Ciencia e Investigación AGRARIA

Cien. Inv. Agr. 40(3):523-535. 2013
www.rcia.uc.cl

CROP PRODUCTION

RESEARCH PAPER

Local influence when fitting Gaussian spatial linear models: an agriculture application

Denise M. Grzegozewski[1], Miguel A. Uribe-Opazo[1], Fernanda De Bastiani[2], and Manuel Galea[3]

[1]Graduate Program in Agricultural Engineering (PGEAGRI), Western Paraná State University – UNIOESTE. Universitária Street 2069 – 85819-110, Cascavel, Paraná, Brazil.

[2]Graduate Program in Statistics, Federal University of Pernambuco, Universitária City 50740-540, Recife, Pernambuco, Brazil.

[3]Facultad de Matemáticas, Pontificia Universidad Católica de Chile. Ave. Vicuña Mackenna 4860, Santiago, Chile.

Abstract

D.M. Grzegozewski, M.A. Uribe-Opazo, F. De Bastiani, and M. Galea. 2013. Local influence when fitting Gaussian spatial linear models: an agriculture application. Cien. Inv. Agr. 40(3): 523-535. Outliers can adversely affect how data fit into a model. Obviously, an analysis of dependent data is different from that of independent data. In the latter, *i.e*., in cases involving spatial data, local outliers can differ from the data in the neighborhood. In this article, we used the local influence technique to identify influential points in the response variables using two different schemes of perturbations. We applied this technique to soil chemical properties and soybean yield. We evaluated the effects of the influential points on the spatial model selection, the parameter estimation by maximum likelihood and the construction of thematic maps by kriging. In the construction of the thematic maps in studies with and without the influential points, there were changes in the levels of nutrients, allowing for the appropriate application of input, generating greater savings for the producer and contributing to the protection of the environment.

Key words: Geostatistical, influence diagnostics, maximum likelihood, outliers, spatial variability.

Introduction

The presence of outliers in a dataset can cause disproportional interference in the analysis. Using geostatistical study, which is based on the theory of regionalized variables, it is possible to define the spatial variability structure of observations. This interference can affect the choice of model fit and, consequently, the parameter estimation.

Cressie and Hawkins (1980) proposed a robust estimator for the semivariance function, but it still can be affected by the presence of outliers in the data set. Using simulations, McBrantney and Webstes (1986) suggested that the semivariance estimator of the function is limited. Genton (1998) commented on the robustness

Received February 6, 2013.Accepted October 9, 2013.
Corresponding author: denisegrzegozewski@gmail.com

of the estimator from a statistical viewpoint. Using a simple linear model for predicting tree diameter from laser-derived tree height and crown diameter measurements, Salas *et al*. (2010) compared the performance of ordinary least squares (OLS), generalized least squares with a non-null correlation structure (GLS), a linear mixed-effects model (LME), and geographically weighted regression (GWR).

In the study of influential points, the literature presents the methodology of diagnostics in global influence, which is based on the elimination of points of the total dataset (Cook, 1977; Paula, 2010), and of diagnostics in local influence, which was proposed by Cook (1986) to study the behavior of some particular measures of influence on the parameter estimates. This technique is performed using appropriate measures of influence to assess the robustness of estimates that are provided by the model from small perturbations in the data (Paula, 2010). This technique does not demand the elimination of observations and allows for simultaneously evaluating the joint influence of all influential points. Christensen *et al*. (1993) studied the methods that are used for georeferenced data and global influence diagnostics based on the elimination of influential points. Uribe-Opazo *et al*. (2012) studied the georeferenced data for the diagnostic methods of local influence in Gaussian spatial linear models using additive perturbation in the response variable.

Borssoi *et al*. (2011) presented a technique of local influence on spatial covariates in Gaussian spatial linear models. Botinha *et al*. (2011) presented a study of spatial local influence on data of soil physical properties and soybean yield, considering the additive perturbation scheme for the response variable as well as the supposition of the Student-t distribution for the model.

Geostatistics, together with precision agriculture, studies factors that affect the spatial variability of attributes relating to the soil and crop, selecting models that best explain the yield and determining the causes of variations in agricultural production. In agriculture, the interaction between soil chemical attributes directly affects the growth and development of crops; furthermore, the characterization and assessment of spatial variability are essential for managing a culture. Using this information, it is possible to map the attributes in question and elaborate prescription maps, which aim to increase agricultural production and decrease the effects of an overdose of inputs on the environment.

This study aimed to detect influential points in the response variable using the local influence technique with two different schemes of perturbations: additive perturbation and Zhu perturbation.

Materials and methods

The data were collected in western Paraná in a commercial area of grain production in Cascavel City; the geographical location of the city is approximately 24.95° south latitude, 53.57° west longitude, and its average altitude is 650 m. The soil is classified as clayey Hapludox (EMBRAPA, 2009). The climate is classified as mesothermal super humid and presents as a mild mesothermal, super humid Cfa Köeppentype climate with moderate temperatures, well-distributed rainfall and hot summers and an annual average temperature of 21 ºC. The data refer to the agricultural year 2010/2011 and reference an area of 167.35 ha. This study used a systematic sampling, known as *lattice plus close pairs*, with a maximum distance of 141 m between points and, in some random locations, shorter distances of 75 and 50 m between points, there by obtaining 89 sampling points. All of the samples were georeferenced and located using the space-based satellite navigation global positioning system (GPS) GeoExplorer3 and GPS Pathfinder trademarks of Trimble Navigation Limited registered in the United Stateswith an accuracy between one and five meters, with system in Universal

Transverse Mercator UTM coordinates, zone 22 and datum WGS 84.

The variables that were measured in the experimental area included Carbon [C] (g dm^{-3}), Calcium [Ca] ($cmol_c$ dm^{-3}), Potassium [K] ($cmol_c$ dm^{-3}), Magnesium [Mg] ($cmol_c$ dm^{-3}), Manganese [Mn] (mg dm^{-3}), Phosphorus [P] (mg dm^{-3}) and soybean yield (productivity) [Prod] (t ha^{-1}). Soybean yield was estimated considering the amount of soybeans harvested in an area of 0.90 m^2, representing the plot. After the screening, the water content was tested, underwent posterior correction to 13% and was converted into t ha^{-1}. Soil sampling to determine the soil chemical properties was performed at each marked point and included four soil subsamples collected close to the points at a depth of 0.0 to 0.2 m. These subsamples were mixed and weighed approximately 500 g, thus composing the representative sample of the plot, and were sent to the COODETEC laboratory where chemical analyses were performed.

Let $\{Z(s_i), s_i \in S\}$ be a stochastic process, with $S \subset \mathcal{R}^2$ being $\mathcal{R}^2$ two-dimensional Euclidean space. We assumed that the elements $Z(s_1), \ldots, Z(s_n)$ of this process are recorded at known spatial locations $s_i,\ i = 1, \cdots, n$ and are generated by the model shown in Equation (1) (Webster and Oliver, 2007)

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

where the deterministic $\mu(s_i)$ and stochastic $e(s_i)$ terms may depend on the spatial location in which $Z(s_i)$ was obtained. It is assumed that the stochastic error e has $E[e(s_i)] = 0$ and that the variation between the points in space is determined by some covariance function $C(s_i, s_u) = Cov[\varepsilon(s_i), \varepsilon(s_u)]$ and, in some known functions of s, such as $X_1(s_i), \ldots, X_p(s_i)$, $\mu(s_i)$, is defined as in Equation (2)

$$\mu(s_i) = \sum_{u=1}^{p} X_u(s_i)\beta_u,\ i = 1, \ldots, n,$$

where $\beta_1, \ldots, \beta_p$ are unknown parameters to be estimated. In matrix notation, we define the spatial linear model in Equation (3) as

$$Z = X\beta + \varepsilon, \quad (3)$$

where ε is the error vector $nx\ 1$, with $E(\varepsilon) = 0$, zero vector $nx\ 1$; X, nxp, is a matrix of covariates of full rank; and Z is the vector of response variables. These data follow an n- varied normal distribution with mean vector $X\beta$ and covariance matrix $\Sigma = [(\sigma_{iu})]$, nxn, being $\sigma_{iu} = C(s_i, s_u)$, *i.e.*, $Z \sim N_n(X\beta, \Sigma)$. It is assumed that Σ is not singular and has a structure as defined in Equation (4)

$$\Sigma = \Sigma(\varphi) = \varphi_1 I_n + \varphi_2 R(\varphi_3), \quad (4)$$

where φ_1, φ_2 and φ_3 are the nugget effect parameters, contribution and function of range, respectively, which define the structure of spatial dependence. I_n is the identity matrix nxn; $R(\varphi_3)$ represents the matrix nxn, which is a function of φ_3 and depends on the fitted model, *i.e.*, $R(\varphi_3) = [(r_{iu})]$ is a symmetric matrix with diagonal elements $r_{iu} = 1,\ i = 1, \ldots, n$, $r_{iu} = \frac{1}{\varphi_2}\sigma_{iu}$ to $\varphi_2 > 0\ \ i \neq u = 1 \ldots, n$.

To study the spatial local influence based on the study of local influence presented by Cook (1986), Uribe-Opazo *et al.* (2012) measured the behavior of the likelihood deviation function in a neighborhood $LD(\omega) = 2\left(l(\hat{\theta}) - l(\hat{\theta}_\omega)\right)$, in which $\hat{\theta}$ is the maximum likelihood (ML) estimator of $\theta = (\beta', \varphi')'$, where $\beta = (\beta_1, \ldots, \beta_p)$ and $\varphi = (\varphi_1, \varphi_2, \varphi_3)'$ of the postulated model and $\hat{\theta}_\omega$ is the ML estimator of the θ model perturbed by ω, where $\omega = (\omega_1, \ldots, \omega_n)'$ is the perturbation vector of the response variable and $\omega_0 = 0_n = (0, \ldots, 0)'$ is the non-perturbed point. The log likelihood function of the estimated parameters is $l(\hat{\theta})$, and $l(\hat{\theta}_\omega) = l(\hat{\theta}/\omega)$ is the perturbed log likelihood function given in Equation (5).

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

We consider two different schemes of perturbation. Scheme 1 is known as additive perturbation and is given by $\boldsymbol{Z_\omega = Z + \omega}$. Scheme 2, given by $\boldsymbol{Z_\omega = Z + \Sigma^{-1/2}\omega}$, was proposed by Zhu *et al*. (2007) and is called Zhu perturbation. These schemes of perturbation detect possible outliers in the dataset that can affect the ML estimator of $\boldsymbol{\theta}$. The goal is to study the local behavior of $\boldsymbol{LD(\omega)}$ around $\boldsymbol{\omega_0 \epsilon \Omega}$ such that $\boldsymbol{l(\theta) = l(\theta/\omega_0)}$. For this goal, the normal curvature Ci of LD(ω) in ω0 in the direction of some unit vector l, defining $\boldsymbol{C_i = 2|l'\Delta' L^{-1}\Delta l|}$, with $\boldsymbol{\|l\| = 1}$, where in L is the hessian matrix, evaluated in Θ=Θ and $\boldsymbol{\Delta}$ is the matrix $\boldsymbol{(3p)x\,n}$, given by $\boldsymbol{\Delta = (\Delta'_\beta, \Delta'_\varphi)'}$ and evaluated in Θ=Θ and in $\boldsymbol{\omega = \omega_0}$, where for the additive perturbation in the response variable $\boldsymbol{Z_\omega = Z + \omega}$,

$\boldsymbol{\Delta_\beta = \frac{\partial^2 l(\theta/\omega)}{\partial\beta\partial\omega'} = X'\Sigma^{-1}}$ and $\boldsymbol{\Delta_\varphi = \frac{\partial^2 l(\theta/\omega)}{\partial\varphi\partial\omega'}}$, with elements $\boldsymbol{\frac{\partial^2 l(\theta/\omega)}{\partial\varphi_j\partial\omega'} = (Z_\omega - X\beta)'\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j}\Sigma^{-1}}$, $\boldsymbol{j = 1,2,3,}$

where, $\boldsymbol{L = \begin{pmatrix} L_{\beta\beta} & L_{\beta\varphi} \\ L_{\varphi\beta} & L_{\varphi\varphi} \end{pmatrix}}$, where in, $\boldsymbol{L_{\beta\beta} = -(X'\Sigma^{-1}X)}$

and $\boldsymbol{L_{\beta\varphi} = \frac{\partial^2 l(\theta)}{\partial\beta\,\partial\varphi'}}$, with elements

$\boldsymbol{\frac{\partial^2 l(\theta)}{\partial\beta\partial\varphi_j} = -X'\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j}\Sigma^{-1}\varepsilon}$ and $\boldsymbol{j = 1,2,3}$ with

$\boldsymbol{\varepsilon = (Z - X\beta);}$

$\boldsymbol{L_{\varphi\beta} = L_{\beta\varphi}'}$ and $\boldsymbol{L_{\varphi\varphi} = \frac{\partial^2 l(\theta)}{\partial\varphi\partial\varphi'}}$ with elements

$$\boldsymbol{\frac{\partial^2 l(\theta)}{\partial\varphi_i\partial\varphi_j} = \frac{1}{2}tr\left\{\Sigma^{-1}\left(\frac{\partial\Sigma}{\partial\varphi_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j} - \frac{\partial^2\Sigma}{\partial\varphi_i\partial\varphi_j}\right)\right\} + \frac{1}{2}\varepsilon'\Sigma^{-1}\left\{\frac{\partial^2\Sigma}{\partial\varphi_i\partial\varphi_j} - \frac{\partial\Sigma}{\partial\varphi_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_j} - \frac{\partial\Sigma}{\partial\varphi_j}\Sigma^{-1}\frac{\partial\Sigma}{\partial\varphi_i}\right\}\Sigma^{-1}\varepsilon.} \quad (6)$$

Zhu *et al*. (2007) proposed a perturbation in the response variable using the covariance matrix $\boldsymbol{Z_\omega = Z + \Sigma^{-1/2}\omega}$, where the matrix $\boldsymbol{\Delta}$ **is** given by $\boldsymbol{\Delta = (\Delta'_\beta, \Delta'_\varphi)'}$ evaluated in $\boldsymbol{\theta = \hat{\theta}}$ and in $\boldsymbol{\omega = \omega_0}$.

In this case, the matrices Δ_β and Δ_ϕ are given by $\boldsymbol{\Delta_\beta = X'\Sigma^{-1}\Sigma^{-1/2}}$ and $\boldsymbol{\Delta_\varphi}$ with elements

$$\boldsymbol{\Delta_{\varphi j} = \varepsilon'_\omega\Sigma^{-1}\left[\frac{\partial\Sigma}{\partial\varphi_j}\Sigma^{-1} + \Sigma^{-1/2}\frac{\partial\Sigma^{1/2}}{\partial\varphi_j}\right]\Sigma^{-1/2} \text{ and } j = 1,2,3,} \quad (7)$$

with $\boldsymbol{\varepsilon_\omega = (Z_\omega - X\beta)}$.

The matrix $\boldsymbol{B = \Delta^T L^{-1}\Delta}$ and $\boldsymbol{C_i = 2|b_{ii}|}$, where $\boldsymbol{b_{ii}}$ represents the main diagonal elements of matrix $\boldsymbol{B}$, can be used to plot $\boldsymbol{C_i}$ versus $\boldsymbol{i}$ (index) as a diagnostic technique to evaluate the existence of influential observations. The plot $\boldsymbol{|L_{max}|}$ versus $\boldsymbol{i}$, where $\boldsymbol{|L_{max}|}$ is the first normalized eigenvector associated with the largest in-module eigenvalue of matrix $\boldsymbol{B}$, may also be used as a diagnostic measure of local influence for detecting influential points.

In this work, the exponential, Gaussian and Matérn family spatial models were used to fit the covariance structure using the ML method of parameter estimation (Mardia and Mashall, 1984; Zimerman and Zimmerman, 1991; Lark, 2000a, 2000b, 2002). The spatial model was chosen by the cross-validation technique and the log likelihood maximum value (*LMV*) (Faraco *et al*., 2008). To investigate the existence of influential points, a diagnostic analysis using the local influence technique was applied using the software R (R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria) version 3.0 and the module geoR (Ribeiro Jr and Diggle, 2013).

Results and discussion

Graphic techniques for local influence diagnostics were applied to the data in this study to detect influential points, which can influence the parameter estimates that define the spatial dependence structure and construction of the matic maps. The results of these analyses are presented in the graphs of coefficients $\boldsymbol{C_i}$ versus $\boldsymbol{i}$ (index)

and $|L_{max}|$ versus i (index), considering the additive perturbation (Figures 1 and 2) and the Zhu perturbation in the response variable (Figures 3 and 4).

Table 1 shows the results of local influence analyses for the response variable using both perturbations chemes. Considering the graphs C_i versus i and $|L_{max}|$ versus i, the analyses revealed the presence of influential points. These points were deleted, and we reanalyzed the new data to verify their influence on the spatial variability structure and construction of thematic maps. For the C content, observation 32 (33.12 g dm^{-3}) was identified as an influential point according to the additive and Zhu perturbations. For the Ca content, observation 46 (11.76 cmol$_c$ dm^{-3}) was identified as influential under the additive perturbation and observation 20 (6.31 cmol$_c$ dm^{-3}) under the Zhu perturbation. For the K content, observation 34 (0.15 cmol$_c$ dm^{-3}) was identified as influential under the additive and Zhu perturbations. For the Mg content, observation 62 (4.15 cmol$_c$ dm^{-3}) was identified as influential under the additive perturbation and observation 79 (2.68 cmol$_c$ dm^{-3}) under the Zhu perturbation. For the Mn content, observation 66 (99.0 mg dm^{-3}) was identified as influential

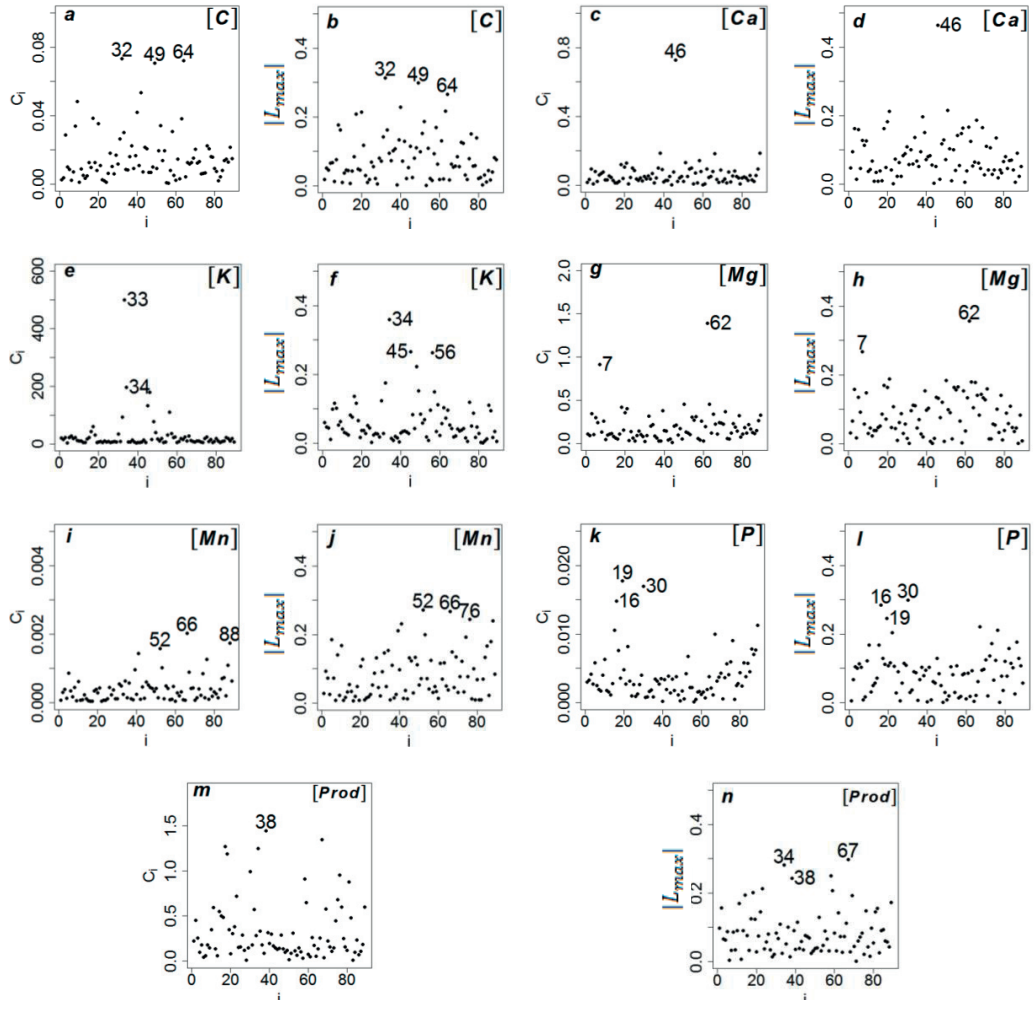

Figure 1. Graphs of the diagnostics C_i vs *i* and $|L_{max}|$ vs *I* for the georeferenced variables using the additive perturbationin the response variable $Z_\omega = Z + \omega$. Prod: Yield (Productivity).

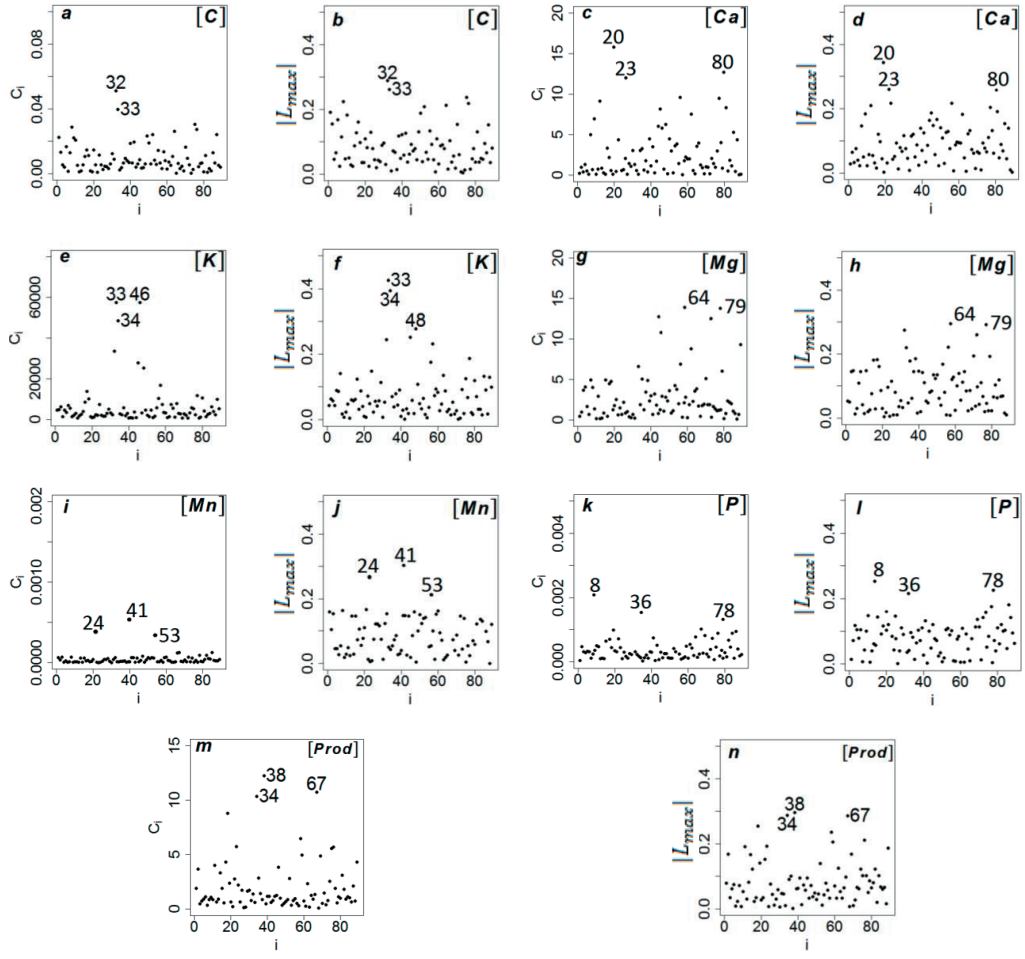

Figure 2. Graphs of the diagnostics C_i vs *I* and $|L_{max}|$ vs *I* for the georeferenced variables using the Zhu perturbation in the response variable $Z_\omega = Z + \Sigma^{-1/2}\omega$. Prod: Yield (Productivity).

under the additive perturbation and observation 41 (58.0 mg dm^{-3}) under the Zhu perturbation. For the P content, the techniques revealed several influential points representing the largest observations among them, corresponding to 30 (24.60 mg dm^{-3}) under the additive perturbation and 36 (13.0 mg dm^{-3}) under the Zhu perturbation. Finally, for soybean yield (Prod), observation 38 (5.13 t ha^{-1}) was identified as influential under the additive and Zhu perturbations.

Table 2 presents descriptive statistics for the variables in this study, analyzed with and without the influential observations detected by both perturbation schemes. Variations in the dispersions and in the homogeneity of the data were observed. The variables Ca, K, Mg, Mn and P had greater dispersions and heterogeneity in the data because they had coefficients of variation CV%> 30%.

Table 3 presents the fitted models and parameters that were estimated by the (ML) method for the models that were selected by the log likelihood maximum value (LMV) and by cross-validation criterion (Table 4) for the original data and for the data without the identified influential points. The estimates of the nugget effect φ_1 in all of the fitted models without the influential observations had reduced values, except for P, which increased slightly when analyzed without observation 36, which was detected by Zhu perturbation (P

Table 1. Influential points detected by the analysis of local influence.

	Additive perturbation $Z_{\omega} = Z + \omega$		Zhu perturbation $Z_{\omega} = Z + \omega^{-1/2}\omega$	
Variables	C_i vs i	$\lvert L_{max} \rvert$ vs i	C_i vs i	$\lvert L_{max} \rvert$ vs i
C	32	32	32	32
Ca	46	46	20	20
K	34	34	34	34
Mg	62	62	79	79
Mn	66	66	41	41
P	30	30	36	36
Prod	38	38	38	38

Table 2. Descriptive statistics of soil properties and yield with and without the influential points.

Variable	n	Min	Average	Max	Q1	Q2	Q3	DP	Var	CV(%)
C	89	19.87	27.18	34.29	25.32	26.88	29.61	3.27	10.67	12.02
C without 32 (A)(Z)	88	19.87	27.11	34.29	25.13	26.88	29.61	3.22	10.38	11.88
Ca	89	2.37	5.22	11.76	4.17	5.05	6.13	1.41	1.99	27.02
Ca without 46 (A)	88	2.37	5.14	8.06	4.16	5.04	6.12	1.23	1.51	23.93
Ca without 20 (Z)	88	2.37	5.21	11.76	4.16	5.05	6.12	1.41	2.00	27.15
K	89	0.08	0.19	0.60	0.14	0.19	0.23	0.081	0.007	41.18
K without 34 (A)(Z)	88	0.08	0.20	0.60	0.14	0.19	0.23	0.080	0.010	41.22
Mg	89	0.86	2.21	4.15	1.68	2.07	2.66	0.68	0.47	30.87
Mg without 62 (A)	88	0.86	2.18	3.72	1.66	2.06	2.65	0.65	0.42	29.87
Mg without 79 (Z)	88	0.86	2.21	4.15	1.66	2.07	2.65	0.68	0.47	31.04
Mn	89	17.00	50.47	107.0	37.00	43.00	62.00	19.96	398.25	39.54
Mn without 66 (A)	88	17.00	49.92	107.0	37.00	43.00	60.50	19.37	375.45	38.81
Mn without 41 (Z)	88	17.00	50.39	107.0	37.00	43.00	62.00	20.05	402.17	39.80
P	89	2.90	10.71	24.60	7.00	10.10	13.20	5.27	27.74	49.15
P without 30 (A)	88	2.90	10.56	24.30	6.97	10.00	13.05	5.08	25.92	48.20
P without 36 (Z)	88	2.90	10.69	24,60	6.98	10.00	13.22	5.29	28.00	49.51
Prod	89	2.12	3.56	5.13	3..23	3.55	3.85	0.54	0.29	15.10
Prod. without 38 (A) (Z)	88	2.12	3.55	4.94	3.23	3.55	3.83	0.51	0.26	14.50

n: number of data points; Min: minimum value; Max: maximum value; Q1: first quartile; Q2: median; Q3: third quartile; SD: standard deviation; CV: coefficient of variation; C (g dm^{-3});Ca ($cmol_c$ dm^{-3}); K ($cmol_c$ dm^{-3}); Mg ($cmol_c$ dm^{-3});Mn (mg dm^{-3}); P (mg dm^{-3}); Prod. (t ha^{-1}); (A) data without the influential point determined by the additive perturbation; (Z) data without the influential point determined by the Zhu perturbation.

without 36 (Z)). Regarding the estimate of the parameter that defines φ_3, the variables C, Ca, K and Prod showed decreases in the estimates without considering the influential observation. The variables Mg and P showed increases in the estimates when they were obtained without the influential observations that were revealed by Zhu perturbation (Mg without 79 (Z) and P without 36 (Z)) and decreases when using the additive perturbation (Mg without 62 (A) and P without 30 (A)); the opposite effects were observed for the Mn estimates.

Except for soybean yield (Prod), which has a weak spatial dependence and almost pure nugget effect, *i.e.*, lack of spatial dependence, the analyzed at-

Table 3. Estimated parameters and standard errors (in parentheses)for the models fitted to the spatial linear model.

Variable	Model	$\hat{\beta}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{a}$	E (%)	LMV
C	Gaussian	27.134 (0.651)	5.194 (1.106)	5.441 (2.126)	217.72 (0.113)	376.82 (0.196)	48.84	-221.4
C without 32	Gaussian	27.082 (0.645)	4.513 (1.0142)	5.766 (2.116)	210.95 (0.102)	365.12 (0.177)	43.91	-220.3
Ca	Gaussian	5.431 (0.419)	1.446 (0.234)	0.663 (0.417)	527.13 (1.591)	912.37 (2.756)	68.58	-150.1
Ca without 46	Exponential	5.343 (0.481)	0.877 (0.272)	0.6610 (0.526)	470.00 (4.790)	1408.00 (14.371)	57.05	-133.8
Ca* without 20	Exponential	5.41 (0.442)	1.120 (0.345)	0.970 (0.527)	370.00 (3.088)	1108.43 (9.265)	53.59	-147.7
K	Gaussian	0.1985 (0.008)	0.0061 (0.002)	0.0005 (0.003)	77.98 (0.042)	134.98 (0.074)	92.42	96.92
K* without 34	Gaussian	0.200 (0.009)	0.0040 (0.004)	0.0030 (0.004)	77.53 (0.085)	134.18 (0.147)	57.14	95.7
Mg	Exponential	2.250 (0.184)	0.349 (0.066)	0.116 (0.089)	480.01 (5.489)	1437.97 (16.467)	75.07	87.69
Mg without 62	Exponential	2.241 (0.174)	0.283 (0.062)	0.145 (0.0941)	360.24 (3.002)	1079.18 (9.008)	66.05	81.85
Mg* without 79	Exponential	2.248 (0.18)	0.3534 (0.06)	0.1147 (0.089)	475.01 (5.469)	1422.99 (16.41)	75.49	-87.21
Mn	Gaussian	53.41 (6.253)	184.85 (5.046)	235.56 (1.112)	184.32 (0.986)	319.03 (1.709)	43.97	-375.4
Mn without 66	Gaussian	53.18 (6.384)	171.09 (4.521)	220.20 (1.053)	206.44 (1.237)	357.30 (2.142)	43.72	-366.6
Mn* without 41	Matérn k=1.0	53.70 (6.315)	166.96 (3.792)	257.65 (1.149)	90.06 (0.536)	427.25 (2.543)	39.32	-371.1
P	Exponential	11.381 (1.286)	20.002 (3.341)	8.393 (4.967)	400.00 (0.729)	1198.31 (2.188)	70.44	-270.1
P without 30	Gaussian	10.875 (1.064)	18.999 (3.315)	7.333 (4.278)	320.14 (0.409)	554.10 (0.708)	72.15	-263.6
P* without 36	Exponential	11.490 (1.573)	20.580 (4.615)	8.370 (5.950)	500.00 (0.855)	1497.88 (2.566)	71.09	-267.3
Prod.	Matérn k=1.5	3.573 (0.064)	0.2663 (0.054)	0.0199 (0.040)	80.838 (0.016)	383.489 (0.076)	93.05	-70.47
Prod without 38	Gaussian	3.558 (0.057)	0.000 (0.134)	0.2610 (0.147)	56.610 (0.001)	97.980 (0.002)	0.00	-64.6

$\hat{\beta}$: Estimated average data; $\hat{\varphi}_1$: Estimated nugget effect; $\hat{\varphi}_2$: Estimated contribution; $\hat{\varphi}_3$: function of the estimated range; $\hat{a}$: Estimated range (m); E = $\hat{\varphi}_1/(\hat{\varphi}_1+\hat{\varphi}_2)$ x 100 Relative nugget effect; *LMV*: Log likelihood maximum value.

tributes have medium spatial dependence (25% <E <75%, Cambardella *et al*., 1994). It can be concluded that the influential points that were identified under both perturbation schemes influence the choice of the best fitted spatial model and parameter estimation.

Using the kriging interpolation technique, which considers the spatial characteristics of the chosen model (model and parameter estimates $\hat{\varphi}_1$, $\hat{\varphi}_2$ and $\hat{\varphi}_3$), thematic maps for the soil chemical properties and soybean yield were creased, with and without the observations that were considered influential under the additive and Zhu perturbations.

Figures 3a, 3b and 3c show the thematic maps of the C content with and without the influential points. According to the classification of Oliveira (2001), the carbon content in the soil is considered high in the range of 20.01 to 35.00 g dm^{-3}, indicating that the entire study area has high levels. Figures 3a, 3b and 3c, with and without the influential points, did not influence a decrease in the C content.

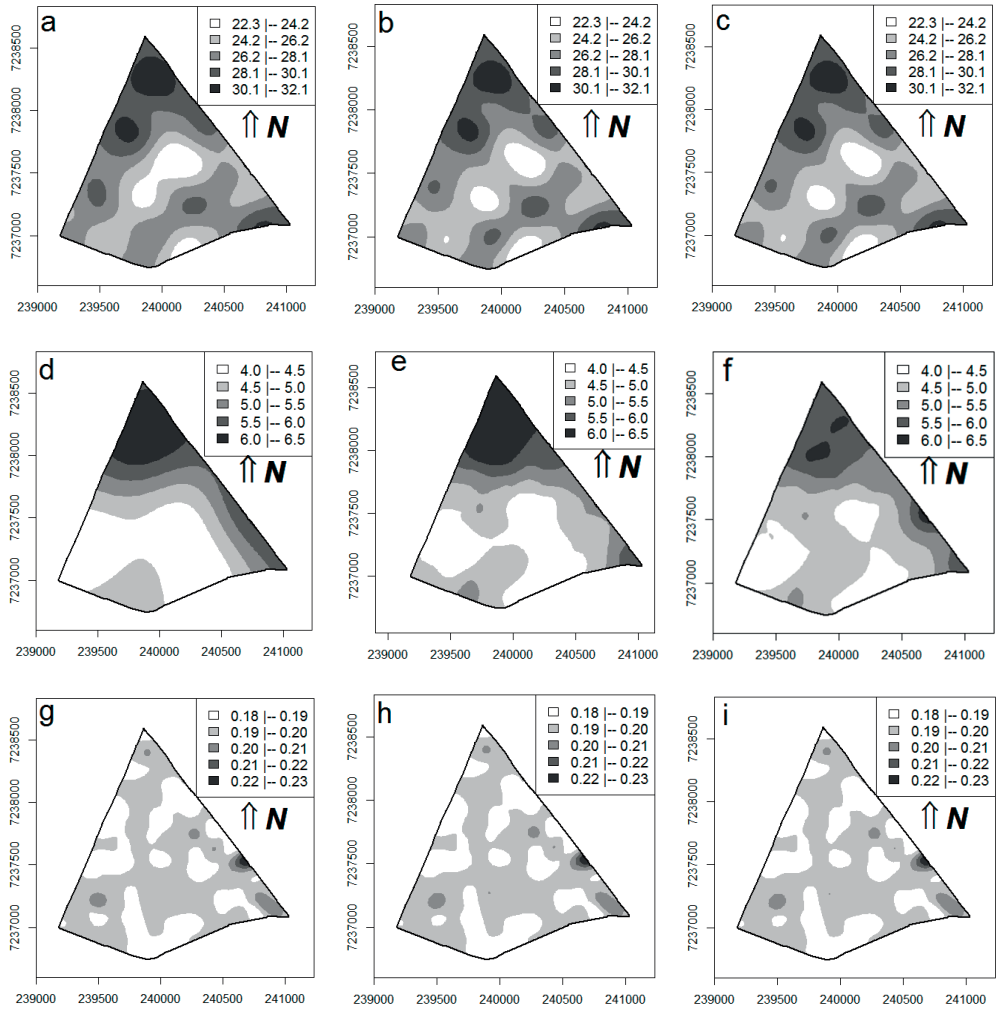

Figure 3. Thematic map of the C content with (a) and without point 32, which was identified as influential under the additive (b) and Zhu perturbations and (c);Thematic map of Ca content with all points (d), without point 46, which was identified as influential under the additive perturbation (e), and without point 20, which was identified as influential under the Zhu perturbation (f);Thematic map of the K content with all points (g) and without point 34, which was identified as influential under the additive (h) and Zhu perturbations (i).

Figures 3d, 3e and 3f show the thematic maps of the Ca content with and without the influential points. There are differences in the maps in the class with lower levels of Ca, located in the central part of the study area. As shown in Figure 3f, when observation 20, defined as influential by the Zhu perturbation, is removed, there is a class change in the northern part of the map, decreasing the Ca content in the soil from 6.0 to 6.5 $cmol_c\ dm^{-3}$ to 5.5 to 6.0 $cmol_c\ dm^{-3}$. Despite these changes, the entire area was ranked with a high level of Ca in the soil (above 4.00 $cmol_c\ dm^{-3}$), which according to Oliveira (2001), is essential for the production of soybean.

According to the maps in Figures 3g, 3h and 3i, the K content nutrient levels are considered medium because the values are between 0.11 and 0.30 $cmol_c\ dm^{-3}$ (Oliveira, 2001). The thematic maps of the K content do not present standards in the nutrient levels due to weak spatial dependence.

In Figures 4a, 4b, and 4c, the highest levels of the Mg content are located in the northern part of the area. According to the classification of Oliveira (2001), the Mg content is considered high when it is above 0.80 $cmol_c\ dm^{-3}$. Few changes are observed any of the classes when analyzed without the influential points (Figures 4b and 4c). It can be concluded that the Mg content influential points detected by both perturbations did not affect the construction of the thematic map.

Figures 4d, 4e, and 4f show the thematic maps of the Mn content with and without the influential points; there is little difference between them. The highest levels of the Mn content are in the northern area, and the largest percentage of the area is where the Mn content is below 40 mg dm^{-3}, which is classified as good according to Oliveira (2001) and does not impair the development of soybean.

Figures 4g, 4h and 4i show the thematic maps of the P content with and without the influential points; differences are evident between the maps. The P content is considered too high for the soybean crop when the content is above 9.00 mg dm^{-3} (Oliveira, 2001).

Finally, Figures 4j, 4k and 4l show the thematic maps of soybean yield (Prod) with and without the influential points. On the map with all of the influential points (Figure 4j), the highest levels of Prod are located in the eastern part of the area. When the influential point is removed (Figure 4k and 4l) and the resulting map is compared with the yield map containing all of the influential points, the entire area becomes homogeneous from 3.5 to 4.0 t ha^{-1} due to the little spatial dependence of the data. The influential points interfere in the building of the thematic map of soybean yield.

To compare the thematic maps of the soil properties and soybean yield with and without the influential points that were identified by the addictive and Zhu perturbations (Figures 3 and 4), the accuracy indices for Global Accuracy (GA) and Kappa (De Bastiani *et al*., 2012) were calculated using the error matrix presented in Table 4. The variables Ca, P and Prod have low levels of accuracy, indicating little similarity among the maps with and without the influential points. Thus, we can conclude the interference of the influential points in the construction of the thematic maps of some of the soil properties and soybean yield in the study area.

We conclude that the local influence technique was efficient in identifying influential points for the analyzed variables. In some cases, both schemes of perturbation identified the same observation, but in other situations, the schemes presented different observations; therefore, we recommend using the Zhu perturbation because it adds information of the structure of spatial dependence. In the construction of the thematic maps, changes in the levels of nutrients were

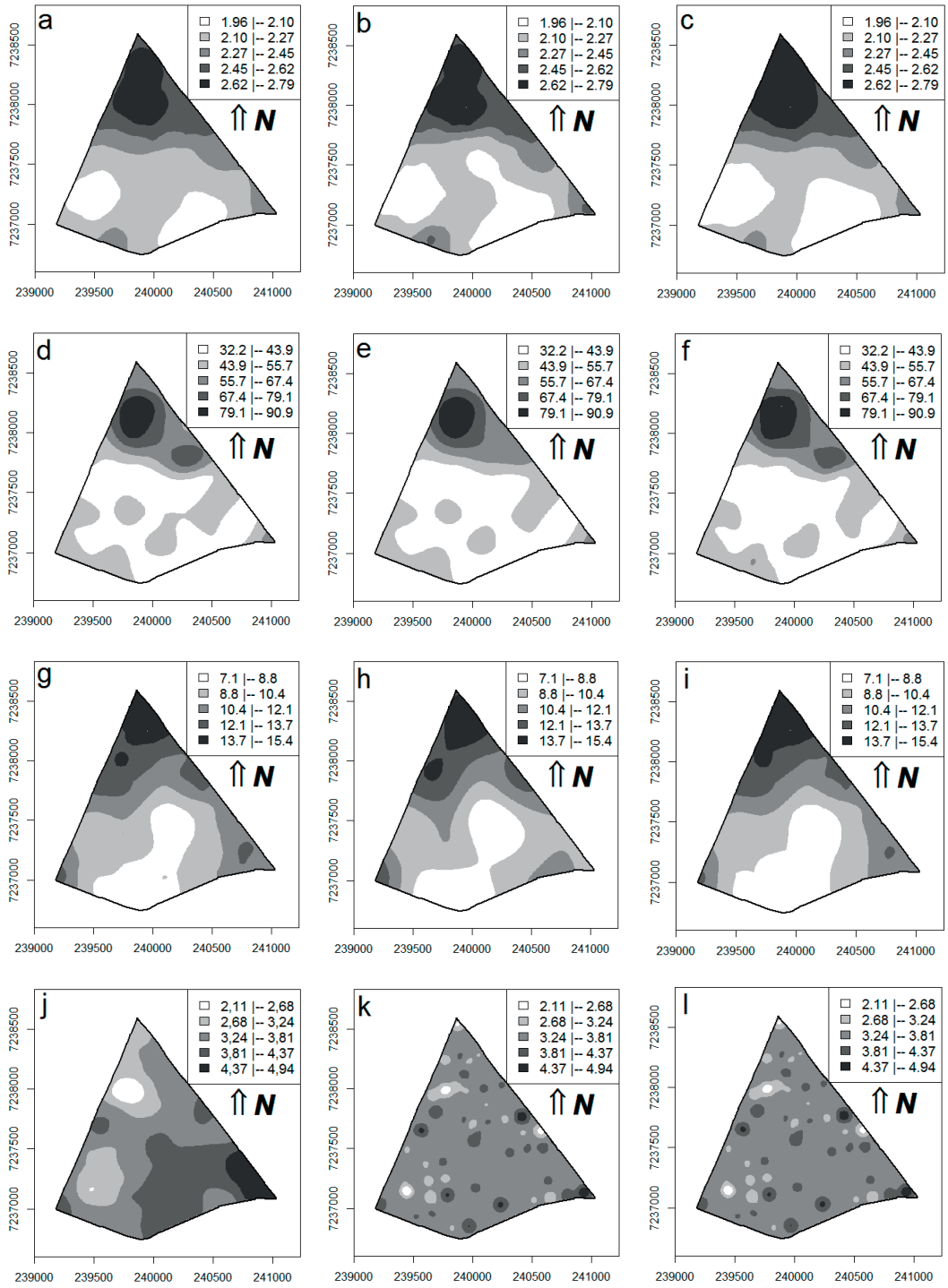

Figure 4. Thematic map of the Mg content with all points (a), without point 62, which was identified as influential under the additive perturbation (b), and without point 79, which was identified as influential under the Zhu perturbation (c); Thematic map of the Mn content with all points (d), without point 66, which was identified as influential under the additive perturbation €, and without point 41, which was identified as influential under the Zhu perturbation (f); Thematic map of the P content with all points (g), without point 30, which was identified as influential under the additive perturbation (h),and without point 36, which was identified as influential under the Zhu perturbation (i); Thematic map of yield (Prod) with all points (j), without point 38, which was identified as influential under the additive perturbation (k),and without point 38, which was identified as influential under the Zhu perturbation (l).

Table 4. Global Accuracy and Kappa Indices.

Variables	Additive perturbation $Z_\omega = Z + \omega$		Zhu perturbation $Z_\omega = Z + \Sigma^{-1/2}\omega$	
	GA	Ka	GA	Ka
C	0.92	0.89	0.92	0.89
Ca	0.58	0.43	0.52	0.38
K	0.93	0.90	0.93	0.90
Mg	0.72	0.63	0.90	0.86
Mn	0.93	0.90	0.90	0.85
P	0.68	0.58	0.90	0.86
Prod	0.35	0.14	0.35	0.14

GA: Global accuracy index; Ka: Kappa index (De Bastiani *et al*., 2012).

observed when working with and without the influential points, providing a differentiated application of inputs. The application of the local influence technique should be part of geostatistical analysis.

Acknowledgements

We acknowledge partial financial support from Capes, CNPq and Facepe, Brazil, and FONDECYT 1110318, Chile.

Resumen

D.M. Grzegozewski, M.A. Uribe-Opazo, F. De Bastiani y M. Galea. 2013. influencia local a modelos espaciales lineales Gaussianos: Aplicación a la agricultura. **Cien. Inv. Agr. 40(3): 523-535.** Los valores discrepantes pueden afectar negativamente el ajuste de un modelo. El análisis de datos dependientes es diferente al de datos independientes. En el primer caso, envuelven datos espaciales que pueden tener valores discrepantes localmente y que tienen algunas características diferentes de los datos vecinos. En este artículo, el objetivo fue detectar los puntos influyentes por medio de la técnica de influencia local en la variable de respuesta, mediante el uso de dos esquemas diferentes de perturbaciones denominados: perturbación aditiva y perturbación de Zhu. Se aplicó esta técnica a las propiedades químicas del suelo y a la productividad de la soja. Se evaluaron los efectos de los puntos influyentes en la elección del modelo, en la estimación de parámetros de máxima verosimilitud y la construcción de mapas temáticos mediante "kriging". En la construcción de mapas temáticos, se pudo observar alteraciones en los niveles de nutrientes al realizar el estudio con y sin los puntos de influencia, de tal forma que permite una aplicación apropiada de los insumos, lo que genera un mayor ahorro para el productor y en la contribución a la protección del medio ambiente.

Palabras clave: Diagnóstico de influencia, geoestadística, máxima verosimilitud, valores discrepantes, variabilidad espacial.

References

Borssoi, J.A., F. De Bastiani, M.A. Uribe-Opazo, and M. Galea. 2011. Local influence of explanatory variables in Gaussian spatial linear models. The Chilean Journal of Statistics 2(2): 29-38.

Botinha, R.A., M.A. Uribe-Opazo, and M. Galea. 2011. Local influence for spatial analysis of soil

physical properties and soybean yield using student's t-distribution. Revista Brasileira de Ciência do Solo 35:1917-1926.

Cambardella, C.A., T.B. Moorman, J.M. Novack, T.B. Parkin, D.L. Karlen, R.F. Turco, and A.E. Knopka. 1994. Field-scale variability of soil proprieties in central Iowa soils. Soil Science Society America Journal 58:1240-1248.

Christensen, R., W. Johnson, and L. Pearson. 1993. Covariance function diagnostics for spatial linear models. Mathematical Geology 25:145-160.

Cook, R.D. 1977. Detection of influential observations in linear regression. Technometrics 19:15-18.

Cook, R.D. 1986. Assessment of local influence Journal of the Royal Statistical Society. Series B Methodological 48 2:133-169.

Cressie, N., and D.M. Hawkins. 1980. Robust estimation of the variogram. Journal of the international Associate Mathematical Geology 12: 115-125.

De Bastiani, F., M.A. Uribe-Opazo, and G.H. Dalposso. 2012. Comparison of maps of spatial variability of soil resistance to penetration constructed with and without covariables using a spatial linear model. Engenharia Agrícola, Jaboticabal 32:394-404.

Embrapa. 2009. Sistema brasileiro de classificação de solos. Empresa Brasileira de investigación Agropecuária- Centro Nacional de Pesquisa de Solos (Embrapa-SPI). Rio de Janeiro. 412 pp.

Faraco, M. A., M.A. Uribe-Opazo, E.A. Silva, J.A. Johann, and J.A. Borssoi. 2008. 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. 32:463-476.

Genton, M.G. 1998. Spatial breakdown point of variogram estimators. Mathematical Geology 30: 213-221.

Lark, R.M. 2000a. Comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science 51:137-157.

Lark, R. M. 2000b. Estimating variograms os soil properties by method-of-moments and maximum likelihood. European Journal of Soil Science 51:717-728.

Lark, R.M. 2002. Otimized sampling of soil for estimation of the variogram by maximum likelihood. Geodema 105:49-80.

McBratney, A.B., and R. Webster. 1986. Choosing functions for semivariograms of soil properties and fitting them to sampling estimates. Journal of Soil Science 37: 617-639.

Mardia, K., and R. Marshall. 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71:135-146.

Oliveira, E.F. 2001. Treinamento: Fertilidade do solo e nutrição das plantas. COODETEC Cooperativa Central de Pesquisa Agrícola, maio. 44 pp.

Paula, G.A. 2010. Modelos de regressãocomapoio computacional. São Paulo - SP: Instituto de Matemática e Estatística (IME), Universidad de São Paulo. 233 pp.

Ribeiro Jr., P.J., and Diggle, P.J. 2001. geoR: A package from geoestatistical analysis. R. News 1:15-18. Available online at: http://cran.r-project.org/doc/Rnews/Rnews_2001-2.pdf (Website accessed: March 02, 2012).

Salas, C., L. Ene, T.G. Regoire, E. Naesset, and T. Gobakken. 2010. Modelling tree diameter from airborne laser scanning derived variables: A comparison of spatial statistical models. Remote Sensing of Environment 114:1277-1285.

Uribe-Opazo, M.A., J.A. Borssoi, and M. Galea. 2012.Influence diagnostics in Gaussian spatial linear models. Journal of Applied Statistics 39:615-630.

Webster, R., and M.A. Oliver. 2007. Geostatistics for Environmental Scientists. Willey. Second edition. John Wiley & Sons, Chichester. 315 pp.

Zhu, H., J.G. Ibrahim, S. Lee, and H. Zhang. 2007. Perturbation selection and influence measures in local influence analysis. The Annals of Statistics 35:2565-2588.

Zimmerman, D.L. and M.B. Zimmerman. 1991. A compararison of spatial semivariogram estimators and corresponding ordinary kriging predictors.Tecnometrics 33:77-91.