

Generalized linear mixed models for the interpolation of non Gaussian spatial data: an application to the study of the soil fertility in countryside of Capitanata⁽¹⁾.

I Modelli Lineari Generalizzati a Effetti Misti per l'Interpolazione di Dati Spaziali Non Gaussiani: un'Applicazione allo studio della Fertilità del Suolo in Capitanata.

Barbara Cafarelli

Dipartimento di Scienze Statistiche, Università di Bari, Via C. Rosalba 53, 70124 Bari,
e-mail: bcafarelli@dss.uniba.it

Riassunto: Il kriging lognormale, trans-gaussiano e l'indicator kriging sono le tecniche di previsione indicate dalla Geostatistica, quando viene meno l'assunzione di normalità dei dati, necessaria affinché il kriging sia un BLUP (best linear unbiased predictor). L'idea di base è quella di trasformare i dati originari per poi applicare il kriging e infine ritrasformare nella scala originaria i risultati ottenuti. Tuttavia, le stime ottenute sono, in generale, distorte. Ciò accade sia perché nelle previsioni non si tiene conto della variabilità delle stime dei parametri della correlazione spaziale sia perché non tutte le trasformazioni inverse consentono di ottenere predittori non distorti. Pertanto, per analizzare una componente della fertilità del suolo, per la quale risulta irrealistica l'ipotesi di normalità, si utilizzano i modelli lineari generalizzati a effetti misti.

Keywords: Generalized Linear Mixed Model, Spatial Variability, Spatial Prediction.

1. Introduction

The main drawbacks related to the use of kriging are concerned with the assumption of Gaussianity of the underlying process and with the underestimation of the prediction variance, due to assuming that the spatial correlation parameters are known instead of estimated in the spatial predictor. The Gaussian assumption is not always realistic as in the case of counts, dicotomic variables or continuous but strongly skewed ones. When working with such data the possibilities suggested by traditional Geostatistics are the lognormal kriging, the trans-Gaussian kriging and the indicator kriging (Cressie, 1993; Gotway & Stroup, 1997). Specific problems connected with these interpolating techniques are the choice of the normalising transformation, the difficulty in interpreting parameter estimates in terms of the transformed scale and the bias of predictors induced by the back-transformation. These remarks have stimulated the use of model based techniques: generalized linear mixed models have been recently proposed (Diggle, Tawn & Moyed, 1998) to analyse and interpolate spatial data. These models, widely used for repeated measurements and correlated data, give a detailed description of the dependence of spatial data, since they explicitly consider an erratic component

⁽¹⁾ This work was supported by MURST grant "Metodi e tecniche statistiche per l'analisi di dati a struttura spaziale e spazio-temporale – COFIN '99".

separately from the spatially correlated and the systematic component of the process. For this reason they suit better to a more effective description of the dependence mechanisms hidden in the spatial phenomena (Berliner, 2000). In this paper estimation and interpolation are performed according to a classical inferential approach.

2. Generalized linear mixed models (GLMMs)

To define a GLMMs for spatial data we consider the spatial stochastic process $\mathbf{Y} = \{\mathbf{Y}(\mathbf{u}) : \mathbf{u} \in D\}$, where $\mathbf{u}$ is the spatial locations vector and $D \subset \mathbf{R}^2$. We assume $\mathbf{S} = (S(\mathbf{u}) : \mathbf{u} \in D)$ to be a second order stationary Gaussian process with null expected value $E[\mathbf{S}] = 0$ and variance covariance matrix $\sigma^2 \mathbf{H}_{11}(\phi)$, where σ^2 is the partial sill, ϕ is the range and $\mathbf{H}_{11}(\phi)$ is the correlation matrix defined by a valid correlation function whose values depend on the distance between locations. We also assume that the distribution of $\mathbf{Y}|\mathbf{S}$ belongs to exponential family. The resulting data vector $\mathbf{y}$ of length n is a realisation of $\mathbf{Y}$, and can be expressed in the following way

$$\mathbf{y} = \mathbf{M} + \mathbf{e} \quad (1)$$

with $\mathbf{M} = E(\mathbf{Y} | \mathbf{S}) = g^{-1}(\mathbf{d}\boldsymbol{\beta} + \mathbf{S})$ and $\text{var}(\mathbf{y} | \mathbf{S}) = \text{var}(\mathbf{e}) = \tau^2 \mathbf{I}$, where g is a monotonic differentiable link function, $\mathbf{d}$ is the design matrix of covariates associated with observed locations, $\boldsymbol{\beta}$ is the fixed effects parameter vector, $\mathbf{d}\boldsymbol{\beta}$ is the spatial trend, and $\mathbf{e}$ is a random error vector, with *iid* components, uncorrelated with random effects, with $E(e_i) = 0$, and constant variance equal to the nugget, $\text{var}(e_i) = \tau^2$.

3. Pseudo-likelihood inference and Spatial interpolation

In the GLMMs context likelihood based inference is complicated by the high dimensionality of the likelihood function. In the last few years various alternative estimation methods have been proposed as the one based on the pseudo-likelihood approach (Wolfinger & O'Connell, 1993) which has the advantage of being fast since it does not require to integrate unobserved random effects out of the likelihood function. The pseudo-likelihood approach (Wolfinger & O'Connell, 1993) is implemented in the SAS macro glimmix and consists in maximising the likelihood function of a linear mixed model for the data of a pseudo variable obtained by a linear transformation of the response variable. In this way, the random effects are treated as unknown parameters, and estimated together with fixed effects $\boldsymbol{\beta}$ and correlation parameters σ^2, ϕ and τ^2 . Let's suppose to predict the behaviour of the process on m unsampled locations and let $\mathbf{Y}_0$ be the corresponding random vector. In this case $E(\mathbf{Y}_0 | \mathbf{S}_0) = \mathbf{M}_0 = g^{-1}(\mathbf{d}_0\boldsymbol{\beta} + \mathbf{S}_0)$, where $\mathbf{d}_0$ and $\mathbf{S}_0$ are the design matrix of covariates and the spatial random effects associated with the unsampled locations and

$$\text{var} \begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}_0 \end{pmatrix} \equiv \mathbf{V} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{10} \\ \boldsymbol{\Sigma}_{01} & \boldsymbol{\Sigma}_{00} \end{bmatrix} \quad (2)$$

where $\boldsymbol{\Sigma}_{11} = \sigma^2 \mathbf{H}_{11}(\phi) + \tau^2 \mathbf{I}$, $\boldsymbol{\Sigma}_{10} = \sigma^2 \mathbf{H}_{10}(\phi)$ and $\boldsymbol{\Sigma}_{00} = \sigma^2 \mathbf{H}_{00}(\phi) + \tau^2 \mathbf{I}$ (Gotway & Stroup, 1997). The covariance between observed and unobserved locations $\boldsymbol{\Sigma}_{10}$ is not supposed to be influenced by the nugget. The GLMMs predictor would be obtained as an extension of the linear mixed model predictor (Christensen, 1991) and the GLM predictor (Gotway & Stroup, 1997)

$$\hat{\mathbf{Y}}_0 = g^{-1}(\mathbf{d}_0 \hat{\boldsymbol{\beta}} + \hat{\mathbf{S}}_0) + \boldsymbol{\Sigma}_{01} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{y} - g^{-1}(\mathbf{d} \hat{\boldsymbol{\beta}} + \hat{\mathbf{S}})) \quad (3)$$

where spatial correlation parameters, $\boldsymbol{\beta}$ and $\mathbf{S}$ can be estimated by the pseudo-likelihood method, but $\mathbf{S}_0$ is not estimable within the classical inferential approach (Diggle et al., 1998). An alternative interpolation over the study area can be obtained by prediction over the sampled locations

$$\hat{\mathbf{Y}} = g^{-1}(\mathbf{d} \hat{\boldsymbol{\beta}} + \hat{\mathbf{S}}) + \sigma^2 \mathbf{H}_{00}(\phi) \boldsymbol{\Sigma}_{00}^{-1} (\mathbf{y} - g^{-1}(\mathbf{d} \hat{\boldsymbol{\beta}} + \hat{\mathbf{S}})) \quad (4)$$

where the covariance matrix $\boldsymbol{\Sigma}_{01}$ is replaced by $\sigma^2 \mathbf{H}_{00}(\phi)$ and is not influenced by the nugget (Christensen, 1991). Notice that $\hat{\mathbf{Y}}$ is not an exact predictor.

4 Analysis of the soil fertility in a countryside of Capitanata

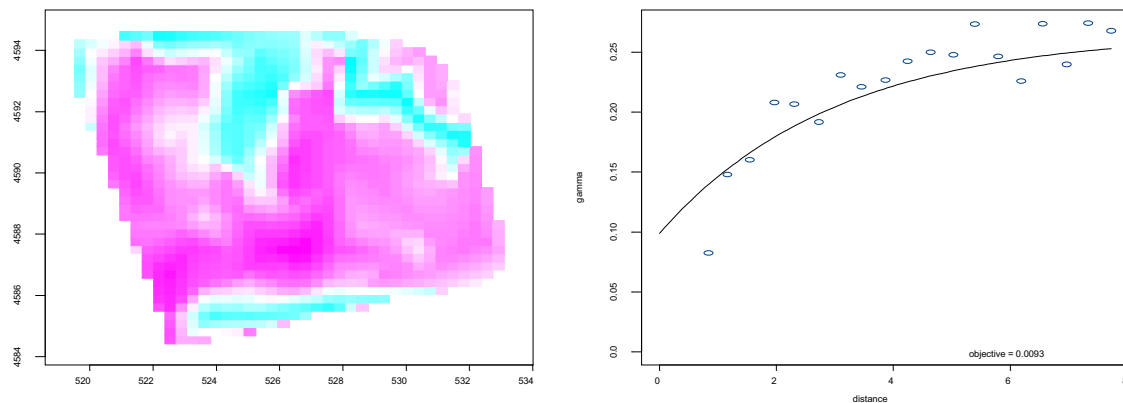
The implementation of precision agriculture in agronomic operations was the object of increasing interest in the last few years by both the agricultural industry and researchers. The strong spatial variability associated with all soil properties causes a loss of homogeneity relevant to the fertilisation practice that should fit to the real local needs of both the soil and the cultivation. The widely used empirical fertilisation practice that has no explicit consideration of the cultivation needs and of macro-nourishments of the soil, gives raise to various kinds of problems like, for examples, plant's nourishment lacks, loss of production, risk of water-bearing pollution. "Istituto Sperimentale Agronomico-Bari" supplied data over a sampled area with an extension of about 10.000 ha, traditionally cropped with durum wheat. It is located in Capitanata (Foggia). The sample was built according to a square grid of about 1.000 m side. The soil samples were taken at a depth of 0-50 cm and then analysed in a laboratory. In order to obtain information about soil fertility, we analysed the spatial distribution of ammoniacal nitrogen, because it is highly soluble in the water, and so, an exceeding quantity of it in the fertiliser can cause the pollution of the water-bearing. Exploratory data analysis showed that ammoniacal nitrogen has a positive skewed marginal distribution similar to a gamma distribution. A logarithmic transformation of the variable was carried out and the interpolation was performed by lognormal kriging. Anyway the results were not satisfactory, because the variogram of the residuals still showed the presence of some spatial dependence. So the analysis and the interpolation were carried out by a GLMM, where ammoniacal nitrogen showed a spatial trend on the longitude. Among the three

estimated models, whose values are showed in table 1, the choice, made according to AIC and -2 Res Log Lik values, fell on the exponential spatial correlation model. Interpolation were carried out (Figure 1) and, in this case, the variogram of the residuals was flat. Finally the fitting of the model obtained by comparing the empirical variogram with the theoretical one (Figure 1) was good.

Table 1: *Model parameter estimates*

Correlazione	σ^2	ϕ	τ^2	-2 Res Log Lik	AIC
Esponenziale	0.1667	2.9989	0.09893	-109.5	-103.5
Sferica	0.1452	3.1771	0.10010	-83.3	-77.3
Gaussiana	0.1528	2.5644	0.11110	-45.5	-41.5

Figure 1: *Ammoniacal nitrogen map, empirical and theoretical variogram*



References

- Berliner L.M. (2000) Hierarchical Bayesian Modelling in the Environmental Sciences. *Technical Report*. Ohio State University.
- Castrignanò A. et al. (2000). Valutazione e descrizione della fertilità di terreni meridionali mediante la Geostatistica Multivariata. *Technical Report*. Istituto Sperimentale Agronomico di Bari.
- Christensen, O.F., Moller, J. e Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo. *Research Report R-00-2009*. Aalborg University.
- Christensen R. (1991). *Linear Models for Multivariate, Time Series and Spatial Data*. Springer Verlag.
- Cressie N.A.C. (1993). *Statistics for Spatial Data*. John Wiley and Sons.
- Diggle P.J., Moyeed R.A., Tawn J.A. (1998). Model-based Geostatistics. *Applied Statistics*, 47, 299-350.
- Gotway C.A., Stroup W.W. (1997). A Generalised Linear Model Approach to Spatial Data Analysis and Prediction. *Journal of Agricultural, Biological and Environmental Statistics*, 2, 157-178.
- Stein M.L. (1999). *Interpolation of Spatial Data – Some Theory for Kriging*. Springer.
- Wolfinger R., O'Connell M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *J. Statist. Comput. Simul*, 48, 233-243.