TÉCNICAS DE DIAGNÓSTICO DE INFLUÊNCIA LOCAL NA ANÁLISE ESPACIAL DA PRODUTIVIDADE DA SOJA

JOELMIR A. BORSSOI[1], MIGUEL A. URIBE-OPAZO[2], MANUEL GALEA[3]

RESUMO: A modelagem da estrutura de dependência espacial pela abordagem da geoestatística é fundamental para a definição de parâmetros que definem esta 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 afetada pela presença de observações atípicas nos dados amostrados. O desenvolvimento deste trabalho teve por objetivo utilizar técnicas de diagnóstico de influência local em modelos espaciais lineares gaussianos, utilizados em geoestatística, para avaliar a sensibilidade dos estimadores de máxima verossimilhança e máxima verossimilhança restrita na presença de dados discrepantes. Estudos com dados experimentais mostraram que tanto a presença de valores atípicos como de valores considerados influentes, pela análise de diagnóstico, pode exercer forte influência nos mapas temáticos, alterando, assim, a estrutura de dependência espacial. As aplicações de técnicas de diagnóstico de influência local devem fazer parte de toda análise geoestatística a fim de garantir que as informações contidas nos mapas temáticos tenham maior qualidade e possam ser utilizadas com maior segurança pelo agricultor.

PALAVRAS-CHAVE: geoestatística, máxima verossimilhança, máxima verossimilhança restrita.

DIAGNOSTIC TECHNIQUES OF LOCAL INFLUENCE IN SPATIAL ANALYSIS OF SOYBEAN YIELD

ABSTRACT: Modeling of spatial dependence structure, concerning geoestatistics approach, is an indispensable tool for fixing parameters that define this structure, applied on interpolation of values in places that are not sampled, by kriging techniques. However, the estimation of parameters can be greatly affected by the presence of atypical observations on sampled data. Thus, this trial aimed at using diagnostics techniques of local influence in spatial linear Gaussians models, applied at geoestatistics in order to evaluate sensitivity of maximum likelihood estimators and restrict maximum likelihood to small perturbations in these data. So, studies with simulated and experimental data were performed. Those results, obtained from the study of real data, allowed us to conclude that the presence of atypical values among the sampled data can have a strong influence on thematic maps, changing, therefore, the spatial dependence. The application of diagnostics techniques of local influence should be part of any geoestatistic analysis, ensuring that the information from thematic maps has better quality and can be used with greater security by farmers.

KEYWORDS: geoestatistic, maximum likelihood, restrict maximum likelihood.

INTRODUÇÃO

Nas últimas décadas, vem aumentando a busca por tecnologias que melhorem a qualidade das informações obtidas com respeito a uma área agrícola seja pela busca de melhores equipamentos, seja pela busca de novas formas de análise de dados.

A geoestatística, que tem como fundamento a teoria das variáveis regionalizadas, é um método que considera a distribuição espacial das medidas, permitindo definir o raio de auto-correlação espacial entre elementos amostrais e, consequentemente, a distância máxima onde se tem dependência espacial. O estudo da variabilidade espacial é considerado de grande importância,

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo- SP, joelmirb@ime.usp.br.
[2] Centro de Ciências Exatas e Tecnológicas, Universidade Estadual do Oeste do Paraná, Cascavel - PR.
[3] Departamento de Estatística, Pontifícia Universidad Católica de Chile, Santiago, Chile.
Recebido pelo Conselho Editorial em: 12-9-2009
Aprovado pelo Conselho Editorial em: 25-2-2011

segundo CRESSIE (1993), e não considerar a variabilidade espacial pode impedir que diferenças reais sejam levantadas, o que impossibilitaria um tratamento diferenciado de acordo com as exigências locais.

Na modelagem da variabilidade espacial, são estimados os parâmetros que definem a estrutura de dependência espacial e que serão utilizados na interpolação de valores, em locais não amostrados, pela técnica de krigagem. Assim, são gerados mapas temáticos que poderão ser utilizados para uma aplicação localizada de insumos ou manejo localizado do solo. A qualidade destes mapas depende da qualidade das inferências dos modelos ajustados. Mas, para que a interpolação produza previsões confiáveis e represente a real variabilidade local, o processo de modelagem deve ser feito com muita cautela, principalmente na presença de pontos discrepantes ou influentes.

Nesta perspectiva, o objetivo deste trabalho foi utilizar técnicas de diagnóstico de influência local em modelos espaciais lineares gaussianos para verificar a existência de observações que possam exercer algum tipo de influência nas inferências sobre os parâmetros que definem a estrutura de dependência espacial e na construção dos mapas temáticos. Para tanto, realizaram-se estudos de influência local, utilizando os métodos de estimação de parâmetros de máxima verossimilhança (MV) e de máxima verossimilhança restrita (MVR) a um conjunto de dados da produtividade da soja, tendo como covariáveis a resistência mecânica do solo à penetração e a densidade do solo em diferentes profundidades.

MATERIAL E MÉTODOS

Modelos espaciais lineares

Para modelar dados com estrutura espacial, considere-se, segundo MARDIA & MARSHALL (1984), um processo estocástico gaussiano $\{Z(s),\ s \in S\}$, com $S \subset \Re^d$, sendo $\Re^d$ o espaço euclidiano, d-dimensional ($d \geq 1$). Suponha-se que os dados, $Z(s_1),...,Z(s_n)$, deste processo sejam registrados em localizações espaciais conhecidas, s_i ($i = 1,...,n$), e gerados pelo seguinte modelo:

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

em que, os termos determinístico $\mu(s_i)$ e estocástico $\varepsilon(s_i)$ podem depender da localização espacial onde $Z(s_i)$ foi obtida.

Assume-se que o erro estocástico $\varepsilon(.)$ tem média zero, $E[\varepsilon(s_i)] = 0$, e que a variação entre pontos no espaço é determinada por alguma função de covariância $C(s_i,s_u) = COV[\varepsilon(s_i), \varepsilon(s_u)]$ e que, para algumas funções conhecidas de s, como $x_1(s),..., x_p(s)$, a média do processo estocástico é:

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

em que,

$\beta_1,...,\beta_p$ - parâmetros desconhecidos e a serem estimados.

De forma equivalente, em notação matricial, tem-se:

$$\boldsymbol{Z} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}. \quad (3)$$

Então,

$E(\varepsilon) = \boldsymbol{0}$ (vetor nulo) e a matriz de covariância de $\boldsymbol{\Sigma} = [(\sigma_{iu})]$, onde $\sigma_{iu} = C(s_i, s_u)$. Assume-se que $\boldsymbol{\Sigma}$ é não singular, que $\boldsymbol{X}$ tem colunas com posto completo e que $\boldsymbol{Z}$ segue uma distribuição gaussiana multivariada com média $\boldsymbol{X\beta}$ e matriz de covariância $\boldsymbol{\Sigma}$, isto é, $\boldsymbol{Z} \sim N_n(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$.

Considerando-se de maneira particular a forma paramétrica da matriz de covariância, segundo

$$\boldsymbol{\Sigma} = \varphi_1 \boldsymbol{I_n} + \varphi_2 \boldsymbol{R}, \quad (4)$$

em que,

φ_1 - efeito pepita, ou erro de variância;

φ_2 - contribuição, ou variância de dispersão;

$\boldsymbol{R}$ - matriz que é função de φ_3, $\boldsymbol{R} = R(\varphi_3) = [(r_{iu})]$, matriz $n \times n$ simétrica com seus elementos da diagonal $r_{ii} = 1$, $i = 1, ..., n$, em que φ_3 é função do alcance (a) do modelo, e

$\boldsymbol{I_n}$ - matriz identidade $n \times n$.

A forma paramétrica da matriz de covariância, dada na Equação 4, ocorre para vários processos isotrópicos, onde a covariância $C(s_i , s_u)$ é definida segundo a função de covariâncias $C(h_{iu}) = \varphi_2 r_{iu}$, em que $h_{iu} = ||s_i - s_u||$ é a distância euclidiana entre os pontos s_i e s_u. Nas funções de covariâncias $C(h_{iu})$, a variância do processo estocástico $\boldsymbol{Z}$ é $C(0) = \varphi_1 + \varphi_2$, e a semivariância pode ser definida como $\gamma(h) = C(0) - C(h)$.

Influência local

Em muitas situações, pode-se observar um conjunto de dados com observações aberrantes ou discrepantes que podem ser consideradas influentes, isto é, que podem mudar algum tipo de decisão na construção de modelos geoestatísticos.

O método de influência local proposto por COOK (1986) avalia o efeito simultâneo de observações sobre os estimadores de MV e MVR sem a necessidade de sua eliminação do conjunto de dados. As técnicas de influência local têm-se constituído num procedimento útil para realizar análises de sensibilidade de modelos estatísticos e têm sido amplamente utilizadas em modelos de regressão linear e não linear. CHRISTENSEN et al. (1993) estudaram métodos de diagnóstico baseado em eliminação de casos em modelos lineares espaciais. BORSSOI et al. (2009) apresentam o gráfico de influência local $|L_{max}|$ versus i (ordem dos dados) como medida de diagnóstico de influência local para a detecção de pontos influentes.

Para um conjunto de dados observados, seja $l(\theta)$ o logaritmo da função verossimilhança do modelo postulado, em que $\theta = (\beta^T, \varphi^T)^T$, e seja $\boldsymbol{\omega}$ um vetor de perturbações pertencente a um espaço de perturbações Ω. Seja $l(\theta/\boldsymbol{\omega})$ o logaritmo da função verossimilhança correspondente ao modelo perturbado por $\boldsymbol{\omega} \in \Omega$. Assume-se que existe um vetor $\boldsymbol{\omega_0} \in \Omega$ tal que $l(\theta) = l(\theta/\boldsymbol{\omega_0})$, para todo θ e que $l(\theta/\boldsymbol{\omega})$ é duas vezes diferenciável em $(\theta^T, \boldsymbol{\omega}^T)^T$.

O logaritmo da função verossimilhança perturbada $l(\theta/\boldsymbol{\omega})$, considerando a perturbação

$\boldsymbol{Z_\omega} = \boldsymbol{Z} + \boldsymbol{\omega}$ para o modelo normal, é dado por

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

A influência da perturbação $\boldsymbol{\omega}$, nos estimadores MV e MVR do vetor de parâmetros θ, pode ser avaliada pelo afastamento da verossimilhança, definido por:

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

em que,

$\hat{\theta}$ - estimador de máxima verossimilhança de θ do modelo postulado, e

$\hat{\theta}_\omega$ - estimador de máxima verossimilhança de θ do modelo perturbado.

COOK (1986) propôs estudar o comportamento local de $LD(\boldsymbol{\omega})$ em torno de $\boldsymbol{\omega_0}$, utilizando a curvatura normal C_l de $LD(\boldsymbol{\omega})$ em $\boldsymbol{\omega_0}$, na direção de algum vetor unitário $\boldsymbol{l}$, definido como sendo:

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

em que,

$\boldsymbol{L}$ - matriz de informação observada, avaliada em $\theta = \hat{\theta}$, e

$\boldsymbol{\Delta}$ - matriz $(p + q) \times n$ dada por $\boldsymbol{\Delta} = (\boldsymbol{\Delta_\beta}^T, \boldsymbol{\Delta_\varphi}^T)^T$, avaliada em $\theta = \hat{\theta}$ e em $\boldsymbol{\omega} = \boldsymbol{\omega_0}$, em que:

$\boldsymbol{\Delta_\beta} = X^T\Sigma^{-1}$ e $\boldsymbol{\Delta_\varphi} = \dfrac{\partial^2 l(\theta/\omega)}{\partial\varphi\partial\omega^T}$, com $\dfrac{\partial^2 l(\theta/\omega)}{\partial\varphi\partial\omega^T} = (Z_\omega - X\beta)^T\Sigma^{-1}\dfrac{\partial\Sigma}{\partial\varphi_i}\Sigma^{-1}$, $i = 1,...,q$.

A matriz de informação $\boldsymbol{L}$ é definida como: $L = \begin{pmatrix} L_{\beta\beta} & L_{\beta\varphi} \\ L_{\varphi\beta} & L_{\varphi\varphi} \end{pmatrix}$, em que,

$L_{\beta\beta} = -\left(X^T\Sigma^{-1}X\right)$; $L_{\beta\varphi} = \dfrac{\partial^2 l(\theta)}{\partial\beta\partial\varphi^T}$, com $\dfrac{\partial^2 l(\theta)}{\partial\beta\partial\varphi_j} = -X^T\Sigma^{-1}\dfrac{\partial\Sigma}{\partial\varphi_j}\Sigma^{-1}\varepsilon$, $j = 1, ..., q$;

$L_{\varphi\beta} = L_{\beta\varphi}^T$, e $L_{\varphi\varphi} = \dfrac{\partial^2 l(\theta)}{\partial\varphi\partial\varphi^T}$, com seus elementos:

$$\frac{\partial^2 l(\theta)}{\partial\varphi_i\partial\varphi_j} = \frac{1}{2}tr\left\{\Sigma^{-1}\left(\frac{\partial\Sigma}{\partial\varphi_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j} - \frac{\partial^2\Sigma}{\partial\varphi_i\partial\varphi_j}\right)\right\} + \frac{1}{2}\varepsilon^T\Sigma^{-1}\left\{\frac{\partial^2\Sigma}{\partial\varphi_i\partial\varphi_j} - \frac{\partial\Sigma}{\partial\varphi_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j} - \frac{\partial\Sigma}{\partial\varphi_j}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_i}\right\}\Sigma^{-1}\varepsilon$$

Considere-se a matriz $\boldsymbol{B} = \boldsymbol{\Delta}^T\boldsymbol{L}^{-1}\boldsymbol{\Delta}$, e a i-ésima observação definida como

$$C_i = 2*|b_{ii}|, \quad (8)$$

em que, b_{ii} são os elementos da diagonal principal da matriz $\boldsymbol{B}$.

Pode-se utilizar o gráfico de C_i versus i (ordem dos dados) como técnica de influência local para avaliar a existência de observações influentes. Sugere-se considerar a i-ésima observação como influente se seu valor for maior que duas vezes a média das medidas C_i.

Para avaliar a influência local no preditor linear $p(s_0, \theta)$, considere-se que $\dot{p}(s_0,\theta)$ é um vetor n x 1:

$$S(l) = l^T\dot{p}(s_0,\theta) \quad (9)$$

$$\dot{p}(s_0,\theta) = \left\{-\Delta^T L^{-1}\frac{\partial p(s_0,\theta)}{\partial\theta}|_{\theta=\hat{\theta}}\right\} \quad (10)$$

Na eq.(10), tem-se que $\dfrac{\partial p(s_0,\theta)}{\partial\theta} = \left(\dfrac{\partial p(s_0,\theta)}{\partial\beta^T}, \dfrac{\partial p(s_0,\theta)}{\partial\varphi^T}\right)^T$,

em que, $\dfrac{\partial p(s_0,\theta)}{\partial\beta} = \omega_0 - \Sigma^{-1}C_0$; $\dfrac{\partial p(s_0,\theta)}{\partial\varphi} = \left[\left(\dfrac{\partial p(s_0,\theta)}{\partial\varphi_j}\right)\right]$, com seus elementos dados por:

$$\frac{\partial p(s_0,\theta)}{\partial\varphi_j} = \left\{\frac{\partial C_0^T}{\partial\varphi_j} - C_0^T\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j}\right\}\Sigma^{-1}(Z - X\beta) \;;\; \frac{\partial C_0^T}{\partial\varphi_j} = \left(\frac{\partial C(h_{10})}{\partial\varphi_j}, ..., \frac{\partial C(h_{n0})}{\partial\varphi_j}\right) \text{ para } j = 1,..., q.$$

A direção de máxima inclinação local é obtida pela eq.(11):

$$l_{p(s_0,\theta)} = \frac{\dot{p}(s_0,\theta)}{\|\dot{p}(s_0,\theta)\|} \quad (11)$$

Pode-se utilizar a medida $l_{p(s_0,\theta)}$ ou simplesmente l_p para analisar a sensibilidade do preditor linear.

Estudo experimental

Foram analisados os dados da variável Produtividade da soja (Prod) [t ha^{-1}] e as covariáveis: Densidade do solo (Des) [mg m^{-3}] e Resistência mecânica do solo à penetração (RSP) [MPa], nas camadas de 0 - 0,10 m, 0,10 - 0,20 m e 0,20 - 0,30 m de profundidade, nos pontos onde foi coletada a produtividade, obtidos no ano agrícola de 2004/2005, em uma área comercial de produção de grãos de 57 ha. A técnica de amostragem utilizada foi a sistemática alinhada, com distância de 75 m entre pontos.

A área está localizada no município de Cascavel, na região oeste do Paraná, aproximadamente 24,95º sul de latitude, 53,57º oeste de longitude e altitude média de 650 m. O solo é classificado como Latossolo Vermelho distroférrico, e o clima da região é temperado mesotérmico e superúmido, tipo climático Cfa (Köeppen), com temperatura anual média de 21 ºC. Todos os elementos amostrais foram georreferenciados com auxílio de um receptor de sinal do GPS (*Global Positioning System*) marca Trimble, modelo GeoExplorer3, num sistema espacial de coordenadas (UTM).

Neste trabalho, utilizaram-se os modelos exponencial, gaussiano e família Matérn, com os métodos de estimação de parâmetros máxima verossimilhança (MV) e máxima verossimilhança restrita (MVR) (MARDIA & MASHALL, 1984). Para escolher o modelo espacial que melhor se ajusta à semivariância, utilizou-se a técnica de validação cruzada (FARACO et al., 2008) e o máximo valor do logaritmo da função verossimilhança (*MLL*). Para investigar a existência de pontos influentes, realizou-se uma análise de diagnóstico por meio da técnica de influência local. Para a análise dos dados, foi utilizado o *software* R (R DEVELOPMENT CORE TEAM, 2009) e o módulo geoestatístico geoR (RIBEIRO JR & DIGGLE, 2009).

RESULTADOS E DISCUSSÃO

Na Tabela 1, são apresentadas as estatísticas descritivas para a variável Produtividade da soja (Prod), Resistência mecânica do solo à penetração nas camadas de 0,0 a 0,10 m, 0,1 a 0,20 m e 0,20 a 0,3 m de profundidade ($RSP_{0,0-0,1}$, $RSP_{0,1-0,2}$ e $RSP_{0,2-0,3}$, respectivamente) e Densidade do solo nas camadas de 0,0 a 0,10 m, 0,10 a 0,20m e 0,20 a 0,30 m de profundidade ($Des_{0-0,1}$, $Des_{0,1-0,2}$ e $Des_{0,2-0,3}$, respectivamente). Verifica-se que o desvio-padrão (DP = 0,38 t ha^{-1}) da produtividade da soja é pequeno comparado com o valor da média, que foi de 3,23 t ha^{-1}, apresentando homogeneidade nos dados da produtividade pelo baixo valor do coeficiente de variação (C.V. < 30%). Observa-se também que existe homogeneidade nos dados de $Des_{0,0-0,1}$, $Des_{0,1-0,2}$, $Des_{0,2-0,3}$, $RSP_{0,0-0,1}$, $RSP_{0,1-0,2}$ e $RSP_{0,2-0,3}$ também pelo baixo valor do coeficiente de variação. Pelo gráfico de pontos espaciais classificados por quartis (*Box-plot*) da produtividade da soja (Figura 1a), observa-se a presença de um ponto discrepante, com valor igual a 2,09 t ha^{-1}, sendo o 13º valor da série de dados localizado na parte inferior esquerda (Figura 1b).

TABELA 1. Estatísticas descritivas para as variáveis Produtividade da soja [t ha^{-1}], Resistência mecânica do solo à penetração [Mpa] e Densidade do solo [mg m^{-3}]. **Descriptive statistics for the variable soybean Productivity [t ha^{-1}], soil Resistance to penetration [Mpa] and soil Density [mg m^{-3}]**.

Variável	n	Média	Mín.	Máx.	Q1	Mediana	Q3	DP	C.V. (%)
Prod	47	3,23	2,09	4,09	3,01	3,19	3,49	0,38	11,79
$RSP_{0,0-0,1}$	47	2,74	1,40	3,69	2,39	2,64	3,14	0,53	19,49
$RSP_{0,1-0,2}$	47	2,44	1,36	3,71	2,14	2,36	2,75	0,49	19,97
$RSP_{0,2-0,3}$	47	2,08	1,34	4,96	1,73	2,05	2,26	0,56	26,88
$Des_{0,0-0,1}$	47	1,09	0,77	1,22	1,05	1,10	1,16	0,08	7,61
$Des_{0,1-0,2}$	47	1,13	0,98	1,55	1,06	1,12	1,17	0,10	8,84
$Des_{0,2-0,3}$	47	1,09	0,95	1,30	1,04	1,08	1,13	0,07	6,59

n: número de dados; Mín.: valor mínimo; Máx.: valor máximo; Q1: primeiro quartil; Q3: terceiro quartil; DP: desvio-padrão; CV: coeficiente de variação.

FIGURA 1. Gráfico *Box-plot* da variável produtividade da soja [t ha^{-1}] (a); Croqui da disposição amostral dos dados na área em estudo, de 57 ha (b). **Box-plot graph of soybean Productivity [t ha^{-1}] (a); Sketch of the sample data layout in the studied area of 57 ha (b)**.

Análises de diagnóstico

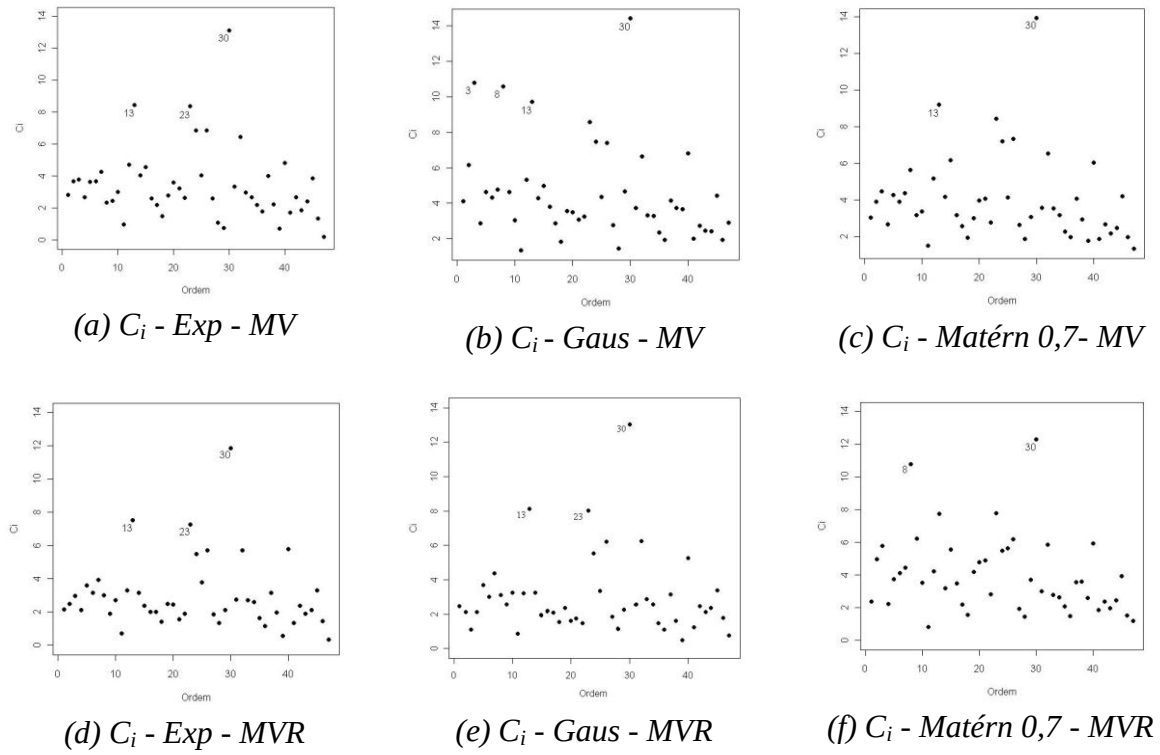
Foram aplicadas técnicas gráficas de diagnóstico de influência local com a finalidade de avaliar, na produtividade da soja, se o ponto discrepante (13º valor) ou alguns outros estariam exercendo algum tipo de influência na estrutura de dependência espacial (afastamento da verossimilhança) e no preditor linear. As técnicas utilizadas foram apresentadas por meio dos gráficos dos coeficientes C_i e l_p obtidos das Equações (8) e (11).

Nas Figuras 2 e 3, são apresentadas as análises de influência local para a variável produtividade, com o total de 47 dados coletados (Prod), utilizando os modelos espaciais ajustados: exponencial, gaussiano e Matérn (com parâmetro *kappa* igual a 0,7) com os métodos de estimação de parâmetros MV e MVR. Na Tabela 2, é apresentado um resumo dos resultados das técnicas de diagnóstico de influência local com as respectivas observações consideradas influentes pelas Figuras 2 e 3, para todas as combinações entre os modelos exponencial, gaussiano e Matérn, e os métodos de estimação de parâmetros MV e MVR.

Pela Tabela 2, verifica-se que a observação número 13, apontada como discrepante pelo gráfico *Box-Plot* (Figura, 1a), foi identificada como influente no afastamento da verossimilhança em todos os casos (técnica C_i), exceto para modelo Matérn (k = 0,7) associado ao método de estimação MVR. Já na avaliação da influência no preditor linear (técnica l_p), ela foi identificada apenas pelo modelo exponencial associado ao método de estimação MV. As observações 3º, 8º, 13º, 23º e 30º foram identificadas com maior frequência na Tabela 2, por isso foram retiradas do conjunto total de dados e novamente realizadas as análises para avaliar suas influências. As observações retiradas da amostra de dados e seus valores são: 3º (3,64 t ha^{-1}), 8º (3,00 t ha^{-1}), 13º (2,09 t ha^{-1}), 23º (3,02 t ha^{-1}) e 30º (2,87 t ha^{-1}). Por simplicidade, definem-se as variáveis: Prod (amostra da produtividade e covariáveis com todas as observações coletadas) e Prod* (amostra da produtividade e covariáveis sem as observações 3º, 8º, 13º, 23º e 30º).

TABELA 2. Resumo das técnicas de influência local e as respectivas observações consideradas influentes. **Summary of the local influence techniques and their observations considered influential**.

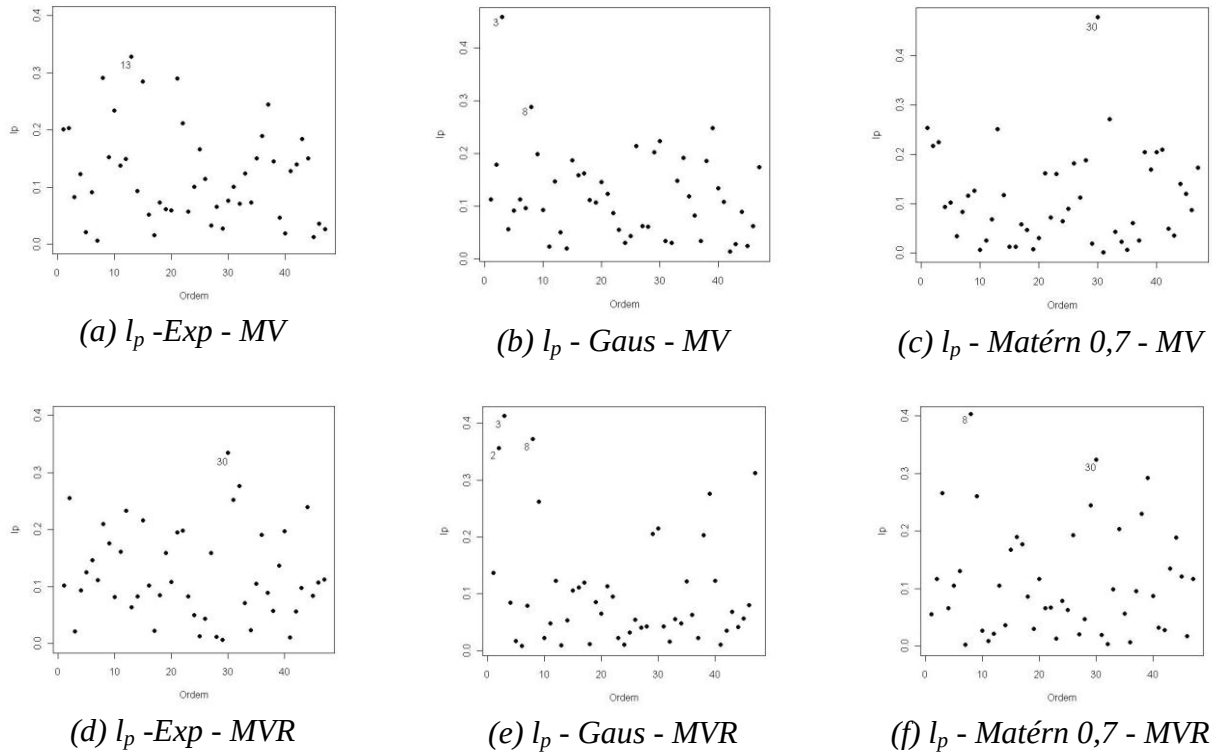
Técnica de Diagnóstico	Método	Observações Influentes		
		Mod. Exponencial	Mod. Gaussiano	Mod. Matérn (k = 0,7)
C_i	MV	13, 23, 30	3, 8, 13, 30	13, 30
	MVR	13, 23, 30	13, 23, 30	8, 30
l_p	MV	13	3, 8	30
	MVR	30	2, 3, 8	8, 30



(a) C_i - Exp - MV *(b) C_i - Gaus - MV* *(c) C_i - Matérn 0,7- MV*

(d) C_i - Exp - MVR *(e) C_i - Gaus - MVR* *(f) C_i - Matérn 0,7 - MVR*

FIGURA 2. Gráficos de diagnóstico C_i para a variável Prod, com os estimadores MV e MVR. **Diagnostic graphs C_i for Prod variable, with the MV and MVR estimators**.



(a) l_p -Exp - MV *(b) l_p - Gaus - MV* *(c) l_p - Matérn 0,7 - MV*

(d) l_p -Exp - MVR *(e) l_p - Gaus - MVR* *(f) l_p - Matérn 0,7 - MVR*

FIGURA 3. Gráficos de diagnóstico l_p, para a variável Prod, com os estimadores MV e MVR. **Diagnostic graphs l_p for Prod variable, with the MV and MVR estimators**.

a) Influência nas medidas descritivas

Pela Tabela 3, observa-se que há pouca variação nas medidas descritivas das variáveis Prod e Prod*. As maiores diferenças são verificadas no valor mínimo (Mín.) e no coeficiente de variação (CV), em que seu valor teve uma redução, passando de 11,79% (Prod) para 10,62% (Prod*), ambos indicando homogeneidade dos dados amostrais.

TABELA 3. Estatísticas descritivas para as variáveis Prod e Prod*, apenas considerando os dados da produtividade. **Descriptive statistics for the variables Prod and Prod*, considering only productivity data**.

Variável	n	Média	Mín.	Máx.	Q1	Med.	Q3	DP	CV (%)
Prod	47	3,23	2,09	4,09	3,01	3,19	3,49	0,38	11,79
Prod*	42	3,26	2,64	4,09	3,03	3,23	3,49	0,35	10,62

n - número de dados; Mín. - valor mínimo; Máx. - valor máximo; Q1 - primeiro quartil; Q3 - terceiro quartil; DP - desvio-padrão; C.V. - coeficiente de variação.

b) Influência nas estimativas dos parâmetros da modelagem espacial da produtividade da soja como função de covariáveis

JOHANN et al. (2004) concluíram que, dentre os atributos físicos dos solos estudados, a resistência mecânica do solo à penetração, na camada de 0,0 a 0,1 m de profundidade, foi a variável que melhor se correlacionou com a produtividade da soja. Assim, da definição de $Z(s_i)$ na eq.(1), e do termo determinístico $\mu(s_i)$ na eq.(2), a média do processo estocástico é representada por uma relação linear com as covariáveis RSP e Des por meio da eq.(12):

$$\mu(s) = \beta_1 + \beta_2 RSP_{0,0\text{-}0,1} + \beta_3 RSP_{0,1\text{-}0,2} + \beta_4 RSP_{0,2\text{-}0,3} + \beta_5 Des_{0,0\text{-}0,1} + \beta_6 Des_{0,1\text{-}0,2} + \beta_7 Des_{0,2\text{-}0,3} \quad (12)$$

em que,

$\beta_1,\ldots, \beta_7$ - parâmetros desconhecidos e a serem estimados.

TABELA 4. Parâmetros β estimados por MV e MVR para as variáveis Prod e Prod*, utilizando os modelos exponencial (Exp), gaussiano (Gaus) e da família Matérn com k = 0,7. **Estimated parameters β by MV and MVR for variables Prod and Prod*, using the models exponential (Exp), Gaussian (Gaus) and Matérn family with k = 0.7**.

Modelo	Método	Variável	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Exp.	MV	Prod	4,564	-0,103	0,063	-0,015	-0,289	0,061	-0,866
			(1,074)	(0,115)	(0,123)	(0,092)	(0,639)	(0,522)	(0,800)
		Prod*	3,266	0,055	-0,203	0,297	-0,172	0,508	-0,613
			(1,088)	(0,120)	(0,157)	(0,179)	(0,577)	(0,620)	(0,751)
	MVR	Prod	4,748	-0,113	0,052	-0,018	-0,295	-0,002	-0,917
			(1,168)	(0,124)	(0,134)	(0,099)	(0,689)	(0,559)	(0,859)
		Prod*	3,238	0,050	-0,207	0,290	-0,139	0,487	-0,609
			(1,207)	(0,129)	(0,169)	(0,194)	(0,624)	(0,669)	(0,809)
Gaus.	MV	Prod	4,449	-0,097	0,061	-0,018	-0,263	0,114	-0,847
			(1,066)	(0,114)	(0,122)	(0,090)	(0,636)	(0,520)	(0,788)
		Prod*	3,300	0,046	-0,211	0,287	-0,145	0,463	-0,576
			(1,069)	(0,117)	(0,154)	(0,176)	(0,565)	(0,600)	(0,733)
	MVR	Prod	4,513	-0,1011	0,054	-0,018	-0,256	0,096	-0,868
			(1,154)	(0,123)	(0,133)	(0,098)	(0,689)	(0,561)	(0,853)
		Prod*	3,294	0,040	-0,211	0,283	-0,117	0,431	-0,567
			(1,175)	(0,127)	(0,167)	(0,192)	(0,613)	(0,649)	(0,795)
Matérn k=0,7	MV	Prod	4,562	-0,103	0,062	-0,015	-0,286	0,063	-0,866
			(1,073)	(0,115)	(0,123)	(0,092)	(0,639)	(0,522)	(0,798)
		Prod*	3,267	0,054	-0,204	0,295	-0,167	0,501	-0,607
			(1,086)	(0,119)	(0,156)	(0,178)	(0,575)	(0,617)	(0,748)
	MVR	Prod	4,792	-0,117	0,051	-0,018	-0,298	-0,016	-0,928
			(1,169)	(0,124)	(0,134)	(0,100)	(0,690)	(0,558)	(0,859)
		Prod*	3,239	0,048	-0,207	0,288	-0,134	0,478	-0,602
			(1,203)	(0,128)	(0,169)	(0,193)	(0,621)	(0,666)	(0,806)

Entre parênteses é apresentado o desvio padrão de cada parâmetro estimado.

Nas Tabelas 4 e 5, são apresentados os resultados das análises de variabilidade espacial da produtividade da soja, por meio da geoestatística, para os dados originais (Prod) e retirando-se os pontos considerados influentes (Prod*).

Pelos resultados da Tabela 4, pode-se observar, para todos os modelos ajustados, quando retiradas as observações 3^o, 8^o, 13^o, 23^o e 30^o (Prod*), influenciaram na estimação dos parâmetros de β_2, β_4, β_5, β_6 e β_7, pois, quando estas foram retiradas (Prod*), os valores das estimativas dos parâmetros aumentaram e o valor de β_3 diminuiu, com relação aos obtidos para Prod.

Observando a Tabela 5, verifica-se que, quando o parâmetro φ_1 (efeito pepita) foi estimado sem as observações influentes (Prod*), em todos os modelos ajustados, houve redução no seu valor, quando comparado com os obtidos para a variável Prod. Um comportamento exatamente ao contrário verifica-se para o parâmetro φ_3 (função do alcance), mostrando que, na estimação de parâmetros, houve uma relação inversa entre φ_1 e φ_3 quando as observações influentes foram retiradas. Quanto à φ_2 (contribuição), dentro de todos os modelos, identifica-se o seguinte comportamento: quando a estimativa é feita por MV, o valor de φ_2 para Prod* é menor do que o obtido para Prod; já por MVR, o valor de φ_2 para Prod* é sempre maior do que o obtido para Prod.

As estimativas para o parâmetro φ_3 para Prod e Prod* apresentam as maiores diferenças, tanto pelo método de MV quanto pelo método de MVR.

TABELA 5. Parâmetros φ estimados por MV e MVR para as variáveis Prod e Prod*, utilizando os modelos exponencial (Exp), gaussiano (Gaus) e família Matérn com k = 0,7. **Estimated parameters φ by MV and MVR for variables Prod and Prod*, using the models exponential (Exp), Gaussian (Gaus) and Matérn family with k = 0.7**.

Modelo	Método	Variável	φ_1	φ_2	φ_3
Exp	MV	Prod	0,079	0,046	142,260
			(0,032)	(0,040)	(1,733)
		Prod*	0,064	0,033	225,410
			(0,022)	(0,031)	(4,782)
	MVR	Prod	0,106	0,065	475,002
			(0,030)	(0,057)	(16,625)
		Prod*	0,080	0,085	825,756
			(0,023)	(0,073)	(55,636)
Gaus	MV	Prod	0,091	0,033	192,354
			(0,024)	(0,027)	(0,509)
		Prod*	0,071	0,029	323,215
			(0,017)	(0,025)	(2,518)
	MVR	Prod	0,108	0,040	220,350
			(0,027)	(0,032)	(0,724)
		Prod*	0,085	0,049	426,438
			(0,020)	(0,041)	(5,286)
Matérn k=0,7	MV	Prod	0,084	0,040	126,733
			(0,071)	(0,044)	(156,500)
		Prod*	0,067	0,031	203,126
			(0,046)	(0,029)	(256,024)
	MVR	Prod	0,111	0,065	430,003
			(0,073)	(0,058)	(519,454)
		Prod*	0,082	0,079	574,95 1
			(0,078)	(0,075)	(690,141)

Entre parênteses, são apresentados o desvio-padrão de cada parâmetro estimado. Alcance a: modelo exponencial: $a = 3\varphi_3$, modelo gaussiano: $a = (3)^{1/2}\varphi_3$; modelo Matérn: $a = \varphi_3$.

c) Influência na validação dos modelos

Na Tabela 6, são apresentados os resultados da validação de modelos ajustados para as variáveis Prod e Prod*, utilizando o critério da validação cruzada (FARACO et al., 2008) e o valor máximo do logaritmo da função verossimilhança (*MLL*).

Para a variável Prod, foi escolhido como melhor ajuste o modelo Matérn (k=0,7) com parâmetros estimados por MVR; e quando foram retiradas as observações consideradas influentes pelas técnicas de influência local, para a variável Prod*, foi escolhido como melhor ajuste o modelo gaussiano, com parâmetros estimados por MVR, utilizando os critérios da validação cruzada, e tendo o indicador *MLL* entre os valores máximos (Tabela 6). Isto mostra que as observações 3º, 8º, 13º, 23º e 30º são influentes também na escolha do melhor modelo espacial ajustado.

TABELA 6. Resultados da validação cruzada e máximo valor do logaritmo da função verossimilhança - *MLL* para as variáveis Prod e Prod*. **Cross validation results and maximum value of the log likelihood function - *MLL* to variables Prod and Prod***.

Variável	Modelo	Método	*EM*	$\overline{ER}$	DP_e	S_{ER}	*EA*	*MLL*
Prod	Exp	MV	-0,00546	-0,00388	0,39311	1,08127	14,29101	-16,180
		MVR	-0,00415	-0,00252	0,38967	1,00024	14,28861	-16,420
	Gaus	MV	-0,00531	-0,00384	0,38918	1,08186	14,05426	-15,860
		MVR	-0,00497	-0,00325	0,38861	0,99977	14,13366	-16,360
	Matérn k=0,7	MV	-0,00541	-0,00385	0,39253	1,08123	14,26218	-16,140
		MVR	-0,00394	-0,00237	0,38915	1,00038	14,29727	-16,390
Prod*	Exp	MV	0,00374	0,00582	0,33281	1,07573	10,67954	-8,655
		MVR	0,00315	0,00452	0,32929	0,98539	10,50028	-9,626
	Gaus	MV	0,00346	0,00545	0,32387	1,07301	10,26349	-8,050
		MVR	0,00300	0,00430	0,32284	0,98373	10,13232	-9,209
	Matérn k=0,7	MV	0,00367	0,00572	0,33138	1,07522	10,61170	-8,575
		MVR	0,00311	0,00446	0,32783	0,98497	10,43333	-9,544

$EM = \frac{1}{n}\sum_{i=1}^{n}\left(Z(s_i) - \hat{Z}(s_{(i)})\right)$; $\overline{ER} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z(s_i) - \hat{Z}(s_{(i)})}{\sigma(s_{(i)})}$; $S_{ER} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\left|Z(s_i) - \hat{Z}(s_{(i)})\right|}{\sigma(s_{(i)})}}$; $EA = \sum_{i=1}^{n}\left|\hat{Z}(s_i) - Z(s_{(i)})\right|$; DP_e: desvio padrão do erro médio.

d) Influência na construção dos mapas temáticos

Na Figura 4, apresentam-se os mapas temáticos das variáveis Prod (Figura 4a) e Prod* (Figura 4b), confeccionados por meio da interpolação por krigagem universal, visto que, neste estudo, foram introduzidas covariáveis. Os mapas foram construídos utilizando-se dos modelos indicados pelo critério da validação cruzada, que é o modelo Matérn com k= 0,7 para Prod e o modelo gaussiano para Prod* (Tabela 6).

Pela Figura 4, observa-se que há grande alteração na distribuição da produtividade da área em estudo, principalmente nas localizações das sub-regiões com produtividade nos intervalos de 3,04 a 3,23 t ha^{-1} e de 3,23 a 3,43 t ha^{-1}.

Estas constatações indicam que as observações 3º, 8º, 13º, 23º e 30º estão exercendo grande influência também na construção dos mapas temáticos, confirmando o que a medida de diagnóstico no preditor linear (gráficos l_p) já havia indicado (Figura 3).

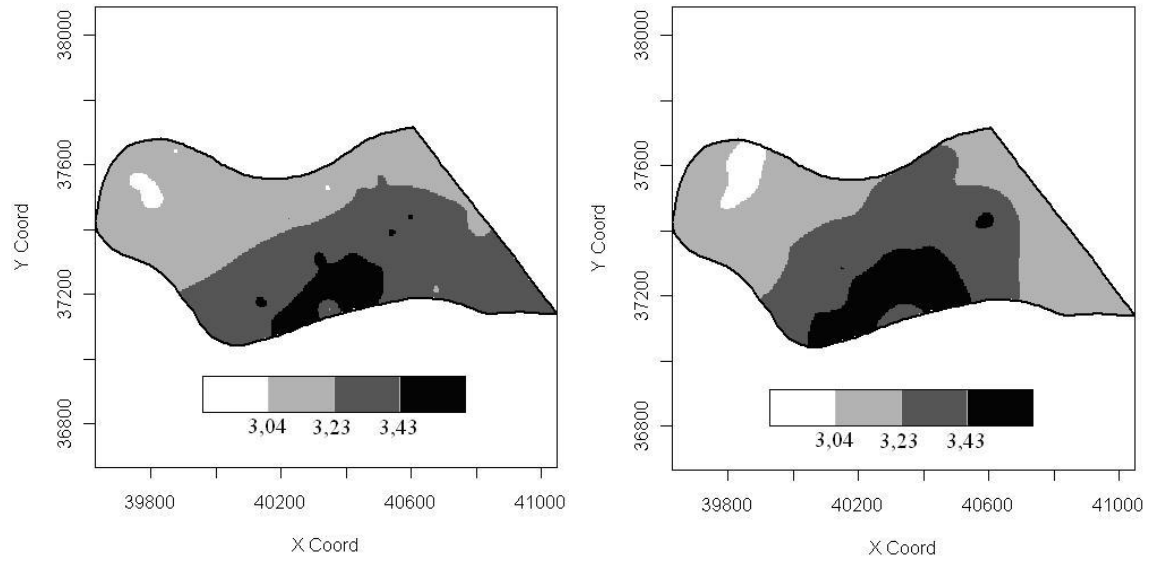


FIGURA 4. Mapas temáticos para a variável produtividade: Prod (a); Prod* (b). **Thematic maps to the yield variable: Prod (a); Prod* (b)**.

CONCLUSÕES

O estudo mostrou que a presença de valores atípicos entre dados amostrados pode exercer forte influência nos mapas temáticos, alterando, assim, a dependência espacial. Assim, o mapa da distribuição espacial da produtividade que seria apresentado ao agricultor, não estaria representando de forma correta as tendências da produtividade na área em estudo. Isso acarretaria em interpretações errôneas no gerenciamento da área quanto às futuras intervenções no trato do solo de forma localizada.

Conclui-se, também, que nem sempre um valor discrepante é influente, ao passo que um valor qualquer, mesmo não discrepante, pode ser considerado influente. Isto é, um valor discrepante não deve ser eliminado da amostra antes de ser realizada uma análise de diagnóstico de influência local.

A aplicação de técnicas de diagnóstico de influência local deve fazer parte de toda análise geoestatística, garantindo que as informações contidas nos mapas temáticos tenham maior qualidade e possam ser utilizadas com maior segurança pelo agricultor.

AGRADECIMENTOS

Ao CNPq, CONAB e Fundação Araucária, pelo apoio financeiro, e à UNIOESTE - Universidade Estadual do Oeste do Paraná**.**

REFERÊNCIAS

CHRISTENSEN, R.; JOHNSON, W.; PEARSON, L. Covariance function diagnostics for spatial linear models. *International Association for Mathematical Geology*, New York, v.25, n.2, p.145-160, 1993.

BORSSOI, J.A.; URIBE-OPAZO, M.A.; GALEA, M. Diagnostic techniques applied in geostatistics for agricultural data analysis. *Revista Brasileira de Ciência do Solo*, Viçosa-MG, v.33, n.6, p. 1.561-1.570, 2009.

COOK, R.D. Assessment of local influence (with discussion). *Journal of the Royal Statistical Society***,** Series B, London, v.48, n.2, p.133-169, 1986.

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

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. *Revista Brasileira de Ciência do Solo*, Viçosa-MG, v.32, n.2, p.463-476, 2008.

JOHANN, J.A.; URIBE-OPAZO, M.A.; SOUZA, E.G.; ROCHA, J.V. Variabilidade espacial dos atributos físicos do solo e da produtividade em uma área experimental de agricultura de precisão com e sem manejo localizado. *Revista Brasileira de Engenharia Agrícola e Ambiental***,** Campina Grande, v.8, n.2, p.1-8. 2004.

MARDIA, K.; MARSHALL, R. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, Alemanha, v.71, p.135-146, 1984.

R DEVELOPMENT CORE TEAM R: *A language and environment for statistical computing. R Foundation for statistical Computing*, Vienna, Austria. ISBN 3-90005107-0. Disponível em: <http://www.R-project.org>. Acesso em: 20 jul. 2009

RIBEIRO JR., P.J.; DIGGLE P. J. geoR: A package for geostatistical analysis. R-NEWS, 01. Disponível em: <http://cran.r-project.org/doc/Rnews. 2001>. Acesso em: 3 jun. 2009.