Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2004/10

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

# Small Area Estimation considering spatially correlated errors: the unit level random effects model

Alessandra Petrucci,
Nicola Salvati

Università degli Studi
di Firenze

*Statistics*

# Small Area Estimation considering spatially correlated errors: the unit level random effects model

Alessandra Petrucci, Nicola Salvati
University of Florence - Department of Statistics "G. Parenti"
Viale Morgagni, 59 - 50134 Florence - Italy;
*e-mail:* alex@ds.unifi.it salvati@ds.unifi.it

**Abstract**

The paper describes an application of a modified small area estimator to the data collected in the Rathbun Lake Watershed in Iowa (USA). Opsomer *et al.* (2003) estimated the average erosion per acre for 61 sub-watersheds within the study region using an empirical best linear unbiased predictor (EBLUP).

The proposed methodology considers an EBLUP estimator with spatially correlated error taking into account the information provided by neighboring areas.

KEY WORDS: Unit level random effect model, EBLUP, Spatial model, Natural resources survey.

## 1 Introduction

Sample survey data are extensively used to provide reliable direct estimates of totals and means for the whole population and large areas or domains. A domain is regarded as "*small*" if the domain-specific sample is not large enough to support direct estimates of adequate precision; they are likely to yield large standard errors due to the unduly small size of the sample in the area. Traditional area-specific direct estimators do not provide adequate precision, then in making estimates for small areas it is necessary to employ indirect estimators that "borrow strength" from related area; in particular, model assisted and model based indirect estimators. They are based on either implicit or explicit models that provide a link to related small areas through auxiliary data. Two types of indirect estimators can be identified: indirect estimators based on implicit models (*models assisted*) include synthetic and composite estimators, while those based on explicit models (*model based*) incorporate area-specific effects .

Small area models make use of explicit linking models based on random area-specific effects that account for between areas variation beyond what is explained by auxiliary variables included in the model. The random area effects are considered independents, but in practice, basically in most of the applications on environmental data, it should be more reasonable to assume that the random area effects between the neighboring areas (for instance the neighborhood could be defined by a contiguity criterium) are correlated and the correlation decays to zero as distance increases. The absence of information about neighborhoods could produced a series of failings at national, local and community level; policies could easily be misdesigned or mistargeted and important trends could be missed by national and local government.

The aim of this work is to estimate the average watershed erosion (Opsomer *et al.*, 2003) taking into account the spatial dimension of the soil erosion data, collected on the Rathbun Lake Watershed (Iowa - USA), adapting a model with spatially correlated errors in the Empirical Best Linear Unbiased Predictor (EBLUP) estimator (Rao, 2003).

The paper is organized as follows: section 2 introduces the small area models that include random area-specific effects and EBLUP estimator is showed. In section 3 the Spatial EBLUP procedure is recalled. Section 4 discusses the results of the application of Spatial EBLUP to estimate the average sub-watershed erosion per acre on the Rathbun Lake Watershed (Iowa - USA).

## 2 Nested Error Unit Level Regression Models

A nested error unit level regression model assumes that unit-specific auxiliary data $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, ...x_{ijp})^T$ are available for each population element $j$ in each small area $i$ and the population mean $\bar{\mathbf{X}}_i$ are known. The relationship among $y_{ij}$, the variable of study, and $\mathbf{x}_{ij}$ is represented through a one-fold nested error linear regression model (Rao, 1994):

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + \epsilon_{ij} \quad i = 1...m \quad j = 1...N_i. \tag{1}$$

The random effects $u_i$ and $\epsilon_{ij}$ are assumed to be mutually independent error terms with zero means and variances $\sigma_u^2$ and $\sigma_\epsilon^2$ respectively (Saei and Chambers, 2003). In addition, it is often assumed normality of the $u_i$'s and $\epsilon_{ij}$'s . The random term $u_i$ represents the joint effect of area characteristics and $\epsilon_{ij}$ is the random effect associated with the $j$-th unit within the $i$-th area. The formula (1) can be write in matrix form as:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + u_i\mathbf{1}_i + \boldsymbol{\epsilon}_i \quad i = 1...m. \tag{2}$$

where $\mathbf{X}_i$ is $N_i \times p$, $\mathbf{y}_i$, $\mathbf{1}_i = (1, ...1)^T$ and $\boldsymbol{\epsilon}_i$ are $N_i \times 1$ vectors.

If the sampling rate $f_i = n_i/N_i$ is negligible, the small area means can be taken as:

$$\bar{Y}_i = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + u_i + \bar{\epsilon}_i \tag{3}$$

where $\bar{\epsilon}_i \cong 0$ is the mean of the $N_i$ errors $\epsilon_{ij}$; then it follows that the estimation of the target parameters $\theta_i$ are approximately equal to $\theta_i = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + u_i$. For known variances $\sigma_u^2$ and $\sigma_\epsilon^2$, the BLUP of $\theta_i$ under the model is:

$$\hat{\theta}_i = \gamma_i[\bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)^T\hat{\boldsymbol{\beta}}] + (1 - \gamma_i)\bar{\mathbf{X}}_i^T\hat{\boldsymbol{\beta}} \;\; with \; i = 1...m, \tag{4}$$

where $\gamma_i = \frac{\sigma_u^2}{(\sigma_u^2+\sigma_\epsilon^2/n_i)}$ is the shrinkage factor, $\hat{\boldsymbol{\beta}}$ is the weighted least squares estimate vector of $\boldsymbol{\beta}$ and $\bar{\mathbf{x}}_i$ is the sample mean of $\mathbf{x}_i$. For area k with no samples, $\hat{\theta}_k = \bar{\mathbf{X}}_k^T\hat{\boldsymbol{\beta}}$. In practice, the variances $\sigma_u^2$ and $\sigma_\epsilon^2$ are seldom known and they are estimated from the sample data, using the method of fitting constants (Battese *et al.*, 1988) or the restricted maximum likelihood (REML) method (Rao, 2001). The resulting predictors are known as the EBPLUP.

Thomsen (in Gosh and Rao, 1994) believes that predictors (4) tend to over-estimate area means with small random effects and under-estimate area means with large effects such that the variation between the predictors is smaller than the variation between the true means (Pfeffermann, 2002). Another critical point for the unit level models is that they assume that the sample values obey the assumed population model, that is, sample selection bias is absent (Rao, 2003).

## 3 Spatial Unit Level Random Effect Models

In order to take into account the correlation between neighboring areas we regarded to the spatial models and how these models could be utilized in small area estimation (Cressie, 1991). In this study a standard linear regression is considered and the spatial dependence has been incorporated in the error structure ($E[v_i, v_j] \neq 0$). It can be specified in a number of different ways, and results in a error variance covariance matrix of the form (Anselin, 1992):

$$E[v_i, v_j] = \Omega(\tau), \tag{5}$$

where $\tau$ is a vector of parameters, such as the coefficient in a Simultaneously Autoregressive (SAR) or Conditional Autoregressive (CAR) error process, and $v_i, v_j$ are the area random effects. A SAR error model is used:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} \tag{6}$$

where $\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u}$, $\rho$ is the spatial autoregressive coefficient, $\mathbf{W}$ is the spatial weight matrix for $\mathbf{y}$, $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I})$ is the error vector and

$$\mathbf{v} \sim \left(\mathbf{0}, \sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\right). \tag{7}$$

Spatial models are a special case of the general linear mixed model. Considering the spatial dimensions of the data, a model with spatially correlated errors could be implemented.

Suppose that the sample data obey the general linear mixed model and let $\mathbf{y}$ be the vector of values of the response variable $n \times 1$ ($n = \sum_i^m n_i$), $\mathbf{X}$ be the matrix of covariates $n \times p$, $\mathbf{v}$ the random area effect vector $m \times 1$ ($m$ is the number of small areas)and $\boldsymbol{\epsilon}$ the error vectors $n \times 1$. Let $\mathbf{Z}$ be the incidence matrix $n \times m$ for the random effect vector $\mathbf{v}$. The model can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \boldsymbol{\epsilon} \tag{8}$$

with the incidence matrix is

$$\mathbf{Z} = \begin{pmatrix} \mathbf{1}_{n_1} & 0 & \ldots & 0 \\ 0 & \vdots & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_{n_m} \end{pmatrix}$$

where $\mathbf{1}_{n_i}$ is a vector of dimension $n_i$ with all elements equal to one. The error vector $\boldsymbol{\epsilon}$ and area effect vector $\mathbf{v}$ are mutually independent error terms with zero mean vectors and covariances matrices given $\sigma_\epsilon^2\mathbf{I}_n$ and $\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}$ respectively. The model (8) can be rewritten as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} + \boldsymbol{\epsilon}. \tag{9}$$

It follows that the covariance matrices of the studied variable is:

$$\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T = \sigma_\epsilon^2\mathbf{I}_n + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T \tag{10}$$

For known $\sigma_u^2$, $\sigma_\epsilon^2$ and $\rho$ the Spatial BLUP estimator of a parameter $\boldsymbol{\theta}_i$ (the small area mean) is:

$$\tilde{\theta}_i^S(\sigma_u^2, \sigma_\epsilon^2, \rho) = \bar{\mathbf{X}}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T\{\sigma_u^2[(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^T)]^{-1}\}\mathbf{Z}^T\{\sigma_\epsilon^2\mathbf{I}_n + \mathbf{Z}\sigma_u^2[(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \tag{11}$$

where $\bar{\mathbf{X}}_i$ are the known population mean, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ and $\mathbf{b}_i^T$ is $1 \times m$ vector $(0, 0, ...0, 1, 0, ...0)$ with 1 in the $i$-th position.

The $MSE[\tilde{\theta}_i^S(\sigma_u^2, \sigma_\epsilon^2, \rho)]$, depending on three parameters $(\sigma_u^2, \sigma_\epsilon^2, \rho)$, can be expressed as:

$$MSE[\tilde{\theta}_i^S(\sigma_u^2, \sigma_\epsilon^2, \rho)] = g_{1i}(\sigma_u^2, \sigma_\epsilon^2, \rho) + g_{2i}(\sigma_u^2, \sigma_\epsilon^2, \rho) \tag{12}$$

with

$$\begin{aligned} g_{1i}(\sigma_u^2, \sigma_\epsilon^2, \rho) = \mathbf{b}_i^T\{\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1} - \sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T \times \\ \{\sigma_\epsilon^2\mathbf{I}_n + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}\mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\}\mathbf{b}_i \end{aligned} \tag{13}$$

and

$$\begin{gathered} g_{2i}(\sigma_u^2, \sigma_\epsilon^2, \rho) = (\bar{\mathbf{X}}_i - \mathbf{b}_i^T\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T \times \\ \{\sigma_\epsilon^2\mathbf{I}_n + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}\mathbf{X}) \times \\ (\mathbf{X}^T\{\sigma_\epsilon^2\mathbf{I}_n + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}\mathbf{X})^{-1} \times \end{gathered}$$

$$(\bar{\mathbf{X}}_i - \mathbf{b}_i^T \sigma_u^2 [(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T \times \\ \sigma_\epsilon^2 \mathbf{I}_n + \mathbf{Z}\sigma_u^2[(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}\mathbf{X})^T. \tag{14}$$

The estimator $\tilde{\theta}_i^S(\sigma_u^2, \sigma_\epsilon^2, \rho)$ depends on the variance components $\sigma_u^2$, $\sigma_\epsilon^2$ and $\rho$, but in practice they will be unknown. Replacing the parameters with asymptotically consistent estimators $\hat{\sigma}_u^2$, $\hat{\sigma}_\epsilon^2$, $\hat{\rho}$, a two stage estimator $\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho})$ is obtained and it is called Spatial EBLUP:

$$\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) = \bar{\mathbf{X}}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T\{\hat{\sigma}_u^2[(\mathbf{I}-\hat{\rho}\mathbf{W})(\mathbf{I}-\hat{\rho}\mathbf{W}^T)]^{-1}\}\mathbf{Z}^T\{\hat{\sigma}_\epsilon^2\mathbf{I}_n + \boldsymbol{Z}\hat{\sigma}_u^2[(\mathbf{I}-\hat{\rho}\mathbf{W})(\mathbf{I}-\hat{\rho}\mathbf{W}^T)]^{-1}\boldsymbol{Z}\}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \tag{15}$$

with $\mathbf{b}_i^T = (0, 0, ...0, 1, 0, ...0)$ and 1 referred to $i$-th area. Assuming normality, $\sigma_u^2$, $\sigma_\epsilon^2$ and $\rho$ can be estimated both by ML and REML procedures. The ML estimators $\hat{\sigma}_{u_{ML}}^2$, $\hat{\sigma}_{\epsilon_{ML}}^2$ and $\hat{\rho}_{ML}$ 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 have a global maximum and some local maximums. The ML estimator obtained with the "scoring" algorithm depend from the selected starting point, while the "Nelder-Mead" method for the maximization of a function of $q$ variables depends on the comparison of function values at the $(q+1)$ vertices of a general simplex; it adapts itself to the local landscape, and contracts on the final maximum. It does not depend from the selected starting point and it is computationally compact but it is not fully efficient: it achieves a point that is close to the global maximum. For this reason it needs to use the "scoring" algorithm selecting as starting point the maximum that has been obtain by the "Nelder-Mead" method.

The "scoring" algorithm can be represented as:

$$\begin{bmatrix} \sigma_u^2 \\ \sigma_\epsilon^2 \\ \rho \end{bmatrix}^{(n+1)} = \begin{bmatrix} \sigma_u^2 \\ \sigma_\epsilon^2 \\ \rho \end{bmatrix}^{(n)} + [\mathcal{I}(\sigma_u^{2^{(n)}}, \sigma_\epsilon^{2^{(n)}}, \rho^{(n)})]^{-1} \cdot s\left[\hat{\boldsymbol{\beta}}(\sigma_u^{2^{(n)}}, \sigma_\epsilon^{2^{(n)}}, \rho^{(n)}), \sigma_u^{2^{(n)}}, \sigma_\epsilon^{2^{(n)}}, \rho^{(n)}\right] \tag{16}$$

where $s\left[\hat{\boldsymbol{\beta}}(\sigma_u^{2^{(n)}}, \sigma_\epsilon^{2^{(n)}}, \rho^{(n)}), \sigma_u^{2^{(n)}}, \sigma_\epsilon^{2^{(n)}}, \rho^{(n)}\right]$ is the vector of the partial derivatives of log-likelihood function with respect to $\sigma_u^2$ $\sigma_\epsilon^2$ and $\rho$, $\mathcal{I}^{-1}(\sigma_u^2, \sigma_\epsilon^2, \rho)$ is the inverse matrix of expected second derivatives minus log-likelihood function with respect to the variance components and $n$ indicates the number of iteration.

The ML procedure to estimate $\sigma_u^2$, $\sigma_\epsilon^2$ and $\rho$ does not consider the loss in degrees of freedom due to estimating $\boldsymbol{\beta}$. This drawback involves the use of REML method (Cressie, 1992). The "Nelder-Mead" method and the "scoring" algorithm are used and at convergence the REML estimators are obtained and the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_R$, $\hat{\sigma}_{u_R}^2$, $\hat{\sigma}_{\epsilon_R}^2$ and $\hat{\rho}_R$ has a diagonal structure $diag\left[\bar{\mathbf{V}}(\hat{\beta}_R), \bar{\mathbf{V}}(\hat{\sigma}_{u_R}^2, \hat{\sigma}_{\epsilon_R}^2, \hat{\rho}_R)\right] \approx diag\left[\bar{\mathbf{V}}(\hat{\beta}_{ML}), \bar{\mathbf{V}}(\hat{\sigma}_{u_{ML}}^2, \hat{\sigma}_{\epsilon_{ML}}^2, \hat{\rho}_{ML})\right]$ with

$$\bar{\mathbf{V}}(\hat{\boldsymbol{\beta}}_R) \approx \bar{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{ML}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \\ \bar{\mathbf{V}}(\hat{\sigma}_{u_R}^2, \hat{\sigma}_{\epsilon_R}^2, \hat{\rho}_R) \approx \bar{\mathbf{V}}(\hat{\sigma}_{u_{ML}}^2, \hat{\sigma}_{\epsilon_{ML}}^2, \hat{\rho}_{ML}) = \mathcal{I}^{-1}(\sigma_u^2, \sigma_\epsilon^2\rho). \tag{17}$$

The ML and REML estimators are robust, in fact they may work well even under non normal distributions (Jiang, 1996).

The MSE of Spatial EBLUP $\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho})$ is:

$$MSE[\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho})] \approx g_{1i}(\sigma_u^2, \sigma_\epsilon^2, \rho) + g_{2i}(\sigma_u^2, \sigma_\epsilon^2, \rho) + g_{3i}(\sigma_u^2, \sigma_\epsilon^2, \rho) \tag{18}$$

where $g_{3i}(\sigma_u^2, \sigma_\epsilon^2, \rho)$ is approximately

$$tr\left\{\begin{bmatrix} \mathbf{b}_i^T\left(\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1})\right) \\ \mathbf{b}_i^T\left(\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{I}_n\mathbf{V}^{-1})\right) \\ \mathbf{b}_i^T\left(\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})\right) \end{bmatrix} \mathbf{V}\times \right. \\ \left. \times \begin{bmatrix} \mathbf{b}_i^T\left(\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1})\right) \\ \mathbf{b}_i^T\left(\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{I}_n\mathbf{V}^{-1})\right) \\ \mathbf{b}_i^T\left(\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})\right) \end{bmatrix}^T \bar{\mathbf{V}}(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho})\right\} \tag{19}$$

with $\mathbf{C} = [(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]$ and $\mathbf{A} = \sigma_u^2[-\mathbf{C}^{-1}(2\rho\mathbf{W}\mathbf{W}^T - 2\mathbf{W})\mathbf{C}^{-1}]$. An estimator of $MSE[\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho})]$ can be expressed as:

$$mse[\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho})] \approx g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) + g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) + 2g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) \tag{20}$$

if $\hat{\sigma}_u^2$, $\hat{\sigma}_\epsilon^2$ and $\hat{\rho}$ are REML estimators. Otherwise, if ML procedure is used, the $mse[\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho})]$ is given by

$$mse[\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho})] \approx g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) - \mathbf{b}_{ML}^T(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) \bigtriangledown g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) + g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) + 2g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_\epsilon^2, \hat{\rho}) \tag{21}$$

with

$$\bigtriangledown g_{1i}(\sigma_u^2, \sigma_\epsilon^2, \rho) = \mathbf{b}_i^T \left\{ \begin{array}{c} (\mathbf{C}^{-1} - [\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\sigma_u^2\mathbf{C}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1})\mathbf{Z}\sigma_u^2\mathbf{C}^{-1} + \\ (-\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{I}_n\mathbf{V}^{-1}) \times \\ (\mathbf{A} - [\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\sigma_u^2\mathbf{C}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})\mathbf{Z}\sigma_u^2\mathbf{C}^{-1} + \end{array} \right.$$

$$\left. \begin{array}{c} +\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}]) \\ Z\sigma_u^2\mathbf{C}^{-1}) \\ +\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}]) \end{array} \right\} \mathbf{b}_i \tag{22}$$

and

$$\mathbf{b}_{ML}^T(\sigma_u^2, \sigma_\epsilon^2, \rho) = \frac{1}{2m} \left\{ \mathcal{I}^{-1}(\sigma_u^2, \sigma_\epsilon^2, \rho) \begin{bmatrix} tr[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1})\mathbf{X}] \\ tr[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(-\mathbf{V}^{-1}\mathbf{I}_n\mathbf{V}^{-1})\mathbf{X}] \\ tr[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})\mathbf{X}] \end{bmatrix} \right\}. \tag{23}$$

If the term $\mathbf{b}_{ML}^T(\hat{\sigma}_u^2, \hat{\rho})\bigtriangledown g_{1i}(\hat{\sigma}_u^2)$ is ignored, the use of ML estimators could lead to underestimation of MSE approximation.

# 4 Data and results

In 2000 a survey designed to estimate the amount of erosion delivered to the streams in the Rathbun Lake watershed was completed. The watershed, located in southern Iowa (USA), covers more than 365000 acres (147715 ha) in six counties and is divided into 61 sub-watersheds.

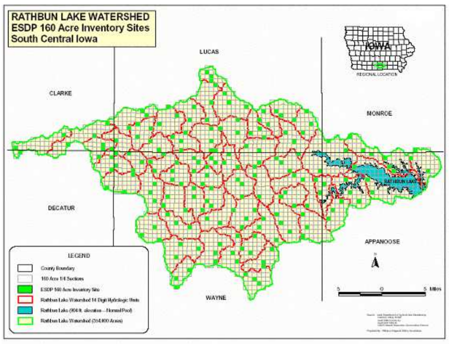
RATHBUN LAKE WATERSHED
ESDP 160 Acre Inventory Sites
South Central Iowa
LUCAS
CLARKE
MONROE
DECATUR
APPANOOSE
WAYNE
IOWA
REGIONAL LOCATION
RATHBUN LAKE
LEGEND
County Boundary
160 Acre 1/4 Sections
ESDP 160 Acre Inventory Site
Rathbun Lake Watershed 14 Digit Hydrologic Units
Rathbun Lake (904 ft. elevation -- Normal Pool)
Rathbun Lake Watershed (364,000 Acres)
5
0
5 Miles

Figure 1: The Watershed of Rathbun Lake

In the application the data are the result of this design: each small area (domain) has been divided in plots (total 2146), each plot has been sequentially labelled and a systematic sampling of plots has been selected. The fractional interval has been fixed in order to select four units from each small area (domain). Not all these $4 \times 61$ units have been included in the sample. From each domain a simple random sample of 3 units has been selected. Then within each sub-watershed, three 160-acre (64 ha) plots were selected, as is showed in Figure 1, and a sample of 183 units was

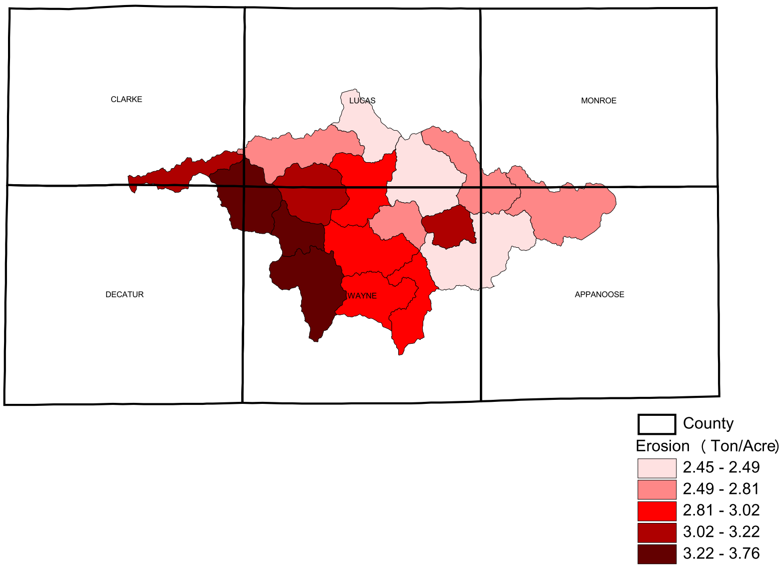

Figure 2: The quantity of the erosion estimated by REML method for the 17 HUC of Rathbun Lake.

obtained. The final sample can be reasonably assimilated to a simple random sample from the domains (for details Opsomer et al., 2003).

Auxiliary data at the plot and sub-watershed level were available for this study: the land use and the topography are considered major determinants of the erosion. Data related to these factor were available for the study region in the form of digital elevation and land use classification coverages. Hence, the Spatial EBLUP method is implemented to this data to estimate the average of watershed erosion in each of the 61 small area within the study region using SAR model. The neighborhood structure W is defined as follows: spatial weight, $w_{ij}$, is 1 if area $i$ shares an edge with area $j$ and 0 otherwise. For the study a new programme, running under the $R$ environment, using the "Nelder-Mead" algorithm and the "scoring" algorithm in sequence was implemented to estimate the parameters $(\sigma^2_u, \sigma^2_\epsilon, \rho)$, to calculate the Spatial EBLUP and its $mse$. The value of the estimated spatial autoregressive coefficient $\hat{\rho}$ is 0.116 ($s.e.$ = 0.0504) with ML procedure and 0.131 ($s.e.$ = 0.0632) with REML method, which suggests a moderate spatial relationship. Table 1 shows the value of estimated parameters and their standard error.

| ***Estimator*** | $\boldsymbol{\hat{\sigma}^2_u}$ | $\boldsymbol{\hat{\sigma}^2_\epsilon}$ | $\boldsymbol{\hat{\rho}_s}$ |
|---|---|---|---|
| $\boldsymbol{\tilde{\theta}^S(\hat{\sigma}^2_{u_{ML}}, \hat{\sigma}^2_{\epsilon_{ML}}, \hat{\rho}_{ML})}$ | 3536.05 (5973.08) | 74503.91 (9344.05) | 0.116 (0.0504) |
| $\boldsymbol{\tilde{\theta}^S(\hat{\sigma}^2_{u_R}, \hat{\sigma}^2_{\epsilon_R}, \hat{\rho}_R)}$ | 3631.75 (5959.45) | 75632.57 (9460.26) | 0.131 (0.0632) |

Table 1: Estimated parameters and their standard error.

To summarize, Figure 2 displays the map of the Rathbun Lake Watershed with the Spatial EBLUP estimates for the average erosion per acre in only 17 small areas, which are an aggregation of sub-watersheds.

In order to asses the achieved results with the introduction of the spatial information in the small area estimation, the direct estimator is also calculated. In Table 2 are reported the average estimated standard errors and its variability per acre of Direct and Spatial EBLUP estimators. Table 2 shows also the average estimated of $mse$ and its decomposition in $g_1$, due to the random effects, $g_2$, which accounts for the variability in the estimator $\hat{\boldsymbol{\beta}}$, $g_3$ due to estimate $\rho$, $\sigma^2_u$ and $\sigma^2_\epsilon$.

The Spatial EBLUP estimator provides estimates with smaller average estimated standard errors than the direct estimator.

| ***Estimator*** | ***A.E.Se.*** | ***A.E.mse*** | ***A.E.*(g₁)** | ***A.E.*(g₂)** | ***A.E.*(g₃)** |
|---|---|---|---|---|---|
| $\tilde{\boldsymbol{\theta}}^{S}(\hat{\boldsymbol{\sigma}}^{2}_{u_{ML}}, \hat{\boldsymbol{\sigma}}^{2}_{\epsilon_{ML}}, \hat{\boldsymbol{\rho}}_{ML})$ | 0.514 | 42.43 | 23.46 | 1.95 | 8.50 |
| $\tilde{\boldsymbol{\theta}}^{S}(\hat{\boldsymbol{\sigma}}^{2}_{u_{R}}, \hat{\boldsymbol{\sigma}}^{2}_{\epsilon_{R}}, \hat{\boldsymbol{\rho}}_{R})$ | 0.537 | 46.48 | 25.60 | 1.82 | 9.52 |
| **DIRECT** $\boldsymbol{\theta}$ | 0.886 | | | | |

Table 2: Average Estimated Standard Errors (A.E.Se.) of Direct and Spatial EBLUP estimators.

As it said above, small area models make use of explicit linking models based on random area-specific effects that account for between area variation beyond that is explained by auxiliary variables included in the model.

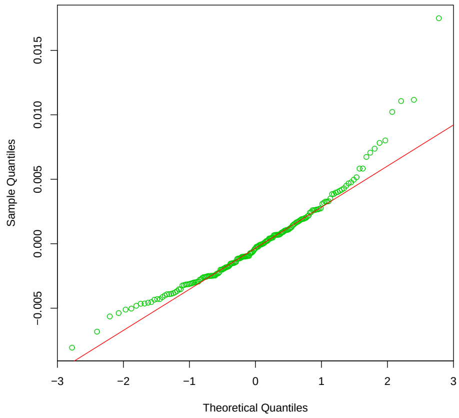

Figure 3: Normal q-q plot to check the normality of the standardized residuals $r_{ij}$

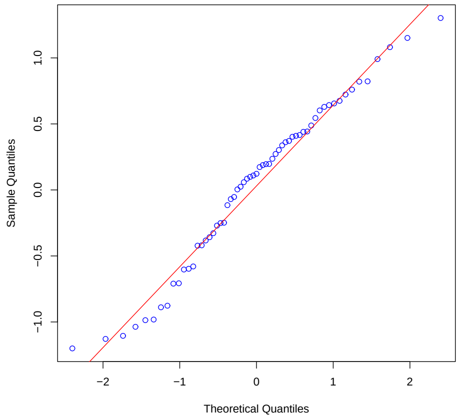

Figure 4: Normal q-q plot to check the normality of the small area effects

Inferences from model based estimators refer to the distribution implied by the assumed model. Model selection and validation play an important role in model based estimation, in fact if the assumed models do not perform a good fit to the data, the estimators will be model biased and can lead to wrong inferences. An evaluation of the resulting model is performed by treating the standardized EBLUP residuals $r_{ij} = (y_{ij} - x_{ij}^{T}\hat{\boldsymbol{\beta}} - \hat{v}_i)/\hat{\sigma}^2_{\epsilon}$. If $r_{ij}$ are approximately iid $N(0,1)$ variables, the model is valid. Residual plots of the standardized residuals show the effects of individual errors (Figure 3). It can be noted that the residuals are lightly skewed: perhaps because of the particular micro-climate which characterizes that region. But this issue should be more developed in order to better evaluate the performances of the model. To check the normality of the small area effects $v_i$ a normal q-q plot is examined (Figure 4). The Shapiro-Wilk W statistic gives value of 0.973 for small area effects, yielding p-value of 0.2048 that suggests no evidence against the hypothesis of normality.

In conclusion, considering the case study, the Spatial EBLUP methodology, which takes into account the SAR spatial model in the small area estimation, suggests a reduction of the width of the confidence interval.

**Acknowledgements**: the author thanks Jean Opsomer for the support providing the data and Prof. Chambers and Dr. Saei for their suggestions.

# References

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

CLIFF, A.D., ORD, J.K. (1981): *Spatial Processes. Models & Applications.* Pion Limited, London.

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. (1992): REML Estimation in Empirical Bayes Smoothing of Census Undercount. *Survey Methodology*, *18*, *1*, 75–94.

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

FAY, R.E., HERRIOT, R.A. (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 (with discussion). *Statistical Science*, *9*, *1*, 55–93.

HENDERSON, C.R. (1975): Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*, 423–447.

JIANG, J. (1996): REML Estimation: Asymptotic Behavior and Related Topics. *Annals of Statistics*, *24*, 255–286.

KACKAR, R.N., HARVILLE, D.A. (1984): Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, *79*, 853–862.

OPSOMER, J.D., BOTTS, C., KIM, J.Y. (2003): Small Area Estimation in Watershed Erosion Assessment Survey. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*, *2*, 139–152.

ORD, K. (1975): Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, *70*, *349*, 120–126.

PETRUCCI, A., SALVATI, N., PRATESI, M. (2003): Stimatore Combinato e Correlazione Spaziale nella Stima per Piccole Aree. *Dipartimento di Statistica e Matematica Applicata all'Economia*, *reports n. 240*, Pisa. (In Italian)

PFEFFERMANN, D. (2002): Small Area Estimation-New Developments and Directions. *International Statistical Review*, *70*, *1*, 125–143.

PRASAD, N., RAO, J.N.K. (1990): The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, *85*, *409*, 163–171.

RAO, J.N.K., YU, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics*, *22*, *4*, 511–528.

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

ROBINSON, G.K. (1991): That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, *6*, *1*, 15–51.

SAEI, A., CHAMBERS, R. (2003): Small Area Estimation: A Review of Methods based on the Application of Mixed Models. *Southampton Statistical Sciences Research Institute*, *WP M03/16*, Southampton.