DIAGNOSTIC TECHNIQUES APPLIED IN GEOSTATISTICS FOR AGRICULTURAL DATA ANALYSIS[1]

Joelmir André Borssoi[2], Miguel Angel Uribe-Opazo[3] & Manuel Galea Rojas[4]

SUMMARY

The structural modeling of spatial dependence, using a geostatistical approach, is an indispensable tool to determine parameters that define this structure, applied on interpolation of values at unsampled points by kriging techniques. However, the estimation of parameters can be greatly affected by the presence of atypical observations in sampled data. The purpose of this study was to use diagnostic techniques in Gaussian spatial linear models in geostatistics to evaluate the sensitivity of maximum likelihood and restrict maximum likelihood estimators to small perturbations in these data. For this purpose, studies with simulated and experimental data were conducted. Results with simulated data showed that the diagnostic techniques were efficient to identify the perturbation in data. The results with real data indicated that atypical values among the sampled data may have a strong influence on thematic maps, thus changing the spatial dependence structure. The application of diagnostic techniques should be part of any geostatistical analysis, to ensure a better quality of the information from thematic maps.

Index terms: local influence, maximum likelihood, restricted maximum likelihood.

---

[1] Parte da Dissertação de Mestrado do primeiro autor. Recebido para publicação em fevereiro de 2008 e aprovado em agosto de 2009.

[2] Professor Assistente do Centro de Ciências Exatas e Tecnológicas – CCET. Universidade Estadual do Oeste do Paraná – UNIOESTE. Rua Universitária 2069, Sala 65, CEP 85819-110 Cascavel (PR). E-mail: jborssoi@yahoo.com.br

[5] Professor Doutor do Centro de Ciências Exatas e Tecnológicas – CCET/UNIOESTE. E-mail: mopazo@unioeste.br

[4] Professor Doutor do Departamento de Estadística, Universidad de Valparaíso, Valparaíso – Chile. E-mail: manuel.galea@uv.cl

RESUMO: *TÉCNICAS DE DIAGNÓSTICO UTILIZADAS EM GEOESTATÍSTICA PARA ANÁLISE DE DADOS AGRÍCOLAS*

A modelagem da estrutura de dependência espacial pela abordagem da geoestatística é de fundamental importância para a definição de parâmetros que definem essa estrutura e que são utilizados na interpolação de valores em locais não amostrados, pela técnica de krigagem. Entretanto, a estimação de parâmetros pode ser muito alterada pela presença de observações atípicas nos dados amostrados. O desenvolvimento deste trabalho teve por objetivo utilizar técnicas de diagnóstico em modelos espaciais lineares gaussianos, empregados em geoestatística, para avaliar a sensibilidade dos estimadores de máxima verossimilhança e máxima verossimilhança restrita a pequenas perturbações nos dados. Foram realizados estudos de dados simulados e experimentais. O estudo com dados simulados mostrou que as técnicas de diagnóstico foram eficientes na identificação da perturbação nos dados. Pelos resultados obtidos com o estudo de dados reais, concluiu-se que a presença de valores atípicos entre os dados amostrados pode exercer forte influência nos mapas temáticos, alterando, assim, a estrutura de dependência espacial. A aplicação de técnicas de diagnóstico deve fazer parte de toda análise geoestatística, a fim de garantir que as informações contidas nos mapas temáticos tenham maior qualidade.

Termos de indexação: influência local, máxima verossimilhança, máxima verossimilhança restrita.

INTRODUCTION

In the last few decades, concepts of monitoring and management of the process of agricultural production have been widely discussed and applied, generating great amounts of information on yield-related factors.

Some of these concepts take the spatial variability of the geo-referenced variable into consideration, mainly those related to the soil, such as soil physical and chemical properties. According to Cressie (1993), not taking the spatial variability into consideration can prevent the perception of real differences, which would make a differentiated treatment according to the local requirements impossible.

The geostatistics, based on the theory of regionalized variable, is a method that considers the spatial distribution of measures, to determine the ray of spatial autocorrelation between its elements and, accordingly, the maximum distance up to which the samples are considered spatially dependent.

For modeling data with spatial structure, according to Mardia & Marshall (1984), a Gaussian random process $\{Z(s_i),\ s_i \in S\}$ is considered, with $S \subset \Re^d$; where $\Re^d$ is the Euclidean space, d-dimensional ($d \geq 1$). It was assumed that the data set $Z(s_1), ..., Z(s_n)$ of this process is registered in known space locations, s_i ($i = 1,..., n$), and generated by the following model:

$$Z(s_i) = \mu(s_i) + \varepsilon(s_i) \qquad (1)$$

where the terms deterministics $\mu(s_i)$, and random $\varepsilon(s_i)$, can depend on the spatial location where $Z(s_i)$ was obtained.

It was assumed that the mean of the random error $\varepsilon(.)$ is zero, $E[\varepsilon(s_i)] = 0$, and that the variation between points in the space is determined by a covariance function $C(s_i, s_u) = Cov[\varepsilon(s_i), \varepsilon(s_u)]$ and that for some known functions of s, $x_1(s),..., x_p(s)$, the mean of the stochastic process is:

$$\mu(s) = \sum_{u=1}^{p} x_u(s)\beta_u \qquad (2)$$

where $\beta_1,..., \beta_p$ are unknown parameters and must be estimated.

Or equivalent, but expressed in matricial notation:

$$Z = X\beta + \varepsilon \qquad (3)$$

Then, $E(\varepsilon) = 0$ and the covariance matrix of ε is $\Sigma = [(\sigma_{iu})]$, where $\sigma_{iu} = C(s_i, s_u)$. It was assumed that Σ is non-singular, that X is of full column rank and that Z follows a multivariate normal distribution with mean $X\beta$ and covariance matrix Σ, that is, $Z \sim N_n(X\beta, \Sigma)$.

A particular parametric form was considered for the covariance matrix:

$$\Sigma = \varphi_1 I_n + \varphi_2 R \qquad (4)$$

where, φ_1: nugget effect, φ_2: sill value; R is a symmetric matrix that depends on φ_3, $R = R(\varphi_3) = [(r_{iu})]$, in the order n x n with its elements of the diagonal $r_{ii} = 1$, $i = 1, ..., n$, where φ_3 is function of the range (a) of the model and I_n is the identity matrix of the order n x n.

The parametric form of the covariance matrix, given in equation (4), applies for several isotropic processes, where the covariance $C(s_i, s_u)$ is defined as $C(h_{iu}) = \varphi_2 r_{iu}$, where $h_{iu} = ||s_i - s_u||$ is the Euclidean distance between the points s_i and s_u. The variance of the stochastic process Z is $C(0) = \varphi_1 + \varphi_2$, and the semivariance can be defined as $\gamma(h) = C(0) - C(h)$.

In many situations a data set with aberrant or discrepant observations can be considered influential, that is, they induce some type of decision in the construction of geostatistical models.

The local influence method proposed by Cook (1986) evaluates the simultaneous effect of observations on the maximum likelihood estimators without the need of its elimination from the data set. Christensen et al. (1993) studied diagnostic methods based on the elimination of cases in linear spatial models. The local influence technique has become known as a procedure to carry out sensitivity analyses of statistical models and has been widely used in linear models and nonlinear regression.

For an observed data set, let $l(\theta)$ be the log-likelihood function of the proposed model, given in equation (3), where $\theta = (\beta^T, \varphi^T)^T$, $\beta = (\beta_1,..., \beta_p)^T$ and $\varphi = (\varphi_1,..., \varphi_\theta)^T$, and let ω be a perturbation vector belonging to a space of perturbations Ω. Let $l(\theta/\omega)$ be the log-likelihood function corresponding to the perturbated model for $\omega \in \Omega$. It was assumed that there is $\omega_0 \in \Omega$ such that $l(\theta) = l(\theta/\omega_0)$, for all θ and that $l(\theta/\omega)$ is twice differentiable in $(\theta^T, \omega^T)^T$.

This study is justified by the importance of the modeling of spatial variability, since this process supplies parameters of spatial dependence structure that are used for the spatial interpolation of kriging.

Based on the interpolation of kriging, thematic maps are generated that could be used for a site-specific application of special inputs or site-specific soil management. The map quality depends on the quality of the inferences of the adjusted models. Therefore, to obtain trustworthy predictions that represent the real local variability by the interpolation produces, the modeling process must be very carefully carried out, mainly in the case of discrepant or influential observations.

The objective of this study was to use diagnostic techniques in Gaussian linear spatial models to evaluate the potential influence of atypical data on the parameter estimates that define the spatial dependence and to indicate the most robust models. For this purpose studies on local influence were conducted using the methods maximum likelihood and restricted maximum likelihood, to study the sensitivity of the models in the presence of influential observations.

MATERIAL AND METHODS

Influence of location

The following perturbation scheme was considered: $Z_\omega = Z + \omega$, with $\omega = (\omega_1, ..., \omega_n)^T$ vector of perturbation of the response and $\omega_0 = (0,...,0)^T$, the point of no perturbation. The objective of this scheme of perturbation was to detect outliers in the data that affect the maximum likelihood estimator θ. Then, the perturbed log-likelihood function $l(\theta/\omega)$, for the normal model is given by

$$l(\theta/\omega) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Z_\omega - X\beta)^T \Sigma^{-1}(Z_\omega - X\beta) \quad (5)$$

The influence of the perturbation ω on the maximum likelihood estimator of θ can be evaluated by the likelihood displacement, defined by:

$$LD(\omega) = 2(l(\hat{\theta}) - l(\hat{\theta}_\omega)) \quad (6)$$

where, $\hat{\theta}$ is the maximum likelihood estimator of θ in the postulated model; and $\hat{\theta}_\omega$ is the maximum likelihood estimator of θ in the perturbed model.

Cook (1986) proposed to study the local behavior of $LD(\omega)$ around ω_0, using the normal curvature C_l of $LD(\omega)$ in ω_0 in the direction of a unit vector l and showed that

$$C_l = 2|l^T \Delta^T L^{-1} \Delta l| \quad (7)$$

with $||l|| = 1$, where, L is the observed information matrix, evaluated in $\theta = \hat{\theta}$; Δ is a $(p + 3)x\ n$ matrix given by $\Delta = (\Delta_\beta^T, \Delta_\varphi^T)^T$, evaluated in $\theta = \hat{\theta}$ and in $\omega = \omega_0$, where, in this case:

$$\Delta_\beta = X^T \Sigma^{-1} \text{ and } \Delta_\omega = \frac{\partial^2 l(\theta/\omega)}{\partial\varphi \partial\omega^T}, \text{ with } \frac{\partial^2 l(\theta/\omega)}{\partial\varphi \partial\omega^T} = (Z_\omega - X\beta)^T \Sigma^{-1} \frac{\partial\Sigma}{\partial\varphi_i} \Sigma^{-1}, \ j = 1, 2, 3.$$

The information matrix L is defined as $L = \begin{pmatrix} L_{\beta\beta} & L_{\beta\varphi} \\ L_{\varphi\beta} & L_{\varphi\varphi} \end{pmatrix}$, where, $L_{\beta\beta} = -(X^T \Sigma^{-1} X)$; $L_{\beta\varphi} = \frac{\partial^2 l(\theta)}{\partial\beta \partial\omega^T}$, with $\frac{\partial^2 l(\theta)}{\partial\beta \partial\varphi_j} = -X^T \Sigma^{-1} \frac{\partial\Sigma}{\partial\varphi_j} \Sigma^{-1}\varepsilon, \ j = 1, 2, 3$; $L_{\varphi\beta} = L^T_{\varphi\beta}$ and $L_{\varphi\varphi} = \frac{\partial^2 l(\theta)}{\partial\varphi \partial\varphi^T}$, with elements; $\frac{\partial^2 l(\theta)}{\partial\varphi_i \partial\varphi_j} = \frac{1}{2}\text{tr}\left\{\Sigma^{-1}\left(\frac{\partial\Sigma}{\partial\varphi_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j} - \frac{\partial^2\Sigma}{\partial\varphi_i \partial\varphi_j}\right)\right\} + \frac{1}{2}\varepsilon^T \Sigma^{-1}\left\{\frac{\partial^2\Sigma}{\partial\varphi_i \partial\varphi_j} - \frac{\partial\Sigma}{\partial\varphi_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j} - \frac{\partial\Sigma}{\partial\varphi_j}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_i}\right\}\Sigma^{-1}\varepsilon$.

Let L_{max} be the eigenvector corresponding to the largest eigenvalue of $B = \Delta^T L^{-1} \Delta$. The graph of the elements of $|L_{max}|$ *versus* i (data order) can reveal which type of perturbation has the greatest influence on $LD(\omega)$, in the neighborhood of ω_0 (Cook, 1986).

Simulation study

Simulation was carried out by Monte Carlo experiments, where simulated data sets were arranged in a regular grid (10 m distance between points),

totalizing 100 points, with known structures of spatial dependence, by means of Gaussian stochastic processes. The simulation of stationary spatial processes of second order was carried out by the Cholesky decomposition method, a matrix operation which, applied to the vector of random numbers generated produces another vector with random numbers with the characteristic of a given correlation matrix between them (Johnson & Wichern, 1982).

The covariance structures of the models used in the simulations were exponential, Gaussian and Matérn, with $\kappa = 0.7$ and 3.0. In all cases, a constant mean of $\mu = \beta = 9.45$ was considered. Four parameter vectors were used for each model $\varphi = (\varphi_1, \varphi_2, \varphi_3)^T$: 1st case: $\varphi = (0, 10, 10)^T$; 2nd case: $\varphi = (0, 10, 15)^T$; 3rd case: $\varphi = (0, 10, 20)^T$; 4th case: $\varphi = (0, 10, 60)^T$.

The perturbation scheme used had been proposed by Ortega et al. (2002), as presented in the equation (8) below

$$z_{max}^* = z_{max} + \sqrt{\|z\|} \quad (8)$$

where, z_{max}^* is the new value maximum of the vector and z_{max} is the maximum value of vector z; $||z|| = \sqrt{z_1^2 + ... + z_n^2}$. Consequently, the perturbation vector ω can be expressed as : $\omega = (0, ..., \sqrt{\|z\|}, ..., 0)^T$.

The structure of spatial dependence was modeled for the simulated data sets, but adding the perturbation vector ω. Diagnostic techniques were applied to evaluate the sensitivity of the models to the perturbation scheme.

The methods of maximum likelihood and restricted maximum likelihood were used for the parameter estimation and in the diagnostic analyses (Mardia & Marshall, 1984 and Christensen et al. 1993). Because the results were very similar, only the graphs based on maximum likelihood are presented.

Experimental study

The data of the soil chemical properties were collected in an area of commercial grain production of 71 ha, in the 2006/2007 growing season, in the county of Cascavel, Western region of Paraná, (lat 24.95 ° S, long 53.57 ° W, average altitude 650 m asl. The soil was classified as dystroferric Red Latosol and the climate of the region is Temperate mesothermal and superhumid, climate type Cfa (Köppen), with an average annual temperature of 21 °C.

A centered systematic sampling with pairs of adjacent points was carried out (*lattice plus close pairs*), with a maximum distance of 141 m between points. At some places the sampling was carried out at distances of 75 and 50 m between points. All samples were geo-referenced with a GPS (*Global Positioning System*) signal receiver.

In the vicinity of a few points four soil sub-samples were randomly collected, from the layer 0.0–0.2 m. The sub-samples of approximately 500 g were mixed and stored in plastic bags, to compose a representative sample of the plot.

Initially, an exploratory statistical analysis of the data was carried out to evaluate the general behavior and to identify to the presence of discrepant points and their possible causes. Among the analyzed chemical properties, discrepant points were observed for soil P. This trait was analyzed in this study.

Thereafter, a spatial data analysis was performed using geostatistical techniques, identifying the structure of spatial dependence by means of the adjustment of some theoretical models, with parameters estimated by maximum likelihood (ML) and restricted maximum likelihood (RML). In this stage, the criteria of model validation and diagnostic techniques were applied for the posterior construction of maps of the study variable. All analyses were carried out using software R (Ihaka & Gentleman, 1996) and the modules: geoR (Ribeiro Júnior & Diggle, 2001) and Splancs (Rowlingson & Diggle, 1993).

RESULTS AND DISCUSSION

Local influence on simulated data

Estimates of maximum likelihood (ML) and restricted maximum likelihood (RML) for $_1$, j_2 and j_3, using the exponential, Gaussian and Matérn covariance functions, are presented in table 1.

The results of the exponential model 0-10-10 (Table 1) show that the methods ML and RML overestimated the parameters φ_1 and φ_3 and underestimated parameter φ_2. A similar behavior was identified in the parameter estimates of the exponential model 0-10-60, except for parameter φ_3 which had been underestimated when it was estimated by MV. The variables simulated by the Gaussian 0-10-10 and Gaussian 0-10-15 model overestimated the parameters φ_1 and φ_3 and underestimated parameter φ_2 by the methods of ML and RML. By the Matérn model 0-10-15 $\hat{e} = 0.7$, the estimated φ_1 values were equal to those supplied in the simulation and the φ_3 values were close to the simulated. However, the values of parameter φ_2 were overestimated. By the Matérn model 0-10-10, k = 3.0, it was observed that φ_2 was clearly overestimated. Hence, the perturbed values in the simulated variables with exponential, Gaussian and Matérn (k = 0.7 and 3.0) covariance functions, had a rather strong influence on parameter estimation, by ML as well as by RML.

Table 1. Estimates of maximum likelihood (ML) and restricted maximum likelihood (RML) for φ_1, φ_2 and φ_3 using the exponential, Gaussian and Matérn covariance functions

Simulated variable	Perturbed observation	φ_1		φ_2		φ_3	
		ML	RML	ML	RML	ML	RML
Exponential 0 -10-10	66	8.111	8.376	3.960	5.759	27.085	54.369
Exponential 0 -10-15	19	0.000	0.000	13.150	14.233	15.386	17.098
Exponential 0 -10-20	21	0.000	0.000	14.010	15.640	17.73 0	20.400
Exponential 0 -10-60	80	1.002	1.091	6.772	9.460	39.163	60.000
Gaussian 0 -10-10	35	4.166	4.229	8.607	8.905	13.949	14.269
Gaussian 0 -10-15	44	2.487	2.564	9.014	9.426	16.920	17.431
Gaussian 0 -10-20	32	0.894	0.906	9.446	9.853	18.009	18.243
Gaussia n 0-10-60	84	1.027	1.030	3.462	4.258	45.314	48.087
Matérn 0 -10-10 $\kappa = 0.7$	42	1.603	2.140	8.332	8.457	9.828	11.669
Matérn 0 -10-15 $\kappa = 0.7$	91	0.000	0.000	14.409	16.170	13.732	15.409
Matérn 0 -10-20 $\kappa = 0.7$	06	0.223	0.552	9.225	9.912	13.473	15.556
Matérn 0 -10-60 $\kappa = 0.7$	31	0.009	0.093	6.825	8.374	23.241	28.945
Matérn 0 -10-10 $\kappa = 3.0$	11	0.232	0.218	93.082	75.689	16.084	15.000
Matérn 0 -10-15 $\kappa = 3.0$	03	0.766	0.781	14.791	21.770	20.278	23.238
Matérn 0 -10-20 $\kappa = 3.0$	99	0.698	0.708	15.346	21.845	21.354	21.020
Matérn 0 -10-60 $\kappa = 3.0$	50	0.814	0.822	3.098	5.410	28.524	35.666

Figures 1 to 8 present the graphs of diagnostic techniques to identify perturbed observations. The results of the local influence analysis showed that the index plots of $|L_{max}|$ highlighted the perturbed observations for all variables, in the parameter estimation by ML as much as RML (not shown here).

Hence, in this simulation study diagnostic measures were effective to detect all perturbed observations.

Spatial analysis of influence of discrepant points on phosphorus content

After sampling, soil samples were sent to the laboratory for chemical analyses. In the first analysis, the P content presented three discrepant values (60.0; 38.6 and 60.0 mg dm^{-3}) in the plots 1, 26 and 45, respectively (Figure 10a). The soil chemical analysis was repeated in these plots and no changes in the values were observed. The data of 1, 26 and 45 are therefore not measurement or laboratory errors; they represent real values of the local soil conditions.

Graphical diagnostic techniques were applied to evaluate whether the discrepant points 1, 26 and 45 or others exert some type of influence on likelihood displacement, on the covariance function and the linear predictor. The index plots of $|L_{max}|$ were used to evaluate the influence.

Diagnostic graphs are presented in figure 9 for the exponential, Gaussian and Matérn ($\kappa = 0.7$ and 3.0) covariance functions, by ML. The index plots of $|L_{max}|$ indicated the observations 1, 26 and 45 as potentially influential.

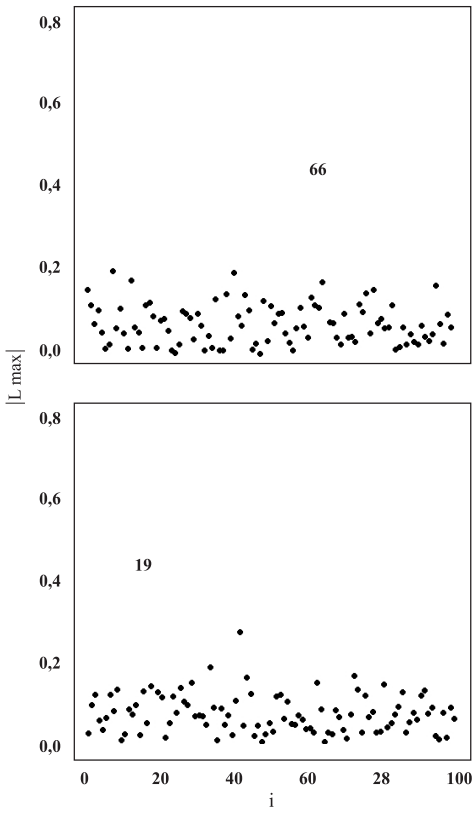

Figure 1. Index plot of $|L_{max}|$ for simulated data using the exponential 0-10-10 (left side) and 0-10-15 (right side) covariance function.

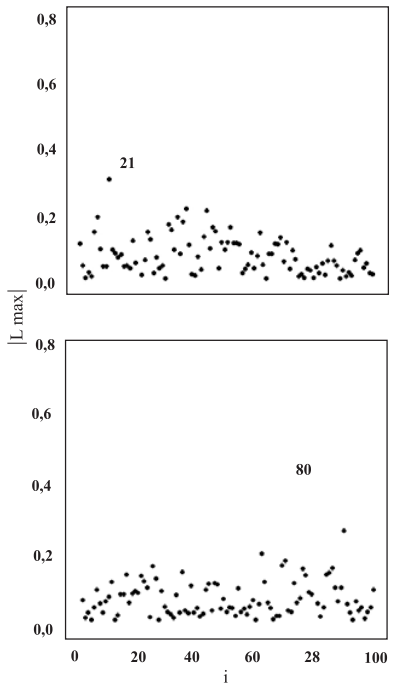

Figure 2. Index plot of $|L_{max}|$ for simulated data using the exponential 0-10-20 (left side) and 0-10-60 (right side) covariance function.

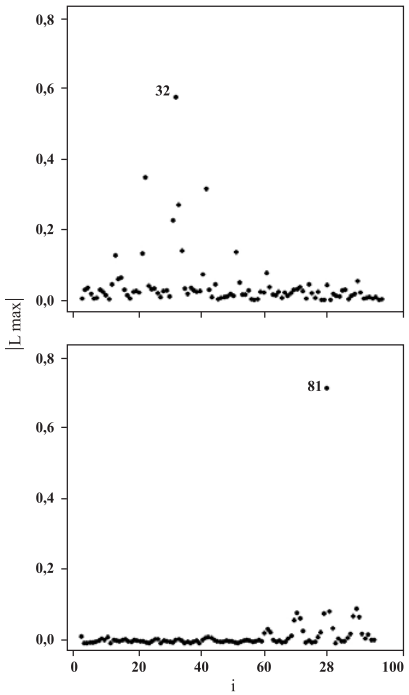

Figure 4. Index plot of $|L_{max}|$ for simulated data using the Gaussian 0-10- 20 (left side) and 0-10-60 (right side) covariance function.

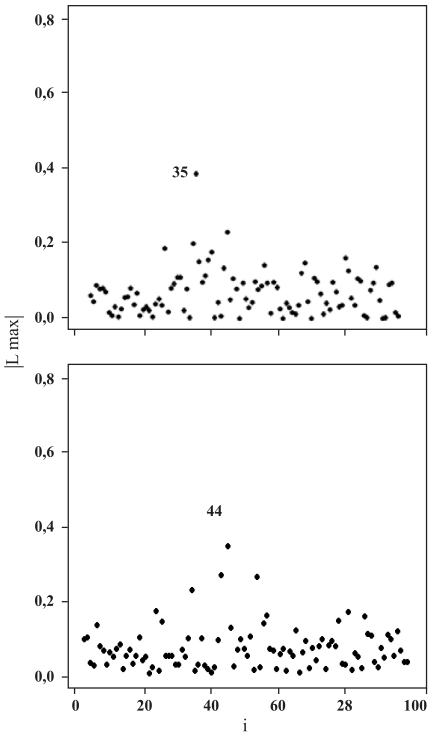

Figure 3. Index plot of $|L_{max}|$ for simulated data using the Gaussian 0-10-10 (left side) and 0-10-15 (right side) covariance function.

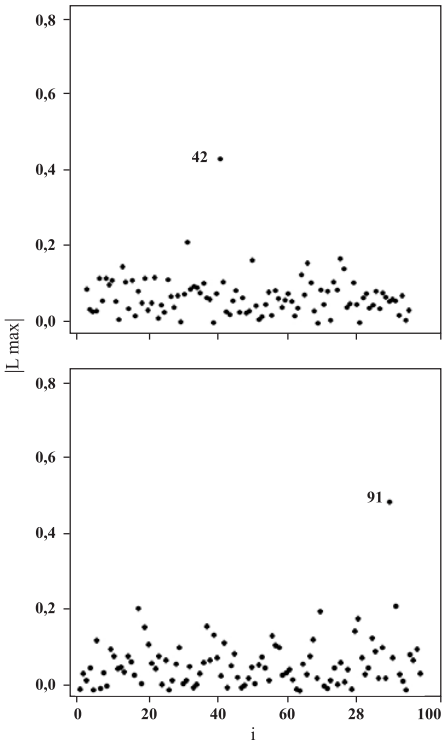

Figure 5. Index plot of $|L_{max}|$ for simulated data using the Matérn 0-10-10 (left side) and 0-10-15 (right side) covariance function, with k =0.7.

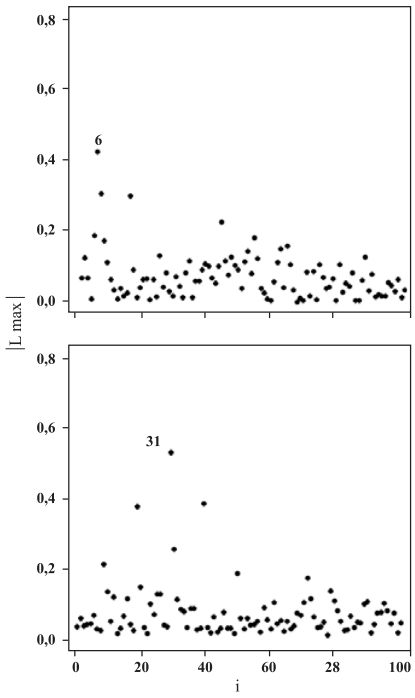

Figure 6. Index plot of $|L_{max}|$ for simulated data using the Matérn 0-10-20 (left side) and 0-10- 60 (right side) covariance function, with = 0.7.

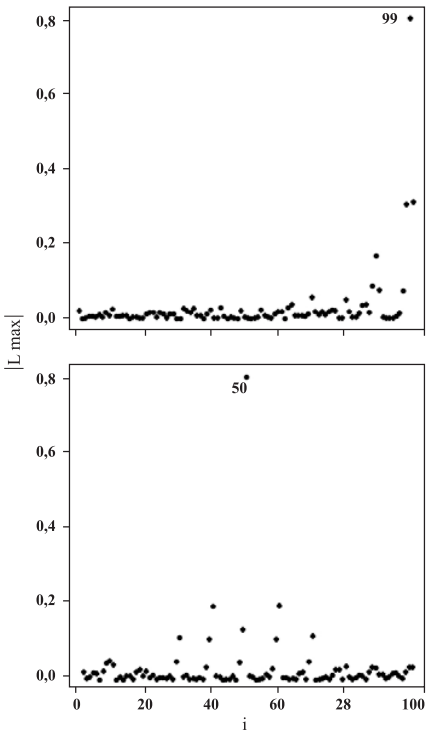

Figure 8. Index plot of $|L_{max}|$ for simulated data using the Matérn 0-10-20 (left side) and 0-10-60 (right side) covariance function, with k = 3.0.

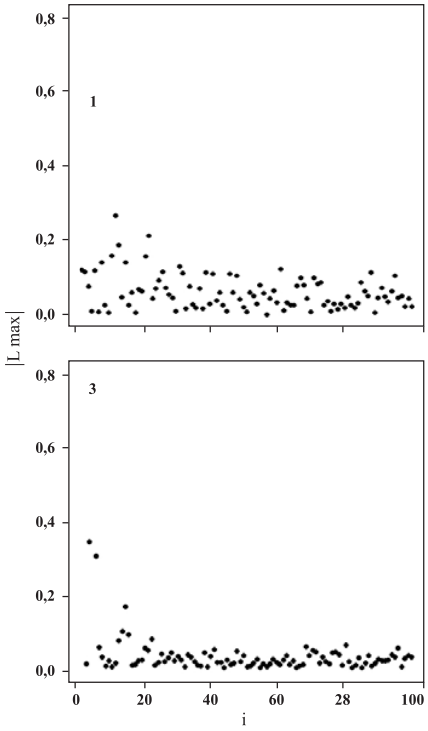

Figure 7. Index plot of $|L_{max}|$ for simulated data using the Matérn 0-10-10 (left side) and 0-10-15 (right side) covariance function, with k = 3.0.

Influence on descriptive analyses

A descriptive summary of the phosphate content was presented (Table 2) with all collected data and without the discrepant observations. It was verified that the mean P content (15.80 mg dm^{-3}) is well above the recommended upper limit (EMBRAPA, 1979). Furthermore, the data were highly heterogeneous, since the variation coefficient was very high (CV = 72.76 %), due to the discrepant values of the observations 1, 26 and 45, identified for the *Box-plot* (Figure 10b).

Since the statistical techniques applied in this paper assume that the data have normal distribution, the Box-Cox was transformed with *lambda* = -0.7. Also, the descriptive analyses are presented for the data set without the observations 1, 26 and 45 (P-1-26-45), verifying the influence in the descriptive analysis. Differences were observed between the means of the variable with all data or when removing influential observations. The same behavior was observed in the standard deviations. The analysis of the coefficient of variation (CV) showed that without the observations 1, 26 and 45 the CV value was much lower than with all data. However, as the CV is > 30 %, the data were still considered heterogeneous (Gomes, 2000).

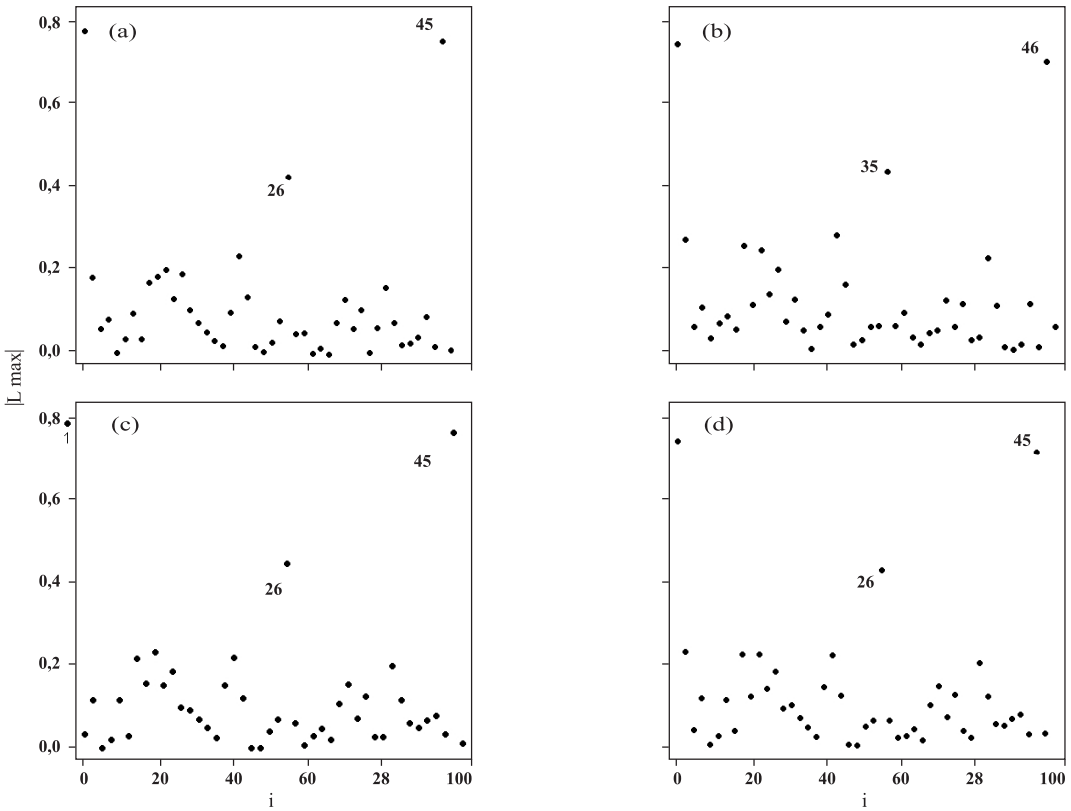

Figure 9. Index plot of $|L_{max}|$ for real data using the exponential (a), Gaussian (b), Matérn- k = 0.7 (c) and Matérn - k = 3.0 (d) covariance functions.

Table 2. Descriptive statistics for phosphorus content [mg dm^{-3}] with all observations (P) and without the observations 1, 26 and 45 (P-1-26-45)

Variable	N	Mean	Min.	Max.	Q1	Med	Q3	SD	CV (%)	*p*-value
P	46	15.80	5.70	60.00	9.70	11.70	18.60	11.50	72.76	< 0.05*
P-1-26-45	43	13.21	5.70	26.90	9.70	11.70	16.55	5.42	40.99	< 0.05*

N: number of observations; Min: minimum value; Max: maximum value; Q1: first quartile; Med: median; Q3: third quartile; SD: standard deviation; CV: coefficient of variation; p-value: For the test of normality of Shapiro & Wilks, at 5 %.

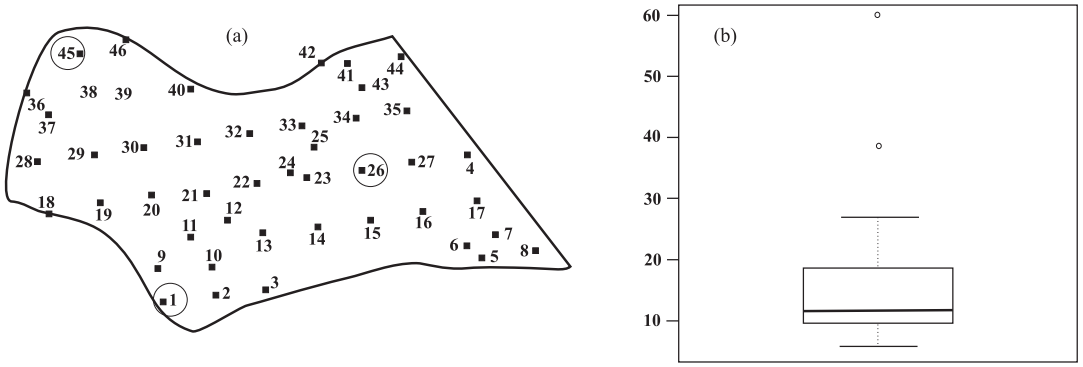

Figure 10. (a) Arrangement of sampled points in the study area of 71 ha. (b) *Box-plot* for P content in full sample

Influence on the parameter estimates

The results of the analyses of spatial variability are presented for the original data and for the data set without the observations 1, 26 and 45 (Table 3). The results of the estimation of the parameters β, φ_1, φ_2 and φ_3, were presented for the exponential, Gaussian and Matérn covariance functions, using ML and RML. The values in brackets stand for the standard deviations of the estimated parameters. Overall, the values of the estimators of β, φ_1, φ_2 without the influential observations were lower than when estimated with the original data. The estimates ML of φ_3, without the influential observations were greater than those obtained with the original data. This was not the case for parameter φ_3, when estimated by RML.

The cross-validation criteria (Faraco et al., 2008) applied to the models in the study, both with all observations and without the observations 1, 26 and 45 indicated the Gaussian model using RML as best fitting.

Influence on the construction of thematic maps

Figure 11 presents the thematic maps for P content with original data and without the observations 1, 26 and 45 based on the interpolation by ordinary kriging. The maps had been constructed using the models indicated for the cross-validation criteria. The variation in the color scale between the maps was considerable. The map for the original data set (Figure 11a) shows that the area comprises regions with a P content of > 18 mg dm^{-3} (Embrapa, 1997). In the map, constructed without the influential observations considered (Figure 11b) no region has values > 18 mg dm^{-3}. This indicates that observations 1, 26 and 45 also exert a strong influence on the construction of the thematic maps.

Thus, if the construction of the thematic maps does not take the diagnostic analyses into consideration, which detects outliers, the distribution map of the P content for producers would overestimate P concentrations in the study area. Consequently, the

Table 3. Estimates of maximum likelihood (ML) and restricted maximum likelihood (RML) for β, φ_1, φ_2 and φ_3 using the exponential, Gaussian and Matérn covariance functions, for phosphorus content [mg dm^{-3}] with all observations (P) and without the observations 1, 26 and 45 (P-1-26-45)

Model	Method	Variable	β	φ_1	φ_2	φ_3
Exp	ML	P	1.1822 (0.01551)	0.0043 (0.00227)	0.0020 (0.00243)	108.6275 (1.08326)
		P-1-26-45	1.1698 (0.01371)	0.0038 (0.00166)	0.0011 (0.00167)	123.4113 (1.77713)
	RML	P	1.1824 (0.02344)	0.0046 (0.00162)	0.0024 (0.00214)	235.0387 (3.86616)
		P-1-26-45	1.1686 (0.16568)	0.0038 (0.30859)	0.0014 (0.19047)	221.8948 (398.20349)
Gaus	ML	P	1.1817 (0.016 49)	0.0052 (0.00137)	0.0012 (0.00126)	214.8181 (0.61264)
		P-1-26-45	1.1690 (0.01512)	0.0042 (0.00109)	0.0008 (0.00094)	243.1303 (0.99819)
	RML	P	1.1817 (0.02169)	0.0052 (0.00128)	0.0018 (0.00150)	292.3174 (1.18161)
		P-1-26-45	1.1681 (0.01851)	0.0041 (0.00104)	0.0012 (0.00111)	292.5954 (1.36237)
Matérn κ =0.7	ML	P	1.1822 (0.01564)	0.0046 (0.00519)	0.0017 (0.00340)	100.9344 (172.08500)
		P-1-26-45	1.1700 (0.01407)	0.0040 (0.00372)	0.0010 (0.00237)	114.3420 (243.15720)
	RML	P	1.1823 (0.02296)	0.0048 (0.00357)	0.0022 (0.00217)	193.7198 (239.18450)
		P-1-26-45	1.1685 (0.01861)	0.0040 (0.00275)	0.0013 (0.00164)	189.6063 (287.11360)
Matérn κ =3.0	ML	P	1.1819 (0.01642)	0.0051 (0.00615)	0.0013 (0.00186)	60.0610 (72.18578)
		P-1-26-45	1.1693 (0.01475)	0.0042 (0.00447)	0.0008 (0.00130)	67.5349 (97.07221)
	RML	P	1.1820 (0.02160)	0.0051 (0.00554)	0.0018 (0.00161)	86.5556 (74.29928)
		P-1-26-45	1.1683 (0.01866)	0.0041 (0.00428)	0.0012 (0.00122)	88.5350 (87.45752)

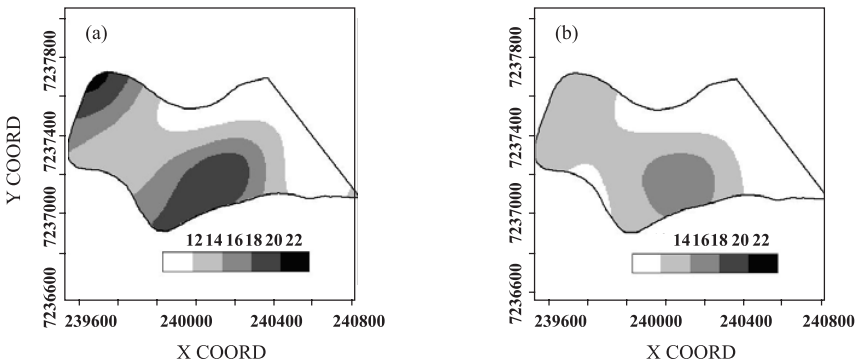

Figure 11. Thematic maps: (a) With original data (P); (b) Elimination of the observations 1, 26 and 45 (P-1-26-45).

principles of the precision agriculture would be disregarded, since the soil correction would not be locally adjusted.

The results showed that the removal of the influential data led to an increase of 10.14 % of the area of the first class; an increase of 17.16 % of the second and a decrease of 7.03 % of the area of the third class in the map (Table 4). However, the fourth and fifth class were not identified in the map when the influential data had been removed.

Table 4. Percentage of the area that each class represents in the maps of phosphorus content (mg dm^{-3})

Phosphorus content	Classe				
	1ª	2ª	3ª	4ª	5ª
	%				
Original data	27.69	31.84	20.20	19.41	0.86
No influential data	37.83	49.00	13.17	-	-
Difference	10.14	17.16	-7.03	-	-

CONCLUSIONS

The study with simulated data showed that the proposed diagnostic techniques were able to identify the perturbed data. The restricted maximum likelihood estimator produced unbiased estimates of the parameters of spatial dependence. For the results obtained with real data, the study concluded that the presence of atypical cases in the data had a strong influence on the thematic maps, due to change in the structure of the spatial dependence. The use of diagnostic techniques should be part of all geostatistical analyses, to ensure the high quality of the information contained in the thematic maps. The elimination of atypical cases can produce maps that are inappropriate.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided of the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), National Council of Scientific and Technological Development (CNPq), National Supply Company (CONAB) and Fundação Araucária, Brazil and Project Dipuv 11/2006, Universidad de Valparaíso, Chile.

LITERATURE CITED

CHRISTENSEN, R.; JOHNSON, W. & PEARSON, L. Covariance function diagnostics for spatial linear models. Inter. Assoc. Mathem. Geol., 25:145-160, 1993.

COOK, R.D. Assessment of local influence (with discussion). J. Royal Statis. Soc., Series B, 48:133-169, 1986.

CRESSIE, N.A.C. Statistics for spatial data. New York, John Wiley & Sons, 1993. 900p.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. Serviço Nacional de Levantamento e Conservação de Solos. Manual de métodos de análise de solo. Rio de Janeiro, Ministério da Agricultura, 1979. 247p.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. Serviço Nacional de Levantamento e Conservação de Solos. Manual de métodos de análise do solo. 2.ed. Rio de Janeiro, Centro Nacional de Pesquisas de Solos, 1997. 212p.

FARACO M.A.; URIBE-OPAZO, M.A.; SILVA, E.A.; JOHANN J.A. & BORSSOI, J.A. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. R. Bras. Ci. Solo, 32:463-476, 2008.

GOMES, P. Curso de estatística experimental. 14.ed. Piracicaba, Degaspari, 2000. 477p.

IHAKA, R. & GENTLEMAN, R. A language for data analysis and graphics. J. Comput. Graphical Statistics, 5:229-314, 1996.

JOHNSON, R.A. & WICHERN, D.W. Applied multivariate statistical analysis. Madison, Prentice Hall International, 1982. 607p.

MARDIA, K. & MARSHALL, R. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71:135-146, 1984.

ORTEGA, E.; BOLFARINE, H. & PAULA, G. Influence diagnostics in generalized log-gamma regression models. Comput. Statistics Data Anal. J., 42:165-186, 2002.

RIBEIRO JUNIOR, P.J. & DIGGLE, P.J. geoR: A package from geoestatistical analysis. R. News, 1:15-18, 2001.

ROWLINGSON, B. & DIGGLE, P.J. Splancs: Spatial point pattern analysis code in S-Plus. Comp. Geosci., 19:627-655, 1993.