Revista Brasileira de Ciência do Solo
ISSN: 0100-0683
revista@sbcs.org.br
Sociedade Brasileira de Ciência do Solo
Brasil

Botinha Assumpção, Rosangela Aparecida; Uribe Opazo, Miguel Angel; Galea, Manuel
LOCAL INFLUENCE FOR SPATIAL ANALYSIS OF SOIL PHYSICAL PROPERTIES AND SOYBEAN YIELD USING STUDENT'S t-DISTRIBUTION
Revista Brasileira de Ciência do Solo, vol. 35, núm. 6, 2011, pp. 1917-1926
Sociedade Brasileira de Ciência do Solo
Viçosa, Brasil

Available in: http://www.redalyc.org/articulo.oa?id=180221446008

LOCAL INFLUENCE FOR SPATIAL ANALYSIS OF SOIL PHYSICAL PROPERTIES AND SOYBEAN YIELD USING STUDENT'S *t*-DISTRIBUTION[1]

Rosangela Aparecida Botinha Assumpção[2], Miguel Angel Uribe Opazo[3] & Manuel Galea[4]

SUMMARY

The modeling and estimation of the parameters that define the spatial dependence structure of a regionalized variable by geostatistical methods are fundamental, since these parameters, underlying the kriging of unsampled points, allow the construction of thematic maps. One or more atypical observations in the sample data can affect the estimation of these parameters. Thus, the assessment of the combined influence of these observations by the analysis of Local Influence is essential. The purpose of this paper was to propose local influence analysis methods for the regionalized variable, given that it has n-variate Student's t-distribution, and compare it with the analysis of local influence when the same regionalized variable has *n*-variate normal distribution. These local influence analysis methods were applied to soil physical properties and soybean yield data of an experiment carried out in a 56.68 ha commercial field in western Paraná, Brazil. Results showed that influential values are efficiently determined with *n*-variate Student's *t*-distribution.

Index terms: geostatistics, EM algorithm, spatial variability.

---

[1] Received for publication in September 9, 2010 and approved August 31, 2011.

[2] D. Eng. in Agricultural Engineering at the State University of West Paraná, Universidade Estadual do Oeste do Paraná – UNIOESTE, Paraná, Brazil. E-mail: rosangela_botinha@hotmail.com

[3] Associate Professor of the Graduate Program in Agricultural Engineering at the State University of West Paraná, UNIOESTE. E-mail: miguel.opazo@pq.cnpq.br

[4] D.S Professor at the Dept. of Mathematics, Catholic University (Universidad Católica) Santiago - Chile. E-mail: mgalea@mat.puc.cl

RESUMO: *INFLUÊNCIA LOCAL PARA ANÁLISE ESPACIAL DOS ATRIBUTOS FÍSICOS DO SOLO E DA PRODUTIVIDADE DA SOJA UTILIZANDO A DISTRIBUIÇÃO* t-*STUDENT*

A modelagem e estimação dos parâmetros que definem a estrutura de dependência espacial de uma variável regionalizada, utilizando métodos geoestatísticos, é de fundamental importância, pois a partir desses parâmetros é realizada a krigagem dos pontos não amostrados para a construção dos mapas temáticos. A presença de uma ou mais observações atípicas nos dados amostrais podem influenciar a estimação desses parâmetros. Assim, torna-se importante a avaliação da influência conjunta destas observações pela análise de Influência Local. Este trabalho tem por objetivo apresentar métodos de análise de influência local para variável regionalizada considerando que ela tenha distribuição t-Student n-*variada e comparar com a análise de influência local considerando que esta mesma variável regionalizada tenha distribuição normal* n-*variada. Esses métodos de análise de influência local foram aplicados a atributos físicos do solo e produtividade da soja obtidos a partir de um experimento realizado em uma área comercial de 56,68 ha da região Oeste do Paraná, Brasil. O estudo mostrou que os valores influentes são identificados eficientemente com a distribuição t-Student* n-*variada.*

Termos de indexação: geoestatística, algoritmo EM, variabilidade espacial.

INTRODUCTION

Geostatistics is an important tool to analyze spatial variability of soil properties and crop yield. Thematic maps based on geostatistics are excellent resources for the analysis of agricultural performance, and are considered the most complete option for the zoning of crop spatial variability (Molin, 2002). However, when the data set contains influential values, the maps diverge from reality and can therefore cause misinterpretation.

The identification of influential observations in the data set is known as diagnostic analysis (Paula, 2004) and was presented by Cook (1986) as a new method called "Local Influence" for diagnostic analysis of the Gaussian process. This analysis was based on the assumption that the model is reliable, and analyzed the power of conclusions for data or model disturbances.

Several studies of spatial analysis have been developed using diagnostic techniques to find observations that influence model sets. Christensen et al. (1992, 1993) worked on studies that analyzed the prediction of parameters in linear spatial models by applying diagnostic techniques to find observations that can influence the estimation of the covariance matrix, used in universal kriging. Militino et al. (2006) tried to identify outliers in multivariate linear spatial models. Borssoi et al. (2009, 2011) conducted diagnostic studies of linear Gaussian spatial models, and developed several measures of local influence.

The purpose of this paper was to propose a method that uses n-variate Student's t-distribution with a fixed degree of freedom and also uses the maximum likelihood function as method of estimating spatial variability parameters of soybean yield. From the estimation of these parameters, a study of local influence is developed, which analyzes the existence of influential observations in the spatial dependence structure (graph C_i) and in the linear predictor (graph l_p), considering the yield itself as explicative variable and soil penetration resistance (*SPR*) and soil density (*Des*) as covariates. Finally, these results were compared with those obtained by using the n-variate normal distribution.

THEORY

Student's t linear spatial models

Let $\{Y(s), s \in S\}$ be a stationary stochastic process, where $S \subset \Re^d$ and $\Re^d$ is a d-dimensional ($d \geq 1$) Eucledian space. The $Y = (Y(s_1),\ldots, Y(s_n))^T$ process has n-variate t-distribution with ν degrees of freedom ($\upsilon \geq 1$ fixed), where μ is the vector of location parameters and Σ the scaling matrix, ie $Y \sim t_n(\mu,\Sigma,\upsilon)$. Consider that $Y = (Y(s_1),\ldots, Y(s_n))^T$ are registered at known locations in space known in s_i and s_j for $i \neq j = 1,\ldots,n$. Each $Y(s_i)$ can be written as:

$$Y(s_i) = \mu(s_i) + \varepsilon(s_i),\ i = 1,\ldots,\ n \qquad (1)$$

where $\mu(s_i)$ is a deterministic and $\varepsilon(s_i)$ a stochastic term, both depending on the parameter space where $Y(s_i)$ operates. It is assumed that the stochastic error ε has zero mean, ie, $(s_i) = E[\varepsilon(s_i)] = 0$, $i = 1,\ldots,\ n$ and that the variation of the points in space is determined by a covariance function $C(.)$ of s_i and s_j. This covariance function is also specified by a parameter vector $\Phi = (\varphi_1,\varphi_2,\varphi_3)^T$, which are the parameters that determine the structure of spatial dependence.

For some known functions of S, such as $X_1(s),\ldots,X_p(s)$, we consider the linear spatial model in the mean of the process $\mu(.)$, as: $\mu(s_i) = \sum_{j=1}^{p}\beta_j X_j(s_i)$, $i = 1,\ldots, n$, whereas X_1 is a vector of um's, $X_2,\ldots,X_p$ are the covariables and $\beta_1,\ldots,\beta_p$ are unknown parameters.

Equation (1) can be written in matrix form,

$$Y = X\beta + \varepsilon \qquad (2)$$

where $Y = (Y(s_1),\ldots,Y(s_n))^T$ and $\varepsilon = (\varepsilon(s_n))^T$ are vectors nx1, $\mu = X\beta$ is a vector nx1, X is a matrix n x p composed by the vector of um's and by (p–1) covariates and $\beta = (\beta_1,\ldots,\beta_p)^T$ is the vector of unknown parameters to be estimated.

Random error vector expectation ε, $E(\varepsilon) = 0$ is a zero vector, and its scaling matrix is $\Sigma = [(\sigma_{ij})]$, where $\sigma_{ij} = C(s_i, s_j)$.

Assuming Σ is not singular and that the matrix X is full rank (rank $(X) = p$), the scaling matrix Σ can assume a spatial structure of:

$$\Sigma = [(\sigma_{ij})] = \varphi_1 l_n + \varphi_2 R \qquad (3)$$

where l_n is the identity matrix nxn, φ_1 is the parameter that determines the nugget effect, φ_2 is the parameter that defines the *sill* or contribution, $R = R(\varphi_3)$ is a symmetric nxn matrix, and the reach a is a function of φ_3, ie $\alpha = g(\varphi_3)$.

The elements i and j of the matrix Σ are $C(s_i,s_j) = C(h_{ij}) = \varphi_2 r_{ij}$, where $h_{ij} = \|s_i - s_j\|$, whereas the elements r_{ij} are from the matrix R of the form $r_{ij} = \frac{1}{\varphi_2} C(h_{ij})$ for $i \neq j$ and $r_{ij} = 1$, for $i = j = 1,\ldots, n$.

The joint probability density function of Y is:

$$f(Y|X\beta,\Sigma,v) = \frac{\Gamma\left(\frac{v+p}{2}\right)}{\Gamma\left(\frac{v}{2}\right)(v\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\ [1+(Y-X\beta)^T\Sigma^{-1}(Y-X\beta)]^{-\left(\frac{v+p}{2}\right)} \qquad (4)$$

where $\Gamma(.)$ is the gamma function. For $v > 1$, $\mu = X\beta$ is the vector of means and for $v > 2$, $\left(\frac{v}{v-2}\right)\Sigma$ is the covariance matrix.

Let $\theta = (\beta^T\ \Phi^T)^T$ be the vector of unknown parameters and Y the vector of the observed data, where $Y \sim t_n(X\beta,\Sigma,v)$ with fixed v. The likelihood function logarithm of the observed data Y in relation to θ is given by

$$L(\theta|Y) = \log\Gamma\left(\frac{v+p}{2}\right) - \frac{1}{2}\log|\Sigma| - \log\Gamma\left(\frac{v}{2}\right) - \frac{p}{2}\log(v\pi) - \left(\frac{v+p}{2}\right)\log[1+(Y-X\beta)^T\Sigma^{-1}(Y-X\beta)]$$

Algorithm for parameter estimation (EM algorithm)

The parameter vector $\theta = (\beta^T\ \Phi^T)^T$ is estimated by the EM algorithm, based on the mixture of normal distribution that obtains Student's t-distribution (Liu & Rubin, 1995).

We consider $Y_c = (Y,Y_m)$ as the complete data set with a parameterized density $f(Y_c|\theta)$ by a vector of n-dimensional parameters, $\theta \in \Theta \subset \Re^s$, with $s=p+q$ (q=3), where, Y and Y_m are the observed and non-observed data, respectively. The maximum likelihood estimate (ML) of θ can be obtained based on the complete data of the likelihood function log and the EM algorithm (Dempster et al., 1997). This algorithm consists of two steps: Expectation (Step E) and Maximization (Step M).

Step E: $Q(\theta|\theta^{(r)}) = E\{L_c(\theta|Y_c)|Y,\theta^{(r)}\}$ is defined; at this step the expectation $Q(\theta|\theta^{(r)})$ results from conditional distribution $f(Y_m|Y,\theta^{(r)})$ in the r^{th} iteration.

Step M: $\theta^{(r+1)} = Arg\ \max_\theta Q(\theta|\theta^{(r)})$; at this step a $\theta^{(r+1)}$ is determined, which maximizes $Q(\theta|\theta^{(r)})$. The sequence derived from the EM algorithm iterations converges to the likelihood maximum estimate of $\hat{\theta}$.

Local influence

A perturbation vector $\omega = (\omega_1,\ldots,\omega_n)^T$ is considered, varying in a region $\Omega \subset \Re^d$, used as additive perturbation of $Y_\omega = Y + \omega$. Let the likelihood function be $L(\theta,\omega|Y)$ in the observed data and $L_c(\theta,\omega|Y_c)$ in the complete data for the perturbed model and assuming that there is ω^0 so that $L(\theta,\omega^0|Y) = L(\theta|Y)$ and $L_c(\theta,\omega^0|Y_c) = L_c(|Y_c)$ for every θ, and also that $\hat{\theta}(\omega^0)$ is the ML estimator of θ for $L(\theta,\omega|Y)$. Cook (1986) considers the displacement of the likelihood function log $LD(\omega) = 2[L(\hat{\theta}|Y) - L(\hat{\theta}(\omega)|Y)]$ to assess the local influence of a small perturbation. Due to the difficulty of its application to complicated models, Zhu & Lee (2001) proposed an option with a shift function for $LD(\omega)$, defined by $f_Q(\omega) = 2[Q(\hat{\theta}|\hat{\theta}) - Q(\hat{\theta}(\omega)|\hat{\theta})]$, where $\hat{\theta}(\omega)$ is the estimate of θ that maximizes $Q(\hat{\theta},\omega|\hat{\theta}) = E[L_c(\theta,\omega|Y_c)Y,\hat{\theta}]$. The local influence graphic of $f_Q(\omega)$ is defined as $\alpha(\omega) = [\omega^T,f_g(\omega)]^T$. In this case, the normal curve C_{fg},l of $\alpha(\omega)$ in ω^0 in the direction of some unit vector l can be used to summarize the local behavior of $f_Q(\omega)$. The normal curvature C_{fQ},l for $\alpha(\omega)$ and ω^0 is defined: $C_{fQ},l = 2[l^T\Delta^T_{\omega 0}\{-\ddot{Q}_\theta(\hat{\theta})\}^{-1}\Delta_{\omega^0}l]$, where $\ddot{Q}_\theta(\hat{\theta}) = \partial^2 Q(\theta,|\hat{\theta})/\partial\theta\partial\theta^T|_{\theta=\hat{\theta}}$ and $\Delta_{\omega^0} = \partial^2 Q(\theta,\omega^0|\hat{\theta})/\partial\theta\partial\omega^T|_{\theta=\hat{\theta}(\omega)}$. Poon & Poon (1999) include the conformal normal curvature B_{fQ},l for ω^0 in the direction of some unit vector l, such as:

$$B_{f_Q,l} = \frac{-2l^T\ddot{Q}_{\omega^0}l}{tr(-2\ddot{Q}_{\omega^0})} \qquad (5)$$

where $\ddot{Q}_{\omega^0} = \partial^2 Q(\hat{\theta}(\omega)|\hat{\theta})/\partial\omega\partial\ \omega^T|_{\omega=\omega^0}$, and B_{fQ},l is a function of one to one of the normal curvature, considering values between [0,1].

Given the matrix $B_{fQ,l}$ and considering

$$\mathrm{Ci} = 2^* \mid b_{ii} \mid \quad (6)$$

where b_{ii} are the elements of the main diagonal of matrix $B_{fQ,l}$, we find graph C_i according to order i, to analyze the existence of influential observations in the spatial dependence structure. Zhu & Lee (2001) stated that for C_i, the i^{th} point is influential in the spatial dependence structure if:

$$b_{ii} > \overline{B}_{f_{Q,l}} + 2S_{B_{f_{Q,l}}}, \text{ for } i,j=1,\ldots, n \quad (7)$$

where b_{ii} is an element of matrix $B_{fQ,l}$, $\overline{B}_{f_{Q,l}} = \frac{1}{n}\sum_{i=j=1}^{n} b_{ij}$ and $S^2_{B_{f_{Q,l}}} = \frac{1}{n}\sum_{i=j=1}^{n}(b_{ij} - \overline{B}_{f_{Q,l}})^2$.

In the study of spatial data, universal kriging has been used as a measure of prediction, to obtain values for the regionalized variable at non-sampled points. Let $Y_0 = Y(s_0)$ be the universal kriging predictor of location $s_0 \in S$. The mean of Y_0 is given by $x_0^T\beta$, where $x_0^T = (x_{01},\ldots,x_{0p})$ and $x_{0j} = x_j(s_0)$ for $j = 1,\ldots,p$.

The predictor of the least mean square error is given by

$$p(s_0,\theta) = x_0^T\beta + C_0^T\Sigma^{-1}(Y - X\beta)$$

thus $C_0^T = (C(h_{10}),\ldots,C(h_{n0}))$, with $h_{i0} = \|s_i - s_0\|$ for $i = 1,\ldots,n$.

Thus, we have $S(l) = l^T\,\dot{p}(s_0,\theta)$, where $\dot{p}(s_0,\theta)$ is a vector n x1 given by

$$\dot{p}(s_0,\theta) = \left\{ -\Delta_\omega^T Q^{-1}\frac{\partial p(s_0,\theta)}{\partial\theta}\right\}_{|\theta=\hat{\theta}} \quad (8)$$

In equation (8) we have $\frac{\partial p(s_0,\theta)}{\partial\theta} = \left(\frac{\partial p(s_0,\theta)}{\partial\beta^T}, \frac{\partial p(s_0,\theta)}{\partial\Phi^T}\right)^T$ where $\frac{\partial p(s_0,\theta)}{\partial\beta^T} = x_0 - X^T\Sigma C_0$ and $\frac{\partial p(s_0,\theta)}{\partial\Phi^T} = \left[\left(\frac{\partial p(s_0,\theta)}{\partial\varphi_j}\right)\right]$, thus $\frac{\partial p(s_0,\theta)}{\partial\varphi_j} = \left\{\frac{\partial C_0^T}{\partial\varphi_j} - C_0^T\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j}\right\}\Sigma^{-1}(Y - X\beta)$ and $\frac{\partial C_0^T}{\partial\varphi_j} = \left(\frac{\partial C(h_{10})}{\partial\varphi_j},\ldots,\frac{\partial C(h_{n0})}{\partial\varphi_j}\right)$ for $j = 1,2,3$.

The direction of maximum local slope for the linear predictor is given by equation (9)

$$l_{p(s_0,\theta)} = \frac{\dot{p}(s_0,\theta)}{\|\dot{p}(s_0,\theta)\|} \quad (9)$$

Graph $l_{p(s_0,\theta)}$ is used to assess the influence on the linear predictor,

MATERIAL AND METHODS

The experimental data were collected in the 2004/2005 growing season, through a research that was conducted in a field of commercial grain production (56.68 ha) in Cascavel, in western Paraná (approx. 24.95º S, 53.57º W, 650 m asl). The soil was classified as a clayey Oxisol (specifically Latossolo Vermelho distroférrico), with a long-term crop succession of oat in the winter and soybean in the summer. The local climate is mild, mesothermal, and super humid, Cfa (Köppen classification), with moderate temperatures and well-distributed rainfall. The mean winter temperature is below 16 °C, with possible frost, and summers are hot, with temperatures above 30 ºC.

The soybean variety COODETEC 216 (CD216) was planted in no-tillage in the experimental area. For the study, 47 plots were marked, all of which were georeferenced using a Trimble GeoExplorer 3 GPS receiver (*Global Positioning System*) and the static method with post-processed differential correction, for the correct location in the spatial system of geographic coordinates *Universal Transverse Mercatur* (UTM), using metric coordinates.

From each of the 47 plots, the following data were collected: soybean yield (*Prod*) [t ha^{-1}], soil density (*Des*) [kg dm^{-3}] in the layers 00–10, 10–20 and 20–30 cm and soil penetration resistance (*SPR*) [MPa] in the same layers.

The *Des* was determined by the volumetric ring method (Kiehl, 1979), where one repetition was applied per depth per sampling. The *SPR* in the layers examined was measured with a penetrometer (SC-60, Soil Control; shaft 600 mm, diameter 9.53 mm), equipped with a cone at the tip (base area 129.3 mm², diameter 12.83 mm, corner angle 30°). A mean value of mechanical resistance to penetration was calculated based on four random replications per location and experimental plot.

Initially, descriptive statistics were obtained for the variables: soybean yield (*Prod*), and penetration resistance, and density in the soil layers 00–10, 10–20, and 20–30 cm deep (RSP_{0-10}, RSP_{10-20}, e RSP_{20-30}; Des_{0-10}, Des_{10-20} e Des_{20-30}, respectively) to assess their behavior, identify the presence of discrepant points and possible causes.

Thereafter, the spatial analysis of soybean yield was performed, considering *SPR* and *Des* as their covariates in a linear spatial model. Then n-variate Student's t-distribution and n-variate normal distribution were taken into account, and the yield spatial dependence structure was determined for both, using the method of maximum likelihood. The linear spatial model parameters of soybean yield as a function of covariates and of covariance structure were estimated by:

$$\mu(s_i) = \beta_1 + \beta_2 RSP_{0-10}(s_i) + \beta_3 RSP_{10-20}(s_i) + \beta_4 RSP_{20-30}(s_i) + \beta_5 Des_{0-10}(s_i) + \beta_6 Des_{10-20}(s_i) + \beta_7 Des_{20-30}(s_i)$$

and

$$\Sigma = [(\sigma_{ij})] = \varphi_1 I_n + \varphi_2 R$$

where $\beta_1,\ldots,\beta_1$ and φ_1,φ_2, φ_3 are the unknown parameters to be estimated.

The cross-validation criterion, Akaike information criterion (Faraco et al., 2008), and the maximum value of the log-likelihood function (LLF) were used for the choice of model space. Local influence analysis was performed to identify influential points. Universal kriging interpolation of the variable under study was the final step along with the creation of thematic maps.

For data analysis *software* R (R Development Core Team, 2005) was used, and its modules geoR (Ribeiro JR & Diggle, 2001) and Splancs (Rowlingson & Diggle, 1993).

RESULTS AND DISCUSSION

The soybean yield mean was 3.23 t ha^{-1}, standard deviation SD = 0.38 t ha^{-1} and the coefficient of variation CV = 11.79 %, identifying data homogeneity. With increasing depth, the covariate of soil penetration resistance (SPR) decreased the values of the mean, first quantile (Q_1), median and trird qualite (Q_3), and increased akewness and kurtosis. It was observed that for SPR, in all layers, there are sample data above 2.60 MPa and that in RSP_{0-10}, 50 % of the data were over 2.64 MPa. According to Canarache (1990), SPR values in the range [1.1 to 2.59] MPa are not very restrictive for roots, but root-limiting in the range [2.6 to 5.0] MPa. It was also observed that the standard deviation of SPR does not vary much in the three layers, remaining within 0.49 and 0.56 MPa. The three layers have homogeneity (coefficient of variation below 30 %). For soil density (*Des*), all statistical measures showed increases from layer 0–10 cm to layer 10–20 cm, and reduction in statistical measures from layer 10–20 cm to layer 20–30 cm, which features a compacter layer below the soil surface, that prevents water infiltration. The mean, standard deviation, and density coefficient of variation do not change much in the three layers. As for *SPR*, homogeneity was observed for *Des* in the three layers (coefficient of variation below 30 %).

Coefficients of skewness and kurtosis were calculated. In covariates RSP_{20-30}, Des_{10-20}, and Des_{20-30}, the calculated coefficients do not belong to the 95 % skewness and kurtosis coefficient confidence intervals constructed by Jones (1969) that characterize normal distribution of probability.

The box-plot graph of soybean yield (Figure 1a), shows an outlier point, with a value of 2.09 t ha^{-1}, which is the 13th value of the data series, located in the experimental area (Figure 1b).

Spatial analysis: Tables 1 and 2 show results of linear spatial model parameter estimates of Equation (10) of soybean yield, setting the following models: exponential (Exp), Gaussian (Gaus), and Matérn with kappa parameter at 0.7. Student's *t*-distribution was considered for data with degrees of freedom equal to 3 (υ = 3) and normal distribution. The Kolmogorov-Smirnov test was used to confirm the hypothesis that the population distribution, from which the given sample was withdrawn, follows a *t*-distribution probability with υ = 3 degrees of freedom.

Table 2 shows that estimates of parameters $\hat{\varphi}_1$ (nugget effect), $\hat{\varphi}_2$ (contribution) and $\hat{\varphi}_3$(function of range) have small variations among the three models, assuming Student's *t*-distribution with υ = 3 or with normal distribution. For both distributions the lowest standard deviation was given by the Gaussian model.

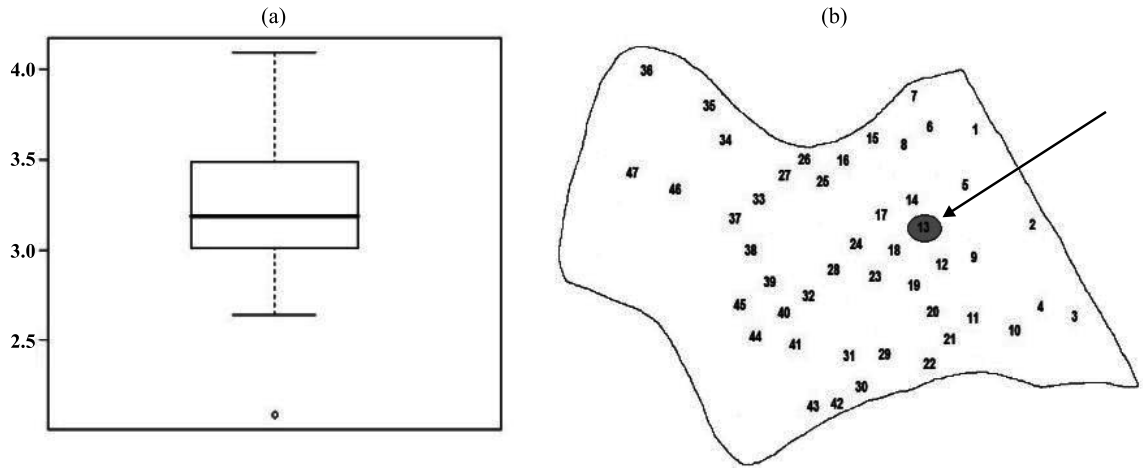

Figure 1. (a) Box-plot graph of soybean yield and (b) Sketch of experimental area.

The degree of spatial variability classified by Canarache (1990) by using the coefficient of relative nugget effect (*RNE*) shows moderate spatial dependence of the variables studied.

Validation of models: Table 3 shows results of the selection criteria for cross-validation models, Akaike information criterion (AIC) and maximum value of the log-likelihood function (LLF) for the soybean yield variate. Considering these results, it appears that both by the Student's t-distribution with $\upsilon = 3$ and by normal distribution, the best-fitting was the Gaussian model.

Local Influence Analysis: Graphs C_i in figure 2a,b, considering the limit of the graph defined in Equation (7), shows that by adopting *t*-distribution with $\upsilon = 3$, elements 13 and 30 are identified as influential for the structure of spatial dependence. However, when adopting normal distribution, elements

Table 1. Parameters estimated by LM assuming Student's *t*-distribution with $\upsilon = 3$ and normal for soybean yield using theoretical models: exponential (Exp), Gaussian (Gaus), and Matérn with k = 0.7 (Matérn)

Model	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		$\hat{\beta}_4$		$\hat{\beta}_5$		$\hat{\beta}_6$		$\hat{\beta}_7$	
	t	N	*t*	N	*t*	N	*t*	N	*t*	N	*t*	N	*t*	N
Exp	4.56 (1.07)	4.37 (0.97)	-0.10 (0.12)	0.01 (0.11)	0.06 (0.12)	0.06 (0.12)	-0.01 (0.09)	-0.02 (0.09)	-0.29 (0.64)	-0.29 (0.64)	0.06 (0.52)	0.06 (0.52)	-0.86 (0.80)	-0.87 (0.80)
Gaus	4.35 (1.07)	4.45 (1.07)	-0.09 (0.12)	-0.10 (0.11)	0.07 (0.12)	0.06 (0.12)	-0.01 (0.09)	-0.02 (0.09)	-0.27 (0.64)	-0.26 (0.64)	0.13 (0.53)	0.11 (0.52)	-0.81 (0.80)	-0.85 (0.79)
Matérn k=0.7	4.73 (1.07)	4.56 (1.07)	-0.11 (0.11)	-0.10 (0.12)	0.05 (0.12)	0.06 (0.12)	-0.02 (0.09)	-0.02 (0.09)	-0.30 (0.64)	-0.29 (0.64)	0.01 (0.52)	0.06 (0.52)	-0.91 (0.80)	-0.87 (0.80)

Student's *t*-distribution with $\upsilon = 3$, N: normal distribution; In brackets, the standard deviations of each parameter estimated.

Table 2. Parameters estimated when adopting Student's *t*-distribution with $\upsilon = 3$ and normal distribution for the yield variable

Model	$\hat{\varphi}_1$		$\hat{\varphi}_2$		$\hat{\varphi}_3$		*RNE*	
	t	N	*t*	N	*t*	N	*t*	N
Exp	0.0784 (0.03)	0.0791 (0.03)	0.0459 (0.04)	0.0455 (0.04)	138.41 (1.64)	142.27 (1.73)	37 %	36 %
Gaus	0.0941 (0.03)	0.0909 (0.02)	0.0367 (0.03)	0.0333 (0.03)	147.14 (0.25)	192.35 (0.51)	28 %	27 %
Matérn k=0.7	0.0786 (0.03)	0.0844 (0.03)	0.0519 (0.05)	0.0403 (0.04)	194.48 (0.75)	126.73 (0.53)	40 %	32 %

Student's *t*-distribution with $\upsilon = 3$, N: normal distribution; In brackets, the standard deviations of each parameter estimated. $EPR(\%) = \varphi_2/(\varphi_1 + \varphi_2)$: coefficient of relative nugget effect.

Table 3. Results of model validation for the soybean yield variate

Dist.	Model	EM	EMR	*S*	S_{ER}	EA	AIC	LLF
t	Exp	-0.0004	-0.0005	0.3484	0.6989	12.04	-94.11	45.95
	Gaus	-0.0004	-0.0005	0.3451	0.6686	11.98	-95.01	44.26
	Matérn k=0.7	-0.0005	-0.0008	0.3503	0.7291	12.01	-93. 60	46.70
N	Exp	-0.0055	-0.0039	0.3931	1.0812	14.29	-302.83	-16.18
	Gaus	-0.0053	-0.0038	0.3892	1.0819	14.05	-305.44	-15.86
	Matérn k=0.7	-0.0054	-0.0038	0.3925	1.0812	14.26	-303.58	-16.14

t: Student's *t*-distribution with $\upsilon = 3$, N: normal distribution; EM: mean error, EMR: reduced mean error; *S*: standard deviation of errors; S_{ER}: standard deviation of mean error and *EA*: absolute error; AIC: Akaike transformation criterium, LLF: maximum value of log-likelihood function.

13, 23, and 30 are identified as influential for the structure of spatial dependence, with greater emphasis on element 30.

According to graphs l_p in figure 2c,d, which assess the influence on the linear predictor, by adopting t-distribution with $\upsilon = 3$, element 13 continues to be considered influential. However, when normal distribution was adopted, element 31 was identified as influential.

Following with the analysis of graphs C_i and l_p, it was decided to remove two observations; 13 (2.09 t ha^{-1}) and 30 (2.87 t ha^{-1}) separately, to identify the influence of these points in the study of spatial variability. To distinguish the new data sets, we considered Prod: total data, Prod (13): data excluding element 13, and Prod (30): data excluding element 30.

Descriptive analysis after the removal of influential values: Table 4 shows that element 13 produced major changes in the descriptive measures besides the removal of element 30. The removal of these elements also altered data skewness and kurtosis, but according to criteria used to assess normality, new data sets still showed normality.

Parameter vector estimate $\hat{\theta} = (\hat{\beta}^T, \hat{\Phi}^T)^T$ after the removal of influential values: The parameters were estimated for the three models, for the t-distribution with $\upsilon = 3$ and for the normal distribution, for each one of the variables Prod, Prod (13), and Prod (30). Considering these results, it appears again that both by the t-distribution with $\upsilon = 3$ and by normal distribution, the best-fitting model is the Gaussian model, ie, the removal of influential points did not alter the choice of the set model.

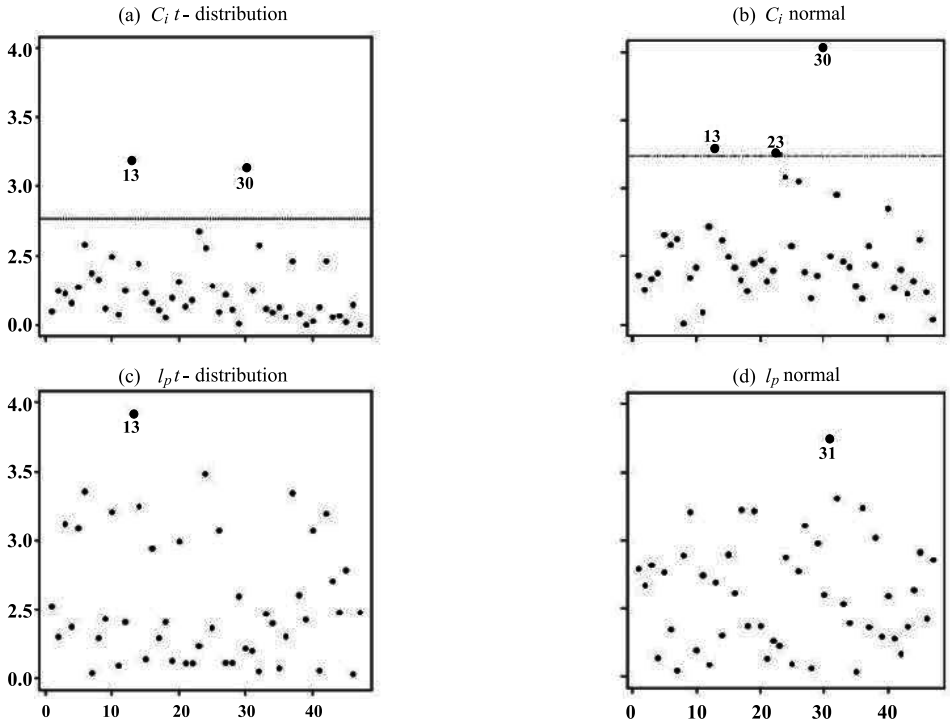

Figure 2. Graphs C_i and l_p for soybean yield.

Table 4. Descriptive statistics for the variables Prod, Prod (13) and Prod (30)

Variable	N	Mean	Mininum	Maximum	Q_1	Medium	Q_3	S	CV(%)	Sk	K
Prod.	47	3.23	2.09	4.09	3.01	3.19	3.49	0.38	11.79	-0.13	0.76
Prod(13)	46	3.25	2.64	4.09	3.02	3.19	3.49	0.34	10.59	0.43	-0.34
Prod(30)	46	3.23	2.09	4.09	3.02	3.19	3.49	0.38	11.79	-0.18	0.83

n: number of data; Min.: minimum value; Max.: maximum value; Q_1: first quartile; Q_3: third quartile; S: standard deviation; CV: coefficient of variation; Sk: Skewness and K: kurtosis.

Comparing the estimates of the parameter vector $\hat{\beta}$ (Table 5) obtained from the data sets Prod (13) and Prod (30), with the estimates obtained with complete data (Prod), it appears that both for the t-distribution (v = 3) and for the normal distribution, estimates $\hat{\beta}_1$ and $\hat{\beta}_3$ decreased, and estimates $\hat{\beta}_2$, $\hat{\beta}_4$ and $\hat{\beta}_5$ increased. Estimates $\hat{\beta}_6$ and $\hat{\beta}_7$ have different variations when adopting t-distribution and normal distribution.

Looking at table 6 and considering the t-distribution (v = 3), it appears that changes in the parameters of the structure of spatial variability after the removal of elements 13 (Prod (13)) and 30 (Prod (30)) were the same, showing reduction of the estimates $\hat{\varphi}_1$ and $\hat{\varphi}_2$. The estimate $\hat{\varphi}_3$ was reduced when obtained with data set Prod (13) and increased when obtained together with the data Prod (30). These changes were more significant for the data set Prod (13) than for Prod (30), when compared to yield (Prod) of the complete data.

However, considering normal distribution, after the removal of element 13 (Prod (13)), there was reduction of $\hat{\varphi}_1$ and $\hat{\varphi}_2$ and increase of $\hat{\varphi}_3$, and after the removal of element 30 (Prod (30)) $\hat{\varphi}_1$ there was an increase and the estimates $\hat{\varphi}_2$ and $\hat{\varphi}_3$ were reduced. But again these changes were more significant for data set Prod (13) than for Prod (30), when compared to yield (Prod) of the complete data. Therefore, it can be said that the points identified as influential affect parameter estimates.

When the coefficient of relative nugget effect is examined in the three data sets, it is possible to note that there is no change of the spatial dependence, it remained moderate.

Figure 3 shows the thematic maps of data sets Prod, Prod (13), and Prod (30), based on universal kriging interpolation, using the covariates for the estimation of parameters $\hat{\theta} = (\hat{\beta}^T, \hat{\Phi}^T)^T$.

The maps were constructed using the Gaussian model adopting t-distribution (v = 3) and normal distribution. Table 7 shows the percentage of each class on the maps of the variable. From the maps of figure 3 and table 7, it can be observed in figure 3a,b that there was a reduction of the area with yield in the first class interval (2.7 and 3.1 t ha^{-1}) changing from 35.84 to 31.81 % of the area, and also an increase in areas of the other classes, more pronounced in the 3.1 and 3.3 class t ha^{-1}, from 29.94 to 32.17 %. Comparing figure 3a,c, it is possible to note a reduction of the area with yield in the interval of the 1st and 4th classes, and increases of the other classes.

Table 5. Parameters estimated when adopting t-distribution with v = 3 and normal for variables Prod, Prod (13), and Prod (30), using Gaussian (Gaus)

Variable	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		$\hat{\beta}_4$		$\hat{\beta}_5$		$\hat{\beta}_6$		$\hat{\beta}_7$	
Distribution	*t*	N	*t*	N	*t*	N	*t*	N	*t*	N	*t*	N	*t*	N
Prod.	4.44 (1.09)	4.45 (1.07)	-0.10 (0.12)	-0.10 (0.11)	0.06 (0.12)	0.06 (0.12)	-0.02 (0.09)	-0.02 (0.09)	-0.27 (0.65)	-0.26 (0.64)	0.12 (0.53)	0.11 (0.52)	-0.84 (0.81)	-0.85 (0.79)
Prod(13)	4.14 (0.97)	4.38 (0.95)	0.02 (0.11)	0.01 (0.11)	-0.03 (0.11)	-0.05 (0.11)	-0.01 (0.08)	0.00 (0.08)	-0.14 (0.58)	-0.17 (0.57)	0.18 (0.48)	0.09 (0.46)	-0.83 (0.72)	-0.88 (0.70)
Prod(30)	4.24 (1.13)	4.15 (1.13)	-0.07 (0.12)	-0.07 (0.12)	0.01 (0.14)	0.01 (0.14)	0.11 (0.17)	0.11 (0.17)	-0.24 (0.64)	-0.24 (0.64)	0.09 (0.52)	0.12 (0.52)	-0.82 (0.80)	-0.79 (0.80)

Student's t-distribution with v = 3, N: normal distribution; In brackets, the standard deviations of each parameter estimated.

Table 6. Spatial parameters estimated when adopting t-distribution (v = 3) and normal for variables Prod, Prod (13), and Prod (30), using Gaussian (Gaus)

Variable	$\hat{\varphi}_1$		$\hat{\varphi}_2$		$\hat{\varphi}_3$		E	
	t	N	*t*	N	*t*	N	*t*	N
Prod.	0.0969 (0.03)	0.0909 (0.02)	0.0367 (0.03)	0.0333 (0.03)	182.18 (0.44)	192.35 (0.51)	27%	27%
Prod(13)	0.0716 (0.02)	0.0739 (0.02)	0.0294 (0.02)	0.0315 (0.02)	145.26 (0.24)	255.19 (1.04)	32%	30%
Prod(30)	0.0944 (0.02)	0.0924 (0.03)	0.0292 (0.03)	0.0309 (0.03)	210.94 (0.71)	182.79 (0.48)	24%	25%

Student's t-distribution with v = 3, N: normal distribution; In brackets, the standard deviations of each parameter estimated. $E(\%) = \varphi_2/(\varphi_1 + \varphi_2)$: coefficient of relative nugget effect.

Table 7. Percentage of class on the maps of soybean yield

Variable	1st class (2.7 |–3.1 t ha^{-1})		2nd class (3.1 |–3.3 t ha^{-1})		3rd class (3.3 |–3.5 t ha^{-1})		4th class (3.5 |–3.9 t ha^{-1})	
	t	N	t	N	t	N	t	N
Prod.	35.84%	22.54%	29.94%	47.75%	22.88%	29.51%	11.34%	0.20%
Prod(13)	31.81%	21.28%	32.17%	45.34%	23.68%	33.38%	12.34%	0.00%
Prod(30)	29.61%	17.10%	34.33%	53.75%	26.24%	28.32%	9.82%	0.83%

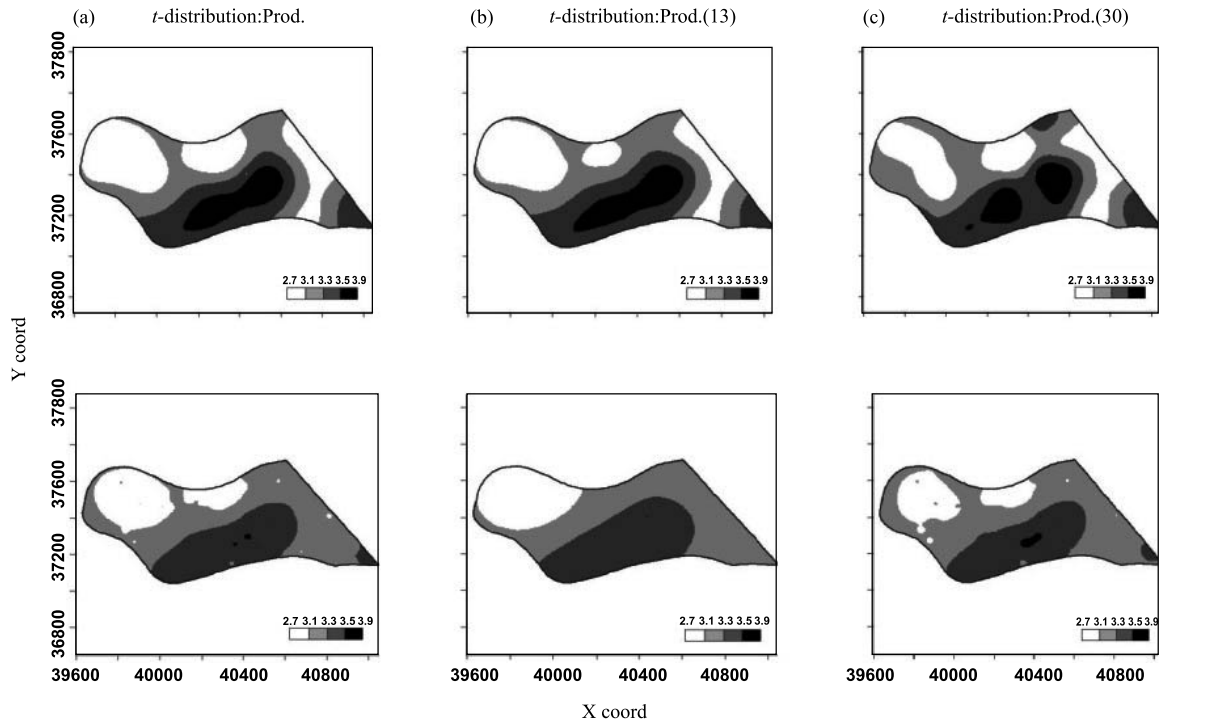

Figure 3. Soybean yield map - Gaussian model.

By comparing figure 3d,e, a reduction in the areas corresponding to the 1st, 2nd and 4th classes is observed, and an increase in the 3rd class. Comparing figure 3d,f there is a reduction in the 1st and 3rd, and an increase in the 2nd and 4th classes. Thus, one can say that the points identified as influential affect the thematic maps underlying the management of the area with a view to future interventions in the soil treatment. It is therefore important to take all factors into consideration that may alter spatial analysis.

CONCLUSIONS

1. Observation 13 is not only an outlier, but also an influential value. This element (13) had greater influence than element 30 on the estimate of the parameters, as well as on the construction of thematic maps, when *t*-distributions or normal distribution is adopted.

2. In a comparison of the maps, it was noted that the map generated by *t*-distribution is less changed with the removal of an influential value (13) than the map generated by normal distribution. Therefore, it is possible to perform spatial analysis from Student's *t*-distribution with little concern about influential values, which do not influence this distribution.

ACKNOWLEDGEMENTS

The authors thank the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), National Council for Scientific and

Technological Development (CNPq) and the Fundação Araucária for financial support.

LITERATURE CITED

BORSSOI, J.A.; URIBE-OPAZO, M.A. & GALEA-ROJAS, M.J. Diagnostic techniques applied in geostatistics for agricultural data analysis. R. Bras. Ci. Solo, 6:1-16, 2009.

BORSSOI, J.A.; URIBE-OPAZO, M.A. & GALEA, M. Técnicas de diagnóstico de influência local na análise espacial da produtividade da soja. Eng. Agríc., 31:376-387, 2011.

CANARACHE, A.P. A generalized semi-empirical model estimating soil resistance to penetration. Soil Tillage Res., 16:51-70, 1990.

CHRISTENSEN, R.; JOHNSON, W. & PEARSON, L.M. Prediction diagnostics for spatial linear models. Biometrika, 79:583-591, 1992.

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

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

DEMPSTER, A.; LAIRD, N. & RUBIN, D. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., Series B, 39:1-38, 1997.

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.

JONES, T. Skewness and kurtosis as criteria of normality in observed frequency distributions. J. Sedim. Res., 39:1622-1627, 1969.

KIEHL, E.J. Manual de edafologia: Relações solo-planta. Piracicaba, Agronômica Ceres, 1979. 264p.

LIU, C. & RUBIN, D.B. ML estimation of the t distribution using EM and its extensions ECM and ECME. Sinica Stat., 5:19-39, 1995.

MILITINO, A.F.; PALACIOS, M.B. & UGARTE, M.D. Outlier detection in multivariate spatial linear models. J. Stat. Planning Infer., 136:125-146, 2006.

MOLIN, J.P. Definição de unidades de manejo a partir de mapas de produtividade. Eng. Agríc., 22:83-92, 2002.

PAULA, G.A. Modelos de regressão com apoio computacional. São Paulo, Instituto de Matemática e Estatística – USP, 2004. 233p.

POON, W.Y. & POON, Y.S. Conformal normal curvature and assessment of local influence. J. Royal Stat. Soc.. Series B (Statistical Methodology), 61:51-61, 1999.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. R Foundation Stat. Computing, ISBN 3-900051-07-0. Available at: <http://www.R-project.org>. Accessed June 3, 2009.

RIBEIRO JR., P.J. & DIGGLE P.J. geoR: A package for geostatistical analysis. R-NEWS, 01. Available at <http://cran.r-project.org/doc/Rnews. 2001>. Accessed June 3, 2009.

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

ZHU, H.T. & LEE, S.Y. Local influence for incomplete-data models. J. Royal Stat. Soc. B, 63:p.111-126, 2001.