Revista Brasileira de Ciência do Solo

Division - Soil in Space and Time | Commission - Pedometry

Slash Spatial Linear Modeling: Soybean Yield Variability as a Function of Soil Chemical Properties

Regiane Slongo Fagundes[1]*, Miguel Angel Uribe-Opazo[2], Luciana Pagliosa Carvalho Guedes[2] and Manuel Galea[3]

[1] Universidade Tecnológica Federal do Paraná, Colegiado de Matemática, Toledo, Paraná, Brasil.
[2] Universidade Estadual do Oeste do Paraná, Centro de Ciências Exatas e Tecnológicas, Programa de Pós-Graduação em Engenharia Agrícola, Cascavel, Paraná, Brasil.
[3] Pontifícia Universidad Católica de Chile, Departamento de Estadística, Macul, Santiago, Chile.

ABSTRACT: In geostatistical modeling of soil chemical properties, one or more influential observations in a dataset may impair the construction of interpolation maps and their accuracy. An alternative to avoid the problem would be to use most robust models, based on distributions that have heavier tails. Therefore, this study proposes a spatial linear model based on the slash distribution (SSLM) in order to characterize the spatial variability of soybean yields as a function of soil chemical properties. The likelihood ratio statistic (LR) was applied to verify the significance of parameters associated with the model. We evaluated the sensitivity of the maximum likelihood estimators by means of local influence analysis for both the soybean response and the linear predictor. In the proposed model, we analyzed data gathered from a commercial grain production area (127.18 ha) located in the western part of the state of Paraná (Brazil). The results showed that the slash distribution allowed us to adjust the high kurtosis of the data set distribution and the LR test confirmed that the soil chemical properties of phosphorus, potassium, pH, and organic matter were significant for the SSLM. Diagnostic analysis indicated that the atypical value of the sample set was not influential in the parameter estimation process. Construction of the interpolation map based on the proposed model is not affected when considering the atypical and/or influential observations. Thus, SSLM becomes a robust alternative in the study of soybean yield variability as a function of soil chemical properties, making it possible to investigate the productive potential of the areas.

Keywords: spatial variability, slash distribution, maximum likelihood, yield estimators.

*** Corresponding author:**
E-mail: regianefagundes@utfpr.edu.br

Received: January 31, 2017

Approved: August 31, 2017

How to cite: Fagundes RS, Uribe-Opazo MA, Guedes LPC, Galea M. Slash spatial linear modeling: soybean yield variability as a function of soil chemical properties. Rev Bras Cienc Solo. 2018;42:e0170030.
https://doi.org/10.1590/18069657rbcs20170030

INTRODUCTION

Knowledge of spatial variability of soybean yield and its relationship to soil chemical properties are essential for proper crop management (Sobjak et al., 2016). However, many precision agriculture users get disappointed trying to find the ideal variable-rate application of the nutrient based on the prescription map, because does not always correspond to the soybean yield map generated after the intervention. One of the main explanations is related to complexity of the soil, which is considered a dynamic system whose functionality arises from interactions between physical, chemical, and biological components that are specific for each crop (Pereira et al., 2016). Thus, a localized management system for soybean requires accurate information on spatial variation and the interaction between soil chemical properties and their relationship to yield (Pagani and Mallarino, 2015; Al-Kaisi et al., 2016; Dalla Nora et al., 2017).

One of the methods used in such characterization is geostatistics, which takes soil spatial variation patterns into account and provides techniques that enable construction of maps associated with one or more soil chemical properties (Cressie, 2015).One of the different methods used in geostatistical studies is spatial linear models, which have been widely evaluated by assuming a Gaussian stochastic process (Uribe-Opazo et al., 2012; Grzegozewski et al., 2013; Nesi et al., 2013; De Bastiani et al., 2015). This modeling enables estimation of spatial dependence parameters through the maximum likelihood method (ML), making inferential studies possible.

However, Assumpção et al. (2014) and Schemmer et al. (2017) reported on the sensitivity of a Gaussian distribution for atypical values and errors with heavier tails than normal and stated that such modeling may generate unrealistic maps. Thus, one of the alternatives to avoid this problem would be to use more robust models based on heavier tails. Following this line of reasoning, multivariate slash distribution is quite attractive because it has an additional parameter, here designated as η ($0 < \eta < 1$) which allows kurtosis adjustment, making the modeling more flexible in the presence of atypical values (Osorio et al., 2009; Alcantara and Cysneiros, 2013). This distribution was first introduced by Lange and Sinsheimer (1993), belonging to the class of scale mixtures of normal distribution. The basic idea behind this distribution class is to insert randomness in the covariance matrix, as well as in the mean vector of the multivariate normal distribution, through a strictly positive mixture variable. This allows generalization of a multivariate normal distribution, preserving the main properties.

Although it creates robust models, spatial modeling based on the slash distribution may be affected by influential observations. Therefore, it is important to evaluate its sensitivity through influence diagnostics, which may be carried out by different evaluation methods, one of them being local influence assessment. The aim of this assessment is to evaluate the goodness of fit of the model, and the robustness of its estimates when small perturbations are introduced in the model and/or in the dataset (Cook, 1986; Zhu and Lee, 2001; Jonathan et al., 2016).

Regarding georeferenced data, Uribe-Opazo et al. (2012) evaluated the sensitivity of covariance function estimators and linear predictors under small dataset perturbations and/or a spatial linear model with a normal distribution. Assumpção et al. (2011) presented techniques for local influence for spatial analysis of soil physical properties and soybean yield using Student's t-distribution. Grzegozewski et al. (2013) considered a Gaussian spatial linear model (GSLM) and information about soil macro- and micronutrients in evaluating the effects of influential observations on spatial model selection, parameter estimation by maximum likelihood, and characterization of spatial continuity of soybean yield. In these studies, the authors emphasized that influential values can modify the interpolated maps and may generate inaccurate predictions.

The aim of this study was to propose a spatial linear model based on the slash distribution (SSLM) to characterize the spatial variability of soybean yields as a function of soil chemical properties.

MATERIALS AND METHODS

Experimental site

We gathered data from a field experiment set up in a commercial grain production area in which the no-tillage system has been practiced since 1994. The area is located in the municipality of Cascavel, in the west of the state of Paraná, Brazil. It lies at the geographical coordinates of approximately 24.95° S, 53.57° W, with an average altitude of 650 m. Soil at the location is classified as a *Latossolo Vermelho Distróférrico* (Santos et al., 2013), which corresponds to Rhodic Hapludox (Soil Survey Staff, 2014), with an average slope of around 4 %. The temperate climate is mesothermal and highly humid, classified as *Cfa* (Köppen system); and mean annual temperature is 21 °C. Samples were taken during the 2014/2015 crop season in a 127.18 ha area. For this study, we used a lattice plus close pairs design (Diggle and Ribeiro Junior, 2007), with a distance of 141 m between points belonging to the regular grid and, in some random locations, shorter distances of 75 and 50 m between points, thereby obtaining 78 locations. All the collected data was georeferenced by a GPS (Global Positioning System) receiver, Geoexplorer® 3 (Trimble®), under a UTM coordinate system, zone 22 south and datum WGS 84 (Figure 1).

The response variable considered in the model was soybean yield, which was estimated by the quantity of grains harvested in an area of 0.90 m² centered on each georeferenced point. After harvest, the grain from each plot was weighed and moisture level was corrected to 13 %; this weight was then converted into t ha^{-1}. As explanatory variables (covariates), we selected P (mg dm^{-3}), K (mg dm^{-3}), pH, and organic matter (OM) (g dm^{-3}). Five samples were collected at a depth of 0.00-0.20 m to determine the chemical contents.

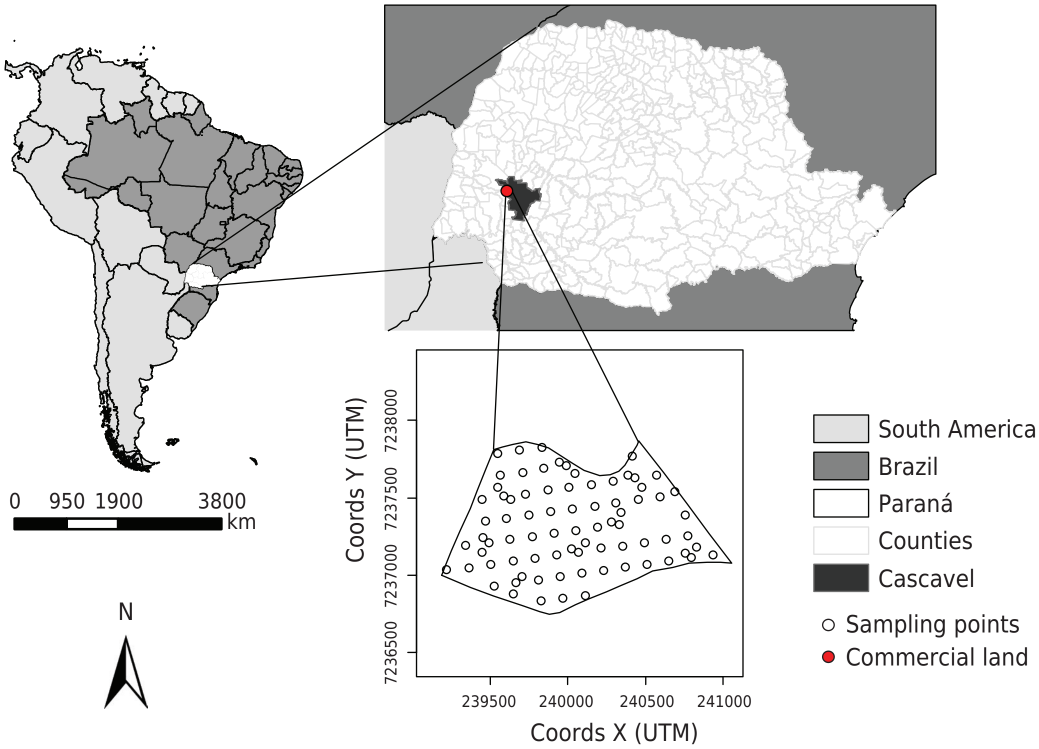

Figure 1. Location map of the study site.

These samples were taken near each georeferenced point and were then mixed in order to have a representative composed sample for each plot. Analysis was carried out in the laboratory of the *Cooperativa Central de Pesquisa Agrícola* (Coodetec, Brazil), whose methodology is available in Costa and Oliveira (2001).

These covariates were chosen based on the following considerations: (i) high kurtosis associated with indexes outside the range of -2 to +2 (Casella and Berger, 2010); (ii) structure of spatial dependence of soil chemical properties, identified from the construction of omnidirectional experimental semivariograms using the Matheron estimator (Soares, 2014; Cressie, 2015); (iii) P increases crop yield at early reproductive stages and is applied in large amounts since Brazilian soils are poor in this element; (iv) K plays an important role in plant photosynthesis and respiration, assists in formation of starch and sugars, and produces vigorous and resistant plants; (v) Brazilian soils are acidic (low pH), which limits crop yield; and (vi) OM provides nutrients for sustainable crop production and increases water infiltration into the soil, reducing erosion (Santos et al., 2013).

Spatial modeling and parameter estimation

In order to build the data model from the spatially correlated variables, we considered a stochastic process $\{Y(s_i), s_i) \in S\}$ defined in a region $S \subset \mathcal{R}^2$, where each element $Y(s_i)$ related to soybean yield has known locations s_i, $i = 1, ..., n$. We also consider the process as stationary, wherein $\boldsymbol{Y} = (Y(s_i), ..., Y(s_n))^T$ follows a multivariate slash distribution (Lange and Sinsheimer, 1993) in which the second moment is finite, being that $\boldsymbol{Y} \sim SL_n(\boldsymbol{X\beta}, \boldsymbol{\Sigma}, \eta)$, where $E(Y) = \boldsymbol{X\beta}$ and $Cov(Y) = \boldsymbol{\Sigma}$. In addition, $\boldsymbol{X}$ is a matrix formed by the vector **1'**s and the covariates P, K, pH, and OM; $\boldsymbol{\beta} = (\beta_0, ..., \beta_q)^T$, the vector of unknown parameters associated with each covariate; and η, the parameter of kurtosis adjustment. Under these conditions, the SSLM was expressed by equation 1:

$$Y(s_i) = \beta_0 + \beta_1 P(s_i) + \beta_2 K(s_i) + \beta_3 pH(s_i) + \beta_4 OM(s_i) + \varepsilon(s_i) \qquad \text{Eq. 1}$$

in which $\varepsilon(s_i)$ is the random $[\boldsymbol{\varepsilon} \sim SL_n(\boldsymbol{0}, \boldsymbol{\Sigma}, \eta)]$.

Spatial dependence was determined by the covariance matrix $\boldsymbol{\Sigma}$, whose parametric form is given by $\boldsymbol{\Sigma} = \phi_1 \boldsymbol{I_n} + \phi_2 \boldsymbol{R}(\phi_3)$, where $\boldsymbol{I_n}$ is the identify matrix; $\phi_1 \geq 0$ is the nugget effect; $\phi_2 \geq 0$ is the partial sill; ϕ_3 is the parameter related to the spatial dependence radius (range); and $\boldsymbol{R}(\phi_3) = [(r_{ij})]$ is the correlation matrix for the soybean yield dataset observed between the points located at s_i and at s_j (Uribe-Opazo et al., 2012).

We considered the Matérn family covariance function (Matérn, 1986) to explain spatial dependence with the smoothness parameter $\kappa \in \{0.5, 1.5, 2.5\}$, whose relationship to the practical range is given by $3\ \phi_3$, $4.75\ \phi_3$, and $5.92\ \phi_3$, respectively (Diggle and Ribeiro Junior, 2007).

We estimate the unknown vector of parameters of the model $\boldsymbol{\theta} = (\boldsymbol{\beta^T}, \boldsymbol{\phi^T})^T$, with $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)^T$, by maximizing the log-likelihood function (ML), whose form is expressed by the equation 2:

$$l(\boldsymbol{\theta}) = -\log(\eta) + \frac{n}{2} \log\left(\frac{c(\eta)}{2\pi}\right) - \frac{1}{2} \log|\boldsymbol{\Sigma}| + \log\left(\int_0^1 v^{a-1} \exp\{-vb\} dv\right) \qquad \text{Eq. 2}$$

in which $a = (n/2 + \eta^{-1})$ and $b = c(\eta)\delta/2$, with $\delta = (\boldsymbol{Y} - \boldsymbol{X\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X\beta})$ and $c(\eta) = (1 - \eta)^{-1}$, for $0 < \eta < 1$, where η stands for the parameter of kurtosis adjustment. Here, the ML method defines the estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ as being the vector that maximizes $l(\boldsymbol{\theta})$ in the parametric space $\boldsymbol{\Theta}$.

The ML estimator was calculated using an expectation-maximization (EM) algorithm, which consists of a computational method that generates approaches iteratively by maximizing expectation of the logarithm of the likelihood function for the completed data set, called the Q-function, which is expressed by the equation 3:

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = -\log(\eta) + \frac{n}{2}\log\left(\frac{c(\eta)}{2\pi}\right) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}w\delta + \left(\frac{n}{2} + \frac{1}{\eta} - 1\right)d, \qquad \text{Eq. 3}$$

in which $w = c(\eta)E\{V|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}\}$ and $d = E\{\log(V|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}\}$. Here, $V \sim Beta(\eta^{-1}, 1)$ and $(\boldsymbol{Y}|V = \nu) \sim N_n(\boldsymbol{X\beta}, c(\eta)\nu^{-1}\Sigma)$. For further details, see Lange and Sinsheimer (1993) and Osorio et al. (2009).

To avoid identifiability problems in ML estimations, the parameters of kurtosis adjustment (η) and smoothing (κ) of the Matérn family model were determined through cross-validation (CrV) (De Bastiani et al., 2015) and Trace criterion (Tr) (Kano et al., 1993).

Significance (hypothesis) testing on the estimated parameters β_s was carried out by the likelihood ratio statistic (LR), given by the equation 4:

$$LR = 2[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})], \qquad \text{Eq. 4}$$

where $l(\hat{\boldsymbol{\theta}})$ is the log-likelihood value of the model with all covariates and $l(\tilde{\boldsymbol{\theta}})$ is the log-likelihood value of the model without a certain covariable. At a significance level α, if $|LR| \geq \chi^2_{(k_1, \alpha/2)}$, we should reject the null hypothesis, meaning that the covariate tested was significant in the composition of the model (Dagenais and Dufor, 1991).

Local influence analysis

After defining the parameters, an interpolation map was built by regression-kriging in order to visualize soybean yield variability as a function of the covariates (P, K, pH, and OM) (Soares, 2014). To investigate the presence of observations that may have interfered in the process of interpolation, a local influence analysis was performed.

As proposed by Cook (1986), we examined possible influential observations on the soybean yield response variable by considering the likelihood displacement (Equation 2) after we insert a perturbation vector $\boldsymbol{\omega}$ (noise). We used the expression $LD(\omega) = 2[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}}_\omega)]$, where $l(\tilde{\boldsymbol{\theta}}_\omega)$ is the log-likelihood value of the model perturbed by $\boldsymbol{\omega}$. High values of $LD(\omega)$ means that $l(\hat{\boldsymbol{\theta}})$ and $l(\tilde{\boldsymbol{\theta}}_\omega)$ differ considerably, with the the i-th influential observation being considered if $|h_{[L]maxi}| > \bar{D} + 2sd(D)$, where $\bar{D}$ denotes the mean of the elements of the vector $|\boldsymbol{h}_{[L]max}|$, and $sd(D)$ is the standard deviation. Here, $|\boldsymbol{h}_{[L]max}|$ corresponds to a normalized unit eigenvector associated with the largest eigenvalue of the matrix $\boldsymbol{B}_L = \boldsymbol{\Delta}^T_{L\omega}[\boldsymbol{L}(\boldsymbol{\theta})^{-1}]\boldsymbol{\Delta}_{L\omega}$, in which $\boldsymbol{L}(\boldsymbol{\theta})$ is a Hessian matrix evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and $\boldsymbol{\Delta}_{L\omega} = \partial l^2(\boldsymbol{\theta},\boldsymbol{\omega})/\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}$ is a perturbation matrix (q + 4) × n evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_\omega$.

In a similar manner, influence analysis was performed on the Q-function (Equation 3), which was considered the reference measure $|\boldsymbol{h}_{[Q]max}|$. For more details, see Zhu and Lee (2001) and Assumpção et al. (2011, 2014).

To evaluate the influence on the linear predictor, we considered the methodology presented by Assumpção et al. (2011) applied to the SSLM (Equation 1). In this case, maximum influence directions were denoted by L_p and Q_p. Analysis was carried out using R software 3.3.2 for Windows (R Development Core Team, 2016).

RESULTS AND DISCUSSIONS

Table 1 shows the exploratory data analysis for the response variable and covariates, i.e., the soil chemical properties (P, K, pH, and OM). Average soybean yield was 2.37 t ha^{-1}, which is lower than the national average (3.03 t ha^{-1}) for the 2014/2015 season (IBGE, 2016). Standard deviation and coefficient of variation were 0.87 t ha^{-1} and 11.48 %, respectively, exhibiting low data dispersion around the mean and homogeneity (Carvalho et al., 2003).

Table 1. Descriptive statistics of soybean yield (Y) and of the soil chemical properties

Variable	Y	P	K	pH(H_2O)	OM
	t ha^{-1} [(1)]	mg dm^{-3}	mg dm^{-3}		g dm^{-3}
Mean	2.37	19.19	0.31	4.82	50.63
Standard deviation	0.07	123.56	0.02	0.17	40.05
Coefficient of variation (%)	11.48	57.91	45.35	8.43	12.50
Skewness	0.51	1.34	0.60	1.08	0.02
Kurtosis	3.11	4.73	2.44	4.31	2.44

[(1)] The units of measure refer only to the mean and standard deviation. P and K: extractor Mehlich-1; pH(H_2O): pH in water at a ratio of 1:2.5 v/v; OM: organic matter (Walkley and Black, 1934).

Quartile analysis indicated an atypical value of 3.18 t ha^{-1}, corresponding to sample 33. Standard analysis of the directional semivariograms (omitted here) towards 0°, 45°, 90°, and 135° showed similar behavior for all directions, indicating that the spatial dependence structure is isotropic.

Descriptive statistics of soil chemical properties (Table 1) indicate very high levels of P and K for soybean (Costa and Oliveira, 2001). Moreover, the standard deviation and coefficient of variation for both P and K suggested average dispersion around the mean and heterogeneity (Warrick and Nielsen, 1980). High variations in P and K contents may be due to continuing fertilization at a fixed rate along the plant row, affecting micro and macro spatial variability (Amado et al., 2009). Another factor may be the mobility of these chemical elements in the soil. For example, as a monovalent cation, K is easily leached absorbed, fixed, adsorbed, or stabilized in the soil solution, generating large variability. Therefore, these variations may reflect the effects of the complex interaction between soil chemical properties and crop management practices (Bottega et al., 2013; Pereira et al., 2016).

Soil pH was classified as highly acidic (from 4.31 to 5.00), with low dispersion around the mean. Organic matter contents, in turn, were rated as high (from 35.01 to 60.00 g dm^{-3}), with a coefficient of variation of 12.50 % (Warrick and Nielsen, 1980; Costa and Oliveira, 2001). Large amounts of OM and high acidity might be associated with the no-tillage system (NTS) implemented in the area since 1994. The amount of straw increases on the soil surface in areas under NTS for long periods, which shifts the quantity and quality of OM and gradually alters pH due to basic cations and soluble organic carbon (Dalchiavon et al., 2013). In addition, NTS may have limited the action of liming, restricting it to the areas of application, i.e., in the surface soil layers (Nunes et al., 2011; Paganiand Mallarino, 2015; Dalla Nora et al., 2017).

All variables showed high kurtosis (Table 1) (Casella and Berger, 2010). A process modeled under this condition may produce a model of interpolation with biased parameters and, consequently, generate overestimated/underestimated regions in the map. Our proposed model based on the slash distribution is able to reduce this impact by determining the shape parameter η, which will allow kurtosis adjustment of the data.

All the samples and the SSLM were considered in studying the dataset, assuming $\boldsymbol{Y}{\sim}SL_n\,(\boldsymbol{X\beta}, \boldsymbol{\Sigma}, \eta)$. For comparison, a Gaussian spatial linear model (GSLM) was also used, in which $\boldsymbol{Y}{\sim}N_n\,(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$. Thus, we could assess the model robustness and the effect of outliers and/or influential values on parameter estimates and mapping. Both criteria, CrV and Tr, indicated that for an SSLM, the adequate value for kurtosis adjustment was η = 0.25, and for the Matérn model, smoothing was κ = 0.25. For a GSLM, an adequate model for explaining spatial dependence was also the Matérn model, with a smoothing parameter of κ = 0.25.

Table 2 shows the parameter estimated by the ML estimator. The asymptotic standard errors (in brackets) were calculated from Fisher’s information matrix. It may be noted

Table 2. Values of parameters for kurtosis adjustment (η) and of smoothing (κ) selected from CrV (Cross-validation) and Tr (Trace) criteria; estimated values of parameters of the Gaussian spatial linear model (GSLM) and slash spatial linear model (SSLM) by maximum likelihood using the EM algorithm with respective asymptotic standard deviations (in brackets)

Model	Structure	Estimated parameter[1]							
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
GSLM	Matérn κ = 2.5	2.030 (0.412)[2]	0.003 (0.003)	-0.121 (0.228)	0.033 (0.076)	0.003 (0.005)	0.053 (0.059)	0.018 (0.059)	52.666 (0.007)
SSLM η = 0.25	Matérn κ = 2.5	1.993 (0.516)	0.003 (0.003)	-0.109 (0.288)	0.034 (0.095)	0.004 (0.006)	0.051 (0.036)	0.026 (0.013)	81.988 (0.006)

[1] β_i: parameters associated with the variable *i* = [soybean yield (Y), P, K, pH, and OM]; $\hat{\phi}_1$ nugget effect, $\hat{\phi}_2$ partial sill, $\hat{\phi}_3$ parameter that defines the spatial dependence radius. [2] The lowest values of the asymptotic standard deviations are underlined.

that the estimated values of the parameter vector β were similar in both models, with the lowest standard deviations observed for the GSLM, except for $\hat{\beta}_1$, which was equal.

It is noteworthy that the large values of coefficient of the parameters were associated with K ($\hat{\beta}_2$) content, but with a negative signal, and pH ($\hat{\beta}_3$) with a positive signal. This relationship may have contributed to the estimation of average soybean yield (t ha^{-1}) showing lower values than the national average, since an increase in soil acidity implies an increase of Al^{3+}, which inhibits root development and, consequently, decreases uptake of nutrients, such as K.

Another issue that may be related to pH is the low value of the $\hat{\beta}_1$ coefficient, since soil acidity decreases release of available P in organic matter of the soil (Santos et al., 2013; Passos et al., 2015).

Comparing values estimated for parameters related to the covariance of the spatial process, relevant differences were found, with the lowest standard deviations being registered in the SSLM. By calculating the relationship $\phi_1/(\phi_1 + \phi_2)$, we found that SSLM had an index of 0.66, while GSLM had an index of 0.75. The smoothing parameter was κ = 0.25 for both models, and this confirms that the practical range of SSLM was 485.37 m, whereas that of GSLM was merely 311.78 m. For Cambardella et al. (1994), the best estimates are reached when models are based on covariance functions with the lowest "nugget effect/sill" ratio and the highest range.

Using data from table 2, we were able to establish the GSLM of the soybean yield at s_i for the area under study, which was expressed by equation 5:

$$Y(s_i) = 2.030 + 0.003\,P(s_i) - 0.121\,K(s_i) + 0.033\,pH(s_i) + 0.003\,OM(s_i) + \varepsilon(s_i) \qquad \text{Eq. 5}$$

with $i = 1, \ldots, 78$ and the covariance matrix given by $\hat{\Sigma} = 0.053\,\boldsymbol{I}_{78} + 0.018\,\hat{\boldsymbol{R}}(52.666)$.

The SSLM was expressed by the equation 6:

$$Y(s_i) = 1.993 + 0.003\,P(s_i) - 0.109\,K(s_i) + 0.034\,pH(s_i) + 0.004\,OM(s_i) + \varepsilon(s_i) \qquad \text{Eq. 6}$$

with $i = 1, \ldots, 78$ and the covariance matrix given by $\hat{\Sigma} = 0.051\,\boldsymbol{I}_{78} + 0.026\,\hat{\boldsymbol{R}}(81.988)$.

In both cases, the correlation matrix elements $\boldsymbol{R}$ were determined by the correlation function of the Matérn family with κ = 0.25.

The hypothesis test to assess the significance of parameter vector $\boldsymbol{\beta}$ was applied jointly and individually for both models. Table 3 displays the results of the *LR* and respective p-values. The null hypothesis $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ was rejected at 5 % probability for both models. Individual testing also showed the parameter values of (β's) as significant. Therefore, the covariates P, K, pH, and OM remained in the modeling.

Local influence was diagnosed to evaluate the sensitivity of estimates from the model and from the predictive process to atypical and/or influential values. For GSLM, we used a perturbation scheme fitted to a normal distribution, as presented by De Bastiani et al. (2015). For the GSLM, the samples exerting influence on the response variables were 15 and 72 (Figure 2a). Conversely, for the linear predictor, the number of influencing samples was higher, namely, samples 15, 58, 65, and 72 (Figure 3a).

The figure regarding local influence on the response variable perturbing the SSLM shows that no observations were identified when considered measure $|\boldsymbol{h}_{[L]max}|$ (Figure 2b). However, sample 71 was influential for measure $|\boldsymbol{h}_{[Q]max}|$ (Figure 2c). Considering perturbation of the linear predictor, samples 1 and 71 were influential (Figure 3b) when measure L_p was used. And, for perturbation of the Q-function, measure Q_p, samples 1, 3, 62, and 71 were identified as influential (Figure 3c).

It is noteworthy that atypical sample 33 was not identified as influential on soybean yield response in any of the cases. These results corroborate those of Uribe-Opazo et al. (2012), who used a GSLM, tested different influence diagnostics techniques, and observed that not every atypical value could have an influence on model determination in a covariate study. Assumpção et al. (2014), based on Student's t-distribution, also found differences between the atypical and the influential values; they highlighted the relevance of the local influence method as opposed to a simple box-plot analysis. Furthermore, the smaller number of influential cases found here when SSLM is fitted is consistent with the results of De Bastiani et al. (2015); these authors used distributions with heavier tails and obtained greater modeling robustness.

Figure 4 displays the post-plot graph of the experimental area, highlighting the position of each influential sample. For GSLM (Figure 4a), the yield value of sample 72 was from

Table 3. Likelihood ratio statistic (*LR*) of the parameter vector **β** of the Gaussian spatial linear model (GSLM) and slash spatial linear model (SSLM) at 5 % probability

Hypothesis[(1)]	GSLM		SSLM	
	LR	p-value	LR	p-value
$\beta_1 = 0$	6.687	0.036*	11.659	0.004*
$\beta_2 = 0$	7.001	0.032*	15.215	0.001*
$\beta_3 = 0$	5.953	0.050*	8.567	0.016*
$\beta_4 = 0$	6.257	0.044*	9.113	0.013*
$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$	7.865	0.039*	13.897	0.003*

[(1)] Parameters associated with the variable: β_1 - P, β_2 - K, β_3 - pH, and β_4 - OM. *: significant at 5 % probability.

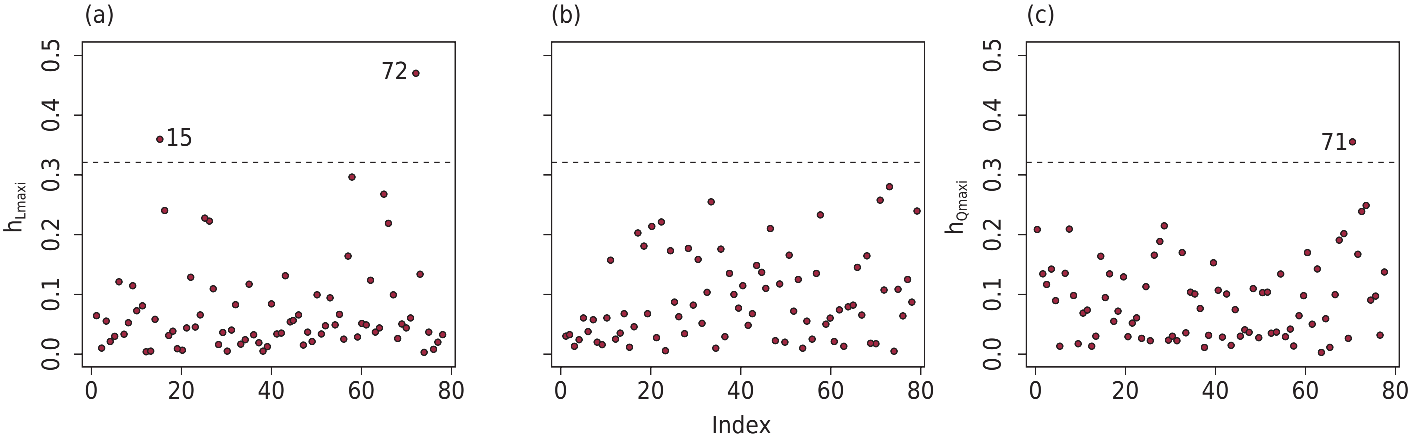

Figure 2. Graphs of local influence perturbing the response variable considering: (a) the Gaussian spatial linear model (GSLM) and measure $|\boldsymbol{h}_{[L]max}|$; (b) the slash spatial linear model (SSLM) and measure $|\boldsymbol{h}_{[L]max}|$; and (c) the slash spatial linear model (SSLM) and measure $|\boldsymbol{h}_{[Q]max}|$. Observations above the dotted line are classified as influential.

1.87 to 2.18 t ha^{-1} and the yield values of samples 58 and 65 were from 2.18 to 2.33 t ha^{-1}. Note that nearest observations of these samples showed soybean yield highs values. Sample 15 was found within a region with high values. When considering SSLM (Figure 4b), we found that the values of samples 1, 62, and 71 were from 2.55 to 3.18 t ha^{-1} and that of sample 3 was from 2.33 to 2.55 t ha^{-1}. Spatial analysis of the samples showed that neighboring observations to the left of sample 71 had values lower than the values of samples 1 and 3, which are close to the contour of the plotting domain.

In diagnostic analysis, when one or more influential values are detected, they are removed from the dataset to understand how their removal affects model selection, parameter estimates, and construction of the interpolation maps (Assumpção et al., 2011). For GSLM, samples 15 and 72 were removed from the dataset since they were identified as influential when applying measures $|\boldsymbol{h}_{[L]max}|$ and L_p. Using the same criterion for SSLM and considering measures $|\boldsymbol{h}_{[L]max}|$, $|\boldsymbol{h}_{[Q]max}|$, L_p, and Q_p, samples 1 and 71 were removed. Table 4 shows the values of η and κ, which were selected by CrV and Tr criteria; this table also shows the estimated values of parameters for both models after sample removal. Furthermore, it is notable that there was no change in the choice of the model

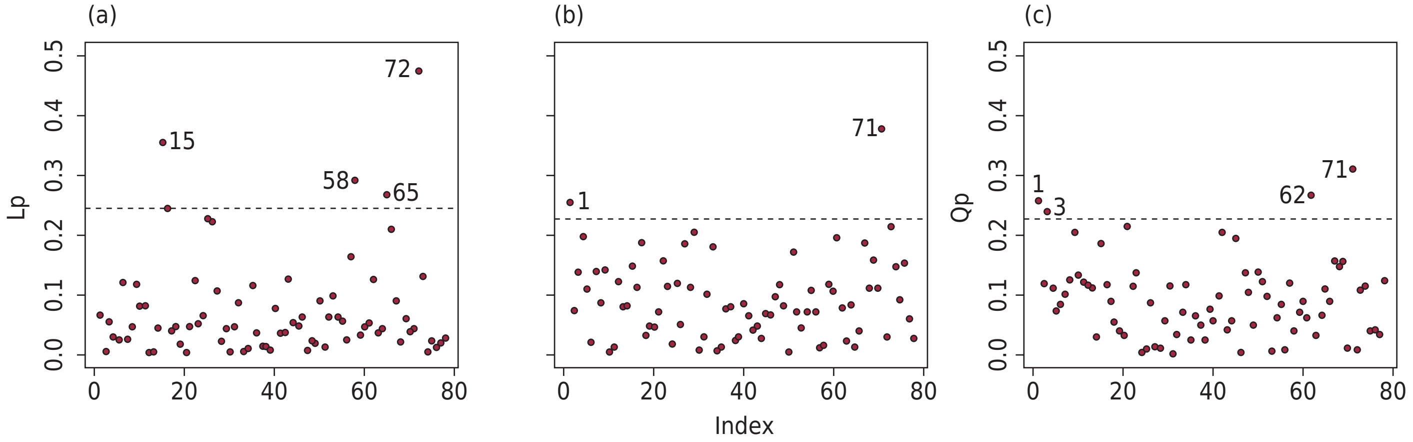

Figure 3. Graphs of local influence perturbing the linear predictor considering: (a) the Gaussian spatial linear model (GSLM) and measure L_p; (b) the slash spatial linear model (SSLM) and measure L_p; and (c) the slash spatial linear model (SSLM) and measure Q_p. Observations above the dotted line are classified as influential.

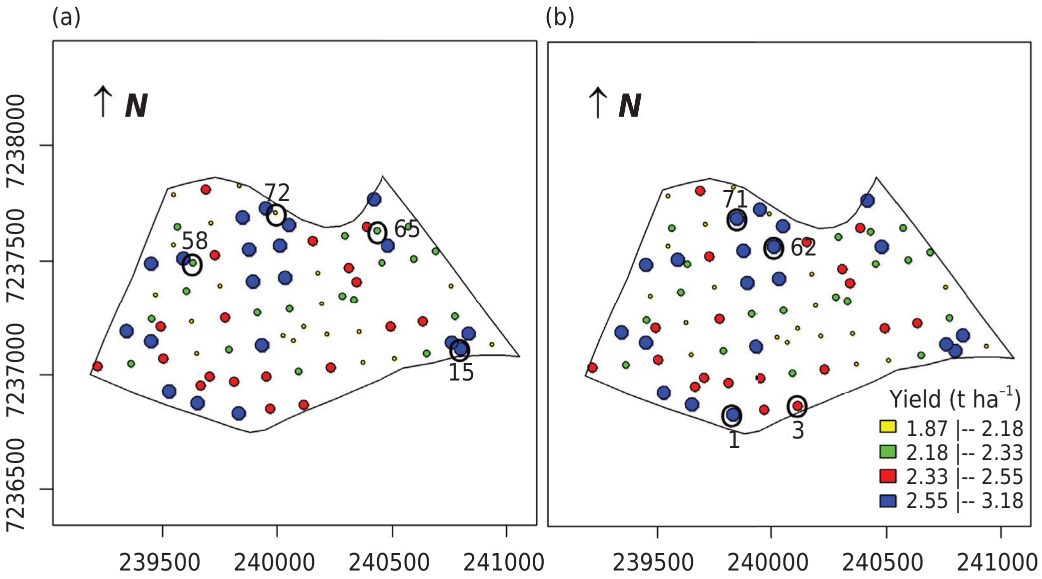

Figure 4. Post-plot graph of soybean yield (t ha^{-1}) in the experimental area identifying influential samples: (a) the Gaussian spatial linear model (GSLM) and (b) the slash spatial linear model (SSLM).

Table 4. Values of the parameter of kurtosis adjustment (η) and of smoothing (κ) selected by CrV (Cross-validation) and Tr (Trace) criteria; estimated values of parameters of the Gaussian spatial linear model (GSLM) and slash spatial linear model (SSLM) by maximum likelihood using the EM algorithm with respective asymptotic standard deviations (in brackets) after removal of influential samples

Model	Structure	Estimated parameter[1]							
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
GSLM	Matérn κ = 2.5	2.066 ($\underline{0.423}$)[2]	0.003 (0.003)	-0.102 ($\underline{0.238}$)	0.039 (0.078)	0.002 (0.006)	0.000 (0.986)	0.074 (0.987)	28.840 (0.003)
SSLM η = 0.20	Matérn κ = 2.5	2.172 (0.493)	0.004 (0.003)	-0.164 (0.277)	0.053 ($\underline{0.009}$)	-0.002 (0.006)	0.050 ($\underline{0.030}$)	0.010 ($\underline{0.007}$)	58.880 (0.003)

[1] β_i: parameters associated with the variable *i* = [soybean yield (Y), P, K, pH, and OM]; $\hat{\phi}_1$: nugget effect, $\hat{\phi}_2$: partial sill, $\hat{\phi}_3$: parameter that defines the spatial dependence radius. [2] The lowest values of the asymptotic standard deviations are underlined.

to describe spatial dependence. Nevertheless, after removing the influential samples, the estimated values of parameters changed, and especially asymptotic standard deviation showed higher values compared to the model without sample exclusion (Table 2). Moreover, we note a reduction in ϕ_1 values and in the estimated spatial dependence radius, which is 170.73 m for GSLM and 348.57 m for SSLM.

Figure 5 shows the interpolated maps of yield as a function of the covariates studied using regression-kriging, and the estimated values of parameters are shown in tables 2 and 4. The maps were generated based on the following scenarios: S1- using the entire dataset and the GSLM (Figure 5a); S2 - using the entire dataset and the SSLM (Figure 5b); S3 - removing the influential observations for GSLM (Figure 5c); and S4 - removing the influential observations for SSLM (Figure 5d). Four classes of the same size were considered to build the interpolated maps; these classes were obtained by dividing the estimated range of yield variation.

Comparing figure 5b to 5a, similar regions could be observed. However, the SSLM generated map (with $\hat{\phi}_1 = 0.051$ and 485.37 m range) exhibited a continuous structure, whereas the GSLM map (with $\hat{\phi}_1 = 0.053$ and 311.78 m range) had image formats of small dimensions. One of the problems related to maps whose modeling generates small regions is the difficulty of homogenizing the area to define management zones for localized application of fertilizers (Al-Kaisi et al., 2016). Soares (2014) affirmed a direct relationship of this effect with the parameters describing spatial dependence structure. According to this author, when the nugget effect is low or null, the influence of all samples is greater, and the map surface becomes smoother for larger ranges. When looking at figure 5c, an interpolated map that was built considering a GSLM without influential samples, a variation could be observed for short distances. This behavior is related to a structure with a sill and no nugget effect, showing a "screen effect" that is characterized by major weights around the point, as well as by almost null or even negative weights due to samples that are remote or beyond the practical limits (Soares, 2014). With respect to figure 5d, empirical analysis indicated that the removal of influential samples caused small alterations in the interpolated map. Thus, when the data set contains influential values, the maps generated using SSLM can prevent problems that cause misinterpretation in defining management zones.

To quantify the similarity between the interpolated maps, the global accuracy (GA) and kappa indexes were used. According to the classification of Krippendorff (2004), maps show low similarity if *kappa* < 0.67, medium similarity if 0.67 ≤ *kappa* ≤ 0.80, and high similarity if *kappa* > 0.80. With respect to GA, as stated by Anderson et al. (1976), maps are similar if the GA index is greater than 0.85. Table 5 shows the values associated with each index. Comparing scenario S_1 to S_2, the *kappa* index indicated medium similarity, whereas the GA index assigned similarity to the maps. However, comparing S_1 to S_3, the index values mark dissimilarity between the maps, i.e., upon removal of samples 15, 58, 65, and 72, the map outlines changed. In the case of comparing S_2 to S_4, the maps were similar.

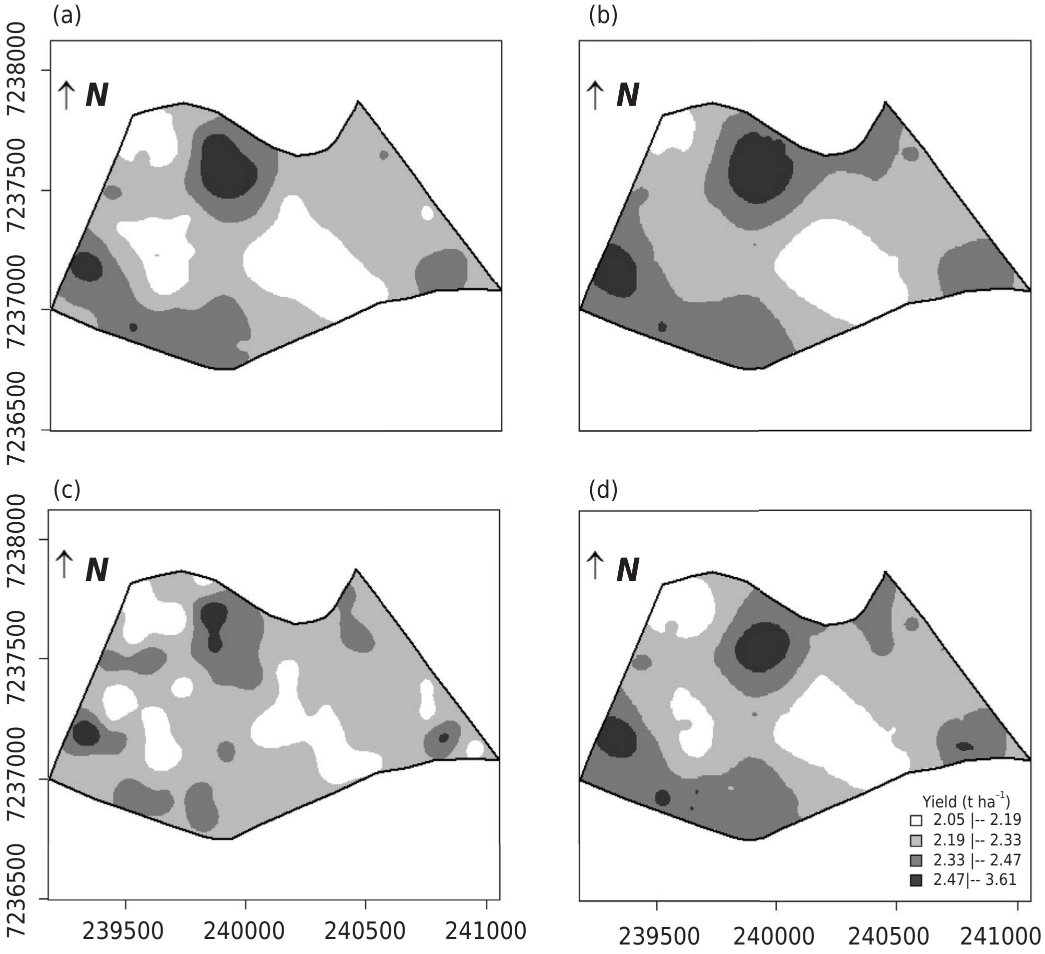

Figure 5. Interpolation maps showing the soybean yield (t ha^{-1}) as a function of soil chemical properties [P, K, pH, and OM] for: (a) S_1: Gaussian spatial linear model (GSLM) with the complete dataset; (b) S_2: slash spatial linear model (SSLM) with the complete dataset; (c) S_3: Gaussian spatial linear model (GSLM) removing samples 15 and 72; (d) S_3: slash spatial linear model (SSLM) removing samples 1 and 71.

Table 5. Index of similarity between the maps after local influence analysis

Maps analyzed	Index	
	kappa	**Global accuracy (GA)**
S_1 against S_2	0.78	0.88
S_1 against S_3	0.59	0.67
S_2 against S_4	0.87	0.86

S_1: using the complete dataset and the GSLM; S_2: using the complete dataset and the SSLM; S_3: removing the influential observations for the GSLM; S_4: removing the influential observations for the SSLM.

Therefore, the accuracy indexes of the maps showed the lower sensitivity of those generated by the slash distribution when removing an influential value, in contrast to a normal distribution. Thus, the SSLM enabled more robust modeling in the presence of influential observations, avoiding unnecessary exclusion of samples. From the interpolated maps, the yield potential of the area and its relationship to sub-regions can be investigated. This allows definition of better soil management and production system strategies, such as acidity correction.

CONCLUSIONS

The map of soybean yield variability as a function of soil chemical properties generated from the SSLM was less sensitive to the high kurtosis of the data set and the presence of influential and/or atypical values, which shows the robustness of the proposed model.

Diagnostic of local influence on the response variable and on the linear predictor based on the SSLM confirmed that an atypical value might not be influential since spatial modeling takes the position of the variable within the space into account. This finding may be a determining factor in the prediction process, avoiding exclusion of samples when using this method of modeling.

ACKNOWLEDGMENTS

We wish to acknowledge partial financial support from the *Fundação Araucária* of the state of Paraná, Capes, and CNPq of Brazil, and Project Fondecyt 1150325 of Chile.

REFERENCES

Al-Kaisi MM, Archontoulis S, Kwaw-Mensah D. Soybean spatiotemporal yield and economic variability as affected by tillage and crop rotation. J Am Soc Agron. 2016;108:1267-80. https://doi.org/10.2134/agronj2015.0363

Alcantara IC, Cysneiros FJA. Linear regression models with slash-elliptical errors. Comput Stat Data An. 2013;64:153-64. https://doi.org/10.1016/j.csda.2013.02.029

Amado TJC, Pes LZ, Lemainski CL, Schenato RB. Atributos químicos e físicos de Latossolos e sua relação com os rendimentos de milho e feijão irrigados. Rev Bras Cienc Solo. 2009;33:831-43. https://doi.org/10.1590/S0100-06832009000400008

Anderson JR, Hardy EE, Roach JT, Witmer, RE. A land use and land cover classification system for use with remote sensor data - Geological survey professional paper 964. Washington, DC: United States Government Printing Office; 1976 [acessed on 7 Sep 2016]. Available at: http://www.pbcgis.com/data_basics/anderson.pdf

Assumpção RAB, Opazo MAU, Galea M. Local influence for spatial analysis of soil physical properties and soybean yield using Student's *t*-distribution. Rev Bras Cienc Solo. 2011;35:1917-26. https://doi.org/10.1590/S0100-06832011000600008

Assumpção RAB, Uribe-Opazo MA, Galea M. Analysis of local influence in geostatistics using Student's *t*-distribution. J Appl Stat. 2014;41:2323-41. http://doi.org/10.1080/02664763.2014.909793

Bottega EL, Queiroz DM, Pinto FAC, Souza CMA. Variabilidade espacial de atributos do solo em sistema de semeadura direta com rotação de culturas no cerrado brasileiro. Rev Cienc Agron. 2013;44:1-9. https://doi.org/10.1590/S1806-66902013000100001

Cambardella CA, Moorman TB, Parkin TB, Karlen DL, Novak JM, Turco RF, Konopka AE. Field-scale variability of soil properties in Central Iowa Soils. Soil Sci Soc Am J. 1994;58:1501-11. https://doi.org/10.2136/sssaj1994.03615995005800050033x

Carvalho CGP, Arias CAA, Toledo JFF, Almeida LA, Kiihl RAS, Oliveira MF, Hiromoto DM, Takeda C. Proposta de classificação dos coeficientes de variação em relação à produtividade e altura da planta de soja. Pesq Agropec Bras. 2003;38:187-93. https://doi.org/10.1590/S0100-204X2003000200004

Casella G, Berger RL. Inferência estatística - tradução da 2a edição norte-americana. São Paulo: Cengage Learning; 2010.

Cook RD. Assessment of local influence. J R Statist Soc B. 1986;48:133-69.

Costa JM, Oliveira EF. Fertilidade do solo e nutrição de plantas. 2. ed rev. Campo Mourão: Coamo - Cascavel: Coodetec; 2001.

Cressie NAC. Statistics for spatial data. 2nd ed. NewYork: Jonh Wiley & Sons;2015.

Dagenais MG, Dufour JM. Invariance, nonlinear models, and asymptotic tests. Econometrica. 1991;59:1601-15. https://doi.org/10.2307/2938281

Dalchiavon FC, Carvalho MP, Montanari R, Andreotti M. Sugarcane productivity correlated with physical-chemical attributes to create soil management zone. Rev Ceres. 2013;60:706-14. https://doi.org/10.1590/S0034-737X2013000500015

De Bastiani F, Cysneiros AHMA, Uribe-Opazo MA, Galea M. Influence diagnostics in elliptical spatial linear models. TEST. 2015;24:322-40. https://doi.org/10.1007/s11749-014-0409-z

Dalla Nora D, Amado TJC, Nicoloso RS, Gruhn EM. Modern high-yielding maize, wheat and soybean cultivars in response to gypsum and lime application on no-till Oxisol. Rev Bras Cienc Solo. 2017;41:e0160504. https://doi.org/10.1590/18069657rbcs20160504

Diggle PJ, Ribeiro Junior PJ. Model-based geostatistics. New York: Springer; 2007.

Grzegozewski DM, Uribe-Opazo MA, De Bastiani F, Galea M. Local influence when fitting Gaussian spatial linear models: an agriculture application. Cien Inv Agr. 2013;40:523-35. https://doi.org/10.4067/S0718-16202013000300006

Instituito Brasileiro de Geografia e Estatística -IBGE. Banco de dados agregados- Sistema IBGE de recuperação automática-SIDRA; 2016 [acesso em 10 nov 2016]. Disponível em: http://www.sidra.ibge.gov.br

Jonathan R, Uribe-Opazo MA, De Bastiani F, Johann JA. Técnicas para detecção de pontos influentes em variáveis contínuas regionalizadas. Eng Agric. 2016;36:152-65. http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v36n1p152-165/2016

Kano Y, Berkane M, Bentler PM. Statistical inference based on pseudo-maximum likelihood estimators in elliptical populations. J Am Stat Assoc. 1993;88:135-43. https://doi.org/10.2307/2290706

Krippendorff K. Content analysis: an introduction to its methodology. 2nd ed. Thousand Oaks: Sage Publications; 2004.

Lange K, Sinsheimer JS. Normal/independent distributions and their applications in robust regression. J Comput Graph Stat. 1993;2:175-98. https://doi.org/10.2307/1390698

Matérn B. Spatial variation - Lecture notes in statistics. 2nd ed. New York: Springer-Verlag; 1986.

Nesi CN, Ribeiro A, Bonat WH, Ribeiro Junior PJ. Verossimilhança na seleção de modelos para predição espacial. Rev Bras Cienc Solo. 2013;37:352-8. https://doi.org/10.1590/S0100-06832013000200006

Nunes RS, Sousa DMG, Goedert WJ, Vivaldi LJ. Distribuição de fósforo no solo em razão do sistema de cultivo e manejo da adubação fosfatada. Rev Bras Cienc Solo. 2011;35:877-88. https://doi.org/10.1590/S0100-06832011000300022

Osorio F, Paula GA, Galea M. On estimation and influence diagnostics for the Grubbs' model under heavy-tailed distributions. Comput Stat Data An. 2009;53:1249-63. https://doi.org/10.1016/j.csda.2008.10.034

Pagani A, Mallarino AP. On-farm evaluation of corn and soybean grain yield and soil pH responses to liming. Agron J. 2015;107:71-82. https://doi.org/10.2134/agronj14.0314

Passos AMA, Rezende PM, Carvalho ER, Ávila FW. Pó de carvão, esterco de curral e cama de frango no cultivo da soja e atributos químicos de um Cambissolo distrófico. Rev Bras Cienc Agrar. 2015;10:382-8. https://doi.org/10.5039/agraria.v10i3a4546

Pereira FCBL, Mello LMM, Pariz CM, Mendonça VZ, Yano EH, Miranda EEV, Crusciol CAC. Autumn Maize Intercropped with Tropical Forages: Crop Residues, Nutrient Cycling, Subsequent Soybean and Soil Quality. Rev Bras Cienc Solo. 2016;40:e0150003. https://doi.org/10.1590/18069657rbcs20150003

R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016 [acessed on 2 Nov. 2016]. Available at: http://www.R-project.org.

Santos HG, Jacomine PKT, Anjos LHC, Oliveira VA, Oliveira JB, Coelho MR, Lumbreras JF, Cunha TJF. Sistema brasileiro de classificação de solos. 3. ed. Rio de Janeiro: Embrapa Solos; 2013.

Schemmer R, Uribe-Opazo MA, Galea M, Assumpção RAB. Spatial variability of soybean yield through a reparameterized t-Student model. Eng Agric. 2017;37:360-70. http://dx.doi.org/10.1590/1809-4430-eng.agric.v37n4p760-770/2017

Soares, A. Geoestatística para as ciências da terra e do ambiente. 3. ed. Portugal: IST Press; 2014.

Sobjak R, Souza EG, Bazzi CL, Uribe-Opazo MA, Betzek NM. Redundant variables and the quality of management zones. Eng Agric. 2016;36:78-93. https://doi.org/10.1590/1809-4430-Eng.Agric.v36n1p78-93/2016

Soil Survey Staff. Keys to soil taxonomy. 12th ed. Washington, DC: United States Department of Agriculture, Natural Resources Conservation Service; 2014.

Uribe-Opazo MA, Borssoi JA, Galea M. Influence diagnostics in Gaussian spatial linear models. J Appl Stat. 2012;39:615-30. https://doi.org/10.1080/02664763.2011.607802

Walkley A, Black I A. An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Sci. 1934;27:29-37. https://doi.org/10.1097/00010694-193401000-00003

Warrick AW, Nielsen DR. Spatial variability of soil physical properties in the field. In: Hillel D, editor. Applications of soil physics. New York: Academic Press; 1980. p. 319-44.

Zhu H-T, Lee S-Y. Local influence for incomplete-data models. J R Statist Soc B. 2001;63:111-26. https://doi.org/10.1111/1467-9868.00279