

Small Area Estimation Considering Spatially Correlated Errors

Stima per Piccole Aree con Errori Spazialmente Autocorrelati

Alessandra Petrucci, Nicola Salvati

Dipartimento di Statistica “G.Parenti”, viale Morgagni 59, 50134 Firenze,

e-mail: alex@ds.unifi.it, salvati@ds.unifi.it

Riassunto: L'esigenza sempre più frequente di ottenere stime relative a zone e domini ridotti che non sono stati considerati tali al momento del disegno dell'indagine, ha determinato lo sviluppo di metodi statistici di stima per piccole aree.

Questo lavoro ha lo scopo di valutare la possibilità di migliorare le stime dei parametri di interesse attraverso l'introduzione di un modello spaziale nello stimatore EBLUP. A differenza dei comuni modelli di stima per piccole aree, che ipotizzano effetti casuali non correlati tra zone, l'applicazione di un modello spaziale permette di considerare la plausibile presenza di correlazione tra valori del fenomeno in “aree vicine” secondo opportuni criteri di distanza. La metodologia sviluppata viene applicata allo studio dell'erosione del suolo di un bacino idrografico sito nello stato dell'Iowa (USA).

Keywords: EBLUP, Spatial model, Natural resources survey.

1. Introduction

Domains may be defined by geographic areas (state, county, municipality, school district, etc.) or socio-demographic groups or other sub-populations (age-sex-race group within a large geographic area). 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. The demand for reliable statistics for small areas, when only reduced sizes of the samples are available, has promoted the development of statistical methods from both the theoretical and empirical point of view. In this perspective, 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, it should be more reasonable to assume that the random effects between the neighboring areas are correlated and the correlation decays to zero as distance increases.

The aim of this paper 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).

2. Spatial area level random effect models

Small area models can be classified into two broad types:

- **area level random effect models**, which are used when auxiliary information is available only at area level. They relate small area direct estimators to area-specific covariates;

- **nested error unit level regression models**, employed originally by Battese *et al.* (1988) for predicting areas under corn and soybeans in 12 counties of the state of Iowa (USA). These models relate the unit values of a study variable to unit-specific covariates (Pfeffermann, 2002).

Among the mixed effects models, a model with random effects at the area level (area level random effects model) is considered and in order to take into account the correlation between neighboring areas, a spatial model is used in small area estimation. Particularly, a standard linear regression is considered and the spatial dependence has been included in the error structure ($E[v_i, v_j] \neq 0$). This approach is appropriate in order to correct the potentially biasing influence of the spatial autocorrelation, due to the use of spatial data.

The Simultaneous AutoRegressive error model (SAR) used (Cressie, 1993) is:

$$y = X\beta + v$$

where X is a known full rank matrix, $v = \rho Wv + u$ is distributed with mean 0 and covariance matrix $G = \sigma_u^2 [I - \rho W]^{-1} [I - \rho W^T]^{-1}$ where I is the identity matrix, ρ is the spatial autoregressive coefficient, W is the spatial weight matrix for y and $u \sim N(0, \sigma_u^2)$ is the direct area effect.

Considering the spatial dimensions of the data (Anselin, 1992), a new area level random effect model with spatially correlated errors could be implemented:

$$\begin{aligned} \mathcal{Y} &= X\beta + (I - \rho W)^{-1}u \\ \hat{\mathcal{Y}} &= \mathcal{Y} + e \end{aligned}$$

where $\mathcal{Y}$ is the parameter of interest (total, mean, etc.), $\hat{\mathcal{Y}}$ is a direct estimator, e represents the sampling error with mean 0 and known diagonal covariance matrix $R = \text{diag}(\sigma_{Di}^2)$. The variance-covariance structure is:

$$V = R + G = \text{diag}(\sigma_{Di}^2) + \sigma_u^2 [I - \rho W]^{-1} [I - \rho W^T]^{-1}$$

as v and e are independently distributed. The Spatial BLUP estimator assumes the form:

$$\begin{aligned} \tilde{\mathcal{Y}} &= X\hat{\beta} + z^T G V^{-1} (\hat{\mathcal{Y}} - X\hat{\beta}) \\ \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} \hat{\mathcal{Y}} \end{aligned}$$

where z is a vector (0,0...0,1,...0) with value 1 referred to the i -th area.

The estimator $\tilde{\mathcal{Y}}$ depends on the variance parameters ρ and σ_u^2 which are unknown in empirical applications. Then the estimate of $\hat{\rho}$ and $\hat{\sigma}_u^2$ are substituted to the Spatial BLUP and the Spatial Empirical BLUP (Spatial EBLUP) is obtained:

$$\begin{aligned} \tilde{\mathcal{Y}} &= X\hat{\beta} + z^T \hat{G} \hat{V}^{-1} (\hat{\mathcal{Y}} - X\hat{\beta}) \\ \hat{\beta} &= (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} \hat{\mathcal{Y}}. \end{aligned}$$

Assuming normality, the parameters can be estimated using maximum likelihood (ML) or the restricted maximum likelihood (REML) by “scoring” algorithm (Rao, 2003):

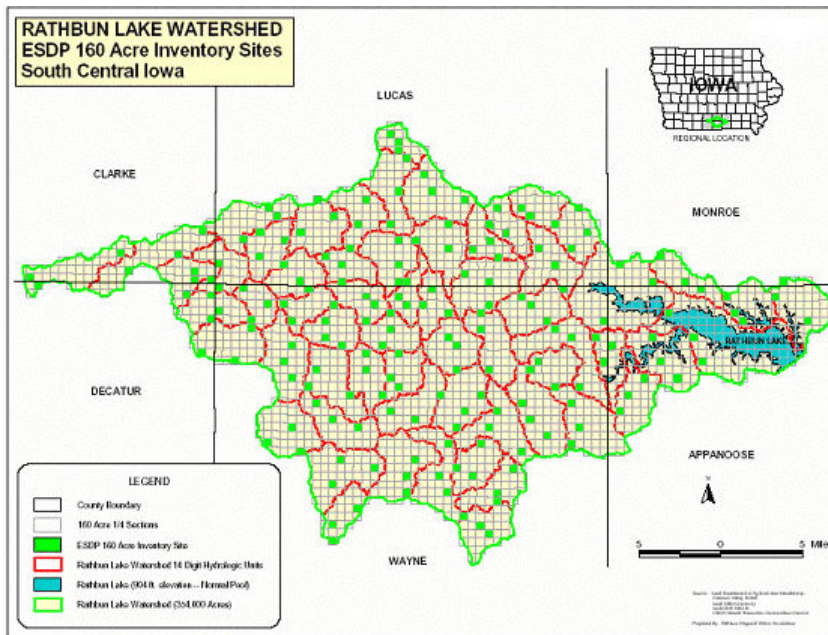
$$\begin{bmatrix} \rho \\ \sigma_u^2 \end{bmatrix}^{(n+1)} = \begin{bmatrix} \rho \\ \sigma_u^2 \end{bmatrix}^{(n)} + [I(\rho^{(n)}, \sigma_u^{2(n)})]^{-1} \cdot d[\hat{\beta}(\rho^{(n)}, \sigma_u^{2(n)}), \rho^{(n)}, \sigma_u^{2(n)}]$$

where $d[\hat{\beta}(\rho^{(n)}, \sigma_u^{2(n)}), \rho^{(n)}, \sigma_u^{2(n)}]$ is the partial derivatives vector of the log likelihood function $l(\beta, \rho, \sigma_u^2)$; $I(\rho^{(n)}, \sigma_u^{2(n)})$ is the matrix of expected second derivatives of $-l(\beta, \rho, \sigma_u^2)$ and n indicates the number of iteration. The partial derivatives and the information matrix are not reported, more details are in Salvati (2004).

3. 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. 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. (Opsomer *et al.*, 2003).

Figure 1: Map of Rathbun Lake watershed with sampled 160-acre plots.



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 total watershed erosion in each of the 61 small area within the study region. 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) which suggests a moderate spatial relationship. In order to assess 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 its decomposition in g_1 , due to the random effects, g_2 , which accounts for the variability in the estimator $\hat{\beta}$, g_3 due to estimate ρ and σ_u^2 .

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

Estimator	A.E.Se.	Var[A.E.Se.]	A.E.MSE	A.E.(g_1)	A.E.(g_2)	A.E.(g_3)
Spatial EBLUP	0.501	0.025	44.33	36.03	5.49	1.38
EBLUP	0.545	0.034	52.76	45.21	5.66	0.92
Direct	0.886	0.321				

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. 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. To assess the accuracy of the Spatial EBLUP estimator a simulation study was carried out. The results (not reported here for lack of space) show a slight improvement of the MSE estimates (Salvati, 2004).

Acknowledgements

The authors are grateful to Jean Opsomer who provided the Rathbun Lake Watershed data.

References

- Anselin, L. (1992) *Spatial Econometrics: Method and Models*, Kluwer Academic Publishers, Boston.
- Battese, G., Hartet, R., and Fuller, W. (1988) An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, *J. of the American Statistical Association*, 83, 28-36.
- Cressie, N. (1993) *Statistics for Spatial Data*, John Wiley & Sons, New York.
- Opsomer, J.D., Botts, C., Kim, J.Y. (2003) Small Area Estimation in Watershed Erosion Assessment Survey, *J. of the Agricultural, Biological, and Environmental Statistics*, 8, 2, 139-152.
- Pfeffermann, D. (2002) Small Area Estimation-New Developments and Directions, *International Statistical Review*, 70, 1, 125-143.
- Rao, J.N.K. (2003) *Small Area Estimation*, John Wiley & Sons, New York.
- Salvati, N. (2004) *La correlazione spaziale nella stima per piccole aree: metodi proposti e casi di studio*, Tesi di Dottorato in Statistica Applicata, Dipartimento di Statistica "G. Parenti", Università degli Studi di Firenze.