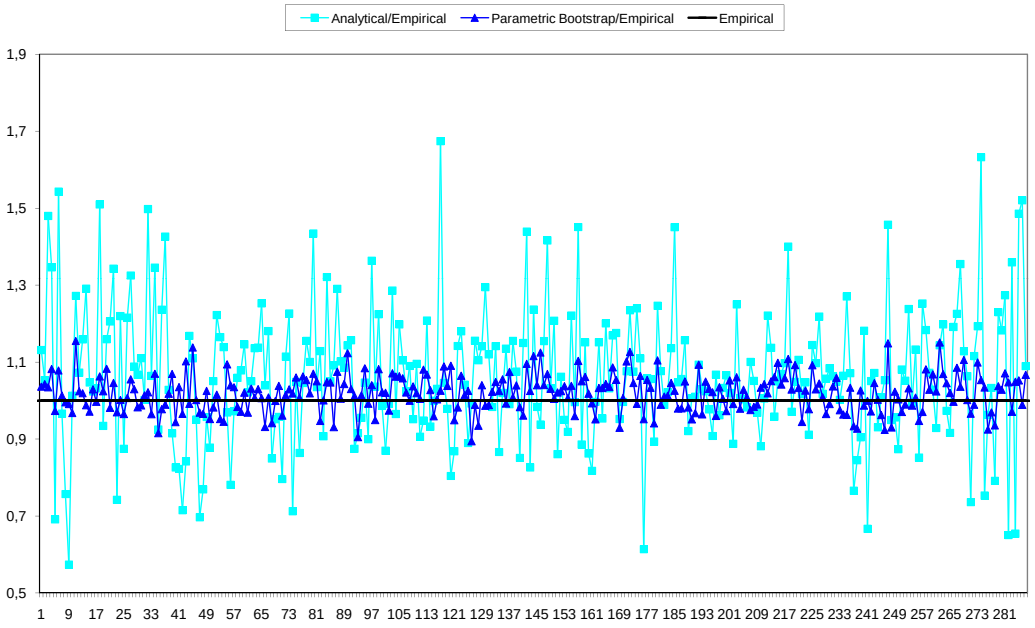

Figure 4: Ratios of the empirical g_{3i}, estimated $\tilde{g}_{3i}$ and parametric bootstrap estimates g^*_{3i} over the empirical values, for the $m = 287$ small areas, with $\rho = 0.5$.

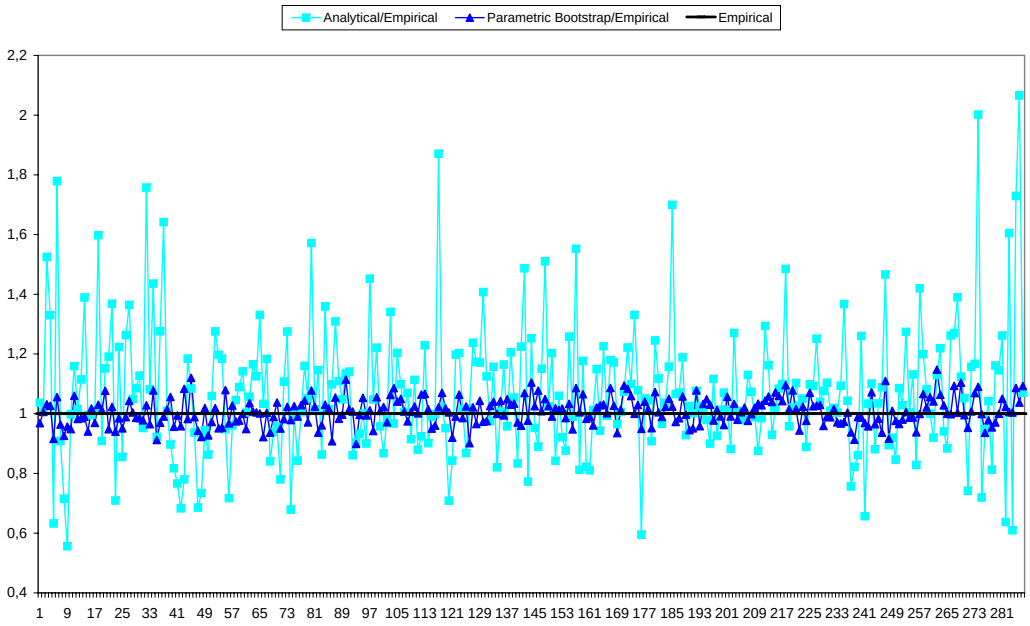

Figure 5: Ratios of the empirical g_{3i}, estimated $\tilde{g}_{3i}$ and parametric bootstrap estimates g^*_{3i} over the empirical values, for the $m = 287$ small areas, with $\rho = 0.25$.

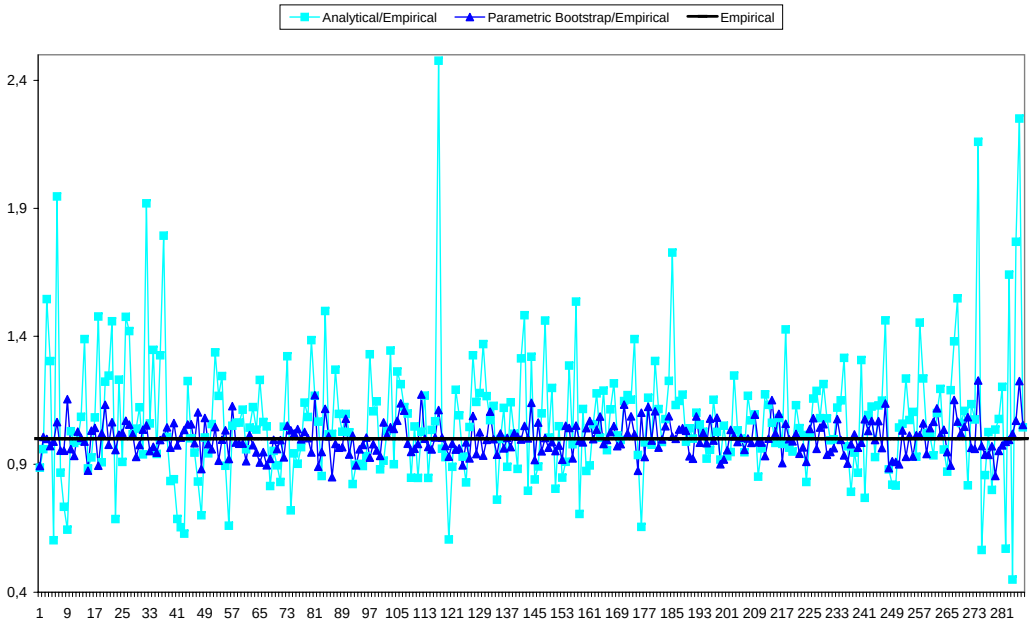

Figure 6: Ratios of the empirical g_{3i}, estimated $\tilde{g}_{3i}$ and parametric bootstrap estimates g^*_{3i} over the empirical values, for the $m = 287$ small areas, with $\rho = -0.5$.

Finally, concerning target (c), both parametric and nonparametric bootstrap procedures were applied to each of $L = 250$ data sets, firstly generated with Normal distribution, and next with Gumbel and Student t_6 distributions. In this case, the GM estimation method was used. Figures 7–9 show the ratios of the empirical values g_{3i}, the parametric bootstrap estimates g_{3i}^{*}, and the nonparametric bootstrap estimates g_{3i}^{**} over the empirical values, for $i = 1, \ldots, m$, and for $\rho = 0.5$. Figure 7 illustrates that under normality of random effects and errors, the nonparametric bootstrap is not less efficient than the parametric bootstrap. Both estimates take similar values, but the right side of the plot indicates a small positive bias for areas with smaller sampling variances. This could be a consequence of the GM estimation method, since this bias is not appreciable in the case of ML (Figure 4). Figure 8 shows that the parametric bootstrap is quite robust to skewness, when the true random effects and the errors follow a Gumbel distribution. Finally, when the data come from a distribution with heavy tails as the Student t_6 (Figure 9), the nonparametric bootstrap seems to perform better.

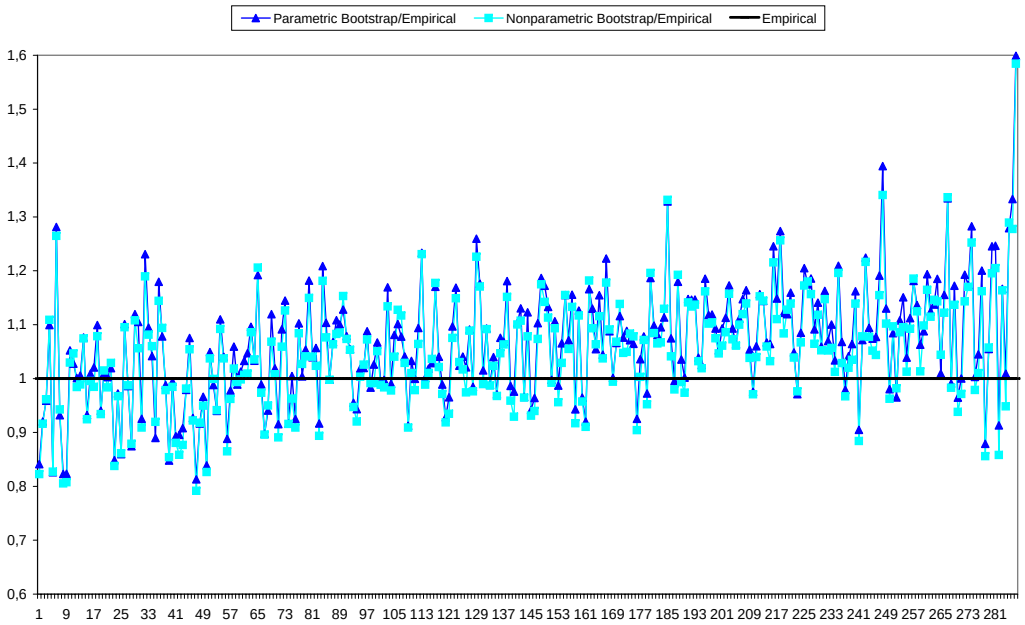

Figure 7: Ratios of the parametric and the nonparametric bootstrap estimates g_{3i}^{*} and g_{3i}^{**} over the empirical values of g_{3i} for the $m = 287$ small areas, for Normal distribution and with $\rho = 0.5$.

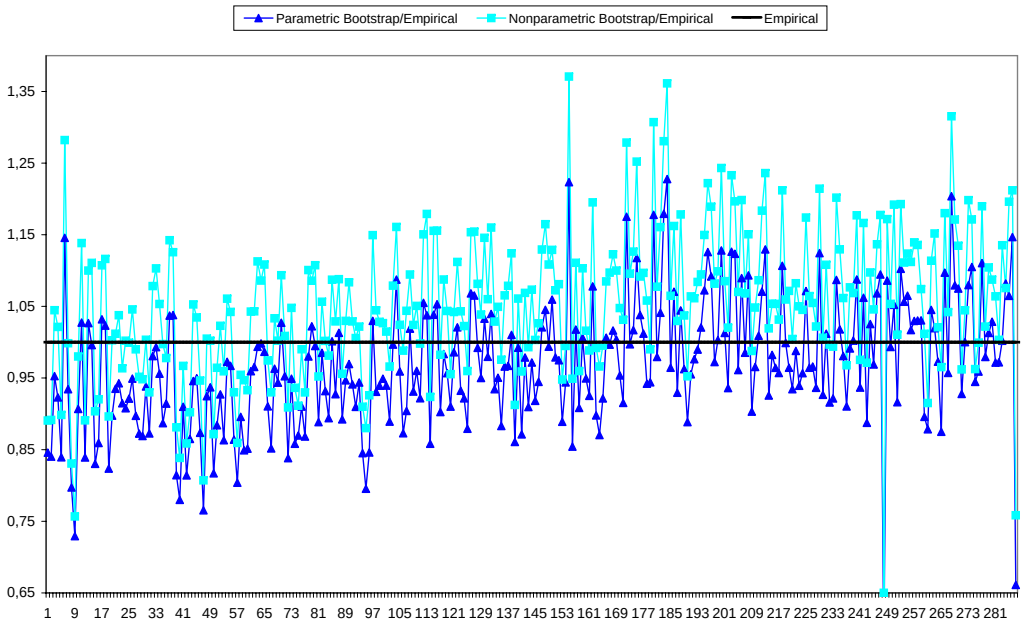

Figure 8: Ratios of the parametric and the nonparametric bootstrap estimates g_{3i}^{*} and g_{3i}^{**} over the empirical values of g_{3i} for the $m = 287$ small areas, for Gumbel distribution and with $\rho = 0.5$.

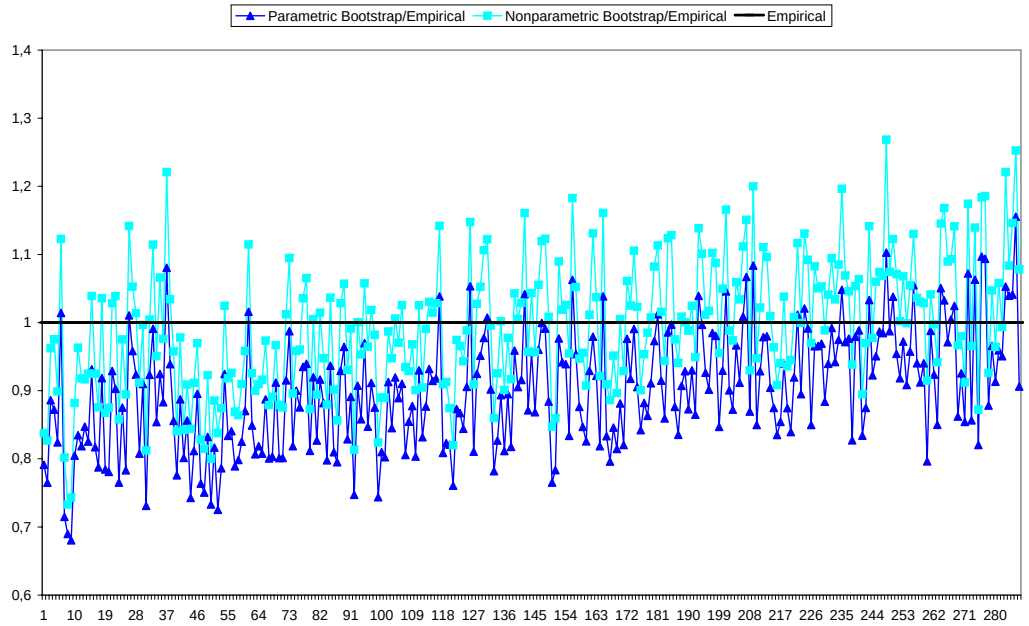

Figure 9: Ratios of the parametric and the nonparametric bootstrap estimates g_{3i}^{*} and g_{3i}^{**} over the empirical values of g_{3i} for the $m = 287$ small areas, for Student t_6 distribution and with $\rho = 0.5$.

6 Two real life applications

This section describes two applications with real data. In the first application, the goal is to estimate the mean production of olives (in quintal units) over the farms in each Agrarian Region (AR) from Tuscany. The data come from the Farm Structure Survey of 2003 (Source: ISTAT). The sample is extracted by a one stage stratified design with self-representation of the larger farms (agricultural holdings). The sample size over all Italy is 55,030 farms. The stratification is carried out in three phases and the optimal allocation of sample size to the strata is obtained by minimizing the sampling error at regional and national level (Ballin & Salvi, 2004).

In the second application, the aim is to estimate the mean per acre erosion (in tons) in each 11 digit Hydrologic Unit Code (HUC) from the Rathbun Lake Watershed in Iowa (U.S.). The data were collected from a team of researchers from Iowa State University and the Chariton Valley Resource Conservation and Development Office, who carried out an environmental health study for the Rathbun Lake Watershed in 1999. Each HUC was divided in plots; in total 2146, and from them 183 plots were selected by systematic sampling. The fractional interval (Särdnal et al. (1992), p. 77) was fixed in order to select four units from each HUC. Then, within each HUC, three 160-acre (64 ha) plots were selected. For details about the sampling design, see Opsomer et al. (2003).

In both applications, the data can be considered as lattice data. The centroid of each area is taken as the spatial reference for all the units (farms for ARs and plots for HUCs) in the same area. Both ARs and HUCs are unplanned domains. They are defined on a geographical basis and are very useful small areas in economic studies of agriculture and land use respectively. The 53 ARs are determined following the administrative boundaries of Municipalities (287 in Tuscany) and the average sample size per AR is $\bar{n} = 45.2$ (s.e.=37.3). The number of HUCs in the Rathbun Lake Watershed is 61 with an average size of 5.800 acres (2.350 ha).

The proximity matrix $\mathbf{W} = (w_{ij})$ is constructed as follows: w_{ij} is equal to 1 if the AR (HUC) i shares an edge with AR (HUC) j, and is equal to 0 otherwise. Afterwards, the rows of $\mathbf{W}$ are standardized so that the row elements sum up to one. Then $\mathbf{W}$ is not symmetric, but it is row stochastic and ρ is called spatial autocorrelation parameter.

The results of both mean per farm production of olives at ARs in Tuscany and mean per acre erosion at HUCs are described in detail in Pratesi & Salvati (2006) and Petrucci & Salvati (2006). Here we are interested in the comparison between the different estimates of g_{3i}. Thus, for the two case studies, we have computed the analytical estimates $\tilde{g}_{3i}(\hat{\sigma}_u^2, \hat{\rho})$ and the two bootstrap estimates g_{3i}^* and g_{3i}^{**}. Additionally, for the first case study, we performed

the semiparametric bootstrap obtained by combining step 2) from the parametric bootstrap with step 3') from the nonparametric procedure. This semiparametric bootstrap assumes normality only for the random effects and not for the errors. The obtained results are plotted in Figures 10 and 11.

In the first case study (mean per farm production of olives) the value of the estimated spatial autoregressive coefficient $\hat{\rho}$ is 0.382 (s.e.=0.271), which means a weak spatial relationship. Observe in Figure 10 that the analytical estimates are much lower than the bootstrap estimates. Thus, it seems that the analytical estimator does not capture the additional variability due to the estimation of the spatial autocorrelation coefficient. Observe also that the nonparametric and the semiparametric bootstrap estimates show a similar behavior. This similarity supports the normality assumption of the random effects. On the other hand, the parametric bootstrap estimates take larger values than the other two bootstrap methods in the areas with larger sampling variance. We deduce from this plot that the distribution of the direct estimators is probably far from normality and in this case the nonparametric bootstrap is more reliable.

In the second case study (mean per acre erosion) the value of the estimated spatial autocorrelation coefficient $\hat{\rho}$ is 0.741 (s.e.= 0.138) using the ML procedure and 0.756 (s.e.= 0.154) with the REML method, which suggests a strong spatial relationship. Figure 11 shows that in this case the estimates are not very different, but the analytical estimates are in general slightly smaller than the bootstrap analogues. This case study suggests that when spatial correlation is stronger, the analytical estimator of g_{3i} can be more reliable.

7 Conclusions

From the results of the simulation experiments and of the two applications with real data, we conclude that in case of spatially correlated data with a spatial autocorrelation parameter $\rho > 0.25$, the analytical estimator of the term g_{3i} of the MSE should be substituted by a bootstrap estimator. Take into account that the term $g_{3i}(\hat{\sigma}_u^2, \hat{\rho})$ is used to approximate the MSE. Since the analytical estimates $\tilde{g}_{3i}(\sigma_u^2, \rho)$ underestimate the true values, using them would lead to too optimistic confidence intervals for predicted values. Alternatively, the bootstrap estimators should result in more satisfactory point estimates of g_{3i} and in more appropriate confidence intervals.

Furthermore, between the bootstrap estimates, we have seen that the nonparametric bootstrap performs well even under normality, and therefore is expected to be reliable regardless of distributional assumptions. It seems reasonable to use the nonparametric bootstrap with a nonparametric estimation method like the GM; however, the ML method

performed better even under Gumbel and Student t_6 distributions. Thus, when there is some evidence of not great deviation from normality, a combination ML-nonparametric bootstrap should work well.

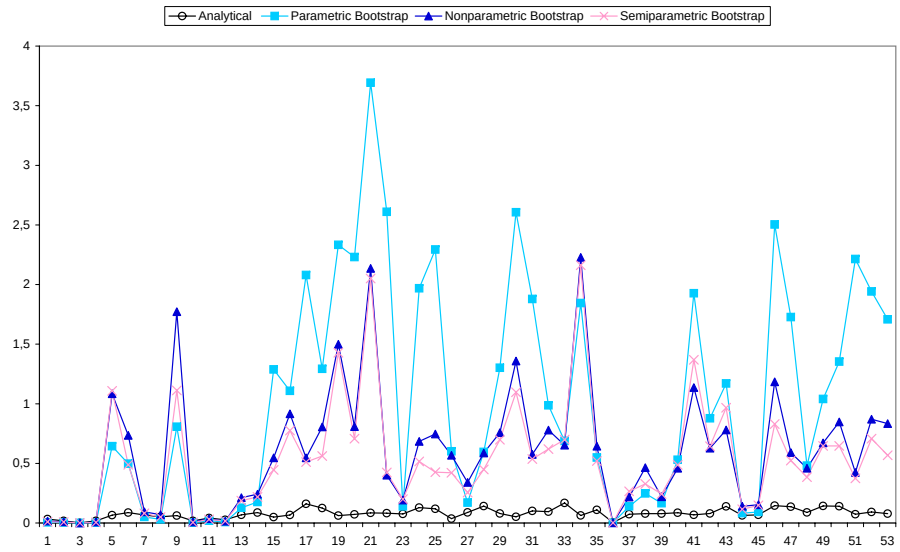

Figure 10: Analytical estimates, parametric and nonparametric bootstrap estimates of g_{3i}. Per farm production of olives at ARs in Tuscany.

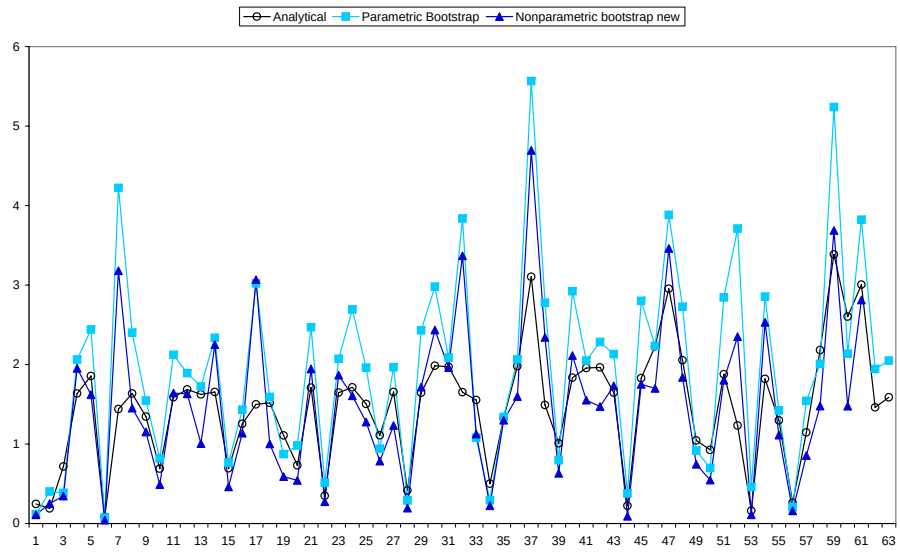

Figure 11: Analytical estimates, parametric bootstrap, and nonparametric bootstrap estimates of g_{3i}. Per acre erosion at HUCs for the Rathbun Lake Watershed in Iowa.

Acknowledgements

The work reported here has been developed under the support of the project PRIN *Metodologie di stima e problemi non campionari nelle indagini in campo agricolo-ambientale* awarded by the Italian Government to the Universities of Florence, Cassino, Pisa and Perugia. It has been also supported by the Spanish grants MTM2006-05693 and SEJ2004-03303.

References

Anselin, L. (1992). *Spatial Econometrics. Method and Models*. Boston: Kluwer Academic Publishers.

Ballin, M. & Salvi, S. (2004). Nota metodologica sul piano di campionamento adottato per l'indagine struttura e produzione delle aziende agricole 2003. *Istat, Servizio Agricoltura*.

Banerjee, S., Carlin, B. & Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. New York: Chapman and Hall.

Cressie, N. (1991). Small-area prediction of undercount using the general linear model. *Proceedings of Statystic Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada*, 93–105.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Datta, G., Rao, J. & Smith, D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika* **92**, 183–196.

Fay, R. & Herriot, R. (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.

Ghosh, M. & Rao, J. N. K. (1994). Small area estimation: An appraisal (Disc: p76-93). *Statistical Science* **9**, 55–76.

González-Manteiga, W., Lombardía, M., Molina, I., Morales, D. & Santamaría, L. (2005). Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. *Working Paper, Statistics and Econometrics Series, Universidad Carlos III de Madrid*, 05–49.

González-Manteiga, W., Lombardía, M., Molina, I., Morales, D. & Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under logistic mixed model. *Computational Statistics and Data Analyis* **51**, 2720–2733.

Griffith, D. & Csillag, F. (1993). Exploring relationships between semi-variogram and spatial autoregressive. *Papers in Regional Science* **72**, 283–296.

Hall, P. & Maiti, T. (2006a). On parametric bootstrap methods for small area prediction. *Journal Royal Statistical Society, Series B* **68**, 221–238.

Hall, P. & Maiti, T. (2006b). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics* **34**, 1733–1750.

Harville, D. & Jeske, D. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association* **87**, 724–731.

Jiang, J. & Lahiri, P. (2002). A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics* **30**, 2720–2733.

Kackar, R. & Harville, D. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association* **79**, 853–862.

Kelejian, H. H. & Prucha, R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* **40**, 509–533.

Opsomer, J. D., Botts, C. & Kim, J. Y. (2003). Small area estimation in watershed erosion assessment survey. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 139–152.

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

Prasad, N. & Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* **85**, 163–171.

Pratesi, M. & Salvati, N. (2005). Small area estimation: the eblup estimator with autoregressive random area effects. *Working Paper, Dipartimento di Statistica e Matematica Applicata all'Economia, Pisa*, 261.

Pratesi, M. & Salvati, N. (2006). Small area estimation in the presence of correlated random area effects. *Working Paper, Dipartimento di Statistica e Matematica Applicata all'Economia, Pisa*, 292.

Rao, J. N. K. (2003). *Small Area Estimation*. London: Wiley.

Salvati, N. (2004). Small area estimation by spatial models: the spatial empirical best linear unbiased prediction (spatial eblup). *Working Paper, Dipartimento di Statistica "G. Parenti", Firenze*, 2004/04.

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

Singh, B., Shukla, G. & Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology* **31**, 183–195.

Zimmerman, D. & Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics* **44**, 27–43.