

## SPATIAL VARIABILITY OF SOYBEAN YIELD THROUGH A REPARAMETERIZED T-STUDENT MODEL

Doi: <http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v37n4p760-770/2017>

ROSANGELA C. SCHEMMER<sup>1</sup>, MIGUEL A. URIBE-OPAZO<sup>2\*</sup>, MANUEL GALEA<sup>3</sup>,  
ROSANGELA A. B. ASSUMPCÃO<sup>4</sup>

<sup>2\*</sup>Corresponding author. Centro de Ciências Exatas e Tecnológica, Universidade Estadual do Oeste do Paraná - UNIOESTE/ Cascavel - PR, Brasil. E-mail: miguel.opazo@unioeste.br

**ABSTRACT:** The t-Student distribution has been used to the spatial dependence modelling of soybean yield as an alternative to the normal distribution, being used for data with heavier tails or discrepant values. However, a usual Student t-distribution does not allow direct comparisons of geostatistical methods with a normal distribution. The aim of this study was to assess the soybean yield spatial variability through a reparameterized t-Student linear model, comparing the results with those of a Gaussian linear model. For parameter estimation, a complete maximum likelihood (CML) method was used through an expectation-maximization (EM) algorithm. The maps constructed with both reparameterized t-Student and normal distributions are dissimilar and present a kappa index ( $K$ ) equivalent to 0.64. The reparameterized t-Student distribution is an alternative in studying data with discrepant values, showing the ability to decrease the influence of these points.

**KEYWORDS:** EM algorithm, spatial dependence, geostatistics, complete maximum likelihood.

### INTRODUCTION

Geostatistics can assist in precision agriculture since its techniques allow constructing maps that determine the spatial dependence structure of yield associated with soil and plant attributes. Thus, it helps the producer to decide on the use of agricultural inputs in appropriate quantities and locations in order to increase yield, reduce losses, and maintain environmental quality. This technique is based on the regionalized variable theory proposed by Matheron, influenced by the observations made by Kriger. According to Vieira (2000), Kriger analyzed gold concentration data in South Africa and observed the impossibility of finding meaning in the variances without taking into account the distance between samples. Therefore, the values of a variable distributed in the space are correlated within a radius of spatial dependence.

In a spatial variability study, the results obtained by geostatistical methods can be influenced by discrepant data, leading to biased predictions (Cressie, 2015). A solution to the presence of discrepant data is the use of robust models, whose parameter estimation is less sensitive to these data. According to Manghi et al. (2016), class models of symmetric distributions allow reducing the influence of discrepant data, incorporating additional parameters that adjust the kurtosis of data distribution. The t-Student distribution belongs to the class of symmetric distributions and exhibits symmetry properties, greater flexibility regarding the degree of kurtosis, and has as additional shape parameter  $\nu > 0$ , which defines the degrees of freedom of distribution (Assumpção et al., 2011; 2014). Lange et al. (1989) propose a reparametrization of the t-Student distribution from a transformation in the shape parameter  $\nu$ , allowing us to assume the existence of the second finite moment and thus a more direct comparison with the normal distribution. This reparametrization is justified by the importance that the spatial dependence modeling represents since the new shape parameter  $\eta$  is limited and this process allows estimating parameters by maximum likelihood (Nesi et al., 2013) and implementing the EM iterative algorithm (Dempster et al., 1977; Assumpção et al., 2014).

<sup>1</sup> Centro de Ciências Exatas e Tecnológica, Universidade Estadual do Oeste do Paraná - UNIOESTE/ Cascavel - PR, Brasil.

<sup>3</sup> Departamento de Estadística, Pontificia Universidad Católica de Chile/ Santiago, Chile.

<sup>4</sup> COMAT, Universidade Tecnológica Federal do Paraná/ UTFPR/ Toledo - PR, Brasil.

Received in: 4-28-2016

Accepted in: 3-8-2017

This study aimed to assess the spatial variability of soybean yield by means of a reparameterized t-Student linear model, comparing the results with a Gaussian linear model. For estimating these model parameters, a complete maximum likelihood (CML) method was used through an expectation-maximization (EM) algorithm.

## THEORETICAL FOUNDATION

### Reparameterized t-Student distribution

Much of the statistical inference involving continuous random variables is based on normal distribution. However, to obtain reasonable inferences, assuming normality, it is necessary to ensure conditions such as symmetry and a certain value of kurtosis. Among the symmetric models alternative to the normal distribution is the t-Student distribution, which presents as an additional parameter the degree of freedom  $\nu$  ( $\nu > 0$ ) that allows kurtosis modeling. A priori, this parameter can be fixed. However, Lange et al. (1989) recommend fixing it at  $\nu = 4$  for a small data set and its estimation for a large data set. This distribution has been widely used in the study with real data because it has tails longer than the normal distribution and allows the discrepant points present in the data set to be encompassed (Lange et al., 1989; Osorio et al., 2007). Galea et al. (2002) suggest the t-Student distribution as an alternative to the normal distribution due to the statistical inference based on the t-Student distribution to combine conceptual and computational simplicity with generality, in addition to being applicable in a great variety of situations. An important feature of t-Student distribution is that when the degree of freedom  $\nu$  increases, the t-Student distribution approaches to the normal probability distribution.

Lange et al. (1989) state that if a random vector  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  has as probability density function multivariate t-Student with a location parameter  $\boldsymbol{\mu}$ , scale matrix  $\mathbf{V}$ , and  $\nu > 0$  degrees of freedom,  $\mathbf{Y} \sim t_n(\boldsymbol{\mu}, \mathbf{V}, \nu)$  is denoted. The expectation of the random vector  $\mathbf{Y}$  is  $E(\mathbf{Y}) = \boldsymbol{\mu}\mathbf{1}$ , where  $\mathbf{1}$  is a vector of 1's of order  $n \times 1$ , for  $\nu > 1$ , and the covariance matrix  $n \times n$  of  $\mathbf{Y}$  is  $Cov(\mathbf{Y}) = \frac{\nu}{\nu - 2} \mathbf{V} = \boldsymbol{\Sigma}$  for  $\nu > 2$ . For values of  $\nu \leq 2$ , the covariance matrix  $Cov(\mathbf{Y})$  is undefined.

Lange et al. (1989) suggest the reparametrization of t-Student distribution for allowing the direct comparison between parameter estimation of the mean vector and the covariance matrix with the model assuming normality. The authors also mention that an improvement of inference is observed when the degree of freedom presents the transformation  $\nu = 1/\eta$ .

$\mathbf{Y} = (Y_1, \dots, Y_n)^T$  is considered a random vector that has reparameterized t-Student distribution with shape parameter  $\eta$  fixed, in which  $0 < \eta < 1/2$ , with covariance matrix  $\boldsymbol{\Sigma}$ , mean vector  $E(\mathbf{Y}) = \boldsymbol{\mu}\mathbf{1}$  if its probability density function is given by [eq. (1)]:

$$f_{\mathbf{Y}}(y) = K_n(\eta) |\boldsymbol{\Sigma}|^{-\frac{1}{2}} [1 + c(\eta)\delta^2]^{-\frac{1}{2\eta}(1+n\eta)}, \quad (1)$$

where,

$$K_n(\eta) = \left(\frac{c(\eta)}{\pi}\right)^{\frac{n}{2}} \frac{\Gamma\left(\frac{1+n\eta}{2\eta}\right)}{\Gamma\left(\frac{1}{2\eta}\right)},$$

Which  $\delta^2 = (\mathbf{Y} - \boldsymbol{\mu}\mathbf{1})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu}\mathbf{1})$  is the Mahalanobis distance,  $c(\eta) = \eta/(1-2\eta)$  and  $0 < \eta < 1/2$ .

It is denoted by  $\mathbf{Y} \sim \tau_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta)$  that the vector  $\mathbf{Y}$  has  $n$ -variate reparameterized t-Student distribution.

### Spatial linear model

For the study of spatial dependence,  $\{Y(s_i), s_i \in \mathbf{S}\}$  is considered a stochastic process of second-order stationary, where  $\mathbf{S} \subset \mathfrak{R}^2$  and  $\mathfrak{R}^2$  is a two-dimensional Euclidean space. Let  $\mathbf{Y} = (Y(s_1), \dots, Y(s_n))^T$  be a vector  $n \times 1$  of the response variable corresponding to spatial locations known in  $s_i$  with  $i = 1, \dots, n$ . The georeferenced variable  $Y(s_i)$  can be written as:

$$Y(s_i) = \mu(s_i) + e(s_i), \quad (2)$$

being the deterministic term  $\mu(s_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ , where  $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$  is a vector  $1 \times p$  of explanatory variables at position  $s_i$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is the vector  $p \times 1$  of unknown parameters to be estimated, and  $e(s_i)$  is a spatially correlated random component.

Equation (2) can be written in a matrix form as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (3)$$

where  $\mathbf{X}$  is a matrix  $n \times p$  of columns with complete rank, with lines  $\mathbf{x}_i^T$  and  $\boldsymbol{\varepsilon} = (e(s_1), \dots, e(s_n))^T$ , with  $i = 1, \dots, n$ . It is assumed that the random errors  $e(s_i)$  have zero mean, i.e.  $E[e(s_i)] = 0$  and the variation between points in space is determined by some covariance function  $Cov[e(s_i), e(s_u)] = Cov[\mathbf{Y}(s_i), \mathbf{Y}(s_u)] = C(s_i, s_u) = \sigma_{iu}$  for  $i, u = 1, \dots, n$ . The spatial modeling given in [eq. (3)] depends on the covariance matrix structure  $\boldsymbol{\Sigma} = [(\sigma_{iu})]$ , where  $\sigma_{iu} = C(s_i, s_u)$  for  $i, u = 1, \dots, n$ , of the stochastic process  $\mathbf{Y}$ . The covariance function  $C(s_i, s_u)$  is used in the study of spatial dependence of the stationary process and it is specified by a three-dimensional vector  $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_3)^T$  of the form given in [eq. (4)] (Uribe-Opazo et al., 2012):

$$\boldsymbol{\Sigma} = \varphi_1 \mathbf{I}_n + \varphi_2 \mathbf{R} \quad (4)$$

where  $\varphi_1$  is the parameter nugget effect ( $\varphi_1 \geq 0$ ),  $\varphi_2$  is the parameter sill ( $\varphi_2 \geq 0$ ),  $\mathbf{R}$  is a symmetric matrix  $n \times n$ , whose elements are as a function of the parameter ( $\varphi_3 > 0$ ) ( $\mathbf{R} = \mathbf{R}(\varphi_3) = [(r_{iu})]$ ) with diagonal elements  $r_{ii} = 1$  and  $r_{iu} = \varphi_2^{-1} C(s_i, s_u)$  for  $\varphi_2 \neq 0$  and  $r_{iu} = 0$  for  $\varphi_2 = 0$ ,  $i \neq u = 1, \dots, n$ , being  $r_{iu}$  dependent on the Euclidian distance  $h_{iu} = \|s_i - s_u\|$  between the points  $s_i$  and  $s_u$ , and  $\mathbf{I}_n$  is the identity matrix  $n \times n$ . The parametric form of the covariance matrix  $\boldsymbol{\Sigma}$ , represented in [eq. (4)], occurs for several stationary and isotropic processes, in which the covariance  $C(s_i, s_u) = C(h_{iu})$  is defined by the covariance function  $C(h_{iu}) = \varphi_2 r_{iu}$ . In the covariance functions  $C(h_{iu})$ , the variance of the stochastic process reparameterized t-Student  $\mathbf{Y}$  is given by  $C(0) = \varphi_1 + \varphi_2$ .

On the assumption that  $\mathbf{Y} \sim \tau_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}, \eta)$ , where  $\eta$  represents the shape parameter, considered fixed and the unknown parameters of the model  $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\varphi}^T)^T$ , with  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  and  $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_3)^T$  can be estimated by maximizing the logarithm of the complete likelihood

function defined by [eq. (5)]:

$$l_c(\hat{\boldsymbol{\theta}}, \mathbf{Y}_c) = \max(l_c(\boldsymbol{\theta}, \mathbf{Y}_c)), \tag{5}$$

being

$$l_c(\boldsymbol{\theta}, \mathbf{Y}_c) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \delta^2 \vartheta + \frac{1}{2\eta} \log\left(\frac{1}{2c(\eta)}\right) - \log\left(\Gamma\left(\frac{1}{2\eta}\right)\right) + \frac{\eta}{2} \log(\vartheta) + \frac{1}{2c(\eta)} [\log(\vartheta) - \vartheta], \tag{6}$$

where  $\delta^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ ,  $c(\eta) = \eta/(1-2\eta)$ ,  $\vartheta > 0$  and  $0 < \eta < 1/2$ .

Maximization of [eq. (6)] is obtained by using an iterative process. In this case, the EM (expectation and maximization) algorithm was applied, being the stopping criterion the relative error (RE), where  $RE^r = \frac{\|\boldsymbol{\theta}^r - \boldsymbol{\theta}^{(r-1)}\|}{\|\boldsymbol{\theta}^r\|} < \epsilon$ , with  $\epsilon = 10^{-5}$ . To determine the shape parameter  $\eta$  considered fixed, the criteria of cross-validation ( $VC(\eta)$ ), presented by De Bastiani et al. (2015), and the trace criterion ( $T_r(\eta)$ ), proposed by Kano et al. (1993), were applied. For the reparameterized t-Student model, cross-validation is given by [eq. (7)]:

$$VC(\eta) = \frac{1}{n} \left[ \sum_{i=1}^n \left( \frac{y(s_i) - \hat{y}_i(s_i)}{1 - h_{ii}} \right)^2 \right], \tag{7}$$

where  $\hat{y}_i(s_i) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_i$ , with  $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$  being the  $i$ -th line of the matrix  $\mathbf{X}$ , is the prediction at the location  $s_i$  without considering the observation  $(y_i, \mathbf{x}_i^T)$ ,  $\hat{\boldsymbol{\beta}}_i$  is the maximum likelihood estimator for  $\boldsymbol{\beta}_i$  without considering the  $i$ -th observation and  $h_{ii}$  is the  $i$ -th diagonal element of the matrix Hat ( $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1}$ ), also called a projection matrix. Trace criterion consists of calculating the trace of the asymptotic covariance matrix of the estimated mean ( $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ ), as a criterion in choosing considering that the shape parameter is obtained by:

$$T_r(\eta) = \left( \frac{(1-2\eta)(1+(n+2)\eta)}{1+n\eta} \right) tr \left[ (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1} \right], \tag{8}$$

Where  $\hat{\boldsymbol{\Sigma}} = \hat{\varphi}_1 \mathbf{I}_n + \hat{\varphi}_2 \hat{\mathbf{R}}$ . For the two criteria, the best shape parameter  $\eta$  is determined by the lowest values of cross-validation ( $VC(\eta)$ ) and trace ( $T_r(\eta)$ ). After choosing the estimation of  $\eta$ , the best Matérn family model was defined with different shape parameters  $\kappa$  (Matérn, 1986) by using the lowest standard error. The map was constructed by means of the regression-kriging method (Michel & Kobiyama, 2015) since it allows the use of covariates. Finally, the maps constructed with the reparameterized t-Student distribution and normal distribution were compared using the Kappa index ( $K$ ) (De Bastiani et al., 2012), used to measure the exactness of thematic classifications, i.e. it provides a measure of agreement between the reference map values and the model map values. This index is recommended as an adequate precision measure because it uses all elements of the error matrix, being defined by [eq. (9)]:

$$K = \frac{\left(N^*\right) \sum_{i=1}^r n_{ii} - \sum_{i=1}^r \left(n_{i+} n_{+i}\right)}{\left(N^*\right)^2 - \sum_{i=1}^r \left(n_{i+} n_{+i}\right)}, \tag{9}$$

where  $N^*$  is the total area,  $n_{ii}$  is the area belonging to class  $i$  of the model and reference maps,  $n_{i+}$  is the area belonging to class  $i$  of the model map,  $n_{+i}$  is the area belonging to class  $i$  of the reference map, and  $r$  is the number of classes. According to Krippendorff (2004) classification,  $K$  is classified with low similarity if  $K < 0.67$ , medium similarity if  $0.67 < K < 0.80$ , and high similarity if  $K > 0.80$ .

## MATERIAL AND METHODS

### Location and characteristics of the study area

Data on soybean yield, plant height, and pods per plant were collected from an experimental area of 47.95 ha located in Cascavel, the western region of Paraná, Brazil, with an approximate location of 24.83° S and 53.60° W, and an average altitude of 650 meters. The soil of this area is classified as a clayey Oxisol (Haplorthox) (EMBRAPA, 2011) and regional climate is a temperate super-humid climate type Cfa (Köppen) with average annual temperature of 21 °C. All samples were georeferenced in the spatial coordinate system (UTM) by using a Trimble GPS25 (Global Positioning System) GEOEXPLORER 3 data receiver. Figure 1 shows the experimental area in a regular grid of 75 × 75 meters, totaling 83 observations for the 2006/2007 agricultural season.

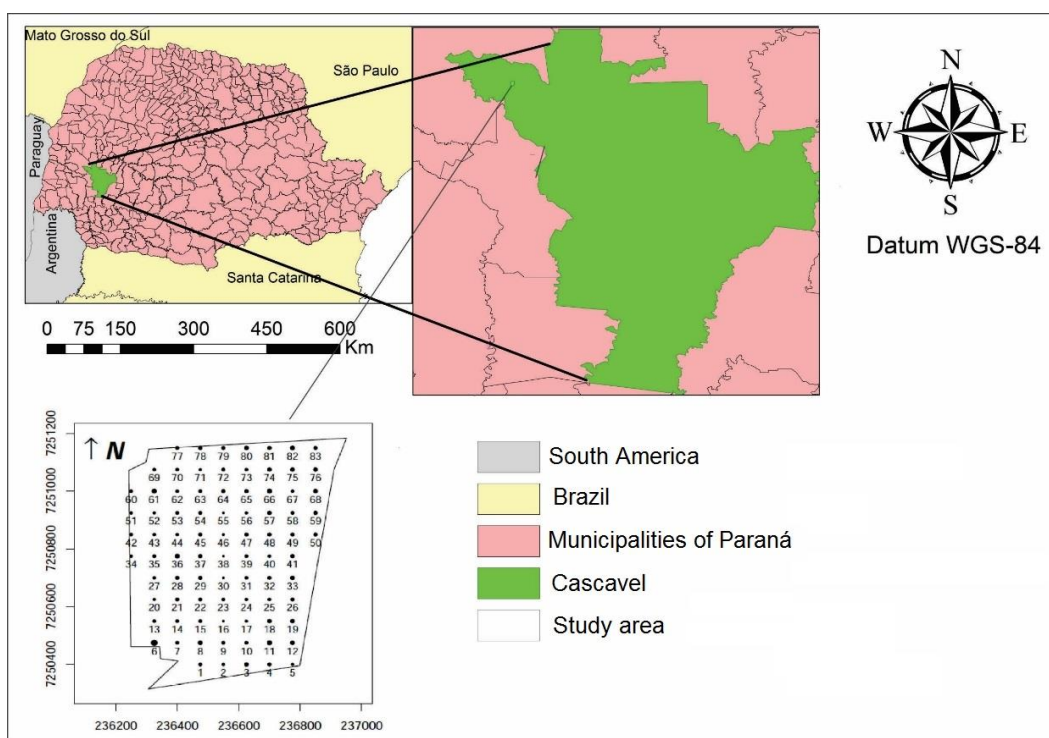


FIGURE 1. Area location in the 2006/2007 agricultural season.

In 2006, soybean was cultivated in this area by means of the no-tillage system. In 2007, data on soybean yield were collected, being estimated by considering the amount of soybean harvested from all plants distributed in two rows over a meter long, representing a plot. Grains were weighed for each plot and the water content was verified for subsequent correction to 13%. Yield value was converted into  $t\ ha^{-1}$ . The estimation of average plant height ( $Hgt$ ), in cm, was performed at soybean vegetative peak by calculating the average of four plants over a linear meter. For the average number of pods per plant ( $N$ ), four plants were chosen at each point and the number of pods was counted per plant at harvest time.

Statistical analyses were performed using the free software R, version 3.2.0 (R Core Team, 2016). The following packages were used: geoR (Ribeiro Junior & Diggle, 2016) for studying the spatial data, map construction by regression kriging interpolation, and comparison of thematic maps; matrixcalc (Novomestky, 2012) for trace calculation; e1071 (Meyer et al., 2015) for calculating the asymmetry and kurtosis; and classInt (Bivand, 2015) for choosing the class intervals for continuous numerical variables.

## RESULTS AND DISCUSSION

Table 1 shows the exploratory analysis of values found for the variables soybean yield (*Prod*) ( $t\ ha^{-1}$ ), average plant height (*Hgt*) (cm), and an average number of pods (*N*). The average soybean yield is  $2.99\ t\ ha^{-1}$ , with a minimum value of  $1.50\ t\ ha^{-1}$  and a maximum value of  $5.53\ t\ ha^{-1}$ . Moreover, 75% of the area presents a yield lower than or equal to  $3.35\ t\ ha^{-1}$ . Soybean yield is classified as heterogeneous since the coefficient of variation (CV) is 21.27%.

TABLE 1. Descriptive statistics for the variable soybean yield (*Prod*), the covariates average plant height (*Hgt*) and an average number of pods (*N*).

Statistics	<i>Prod</i>	<i>N</i>	<i>Hgt</i>
Number of data	83	83	83
Minimum	1.50	34.00	23.00
Maximum	5.53	60.20	61.40
Mean	2.99	42.62	38.20
Median	2.94	41.40	37.00
Standard deviation	0.64	5.11	8.26
Q <sub>1</sub>	2.60	38.90	32.00
Q <sub>3</sub>	3.35	46.00	44.00
CV (%)	21.27	11.99	21.58
Asymmetry	0.54	0.79	0.56
Kurtosis	2.04	0.36	-0.32

Q<sub>1</sub>: 1st quartile; Q<sub>3</sub>: 3rd quartile; CV: coefficient of variation.

The boxplot graph presented in Figure 2a detected a single discrepant point, which corresponds to the sample element 6, with coordinates (236325, 7250475), referring to the maximum yield value in the data set, being equivalent to  $5.53\ t\ ha^{-1}$ . According to the Postplot graph shown in Figure 2b, observation 6 is in a region where the nearest neighbors have a soybean yield between 2.60 and 2.94  $t\ ha^{-1}$ .

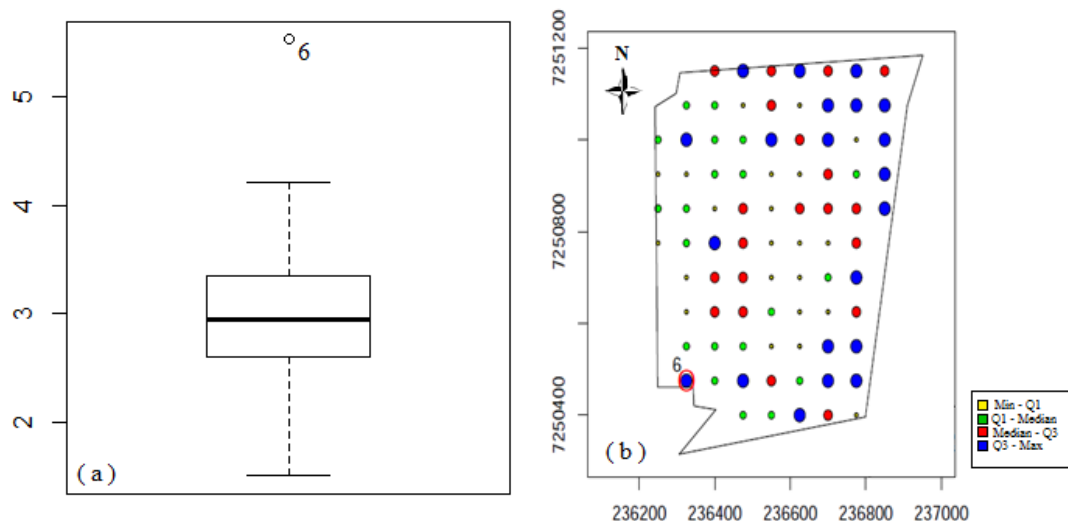


FIGURE 2. Box plot (a) and Postplot (b) graphs for soybean yield data.

In order to identify the spatial dependence structure of soybean yield as a function of the average plant height ( $Hgt$ ) and an average number of pods per plant ( $N$ ), the average soybean yield  $\mu(s_i)$  in the position  $s_i \in S \subset \mathfrak{R}^2$  was considered as a spatial linear regression model given by:

$$\mu(s_i) = \beta_1 + \beta_2 Hgt(s_i) + \beta_3 N(s_i), \quad i = 1, \dots, n, \tag{10}$$

where  $\beta_1, \beta_2$ , and  $\beta_3$  are the unknown parameters to be estimated.

Parameter estimation studies were performed by complete maximum likelihood (CML) using the EM algorithm of the spatial linear model defined in [eq. (10)] and parameters of the spatial dependence structure  $\Sigma$  given in [eq. (4)], considering the Matérn family with parameters  $\kappa = 0.5, 1.0, 2.0, 5.0, 10$  and  $20$  associated to shape parameters of the reparameterized t-Student  $\eta = 0.05, 0.067, 0.1, 0.143$ , and  $0.2$ .

Table 2 shows the determination of the best shape parameter  $\eta$  of the reparameterized t-Student distribution associated to each shape parameter  $\kappa$  of the Matérn family using the cross-validation criterion and trace defined by Equations (7) and (8). In bold is presented the choice of each parameter  $\eta$  for each  $\kappa$  with the lowest values of cross-validation ( $VC(\eta)$ ) and trace ( $T_r(\eta)$ ).

TABLE 2. Cross-validation and trace for the choice of the best shape parameter  $\eta$ .

$\kappa$	$\eta$	$VC(\eta)$	$T_r(\eta)$	$\kappa$	$\eta$	$VC(\eta)$	$T_r(\eta)$
0.5	<b>0.050</b>	<u>0.412220</u>	<u>2.115664</u>	1.0	<b>0.050</b>	<u>0.4129935</u>	<u>2.1301236</u>
	0.067	0.412234	2.1760428		0.067	0.4130204	2.1530738
	0.100	0.412260	2.2758176		0.100	0.4130351	2.1797482
	0.143	0.412285	2.3646624		0.143	0.4130009	2.6215253
	0.200	0.412308	2.4774702		0.200	0.4130525	2.2530225
2.0	0.050	0.413862	2.2563595	5.0	<b>0.050</b>	<u>0.4144359</u>	<u>2.2792236</u>
	0.067	0.413766	2.330967		0.067	0.4144402	2.3613108
	0.100	0.413772	2.4853913		0.100	0.4145646	2.3221812
	0.143	0.413857	2.1550879		0.143	0.4144596	2.7789105
	<b>0.200</b>	<u>0.413763</u>	<u>2.1447926</u>		0.200	0.4144739	3.1683194
10	<b>0.050</b>	<u>0.414694</u>	<u>2.2676439</u>	20	<b>0.050</b>	<u>0.4148278</u>	<u>2.2558034</u>
	0.067	0.414715	2.3462043		0.067	0.4148312	2.3450125
	0.100	0.414790	2.484116		0.100	0.4148545	2.5380225
	0.143	0.414886	2.5932054		0.143	0.4149831	2.7765308
	0.200	0.414731	3.2988219		0.200	0.4151899	3.0632745

$\eta$ : shape parameter;  $VC(\eta)$ : cross-validation;  $T_r(\eta)$ : trace;  $\kappa$ : shape parameter of the Matérn family. In bold is the best shape parameter  $\eta$ ; underlined is the lowest value of cross-validation and trace.

Figure 3 shows the cross-validation  $VC(\eta)$  and trace  $T_r(\eta)$  graphs for each  $\kappa$  value of the Matérn family model related to those chosen in Table 3. For  $\kappa = 0.5$  and  $20$ ,  $VC(\eta)$  and  $T_r(\eta)$  values increase as  $\eta$  value increases. For the other cases, when  $\eta$  values increase,  $VC(\eta)$  and  $T_r(\eta)$  values oscillate.

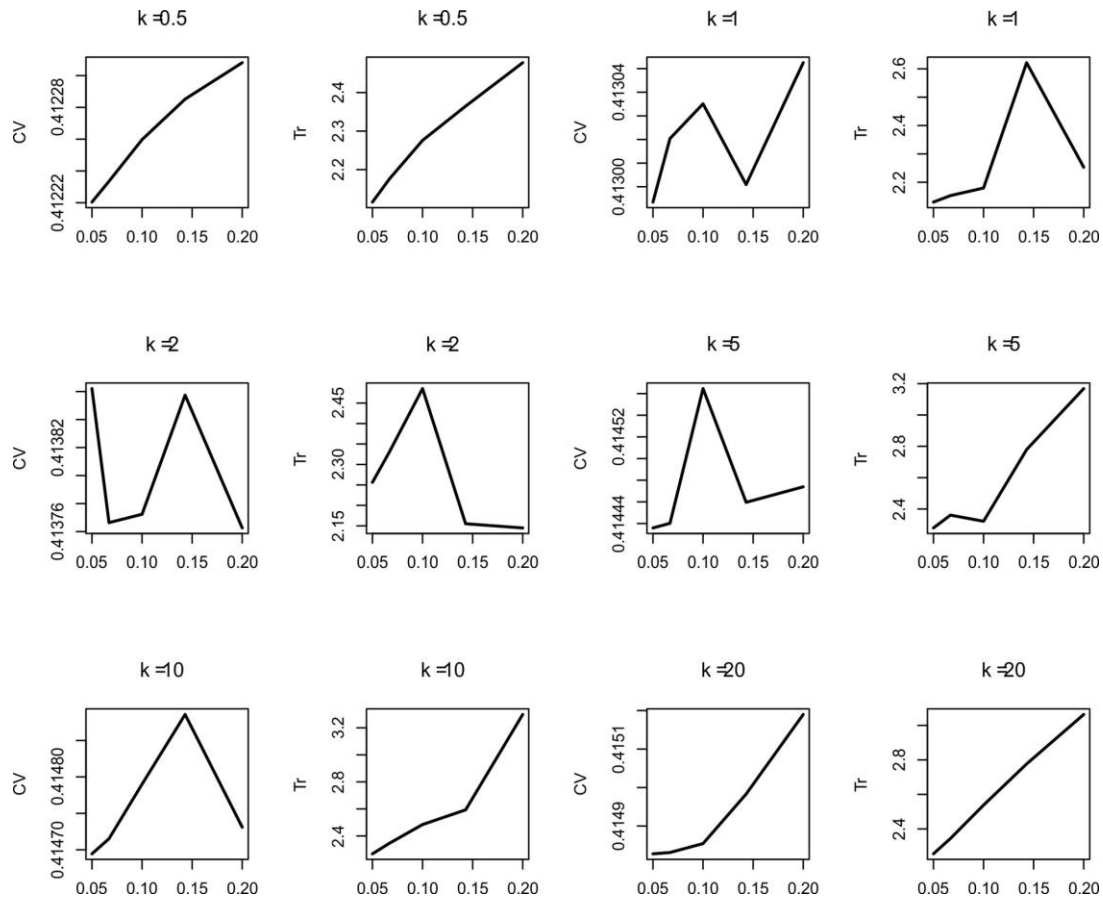


FIGURE 3. Graphs of cross-validation  $VC(\eta)$  and trace  $T_r(\eta)$ .

Table 3 shows the results of parameter estimation and the respective standard deviations considering the  $\eta$  values for each  $\kappa$  selected in Table 2. The lowest standard deviations of estimators correspond to the estimated values of  $\eta = 0.050$  and  $\kappa = 0.5$ , whose estimates are  $\hat{\beta}_1 = 0.993$ ,  $\hat{\beta}_2 = 0.021$ ,  $\hat{\beta}_3 = 0.030$ ,  $\hat{\phi}_1 = 0.248$ ,  $\hat{\phi}_2 = 0.121$ , and  $\hat{\phi}_3 = 112.8$ , with a practical range of approximately 338.0 m.

TABLE 3. Estimation of the parameters  $\beta$  and  $\phi$  via EM algorithm for different  $\kappa$  and  $\eta$ .

$\kappa$	$\eta$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
0.5	0.050	0.993	0.021	0.030	0.248	0.121	112.8
		<b>(0.7216)</b>	<b>(0.0144)</b>	<b>(0.0077)</b>	<b>(0.1189)</b>	<b>(0.8792)</b>	<b>(0.0003)</b>
1.0	0.050	0.978	0.022	0.030	0.270	0.087	96.88
		(0.7319)	(0.0146)	(0.0080)	(0.1141)	(0.8905)	(0.0003)
2.0	0.200	0.965	0.022	0.030	0.273	0.072	76.69
		(0.7433)	(0.0148)	(0.0080)	(0.3028)	(1.6411)	(0.0005)
5.0	0.050	0.959	0.022	0.031	0.300	0.072	49.09
		(0.7478)	(0.0149)	(0.0080)	(0.1214)	(0.9141)	(0.0002)
10	0.050	0.956	0.022	0.031	0.302	0.070	34.68
		(0.7483)	(0.0149)	(0.0080)	(0.1219)	(0.9190)	(0.0003)
20	0.050	0.956	0.022	0.031	0.303	0.069	24.39
		(0.7489)	(0.0148)	(0.0080)	(0.1223)	(0.9216)	(0.0005)

$\hat{\beta}$ : estimated parameters of the spatial linear regression model;  $\hat{\phi}$ : estimated spatial parameters.

Figure 4a shows the soybean yield map constructed by means of regression kriging interpolation considering that the data have a reparameterized t-Student distribution with  $\eta = 0.05$  and shape parameter of the Matérn model  $\kappa = 0.5$  with the following parameters estimated by CML



of the spatial linear regression model:  $\hat{\beta}_1 = 0.993$ ,  $\hat{\beta}_2 = 0.021$ ,  $\hat{\beta}_3 = 0.030$ ,  $\hat{\phi}_1 = 0.248$ ,  $\hat{\phi}_2 = 0.121$ , and  $\hat{\phi}_3 = 112.8$ , with a practical range of 328.0 m. Figure 4b shows the soybean yield map considering that the data have a normal distribution and shape parameter of the Matérn model  $\kappa=0.5$  with the following parameters estimated by the maximum likelihood of the spatial linear regression model:  $\hat{\beta}_1 = 0.957$ ,  $\hat{\beta}_2 = 0.023$ ,  $\hat{\beta}_3 = 0.030$ ,  $\hat{\phi}_1 = 0.298$ ,  $\hat{\phi}_2 = 0.132$ , and  $\hat{\phi}_3 = 133.4$ , with a practical range of 400.20 m.

An increase in area percentage was observed in the 1st, 2nd, and 5th classes of the map constructed with a normal distribution (Map 2) when compared to the map constructed with the reparameterized t-Student distribution (Map 1) (Figure 4 and Table 4). Consequently, the 3rd and 4th classes presented a reduction, with the 3rd class obtaining a greater reduction, equivalent to 6.08%, decreasing from 37.33 to 31.25% of the area.

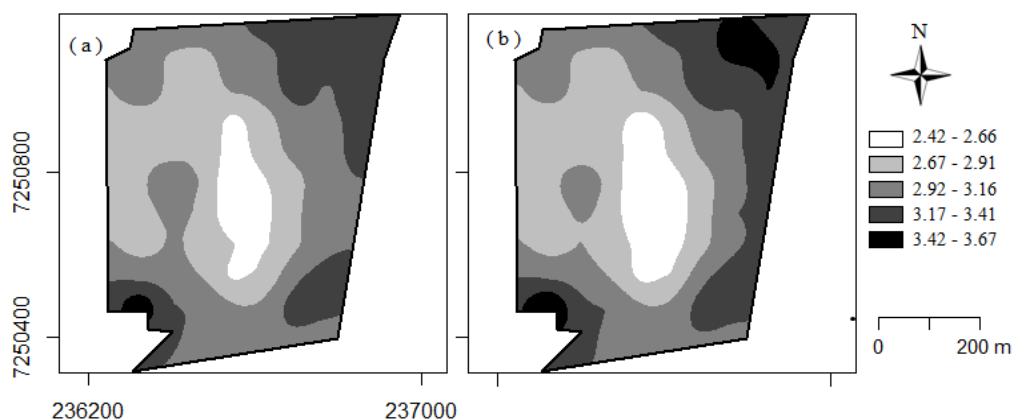


FIGURE 4. (a) Map 1: soybean yield with reparameterized t-Student distribution with  $\eta = 0.050$  and Matérn family model with shape parameter  $\kappa = 0.5$ ; (b) Map 2: soybean yield with normal distribution and Matérn family model with shape parameter  $\kappa = 0.5$ .

TABLE 4. Area percentage at each map class of soybean yield constructed with the reparameterized t-Student distribution and normal distribution.

Class (t ha <sup>-1</sup> )	Map 1 % area	Map 2 % area	Difference between maps (%)
2.42–2.66	8.07	10.33	2.26
2.67–2.91	30.68	30.80	0.12
2.92–3.16	37.33	31.25	6.08
3.17–3.41	23.25	23.20	0.05
3.42–3.67	0.67	4.31	3.64

Map 1 related to Figure 4a with the reparameterized t-Student distribution; Map 2 related to Figure 4b with the normal distribution.

For comparison between maps, the kappa accuracy index ( $K$ ) was calculated. This index is considered an appropriate measure by Anderson et al. (2001) since it uses all elements of the error matrix constructed from omission errors and designation between maps (De Bastiani et al., 2012). The obtained value of  $K = 0.64$  allows classifying it as a low similarity. Consequently, the maps constructed with reparameterized t-Student and normal distributions are dissimilar due to the influence of the discrepant point.

As a complementary analysis, a new geostatistical study was carried out by removing the point 6, which was considered as discrepant and assuming that the data presented reparameterized t-Student distribution and normal distribution. The maps constructed without the discrepant point are shown in Figure 5. The kappa accuracy index for comparison between the new maps was  $K = 0.89$ , indicating a high similarity between maps (Krippendorff, 2004). Therefore, the interference of this

discrepant point in mapping is relevant.

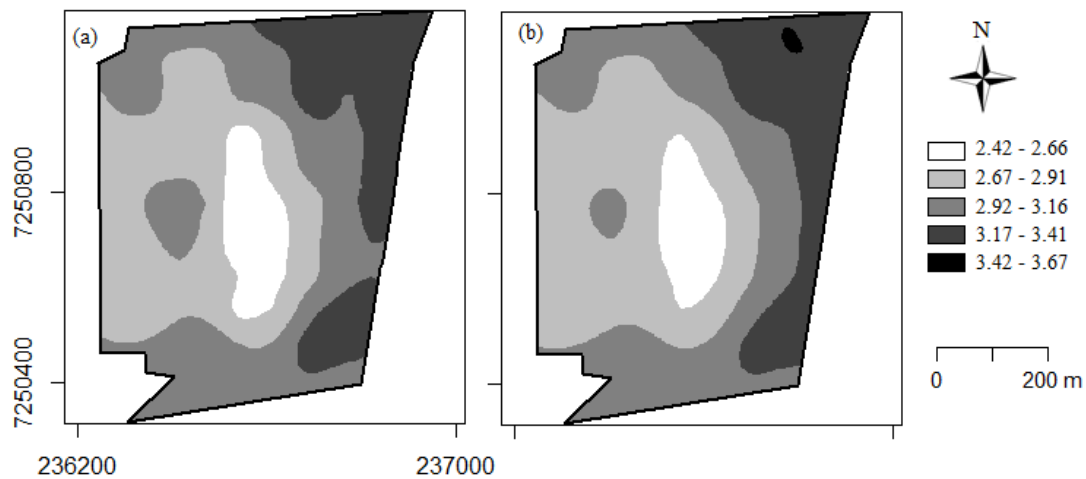


FIGURE 5. (a) Map 1: soybean yield with reparameterized t-Student distribution with  $\eta = 0.050$  and Matérn family model with shape parameter  $\kappa = 0.5$  without point 6; (b) Map 2: soybean yield with normal distribution and Matérn family model with shape parameter  $\kappa = 1.0$  without point 6.

## CONCLUSIONS

When applying the methodology proposed in this study for soybean yield data with the covariates average height and an average number of pods per plant, the parameters estimated by complete maximum likelihood using the reparameterized t-Student distribution presented differences in the estimates of parameters that define the spatial dependence structure when compared to those obtained from a normal distribution. Consequently, differences were observed in soybean yield maps obtained from the different methods. Thus, the use of reparameterized t-Student distribution is an alternative in studying data with discrepant values, showing the ability to decrease the influence of these points.

## ACKNOWLEDGEMENTS

To the CNPq, CAPES, Araucária Foundation of the Paraná state and project FONDECYT 1150325 Chile, for the financial support to develop this research.

## REFERENCES

- Anderson J, Hardy E, Roach J, Witmer R (2001) A land use and land cover classification system for use with remote sensor data.. US Geological Survey Professional, Washington, DC, US Geological Survey Professional. 41p. (Technical Report Paper 964)
- Assumpção RAB, Uribe-Opazo MA, Galea M (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(6):1917-1926. DOI: <http://dx.doi.org/10.1590/S0100-06832011000600008>
- Assumpção RAB, Uribe-Opazo MA, Galea M (2014) Analysis of local influence in geostatistics using Students t-distribution. *Journal of Applied Statistics* 41(3):615-630. DOI: <http://dx.doi.org/10.1080/02664763.2014.909793>
- Bivand, R (2015) classInt: Choose Univariate Class Intervals. R package version 0.1-23. Available: <https://cran.r-project.org/package=classInt>.
- Cressie NAC (2015) *Statistics for spatial data*. New-York, Jonh Willey & Sons.

- De Bastiani F, Uribe-Opazo MA, Dalposso GH (2012) Comparison of maps of spatial variability of soil resistance to penetration constructed with and without covariables using a spatial linear model. *Engenharia Agrícola* 32(2):394-404. DOI:<http://dx.doi.org/10.1590/S0100-69162012000200019>
- De Bastiani F, Cysneiros AHMA, Uribe-Opazo MA, Galea M (2015) Influence diagnostics in elliptical spatial linear models. *Test* 24(2):322-340. DOI:<http://dx.doi.org/10.1007/s11749-014-0409-z>
- Dempster A, Laird N, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39(1):1-38.
- EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária (2011) Manual de métodos de análise de solo. Rio de Janeiro, Embrapa Solos, 2 ed. p212.
- Galea M, Bonfarine H, Labra FV (2002) Influence diagnostics in structural errors-in-variables model under Student-t-distribution. *Journal of Applied Statistics* 29(8):1191-1204. DOI:  
<http://dx.doi.org/10.1080/0266476022000011265>
- Kano Y, Berkane M, Bentler P (1993) Statistical inference based on pseudo-maximum likelihood estimators in elliptical populations. *Journal American Statistical Association* 88(421):135-143. DOI:  
<http://dx.doi.org/10.2307/2290706>
- Krippendorff K (2004) Content analysis: an introduction to its methodology. Beverly Hills, Sage Publications.
- Lange KL, Little RJA, Taylor JMG (1989) Robust statistical modeling using the t distribution. *Journal of the American Statistics* 84(408):881-896. DOI:<http://dx.doi.org/10.2307/2290063>
- Manghi RF, Paula GA, Cysneiros FJA (2016) On elliptical multilevel models. *Journal of Applied Statistics* 43(12):2150-2171. DOI:<http://dx.doi.org/10.1080/02664763.2015.1134445>
- Matérn B (1986) Lecture notes in statistics. Springer, New York, p68-106.
- Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2015) e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. Available: <https://CRAN.R-project.org/package=e1071>.
- Michel PG, Kobiyama M (2015) Estimativa da profundidade do solo: parte 2- métodos matemáticos. *Revista Brasileira de Geografia Física* 8(4):1225-1243.
- Nesi CN, Ribeiro A, Bonat WH, Ribeiro Jr PJ (2013) Verossimilhança na seleção de modelos para predição espacial. *Revista Brasileira de Ciência do Solo* 37(2):352-358. DOI:<http://dx.doi.org/10.1590/S0100-06832013000200006>.
- Novomestky F (2012) matrixcalc: Collection of functions for matrix calculations. R package version 1.0-3. Available: <https://CRAN.R-project.org/package=matrixcalc>.
- Osorio F, Paula GA, Galea M (2007) Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics & Data Analysis Journal* 51(9):4354-4368.
- R Core Team. (2016) A language and environment for statistical computing. Vienna, Foundation for Statistical Computing.
- Ribeiro Junior PJ, Diggle PJ (2016) geoR: Analysis of Geostatistical Data. R package version 1.7-5.1. Available: <https://CRAN.R-project.org/package=geoR>.
- Uribe-Opazo MA, Borssoi JA, Galea M (2012) Influence Diagnostics in Gaussian Spatial Linear Models. *Journal of Applied Statistics* 39(3):615-630. DOI:<http://dx.doi.org/10.1080/02664763.2011.607802>
- Vieira SR (2000) Geoestatística em estudos de variabilidade espacial do solo. Tópicos em Ciências do Solo. *Revista Brasileira de Ciência do Solo* 1:1-54. DOI:<http://dx.doi.org/10.1590/S0100-06832005000200002>