

## COMPARISON MEASURES OF MAPS GENERATED BY GEOSTATISTICAL METHODS

GUSTAVO H. DALPOSSO<sup>1</sup>, MIGUEL A. URIBE-OPAZO<sup>2</sup>, ERIVELTO MERCANTE<sup>3</sup>,  
JERRY A. JOHANN<sup>4</sup>, JOELMIR A. BORSSOI<sup>5</sup>

**ABSTRACT:** This study uses several measures derived from the error matrix for comparing two thematic maps generated with the same sample set. The reference map was generated with all the sample elements and the map set as the model was generated without the two points detected as influential by the analysis of local influence diagnostics. The data analyzed refer to the wheat productivity in an agricultural area of 13.55 ha considering a sampling grid of 50 x 50 m comprising 50 georeferenced sample elements. The comparison measures derived from the error matrix indicated that despite some similarity on the maps, they are different. The difference between the estimated production by the reference map and the actual production was of 350 kilograms. The same difference calculated with the mode map was of 50 kilograms, indicating that the study of influential points is of fundamental importance to obtain a more reliable estimative and use of measures obtained from the error matrix is a good option to make comparisons between thematic maps.

**KEYWORDS:** error matrix, confusion matrix, local influence.

## MEDIDAS DE COMPARAÇÃO DE MAPAS GERADOS POR MÉTODOS GEOESTATÍSTICOS

**RESUMO:** Este trabalho utiliza diversas medidas derivadas da matriz dos erros para comparar dois mapas temáticos gerados com o mesmo conjunto amostral. O mapa de referência foi gerado com todos os elementos amostrais e o mapa definido como mapa modelo foi gerado sem dois pontos detectados como influentes por meio da análise de diagnósticos de influência local. Os dados analisados referem-se à produtividade de trigo em uma área agrícola de 13,55 há, considerando uma grade amostral de 50 x 50 m totalizando 50 elementos amostrais georreferenciados. As medidas de comparação obtidas a partir da matriz dos erros indicaram que, apesar de os mapas apresentarem certa semelhança, eles são diferentes. A diferença entre a produção estimada pelo mapa de referência e a produção real foi de 350 quilogramas. A mesma diferença calculada com o mapa-modelo foi de 50 quilogramas, indicando que o estudo de pontos influentes é de fundamental importância para obter uma estimação mais precisa, e a utilização de medidas obtidas da matriz dos erros é uma adequada opção para realizar comparações entre mapas temáticos.

**PALAVRAS-CHAVE:** matriz dos erros, matriz de confusão, influência local.

<sup>1</sup> Matemático, Prof. Assistente, UTFPR, Toledo - PR, Fone: (0xx45) 3379-6800. gustavodalposso@hotmail.com.

<sup>2</sup> Estatístico. Prof. Associado, PGEAGRI/CCET/ UNIOESTE, Cascavel-PR, Fone: (0xx45) 3220-3228. miguel.opazo@unioeste.br.

<sup>3</sup> Eng. Agrícola, Prof. Adjunto, PGEAGRI/CCET/UNIOESTE, Cascavel - PR, Fone: (0XX45) 3220-3196, eriveltomercante@yahoo.com.br.

<sup>4</sup> Eng. Agrícola, Doutorando na FEAGRI/UNICAMP, Prof. Assistente CCET/UNIOESTE, Cascavel-PR (0XX45)3220-2028, jerry.johann@hotmail.com.

<sup>5</sup> Matemático, Doutorando em Estatística, IME/USP, São Paulo - SP, Fone: (0XX11) 3091-612, joelmirb@ime.usp.br.

Recebido pelo Conselho Editorial em: 14-6-2010

Aprovado pelo Conselho Editorial em: 26-11-2011

## INTRODUCTION

In the early 90s, began to be developed technologies to manage the spatial variability associated with aspects of agricultural production. Productivity varies spatially, and to determine the causes of this variation is the challenge that faces the precision agriculture. The processing and the integration of data, in precision agriculture are usually made in order to select and model the variables that best explain the productivity (JOHANN et al., 2002).

The techniques of localized application and varying rates of agricultural inputs with the help of geostatistics has been one of the areas to focus efforts in recent research (CORA & BERALDO, 2006, SOUZA et al., 2007), and facing this, it is desirable the study of methodologies that allows the comparison of thematic maps generated by different techniques, because only a visual comparison is subjective and important differences between maps can be ignored.

The aim of this study is to use measures derived from the error matrix to compare two thematic maps constructed by geostatistical methods, being one map generated with all samples and defined as a reference map and the other map is generated without two points detected as influential by the diagnostic technique of local influence and defined as model.

## MATERIALS AND METHODS

The experimental area covers a grain production field with 13.55 hectares, located in the city of Salto do Lontra - PR, which was cultivated with the wheat cultivar IAPAR 78. In the monitored area, was defined a regular sampling grid of 50 x 50 m a total of 50 georeferenced locations (Figure 1 (A)).

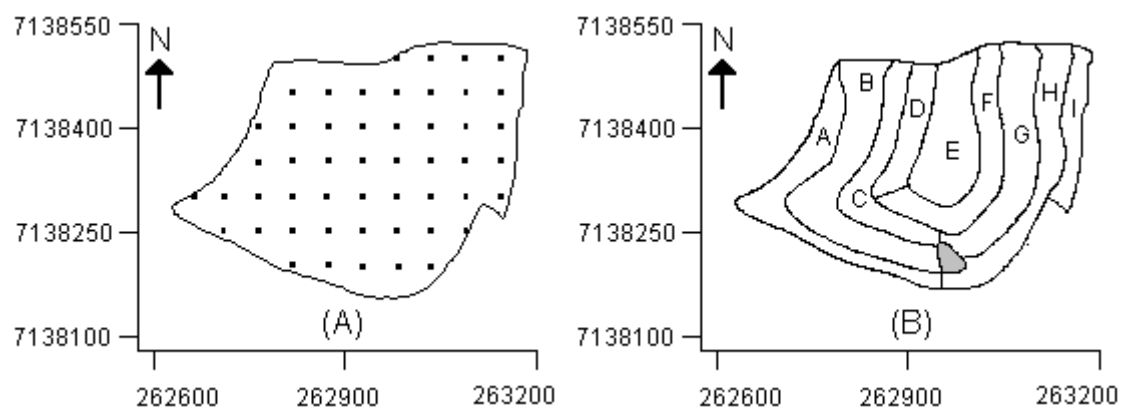


FIGURE 1. (A) Location of the 50 sampling points and (B) Area plots division, the gray area was not used in the experiment.

In Figure 1 (B) it is observed that the area was divided into nine plots to facilitate mechanical harvesting, making possible to determine the wheat real production. The area shaded in Figure 1 (B) represents a region of rocks that is not monitored in the experiment. In the sampling points were collected the wheat plots in an area of 1 m<sup>2</sup>. These plots were threshed, weighed and converted into t ha<sup>-1</sup>. To perform the geostatistical analysis it was used the software R (R DEVELOPMENT CORE TEAM, 2009) and its GeorR package (RIBEIRO JUNIOR & DIGGLE, 2001).

To model data with a spatial structure, it was considered a Gaussian stochastic process  $\{Z(s), s \in S\}$ , with  $S \subset \mathbb{R}^2$ , being  $\mathbb{R}^2$  two-dimensional Euclidean space. Supposing that the data,  $Z(s_1), \dots, Z(s_n)$ , of this process are recorded in known spatial locations  $s_i$  ( $i = 1, \dots, n$ ), and generated by the model  $Z(s_i) = \mu(s_i) + \varepsilon(s_i)$ , where the deterministic terms  $\mu(s_i)$  and stochastic  $\varepsilon(s_i)$  may depend on the spatial location where  $Z(s_i)$  was obtained. It is assumed that the stochastic error  $\varepsilon(\cdot)$  has mean zero,  $E[\varepsilon(s_i)] = 0$ , that the variation between points in the space is determined by some covariance function  $C(s_i, s_u) = \text{COV}[\varepsilon(s_i), \varepsilon(s_u)]$  and that for some known functions of  $s$ , as  $x_1(s), \dots, x_p(s)$ , the

mean is defined as  $\mu(s) = \sum_{u=1}^p x_u(s) \beta_u$ , being,  $\beta_1, \dots, \beta_p$  unknown parameters to be estimated. In an equally form and in a matrix form, we have  $Z = X\beta + \varepsilon$ , being,  $E(\varepsilon) = 0$  (null vector) and the covariance matrix of  $\Sigma = [(\sigma_{iu})]$ , in which  $\sigma_{iu} = C(s_i, s_u)$ . It is assumed that  $\Sigma$  is nonsingular that  $X$  has columns with full rank and that  $Z$  follows abnormal  $n$ -variant distribution with mean vector  $X\beta$  and covariance matrix  $\Sigma$ .

Considering the parametric form of the covariance matrix  $\Sigma = \varphi_1 I_n + \varphi_2 R$ , (MARDIA & MARSHALL, 1984), where  $\varphi_1$  is the parameter known as the nugget effect,  $\varphi_2$  is known as the contribution parameter,  $R = R(\varphi_3) = [r_{iu}]$  is an  $n \times n$  symmetric matrix, with diagonal elements  $r_{ii} = 1$ ,  $i = 1, \dots, n$ , where  $\varphi_3$  is the range function ( $a$ ) of the model, and  $I_n$  is an  $n \times n$  identity matrix. The parametric form of the covariance matrix  $\Sigma$ , occurs for various isotropic processes, where the covariance  $C(s_i, s_u)$  is defined according to the covariance function  $C(h_{iu}) = \varphi_2 r_{iu}$ , where  $h_{iu}$  is the Euclidean distance between the points  $s_i$  and  $s_u$ . In the covariance functions  $C(h_{iu})$ , the variance of the stochastic process  $Z$  is  $C(0) = \varphi_1 + \varphi_2$ , and can be defined as semivariance  $\gamma(h) = C(0) - C(h)$ .

In this study for the adjustment of a spatial model to the experimental semivariogram it was used the exponential, spherical and Gaussian models, for the parameters estimation it was used the ordinary least squares (OLS), the weighted least squares (WLS1), the maximum likelihood (ML) and the restricted maximum likelihood (RML) estimation methods (MARDIA & MASHALL, 1984). To choose the space model that best suits the semivariances it was used the cross-validation technique (VAUCLIN et al., 1983; FARACO et al., 2008) and to investigate the existence of influential points it was carried out a diagnostic analysis of local influence (BORSSOI et al., 2009).

For a set of observed data  $l(\theta)$  is the log-likelihood function of the postulated model, where  $\theta = (\beta, \varphi_1, \varphi_2, \varphi_3)^T$  and that  $\omega$  a disturbance vector belonging to a disturbance space  $\Omega$ . And that  $l(\theta/\omega)$  is the logarithm of the likelihood function corresponding to the perturbed model by  $\omega \in \Omega$ . It is assumed that there is a  $\omega_0 \in \Omega$  such that  $l(\theta) = l(\theta/\omega_0)$ , for all  $\theta$  and that  $l(\theta/\omega)$  is twice differentiable on  $(\theta^T, \omega^T)^T$ . Consider the following perturbation scheme:  $Z_\omega = Z + \omega$ , with  $\omega = (\omega_1, \dots, \omega_n)^T$  response perturbation vector and  $\omega_0 = (0, \dots, 0)^T$  the point of no perturbation. With this perturbation scheme is intended to detect possible outliers in the data that affect the maximum likelihood estimator of  $\theta$ . The  $\omega$  perturbation influence in the ML estimator of the  $\theta$  parameter vector can be evaluated by the likelihood difference, defined by  $LD(\omega) = 2(l(\hat{\theta}) - l(\hat{\theta}_\omega))$ , where  $\hat{\theta}$  is the ML estimator of  $\theta$  of the assumption model and  $\hat{\theta}_\omega$  is the ML estimator of  $\theta$  of the disturbed model. COOK (1986) proposed to study the local behavior of the LD ( $\omega$ ) around  $\omega_0$ , using the normal curve of LD ( $\omega$ ) in  $\omega_0$  in the direction of some unitary vector  $l$ , defined  $C_l = 2|l^T \Delta^T L^{-1} \Delta l|$ , with  $\|l\| = 1$ , where  $L$  is the observed information matrix, evaluated in  $\theta = \hat{\theta}$ ,  $\Delta$  is a matrix  $(p + q) \times n$  given by

$$\Delta = (\Delta_\beta^T, \Delta_\varphi^T)^T, \text{ evaluated in } \theta = \hat{\theta} \text{ and in } \omega = \omega_0, \quad (01)$$

In witch,

$$\Delta_\beta = X^T \Sigma^{-1} \text{ and } \Delta_\varphi = \frac{\partial^2 l(\theta/\omega)}{\partial \varphi \partial \omega^T} \quad (02)$$

$$\text{with, } \frac{\partial^2 l(\theta/\omega)}{\partial \varphi \partial \omega^T} = (Z_\omega - X\beta)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} \quad (03)$$

$$\text{Where the observed information matrix is } L = \begin{pmatrix} L_{\beta\beta} & L_{\beta\varphi} \\ L_{\varphi\beta} & L_{\varphi\varphi} \end{pmatrix} \quad (04)$$

$$\text{that is, } L_{\beta\beta} = -(X^T \Sigma^{-1} X) \quad (05)$$

$$L_{\beta\varphi} = \frac{\partial^2 l(\theta)}{\partial \beta \partial \varphi^T}, \text{ with } \frac{\partial^2 l(\theta)}{\partial \beta \partial \varphi_j} = -X^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_j} \Sigma^{-1} \varepsilon, \text{ being } \varepsilon = (Z_\omega - X\beta)^T \quad (06)$$

$$L_{\varphi\beta} = L_{\beta\varphi}^T \text{ and } L_{\varphi\varphi} = \frac{\partial^2 l(\theta)}{\partial \varphi \partial \varphi^T} \quad (07)$$

with

$$\frac{\partial^2 l(\theta)}{\partial \varphi_i \partial \varphi_j} = \frac{tr}{2} \left\{ \Sigma^{-1} \left( \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_j} - \frac{\partial^2 \Sigma}{\partial \varphi_i \partial \varphi_j} \right) \right\} + \frac{\varepsilon^T \Sigma^{-1}}{2} \left\{ \frac{\partial^2 \Sigma}{\partial \varphi_i \partial \varphi_j} - \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_j} - \frac{\partial \Sigma}{\partial \varphi_j} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \right\} \Sigma^{-1} \varepsilon \quad (08)$$

$L_{\max}$  is the eigenvector, normalized, associated with the largest eigenvalue, in modulus, from the matrix  $B = \Delta^T L^{-1} \Delta$ . The graphic elements  $|L_{\max}|$  versus  $i$  (order) may reveal what kind of perturbation has the greatest influence in LD ( $\omega$ ) in the vicinity of  $\omega_0$  (COOK, 1986; BORSSOI et al., 2009).

Table 1 shows the generic form of an error matrix (GRINAND et al., 2008). In this matrix, the pixels of the reference map are quantified in columns while the pixels of the model map are quantified in the lines. Each matrix element represents the number of pixels belonging to the class  $C_i$ ,  $i=1, \dots, k$ , of the model map and the class  $C_j$ ,  $j=1, \dots, k$ , of the reference map.

TABLE 1. Generic error matrix of  $k \times k$  order.

Pixels of the model map	Pixels of the reference map			
	Classes	$C_1$	$C_2$	... $C_k$
				Total per line
				$n_{i\bullet}$
	$C_1$	$n_{11}$	$n_{12}$	... $n_{1k}$
	$C_2$	$n_{21}$	$n_{22}$	... $n_{2k}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$C_k$	$n_{k1}$	$n_{kk}$	... $n_{kk}$
	Total per column	$n_{\bullet 1}$	$n_{\bullet 2}$	... $n_{\bullet k}$
		$n_{\bullet i}$		Grand total
				$n$

$K$ : class number;  $C_i$ : class  $i$ ;  $n_{i\bullet}$ : total of pixels in  $C_i$  class of the model map;  $n_{\bullet i}$ : total of pixels in  $C_i$  class of the reference map.

The elements of the main diagonal (when  $i = j$ ) represent cases where the pixels have the same classification in both maps while the elements out of the diagonal represent the main erroneous classifications. Using the error matrix elements, the following indexes were calculated; the Global Accuracy (GA) (BACH et al., 2006), the user accuracy ( $UA_i$ ) (ELLIS & PORTER-BOLLAND, 2008), the producer accuracy ( $PA_i$ ) LIU et al., 2007), the Kappa ( $\hat{K}$ ) (JENNESS & WYNNE, 2005, LI et al, 2008) and the Tau ( $T$ ) (RIEGL & PURKIS, 2005), presented respectively from Equations (9) to (13).

$$GA = \frac{\sum_{i=1}^k n_{ii}}{n} \quad (09)$$

$$UA_i = \frac{n_{ii}}{n_{i\bullet}} \quad (10)$$

<sup>1</sup>Assemblage of the words Picture and Elements. Represents the smallest element of a map to which is possible to assign a value. In this study, the pixels will be defined as the matrix elements of the Kriger values.

$$PA_i = \frac{n_{ii}}{n_{\bullet i}} \quad (11)$$

$$\hat{K} = \frac{n \sum_{i=1}^k n_{ii} - \sum_{i=1}^k n_{i\bullet} n_{\bullet i}}{n^2 - \sum_{i=1}^k n_{i\bullet} n_{\bullet i}} \quad (12)$$

$$T = \frac{\frac{\sum_{i=1}^k n_{ii}}{n} - p_k}{1 - p_k} \quad (13)$$

The global accuracy (*GA*) is a statistical assessment used to measure the similarity between something real or of reference and an adjusted model, however, this measure is to be presented in association with other measures that take into account other values contained in the error matrix and not only the main diagonal elements. The user accuracy index (*UA<sub>i</sub>*) represents the ratio between the number of pixels correctly classified in class *i* over the total number of pixels classified in class *i* of the model map and the producer accuracy index (*PA<sub>i</sub>*) is an statistic that gives the probability of a pixel to be classified in class *i* if it really belong s to class *i*.

The Kappa index ( $\hat{K}$ ) usually ranges between 0 and 1, with values close to 1 reflecting greater agreement (ALMEIDA & VIEIRA, 2008; LI et al, 2008). According to the classification of KRIPENDORFF (1980),  $\hat{K}$  is classified with low accuracy if  $\hat{K} < 0.67$ , average accuracy if  $0.67 \leq \hat{K} < 0.80$  and high accuracy if  $\hat{K} \geq 0.80$ . The Tau index (*T*) is similar to Kappa, however, differentiates by using in its formulation the probabilities priori for each class. When these probabilities a priori are equal, we have  $p_i = 1/k$ , where *k* represents the number of classes. Another way to compare thematic maps is provided by measurements obtained from the confusion matrix per class (JENNESS & WYNNE, 2005), this matrix consists of four elements,  $a_k = n_{kk}$ ,  $b_k = n_{\bullet k} - n_{kk}$ ,  $c_k = n_{k\bullet} - n_{kk}$  and  $d_k = n - (a_k + b_k + c_k)$ , obtained from the error matrix. The sensitivity index  $S_k = a_k/(a_k + c_k)$  is a measure that indicates the probability of a pixel in the model map be classified as belonging to class *k* if it actually belongs to class *k* in the reference map and is therefore an equivalent measure to the producer accuracy. The specificity index  $E_k = d_k/(b_k + d_k)$  indicates the probability of a pixel not belonging to the class *k* in the reference map being classified as not belonging to class *k* in the model map. The false positive rate  $TFP_k = b_k/(b_k + d_k)$  represents the commission errors and indicates the proportion of pixels that do not belong to class *k* in the reference map that are classified as belonging to class *k* in the model map. The false negative rate  $TFN_k = c_k/(a_k + c_k)$  represents the omission errors and indicates the proportion of pixels that belong to the class *k* of the reference map and were classified in other classes in the model map. Based on these measures, it is possible to compare individually the classes of the model map with the classes of the reference map. For a global comparison of the maps, JENNESS & WYNNE(2005) present

the total confusion matrix, in which the values  $a = \sum_{i=1}^k n_{ii}$  (True Positive)

and  $d = \sum_{i=1}^k \sum_{i \neq j} \sum_{j \neq i} n_{ij}$  (True Negative) are correct estimates, whereas the values  $b = \sum_{i=1}^k \sum_{j \neq i} n_{ij}$  (False

Positive) and  $c = \sum_{j=1}^k \sum_{i \neq j} n_{ij}$  (False Negative) are considered estimation errors. The errors of type *b* are

also known as commission or overestimation errors, while errors of type *c* are known as omission errors.

Table 2 presents some measurements of the total confusion matrix. The total sensitivity ( $S$ ) and the total specificity ( $E$ ) index are analogous in its versions per class; with the difference that now they show measures for the full map. The Matthews correlation coefficient (MCC) is a discrete version of the Pearson correlation coefficient and its values are found in the interval  $[-1, 1]$ , that is the value 1 is equivalent to a perfect prediction, 0 is equivalent to a random prediction and -1 means an inverse prediction.

TABLE 2. Measures derived from total confusion matrix.

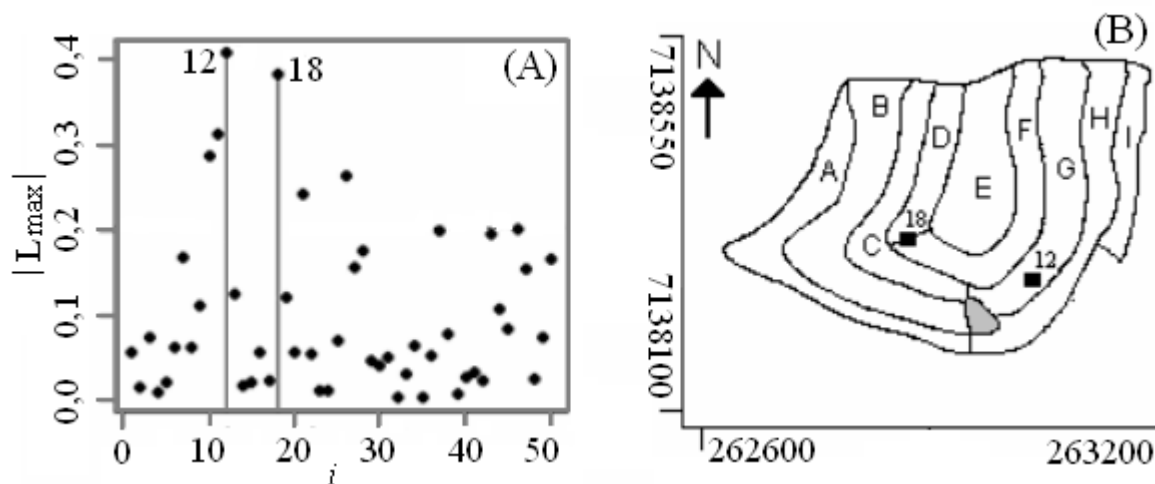
Measures	Equation
$S$	$a/(a+c)$
$E$	$d/(b+d)$
CCM	$((a \cdot d) - (b \cdot c)) / ((a+b) \cdot (a+c) \cdot (d+b) \cdot (d+c))^{1/2}$

## RESULTS AND DISCUSSION

The samplings exploratory analysis had an wheat average yield of  $1.31 \text{ t ha}^{-1}$ , and the minimum production of  $0.30 \text{ t ha}^{-1}$  and a maximum of  $2.40 \text{ t ha}^{-1}$  with standard deviation of  $0.54 \text{ t ha}^{-1}$  and coefficient of variation (CV) of 41.40%, considering an heterogeneity among the sample values of the wheat productivity.

A variographic study was carried out using the whole sample and was concluded by the cross-validation technique that the best fitted model was an exponential with a parameter estimation method of maximum likelihood (ML).

It was conducted a variographic study using the whole sample and found the cross-validation technique that best fitted model was an exponential with parameter estimation method of maximum likelihood (ML). Picture 2 (A) shows a graph of influence of the self-location vector  $|L_{\max}|$  versus  $i$  that it is found that the sampling elements 12 and 18 were considered influential, and thus can change some kind of decision in the construction of models geostatistical and /or construction of thematic maps. The figure 2(B) shows the location of influential points in the monitored area.

FIGURE 2. Diagnostic graph  $|L_{\max}| \times i$ ; (B) Localization of the influential points.

Thus, it was decided to make a new geostatistical analysis by removing the points considered influential. The criterion of cross-validation indicated that the best adjusted model was the Gaussian with estimation method parameters restricted maximum likelihood (RML). Figure 3 shows the semivariograms adjusted with and without the influential points and their parameters ( $\phi_1$ ,  $\phi_2$ ,  $a$ ).

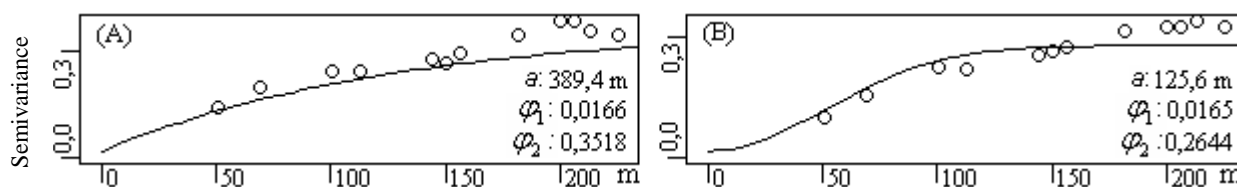


FIGURE 3. Adjusted semivariograms: (A) Complete set: exponential model – ML and (B) Set without influential points: Gaussian model – RML. Parameters –  $a$ : range,  $\phi_1$ : nugget effect and  $\phi_2$ : contribution.

It is noteworthy that despite the exponential model-ML present a greater range of spatial dependence (389.4 m) compared to the Gaussian model-RML (125.6 m), both models show a strong spatial dependence. Figure 4 (A) shows the thematic map generated using the exponential model-ML and Figure 4 (B) shows the thematic map generated using the Gaussian model-RML. Both maps were generated by ordinary kriging interpolation using a grid composed of 5423 pixels classified as the same caption using classes  $C_1 = [0,371; 0,784]$ ,  $C_2 = [0,784; 1,198]$ ,  $C_3 = [1,198; 1,611]$ ,  $C_4 = [1,611; 2,024]$  e  $C_5 = [2,024; 2,437]$ .



FIGURE 4. (A) Thematic map of the wheat's productivity using 50 sampling units in the interpolation and (B) Thematic map of the wheat's productivity using 48 sampling units in the interpolation.

The maps were constructed using five classes, this number being determined in such a manner that enables both a visual identification of areas of productivity as convenience to perform localized applications of inputs, for a large number of intervals results in very small regions, making handling found unfeasible. It is observed in Figure 4 that the regions of greatest difference are the regions of influential points, however, various other regions differ, indicating that the withdrawal of influential points affected the maps as a whole. To better compare the maps, it is appropriate to quantify their 5423 pixels in an array of errors (Table 3), defining the map generated with all the points as the reference map and the map generated without influential points as the model map.

TABLE 3. Error matrix of wheat's productivity maps.

		Pixels of reference map					TOTAL
		$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	
Pixels of model map	$C_1$	485	272	9	0	0	766
	$C_2$	45	1202	124	18	0	1389
	$C_3$	7	206	1218	99	19	1549
	$C_4$	0	0	135	616	241	992
	$C_5$	0	0	0	135	592	727
TOTAL		537	1680	1486	868	852	5423

In the matrix of the errors, the pixels of the reference map are measured in columns and the pixels of the model map are quantified in the lines. Thus, it has that from the 5423 pixels in each

map 537 pixels are in the first class of the reference map and 766 pixels are in the first class of the model map. This first information allows realizing an estimate of the wheat production in the area to be compared with the actual grain yield, collected and weighed by the manufacturer, which was 18.49 tons. The total area of 13.551 hectares was monitored and corresponds to 5423 pixels. Performing an arithmetic operation, it is found that the 537 pixels of the first class of the reference map represent an area of 1.3419 hectares. Multiplying this area by the midpoint of the first class (0.5778), it is obtained the estimate of 0.78 tons. Following the same procedure for the other classes, it is obtained the estimated production in tons of the area in Table 4.

TABLE 4. Estimated production in each subtitle class.

Map	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	Total
Reference	0.78	4.16	5.21	3.94	4.75	18.84
Model	1.11	3.44	5.43	4.50	4.05	18.54

It is noted in Table 4 that the production estimated by reference map was 18.84 tons, 350 kg above the real production and estimated production using the map model was 18.54 tons, 50 kg more than actual production. Thus, it is emphasized that the withdrawal of influential points was of great importance in the analysis because it allowed the estimated production with better precision. The matrix of errors is obtained from Table 3 that the EG index = 0.76, ie the overall accuracy was 76%. This value indicates that the thematic maps differ, because if they were equal, the overall accuracy would be 1. Accuracy can be investigated for each of the five classes, using the contents of Table 5.

TABLE 5. User ( $AU_i$ ) and producer ( $AP_i$ ) accuracy indexes

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$
$AU_i$	0.6432	0.8654	0.7863	0.6210	0.8143
$AP_i$	0.9032	0.7155	0.8197	0.7097	0.6948

It is noted from Table 5 that the  $AU_4$  index indicates that 37.9% of the pixels that belong to the class  $C_4$  on the model map belong to other classes in the reference map and the index  $AP_4$  indicates that the probability of a pixel to be classified in class  $C_4$  of the model map if it really belongs to this class in the reference map is 0.7097. Using the elements of Table 3 calculated the Kappa index = 0.69 and considering the equality of the a priori probabilities of each class it was calculated the index Tau T = 0.70. These indexes were close, indicating an average accuracy of the maps, according to the classification of the Kappa (KRIPENDORFF, 1980). Table 6 presents the measurements of the confusion matrices of each class  $C_i$ . It stands out with the class  $C_5$  is the lowest sensitivity and highest specificity, so the probability of a pixel to be classified in class  $C_5$  on the model map if this pixel belongs to the class  $C_5$  of the reference map, is the lowest if compared to other classes and the probability of a pixel of the model map does not present yield in the range [2.024, 2.437] as the pixel has no productivity in the same range reference map is greater when compared with other classes.

TABLE 6. Measures obtained from the confusion matrix of each class.

Classes	$S_k$	$E_k$	$TFP_k$	$TFN_k$
$C_1$	0,9032	0,9425	0,0575	0,0968
$C_2$	0,7155	0,9500	0,0500	0,2845
$C_3$	0,8197	0,9159	0,0841	0,1803
$C_4$	0,7097	0,9175	0,0825	0,2903
$C_5$	0,6948	0,9705	0,0295	0,3052

$S_k$ : Sensitivity,  $E_k$ : Specificity,  $TFP_k$ : False positive rate e  $TFN_k$ : False negative rate.



It is observed that the biggest mistake of over estimating ( $TFP_k$ ) occurs in the class  $C_3$  and the largest error of omission ( $TFN_k$ ) occurs in the class  $C_5$ . Using the elements of the matrix of errors in Table 3, it is obtained the following complete confusion matrix:  $a = 4113$ ,  $b = 1310$ ,  $c = 1310$  and  $d = 20382$ . Using these values obtains  $S = 0.76$ ,  $E = 0.94$  and  $CCM = 0.70$ . An important feature of the total sensitivity index (S) is that it equals the rate of global accuracy (GA), so the overall capacity on the model map correctly classified pixels is 0.76. The total level of specificity (E) indicates that the capacity on the model map to avoid incorrect treating is 0.94. The correlation of Matthews's coefficient (CCM) obtained was 0.70 and as this index is value 1 if the maps are the same, it appears that the model map differs as compared with the generated map with all points.

An important issue that must be addressed is to define a priori the number of classes, as the increase of these implies a decrease in the similarity between the maps. For example, considering 7, 10, 15 and 20 classes, it is obtained respectively indices of  $GA = 0.72$ ,  $GA = 0.61$ ,  $GA = 0.50$ ,  $GA = 0.40$  that are smaller than the overall accuracy considering five classes. Considering that one of the functions of the thematic map is to indicate locations for a localized management, it is important to choose a number of classes that makes this work possible. MONMOIER (1993) recommends that the number of ideal classes is around five classes and never more than seven classes, because it is preferable to represent with few groups with symbols of areas graphically stable rather than risk a fine representation of difficult visualization.

## CONCLUSIONS

The results showed that the measures derived from the error matrix if the confusion matrix is appropriate to compare thematic maps, as they provide global values and still allow making comparisons by class. Considering that the wheat production is estimated by the model map close to actual production obtained, it is concluded that it is fundamentally important to carry out research on influential points, and the technique of local influences appropriate for that function.

## ACKNOWLEDGEMENTS

CNPq, CAPES and Fundação Araucária by the financial support.

## REFERENCES

- ALMEIDA, A. S.; VIEIRA, I. C. G. Dinâmica da cobertura vegetal e uso da terra no município de São Francisco do Pará (Pará, Brasil) com o uso da técnica de sensoriamento remoto. *Boletim do Museu Paraense Emílio Goeldi de Ciências Naturais*, Belém, v.3, n.1, p.81-92, 2008.
- BACH M.; BREUER L.; FREDE H. G.; HUISMAN J. A.; OTTE A.; WALDHARDT R. Accuracy and congruency of three different digital land-use maps. *Landscape and urban planning*. Amsterdam, v.78, n.4, p. 289-299, 2006.
- BORSSOI, J. A.; URIBE-OPAZO, M. A.; GALEA, M. Diagnostic techniques applied in geostatistics for agricultural data analysis. *Revista Brasileira de Ciência do Solo*, Viçosa-MG, v. 33, n.6, p. 1561-1570, 2009.
- COOK, R. D. Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, London, v. 48, n.2, p. 133-169, 1986.
- CORA, J. E.; BERALDO, J. M. G. Variabilidade espacial de atributos do solo antes e após calagem e fosfatagem em doses variadas na cultura de cana-de-açúcar. *Engenharia Agrícola*, Jaboticabal, v. 26, n. 2, p. 374-387, 2006.
- ELLIS E. A.; PORTER-BOLLAND L. Is community-based forest management more effective than protected areas? A comparison of land use/land cover change in two neighboring study areas of the Central Yucatan Peninsula, Mexico. *Forest Ecology and Management*, Melbourne, v. 256, n. 11, p. 1971-1983, 2008.

- FARACO, M. A.; URIBE-OPAZO, M. A.; SILVA E. A.A.; JOHANN, J. A.; BORSSOI, J. A. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. *Revista Brasileira de Ciência do Solo*, Viçosa-MG, v.32, n.2, p. 463-476, 2008.
- GRINAND, C.; ARROUAYS D.; LAROCHE B.; MARTIN M. P. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *GEODERMA*, Amsterdam, v.143, n.1-2, p. 180-190, 2008.
- JENNESS, J.; WYNNE, J. J. *Cohen's Kappa and classification table metrics 2.0: an ArcView 3x extension for accuracy assessment of spatially explicit models*. Flagstaff: U.S. Geological Survey, 2005. 86 p.
- JOHANN, J. A.; URIBE-OPAZO, M. A.; ROCHA, J.V.; SOUZA, E.G. de . Modelagem da correlação espacial entre mapas de produtividade da soja e mapas de atributos do solo por meio de análise de componentes principais. *Engenharia Agrícola*, Jaboticabal, v. 22, n. 3, p. 383-390, 2002.
- KRIPPENDORFF, K. *Content analysis: an Introduction to its Methodology*. Beverly Hills: Sage Publications. 1980. 189p.
- LI, Y.; LINDE, K.; HAGER, S.; DAÍ, J.; MELCHART, D. Assessing morphological characteristics within traditional Chinese tongue diagnosis—A reliability study. *European Journal of Integrative Medicine*, Netherlands, v. 1, Supp. 1, p. 10, 2008.
- LIU C.; FRAZIER P.; KUMAR L. Comparative assessment of the measures of thematic classification accuracy. *Remote sensing of environment*, New York, v. 107, n. 4, p. 606-616, 2007.
- MARDIA, K.; MARSHALL, R. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, Great Britain, v.71, n.1, p.135-146, 1984.
- MONMONIER, M. *Mapping it out: expository cartography for the humanities and social sciences*. Chicago: University of Chicado Press, 1993. 301p.
- RIEGL, B. M.; PURKIS, S. J. Detection of shallow subtidal corals from IKONOS satellite and QTC View (50, 200 kHz) single-beam sonar data (Arabian Gulf; Dubai, UAE). *Remote Sensing of Environment*, New York, v. 95, n. 1, p.96-114, 2005.
- R DEVELOPMENT CORE TEAM . *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2009. Disponível em: <<http://www.R-project.org>>.
- RIBEIRO JUNIOR, P. J.; DIGGLE, P. J. GeoR: a package from geoestatistical analysis. *R-New*, v. 1, n. 2, p. 15-18, 2001.
- SOUZA, Z. M.; BARBIERI, D.M.; MARQUES JR, J.; PEREIRA, G.T.; CAMPOS, M. C. Influência da variabilidade espacial de atributos químicos de um latossolo na aplicação de insumos para cultura de cana-de-acúcar. *Ciência e Agrotecnologia*, Lavras, v.31, n.2, p. 371-377, 2007.
- VAUCLIN, M.; VIEIRA, S. R.; VACHAUD, G.; NIELSEN, D. R. The use of cokriging with limited field soil observations. *Soil Science Society of America Journal*, Madison, v. 47, n.1, p.175-184, 1983.