ea engenharia agrícola
Journal of the Brazilian Association of Agricultural Engineering
ISSN: 1809-4430 (on-line)

TÉCNICAS PARA DETECÇÃO DE PONTOS INFLUENTES EM VARIÁVEIS CONTÍNUAS REGIONALIZADAS

Doi:http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v36n1p152-165/2016

JONATHAN RICHETTI[1], MIGUEL A. URIBE-OPAZO[2], FERNANDA DE BASTIANI[3], JERRY ADRIANI JOHANN[4]

RESUMO: Na análise de dados espaciais em agricultura, a presença de pontos influentes pode alterar consideravelmente os resultados das análises de dependência espacial e, consequentemente, a construção dos mapas. Quando se referem a atributos físico-químicos do solo e da produtividade, os mapas devem representar uma estimativa eficiente das condições reais do campo, já que são importantes informações utilizadas para a manutenção de um sistema agrícola de manejo localizado, com a otimização da aplicação de insumos agrícolas, visando à maior produtividade. Este trabalho teve por objetivo apresentar as técnicas gráficas *hair-plot*, de influência local (C_i e $|L_{max}|$) de identificação de observações influentes em dados contínuos espaciais georreferenciados, coletados em uma área experimental de cultivo comercial, com 167,35 hectares, onde o sistema agrícola de manejo localizado é adotado. Como resultados apresentam-se os pontos potencialmente influentes e os mapas construídos com e sem eles. Na comparação entre os mapas com e sem estes pontos, as métricas de comparação dos mapas mostraram a importância da identificação dos pontos influentes em uma base de dados espaciais. Sendo assim, a existência de pontos influentes deve ser investigada para entender o motivo de seu comportamento atípico, já que eles modificam, consideravelmente, os mapas gerados.

PALAVRAS-CHAVE: geoestatística, *hair-plot*, influência local, máxima verossimilhança, atributos químicos.

TECHNIQUES FOR DETECTION OF INFLUENCING POINTS IN REGIONALIZED CONTINUOUS VARIABLES

ABSTRACT: Influencing points in agricultural spatial analysis may change considerably results on spatial dependence and hence map building. With regards to physico-chemical soil properties and crop yield, such maps should efficiently estimate current field conditions, being important for an agricultural site-specific management, optimizing thus input applications in order to increase yields. This study aimed to analyze hair-plot graphic techniques, with local influence (Ci and $|L_{max}|$) to identify influencing points within a set of georeferenced spatial continuous data. These information were gathered from an experimental area with 167.35 hectares, wherein an agricultural site-specific management has been adopted. As a result, we obtained potentially influencing points and then outlined maps with and without the use of them. By comparing both maps, we could note by metric comparison that it is of major importance to identify those points on a spatial database. Thus, such investigations must be carried out to understand cases of unusual performance, since they considerably modify the generated maps.

KEYWORDS: geostatistics, hair-plot, local influence, maximum likelihood, chemical soil properties.

[1] Engº Agrícola, Ms. em Engenharia Agrícola, Doutorando do Programa de pós-graduação em Engenharia Agrícola - PGEAGRI, UNIOESTE, Cascavel-PR, Fone: (45) 3220-7320, j_richetti@hotmail.com
[2] Estatístico, Dr. em Estatística, Prof. Associado do PGEAGRI, UNIOESTE, Cascavel-PR, Fone: (45) 3220-3228, Pesquisador de Produtividade do CNPq, miguel.opazo@unioeste.br
[3] Matemática, Dr. em Estatística, UFPE, Recife-PE, Fone: (45) 3220-3228, fernandadebastiani@gmail.com
[4] Engº Agrícola, Dr. em Engenharia Agrícola, Prof. Adjunto do PGEAGRI, UNIOESTE, Cascavel-PR, Fone: (45) 3220-7320, Pesquisador de Produtividade da Fundação Araucária, jerry.johann@hotmail.com
Recebido pelo Conselho Editorial em: 27-3-2014
Aprovado pelo Conselho Editorial em: 01-12-2015

INTRODUÇÃO

Na modelagem da dependência espacial de variáveis contínuas regionalizadas, existem diferentes procedimentos estatísticos que são utilizados para a estimação dos parâmetros que definem a estrutura de dependência espacial e, que são empregados na interpolação espacial pela técnica de krigagem, gerando mapas que representam as estimativas das condições do solo e a produtividade. Além disso, a qualidade destes mapas depende das técnicas de ajuste dos modelos utilizados. No entanto, para que a interpolação produza estimações confiáveis e represente a real variabilidade local, a modelagem deve ser feita com muita cautela, principalmente na presença de *outliers* e de pontos influentes.

Outliers em dados espaciais podem ocorrer com características distintas, dependendo dos pontos próximos, determinados pelo raio de dependência espacial, ou ainda pelo modelo que define a estrutura de dependência espacial. Técnicas de mineração de dados não eram capazes de detectar *outliers* espaciais; contudo, com sua definição atual, eles servem para identificar observações extremas globais como *outliers* espaciais (SHEKHAR et al., 2003). Tais pontos representam uma instabilidade local, pois as observações são extremas em relação aos seus vizinhos, mesmo quando estes não aparentam diferença da população em geral.

É possível que uma única observação possua grande influência nos resultados de uma análise de dados espaciais, isto é, pode alterar consideravelmente os resultados que definem a estrutura de dependência espacial, e consequentemente alterar a construção dos mapas que a representam. É, pois, importante que se tenha cuidado com a possibilidade de observações influentes, e estas devem ser consideradas na interpretação dos resultados.

No estudo de *outliers*, CRESSIE & HAWKINS (1980) apresentaram um estimador da função semivariância, denominado robusto, para o estudo de dependência espacial, quando se tem a presença de *outliers*. Porém, GENTON (1998) sugeriu que este estimador da função semivariância tem limitações na robustez, do ponto de vista estatístico. Um *outlier* pode ser influente na escolha do modelo a ser ajustado à função semivariância, na estimatição dos parâmetros e na construção de mapas. No estudo de pontos influentes, existem aqueles que não são *outliers*, podendo ser detectados por meio da análise de diagnóstico dos dados.

Na literatura, existem diferentes técnicas de diagnóstico espacial para detectar pontos influentes. MILITINO et al. (2004) estudaram métodos de diagnóstico de influência global baseados na eliminação de casos em modelos lineares espaciais. URIBE-OPAZO et al. (2012) desenvolveram estudos de diagnóstico de influência local para modelos espaciais lineares gaussianos e DE BASTIANI et al. (2015) generalizaram resultados desenvolvendo estudos de diagnóstico de influência local para modelos espaciais lineares da família contornos elípticos. GENTON & RUIZ-GAZEN (2010) introduziram o gráfico *hair-plot*, que consiste em todas as trajetórias do valor de um estimador quando cada ponto é modificado por uma perturbação aditiva.

O objetivo deste trabalho foi aplicar, em conjunto, a técnica *hair-plot* e as técnicas de diagnóstico de influência local em modelos espaciais lineares, considerando a estrutura na matriz de covariância os modelos exponencial, Gaussiano e família Matérn com diferentes parâmetros de forma, em atributos químicos do solo e da produtividade da soja, para detectar a existência de pontos que possam exercer algum tipo de influência na análise da variabilidade espacial e, consequentemente, na construção dos mapas.

MATERIAL E MÉTODOS

Área Experimental e Dados Analisados

A coleta de dados foi realizada em uma área comercial de produção de grãos de 167,35 ha, na região oeste do Paraná, especificamente na microrregião de Cascavel, cuja localização geográfica é, aproximadamente, latitude de 24,95º S, longitude de 53,57º O, e altitude média de 650 m (Figura 1). O solo desta região é classificado como Latossolo Vermelho Distroférrico, com textura argilosa (EMBRAPA, 2013). O clima da região é classificado como temperado mesotérmico e superúmido,

tipo climático Cfa (Koeppen), com temperatura anual média e de 21 ºC. Os dados foram coletados no ano agrícola de 2010/2011. Foi realizada uma amostragem sistemática centrada com pares de pontos próximos (*lattice plus close pairs*), com distância máxima de 141 m entre pontos, e em alguns locais escolhidos de forma aleatória, a amostragem foi realizada com distâncias de 75 m (ex.: ponto 27) e 50 m (ex.: ponto 26) entre pontos, totalizando 102 pontos amostrais na área em estudo (Figura 1). Todas as amostras foram georreferenciadas e localizadas com auxílio de um aparelho receptor de sinal com o sistema de posicionamento global (GPS), da marca Trimble, modelo GeoExplorer III, com precisão de 2 a 5 metros.

A amostragem sistemática foi feita sobre os nós de uma malha regular definida. AUNE-LUNDBERG & STRAND (2014) mostraram vantagens em se usar amostragem sistemática aleatória ao invés de amostragem aleatória simples, e esta melhora estava associada à presença da correlação espacial. O estudo também mostrou que estimadores de variância local são superiores à variância estimada, quando empregada a amostragem sistemática em comparação à amostragem aleatória.

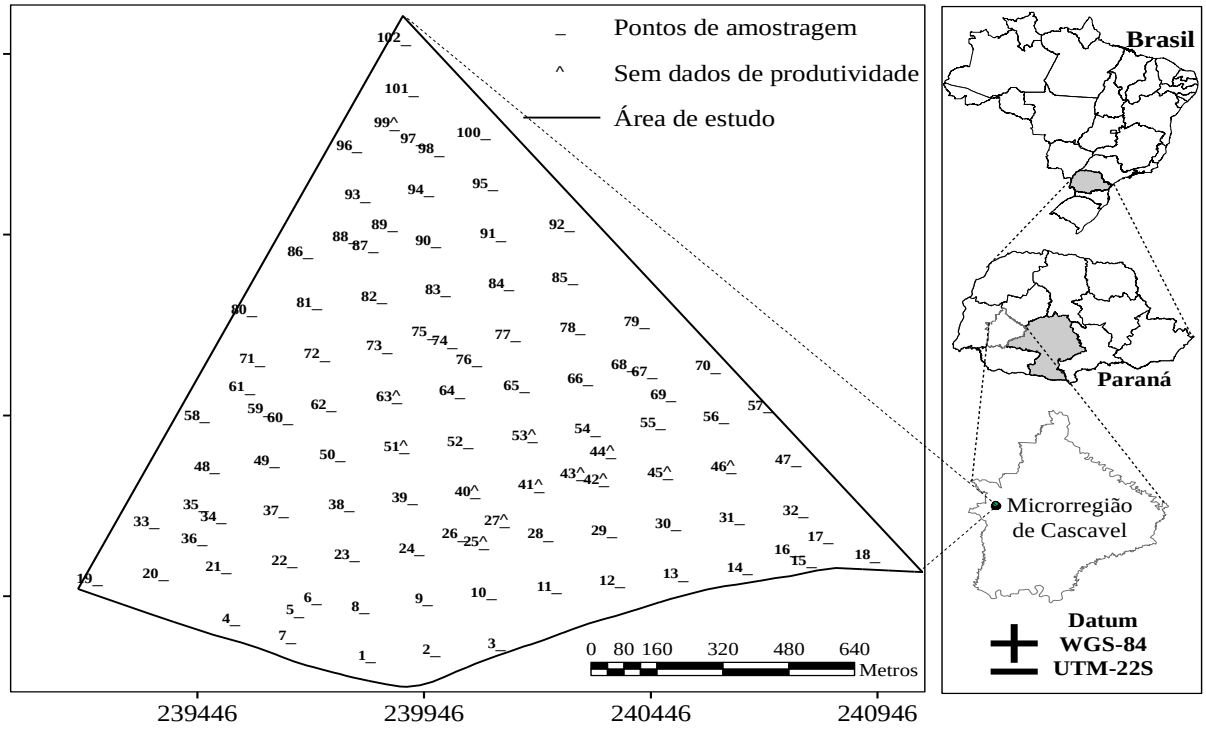

FIGURA 1. Mapa de localização da área de estudo. **Study area location.**

A amostragem do solo para a determinação dos atributos químicos foi realizada em todos os pontos amostrais (102). Em cada ponto amostral, foram coletadas quatro subamostras de solo, de 0,0 a 0,2 m de profundidade, que depois de misturadas continham aproximadamente 500 g de solo, compondo, assim, a amostra representativa de cada ponto amostral. Entretanto, para este estudo, foram utilizados os atributos químicos teor de fósforo-P (mg dm^{-3}) e teor de cálcio-Ca ($cmol_c$ dm^{-3}) por apresentarem pontos *outliers* na área em estudo.

A produtividade da soja (Prod) foi coletada em 89 pontos amostrais, pois 13 pontos amostrais (destacadas na Figura 1) estavam sobre áreas experimentais cultivadas com milho. Para a determinação da produtividade da soja, a área de 1,0 m² foi colhida em cada ponto amostral. Após ajustar a umidade dos grãos para 13%, fez-se a conversão em t ha^{-1} para cada ponto amostral.

Análises Estatísticas – *Hair-plot*

Seja $\mathbf{Z} = (Z_1,...,Z_n)^T$ é uma amostra espacialmente georreferenciada de dimensão *n*, então $Z_i = Z(s_i)$ representa uma observação Z no local $s_i \in S$, com $S \subset \Re^2$, sendo $\Re^2$ espaço euclidiano, bi-dimensional. O *hair-plot* é um gráfico proposto por GENTON & RUIZ-GAZEN

(2010), que é utilizado como um método para detectar pontos influentes, considerando suas dependências espaciais. O *hair-plot* é formado por linhas que possuem um ponto central para o qual as linhas convergem. Para a construção deste gráfico, os autores propõem que, associado ao *hair-plot*, existam duas medidas de interesse. Uma medida baseada na influência local de cada ponto, na estimação de uma estatística $\tau_i(\hat{\theta}, \mathbf{Z})$, para $i = 1,...,n$, sendo $\hat{\theta}$ o estimador do parâmetro de dependência espacial, e a outra medida baseada na influência anisotrópica de um ponto na estatística $v_i(\hat{\theta}, \mathbf{Z})$, para $i = 1,...,n$, definidas por GENTON & RUIZ-GAZEN (2010).

Para os casos em que a dependência espacial é tratada, o parâmetro (θ), que mede a dependência espacial, foi obtido por meio do índice de Moran $I(\mathbf{Z})$ (MORAN, 1950), definido na [eq. (1)]:

$$I(\mathbf{Z}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\mathbf{Z_i} - \bar{\mathbf{Z}})(\mathbf{Z_j} - \bar{\mathbf{Z}})}{\sum_{i=1}^{n}(\mathbf{Z_i} - \bar{\mathbf{Z}})^2}, \quad (1)$$

em que, w_{ij} representa um elemento *ij* da matriz de proximidade espacial $\boldsymbol{W}$, *n* x *n*, e possui valores não negativos que descrevem os pesos entre os vizinhos no plano, a matriz $\boldsymbol{W}$ pode ter diferentes esquemas de proximidade (ANSELIN, 2002). Para este estudo considerou-se $w_{ij} = 1/(1+D_{ij})$, sendo D_{ij} a distância entre os pontos Z_i e Z_j.

Assim, GENTON & RUIZ-GAZEN (2010) reescrevem $v_i(I, \mathbf{Z})$, como definido na [eq. (2)], que indica a influência do ponto *i* nos dados.

$$v_i(I, \mathbf{Z}) = \frac{1}{n-1}\left(\frac{w..}{n - w_{.i} - w_{i.}}\right), \quad (2)$$

em que,

$$w.. = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij};\ w_{.i} = \sum_{j=1}^{n} w_{ji}, \text{ e } w_{i.} = \sum_{j=1}^{n} w_{ij}.$$

Análises Estatísticas – Influência

O estudo de *outliers* e de detecção de pontos influentes é um importante passo em análise de conjuntos de dados espaciais. A eliminação de pontos é um método de diagnóstico que consiste em avaliar o impacto da retirada de uma observação nas estimativas dos parâmetros do modelo. No estudo de pontos influentes, a literatura apresenta em destaque as metodologias de diagnósticos em influência global propostas por COOK (1977), que é baseada na eliminação de pontos do conjunto total de dados. Uma medida análoga a distância de Cook foi apresentada por URIBE-OPAZO et al. (2012), chamada de influência global espacial sobre o estimador de máxima verossimilhança. O método de influência local proposto por COOK (1986) avalia o efeito simultâneo de observações sobre os estimadores de máxima verossimilhança (ML), sem a necessidade de sua eliminação do conjunto de dados, como é realizado no estudo de influência global. DE BASTIANI et al. (2015) estudaram a influência local em modelos espaciais lineares de contorno elípticos, estacionários e isotrópicos (WEBSTER & OLIVER, 2007; GUEDES et al., 2013) e obtiveram expressões explícitas para implementar as técnicas gráficas C_i e $|L_{max}|$, utilizando as funções de covariância exponencial, Gaussiano e família Matérn (com parâmetros de forma k = 0,7; 1,0 e 1,5).

A estimação do vetor de parâmetros $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_3)^T$, que definem a estrutura de dependência espacial $\boldsymbol{\Sigma} = \varphi_1 \boldsymbol{I}_n + \varphi_2 \boldsymbol{R}(\varphi_3)$, em que, φ_1 é o efeito pepita ou erro de variância; $\boldsymbol{I}_n$ é uma matriz identidade *n* x *n* ; φ_2 é a contribuição ou variância de dispersão (*sill*); φ_3 é uma função do alcance (*a*) do modelo; $\boldsymbol{R}(\varphi_3)$ é uma matriz *n* x *n*, que é função de φ_3. (MARDIA & MARSHALL, 1984), foi

realizada utilizando o método de ML (LARK, 2002), com auxílio do algoritmo L-BFGS-B (BYRD et al., 1995).

A partir dos modelos ajustados, foram construídos mapas da variabilidade espacial das variáveis regionalizadas da área em estudo, utilizando a técnica de interpolação por krigagem numa grade de 5 x 5 metros. Para comparar os mapas gerados construídos com e sem os pontos influentes, utilizaram-se medidas de acurácia: exatidão global (EG), Kappa (Ka) e Tau (T) (quanto mais próximo de um, maior acurácia existe entre os mapas) (FOODY, 2010), obtidas a partir da construção da matriz de erros (DE BASTIANI et al., 2012). Além disso, para a escolha de cada modelo, os critérios de seleção utilizados foram os de validação cruzada e máximo valor do logaritmo da função verossimilhança-*MLL* (FARACO et al., 2008), de forma que os dados (teor de fósforo, teor de cálcio e produtividade da soja) foram krigados de acordo com o modelo com melhor ajuste.

Dentre as observações não detectadas como potencialmente influentes, foram selecionados pontos aleatoriamente, com o objetivo de validar os resultados. Foi avaliado se a retirada de algum ponto apresenta variação na estimação da estrutura de dependência espacial.

Para a análise dos dados, foi utilizado o software R (R DEVELOPMENT CORE TEAM, 2013), versão 2.12.2 e o módulo geoR (RIBEIRO JÚNIOR & DIGGLE, 2001) e a rotina do *hair-plot* cedida por GENTON & RUIZ-GAZEN (2010).

RESULTADOS E DISCUSSÃO

A Figura 2 apresenta o *box-plot*, *hair-plot* e os gráficos de influência local C_i e $|L_{max}|$ para detecção de pontos influentes na variável teor de fósforo. De acordo com o *box-plot*, os pontos 44 e 46 são considerados *outliers* superiores. A interpretação do *hair-plot*, segundo GENTON & RUIZ-GAZEN (2010), é que cada linha descreve o efeito no estimador para dada observação. Assim, a linha possui uma grande inclinação em zero ou comporta-se diferentemente das outras linhas e corresponderá a uma observação com maior influência, significando que o *hair-plot* é uma ferramenta descritiva, requerendo interpretação.

Assim, observa-se na Figura 2 que o *hair-plot* indica que a observação 46 é influente, já que a linha que representa o ponto amostral 46 começa movendo-se levemente das outras, apresentando uma inclinação e cruzando o grupo de linhas vizinhas, que representam os outros pontos amostrais, ou seja, a linha 46 apresenta um comportamento completamente diferente das demais, cruzando-as, afastando-se, fazendo uma curva mais acentuada e apresentando um pico superior de autocorrelação, quando a variação de ζ é negativa. De acordo com os gráficos de influência local C_i e $|L_{max}|$, o ponto 46 também é indicado como ponto influente.

Para analisar a influência espacial do ponto 46, detectado pelas técnicas hair-plot, C_i e $|L_{max}|$, construíram-se dois mapas para o teor de fósforo (P): o primeiro mapa (Figura 3a) foi construído com todos os pontos amostrais, e o segundo mapa (Figura 3b) foi construído sem o ponto 46, considerado influente. Para a construção dos mapas de P (mg dm^{-3}), segundo os critérios de seleção de validação cruzada e máximo valor do logaritmo da função verossimilhança-*MLL*, o modelo Gaussiano foi o que apresentou melhores ajustes. Os parâmetros espaciais estimados foram $\hat{\varphi}_1$ = 13,159, com desvio-padrão DP = 8,5384; $\hat{\varphi}_2$ = 59,256, com DP = 9,6422, e $\hat{\varphi}_3$ = 947,744 com DP = 9,6233, sendo o raio de dependência espacial (alcance) de 1640,38 m para todos os pontos; e $\hat{\varphi}_1$ = 20,953, com DP = 3,8787; $\hat{\varphi}_2$ = 7,069, com DP = 4,1536, e $\hat{\varphi}_3$ = 199,956, com DP = 0,110, com raio de dependência espacial (alcance) de 346,09 m para os dados, sem o ponto influente. Observam-se diferenças nas estimações dos parâmetros que definem a estrutura de dependência espacial com e sem o ponto influente, em especial na variância estimada, denominada patamar ($\hat{\varphi}_1 + \hat{\varphi}_2$), e nos valores estimados dos parâmetros que definem os raios de dependência espacial ($\hat{\varphi}_3$).

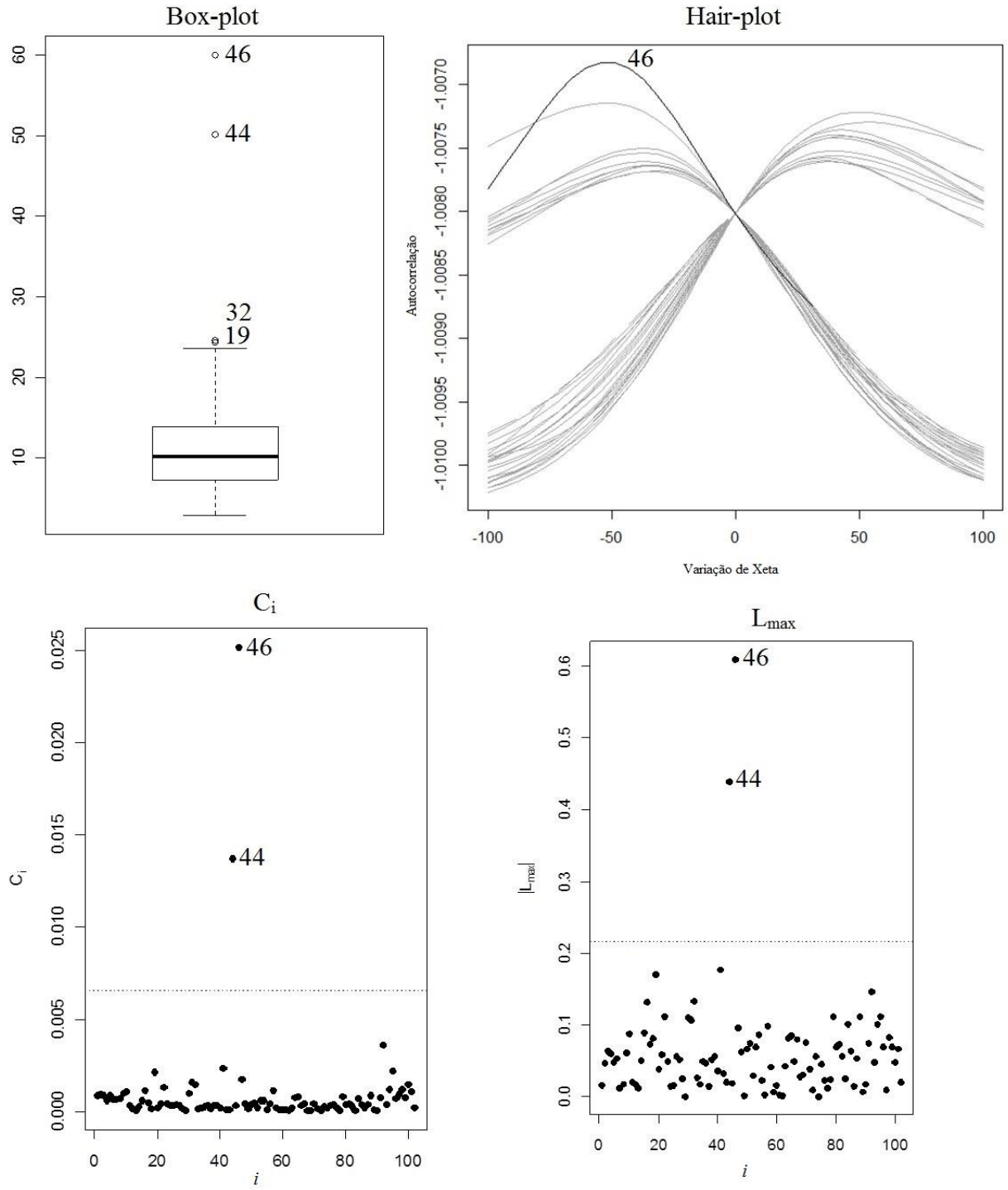

FIGURA 2. Box-plot, *hair-plot*, gráficos C_i e $|L_{max}|$ para identificação de *outliers* e pontos influentes para o teor de fósforo (mg dm^{-3}). **Box-plot, hair-plot, C_i and $|L_{max}|$ graphs generated to identify outliers and influent points for phosphorus content (mg dm^{-3})**.

Pelos mapas (Figura 3), observa-se que os níveis do teor de fósforo (mg dm^{-3}) são mais altos na região norte da área e vão decrescendo na direção norte-sul. Os mapas mostraram diferenças espaciais do teor de fósforo, especialmente na região sul da área onde foi detectado o ponto influente**.**

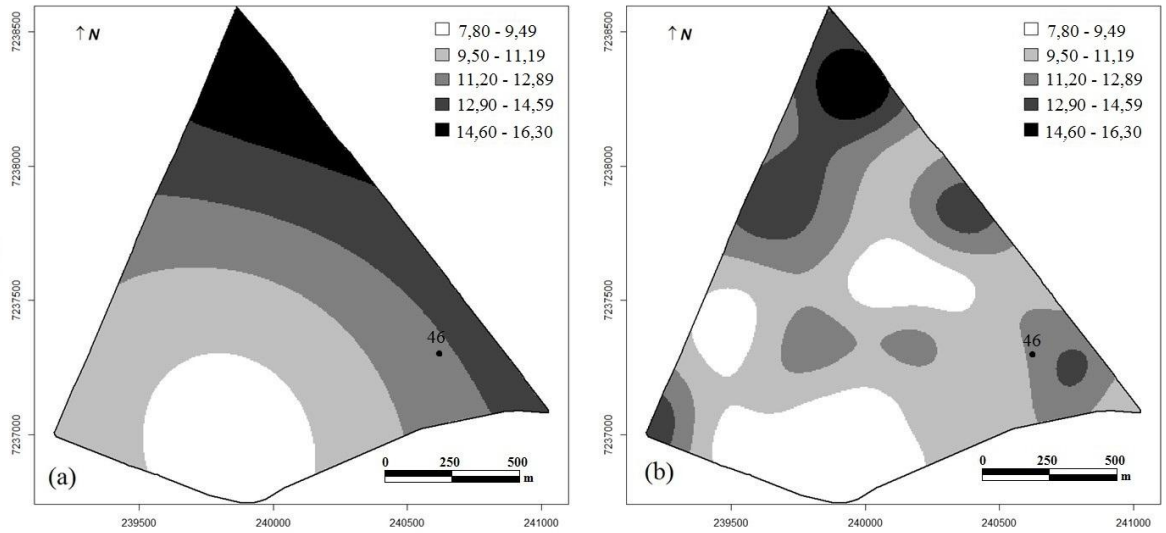

FIGURA 3. Mapas dos teores de fósforo (P) (mg dm^{-3}) com todos os pontos - CPI (a) e sem o ponto influente 46 - SPI (b). **Phosphorus content map (mg dm^{-3}) with all points – CPI (a) and without 46 influent point – SPI (b)**.

Os mapas (Figuras 3a e 3b) foram construídos com a mesma escala para sua comparação. Como resultados, foi possível a quantificação das proporções de áreas por classe de nível de teor de fósforo (Tabela 1). Observaram-se diferenças de área que variaram entre 1,25% (11,20 – 12,89 mg dm^{-3}) e 9,72% (9,50 – 11,19 mg dm^{-3}) nas classes dos mapas construídos com (CPI) e sem (SPI) o ponto influente do teor de fósforo, evidenciando que a existência do ponto influente altera a distribuição espacial do teor de fósforo.

TABELA 1. Proporção de áreas por classes dos mapas do teor de fósforo (mg dm^{-3}) com (CPI) e sem (SPI) o ponto influente. **Area proportion by class of the phosphorus content (mg dm^{-3}) with and without the influent point.**

Classes		7,80- 9,49	9,50-11,19	11,20-12,89	12,90-14,59	14,60-16,30	Total
Área (ha)	(CPI)	27,46	50,53	37,93	34,41	17,03	167,35
	(SPI)	38,08	66,79	35,83	21,07	5,58	
Área (%)	(CPI)	16,41	30,19	22,66	20,56	10,18	100
	(SPI)	22,76	39,91	21,41	12,59	3,34	
Diferença (%)		6,35	9,72	1,25	7,97	6,84	-

CPI: com ponto influente (Figura 3a); SPI: sem ponto influente (Figura 3b).

A partir da matriz de erros (Tabela 2), determinaram-se os valores dos índices exatidão global (EG), Kappa (Ka) e Tau (T) (DE BASTIANI et al., 2012), para a comparação dos mapas com e sem o ponto influente (Figura 3). Os valores encontrados de EG = 0,3097, Ka = 0,0775 e T = 0,1371 mostram diferenças na variabilidade espacial dos mapas do teor de fósforo, construídos com e sem o ponto influente, segundo a classificação apresentada por FOODY (2010), indicando assim a forte influência do ponto 46 sobre o mapa temático (Figura 3b).

TABELA 2. Matriz de erros para o teor de fósforo (mg dm^{-3}). **Matrix of errors for phosphorus content (mg dm^{-3}).**

	Classes	**Mapa Fósforo CPI (Figura 3 a)**					
		7,80 –9,49	9,50 - 11,19	11,20 – 12,89	12,90 - 14,59	14,60 - 16,30	Total
Mapa Prod. SPI (Figura 3 b)	7,80 - 9,49	1691	1023	849	334	0	3897
	9,50 - 11,19	5770	10508	7003	4750	1387	29418
	11,20 - 12,89	3423	7971	6623	7107	3639	28763
	12,90 - 14,59	104	381	587	1015	901	2988
	14,60 - 16,30	0	314	112	541	892	1859
	Total	10988	20197	15174	13747	6819	66925

CPI: com ponto influente (Figura 3a); SPI: sem ponto influente (Figura 3b).

A Figura 4 apresenta o *box-plot, hair-plot* e gráficos de influência local C_i e $|L_{max}|$ na identificação de pontos *outliers* e pontos influentes na variável teor de cálcio (Ca). De acordo com as técnicas estudas e a interpretação análoga à da variável teor de fósforo (P), todos os gráficos indicaram que a observação 57 é um ponto influente.

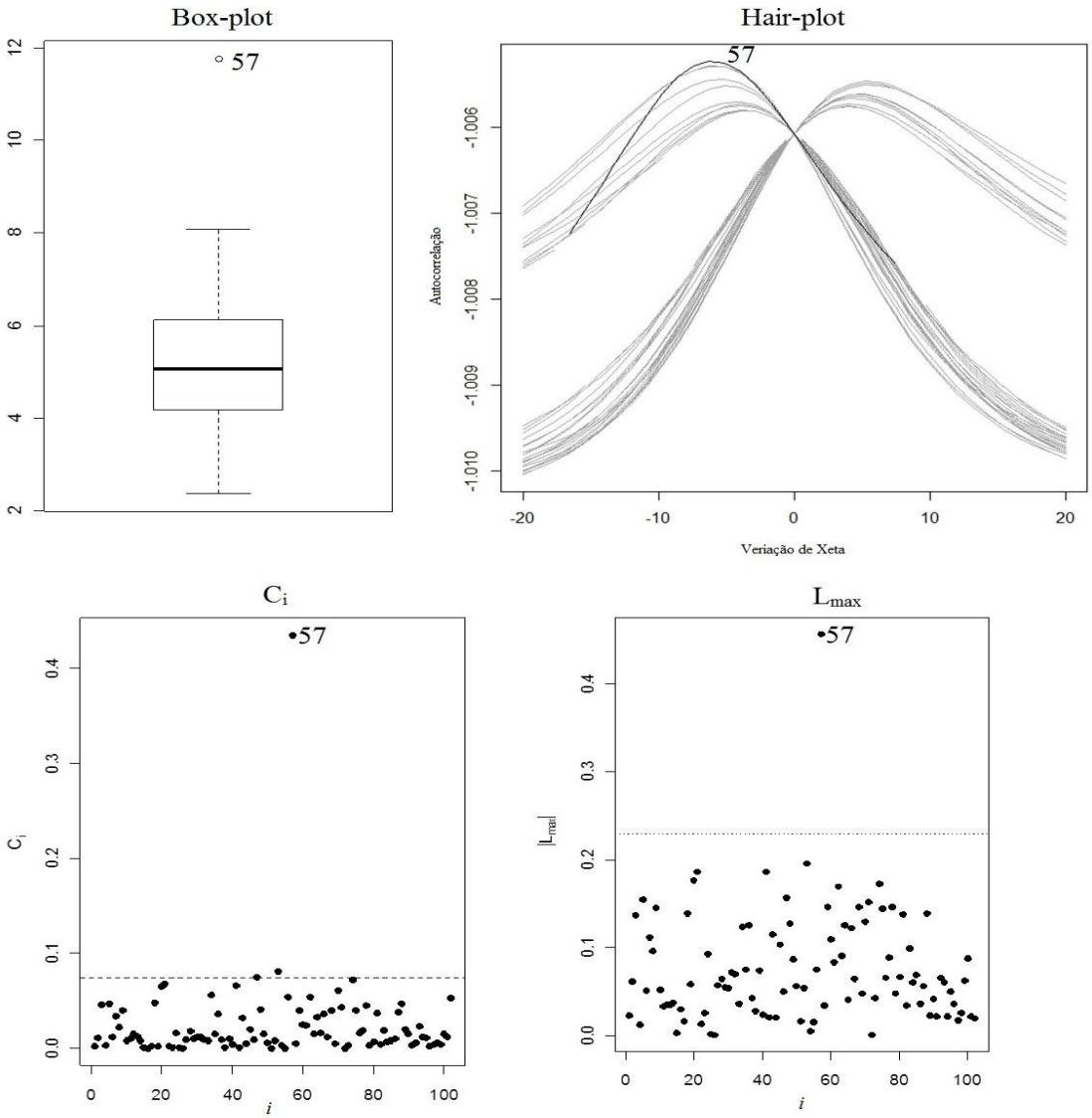

FIGURA 4. *Box-plot, Hair-plot, C_i* e $|L_{max}|$ na identificação de *outliers* e pontos influentes para a variável teor de cálcio (Ca) ($cmol_c\ dm^{-3}$). **Box-plot, Hair-plot, C_i and $|L_{max}|$ graphs to identify outliers and influent points for calcium content ($cmol_c\ dm^{-3}$).**

A construção dos mapas para o teor de cálcio (Ca) ($cmol_c\ dm^{-3}$), com todos os pontos amostrais e sem o ponto 57 (Figura 5a e 5b, respectivamente), levou em consideração os mesmos critérios já discutidos, e também o modelo gaussiano foi o que apresentou melhores ajustes por validação cruzada e *MLL*. Os parâmetros espaciais estimados foram: $\hat{\varphi}_1$ = 1,0577, com desvio padrão DP = 0,1565; $\hat{\varphi}_2$ = 0,4768, com DP = 0,3248, e $\hat{\varphi}_3$ = 689,3844, com DP = 3,2339 e raio de dependência espacial (alcance) de 1193,19 m para todos os pontos; e $\hat{\varphi}_1$ = 1,4176, com desvio- - padrão DP = 0,2099; $\hat{\varphi}_2$ = 0,5873, com DP = 0,3889 e $\hat{\varphi}_3$ = 615,0229 com DP = 2,2966 e raio de dependência espacial (alcance) de 1064,49 m. Observam-se, nos modelos, pequenas diferenças nas estimações dos parâmetros que definem a estrutura de dependência espacial com e sem o ponto influente 57. Pelos mapas (Figura 5), observa-se que os níveis do teor de Ca ($cmol_c\ dm^{-3}$), a exemplo do que ocorreu para o teor de fósforo (Figura 3), são mais altos na região norte e decrescem em direção ao sul da área.

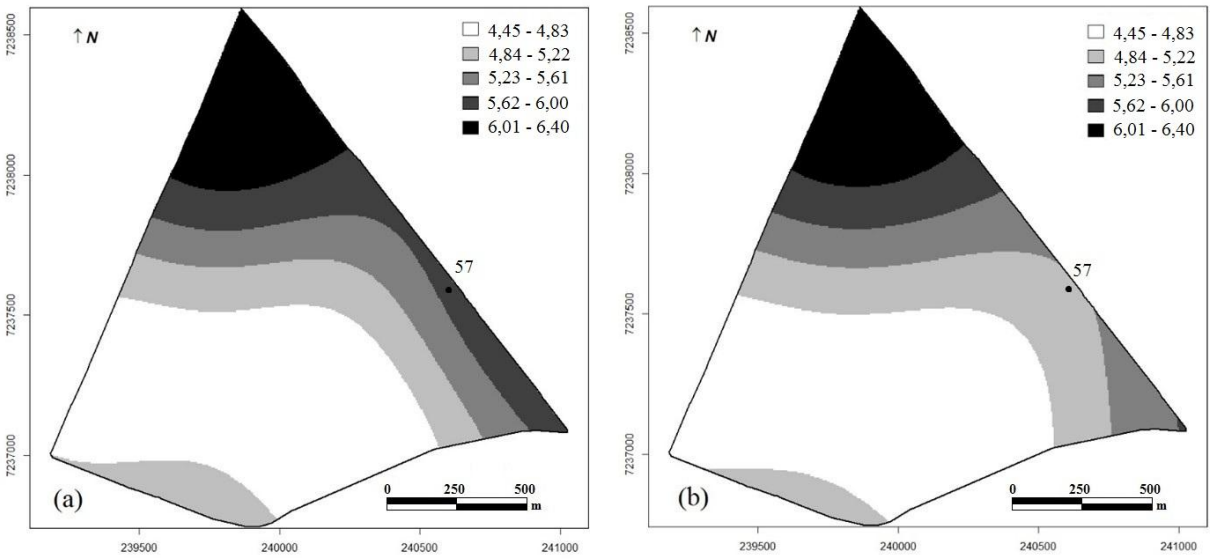

FIGURA 5. Mapas dos teores de Cálcio ($cmol_c\ dm^{-3}$) com todos os pontos - CPI (a) e sem o ponto influente 57 - SPI (b). **Calcium content maps ($cmol_c\ dm^{-3}$) with all points – CPI (a) and without 57 influent point– SPI (b).**

Observaram-se diferenças de área inferior a 4% entre os mapas construídos com e sem o ponto influente do teor de cálcio (Tabela 3). Na comparação espacial, através da matriz de erros (Tabela 4), dos mapas do teor de Ca ($cmol_c\ dm^{-3}$) com (Figura 5a) e sem o ponto influente (Figura 5b), encontraram-se EG = 0,5306, Ka = 0,4077 e T = 0,4132, mostrando menores diferenças entre os mapas do teor de cálcio quando comparados aos do teor de fósforo, onde os índices foram maiores.

TABELA 3. Proporção de áreas por classes dos mapas do teor de cálcio ($cmol_c\ dm^{-3}$) com e sem o ponto influente. **Area proportion by class of calcium content ($cmol_c\ dm^{-3}$) with and without the influent point.**

Classes		4,45 - 4,83	4,84 - 5,22	5,23 - 5,61	5,62 - 6,00	6,01 - 6,40	Total
Área (ha)	(CPI)	75,22	37,54	20,73	12,53	21,33	167,35
	(SPI)	69,54	31,12	23,27	21,58	21,85	
Área (%)	(CPI)	44,95	22,43	12,39	7,49	12,74	100
	(SPI)	41,55	18,59	13,90	12,90	13,06	
Diferença (%)		3.39	3,84	1,51	5,41	0,31	-

CPI: com ponto influente (Figura 5a); SPI: sem ponto influente (Figura 5b)

TABELA 4. Matriz de erros para do teor de cálcio ($cmol_c\ dm^{-3}$). **Matrix of errors for calcium content ($cmol_c\ dm^{-3}$).**

	Classes	Mapa do teor de cálcio CPI (Figura 5 a)					
		4,45 - 4,83	4,84 - 5,22	5,23 - 5,61	5,62 - 6,00	6,01 - 6,40	Total
Mapa do teor de Calcio SPI (Figura 5 b)	4,45 - 4,83	10775	0	0	0	0	10775
	4,84 - 5,22	17053	12461	5143	1674	0	36331
	5,23 - 5,61	0	0	4156	4409	0	8565
	5,62 - 6,00	0	0	0	2545	3141	5686
	6,01 - 6,40	0	0	0	0	5598	5598
	Total	27828	12461	9299	8628	8739	66955

CPI: com ponto influente (Figura 5a); SPI: sem ponto influente (Figura 5b).

A Figura 6 apresenta o *box-plot*, *hair-plot* e os gráficos de influência local C_i and $|L_{max}|$ na identificação de pontos *outliers* e influentes na produtividade da soja. Neste caso, de acordo com as técnicas *hair-plot* e C_i e $|L_{max}|$, foram considerados os pontos 32 e 87 como influentes.

Diferentemente do que ocorreu com o teor do fósforo e o teor de cálcio, os pontos *outliers* detectados pelo *box-plot* da produtividade da soja (pontos 36; 18; 47 e 79) não foram considerados influentes para a geração dos mapas. Uma explicação para isto é que a maioria destes pontos (18; 47 e 79, que foram *outliers* superiores) estão espacialmente localizados a leste, na borda limite da área em estudo (Figura 1), não sendo assim considerados como pontos influentes, pelas técnicas propostas, para a geração dos mapas. Em função disto, fica evidente a importância do uso destas técnicas que consideram a localização das observações, em detrimento do simples uso do *box-plot*, como ferramenta para detecção de dados que possam interferir na construção dos mapas, já que estes são utilizados pelos agricultores para identificar regiões com maiores e menores produtividades, bem como para executar possíveis manejos diferenciados na área.

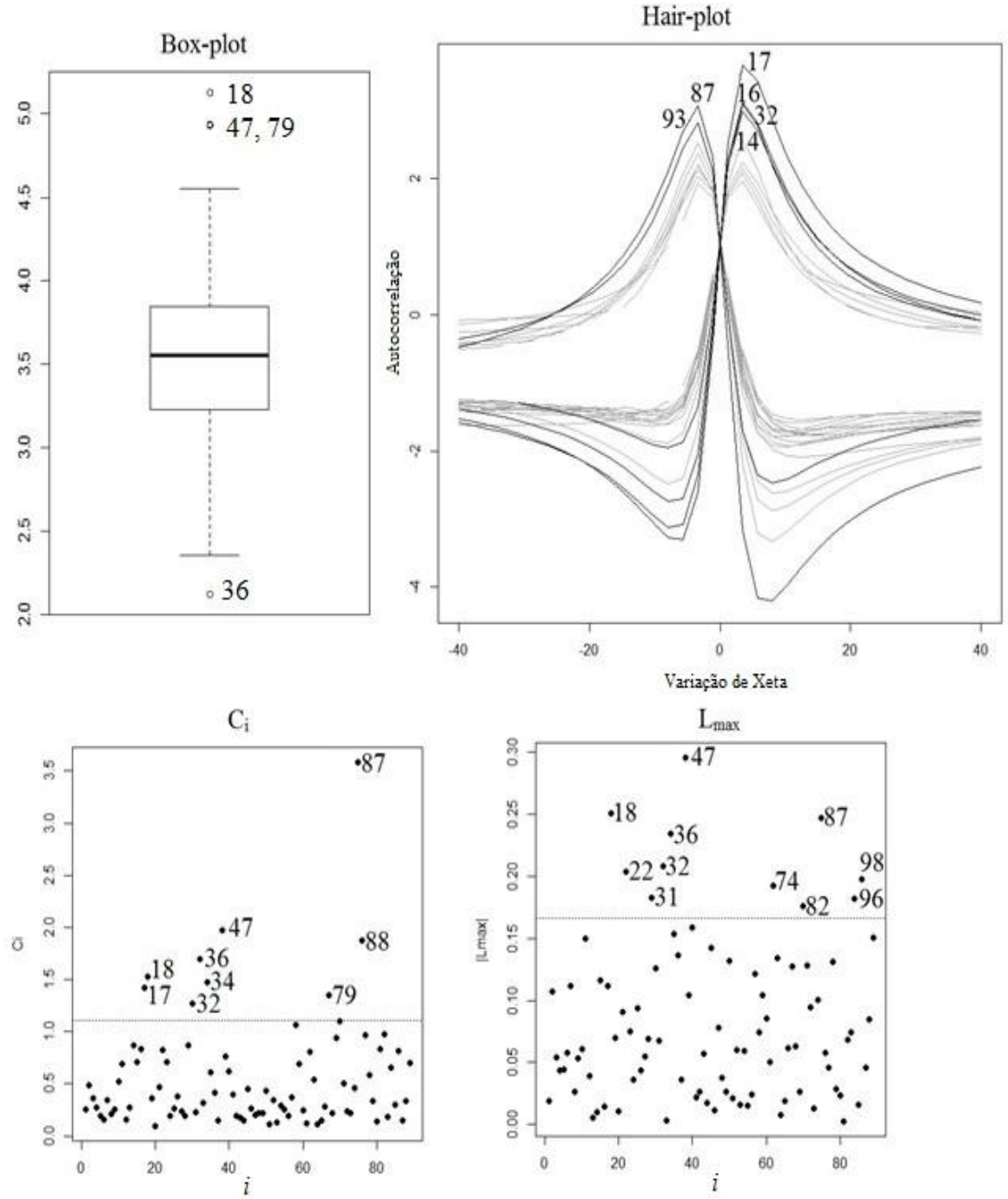

FIGURA 6. *Box-plot*, *Hair-plot*, C_i e $|L_{max}|$ na identificação de pontos *outliers* e influentes para a produtividade da soja. **Box-plot, hair-plot, C_i and $|L_{max}|$ graphs to identify outliers and influent points for soybean yield.**

Para o estudo de variabilidade espacial da produtividade da soja (t ha^{-1}), com todos os pontos amostrais e sem os pontos influentes 32 e 87, segundo os critérios de validação cruzada e *MLL*, o modelo que melhor se ajustou foi o modelo Matérn com parâmetro de forma k = 1,5 para todos os pontos amostrais (Figura 7a) e k = 1,0 para os dados sem os pontos influentes (Figura 7b). Os parâmetros de variabilidade espacial estimados foram: $\hat{\varphi}_1$ = 0,2663, com desvio-padrão DP = 0,1119; $\hat{\varphi}_2$ = 0,0199, com DP = 0,1035; e $\hat{\varphi}_3$ = 80,8388 com DP = 0,0002, com raio de dependência espacial (alcance) de 383,4885 m para todos os pontos; e $\hat{\varphi}_1$ = 0,2696, com DP = 0,1304; $\hat{\varphi}_2$ = 0,0220, com DP = 0,1457, e $\hat{\varphi}_3$ = 115,7761, com DP = 0,0019 e raio de dependência espacial (alcance) de 462,9333 m para os ajustes sem os pontos influentes. Observam-se nos modelos diferenças nas estimativas dos parâmetros que definem a estrutura de variabilidade espacial com e sem os pontos influentes, em especial no alcance de dependência espacial que aumentou cerca de 21% com a retirada dos pontos influentes.

Os mapas para a produtividade da soja (t ha^{-1}) são apresentados nas Figuras 7a e 7b, respectivamente, com todos os pontos amostrais e sem os pontos influentes 32 e 87. As Figuras ilustram que a retirada destes pontos resultou em alterações na variabilidade espacial da produtividade da soja.

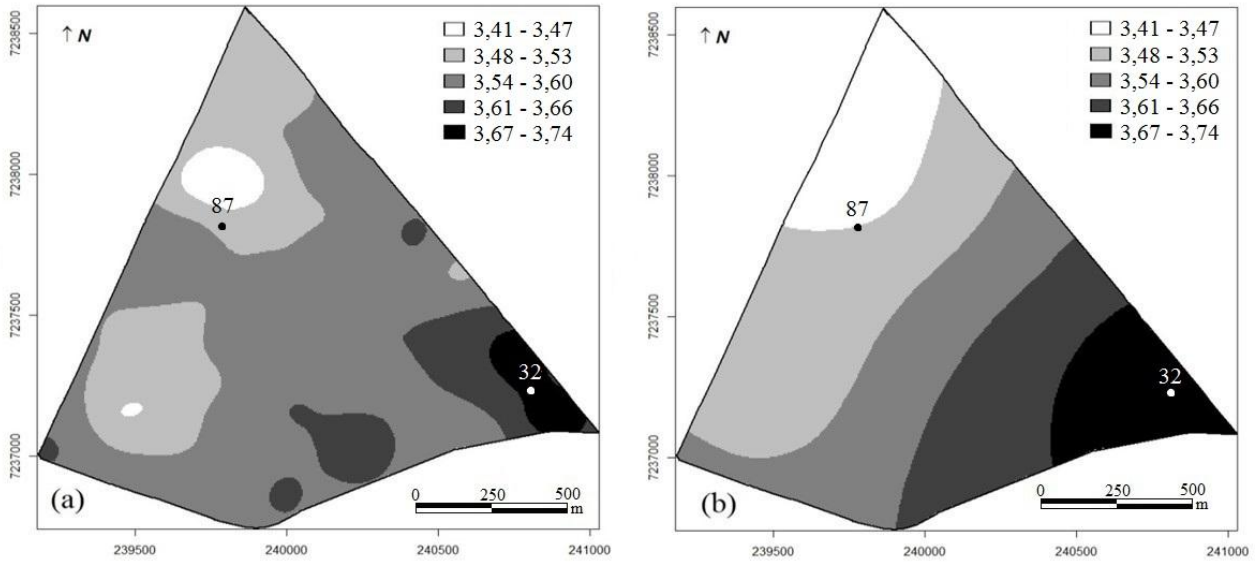

FIGURA 7. Mapas da produtividade da soja (t ha^{-1}) com todos os pontos (CPI) (a) e sem (SPI) os pontos influentes 32 e 87 (b). **Soybean yield maps (t ha^{-1}) with all points (CPI) (a) and without 32 and 87 influent points (SPI) (b).**

Na Tabela 5, observa-se uma diferença entre os mapas da produtividade da soja superior a 8% em praticamente 97% da área (segunda a quinta classes), ressaltando a diferença de 34,10% de área com soja, no mapa construído sem os pontos influentes, na terceira classe, e a diferença de 10,22% da área com soja, quando se trabalhou sem os pontos influentes na segunda classe.

TABELA 5. Diferença e número de pixels dos mapas da produtividade da soja (t ha^{-1}). **Difference and number of pixels of the soybean yield maps (t ha^{-1})**.

Classes		3,41 - 3,47	3,48 - 3,53	3,54 - 3,60	3,61 - 3,66	3,67 - 3,74	Total
Área (ha)	(CPI)	8,20	31,14	100,13	19,64	8,25	167,35
	(SPI)	6,99	29,95	61,04	41,14	28,24	
Área (%)	(CPI)	3,56	23,24	60,63	8,98	3,58	100
	(SPI)	3,04	13,02	26,53	17,88	12,27	
Diferença (%)		0,52	10,22	34,10	8,90	8,69	-

CPI: com ponto influente (Figura 7a); SPI: sem ponto influente (Figura 7b).

Na comparação espacial, através da matriz de erros (Tabela 6), dos mapas da produtividade da soja com (Figura 7a) e sem os pontos influentes (Figura 7b), obtiveram-se EG = 0,5634; Ka = 0,3735, e T= 0,4542, evidenciando diferenças visuais já observadas entre os mapas da produtividade da soja (t ha^{-1}) construídos com e sem os pontos influentes.

TABELA 6. Matriz de erros para a produtividade de soja (t ha^{-1}). **Matrix of errors for soybean yield (t ha^{-1}).**

	Classes	Mapa Produtividade CPI (Figura 7 – a)					
		3,41 - 3,47	3,48 - 3,53	3,54 - 3,60	3,61 - 3,66	3,67 - 3,74	Total
Mapa	3,41 - 3,47	2031	6157	614	0	0	8802
Prod. SPI	3,48 - 3,53	350	7376	7533	0	0	15259
(Figura 7 –	3,54 - 3,60	0	1930	22494	499	0	24923
b)	3,61 - 3,66	0	89	8895	3411	0	12395
	3,67 - 3,74	0	0	1052	2106	2398	5556
	Total	2381	15552	40588	6016	2398	66935
Diferença (%)		0,52	10,22	34,10	8,90	8,69	-

CPI: com ponto influente (Figura 7a); SPI: sem ponto influente (Figura 7b).

De acordo com este estudo, todas as técnicas gráficas indicaram que vários pontos foram considerados como influentes. Os pontos conjuntamente influentes são apresentados em negrito para o teor de fósforo (P), teor de cálcio (Ca) e produtividade da soja (Prod), na Tabela 7.

TABELA 7. Pontos influentes detectados pelos gráficos *hair-plot* e C_i e $|L_{max}|$ de Influência local. **Influent points detected by hair-plot, C_i and $|L_{max}|$ local influence graphs**.

	Box-plot	Hair-plot	Influência Local	
			C_i	$\|L_{max}\|$
P	19-32-44-**46**	**46**	44 - **46**	44 - **46**
Ca	**57**	**57**	**57**	**57**
Prod	18-47-79-36	14-16-17-**32**-**87**-93	17-18-**32**-34-36-47-79-**87**-88	18-22-31-**32**-36-47-82-**87**-96-98

Finalmente, para validar os resultados de cada variável, foi analisado o impacto da retirada de um ponto escolhido aleatoriamente. Para o teor de fósforo, retirou-se o ponto 26, para o teor de cálcio retirou-se o ponto 72, e para a produtividade da soja, o ponto 51. Foram construídos os mapas sem estes pontos, e comparados com seu respectivo mapa, com todos os pontos. Observou-se que a retirada destes pontos escolhidos aleatoriamente, dentre os pontos não detectados pelos gráficos de influência, não interfere na estimação da estrutura de dependência e, consequentemente, também não interferem na construção dos mapas. Pois, os valores de EG variaram de 0,9316 a 0,9972, de Ka variam entre 0,8795 e 0,9962 e de T entre 0,9145 e 0,9965, o que caracteriza alta similaridade entre os mapas.

CONCLUSÕES

Embora comumente a identificação de *outliers* seja realizada numa análise estatística descritiva através do gráfico *box-plot*, verificou-se, no estudo espacial, que nem sempre *outliers* interferem na análise de dados com dependência espacial.

Através deste estudo, ficou evidente que a aplicação em conjunto de técnicas de diagnóstico de influência local e o gráfico *hair-plot* permitem identificar espacialmente quais são os dados que interferem na dependência espacial. Sendo assim, estas técnicas devem fazer parte de toda análise geoestatística. Isto permite avaliar estes pontos com cuidado, garantindo que as informações contidas nos mapas tenham maior qualidade e possam ser utilizadas com maior precisão pelo agricultor, evitando, no contexto do manejo localizado, que determinadas regiões recebam adubação

em excesso ou déficit, o que, por consequência, significa desperdício de recursos financeiros bem como passivos ambientais e produtividades inferiores ao potencial da cultura.

Os pontos apontados como potencialmente influentes interferem na estimação da estrutura de dependência espacial, de tal forma que os mapas construídos sem estes pontos se caracterizam com baixa similaridade em relação aos mapas construídos com todos os pontos, enquanto a retirada aleatória de um ponto não detectado pelos gráficos de influência não apresentou mapas com alta similaridade em relação ao mapa com todos os pontos. Portanto, os métodos de diagnósticos de influência utilizados mostraram-se efetivos.

AGRADECIMENTOS

Ao CNPq, à CAPES, à Fundação Araucária e à FACEPE, pelo apoio financeiro, e aos professores Doutores Genton e Ruiz-Gazen, por cederem as rotinas computacionais do *hair-plot*.

REFERÊNCIAS

ANSELIN, L. Under the hood: issues in the specification and interpretation of spatial regression models. **Agricultural Economics**, Amsterdam, v.27, p.247-267, 2002.

AUNE-LUNDBERG, L.; STRAND, G.H. Comparison of variance estimation methods for use with two-dimensional systematic sampling of land use/land cover data. **Environmental Modelling & Software**, Oxford, v.61, p.87-97, 2014.

BYRD, R.H.; LU, P.; NOCEDAL, J.; ZHU, C. A limited memory algorithm for bound constraints optimization. **SIAM Journal on Scientific Computing**, Philadelphia, v.16, n.5, p.1190-1208, 1995.

COOK, R.D. Detection of influential observations in linear regression. **Technometrics**, Rochester, v.19, n.1, p.15-18, 1977.

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.; HAWKINS, D.M. Robust estimation of the variogram: I. **Journal of the international Association for Mathematical Geology**, New York, v.12, n.2, p.115-125, 1980.

DE BASTIANI, F.; URIBE-OPAZO, M.A.; DALPOSSO, G.H. Comparison of maps of spatial variability of soil resistance to penetration constructed with and without covariables using a spatial linear model. **Engenharia Agrícola**, Jaboticabal, v.32, n.2, p.394-404, 2012.

DE BASTIANI, F.; CYSNEIROS, A.H.M.A.; URIBE-OPAZO, M.A.; GALEA, M. Influence diagnostics in elliptical spatial linear models. **TEST**, Berlin, v.24, n.2, p.322-340, 2015.

EMBRAPA – Centro Nacional de Pesquisa de Solos. **Sistema brasileiro de classificação de solos**. 3. ed. Rio de Janeiro: EMBRAPA – SPI, 2013. 412p.

FARACO, A.M.; URIBE-OPAZO, M.A.; SILVA, E.A.A.; JOHANN, J.J.; 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 Ciências do Solo**, Viçosa, v.32, n.2, p.463-479, 2008.

FOODY, G.M. Assessing the accuracy of land cover change with imperfect ground reference data. **Remote Sensing of Environment**, New York, v.114, n.10, p.2271-2285, 2010.

GENTON, M.G. Spatial breakdown point of variogram estimators. **Mathematical Geology**, v.30, n.7, p. 853-871, 1998.

GENTON, G.M.; RUIZ-GAZEN, A. Visualizing influential observations in dependent data. **Journal of Computational and Graphical Statistics**, Alexandria, v.19, n.4, p.808-825, 2010.

GUEDES, L.P.C.; URIBE-OPAZO, M.A.; RIBEIRO JÚNIOR, P.J. Influence of incorporating geometric anisotropy on the construction of thematic maps of simulated data and chemical attributes of soil. **Chilean Journal of Agricultural Research**, Santiago, v.73, n.4, p.414-423, 2013.

LARK, R.M. Optimized sampling of soil for estimation of the variogram by maximum likelihood. **Geoderma**, Amsterdam, v.105, n.1-2, p.49-80, 2002.

MARDIA, K.V.; MARSHALL, R.J. Maximum likelihood models for residual covariance in special regression. **Biometrika**, Oxford, v.71, n.1, p.319-332. 1984.

MILITINO, A.F.; PALACIUS, M.B.; UGARTE, M.D. Outliers detection in multivariate spatial linear models. **Journal of Statistical Planning and Inference**, Amsterdam, v.136, n.1, p.125-146, 2004.

MORAN, P.A.P. Notes on continuous stochastic phenomena. **Biometrika**, Cambridge, v.37, n.1-2, p.17–23, 1950.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, 2013. Disponível em: <http://www.R-project.org>. Acesso em: 25 out. 2014.

RIBEIRO JÚNIOR, P.J.; DIGGLE P.J. geoR: A package for geostatistical analysis. **R News**, v.1-2, p.15-18, June 2001.

SHEKHAR, S.; LU, C-T.; ZHANG, P. A unified approach to detecting spatial outliers. **Geoinformatica**, Dordrecht, v.7, n.2, p.139-166, 2003.

URIBE-OPAZO, M. A.; BORSSOI, J. A.; GALEA, M. Influence diagnostics in Gaussian spatial linear models. **Journal of Applied Statistics**, London, v.39, n.3, p.615-630, mar. 2012.

WEBSTER, R.; OLIVER, M.A. **Geostatistics for environmental scientists**. Iowa: Willey, 2. ed., 2007. 315p.