Engenharia Agrícola

ISSN: 1809-4430 (on-line)

www.engenhariaagricola.org.br

Doi: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v39nep56-65/2019

Special Issue: Precision Agriculture

REDUCTION OF SAMPLE SIZE IN THE ANALYSIS OF SPATIAL VARIABILITY OF NON-STATIONARY SOIL CHEMICAL ATTRIBUTES

Tamara C. Maltauro[1*], Luciana P. C. Guedes[2], Miguel A. Uribe-Opazo[2]

[1*]Corresponding author. Universidade Estadual do Oeste do Paraná/ Cascavel - PR, Brasil.
E-mail: tamara_ma02@hotmail.com | ORCID ID: http://orcid.org/0000-0003-2682-8159

KEYWORDS

Fisher information matrix, genetic algorithm, geostatistics, spatial dependence.

ABSTRACT

In the study of spatial variability of soil attributes, it is essential to define a sampling plan with adequate sample size. This study aimed to evaluate, through simulated data, the influence of parameters of the geostatistical model and sampling configuration on the optimization process, and resize and reduce the sample size of a sampling configuration of a commercial area composed of 102 points. For this, an optimization process called genetic algorithm (GA) was used to optimize the efficiency of the geostatistical model estimation based on the Fisher information matrix. The simulated data evidenced that the variation of the nugget effect or practical range did not significantly alter the sample size. GA was efficient in reducing the sample size, determining for soil chemical attributes a sample size between 30 and 40 points (29.41 to 39.22% of the initial sampling grid). The presence of spatial dependence was observed for all soil chemical attributes in the two sampling configurations (initial and optimized). The optimized sampling configuration evidenced an increase in trend intensity in the north direction and a more efficient estimation of parameters of the linear spatial regression model.

INTRODUCTION

Soil quality is essential to sustainable development and preservation of ecosystems and biodiversity, and the variability of soil chemical attributes is influenced by differences in interactions between soil formation factors and processes, which contribute to the existence of spatial variability of crops (Artur et al., 2014). In addition, it is important for the cultivation system to reduce costs of applying inputs and possibilities of environmental problems in order to improve the management of the production process and maximize the profitability of production (Bernardi et al., 2014).

Geostatistical techniques allow studying the spatial variability of georeferenced attributes (Cressie, 2015). Thus, understanding the spatial variability of soil is important for planning a soil sampling configuration and crop management (Cherubin et al., 2014).

A reduced sampling plan, i.e., a sampling configuration with the smallest possible size, is important in experiments that involve the spatial variability, allowing a reduction of operational costs and minimization of quality loss of the obtained results (Guedes et al., 2014; Siqueira et al., 2014).

There are several traditional methodologies of spatial sampling that can be used to study the spatial variability of soil and select a sample size, such as stratified (Wang et al., 2012), systematic (Guedes et al., 2011; Wang et al., 2012; Cherubin et al., 2015), random (Guedes et al., 2011; Wang et al., 2012), lattice plus close pairs (Chipeta et al., 2017), and lattice plus infill samplings (Chipeta et al., 2017; Cheng et al., 2018). In contrast to traditional samplings that use a fixed number of samples, there is the sequential sampling in which the sample size increases item by item until it reaches a conclusion in order to accept or reject a hypothesis (Santos et al., 2017).

Moreover, the choice of a configuration and a sample size can be defined as an optimization problem. This methodology is used in the context of redefinition of a sampling configuration obtained from known information of an initial sampling configuration, in which a sampling configuration that minimizes the loss of information on the

[2] Universidade Estadual do Oeste do Paraná/ Cascavel - PR, Brasil.

Area Editor: Fabio Henrique Rojo Baio
Received in: 3-26-2019
Accepted in: 6-4-2019

results of the analyses should be chosen (Guedes et al., 2014). One of these optimization processes is called genetic algorithm (GA), which consists of a search technique based on the process of evolution and adaptation of individuals of a population so that the fit ones remain in this population (Pessoa et al., 2015).

In addition, the process of resizing sampling configurations must consider a search criterion known as the objective function, which is minimized or maximized and expresses the optimization efficiency. There are criteria of optimization efficiency based on spatial prediction (mean or weighted variance, sum of the quadratic error, measure of accuracy, overall accuracy, etc.) (Guedes et al., 2011; Guedes et al., 2016; Szatmári et al., 2018), as well as criteria that consider the efficiency as the geostatistical model estimation, such as the objective function based on the inverse-Fisher information matrix (Zhu & Stein, 2005).

Previous studies involving optimized sampling configurations only optimized the sample size or sampling configuration. A methodology to simultaneously optimize sample size and sampling configuration was obtained by Guedes et al. (2014), who used the optimization algorithm called simulated annealing. However, the simulated annealing has as a disadvantage the direct relationship of the computational cost and the number of samples in the initial configuration.

Moreover, these studies have used georeferenced stationary variables, i.e., the average of the georeferenced variable throughout the area is constant. However, stationarity is not a characteristic not always identified in soil properties (Szatmári et al., 2018).

Considering non-stationary simulated and real data (soil chemical attributes), this study aimed (a) to evaluate the influence of parameters of the geostatistical model and the initial sampling configuration used in the optimization process; and (b) to propose and evaluate the resizing of a sampling configuration, aiming at reducing its sample size for a commercial area of soybean cultivation.

MATERIAL AND METHODS

Initially, a simulation study was carried out to reproduce a set of possibilities in the real data to be evaluated in this research, as well as to extend the theoretical-practical knowledge on the optimization of size and sampling configuration in soil chemical properties with non-stationary spatial dependence structure.

Study of simulations

Nine non-stationary simulated data sets were generated to combine parameters of the geostatistical model with low, medium, and high radius (range) and intensity (relationship between the nugget effect and sill) of spatial dependence. Simulations were generated with reference to the sampling configuration of the agricultural area considered in the practical study. The lattice plus close pairs configuration, composed of 100 sample points distributed in a 9 × 9 regular sampling grid with addition of 19 nearby points, which were randomly added to the regular grid, showing lower distances with some grid points than that between points of the regular grid, wad used. For this, a square area with x and y coordinates ranging from 0 to 1 was used, which represented a discretization of the study area (Figure 1).

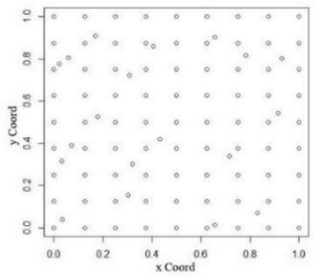

FIGURE 1. Example of a lattice plus close pairs configuration.

The values of the regionalized variables were simulated for each simulated data set by a Monte Carlo experiment, which represented stochastic process realizations $\{Z(\boldsymbol{s_i}), \boldsymbol{s_i} \in S\}$, where $Z(\boldsymbol{s_1}), \dots, Z(\boldsymbol{s_n})$ are observations of the georeferenced variable at $\boldsymbol{s}_i = (x_i, y_i)^T$ $(i = 1, \dots, n)$ sampled spatial locations, where $S \subset \mathcal{R}^2$ and $\mathcal{R}^2$ is the two-dimensional Euclidean space (Mardia & Marshall, 1984). The georeferenced variable was expressed by a Gaussian linear spatial model (Uribe-Opazo et al., 2012) described in matrix notation by $\boldsymbol{Z} = \mathbf{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{Z} \sim N_n(\mathbf{X\beta}, \sum)$, and the random error vector $\boldsymbol{\epsilon}$ has $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ (null vector $n \times 1$) and covariance matrix $\sum = [(\sigma_{ij})]$, $n \times n$, with elements $\sigma_{ij} = C(\boldsymbol{s}_i, \boldsymbol{s}_j)$, i,j = 1,...,n (Mardia & Mashall, 1984; Uribe-Opazo et al., 2012).

The georeferenced variable was considered non-stationary, and the vector of mean ($\boldsymbol{\mu} = \mathbf{X\beta}, n \times 1$) represented a directional trend of the georeferenced variable expressed by the model $\mu = \beta_0 + \beta_1 y$, where $\boldsymbol{\beta} = (\beta_o, \beta_1)^T$ is a vector of unknown parameters, such that β_o and β_1 need to be estimated and $\mathbf{X}$ is the full-rank delineation matrix (Cressie, 2015).

Furthermore, it was assumed that $\sum$ is the non-singular covariance matrix, such that $\sum = \varphi_1 \mathbf{I_n} + \varphi_2 \mathbf{R}(\varphi_3)$, where φ_1 is the nugget effect, $\boldsymbol{I_n}$ is the identity matrix $n \times n$, φ_2 is the contribution, φ_3 is the range function of the model, where the practical range $(a = g(\varphi_3))$ is the radius of spatial dependence, and $\mathbf{R}(\varphi_3)$ is a matrix $n \times n$, which is a function of φ_3 (Uribe-Opazo et al., 2012; De Bastiani et al., 2015).

Simulations were carried out at each test considering $\beta_0 = 10$ and $\beta_1 = 3$ and an exponential model to define the covariance with the parameter contribution (φ_2) equal to 1 and all combinations of the following values for the practical range parameters (a = 0.45, 0.60, and 0.90) and nugget effect ($\varphi_1 = 0$, 0.5, and 0.8).

An iterative optimization process of configuration and sample size was applied for each simulation of each test. This optimization process consists of two nested phases: the "external" and "internal" phases. A sampling plan with an established sample size was performed in the external process.

The internal phase, based on the methodology of the genetic algorithm, was applied in this sample size. This algorithm always seeks to obtain modifications in the optimization process, i.e., changes in the individuals of the population, always seeking an improvement in the objective function (Equation 1) (Zhu & Stein, 2005; Cressie, 2015). The flowchart shown in Figure 2 exemplifies the optimization process. The configuration and optimized sample size were obtained at the end of this process.

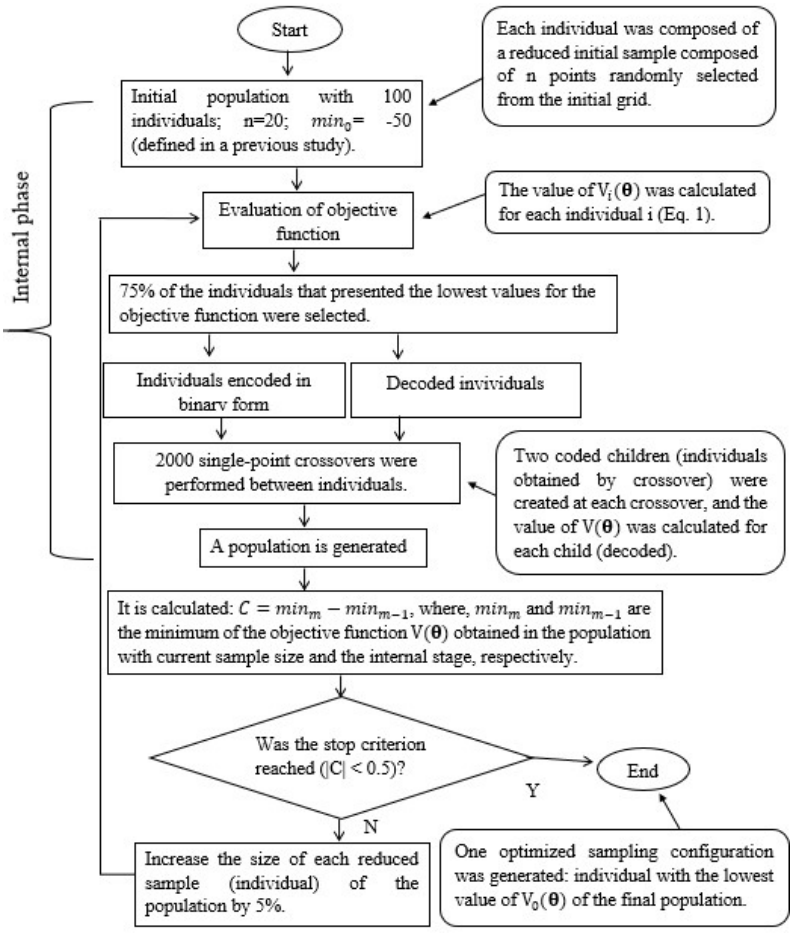

FIGURE 2. Flowchart of the optimization process of configuration and sample size.

$$V_0(\boldsymbol{\theta}) = -\log|\mathbf{I}_F(\boldsymbol{\theta})| = \log|(\mathbf{I}_\mathbf{F}(\boldsymbol{\theta}))^{-1}| \quad (1)$$

Where,

$\mathbf{I}_F(\boldsymbol{\theta})$ is the Fisher information matrix with a dimension that depends on the number of parameters of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\varphi}^T)^T$, with $\boldsymbol{\beta}^T = (\beta_0, \beta_1)$ and

$\boldsymbol{\varphi}^T = (\varphi_1, \varphi_2, \varphi_{3,})$ in such a way that the parameters of the vector $\boldsymbol{\theta}$ were estimated by the maximum likelihood method. More details on this matrix are found in Uribe-Opazo et al. (2012).

Practical study

Soil chemical properties were observed in the 2010/2011 cropping season in a commercial area of soybean production with 167.35 ha located at Fazenda Agassiz in Cascavel, PR, with minimum and maximum limits for the geographical coordinates of 24°57′30″ and 24°56′45″ South latitude and 53°35′ and 53°34′ West longitude, Datum WGS84, and an average elevation of 650 m. The soil is classified as a Dystroferric Red Latosol with clay texture. A total of 102 soil sample points of a lattice plus close pairs configuration were collected (Chipeta et al., 2017), with a minimum distance between regular grid points of 141 meters, and in some randomly selected places, sampling was performed with smaller distances (75 and 50 meters between pairs of points) (Figure 3). The samples were located and georeferenced by a Global Positioning System (GPS) signal receiver in a Datum coordinate system WGS84, UTM (Universal Transverse Mercator) projection.

Soil samples were taken at each demarcated point (Figure 3). Four soil subsamples were collected near these points at a depth of 0.0 to 0.2 m, mixed and stored in plastic bags, with samples of approximately 500 g, thus composing the sample representative of the plot. Chemical analyses were performed using the Walkley-Black method (Walkley & Black, 1934).

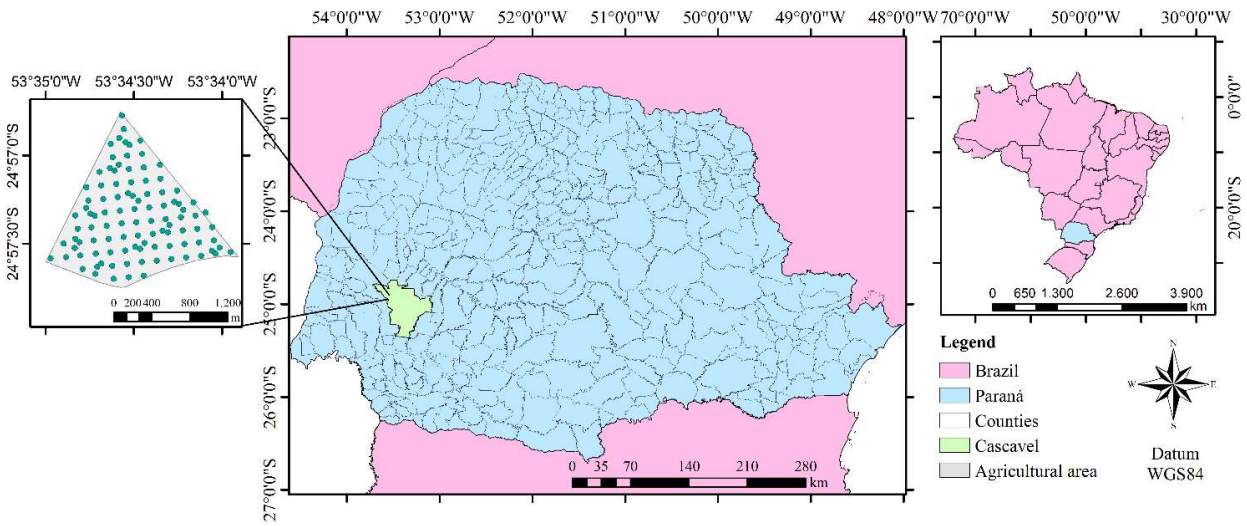

FIGURE 3. Map with the location of the study area and sampling configuration.

The following soil chemical attributes were determined in the chemical analysis: aluminum (Al, $cmol_c$ dm^{-3}), calcium (Ca, $cmol_c$ dm^{-3}), carbon (C, g dm^{-3}), copper (Cu, $cmol_c$ dm^{-3}), iron (Fe, mg dm^{-3}), phosphorus (P, g dm^{-3}), H+Al ($cmol_c$ dm^{-3}), magnesium (Mg, $cmol_c$ dm^{-3}), manganese (Mn, $cmol_c$ dm^{-3}), potassium (K, $cmol_c$ dm^{-3}), zinc (Zn, mg dm^{-3}), and pH. Among them, only the chemical attributes that had spatial dependence were selected: Ca, C, Cu, Mn, and pH.

Descriptive and geostatistical analyses were performed for each soil chemical attribute. The existence of anisotropy using the non-parametric test of Maity & Sherman (2012) (MS) was evaluated at 5% significance level. The following models of the semivariance function were estimated by the maximum likelihood method (Uribe-Opazo et al., 2012; Cressie 2015): exponential, Gaussian, and Matérn family with shape parameters k = 1, 1.5, and 2. The choice of the model was performed by the cross-validation technique (leave one out) (Faraco et al., 2008; Cressie, 2015). Subsequently, the spatial prediction by kriging of each soil chemical attribute was carried out in a grid of non-sampled locations in the agricultural area under study (Figure 3). The thematic map of each attributed was constructed considering this spatial prediction.

Subsequently, the GA was applied to each soil chemical attribute taking into account the same phases and criteria applied in the simulations (Figure 2). A small sample size configuration was obtained for each soil chemical attribute at the end of the optimization process, and exploratory and geostatistical analyses were carried out again.

Furthermore, the initial and optimized sample configurations were compared. The purpose of this comparison was to identify which one provided a better estimation of the variable in non-sampled locations. For this, the following measures were used: the mean of the kriging variance, overall accuracy (OA), and Kappa (Kp) and Tau (T) concordance indices. The studies of Guedes et al. (2014) and Landis & Koch (1977) are recommended for further details of the indices.

Simulations, GA implementation, and statistical and geostatistical analyses were performed in the software R (R Development Core Team, 2018) using the packages geoR (Ribeiro Jr. & Diggle, 2001) and sm (Maity & Sherman, 2012).

RESULTS AND DISCUSSION

Study of simulations

The estimated values for the logarithm function of the determinant of the inverse-Fisher information matrix ($V_0(\theta)$), obtained at the end of the optimization process, are very close and have a low dispersion of these minimum values in all simulations (Table 1). A relevant mean decrease in the estimated value of $V_0(\theta)$ (ranging from 68 to 108%) was observed when the estimated values of $V_0(\theta)$ were compared at the beginning and end of the optimization process, indicating an efficiency in the minimization of $V_0(\theta)$ (Table 1). In addition, all the simulations presented a low variability of the estimated values of $V_0(\theta)$, which means that the optimization process determined a reduced size sample configuration with a higher minimization of $V_0(\theta)$, thus showing the efficiency of the process.

TABLE 1 Mean values according to the simulated practical range and the simulated nugget effect of the estimated value of the logarithm of the determinant of the inverse-Fisher matrix information ($V_0(\theta)$) of the optimized sample, the percentage of decrease of ($V_0(\theta)$) (Δ (%)) in relation to the beginning and end of the optimization process, and the reduced sample size (N). In parentheses is the standard deviation of these values.

Nugget effect	Statistics	Practical range		
		$a = 0.45$	$a = 0.60$	$a = 0.90$
$\varphi_1 = 0$	$V_0(\theta)$	−22.30 (2.51)	−25.31 (2.06)	−23.34 (2.58)
	Δ (%)	73% (20.52)	68% (17.40)	74% (18.10)
	N	36.32 (11.15)	37.40 (10.46)	34.90 (8.50)
$\varphi_1 = 0.5$	$V_0(\theta)$	−19.79 (2.30)	−20.14 (2.28)	−20.34 (2.57)
	Δ (%)	83% (30.40)	89% (36.24)	96% (32.48)
	N	39.11 (11.52)	36.04 (10.15)	36.65 (10.27)
$\varphi_1 = 0.8$	$V_0(\theta)$	−19.60 (2.22)	−19.31 (2.47)	−19.93 (2.53)
	Δ (%)	97% (28.05)	97% (28.12)	108% (34.04)
	N	36.85 (11.20)	39.15 (11.34)	35.80 (11.02)

a is the simulated practical range (km); φ_1 is the simulated nugget effect; and $\Delta\ (\%) = \frac{\text{Min. final}\,(V_0(\theta)) - \text{Min. initial}(V_0(\theta))}{\text{Min. initial}\,(V_0(\theta))} * 100$.

The simulation study showed no relationship between the estimated values of $V_0(\theta)$ (obtained at the end of the optimization process) and values of the nugget effect and practical range (Table 1). However, according to Landim (2006), better estimates of parameters of the geostatistical model are obtained when these models are based on semivariograms that show the lowest ratio between the nugget effect and sill and highest practical range.

In most cases, when the nugget effect or the practical range varied, no relevant change was observed in the reduced sample size, and its lowest dispersion was obtained for the simulation with the lowest nugget effect and highest practical range ($\varphi_1 = 0$ and $a = 0.90$). Considering all the simulations, the best sample configurations obtained by the optimization process had, on average, 35 to 39 points, thus reducing the number of sampling points by 62 to 66% in relation to the initial grid (Table 1).

On average, the smallest sample size was obtained with the lowest value of the nugget effect and the highest value of practical range ($\varphi_1 = 0$ and $a = 0.90$), while the largest sample size was obtained with the highest value of nugget effect and the second largest value of practical range ($\varphi_1 = 0.80$; $a = 0.60$). No pattern was identified for the arrangement of chosen points in all simulated cases when comparing the layout of points of the optimized sample grid (for an example of each simulation – Figure 4).

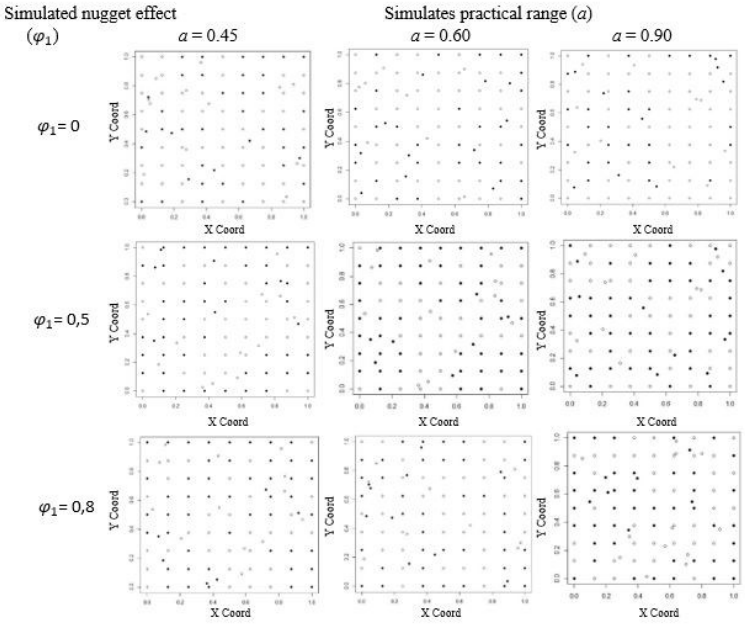

FIGURE 4. Location of the 100 initial points arranged in a lattice plus close pairs configuration (○) and selected points (●) for an example of each simulation.

Practical study

The commercial area was initially composed of 102 sampling points. A minimum sample size ranging from 30 to 40 points was obtained after applying GA for each soil chemical attribute (Table 2), being the highest and lowest number of points in the reduced sample found for Mn and C content, respectively. This reduced sample size corresponded, respectively, to 39.22 and 29.41% of the number of points in the initial sample grid, i.e., a reduction of 60.78 to 70.59% in the initial grid and, consequently, in the cost with laboratory analysis of future studies.

These results were similar to those obtained in simulations and lower than the sample size optimized by simulated annealing proposed by Guedes et al. (2014) or the fixed sample size (50% of the initial grid) in the optimization of a sample configuration proposed by Guedes et al. (2011) using a hybrid genetic algorithm and considering the efficiency of spatial prediction. In addition, these results corroborate the findings of Dias et al. (2018), who evaluated the effect of sample densities and observed that a reduction in an interval of 60 to 80% of the sample grid allowed the identification of spatial variability.

There is no consensus in the literature regarding the number of samples to be collected per hectare. The results of the present study show a variation of one sample for every 4 to 6 ha, which is in line with the amplitude of the sample density found in the literature (Cherubin et al., 2014; Siqueira et al., 2014; Zonta et al., 2014) (Table 2).

Table 2 shows that even with sample reduction, the descriptive statistics of soil chemical attributes obtained for the sampling configuration optimized by GA were similar to the results of the initial sampling configuration.

All soil chemical attributes showed a decrease in the value of the coefficient of variation (CV) when comparing the initial and optimized sampling configurations (Table 2). According to Schmidt et al. (2002), attributes with a high dispersion are theoretically better to evidence some locations than attributes with lower dispersion.

TABLE 2. Descriptive statistics and Pearson's linear correlation coefficient of the soil chemical attributes Ca ($cmol_c\ dm^{-3}$), C ($g\ dm^{-3}$), Cu ($mg\ dm^{-3}$), Mn ($cmol_c\ dm^{-3}$), and pH, considering the original and small-sized sampling configurations.

Statistics	Sampling configuration	Ca	C	Cu	Mn	pH
Mean	Original n = 102	5.20	26.93	2.95	49.32	5.10
Minimum		2.37	19.87	1.10	17.00	4.40
Median		5.08	26.88	2.80	43.00	5.10
Maximum		11.76	34.29	4.90	107.00	6.70
CV (%)		26.41	12.06	27.86	39.50	7.58
Coef. X (r)		0.14	−0.08	−0.08	−0.08	0.14
Coef. Y (r)		0.40	0.36	−0.56	0.59	0.35
n^*	Optimized with reduced size	35	30	35	40	35
No. of samples/No. hectares		1/5	1/6	1/5	1/4	1/5
Mean		5.04	27.11	2.92	45.75	5.12
Minimum		3.61	24.55	1.80	26.00	4.60
Median		4.92	26.88	2.70	42.50	5.20
Maximum		6.87	30.39	4.30	75.00	5.60
CV (%)		16.14	6.20	24.05	25.79	5.44
Coef. X (r)		−0.04	0.20	0.06	−0.25	0.26
Coef. Y (r)		0.82	0.85	−0.80	0.80	0.63

CV is the coefficient of variation; Coef. X (r), Coef. Y (r): Person's linear correlation coefficient between soil chemical attributes and X and Y coordinates; and n and n^* are the number of points of the original and optimized grid, respectively.

The presence of trend in the north direction (Y coordinate) for each soil chemical attribute was intensified when the optimized sampling configuration was used, which is due to an increase in the values of the Pearson's linear correlation coefficient (r) of each soil chemical attribute with the coordinates Y (coef. Y (r)). All soil chemical attributes showed no trend in the south direction (X coordinate) for both sampling configurations (initial and optimized) (Table 2).

The null hypothesis that the spatial dependence structure is isotropic was not rejected (p-value> 0.05) in the non-parametric MS test of isotropy applied for each soil chemical attribute. In addition, for both sampling configurations, the best model of the semivariance function was the Matérn family model with k = 2 for Ca and C contents and the exponential model for Cu and Mn contents and pH.

All soil chemical properties in all sampling configurations presented spatial dependence when the estimated value of the relative nugget effect (RNE) was evaluated (Table 3) (Cambardella et al., 1994). The ratio between the nugget effect and sill (RNE), which characterizes the spatial dependence, decreased for most soil chemical attributes with a reduction in the number of points, a result that has also been found in the literature when considering different intensities of regular soil sampling (Souza et al., 2014). Thus, the lower the ratio between the nugget effect and sill is, the lower the variance of the estimate and hence the higher the confidence in the estimation (Landim, 2006).

An increase in the spatial dependence radius ($\hat{a}$) and a reduction in the estimated value of the nugget effect were observed for all soil chemical attributes when using the optimized sampling configuration in the estimation of the geostatistical model, which evidenced a difference in the estimated values of parameters of the geostatistical model. This reduction is related to a small reduction of randomness as the number of samples decreased, i.e., a sampling grid with a higher number of samples is associated with a higher variability in the measured values and also with a higher presence of sampling or measurement noise (Porto et al., 2011; Souza et al., 2014).

TABLE 3. Estimated values of the parameters of the adjusted geostatistical model and objective function obtained by GA for the soil chemical attributes Ca ($cmol_c\ dm^{-3}$), C ($g\ dm^{-3}$), Cu ($mg\ dm^{-3}$), Mn ($cmol_c\ dm^{-3}$), and pH, considering the original and optimized sampling configurations.

Attribute	Configuration	$\mu = \beta_0 + \beta_1 y$	$\sum = \varphi_1 \mathbf{I_n} + \varphi_2 \mathbf{R}(\varphi_3)$	$\hat{a}$ (km)	$\widehat{EPR}$ (%)	$\widehat{V_0}(\theta)$
Ca		$\mu = -9.9 \cdot 10^3 + 1.37y$	$\sum = 1.14\mathbf{I_n} + 0.43\mathbf{R}(0.05)$	0.25	72.43	−17.48
C		$\mu = -2.1 \cdot 10^4 + 2.95y$	$\sum = 4.50\mathbf{I_n} + 4.58\mathbf{R}(0.07)$	0.38	49.53	−6.89
Cu	Original	$\mu = 6.3 \cdot 10^3 - 0.87y$	$\sum = 0.09\mathbf{I_n} + 0.36\mathbf{R}(0.27)$	0.81	20.71	−20.04
Mn		$\mu = -1.7 \cdot 10^5 + 23.74y$	$\sum = 95.25\mathbf{I_n} + 146.10\mathbf{R}(0.15)$	0.45	39.47	16.13
pH		$\mu = -2.4 \cdot 10^3 + 0.34y$	$\sum = 0.10\mathbf{I_n} + 0.03\mathbf{R}(0.10)$	0.30	73.75	−29.37
Ca		$\mu = -9.1 \cdot 10^3 + 1.26y$	$\sum = 0.0006\mathrm{I_n} + 0.21\mathbf{R}(0.05)$	0.28	0.30	−28.51
C		$\mu = -2.0 \cdot 10^4 + 2.73y$	$\sum = 0.51\mathbf{I_n} + 0.26\mathbf{R}(0.14)$	0.75	66.23	−20.15
Cu	Optimized	$\mu = 7.8 \cdot 10^3 - 1.07y$	$\sum = 0.02\mathbf{I_n} + 0.17\mathbf{R}(0.46)$	1.37	10.53	−20.60
Mn		$\mu = -1.6 \cdot 10^5 + 21.60y$	$\sum = 33.84\mathbf{I_n} + 18.36\mathbf{R}(0.42)$	1.26	64.83	13.05
pH		$\mu = -2.9 \cdot 10^3 + 0.41y$	$\sum = 0.04\mathbf{I_n} + 0.01\mathbf{R}(0.32)$	0.96	81.09	−30.64

μ : mean, where, $\hat{\beta}_0, \hat{\beta}_1$: are the estimated values of parameters of the regression model, y: coordinate Y, $\sum$: matrix of covariance, where φ_1 is the nugget effect; $\boldsymbol{I_n}$ is the identity matrix $n \times n$; φ_2 is the contribution; φ_3 is the function of the range of the model; $\mathbf{R}(\varphi_3)$ is the matrix $n \times n$ which is a function of φ_3, $\hat{a}$: estimated practical range, RNE: estimated value of the relative nugget effect ($\widehat{RNE} = \hat{\varphi}_1/\hat{\varphi}_1 + \hat{\varphi}_2$) (%), and $\widehat{V_0}(\theta)$: objective function.

Soil chemical attributes showed an estimated value of the spatial dependence radius ($\hat{a}$) ranging from 250 to 880 m when considering the initial sampling configuration and values from 280 to 1370 m when considering the reduced sampling configuration (Table 3). The increase in the range value produces a thematic map with more continuous structures, without the formation of small subregions, which facilitates the agricultural management. However, the thematic maps become less attenuated as the value of the nugget effect decreases, with a higher influence of neighboring samples to points to be estimated, which leads a higher precision of the neighborhood (Cressie, 2015).

Considering the optimized sampling configuration, all soil chemical attributes showed a reduction in the estimated values of $V_0(\theta)$ (from 4 to 192%) when compared with the estimated values of $V_0(\theta)$, obtained with the original sample. The lowest reduction was obtained for pH, which presented the highest value for RNE, indicating it is close to the threshold defined as weak spatial dependence for original and optimized sampling configurations (RNE > 75%, Cambardella et al., 1994). Moreover, the highest reduction was obtained in the attribute that presented the smallest sample size in the reduced sampling configuration (Table 3).

The estimated values of the standard deviation of the parameters indicated a decrease for most of soil chemical attributes associated with parameters of the regression model, which explains the mean, nugget effect, practical range, and contribution when comparing the reduced and original sampling configurations. It shows that model estimation in the optimized configuration was more efficient than in the original configuration (Table 4) (Pigoto & Barreto, 2004).

TABLE 4. Values of estimated standard deviations of the model parameters adjusted for the soil chemical attributes Ca ($cmol_c\ dm^{-3}$), C ($g\ dm^{-3}$), Cu ($mg\ dm^{-3}$), Mn ($cmol_c\ dm^{-3}$), and pH considering the original and optimized configurations.

Attribute	Configuration	$D(\hat{\beta}_0)$	$D(\hat{\beta}_1)$	$D(\hat{\varphi}_1)$	$D(\hat{\varphi}_2)$	$D(\hat{a})$
Ca		2.9×10^{-9}	2.1×10^{-5}	0.11	0.11	0.02
C		1.1×10^{8}	7.7×10^{-5}	2.7×10^{-4}	3.9×10^{-5}	0.02
Cu	Original	4.1×10^{-9}	2.9×10^{-5}	0.06	0.13	0.15
Mn		5.9×10^{-8}	4.0×10^{-4}	8.8×10^{-5}	2.8×10^{-5}	0.06
pH		8.7×10^{-10}	6.3×10^{-6}	0.03	0.03	0.02
Ca		1.9×10^{-9}	3.0×10^{-4}	0.04	0.07	0.02
C		4.7×10^{-9}	3.0×10^{-5}	0.11	0.02	0.02
Cu	Optimized	3.8×10^{-9}	2.8×10^{-5}	0.03	0.09	0.01
Mn		4.4×10^{-8}	3.0×10^{-4}	0.01	1.0×10^{-3}	0.06
pH		8.7×10^{-10}	7.0×10^{-6}	0.01	0.01	1.0×10^{-3}

$\hat{\beta}_0, \hat{\beta}_1$: are the estimated values of the parameters of the regression model, which explain the mean, where $\mu = \beta_0 + \beta_1 y$; $\hat{\varphi}_1$: estimated nugget effect, $\hat{\varphi}_2$: contribution, â: estimated practical range (km), D(•): standard deviation.

The maps of all soil chemical attributes constructed using the optimized and initial sampling configuration did not present visual similarities, which was confirmed by measurements of overall accuracy and Kappa and Tau concordance indices (OA<0.85; K<0.67; T<0.67) (Figure 5) (Landis & Koch, 1977, Guedes et al., 2014). In relation to spatial prediction, this dissimilarity can be considered a disadvantage for the optimization process, which considered only one criterion associated with the efficiency of the estimation quality of the geostatistical model.

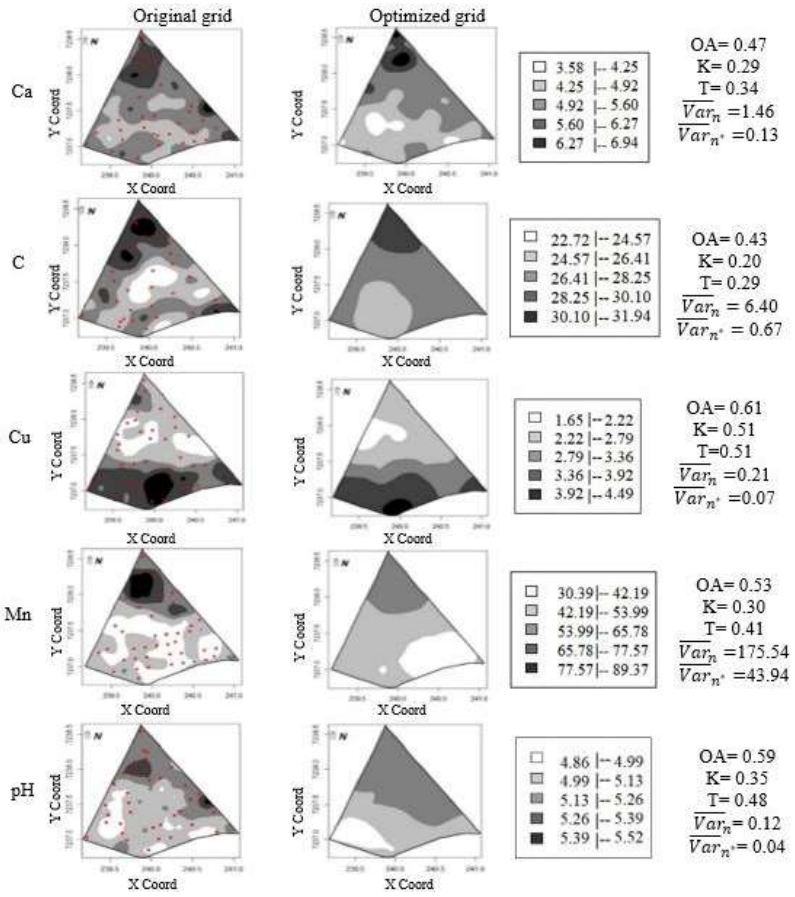

FIGURE 5. Thematic maps of soil chemical attributes constructed with the original and optimized sampling grid. ● represents each selected location in the original grid to compose the optimized grid. Estimated values of OA (overall accuracy), K, and T (Tau and Kappa concordance indices), mean of kriging variance of the original configuration ($\overline{\mathbf{Var}_\mathbf{n}}$) and of the optimized ($\overline{\mathbf{Var}_{\mathbf{n}^*}}$).

However, criteria associated with the quality of the geostatistical model estimation or spatial prediction are not necessarily concomitant (Muller, 2007). Also, these measurements do not identify which sampling configuration is the best in the spatial prediction, but only indicate the similarity present between maps.

All soil chemical attributes presented a reduction in the mean value of the kriging variance when the reduced sampling configuration was used in the spatial prediction. Thus, kriging produced better estimates of the georeferenced variable in non-sampled locations when the optimized sampling configuration was used (Figure 5).

For all soil chemical attributes, the visual analysis of the layout of selected locations (red dots in Figure 5) shows that the optimization process sought to select points of heterogeneous sub-regions or points close to each sub-area described by the thematic map of the original grid. A higher scattering of selected points was observed mainly for C, Cu, and pH, which presented intermediate practical range values when compared to other attributes (Figure 5).

Therefore, the algorithm sought a total area coverage, tending not to select contiguous samples, which produces better results regarding the analysis of spatial variability (Fattorini et al., 2015). Thus, in general, a scattering of the chosen sample points was observed throughout the study area.

CONCLUSIONS

The optimization process was efficient for the simulated and real data and resized the sample grid that involves the experiment, reducing its sample size and improving the estimates of the Gaussian spatial linear model.

The new optimized sampling configuration varied from 30 to 40 points for all soil chemical attributes, which corresponds respectively to 29.41 to 39.22% of the original grid. Thus, one sample at every 4 or 6 hectares would be required for the composition of the sampling configuration. These conclusions were obtained from an optimization process that considers previously known information, such as an initial sampling configuration and spatial dependence structure of the already estimated attributes. Thus, the implementation of an initial sampling configuration composed of 30 to 40 points and efficient in obtaining the results of the spatial variability analysis would be difficult.

Regarding the estimation of the spatial dependence structure, all soil chemical attributes in both sampling configurations presented moderate or strong spatial dependence when the relative nugget effect and practical range were simultaneously evaluated. Relevant differences were observed for all soil chemical properties between thematic maps constructed considering the configuration of sampling points of the original grid and reduced size. However, the values of the mean of kriging variance and deviations of model estimates showed that the optimized sampling configuration produced a better quality in describing the spatial dependence structure.

Regarding the simulated data, the variation in the nugget effect or practical range did not provide any relevant change in the reduced sample size in most cases.

ACKNOWLEDGMENTS

The authors are grateful for the partial financial support from the Foundation Araucária of Paraná State-Brazil, Coordination for the Improvement of Higher Education Personnel, Brazil (CAPES), and National Council for Scientific and Technological Development (CNPq).

REFERENCES

Artur AG, Oliveira DP, Costa MCG, Romero RE, Silva MVC, Ferreira TO (2014) Variabilidade espacial dos atributos químicos do solo, associada ao microrrelevo. Revista Brasileira de Engenharia Agrícola e Ambiental 18(2):141-149. DOI: http://dx.doi.org/10.1590/S1415-43662014000200003

Bernardi ACC, Rabello LM, Inamasu RY, Grego CR, Andrade RG (2014) Variabilidade espacial de parâmetros físico-químicos do solo e biofísicos de superfície em cultivo de sorgo. Revista Brasileira de Engenharia Agrícola e Ambiental 8 (6): 623-630. DOI: http://dx.doi.org/10.1590/S1415-43662014000600009

Cambardella CA, Moorman TB, ParkiN TB, Novack JM, Karlen DL, Turco RF, Knopka AE (1994) Field-scale variability of soil properties in Central Iowa Soils. Soil Science Society America Journal, Medison 58(4):1501-1511. DOI: http://dx.doi.org/10.2136/sssaj1994.03615995005800050033x

Cheng L, Liu J, To AC (2018) Concurrent lattice infill with feature evolution optimization for additive manufactured heat conduction design. Structural and Multidisciplinary Optimization 58(2):511-535. DOI: https://doi.org/10.1007/s00158-018-1905-7

Cherubin MR, Santi AL, Eitelwein MT, Menegol DR, Ros COD, Pias OHC, Bergjetti J (2014) Eficiência de malhas amostrais utilizadas na caracterização da variabilidade espacial de fósforo e potássio. Ciência Rural 44(3):425-432. DOI: http://dx.doi.org/10.1590/S0103-84782014000300007

Cherubin MR, Santi AL, Eitelwein MT, Amado TJC, Simon DH, Damian JM (2015) Dimensão da malha amostral para caracterização da variabilidade espacial de fósforo e potássio em Latossolo Vermelho. Pesquisa Agropecuária Brasileira 50(2):168-177. DOI: 10.1590/S0100 204X2015000200009

Chipeta MG, Terlouw DJ, Phiri KS, Diggle PJ (2017) Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. Environmetrics 28(1):e2425. DOI: https://doi.org/10.1002/env.2425

Cressie NAC (2015) Statistics for spatial data. John Wiley & Sons, 928 p.

De Bastiani F, Cysneiros AHMA, Uribe-Opazo MA, Galea M (2015) Influence diagnostics in elliptical spatial linear models. Sociedad de Estadística e Investigación Operativa, TEST 24 (2): 322-340. DOI: https://doi.org/10.1007/s11749-014-0409-z

Dias FPM, Paes EC, Nunes FJ, Nonato ACR, Silva ND, Oliveira FOP, Fereira LG, Nóbrega JCA (2018) Amostral Optimization of Mechanical Resistance to the Penetration of a Yellow Oxisol Under Pasture. Journal of Agricultural Science 10(9):275-283. DOI: https://doi.org/10.5539/jas.v10n9p275

Faraco MA, Uribe-Opazo MA, Silva EA, Johann JÁ, Borssoi JA (2008) Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. Revista Brasileira de Ciência do Solo 32(2):463-476.

Fattorini L, Corona P, Chirici G, Pagliarella MC (2015) Design-based strategies for sampling spatial units from regular grids with applications to forest surveys, land use, and land cover estimation. John Wiley & Sons, Ltd. 26(3):216-228. DOI: http://dx.doi.org/10.1002/env.2332

Guedes LPC, Ribeiro Jr PJ, Piedade SMDS, Uribe-Opazo MA (2011) Optimization of spatial sample configurations using hybrid genetic algorithm and simulated annealing. Chilean Journal of Statistics 2(2):39-50.

Guedes LPC, Uribe-Opazo MA, Ribeiro Jr PJ (2014) Optimization of sample design size and shapes for regionalized variables using simulated annealing. Ciência e Investigación Agraria 41(1): 33-48. DOI: http://dx.doi.org/10.4067/S0718-16202014000100004

Guedes LPC, Ribeiro Jr PJ, Uribe-Opazo MA, De Bastiani F (2016) Soybean yield maps using regular and optimized sample with different configurations by simulated annealing. Engenharia Agrícola 36(1):114-125. DOI: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v36n1p114-125/2016

Landim PMB (2006) Sobre Geoestatística e mapas. Terra e Didatica 2 (1): 19-33.

Landis JR, Koch GG (1977) The Measurement of observer agreement for categorical data. Biometrics 33(1):159-174.

Maity A, Sherman M (2012) Testing for spatial isotropy under general designs. Journal of Statistical Planning and Inference 142(5):1081-1091. DOI: http://dx.doi.org/10.1016/j.jspi.2011.11.013

Mardia KV, Marshall RJ (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71(1):135-146. DOI: http://dx.doi.org/10.1093/biomet/71.1.135

Muller WG (2007) Collecting spatial data. Berlin, Springer-Verlag, 3 ed.

Pessoa ALS, Ulisses PHC, Branco HMGC, Rabêlo RAL (2015) Uma aplicação de algoritmos genéticos simples e compacto para estimação de componentes harmônicas. Revista Brasileira de Computação Aplicada 7(2):77-91. DOI: http://dx.doi.org/10.5335/rbca.2015.4624

Pigoto FJr, Barreto MCM (2004) Desempenho de estimadores da média populacional de distribuições assimétricas baseados em amostragem por conjuntos ordenados. Revista de Matemática e Estatística 21(2):19-29.

Porto AL, Soares JA, Monteiro VED (2011) Otimização da malha de amostragem de compostos orgânicos voláteis no solo através de krigagem. Águas Subterrâneas 25(1):57-73.

R Development Core Team (2018) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available: http://www.R-project.org.

Ribeiro Jr PJ, Diggle PJ (2001) geoR: A package for geostatisticalanalysis. R-NEWS 1(2):15-18.

Santos WM, Souza RMS, Souza ES, Almeida AQ, Antonino ACD (2017) Variabilidade espacial da sazonalidade da chuva no semiárido brasileiro. Journal of Environmental Analysis and Progress 2(4):368-376. DOI: http://dx.doi.org/10.24221/jeap.2.4.2017.1466.368-376

Siqueira DS, Marques Jr J, Pereira GT, Barbosa RS, Teixeira DB, Peluco RG (2014) Sampling density and proportion for the characterization of the variability of Oxisol attributes on different materials. Geoderma 232(234):172-182. DOI: https://doi.org/10.1016/j.geoderma.2014.04.037

Schmidt JP, Taylor RK, Milliken GA (2002) Evaluating the potential for site-specific phosphorus applications without high-density soil sampling. Soil Science Society of America 66(1):276-283.

Souza ZM, Souza GS, Marques Jr J, Pereira GT (2014) Número de amostras na análise geoestatística e na krigagem de mapas e atributos do solo. Ciência Rural 44:261-268. DOI: http://dx.doi.org/10.1590/S0103-84782014000200011

Szatmári G, László P, Takács K, Szabó J, Bakacsi Z, Koós S, Pásztor L (2018) Optimization of second-phase sampling for multivariate soil mapping purposes: Case study from a wine region, Hungary. Geoderma 7:1-12. DOI: https://doi.org/10.1016/j.geoderma.2018.02.030

Uribe-Opazo MA, Borssoi JA, Galea M (2012) Influence diagnostics in Gaussian spatial linear models. Journal of Applied Statistics 39(3):615-630.

Walkley A, Black IA (1934) An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Science 37:29-38.

WangJF, Stein A, Gao BB, Ge Y (2012) A review of spatial sampling. Spatial Statistics 2:1-14. DOI: https://doi.org/10.1016/j.spasta.2012.08.001

Zonta JH, Brandão ZN, Medeiros JC, Sana RS, Sofiatti (2014) Variabilidade espacial da fertilidade do solo em área cultivada com algodoeiro no Cerrado do Brasil. Revista Brasileira de Engenharia Agrícola e Ambiental 18(6):595-602

Zhu Z, Stein ML (2005) Spatial sampling for desing for parameter estimation of the covariance function. Journal of Statistical Planning and Inference 134(2):583-603. DOI: https://doi.org/10.1016/j.jspi.2004.04.017