Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2004/02

Small Area Estimation Using Spatial Information. The Rathbun Lake Watershed Case Study

Alessandra Petrucci,
Nicola Salvati

Università degli Studi
di Firenze

Statistics

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

Small Area Estimation Using Spatial Information. The Rathbun Lake Watershed Case Study

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) and a composite estimator.

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

KEY WORDS: Small area models; watershed erosion; spatial models; spatial EBLUP.

1 Introduction

The previous study (Opsomer *et al.*, 2003) discussed 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 independent, but in practice, especially 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 aim of this article is to estimate the average sub-watershed erosion taking into account the spatial dimension of the soil erosion data, collected on the Rathbun Lake Watershed, adapting a model with spatially correlated errors in the EBLUP estimator. As well the paper considers the possible gains from modelling the spatial correlation among small area random effects used to represent the unexplained variation of the small area target quantities are examined.

Section 2 introduces the small area models that include random area-specific effects and EBLUP estimator is showed. Section 3 reports the Spatial EBLUP. Section 4 shows the data and 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 Area Level Random Effect Models

Area level random effect models are used when auxiliary information is available only at area level. The basic area level model includes random area specific effects and the area specific auxiliary covariates $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ are related to the parameters of inferential interest θ_i (totals y_i, means $\bar{y}_i$):

$$\theta_i = \mathbf{x}_i\boldsymbol{\beta} + z_i u_i \quad with\ i = 1...m \tag{1}$$

where z_i are known positive constants, $\boldsymbol{\beta}$ is the regression parameters vector $p \times 1$, u_i are independent and identically distributed random variables with mean 0 and variance σ_u^2. Moreover it assumes that direct estimators $\hat{\theta}_i$ are available and design-unbiased:

$$\hat{\theta}_i = \theta_i + e_i \tag{2}$$

where e_i are independent sampling errors with mean 0 and known variance ψ_i. Combining (1) and (2) the obtained model is:

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

that is a special case of the general linear mixed model with diagonal covariance structure. The covariance matrices $m \times m$ of u and e are:

$$\mathbf{G} = \sigma_u^2\mathbf{I} \tag{4}$$

and

$$\mathbf{R} = diag(\psi_i) \tag{5}$$

with $\mathbf{I}$ is an identity matrix. Then the covariance matrix of the studied variable is:

$$\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T. \tag{6}$$

The Best Linear Unbiased Predictor (BLUP) estimator of θ_i is:

$$\tilde{\theta}_i(\sigma_v^2) = \mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{7}$$

where $\mathbf{b}_i^T$ is $1 \times m$ vector $(0, 0, ...0, 1, 0, ...0)$ with 1 referred to i-th area and $\boldsymbol{\beta}$ are estimated by generalized least squares: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$.

The BLUP estimator is a weighted average of the design-based estimator $\hat{\theta}_i$, and the regression-synthetic estimator $\mathbf{x}_i\hat{\boldsymbol{\beta}}$; it can be given by:

$$\tilde{\theta}_i(\sigma_v^2) = \gamma_i\hat{\theta}_i + (1 - \gamma_i)\mathbf{x}_i\hat{\boldsymbol{\beta}} \tag{8}$$

where $\gamma_i = \sigma_u^2/\sigma_u^2 + \psi_i$ is a weight $(0 \leq \gamma_i \leq 1)$, it is called shrinkage factor and it measures the uncertainty in modelling the θ_i (Ghosh and Rao, 1994).

The $MSE[\tilde{\theta}_i(\sigma_u^2)]$ depends on a variance parameter σ_u^2 and it is:

$$MSE[\tilde{\theta}_i(\sigma_u^2)] = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) \tag{9}$$

with

$$g_{1i}(\sigma_u^2) = \mathbf{b}_i^T(\mathbf{G} - \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{G})\mathbf{b}_i = \sigma_u^2 z_i^2\psi_i(\sigma_u^2 z_i^2 + \psi_i)^{-1} = \gamma_i\psi_i \tag{10}$$

and

$$g_{2i}(\sigma_u^2) = (\mathbf{x}_i - \mathbf{b}_i^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{x}_i - \mathbf{b}_i^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{X})^T =$$

$$= (1 - \gamma_i)^2\mathbf{x}_i\left[\frac{\sum_{i=1}^m \mathbf{x}_i^T\mathbf{x}_i}{(\sigma_u^2 z_i^2 + \psi_i)}\right]^{-1}\mathbf{x}_i^T \tag{11}$$

where $g_{1i}(\sigma_u^2)$ due to the random effects and $g_{2i}(\sigma_u^2)$ accounts for the variability in the estimator $\hat{\boldsymbol{\beta}}$ (Rao, 2003).

In practical applications σ_u^2 is unknown and it is replaced by an estimator $\hat{\sigma}_u^2$, a two stage estimator $\tilde{\theta}(\hat{\sigma}_u^2)$ is obtained and it is called Empirical BLUP (EBLUP). It has some properties:

1. it is unbiased for θ;
2. $E[\tilde{\theta}(\hat{\sigma}_u^2)]$ is finite;
3. $\hat{\sigma}_u^2$ is any translation invariant estimator of σ_u^2 (Kackar and Harville, 1984).

The variance of random effects can be estimated either by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) methods, assuming normality, or by the method of fitting constants. The MSE of EBLUP estimator appears to be insensitive to the choice of the estimator $\hat{\sigma}_u^2$. Under normality of random effects

$$MSE[\tilde{\theta}_i(\hat{\sigma}_u^2)] = MSE[\tilde{\theta}_i(\sigma_u^2)] + E[\tilde{\theta}_i(\hat{\sigma}_u^2) - \tilde{\theta}_i(\sigma_u^2)]^2 \tag{12}$$

where the last term is obtained as an approximation because is generally intractable:

$$E[\tilde{\theta}_i(\hat{\sigma}_u^2) - \tilde{\theta}_i(\sigma_u^2)]^2 \approx tr\left\{\left[\frac{\partial \mathbf{b}_i^T \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}}{\partial \sigma_u^2}\right]\mathbf{V}\left[\frac{\partial \mathbf{b}_i^T \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}}{\partial \sigma_u^2}\right]^T \bar{V}(\sigma_u^2)\right\} = g_{3i}(\sigma_u^2) =$$

$$= \psi_i^2 z_i^4 (\psi_i + \sigma_u^2 z_i^2)^{-3} \bar{V}(\sigma_u^2) \tag{13}$$

with $\bar{V}(\sigma_u^2)$ denoting the asymptotic variance of σ_u^2 which can be approximated as $\bar{V}(\hat{\sigma}_u^2)$. An approximation to the $MSE[\tilde{\theta}_i(\hat{\sigma}_u^2)]$ is

$$MSE[\tilde{\theta}_i(\hat{\sigma}_u^2)] \approx g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) \tag{14}$$

with $g_{2i}(\sigma_u^2)$ and $g_{3i}(\sigma_u^2)$ are of lower order than the term $g_{1i}(\sigma_u^2)$.

In practical application the estimator $\tilde{\theta}_i(\hat{\sigma}_u^2)$ has to be associated with an estimator of $MSE[\tilde{\theta}_i(\hat{\sigma}_u^2)]$. An approximately unbiased estimator of this mean square error is computed using the following expression:

$$mse[\tilde{\theta}_i(\hat{\sigma}_u^2)] \approx g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2) \tag{15}$$

when $\hat{\sigma}_u^2$ is obtained by REML method. Otherwise, if a ML procedure is used

$$mse[\tilde{\theta}_i(\hat{\sigma}_u^2)] \approx g_{1i}(\hat{\sigma}_u^2) - b_{ML}^T(\hat{\sigma}_u^2) \bigtriangledown g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2) \tag{16}$$

where $b_{ML}^T(\hat{\sigma}_u^2) \bigtriangledown g_{1i}(\hat{\sigma}_u^2)$ is an extra term due to the bias $g_{1i}(\hat{\sigma}_u^2)$ and it is of the same order as $g_{2i}(\sigma_u^2)$ and $g_{3i}(\sigma_u^2)$.

The area basic model considers the random area effects as independent. In practice, it should be more reasonable to assume that the random effects between the neighboring areas (for instance the neighborhood could be define by a distance criterium) are correlated and the correlation decays to zero as distance increases. Considering the spatial dimension of the data, a model with spatially autocorrelated errors has to be implemented, as it is shown in the next section.

3 Spatial Area 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:

$$E[v_i, v_j] = \Omega(\tau), \tag{17}$$

where τ 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:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} \tag{18}$$

where $\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u}$, ρ 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 direct area effect and

$$\mathbf{v} \sim \left(\mathbf{0}, \sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\right). \tag{19}$$

Spatial models are a special case of the general linear mixed model. Considering the spatial dimensions of the data, a new model with spatially correlated errors could be implemented and in matrix form it is:

$$\begin{aligned}\boldsymbol{\theta} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} \\ \hat{\boldsymbol{\theta}} &= \boldsymbol{\theta} + \mathbf{e}\end{aligned} \tag{20}$$

where $\boldsymbol{\theta}$ is the parameter of inferential interest, $\mathbf{X}$ is the matrix of area auxiliary information, $\boldsymbol{\beta}$ is the regression parameters vector $p \times 1$, $\mathbf{Z}$ is a matrix of known positive constants, $\mathbf{v}$ is defined as in (18), $\hat{\boldsymbol{\theta}}$ is the vector of the direct estimators, $\mathbf{e}$ represents the sampling errors with mean $\mathbf{0}$ and known variance $diag(\psi_i)$, $\mathbf{u}$ is a vector of independent and identically distributed random variables with mean $\mathbf{0}$ and variance $\sigma_u^2\mathbf{I}$ and m is the number of small areas. The covariance matrices $m \times m$ of $\mathbf{v}$ and $\mathbf{e}$ are:

$$\mathbf{G} = \sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1} \tag{21}$$

and

$$\mathbf{R} = diag(\psi_i). \tag{22}$$

Then the covariance matrix of the studied variable is:

$$\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T = diag(\psi_i) + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T \tag{23}$$

with $\mathbf{v}$ and $\mathbf{e}$ independently distributed. Combining the first and the second model in formula (20) the Spatial BLUP estimator of θ_i is:

$$\tilde{\theta}_i^S(\sigma_u^2, \rho) = \mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T\{\sigma_u^2[(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^T)]^{-1}\}\mathbf{Z}^T\{diag(\psi_i)+\sigma_u^2[(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^T)]^{-1}\}^{-1}(\hat{\boldsymbol{\theta}}-\mathbf{X}\hat{\boldsymbol{\beta}}) \tag{24}$$

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

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

$$MSE[\tilde{\theta}_i^S(\sigma_u^2, \rho)] = g_{1i}(\sigma_u^2, \rho) + g_{2i}(\sigma_u^2, \rho) \tag{25}$$

with

$$\begin{aligned}g_{1i}(\sigma_u^2, \rho) = \mathbf{b}_i^T\{\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1} - \sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T \times \\ \times\{diag(\psi_i) + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}\mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\}\mathbf{b}_i\end{aligned} \tag{26}$$

and

$$\begin{aligned}g_{2i}(\sigma_u^2, \rho) = (\mathbf{x}_i - \mathbf{b}_i^T\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T \\ \{diag(\psi_i) + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}\mathbf{X})\times \\ \times(\mathbf{X}^T\{diag(\psi_i) + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}\mathbf{X})^{-1}\times \\ (\mathbf{x}_i - \mathbf{b}_i^T\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T \\ \{diag(\psi_i) + \mathbf{Z}\sigma_u^2[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}\mathbf{Z}^T\}^{-1}\mathbf{X})^T.\end{aligned} \tag{27}$$

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

$$\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\rho}) = \mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T\{\hat{\sigma}_u^2[(\mathbf{I}-\hat{\rho}\mathbf{W})(\mathbf{I}-\hat{\rho}\mathbf{W}^T)]^{-1}\}\mathbf{Z}^T\{diag(\psi_i)+\hat{\sigma}_u^2[(\mathbf{I}-\hat{\rho}\mathbf{W})(\mathbf{I}-\hat{\rho}\mathbf{W}^T)]^{-1}\}^{-1}(\hat{\boldsymbol{\theta}}-\mathbf{X}\hat{\boldsymbol{\beta}}) \tag{28}$$

with $\mathbf{b}_i^T = (0, 0, ...0, 1, 0, ...0)$ and 1 referred to i-th area. Assuming normality, σ_u^2 and ρ can be estimated both by ML and REML procedures. The ML estimators $\hat{\sigma}_{u_{ML}}^2$ and $\hat{\rho}_{ML}$ can be obtained iteratively using the "scoring" algorithm:

$$\begin{bmatrix}\sigma_u^2 \\ \rho\end{bmatrix}^{(n+1)} = \begin{bmatrix}\sigma_u^2 \\ \rho\end{bmatrix}^{(n)} + [\mathcal{I}(\sigma_u^{2^{(n)}}, \rho^{(n)})]^{-1} \cdot s\left[\hat{\boldsymbol{\beta}}(\sigma_u^{2^{(n)}}, \rho^{(n)}), \sigma_u^{2^{(n)}}, \rho^{(n)}\right] \tag{29}$$

where $s\left[\hat{\boldsymbol{\beta}}(\sigma_u^{2^{(n)}},\rho^{(n)}),\sigma_u^{2^{(n)}},\rho^{(n)}\right]$ is the vector of the partial derivatives of log-likelihood function with respect to σ_u^2 and ρ, $\mathcal{I}^{-1}(\sigma_u^2,\rho)$ 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 σ_u^2 and ρ 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 "scoring" algorithm (29) is used and at convergence the REML estimators are obtained and the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_R$, $\hat{\sigma}^2_{u_R}$ and $\hat{\rho}_R$ has a diagonal structure $diag\left[\bar{\mathbf{V}}(\hat{\beta}_R),\bar{\mathbf{V}}(\hat{\sigma}^2_{u_R},\hat{\rho}_R)\right] \approx diag\left[\bar{\mathbf{V}}(\hat{\beta}_{ML}),\bar{\mathbf{V}}(\hat{\sigma}^2_{u_{ML}},\hat{\rho}_{ML})\right]$ with

$$\bar{\mathbf{V}}(\hat{\boldsymbol{\beta}}_R) \approx \bar{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{ML}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$$

$$\bar{\mathbf{V}}(\hat{\sigma}^2_{u_R},\hat{\rho}_R) \approx \bar{\mathbf{V}}(\hat{\sigma}^2_{u_{ML}},\hat{\rho}_{ML}) = \mathcal{I}^{-1}(\sigma_u^2,\rho). \tag{30}$$

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 $\tilde{\theta}_i^S(\hat{\sigma}_u^2,\hat{\rho})$ is:

$$MSE[\tilde{\theta}_i^S(\hat{\sigma}_u^2,\hat{\rho})] \approx g_{1i}(\sigma_u^2,\rho) + g_{2i}(\sigma_u^2,\rho) + g_{3i}(\sigma_u^2,\rho) \tag{31}$$

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

$$tr\left\{\begin{bmatrix}\mathbf{b}_i^T\left(\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1}+\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1})\right)\\ \mathbf{b}_i^T\left(\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1}+\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})\right)\end{bmatrix}\mathbf{V}\times\right.$$

$$\left.\times\begin{bmatrix}\mathbf{b}_i^T\left(\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1}+\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1})\right)\\ \mathbf{b}_i^T\left(\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1}+\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})\right)\end{bmatrix}^T\bar{\mathbf{V}}(\hat{\sigma}_u^2,\hat{\rho})\right\} \tag{32}$$

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 $MSE[\tilde{\theta}_i^S(\hat{\sigma}_u^2,\hat{\rho})]$ can be expressed as:

$$mse[\tilde{\theta}_i^S(\hat{\sigma}_u^2,\hat{\rho})] \approx g_{1i}(\hat{\sigma}_u^2,\hat{\rho}) + g_{2i}(\hat{\sigma}_u^2,\hat{\rho}) + 2g_{3i}(\hat{\sigma}_u^2,\hat{\rho}) \tag{33}$$

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

$$mse[\tilde{\theta}_i^S(\hat{\sigma}_u^2,\hat{\rho})] \approx g_{1i}(\hat{\sigma}_u^2,\hat{\rho}) - \mathbf{b}_{ML}^T(\hat{\sigma}_u^2,\hat{\rho}) \bigtriangledown g_{1i}(\hat{\sigma}_u^2,\hat{\rho}) + g_{2i}(\hat{\sigma}_u^2,\hat{\rho}) + 2g_{3i}(\hat{\sigma}_u^2,\hat{\rho}) \tag{34}$$

with

$$\bigtriangledown g_{1i}(\sigma_u^2,\rho) = \mathbf{b}_i^T\left\{\begin{array}{c}(\mathbf{C}^{-1} - [\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\sigma_u^2\mathbf{C}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1})\mathbf{Z}\sigma_u^2\mathbf{C}^{-1}+\\ (\mathbf{A} - [\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\sigma_u^2\mathbf{C}^{-1} + \sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})\mathbf{Z}\sigma_u^2\mathbf{C}^{-1}+\\ +\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}])\\ +\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}])\end{array}\right\}\mathbf{b}_i \tag{35}$$

and

$$\mathbf{b}_{ML}^T(\sigma_u^2,\rho) = \frac{1}{2m}\left\{\mathcal{I}^{-1}(\sigma_u^2,\rho)\begin{bmatrix}tr[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{V}^{-1})\mathbf{X}]\\ tr[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})\mathbf{X}]\end{bmatrix}\right\}. \tag{36}$$

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.

The main sources of agricultural erosion are sheet and rill, ephemeral gullies, gullies, and streambanks. The sheet and rill erosion was expected to be a major contributor to total erosion.

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 4×61 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 obtained. The final sample can be reasonably assimilated to a simple random sample from the domains and the sampling variance ψ_i at the domain level can be estimated by $\left\{(1 - \frac{n_i}{N_i})\frac{\hat{\sigma}_i^2}{n_i}\right\}$, where $n_i = 3$ and N_i is the number of plots in the i-th area (for details Opsomer et al., 2003). The estimated variance $\hat{\psi}_i$ is then treated as a proxy to ψ_i. As result the $mse[\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\rho}, \hat{\psi}_i)]$ is greater than $mse[\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\rho}, \psi_i)]$.

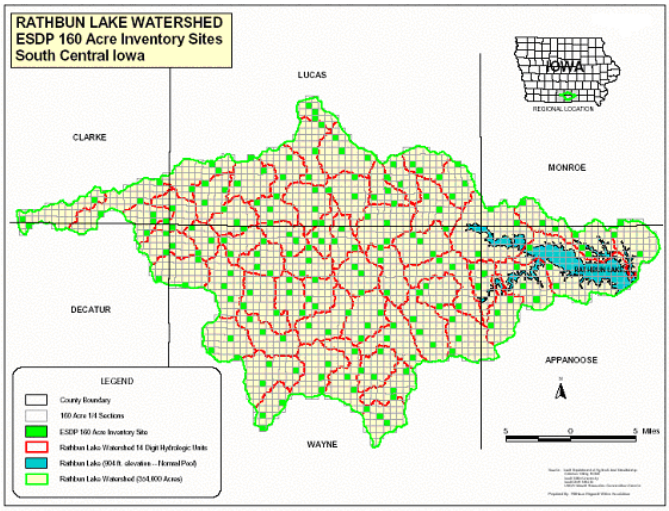

Figure 1: The Rathbun Lake

Auxiliary data at the sub-watershed level were the land use and the topography that 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. The value of the estimated spatial autoregressive coefficient $\hat{\rho}$ is 0.132 ($s.e. = 0.0258$) with ML procedure and 0.136 ($s.e. = 0.0288$) with REML method, which suggests a moderate spatial relationship. 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 EBLUP estimator and the direct estimator are also calculated. In Table 1 are reported the average estimated standard errors and its variability per acre of Direct, EBLUP and Spatial EBLUP estimators. Table 1 shows also the average estimated of MSE and

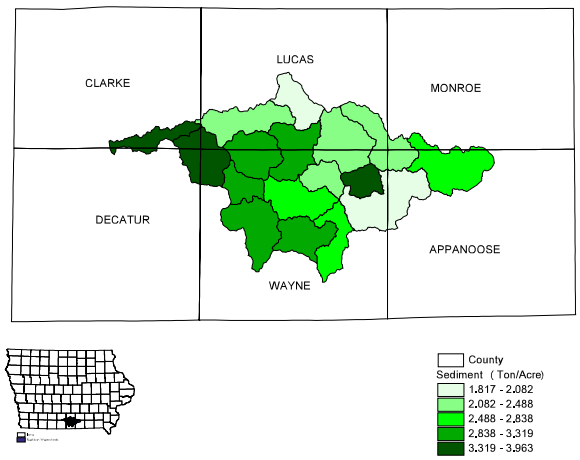

Figure 2: The 17 HUC of Rathbun Lake

Estimator	***A.E.Se.***	***V[A.E.Se.]***	***A.E.MSE***	***A.E.***$(\mathbf{g_1})$	***A.E.***$(\mathbf{g_2})$	***A.E.***$(\mathbf{g_3})$
$\tilde{\boldsymbol{\theta}}^{\boldsymbol{S}}(\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{u_{ML}}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{ML}})$	0.501	0.025	44.33	36.03	5.49	1.38
$\tilde{\boldsymbol{\theta}}^{\boldsymbol{S}}(\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{u_R}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{R}})$	0.510	0.027	45.96	36.92	5.65	1.68
$\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{u_{ML}}})$	0.545	0.034	52.76	45.21	5.66	0.92
$\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{u_R}})$	0.554	0.036	54.84	47.09	5.75	1.00
DIRECT $\boldsymbol{\theta}$	0.886	0.321				

Table 1: Average Estimated Standard Errors (A.E.Se.) of Direct, EBLUP and Spatial EBLUP estimators.

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 ρ and σ_u^2.

The Spatial EBLUP method provides estimates with smaller average estimated standard errors than the direct and the EBLUP estimators. Moreover the Spatial EBLUP presents the smallest variability. The estimate of the total watershed erosion in each of the 61 small area is reported in Annex A.

An evaluation of the resulting model is performed by treating the standard residuals $r = \tilde{\boldsymbol{\theta}}^S(\hat{\sigma}_u^2, \hat{\rho}) - \mathbf{X}\boldsymbol{\beta}/(diag(\mathbf{V}))^{1/2}$ as iid $N(0,1)$. In particular, to check the normality of the standardized residuals r and to detect outlier r, a normal q-q plot is examined (Figure 3). It can be noted that the outliers r are few, which correspond to neighboring areas in the north-west of the watershed; they can be originated from a particular micro-climate which characterizes that region. Nothing else significant departures from the assumed model are observed.

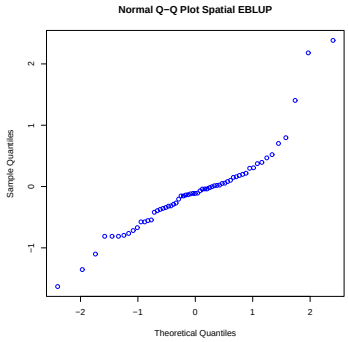

Figure 3: Normal q-q plot to check the normality of the standardized residuals r

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

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

A Annex

Area Code	$\tilde{\theta}^S(\hat{\sigma}^2_{u_R}, \hat{\rho}_R)$	$\widehat{\text{s.q.m.}}$	$\tilde{\theta}(\hat{\sigma}^2_{u_R})$	$\widehat{\text{s.q.m.}}$	$\hat{\theta}$	$\widehat{\text{s.q.m.}}$
10280201040010	2.831	0.217	2.779	0.218	2.796	0.225
10280201040020	4.167	0.249	4.166	0.253	4.364	0.264
10280201040030	3.228	0.368	3.122	0.379	3.306	0.419
10280201040040	4.306	0.494	4.074	0.541	5.085	0.701
10280201040050	4.144	0.606	3.540	0.654	3.572	0.992
10280201040060	4.436	0.188	4.404	0.189	4.512	0.193
10280201040070	4.035	0.633	2.713	0.766	3.404	2.525
10280201040080	2.711	0.496	2.213	0.546	1.939	0.714
10280201040090	3.258	0.467	3.345	0.496	3.882	0.605
10280201040100	2.037	0.373	1.977	0.388	1.786	0.434
10280201040110	2.562	0.516	2.357	0.565	2.148	0.761
10280201040120	2.098	0.530	2.109	0.586	1.754	0.809
10280201040130	3.625	0.512	3.497	0.547	3.680	0.712
10280201040140	3.033	0.687	2.806	0.722	3.765	1.542
10280201040150	3.584	0.388	3.560	0.396	3.704	0.441
10280201040160	2.001	0.444	1.968	0.475	1.656	0.569
10280201040170	2.732	0.610	2.961	0.723	4.455	1.565
10280201040180	5.141	0.652	5.178	0.650	4.953	0.700
10280201040190	3.241	0.429	3.300	0.448	3.543	0.525
10280201040200	2.234	0.377	2.419	0.393	2.408	0.439
10280201040210	2.307	0.512	2.413	0.574	2.154	0.779
10280201040220	1.267	0.300	1.230	0.307	1.016	0.327
10280201040230	1.868	0.543	2.021	0.596	1.489	0.841
10280201040240	2.486	0.606	2.598	0.650	2.668	1.043
10280201040250	2.076	0.509	2.383	0.546	2.009	0.710
10280201040260	1.816	0.441	1.766	0.463	1.437	0.547
10280201040270	1.815	0.551	1.808	0.586	1.160	0.806
10280201040280	3.006	0.306	3.133	0.311	3.221	0.332
10280201040290	2.184	0.529	2.409	0.577	2.313	0.790
10280201050010	2.760	0.606	2.927	0.699	2.766	1.273
10280201050020	2.583	0.559	2.851	0.629	3.183	0.959

Table A-1: Estimate of the total watershed in each small area and estimated Standard Errors (E.Se.) of Spatial EBLUP, EBLUP and Direct estimators. REML estimators

(Continued)

Area Code	$\tilde{\theta}^S(\hat{\sigma}^2_{u_R}, \hat{\rho}_R)$	$\widehat{\text{s.q.m.}}$	$\tilde{\theta}(\hat{\sigma}^2_{u_R})$	$\widehat{\text{s.q.m.}}$	$\hat{\theta}$	$\widehat{\text{s.q.m.}}$
10280201060010	3.166	0.666	2.951	0.747	4.366	1.961
10280201060020	2.738	0.492	2.695	0.524	2.721	0.666
10280201060030	3.656	0.350	3.603	0.351	3.615	0.366
10280201060040	3.281	0.501	3.167	0.523	2.801	0.644
10280201060050	4.197	0.584	4.358	0.617	4.500	0.787
10280201060060	3.012	0.600	3.118	0.690	3.462	1.257
10280201060070	3.098	0.665	3.117	0.747	4.498	1.865
10280201060080	3.028	0.430	2.974	0.438	2.985	0.504
10280201060090	3.236	0.671	3.531	0.777	4.780	1.608
10280201060100	2.801	0.586	2.710	0.659	2.316	1.091
10280201060110	2.566	0.523	2.659	0.551	2.556	0.726
10280201060120	2.510	0.642	2.801	0.702	3.447	1.359
10280201060130	2.132	0.256	2.117	0.259	2.044	0.270
10280201060140	2.271	0.618	2.761	0.666	2.939	1.130
10280201060150	3.107	0.572	3.097	0.632	2.819	0.952
10280201060160	2.727	0.642	2.869	0.745	4.093	1.903
10280201060170	3.916	0.664	4.264	0.750	5.931	1.518
10280201060180	2.275	0.428	2.338	0.447	2.121	0.521
10280201060190	1.849	0.394	1.777	0.410	1.423	0.464
10280201060200	2.300	0.564	2.491	0.619	2.421	0.911
10280201060210	4.584	1.311	4.927	1.407	6.024	3.131
10280201060220	2.802	0.245	2.877	0.248	2.870	0.257
10280201060230	1.900	0.549	2.130	0.617	1.633	0.910
10280201060240	1.971	0.426	1.999	0.453	1.728	0.533
10280201060250	1.758	0.270	1.750	0.274	1.621	0.288
10280201070010	2.274	0.457	2.313	0.476	2.083	0.571
10280201070020	3.161	0.552	3.308	0.606	3.836	0.878
10280201070030	2.786	0.616	2.689	0.703	2.776	1.376
10280201070040	2.635	0.566	2.618	0.625	2.800	0.937
10280201070050	2.415	0.594	2.542	0.670	2.391	1.134

References

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

ARBIA, G., ESPA, G. (1996). *Statistica Economica Territoriale*. CEDAM, Padova.

BAILEY, T.C., GATRELL, A.C. (1995): *Interactive Spatial Data Analysis*. Longman, London.

BATTESE, G.E., HARTER, R.M., 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*, *401*, *28–36*.

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

COWLING, A., CHAMBERS, R., LINDSAY, R., PARAMESWARAN, B. (1996): Applications of Spatial Smoothing to Survey Data. *Survey Methodology*, *22*, *2*, 175–183.

CRESSSIE, 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.

CRESSSIE, 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.

ESTEVAO, V.M., SÄRDNAL, C.E. (1999): The use of Auxiliary Information in Design-Based Estimation for Domains. *Survey Methodology*, *25*, *2*, 213–221.

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.

NOBLE, A., HASLETT, S., ARNOLD, G. (2002): Small Area Estimation via Generalized Linear Models. *Journal of Official Statistics*, *18*, *1*, 45–60.

OPENSHAW, S., TAYLOR, P. (1981): The Modifiable Unit Problem. In: N. Wrigley: (Eds.) *Quantitative Geography*. Pion, London, 127–144.

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. (1998): Small Area Estimation. *Encyclopedia of Statistical Sciences*, *2*, 621–628.

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

REY, S.J., MONTOURI, B.D. (1999): U.S. Regional Income Convergence: A Spatial Econometric Perspective. *Regional Studies*, *33.2*, 143–156.

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.

SINGH, A.C., MANTEL, H.J., THOMAS, B.W. (1994): Time Series EBLUPs for Small Areas Using Survey Data. *Survey Methodology*, *20*, *1*, 33–43.

UPTON, G.J.G., FINGLETON, B. (1985): *Spatial Data Analysis by Example*. John Wiley & Sons, New York.

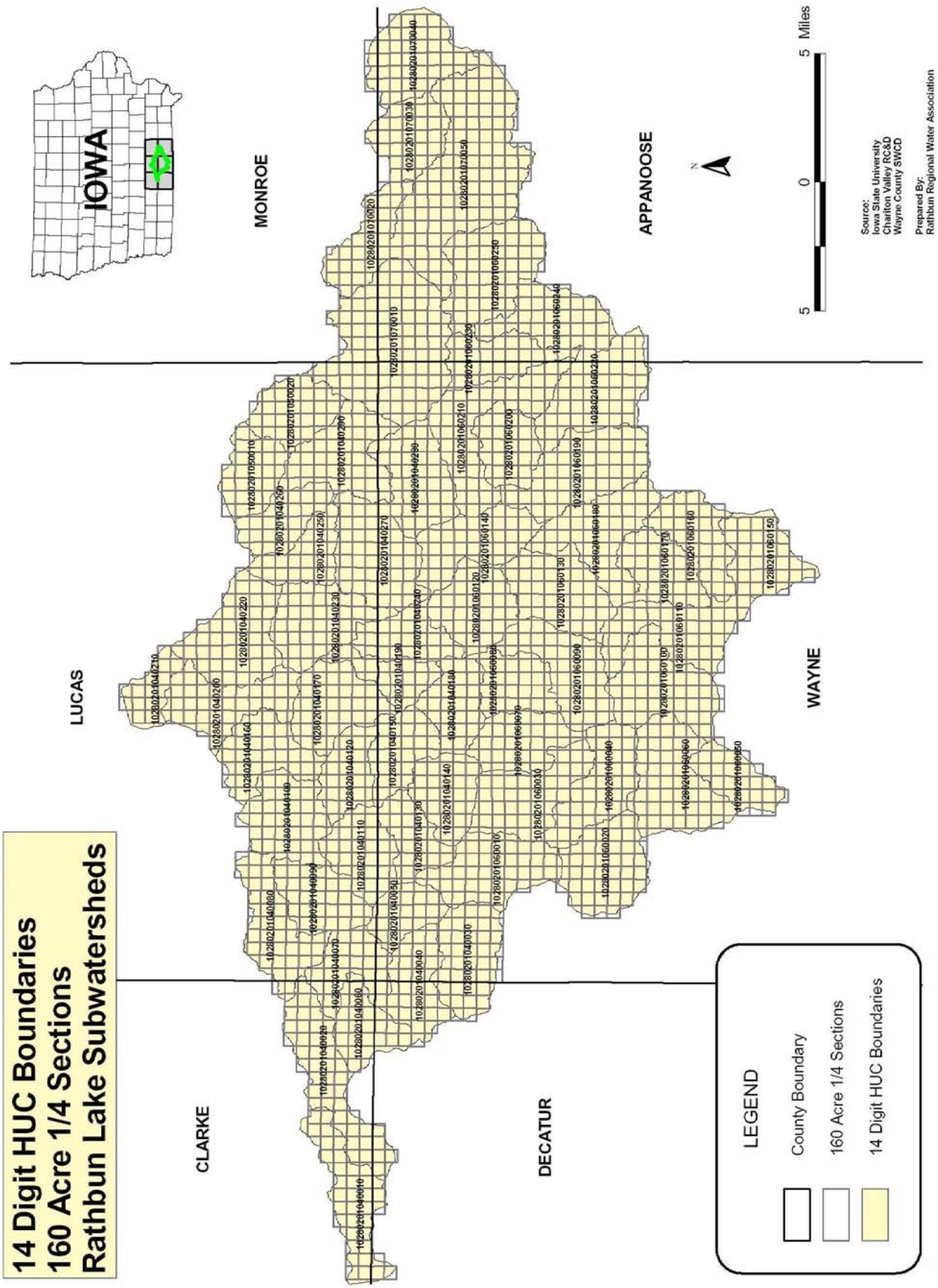

Figure A-1: Map of the area code in the Rathbun Lake Wathershed