Spatial analysis of rainfall erosivity in a Spanish Mediterranean area

J. Mateu[1], P. Juan[1], C. Añó[2] & C. Antolín[2,3]
[1]*Department of Mathematics, Universitat Jaume I, Castellon, Spain*
[2]*Land Use Planning Department. Centro de Investigaciones sobre Desertificacion (CIDE), Valencia, Spain*
[3]*Department of Vegetal Biology, University of Valencia, Valencia, Spain*

Abstract

The problem of estimation and prediction of a spatial stochastic process, observed at irregular locations in space, is considered. Geostatistical techniques analyze and describe the spatial dependence and quantify the scale and intensity of the spatial variation providing the essential spatial information for local estimation. Environmental variables usually show spatial dependence among observations, which is an important drawback to traditional statistical methods. The statistical model proposed is a Gaussian spatial linear mixed model (GSLMM) which introduces in the model various interesting terms under a hierarchical structure: a trend part, a lying signal spatial process derived through latent processes and a residual spatial term. Uncertainty in the model parameters can be evaluated using Bayesian statistical tools. The proposed methodology is applied to model the spatial process of rainfall erosivity in a Mediterranean region.

1 Introduction

A feature common to the earth sciences is the nature of their data. Most of the properties of interest vary continuously in space and cannot be measured or recorded everywhere. Thus, to represent their variation the values of individual variables or class types at unsampled locations must be estimated from information recorded at sample sites. The need to define spatial variation precisely is clear, and geostatistics is largely the application of this theory. It embraces a set of stochastic techniques that take into account both the random and structured nature of spatial variables, the spatial distribution of sampling sites and the uniqueness of any spatial observation (Journel & Huijbregts [1], Goovaerts [2]). Geostatistical methods find wide applications, for example in soil science, meteorology, hydrology and ecology. A methodological framework for dealing with problems of this kind was motivated by problems in the mining industry.

An important tool in geostatistics is the kriging predictor. The term kriging refers to a least square linear predictor which, under certain stationarity assumptions, requires at least the knowledge of the covariance parameters and the functional form for the mean of the underlying random function. In practical grounds, the parameters are usually not known. The kriging predictor does not take their uncertainty into account, but uses plug-in estimates as if they were true. Bayesian inference provides a way to incorporate parameter uncertainty in the prediction by treating the parameters as random variables and integrating over the parameter space to obtain the predictive distribution of any quantity of interest (Ribeiro & Diggle, [3]).

The objectives of a geostatistical analysis are broadly of two kinds: estimation and prediction. Estimation refers to inference about the parameters of a stochastic model for the data. Prediction refers to inference about the realization of the unobserved signal $S(\mathbf{u})$.

2 Gaussian spatial linear mixed models

2.1 Data structure

Consider a finite set of spatial sample locations $u_1, u_2, \ldots, u_n$, within a region D and denote $\mathbf{u} = (u_1, u_2, \ldots, u_n)$. Geostatistical data consist of measurements taken at the sample locations $\mathbf{u}$. The data vector is denoted by $\mathbf{y}(\mathbf{u}) = (y(u_1), \ldots, y(u_n))$, and the data are regarded as being a realization of a spatial stochastic process $\{\mathrm{Y}(u); u \in D\}$. An arbitrary location is denoted by $\mathbf{u}$ and the region D is a fixed subset of $\Re^d$ with positive d-dimensional volume. We assume that $\mathbf{u}$ varies continuously throughout the region D.

2.2 A conditional model specification

The model assumed here considers that the variable Y is a noisy version of a latent spatial process, the signal $S(\mathbf{u})$. The noises are assumed to be Gaussian and conditionally independent given $S(\mathbf{u})$. The model is specified by:

(1) Covariates: the mean part of the model is described by the term $X(u_i)\beta$. $X(u_i)'$ denotes a vector of spatially referenced non-random variables at location u_i and β is the mean parameter.

(2) The underlying spatial process $\{S(\mathbf{u}) : \mathbf{u} \in \Re^d\}$ is a stationary Gaussian process with zero mean, variance σ^2 and correlation function $\rho(\mathbf{h}; \phi)$, where

Transactions on Ecology and the Environment vol 46, © 2001 WIT Press, www.witpress.com, ISSN 1743-3541

ϕ is the correlation function parameter and **h** is the vector distance between two locations.

(3) Conditional independence: the variables $Y(u_i)$ are assumed to be Gaussian and conditionally independent given the signal,

$$Y(u_i) \mid S \sim N(X(u_i)'\beta + S(u_i), \tau^2)\,. \tag{1}$$

2.3 A conditional model specification: a hierarchical structure

In some applications we may want to consider a decomposition of the signal $S(\mathbf{u})$ into a sum of latent processes $T_k(\mathbf{u})$ scaled by σ_k^2. Then, the model can be rewritten in a first level as:

$$\mathbf{Y}(\mathbf{u}) = \mathbf{X}(\mathbf{u})\beta + S(\mathbf{u}) + \varepsilon(\mathbf{u}) = \mathbf{X}(\mathbf{u})\beta + \sum_{k=1}^{K} \sigma_k \mathbf{T}_k(\mathbf{u}) + \varepsilon(\mathbf{u})\,. \tag{2}$$

In a second level $\mathbf{T}_k(\mathbf{u}) \sim N(0, R_k(\phi_k))$, $\mathbf{T}_1, \ldots, \mathbf{T}_K$ mutually independent and $\varepsilon(\mathbf{u}) \sim N(0, \tau^2 I)$. Finally, level 3 defines the prior distribution for the parameters.

The model components are: (a) $\mathbf{Y}(\mathbf{u})$ is a random vector related to the measurements at sample locations; (b) $\mathbf{X}(\mathbf{u})\beta = \mu(\mathbf{u})$ is the expectation of $\mathbf{Y}(\mathbf{u})$. $\mathbf{X}(\mathbf{u})$ is a matrix of fixed covariates measured at sample locations **u**; (c) $\mathbf{T}_k(\mathbf{u})$ is a random vector at sample locations, of a standardized latent stationary spatial process $\mathbf{T}_k$. It has zero mean, variance one and correlation matrix $R_k(\phi_k)$. The signal S is defined by the sum of scaled latent processes $S(u) = \sum_{k=1}^{K} \sigma_k T_k(u)$; (d) σ_k is a scale parameter; (e) $\varepsilon(\mathbf{u})$ denotes a spatially independent process (spatial white noise) with zero mean and variance τ^2.

3 Spatial prediction

In geostatistical problems, often the main interest is not parameter estimation but prediction of the variable at a set of locations. Denote by $Y(\mathbf{u}_0)$ (Y_0) the variable to be predicted at locations u_0.

The optimal point predictor, defined as the one which minimizes the prediction mean square error (MSE), is given by

$$\hat{Y}_0 = E[Y_0 \mid Y] \tag{3}$$

This predictor is called the *least squares predictor* and its prediction variance is given by $Var[Y_0 \mid Y]$. Finding the conditional expectation (3) or an approximation to it, is a central problem in geostatistics, and several methods have been proposed.

3.1 Lineal predictor assuming known parameters

The linear predictor which minimizes the MSE is called *simple kriging* (SK) predictor. The SK predictor requires knowledge of the mean and covariance parameters, i.e. the parameters of the trend, signal and noise should be provided. The SK predictor is of the form

$$\hat{Y}_{\mathrm{SK}}(u_0) = \lambda_0 + \sum_i \lambda_i Y(u_i) \tag{4}$$

The weights λ_i are such that the prediction MSE is minimum. Under the Gaussian model, and if all the parameters are known, the SK predictor coincides with the conditional expectation (3), and therefore is optimal.

3.2 Lineal predictor filtering the mean and assuming known covariance parameters

The *ordinary kriging* (OK) predictor filters a constant mean requiring only the knowledge of the covariance parameters. The OK predictor is of the form

$$\hat{Y}_{\mathrm{OK}}(u_0) = \sum_i \lambda_i (u_i) \tag{5}$$

The weights λ_i are such that the prediction MSE is minimum under the constraint $\Sigma\lambda_i=1$. This constraint ensures the unbiasedness of the estimator. The results provided by OK coincide with those obtained by SK with the scalar mean parameter β given by its generalized least squares estimator.

3.3 A model-based approach

If complete parametric specification for the model components is assumed the conditional expectation (3) can be assessed (Diggle et al. [4], Ribeiro & Diggle [3]). Consider, for example, the Gaussian model specified in (2) extended to include both Y and Y_0. The joint distribution is given by

$$(Y, Y_0 \mid \beta, \sigma^2, \phi, \tau^2) \sim N\left(\begin{bmatrix} X \\ X_0 \end{bmatrix} \beta ; \tau^2 I + \begin{bmatrix} V_y(\sigma^2,\phi) & v(\sigma^2,\phi) \\ v'(\sigma^2,\phi) & V_0(\sigma^2,\phi) \end{bmatrix} \right) \tag{6}$$

4 Bayesian inference for a geostatistical model

Let us focus now on parameter estimation and prediction results for a Bayesian analysis of geostatistical data. For this aim consider a simpler model than (2) defined as $\mathbf{Y}(\mathbf{u}) = \mathbf{X}(\mathbf{u})\beta + \sigma T(\mathbf{u})$, where $T_{\mathbf{u}} \sim (N, R_y(\phi))$ and $pr(\beta, \sigma^2, \phi)$, a prior distribution.

Transactions on Ecology and the Environment vol 46, © 2001 WIT Press, www.witpress.com, ISSN 1743-3541

4.1 Uncertainty in the mean parameter

In this case only the mean parameter β is unknown. The covariance parameters are known and the covariance matrix is written as $V(\sigma_*^2, \phi_*) = \sigma_*^2 R(\phi_*)$, and denoted by $\sigma_*^2 R$. The model considered here corresponds to the common situation in geostatistics where the mean is filtered and the covariance parameters are estimated by some method and plugged-in for predictions.
The joint probability distribution for (Y, Y_0) is a simpler version of (6), without the nugget effect and with only one latent process

$$(Y, Y_0 \mid \beta, \sigma_*^2, \phi_*) \sim N\left(\begin{bmatrix} X \\ X_0 \end{bmatrix}\beta; \sigma_*^2 + \begin{bmatrix} R_y & r \\ r' & R_0 \end{bmatrix}\right) \qquad (7)$$

and the associated marginal and conditional distributions are

$$(Y \mid \beta, \sigma_*^2, \phi_*) \sim N(X\beta; \sigma_*^2 R_y) \qquad (8)$$

and

$$(Y_0 \mid Y, \beta, \sigma_*^2, \phi_*) \sim N(X_0\beta + r'R_y^{-1}(y - X\beta); \sigma_*^2(R_0 - r'R_y^{-1}r)) \qquad (9)$$

4.1.1 Predictive distribution for a Conjugate prior

Assuming a Normal prior for the mean parameter

$$(\beta \mid Y, \sigma_*^2, \phi_*) \sim N(m_\beta; \sigma_*^2 V_\beta) \qquad (10)$$

the mean and variance of the predictive distribution will be

$$\begin{aligned} E[T_0 \mid Y] = {} & (X_0 - r'R_y^{-1}X)(V_\beta^{-1} + X'R_y^{-1}X)^{-1}V_\beta^{-1}m_\beta \\ & + [r'R_y^{-1} + (X_0 - r'R_y^{-1}X)(V_\beta^{-1} + X'R_y^{-1}X)^{-1}X'R_y^{-1}]y \end{aligned} \qquad (11)$$

$$\begin{aligned} Var[T_0 \mid Y] = \sigma_*^2[& R_0 - r'R_y^{-1}r + (X_0 - r'R_y^{-1}X) \\ & (V_\beta^{-1} + X'R_y^{-1}X)^{-1}(X_0 - r'R_y^{-1}X)'] \end{aligned} \qquad (12)$$

4.1.2 Predictive distribution for a Flat prior

Assuming a flat prior for the mean parameter, i.e. $p(\theta) \propto 1$, the mean and variance of the predictive distribution can be calculated from (11) and (12) with $V_\beta^{-1} \equiv 0$,

$$E[T_0 \mid Y] = (X_0 - r'R_y^{-1}X)\hat{\beta} + r'R_y^{-1}y \qquad (13)$$

$$Var[T_0 \mid Y] = \sigma_*^2[R_0 - r'R_y^{-1}r + (X_0 - r'R_y^{-1}X)(X'R_y^{-1}X)^{-1}(X_0 - r'R_y^{-1}X)'] \qquad (14)$$

Finally, the posterior for known mean parameter β can also be obtained from (11) and (12) considering $V_\beta \gg X'R_y^{-1}X$ or $V_\beta \equiv 0$.

5 Application to real data

5.1 Introduction to the data set

To evaluate the rainfall erosivity in the Mediterranean area, many parameters have been studied and the most famous is the R factor, the rain erosion agressivity, included in the Universal Soil Loss Equation (USLE). Here the importance of the R factor in soil erosion is the kinetic energy of each storm and its maximum intensity. Therefore, this factor follows the expression $R = E \times I30$, where E is the total kinetic energy liberated in the rain, and $I30$ is the maximum intensity produced in the storm in 30 minutes. In Spain it is very difficult to get the values of the factor R because there are few observatories and they normally only analyze the rain in 24 hours, without information on the temporal series. To bypass this problem, several expressions were evaluated to calculate R as a function of the more common weather variables. In particular, for the Mediterranean area the following expression was adopted (Antolín et al., |5|): $R = 2.375P_{24}^2 + 0.513P_m ex - 94.4 - 81Z_1 + Z_2 + 37Z_3 + 84Z_4$, where P_{24}^2 is the maximum daily rainfall in a period of two years, $P_m ex$ is the interannual mean of rainfall of the maximum rainfall month in the year, Z_1 is Zone 1 (the nearest zone to Grazalema), Z_2 stands for Zone 2 (south of Spain and the Segura river basin), Z_3 is Zone 3 (the rest of the Mediterranean area) and Z_4 is the Oriental Pirinean basin. This formula was applied using the weather values of 199 stations (observatories) and then, the maximum rainfall in 24 hours in a period of two years and the interannual mean rainfall in the most rainfall month were calculated in each station with the Gumbel method. In this paper we analyze 295 spatial locations in the Province of Castellon (Spain) where the R factor together with elevation and distance to the sea were recorded. The aim is to build a statistical model to predict the rainfall erosivity all over the study region.

5.2 Analysis of results and conclusions

Figure 1 shows the spatial locations where the data was recorded together with the locations of the 41 automatic stations used to analyze the goodness-of-fit of the results. The spherical variogram model was fitted to the data using maximum likelihood. Note that the fitted variogram lies within the simulated variogram envelopes. Traditional simple kriging provided similar results in terms of prediction and standard errors than ordinary kriging (see Figure 2). To perform uncertainty in the model, we evaluated the model-based approach to analyze the parameter β. At a first step, let us suppose that there is no trend to evaluate the factor R. Figure 3 shows the estimates given by traditional OK and model-based approach for the 41 automatic stations. Both procedures do not predict quite well the extreme cases, which could be due to the fact that a trend model should be considered. Then, in a further step, we evaluated the trend by means of linear and loess models that predict factor *R* from distance to the sea and elevation. Figure 4 shows the estimates given by both approaches for the automatic stations. There are slightly differences between both methods but in any case when using a trend model, the results seem to improve.

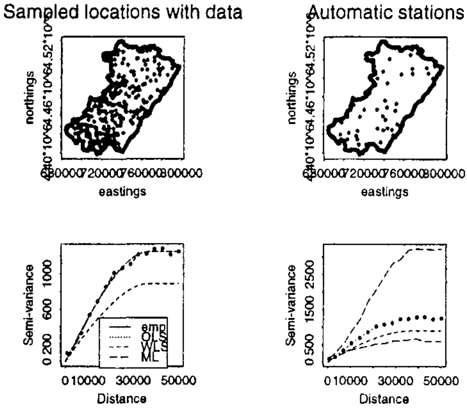

Figure 1: Sampled locations in the study region (left) and locations for 41 automatic stations (right). Estimated variograms and envelopes.

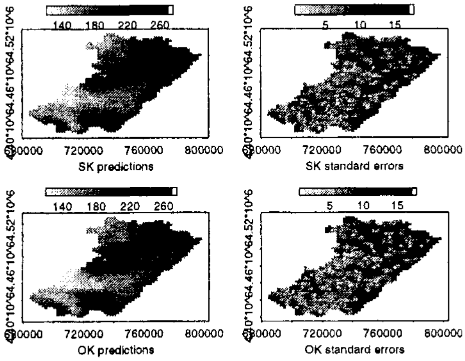

Figure 2: Kriged *R* factor using SK and OK methods and corresponding standard errors (s.e.).

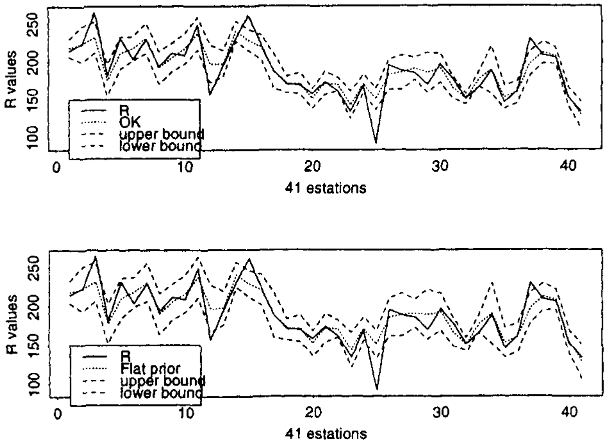

Figure 3: Prediction of *R* at the automatic stations using traditional OK and a Bayesian approach with a flat prior for parameter β.

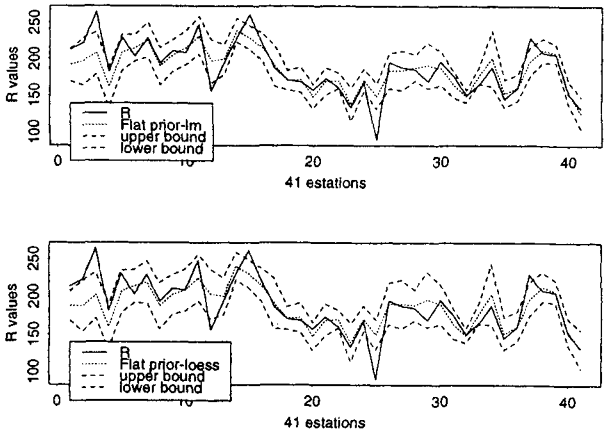

Figure 4: Prediction of *R* at the automatic stations using a Bayesian approach with a flat prior for parameter β with trend evaluated under linear and loess models.

References

[1] Journel, A. & Huijbregts, C., *Mining Geostatistics*, Academic Press, 1978.

[2] Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

[3] Ribeiro, P.J. & Diggle, P.J. *Bayesian inference in Gaussian model-based geostatistics.* Technical Report, Lancaster University, 2001.

[4] Diggle, P.J., Moyeed, R.A. & Tawn, J.A., Non-gaussian geostatistics. *Applied Statistics*, **47**, pp. 299-350, 1998.

[5] Antolín, C., Carbó, E. & Alvarez, D., Aplicación de la ecuación universal de pérdida de suelo en la Comunidad Valenciana. In: *El suelo como recurso natural en la Comunidad Valenciana. Territori 8.* Generalitat Valenciana, Valencia, pp. 136-165, 1998.