RESEARCH ARTICLE

A genetic algorithm for resizing and sampling reduction of non-stationary soil chemical attributes optimizing spatial prediction

Tamara C. Maltauro, Luciana P. C. Guedes, Miguel A. Uribe-Opazo and Letícia E. D. Canton

Western Paraná State University (UNIOESTE), 2069 Universitária Street, 85819-110 Cascavel, Paraná, Brazil.

Abstract

Aim of study: To evaluate the influence of the parameters of the geostatistical model and the initial sample configuration used in the optimization process; and to propose and evaluate the resizing of a sample configuration, reducing its sample size, for simulated data and for the study of the spatial variability of soil chemical attributes under a non-stationary with drift process from a commercial soybean cultivation area.

Area of study: Cascavel, Brazil

Material and methods: For both, the simulated data and the soil chemical attributes, the Genetic Algorithm was used for sample resizing, maximizing the overall accuracy measure.

Main results: The results obtained from the simulated data showed that the practical range did not influence in a relevant way the optimization process. Moreover, the local variations, such as variance or sampling errors (nugget effect), had a direct relationship with the reduction of the sample size, mainly for the smaller nugget effect. For the soil chemical attributes, the Genetic Algorithm was efficient in resizing the sampling configuration, since it generated sampling configurations with 30 to 35 points, corresponding to 29.41% to 34.31% of the initial configuration, respectively. In addition, comparing the optimized and initial configurations, similarities were obtained regarding spatial dependence structure and characterization of spatial variability of soil chemical attributes in the study area.

Research highlights: The optimization process showed that it is possible to reduce the sample size, allowing for lesser financial investments with data collection and laboratory analysis of soil samples in future experiments.

Additional key words: geostatistics; overall accuracy; sample size; spatial dependence; simulation

Abbreviations used: ESS (Effective Sample Size); GA (Genetic Algorithm); GPS (Global Positioning System); Kp (Kappa); OA (Overall Accuracy); SA (Simulated Annealing); T (Tau); UTM (Universal Transverse Mercator).

Introduction

Precision agriculture differs from traditional agriculture for allowing the localized application of inputs, according to the needs of the soil in each location, thus avoiding environmental risks and high costs to the producer, and also enabling an optimal development of the crop (Inamasu *et al.*, 2011).

In agriculture, the interaction between the soil chemical attributes directly affects the growth and development of crops (Dalmago *et al.*, 2014). Thereby, studying the spatial variability of soil chemical attributes and cultures, represented by continuous georeferenced variables in a cultivated area, is pointed out as the basic principle for precise management in agricultural areas, at any scale, thus strengthening the concepts of

sustainable agriculture (Nanni *et al.*, 2011; Grzegozewski *et al.*, 2013).

For the spatial variability analysis of soil chemical attributes, the sampling configuration and the sample size are factors that must be carefully defined, since the number of soil samples used for the spatial variability analysis by geostatistical techniques is directly linked to the costs for experimenting (Catapatti *et al.*, 2008; Cressie, 2015). Therefore, in an experiment involving the spatial variability analysis, it is essential to define a sampling configuration that has an adequate sample size, which allows minimizing operational costs and maximizing quality of the spatial prediction results (Di *et al.*, 1989; Siqueira *et al.*, 2014).

Thus, the sampling scheme is essential in the investigation of the spatial variability of soil properties in Soil Science studies. However, insufficient sampling intensity can be an important limiting factor for precision agriculture and has been the subject of several studies (Montanari *et al.*, 2012).

A traditional sampling methodology can determine the reduced sampling configuration. Using systematic sampling, Kestring *et al.* (2015) compared different sample sizes to verify whether the sample density between points showed interference in the estimates of the parameters of the model that defines the structure of spatial variability, as well as in the estimation of values in unsampled locations of soybean productivity. These authors concluded that maps generated with a mean reduction of 66% of the study area did not generate satisfactory values for the similarity measure Overall Accuracy (OA). However, Ribeiro *et al.* (2016) analyzed the impact from different sample designs on the results of spatial inference by Ordinary Kriging, applying the techniques of simple random sampling, systematic sampling, and stratified sampling. These authors concluded that a gradual reduction in the initial sample size ($\approx 87\%$) did not change the inference result.

Another methodology for reducing the sample size is the calculation of the effective sample size (ESS) (Griffith, 2005; Vallejos & Osorio, 2014). The ESS considers the spatial dependence structure of the variable to calculate the effect of spatial autocorrelation between the sampled points collected. In this way, the sample size is reduced by removing redundant information.

The choice of a configuration and an efficient sample size can also be an optimization problem based on three elements: the contained information of a known sampling, the choice of a search criterion or method, as well of an objective function that expresses optimization efficiency. This objective function is minimized or maximized, according to the aim of the problem (Linden, 2012). The objective functions can consider spatial prediction efficiency (Ferreyra *et al.*, 2002; Nunes *et al.*, 2006; Guedes *et al.*, 2014); other measures associated with the efficien-

cy of the prediction (Ferreyra *et al.*, 2002; Guedes *et al.*, 2011; Szatmári *et al.*, 2018) and some measures compare maps (Guedes *et al.*, 2014). However, the efficiency of the geostatistical model estimation can still be considered in the optimization process, such as the objective function based on the inverse-Fisher information matrix (Zhu & Stein, 2005; Maltauro *et al.*, 2019). In addition, there are mixed objective functions, which consider more than one parameter in the optimization process (Yang *et al.*, 2006).

There are several search methods, which can be conducted by studying all possible combinations of solutions or by sequential search methods (Santos *et al.*, 2011). However, these methods, for large samples, become exhaustive and computationally unviable. Thus, another alternative is the meta-heuristic methodologies used in artificial intelligence, such as Simulated Annealing (SA) (Guedes *et al.*, 2014) and the Genetic Algorithm (GA) (Maltauro *et al.*, 2019). Specifically, the GA is a technique inspired by the natural evolution of species and consists of an iterative process of searching for the optimal solution (Linden, 2012).

The GA shows a great advantage in solving a combinatorial optimization problem, given its characteristic that has high efficiency and suitability for the practical application, mainly for tackling problems with extremely large search spaces. This is because this algorithm does not use a completely random search method, is capable of handling many types of functions, and operates in a space of coded solutions. In addition, this method works with the concept of population with possible solutions, unlike other methods that consider only one solution, and only needs information about the objective function value for each solution of the population. Thus, a population of adapted solutions reduces the possibility of reaching a false optimum (Min & Cheng, 1999; Linden, 2012).

Sexton *et al.* (1999), Min & Cheng (1999) and Guedes *et al.* (2011) concluded that the GA obtains superior solutions to SA for optimizing neural networks; for a larger scale identical parallel machine scheduling problem for minimizing the makespan; and in a problem of choosing the spatial sampling configuration.

The stationary hypothesis is a usual assumption in Geostatistics. The stationarity assumption can refer to a constant mean, homoscedasticity of variance, and a fixed probability distribution function within the domain (Manchuck & Deutsch, 2012). Non-stationary processes in the mean have been applied in Geostatistics for many years. In fact, Universal Kriging involves an unknown, but not stationary mean. More commonly, where the local mean value varies with the location, it can be modeled with a trend (Atkinson & Lloyd, 2007), or with a drift (Manchuck & Deutsch, 2012), by the explicit and implicit trend modeling techniques. An implicit technique is to directly consider a function that expresses the mean in the geostatistical model, while the explicit technique is, for

example, the estimation of the trend by least squares and the use of residues in geostatistical modeling (Manchuck & Deutsch, 2012).

Machuca-Mory & Deutsch (2013) considered a technique that assumes local stationarity through an integrated approach to incorporate local changes in the cumulative distribution function and all necessary statistics, concluding that the resulting maps are rich in non-stationary spatial features. This technique requires higher computational effort but, if data availability allows for a reliable inference of the local distributions and statistics, higher accuracy of the estimates can be achieved.

Studies developed by Sexton *et al.* (1999), Min & Cheng (1999), and Guedes *et al.* (2011, 2014) use stationary georeferenced variables in their development. However, stationarity is not a characteristic always identified in soil chemical attributes (Szatmari *et al.*, 2018). The assumption of the non-stationary with drift process is considered, for example, in a study by Maltauro *et al.* (2019), using an objective function of the efficiency in estimating the geostatistical model. Nevertheless, there are no studies about the optimization of configuration and sample size that consider the absence of stationarity and the efficiency of spatial prediction as well.

In this way, the objective of the study was to evaluate the influence of the parameters of the geostatistical model and the initial sample configuration used in the optimization process; and to propose and evaluate the resizing of a sample configuration, reducing its sample size, for simulated data and for the study of the spatial variability of soil chemical attributes under a non-stationary with drift process from a commercial soybean cultivation area, to resize the sampling for future experiments.

Material and methods

Simulation study

Firstly, a simulation study was carried out to reproduce a set of possibilities existing in the real data to be evaluated in this study (Shiflet & Shiflet, 2014), as well as to improve the theoretical and practical knowledge about size optimization and the sampling configuration in soil chemical attributes, with a non-stationary spatial dependence structure (Cressie, 2015).

Thus, the simulated data sets were obtained through Monte Carlo simulations, using Cholesky decomposition (Cressie, 2015). These sets represent realizations of stochastic processes $\{Z(s_i), s_i \in S\}$, where $s_i = (x_i, y_i)^T$ is the vector that represents a particular location in the study area, with $S \subset R^2$, where R^2 is the two-dimensional Euclidean space (Mardia & Marshall, 1984). Supposing that the $Z(s_1),...,Z(s_n)$ data of this process are observations of the random variable under study at s_i ($i = 1,...,n$)

known spatial locations and, generated by the Gaussian linear spatial model given in matrix notation by Eq. 1 (Uribe-Opazo *et al.*, 2012):

$$Z = X\beta + \varepsilon, \qquad (1)$$

where $X\beta$ is the deterministic term and ε is the stochastic term; such that both depend on the spatial location in which $Z(s_i)$; ZN_n $(X\beta,\sum)$ is observed, the random error vector ε has E(ε)=0 (null vector, $n \times 1$) and the covariance matrix $\sum[(\sigma_{ij})]$, $n \times n$, with $\sigma_{ij} = C(s_i, s_j)$ $i, j = 1,...,n$ (Mardia & Mashall, 1984).

In this study, the georeferenced variable was considered non-stationary (due to the fact that the real data are non-stationary in the mean), and the deterministic term that represents the mean vector $\mu = X\beta$, $n \times 1$, is a directional trend of the georeferenced variable, in which a linear trend of the increase was pre-defined of the mean value in the North direction, that is, the process mean value equals to $\mu = \beta_0 + \beta_1 y$, where β_o and β_1 are unknown parameters and need to be estimated, y represents the directional trend identified, and X ($n \times (p+1)$) is a complete station design matrix, whose i-th row is given by $x_i^T = (1, x_{i1}, ..., x_{ip})$, and $x_{il} = x_l$ (si), where $i = 1,...,n; l = 1,...,p$ represents the value of the l-th covariate taken in the n-th position (Cressie, 2015).

In addition, we have that $\sum$ is the non-singular covariance matrix (Uribe-Opazo *et al.*, 2012), given by Eq. 2:

$$\sum = \varphi_1 I_n + \varphi_2 R(\varphi_3), \qquad (2)$$

where $\varphi_1 \geq 0$ is the nugget effect; I_n is the $n \times n$ identity matrix; $\varphi_2 \geq 0$ is the contribution or the dispersion variance; $\varphi_3 > 0$ is s the range function of the model, the practical range ($a = g(\varphi_3)$) is the limit distance of spatial dependence (Cressie, 2015), and $R(\varphi_3)$ is an $n \times n$ matrix which is a function of φ_3 , where $R(\varphi_3) = [(r_{ij})]$ a symmetric $n \times n$ matrix, where r_{ij} depends only on the Euclidean distance between s_i and s_j ($h_{ij} = \|s_i - s_j\|$), with diagonal elements $r_{ij} = 1$; for $i = j$, $r_{ij} = \varphi_2^{-1} C(s_i, s_j)$ for $\varphi_2 \neq 0$ and $i \neq j$; and $r_{ij} = 0$ for $\varphi_2 = 0$, $i \neq j (i, j = 1,...,n)$ (De Bastiani *et al.*, 2015).

For each sampling configuration, nine trials were considered (Fig. 1). In each trial, the simulations were performed considering the Gaussian spatial linear model (Eq. 1), and an exponential model (simple model and with good estimation; Cressie, 2015) to define the covariance, considering a combination of the parameter values of the Gaussian spatial linear model (Fig. 1), and the choice of these parameters to contemplated different intensities of spatial dependence (weak, moderate and strong; Cambardella *et al.*, 1994), as for the relative nugget effect $\left(E = \frac{\varphi_1}{\varphi_1 + \varphi_2}\right)$, and as for the practical range (a).

The methodological scheme occurred in two phases, the external and the internal phase. The external phase
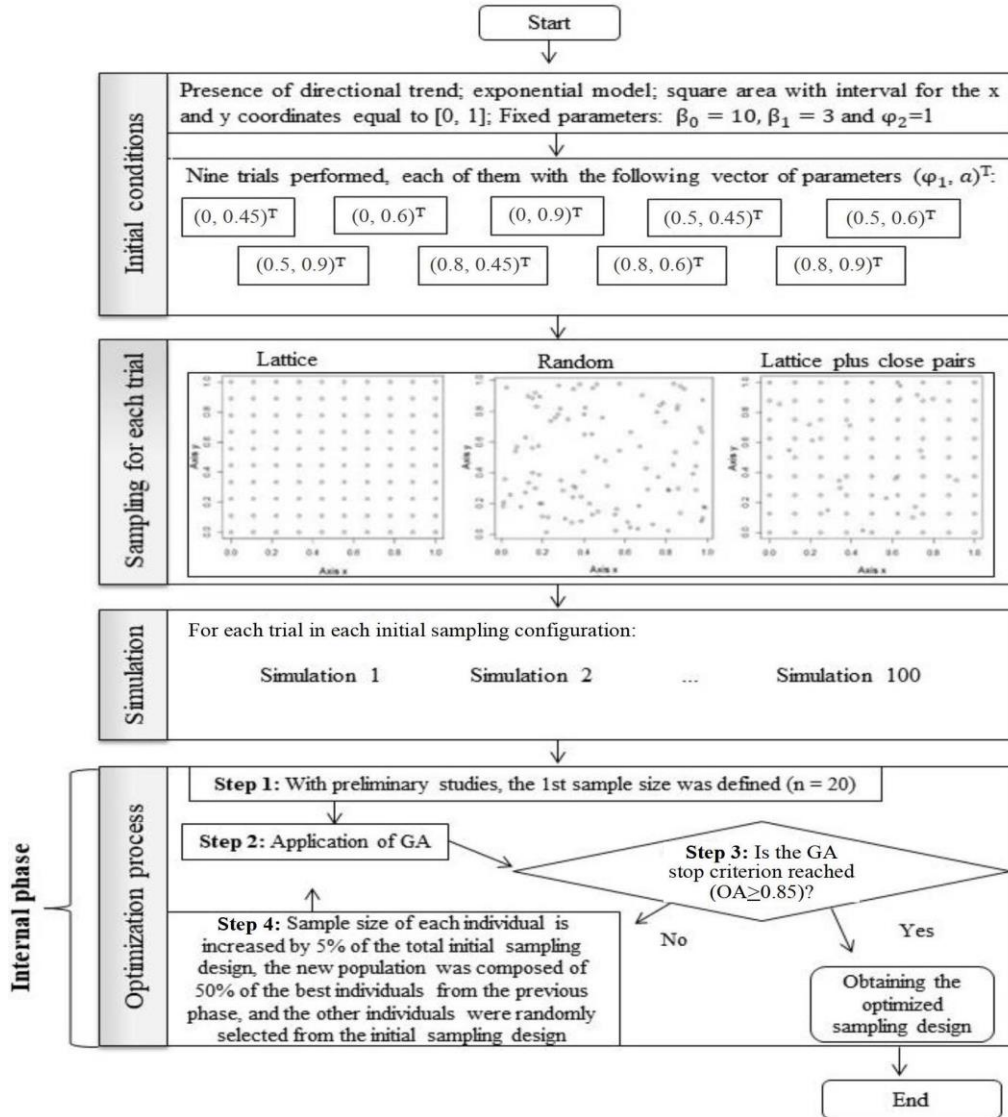
Figure 1. Methodological scheme for the choice of a configuration and a sample size.

begins with an established sample size - while the optimization of this size is also sought - and aims to carry out a sampling plan in this sample size.

In the internal phase the GA optimization process that considered the established sample size in the external phase took place. At each iteration of the GA, the algorithm seeks changes in its individuals (each individual is a sampling configuration, with a fixed size by the general optimization process), searching for an improvement in the objective function (maximizing OA; Cressie, 2015).

The optimized sampling configuration does not follow a traditional sampling pattern. However, the points chosen in the optimization process belong to the initial sampling configuration. Fig. 1 shows the entire simulation study scheme, as well as the optimization process, mainly the description of the external phase.

The objective function was the similarity measure called OA (Anderson *et al.*, 2001), calculated for each individual in the population according to Eq. 3. This measu-

re compares the predicted values obtained by the initials and optimizes configurations, through the error matrix, whose unit of measure is the pixel. Each element of this matrix represents the total number of pixels or the total area belonging to class k of the model map and class l of the reference map. In this case, in the geostatistical analysis, the elaboration of the model map and the reference map considered the initial sampling configuration and the optimized sampling configuration, respectively.

$$OA = \frac{\sum_{k=1}^{t} n_{kk}}{N}, \qquad (3)$$

where n_{kk} are the main diagonal elements of the error matrix, which represent the number of pixels that had the same classification for the two maps compared; t is the number of classes, and N is the total pixels or the total area.

We applied the GA successively until the sample optimized by the algorithm reached the minimum value

of OA equal to 0.85 because OA values ≥ 0.85 indicate similarity between the maps that were prepared considering the two sampling configurations (Anderson *et al.*, 2001). The flowchart shown in Fig. 2 exemplifies the GA optimization process. And, at the end of this optimization process, the optimized sampling configuration of reduced size was obtained.

Practical study

The data set used in this research refers to a commercial area of soy production with 167.35 ha, in which precision agriculture is developed, with the localized application of inputs. The area consists of 102 sample points, located at Fazenda Agassiz in Cascavel (PR, Brazil), which has approximated geographic coordinates of 24.95° S and 53.57° W, and an average elevation of 650 m (Fig. 3). The soil is dystrophic Red Latosol, with a clay texture. The climate in the region is mesothermal and super-humid temperate, climatic type Cfa (Koeppen) and the average annual temperature is 21º C (Embrapa, 2013). The data belongs to the database of the Laboratory of Spatial Statistics and the Laboratory of Applied Statistics of the Western Paraná State University (UNIOESTE), Cascavel/Brazil.

The defined sampling configuration in this area was lattice plus close pairs ($k \times k$, m, δ), being that it consists of a regular $k \times k$ grid with Δ spacing adding other m points, each of which are located uniformly at random within a circumference of radius δ, whose center is always at a randomly selected grid point (Chipeta *et al.*, 2017). These designs contained a regular grid, with a minimum distance between regular grid points of 141 m, and in some randomly selected places of this regular grid, we added 19 sample points, with smaller distances with the regular grid (75 m and 50 m between pairs of points) (Fig. 3). The samples were located and georeferenced by a Global Positioning System (GPS) signal receiver in a WGS84Datum coordinate system, UTM (Universal Transverse Mercator) projection.

In addition to the organic macronutrients provided by the atmosphere, soy needs nutrients provided by the soil for optimal crop development. The following soil chemical attributes, observed in the 2010/2011 cropping season and that showed spatial dependence and non-stationary were studied: calcium (Ca, cmolc dm^{-3}), carbon (C, g dm^{-3}), copper (Cu, cmolc dm^{-3}), manganese content (Mn, cmolc dm^{-3}), and pH. The analysis of these chemical attributes of the soil is important because imbalance of
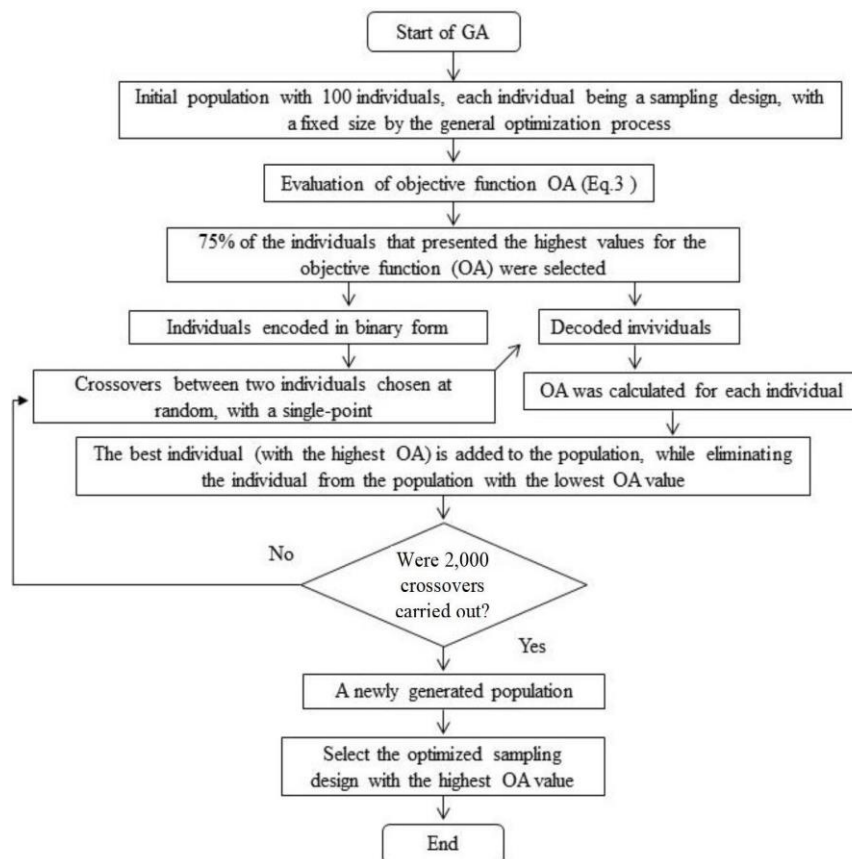

Figure 2. Methodological scheme for the application of the Genetic Algorithm (GA) for the choice of a configuration and a sample size.
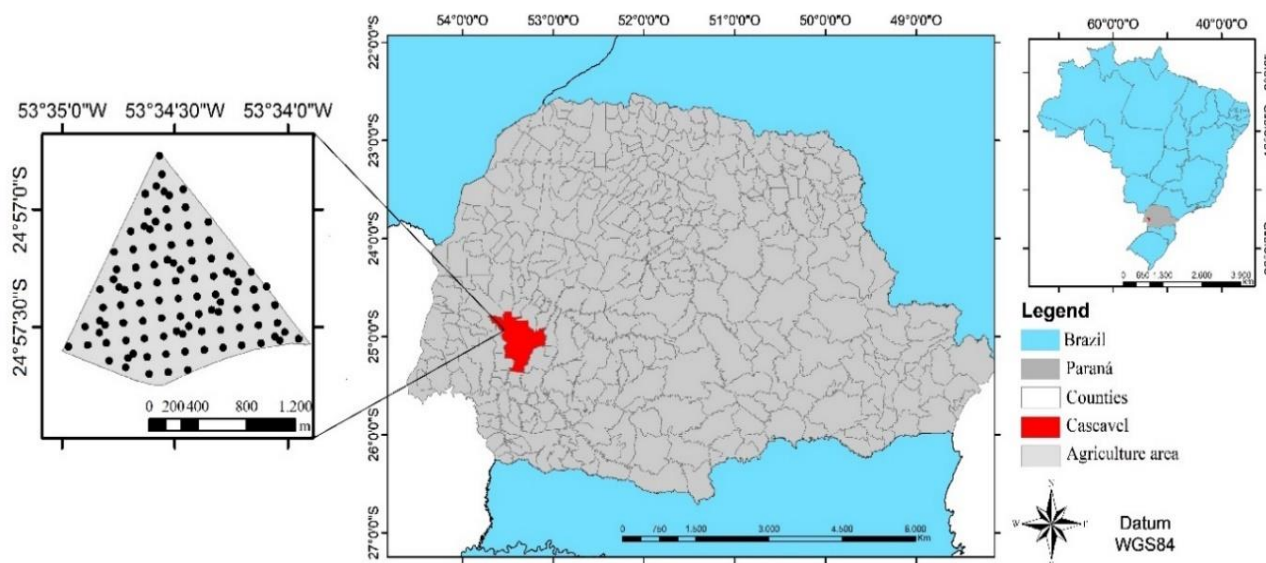
Figure 3. Map of the location of the study area and the sampling configuration.

the content of these attributes in the soil can modify the growth and development phases of the plant, thus affecting the grain and, consequently, the soybean productivity (Taiz *et al.*, 2017).

Soil sampling, to determine the levels of soil chemical attributes, was performed at each demarcated point (Fig. 3). In these, four soil subsamples (Filizola *et al.*, 2006) were collected near these points at a depth of 0.0 to 0.2 m (Arruda *et al.*, 2014), mixed and stored in plastic bags, with samples of approximately 500 g (Alloway, 1995), thus composing the representativeness of the plot, in which the values of micronutrients Cu and Mn were extracted by the Mehlich-1 method, C by Walkley Black, Ca by 1 mol L^{-1} KCI, and pH by CaCl 1:2.5 (w/v), in the Laboratory of the Central Agricultural Research Cooperative (COODETEC), Cascavel-PR, Brazil.

Descriptive and geostatistical analyses were performed for each soil chemical attribute. The existence of anisotropy, using the non-parametric test by Maity & Sherman (2012), was evaluated at a 5% significance level. The non-stationarity of the data was verified by Pearson's linear correlation coefficients between the soil chemical attributes with the X-axis and Y-axis coordinates, indicating the presence of directional trend (Callegari-Jacques, 2003).

Afterward, models of the semivariance function were estimated by the maximum likelihood method (Cressie, 2015) considering the following models: exponential, Gaussian, and the Matérn family with shape parameter k = 1, 1.5, and 2 (Uribe-Opazo *et al.*, 2012). The cross-validation technique selected the best model (Faraco *et al.*, 2008). 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, in a fine grid (5 m in 5 m); and thematic maps of the estimated values of each attribute were constructed (Cressie, 2015).

Subsequently, for each soil chemical attribute, the GA was applied in the same way and with the same criteria applied in the simulations. At the end of the optimization process and for each soil chemical attribute, a small sample size configuration was obtained. And for this new sampling configuration, descriptive and geostatistical analyses were carried out again. For each of the attributes, to compare the thematic maps generated by the geostatistical analysis, considering the initial and optimized sampling configurations, the OA similarity measure and the Kappa (Kp) and Tau (T) agreement indexes were calculated (Krippendorff, 2013).

Simulations, GA routines, and statistical and geostatistical analyses were performed in the R software (R Development Core Team, 2021) using the geoR (Ribeiro Jr & Diggle 2001) and sm (Bowman & Azzalini, 2015) packages.

Results

Simulation study

For all simulations performed in all sampling configurations (Tables S1 to S3), at the end of the optimization process, on average, the minimum and maximum OA values in the populations are very close. The estimated OA value, obtained at the end of the optimization, did not suffer a relevant influence due to the variation of the nugget effect and practical range parameters, which ranged from 0.8606 to 0.8652. This fact was a consequence of the GA stopping criterion, which stops the optimization process when the estimated OA value is greater than or equal to 0.85 (Tables S1 to S3).

These results indicate that the search space for the process solution was reduced by a few values at the end of the optimization process, indicating that the optimization process was efficient and stable.

At the end of the process, there was a low standard deviation value for the estimated OA value. There was also, on average, a relevant increase in OA ranging from 44.42% to 90.35%, compared to the beginning and the end of the optimization process (Tables S1 to S3).

For all the trials and simulated sampling configurations, the best small sampling configurations presented, on average, results between 34 and 53 points (Tables S1 to S3). Therefore, there was a mean reduction in the number of points, which varied from 48% to 67%, concerning the initial sampling configuration. In addition, for all simulated sampling configurations, the smallest sample size was obtained with the smallest value for the nugget effect ($\varphi_1 = 0$) and the highest practical range ($a = 0.90$). The highest sample size occurred in the simulations performed with the highest nugget effect value ($\varphi_1 = 0.80$) and the lowest practical range value ($a = 0$). For almost all trials, systematic sampling showed, on average, the highest optimized sample size.

Besides, for all simulated sampling, the reduced sample size showed a direct trend with the nugget effect and an inverse trend with the practical range, mainly for the smaller nugget effect (Tables S1 to S3).

Sample resizing considering soil chemical attributes

Initially, the commercial area had 102 points. After using the GA, there were 35 points for the sampling configuration of the soil chemical attributes Ca and Mn (corresponding to 34.31% of the number of points in the initial sampling configuration) (Table 1), and 30 points for the attributes C, Cu, and pH (corresponding to 29.41% of the number of points in the initial sampling configuration) (Table 1).

Regarding the variation coefficient, only Mn content showed a very high dispersion in both sampling configurations. Ca and Cu contents showed very high dispersion only in the optimized sampling configuration (Pimentel Gomes, 1985).

In the reduced sample, the Ca, C, Cu, and pH attributes showed the same directional trend behaviors observed in the initial sampling configuration, with values close to Pearson's linear correlation coefficient (Table 1), presenting a moderate linear association of their respective values with the Y-axis coordinates (Callegari-Jacques, 2003). Mn content, on the other hand, presented a moderate linear association of its respective values with

Table 1. Descriptive statistics and Pearson's linear correlation coefficient of the soil chemical attributes Ca [cmolc dm^{-3}], C [g dm^{-3}], Cu [mg dm^{-3}], Mn [cmolc dm^{-3}], and pH, considering the initial and optimized sampling configurations.

Statistic	Sampling configuration	Ca	C	Cu	Mn	pH
n	Initial	102	102	102	102	102
Mean		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
Variance		1.88	10.55	0.68	379.51	0.15
CV (%)		26.41	12.06	27.86	39.50	7.58
Coef. X		0.14	-0.08	-0.08	-0.08	0.14
Coef. Y		0.40	0.36	-0.56	0.59	0.35
n*	Optimized	35	30	30	35	30
No. of samples/No. ha		1/5	1/6	1/5	1/5	1/6
Mean		5.17	26.88	3.01	50.94	5.06
Minimum		2.51	20.65	1.10	23.00	4.40
Median		4.93	26.88	2.90	47.00	5.05
Maximum		11.76	33.90	4.70	93.00	5.90
Variance		3.11	11.52	0.87	414.11	0.13
CV (%)		34.09	12.62	31.07	39.95	7.25
Coef. X		0.13	0.06	-0.01	-0.10	0.17
Coef. Y		0.50	0.39	-0.57	0.62	0.45

CV: coefficient of variation. Coef.: coefficient of Pearson's linear correlation between the soil chemical attributes and X and Y coordinates. n: initial sample size. n*: number of points obtained by the optimization.

the Y-axis coordinates for the initial configuration and a strong linear association of its respective values with the Y-axis coordinates for the reduced sampling configuration (Callegari-Jacques, 2003).

In all soil chemical attributes in both sampling configurations (initial and optimized), the spatial dependence structure was isotropic, because it was not possible to reject the null hypothesis isotropy ($p>0.05$), by the non-parametric MS isotropy test (Maity & Sherman, 2012).

Spatial dependence expresses the behavior of the correlation between observed data, being essential in geostatistical analyses. Regarding the intensity of spatial dependence, all soil chemical attributes, and in both sampling configurations, identified the presence of spatial

dependence. The Ca, C, Mn, and pH attributes showed moderate spatial dependence, while Cu content showed strong spatial dependence (Cambardella *et al.*, 1994) (Table 2).

For both sampling configurations (initial and optimized), the best model of the semivariance function, obtained by the maximum likelihood method by cross-validation was the following: the Matérn family model with $k = 2$ for Ca and C contents; and the exponential model for the Cu, Mn content, and pH attributes (Table 2).

In the initial sampling configuration and the soil chemical attributes Ca, C, Cu, Mn, and pH, an estimated value for the spatial dependence radius (practical range) was observed, ranging from 249.14 m to 805.21 m (Table 2).

Table 2. Estimated values of the parameters of the best adjusted models, with their respective standard deviations, for the soil chemical attributes Ca, C, Cu, Mn, and pH, considering the initial and optimized sampling configuration.

Model	Sampling configuration	Ca	C	Cu	Mn	pH
		Matérn k=2		Exp.		
$\hat{\beta}_0$	Initial	-9,916.96	-21,325.50	6,283.82	$-1.7*10^5$	-2,455.07
$\hat{\beta}_1$		0.001	0.003	-0.001	0.023	0.0003
$\hat{\phi}_1$		1.14	4.50	0.09	95.25	0.10
$\hat{\phi}_1 + \hat{\phi}_2$		1.56	9.08	0.45	241.4	0.13
$\hat{a} = g(\hat{\phi}_3)$		249.14	383.47	805.21	453.49	300.96
$E\hat{P}R\,(\%)$		72.43	49.53	20.71	39.47	73.75
SD ($\hat{\beta}_0$)		$3.0*10^{-9}$	$1.1*10^8$	$4.1*10^{-9}$	$5.9*10^{-8}$	$8.7*10^{-10}$
SD ($\hat{\beta}_1$)		$2.1*10^{-5}$	$7.7*10^{-5}$	$2.9*10^{-5}$	0.00043	$6.3*10^{-6}$
SD ($\hat{\phi}_1$)		0.11223	0.00027	0.06244	$8.8*10^{-5}$	0.02528
SD ($\hat{\phi}_2$)		0.11195	$3.9*10^{-5}$	0.12878	$2.8*10^{-5}$	0.02616
SD (a)		0.02242	0.01961	0.14615	0.05603	0.01952
$\hat{\beta}_0$	Optimized	-16,370.00	-28,110.00	8,299.05	$-1.8*10^5$	-3,346.60
$\hat{\beta}_1$		0.0002	0.004	-0.001	0.025	0.0005
$\hat{\phi}_1$		1.71	3.43	0.09	74.41	0.06
$\hat{\phi}_1 + \hat{\phi}_2$		2.27	9.16	0.56	247.81	0.10
$\hat{a} = g(\hat{\phi}_3)$		311.81	435.15	625.84	579.09	237.42
$E\hat{P}R\,(\%)$		71.09	37.46	14.30	30.02	54.76
SD ($\hat{\beta}_0$)		$5.5*10^{-9}$	$1.5*10^8$	$4.6*10^{-9}$	$8.2*10^{-8}$	$1.2*10^{-9}$
SD ($\hat{\beta}_1$)		$4.0*10^{-5}$	0.00011	$3.3*10^{-5}$	0.00059	$8.9*10^{-6}$
SD ($\hat{\phi}_1$)		0.00169	0.00053	0.07800	0.00019	0.01365
SD ($\hat{\phi}_2$)		0.00113	$8.6*10^{-5}$	0.10143	$6.0*10^{-5}$	0.01325
SD (a)		0.05742	0.03690	0.12970	0.11130	0.00037

Exp.: exponential model. $\hat{\beta}_0$, $\hat{\beta}_1$: estimated values of the parameters of the regression model, which explain the mean, where $\mu = \beta_0 + \beta_1 Y_1$, and Y_1 represents the directional trend identified. φ_1: estimated value of the nugget effect. φ_2: estimated value of the contribution. $\hat{a}$: estimated value of the range (m). $E\hat{P}R = \varphi_1/\varphi_1 + \varphi_2$: estimated value of the relative nugget effect (%). $\varphi_1 + \varphi_2$: estimated value of the sill. SD: standard deviation.

While for the referred soil chemical attributes, considering the optimized sampling configuration, the practical range varied from 237.42 m to 625.84 m (Table 2).

For both sampling configurations, the standard deviations of the parameters were estimated. And, for most of the soil chemical attributes, there was an increase in the estimated standard deviation of the regression model parameters, which explains the mean. The same happened for the nugget effect and practical range. In addition, the-

re was a reduction in the estimated standard deviation of the contribution parameter, when comparing the reduced sampling configuration with the initial sampling configuration (Table 2). The lower standard deviation of the parameters, in relation to its mean, the better the model estimative.

The maps constructed using kriging (Fig. 4) for the optimized sampling configuration presented visual similarities for all soil chemical attributes when compared to
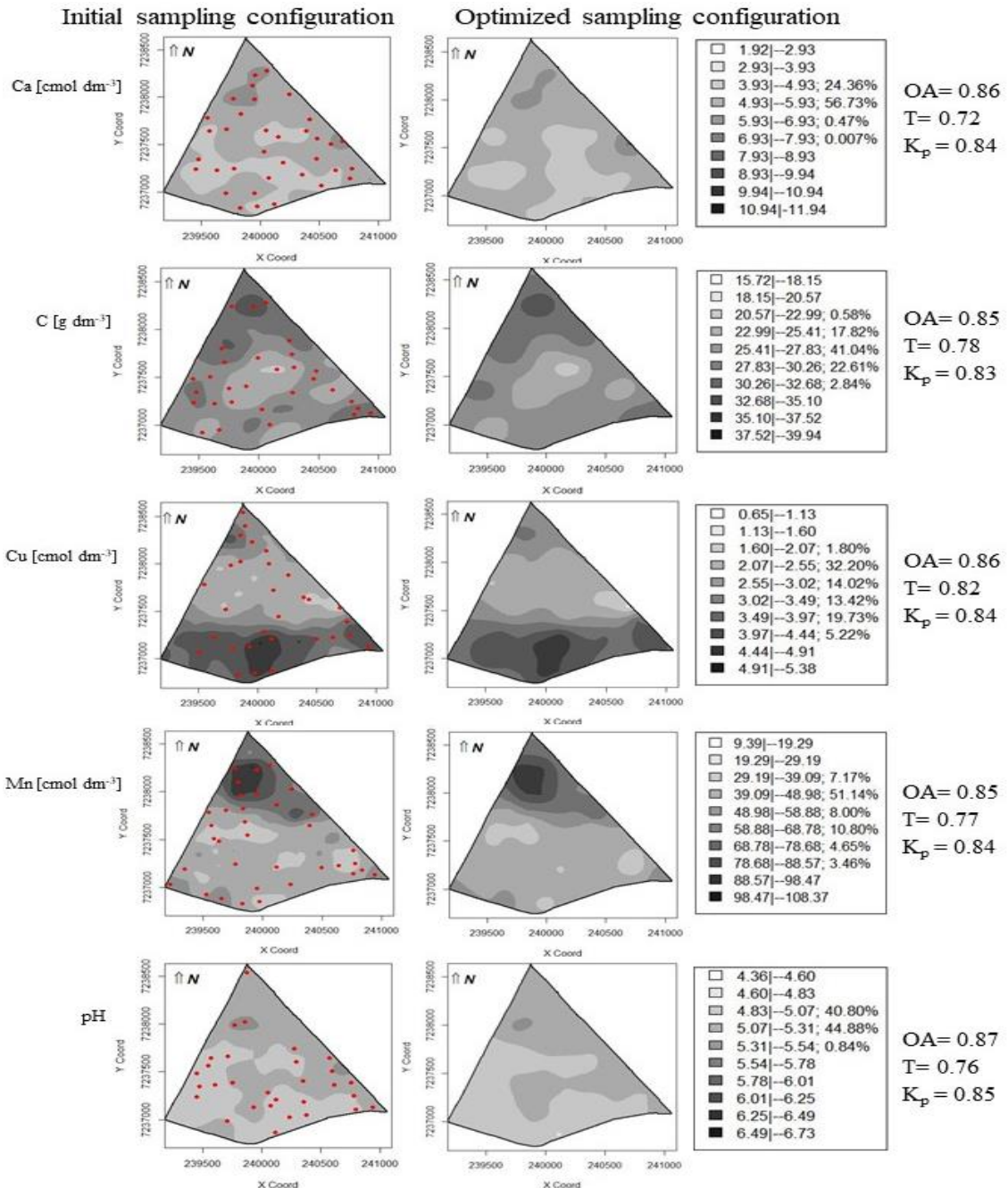

Figure 4. Thematic maps of soil chemical attributes elaborated considering the initial and optimized sampling configuration, where ● represents the location of the selected points in the optimization process for each attribute; the percentage value represents the percentage of pixels with the same classification in each class; and estimated values of the Overall Accuracy (OA), Tau (T), and Kappa (Kp) similarity measures.

the same maps constructed considering the initial sampling configuration. Besides, the highest values of Ca and Mn contents were concentrated in the northern region of the study area, while that of the Cu content presented a concentration of the highest values in the southern region (Fig. 4). The soil chemical attributes Ca content and pH showed more scattered concentrations around the study area (Fig. 4).

For the soil chemical attributes Ca, Cu, Mn, and pH, the highest percentage of pixels that had the same classification are in class four, with the following respective percentages: 56.73%, 32.20%, 51.14%, and 44.88%; while C in class five, with 41.04%, presented pixels classified as equal in both maps.

In the comparison of the thematic maps generated and considering the initial and optimized sampling configurations, the soil chemical attributes Ca, C, Cu, Mn, and pH presented estimated OA values from 0.85 to 0.87.

Discussion

Simulation study

For all simulated samplings, the reduced sample size showed a direct trend with the nugget effect and an inverse trend with the practical range, mainly for the smaller nugget effect. This occurs because the nugget effect is associated with the sampling or analysis errors, which indicate that two observations fairly close together have very different values. In addition, a high nugget effect indicates a low spatial dependence and leads to estimates around the sample mean (Gallardo & Paramá, 2007; Alencar *et al.*, 2019).

These results were similar to those obtained by Guedes *et al.* (2011), who determined only the sampling configuration, with a reduced sample size previously fixed, studying stationary models. Considering stationary models and ESS as a methodology to reduce the sample size, Griffith (2005), Vallejos & Osorio (2014), and Dal Canton *et al.* (2021) found similar results on the inverse trend of the sample size and the value of the practical range. The practical range consists of the spatial dependence radius. Because of this, a greater spatial dependence radius implies correlated samples over a longer distance, which provides redundant information over a longer distance and, consequently, the possibility to use a smaller number of samples to explain spatial variability (Griffith, 2005; Vallejos & Osorio, 2014).

This shows that results obtained with non-stationary data are similar to those obtained considering the stationarity of the data. Atkinson & Lloyd (2007) show that non-stationary models of the semivariogram can lead to greater efficiency in the sampling design, combining sampling density with the character of the local spatial variation.

Sample resizing considering soil chemical attributes

In general, the results obtained with the GA indicated that there was a greater reduction in the number of sample points, when compared to the results obtained in the simulations and in sample resizing with objective function also associated with the efficiency of the spatial prediction proposed by Guedes *et al.* (2011; 2014). These authors considered a fixed reduced sample size and optimization of the sample reduction by the SA algorithm as well and obtained a 40%-50% reduction of the initial sampling configuration. Thus, this study found a greater sample reduction than in similar studies.

On the other hand, the percentage of sample reduction observed in this study is similar to those found by Maltauro *et al.* (2019) optimizing the efficiency of the geostatistical model estimation, such as the objective function based on the inverse-Fisher information matrix; and Domenech *et al.* (2017), who, considering information on auxiliary variables, concluded that 30% of the samples of the area were already sufficient to obtain efficient thematic maps in terms of the representativeness of the spatial variability of the soil depth.

There is no consensus in the literature regarding the number of samples to be collected per hectare. There are studies using grids with a sample every 1 ha or even every 9 ha (Nanni *et al.*, 2011; Cherubin *et al.*, 2014; Siqueira *et al.*, 2014). Thus, the results obtained in this study (Table 1) are within the variation proposed by the literature.

For each soil chemical attribute, there was similarity in the descriptive statistics between the different sample sizes (Table 1), indicating that these sampling configurations are similar, concerning the sets of values of the variables. This fact also occurred in Kestring *et al.* (2015), who analyzed different intensities of soil sampling concerning precision in geostatistical analysis and map interpolation, for precision agriculture, in areas planted with sugar cane and soy, respectively.

Regarding the intensity of spatial dependence, the Ca, C, Mn and pH attributes showed moderate spatial dependence, while Cu content showed strong spatial dependence (Cambardella *et al.*, 1994) (Table 2). This fact makes thematic maps more accurate than maps generated with weak spatial dependence, all of this since there is a lesser contribution of the random component in data variability (Di *et al.*, 1989; Kravchenko, 2003).

Both sampling configurations presented an estimated value of the spatial dependence radius similar for each soil chemical attribute, except for Cu and Mn, which showed a greater difference when compared to the spatial dependence radius considering the initial sampling configuration, with that, obtained considering the optimized. As for the soil chemical attributes Ca, C and Mn, the spatial dependence radius had a slight increase, and a slight

decrease for the Cu and pH attributes (13.5% to 27.7%; 21% to 22.3%, respectively) (Table 2).

The maps elaborated considering both configurations are similar in terms of the content distribution of the attributes in the study area (Fig. 4). This fact was also found for the simulations and is due to the consequence of the algorithm stop criterion (estimated OA value is ≥ 0.85). In addition, according to the estimated values of the Kp and T concordance indexes, all soil chemical attributes presented medium ($0.67 \leq$ Kp < 0.80) or high (Kp ≥ 0.80) accuracy (Krippendorff, 2013), with values estimated between 0.72 to 0.82.

The arrangement of the points obtained by the optimization process (points in red in Fig. 4), in all soil chemical attributes, showed that the optimization algorithm chose the sampling points of the optimized sampling configuration to select points from each subarea or points next to each subarea described by the thematic map of the initial sampling configuration. In the optimized sampling configuration, it is also observed that the corresponding number of points in each subarea is related to the proportion that this subarea represents the agricultural area. This result showed the concentration of points in the regions where there is greater variability, as well as a decrease of the sample density in the most uniform places (Grego *et al.*, 2014). It is also clear that the soil chemical attributes C and pH showed the formation of some small clusters of chosen points, but the position of these clusters did not follow a trend (Fig. 4).

The algorithm sought a total coverage of the study area, that is, it defined a spatial sampling scheme with samples scattered throughout the study area, showing that the sampling scheme tends not to select contiguous samples, this being an important fact for general coverage of the study area (Fattorini *et al.*, 2015). This fact also occurred in a study with sample optimization methodologies, with fixed size and based on space coverage, called: space-filling design (Pronzato, 2017) and cover design (Johnson *et al.*, 1990).

Furthermore, a sampling procedure with greater area coverage, not considering any optimization process, can be obtained through stratified sampling, which in turn can consider the division of the area into several heterogeneous extracts, although with homogeneous samples within each stratum, collecting samples at random within each extract (Wang *et al.*, 2012).

The optimization process used in this study showed that it is possible to reduce the sample size, with a minimization of the loss of similarity of the spatial prediction. There are not similar results in the reduction of the sample size using traditional sampling methodologies such as the systematic sampling used by Kestring *et al.* (2015), whose maps generated with a mean reduction of 66% of the study area did not generate satisfactory values for OA.

References

Alencar NM, dos Santos AC, de Paula Neto JJ, Rodrigues MOD, de Oliveira LBT, 2019. Variabilidade das perdas de solo em Neossolo Quartzarênico sob diferentes coberturas no ecótono Cerrado-Amazônia. Agrarian 12 (43): 71-78. https://doi.org/10.30612/agrarian.v12i43.8081

Alloway BJ, 1995. Heavy metals in soils, 2nd ed. Blackie A & P, Glasgow. https://doi.org/10.1007/978-94-011-1344-1

Anderson JR, Hardy EE, Roach JT, Witmer RE, 2001. A land use and land cover classification system for use with remote sensor data. U.S. Government Print Office. Washington DC. 41 pp.

Arruda MR, Moreira A, Pereira JCR, 2014. Amostragem e cuidados na coleta de solo para fins de fertilidade. Embrapa Amazônia Ocidental Manaus.

Atkinson PM, Lloyd CD, 2007. Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data. Comput Geosci 33 (10): 1285-1300. https://doi.org/10.1016/j.cageo.2007.05.011

Bowman AW, Azzalini A, 2015. R package sm: nonparametric smoothing methods (vers 2.2-5.4). University of Glasgow, UK, & Università di Padova, Italy.

Callegari-Jacques SM, 2003. Bioestatística: princípios e aplicações. Artemed, Porto Alegre.

Cambardella CA, Moorman T, Parkin T, Karlen D, Novak J, Turco R, Konopka A, 1994. Field-scale variability of soil properties in central Iowa soils. Soil Sci Soc Am J 58: 1501-1511. https://doi.org/10.2136/sssaj1994.03615995005800050033x

Catapatti TR, Goncalves MC, Silva Neto MR, Sobroza R, 2008. Sample size and number of replications for assessment of agronomic characters in popcorn. Ciênc Agrotec 32 (3): 855-862. https://doi.org/10.1590/S1413-70542008000300023

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. Ciênc Rural 44 (3): 425-432. https://doi.org/10.1590/S0103-84782014000300007

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. https://doi.org/10.1002/env.2425

Cressie NAC, 2015. Statistics for spatial data, rev. ed. John Wiley & Sons, NY. 928 pp.

Dal Canton LE, Guedes LPC, Uribe-Opazo MA, 2021. Reduction of sample size in the soil physical-chemical attributes using the multivariate effective sample size. J Agr Stud 9 (1): 357-376. https://doi.org/10.5296/jas.v9i1.17473

Dalmago GA, da Cunha GR, Pires JLF, Santi A, Fochesatto E, 2014. Potencial de aplicação da agrometeorologia em agricultura de precisão para produção de grãos. In: Agricultura de precisão: resultados de um novo olhar; Bernardi AC de C, Naime J de M, de Rezende AV, Bassoi LH, Inamasu RY (eds.). pp. 331-337. Embrapa, Brasília.

De Bastiani F, Cysneiros AFJ, Cysneiros AHM, Uribe-Opazo MA, Galea M, 2015. Influence diagnostics in elliptical spatial linear models. Test 24 (2): 322-340. https://doi.org/10.1007/s11749-014-0409-z

Di HJ, Kemp RA, Trangmar BB, 1989. Use of geoestatistics in designing sampling strategies for soil survey. Soil Sci Soc Am J 53 (4): 1163-1167. https://doi.org/10.2136/sssaj1989.03615995005300040028x

Domenech MB, Castro-Franco M, Costa JL, Amiotti NM, 2017. Sampling scheme optimization to map soil depth to petrocalcic horizon at field scale. Geoderma 290: 75-82. https://doi.org/10.1016/j.geoderma.2016.12.012

Embrapa, 2013. Sistema brasileiro de classificação de solos, 3ed. Empresa Brasileira de Pesquisa Agropecuária, Centro Nacional de Pesquisa de Solos, Brasília. 306 pp.

Faraco MA, Uribe-Opazo MA, Silva EA, Johann JA, Borssoi JA, 2008. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. Rev Bras Cienc Solo 32 (2): 463-476. https://doi.org/10.1590/S0100-06832008000200001

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. Environmetrics 26 (3): 216-228. https://doi.org/10.1002/env.2332

Ferreyra RA, Apezteguía HP, Sereno R, Jones JW, 2002. Reduction of soil water spatial sampling density using scaled semivariograms and simulated annealing. Geoderma 110 (3-4): 265-289. https://doi.org/10.1016/S0016-7061(02)00234-3

Filizola HF, Gomes MAF, Souza MD, 2006. Manual de procedimentos de coleta de amostras em áreas agrícolas para análise da qualidade ambiental: solo, água e sedimentos. Embrapa Meio Ambiente, Jaguariúna - SP.

Gallardo A, Paramá R, 2007. Spatial variability of soil elements in two plant communities of NW Spain. Geoderma 139: 199-208. https://doi.org/10.1016/j.geoderma.2007.01.022

Grego CR, Oliveira RP, Vieira SR, 2014. Geoestatística aplicada a agricultura de precisão. In: Agricultura de precisão: resultados de um novo olhar; Bernardi AC de C, Naime J de M, de Rezende AV, Bassoi LH, Inamasu RY (Eds.). pp. 350-360. Embrapa, Brasília.

Griffith DA, 2005. Effective geographic sample size in the presence of spatial autocorrelation. Ann Assoc Am Geogr 95 (4): 740-760. https://doi.org/10.1111/j.1467-8306.2005.00484.x

Grzegozewski DM, Uribe-Opazo MA, De Bastiani F, Galea M, 2013. Influencia local a modelos espaciales lineales Gaussianos: Aplicación a la agricultura. Cienc Inv Agr 40 (3): 537-545. https://doi.org/10.4067/S0718-16202013000300006

Guedes LPC, Ribeiro Jr PJ, Piedade SMS, Uribe-Opazo MA, 2011. Optimization of spatial sample configurations using hybrid genetic algorithm and simulated annealing. Chil J Stat 2 (2): 39-50.

Guedes LPC, Uribe-Opazo MA, Ribeiro Jr PJ, 2014. Optimization of sample design sizes and shapes for regionalized variables using simulated annealing. Cienc Inv Agr 41 (1): 33-48. https://doi.org/10.4067/S0718-16202014000100004

Inamasu RY, Bernardi ADC, Vaz CMP, Naime JDM, Queiros LR, Resende AV, et al., 2011. Agricultura de precisão para a sustentabilidade de sistemas produtivos do agronegócio brasileiro. Embrapa Instrumentação, 2011. p. 14-26.

Johnson ME, Moore LM, Ylvisaker D, 1990. Minimax and maximin distance designs. J Stat Plan Inference 26 (2): 131-148. https://doi.org/10.1016/0378-3758(90)90122-B

Kestring F, Guedes LPC, De Bastiani F, Uribe-Opazo MA, 2015. Thematic maps comparison of different sampling grids for soybean productivity. Eng Agr 35: 733-743. https://doi.org/10.1590/1809-4430-Eng.Agric.v35n4p733-743/2015

Kravchenko AN, 2003. Influence of spatial structure on accuracy of interpolation methods. Soil Sci Soc Am J 67 (5): 1564-1571. https://doi.org/10.2136/sssaj2003.1564

Krippendorff K, 2013. Content analysis an introduction to its methodology, 2nd ed. Sage Publ Ltd, California. 412 pp.

Linden R, 2012. Algoritmos genéticos. Ciência Moderna Ltd, Rio de Janeiro. 496 pp.

Machuca-Mory DF, Deutsch CV, 2013. Non-stationary geostatistical modeling based on distance weighted statistics and distributions. Math Geosci 45 (1): 31-48. https://doi.org/10.1007/s11004-012-9428-z

Maity A, Sherman M, 2012. Testing for spatial isotropy under general designs. J Stat Plan Inf 142 (5): 1081-1091. https://doi.org/10.1016/j.jspi.2011.11.013

Maltauro TC, Guedes LPC, Uribe-Opazo MA, 2019. Reduction of sample size in the analysis of spatial variability of non-stationary soil chemical attributes. Eng Agr 39 (s): 56-65. https://doi.org/10.1590/1809-4430-eng.agric.v39nep56-65/2019

Manchuk JG, Deutsch CV, 2012. Note on working with trends in geostatistics. CCG Annual Report 14: 128-1/128-8.

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

Min L, Cheng, W, 1999. A genetic algorithm for minimizing the makespan in the case of scheduling identical parallel machines. Artif Intell Eng 13 (4): 399-403. https://doi.org/10.1016/S0954-1810(99)00021-7

Montanari R, Souza GSA, Pereira GT, Marques Jr. J, Siqueira DS, Siqueira GM, 2012. The use of scaled semivariograms to plan soil sampling in sugarcane fields. Precis Agric 13:1-11. https://doi.org/10.1007/s11119-012-9265-6

Nanni MR, Povh FP, Demattê JAM, Oliveira R.B, Chicati ML, Cezar E, 2011. Optimum size in grid soil sampling for variable rate application in site-specific management. Sci Agric 68 (3): 386-392. https://doi.org/10.1590/S0103-90162011000300017

Nunes LM, Caiero S, Cunha MC, Ribeiro L, 2006. Optimal estuarine sediment monitoring network design with simulated annealing. J Environ Manage 78 (3): 294-304. https://doi.org/10.1016/j.jenvman.2005.04.024

Pimentel Gomes F, 1985. Curso de Estatística Experimental, 12th. ed. Nobel, São Paulo. 451 pp.

Pronzato L, 2017. Minimax and maximin space-filling designs: some properties and methods for construction. J Soc Fr Statistique 158: 7-36.

R Development Core Team, 2021. R: A language and environment for statistical computing, vers 4.0.0. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr PJ, Diggle PJ, 2001. geoR: a package for geostatistical analysis. R-NEWS1. 15-18.

Ribeiro GG dos S, Tachibana VM, Galo MDLBT, 2016. Influência do delineamento amostral na inferência espacial por geoestatística aplicada a dados de clorofila-a adquiridos em transectos. Rev Bras Cartogr 68 (4): 745-758.

Santos PCD, Santana ACD, Barros PLCD, Queiroz JCB, Vieira TDO, 2011. O emprego da geoestatística na determinação do tamanho "ótimo" de amostras aleatórias com vistas à obtenção de estimativas dos volumes dos fustes de espécies florestais em Paragominas, estado do Pará. Acta Amazon 41 (2): 213-222. https://doi.org/10.1590/S0044-59672011000200005

Sexton RS, Dorsey RE, Johnson JD, 1999. Optimization of neural networks: A comparative analysis of the genetic algorithm and simulated annealing. Eur J Oper Res 114 (3): 589-601. https://doi.org/10.1016/S0377-2217(98)00114-3

Shiflet AB, Shiflet GW, 2014. Introduction to computational science: modeling and simulation for the sciences, 2nd ed. Princeton Univ Press. 857 pp.

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. https://doi.org/10.1016/j.geoderma.2014.04.037

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 352: 373-384. https://doi.org/10.1016/j.geoderma.2018.02.030

Taiz L, Zeiger E, Møller IM, Murphy A, 2017. Fisiologia e desenvolvimento vegetal, 6th ed. Artmed Editora. 888 pp.

Uribe-Opazo MA, Borssoi JA, Galea M, 2012. Influence diagnostics in Gaussian spatial linear models. J Appl Stat 39 (3): 615-630. https://doi.org/10.1080/02664763.2011.607802

Vallejos R, Osorio F, 2014. Effective sample size of spatial process models. Spat Stat 9: 66-92. https://doi.org/10.1016/j.spasta.2014.03.003

Wang JF, Stein A, Gao BB, Ge Y, 2012. A review of spatial sampling. Spat Stat 2: 1-14. https://doi.org/10.1016/j.spasta.2012.08.001

Yang CT, Sung TC, Weng WC, 2006. An improved tabu search approach with mixed objective function for one-dimensional cutting stock problems. Adv Eng Soft 37 (8): 502-513. https://doi.org/10.1016/j.advengsoft.2006.01.005

Zhu Z, Stein ML, 2005. Spatial sampling for design for parameter estimation of the covariance function. J Stat Plan Infer 134 (2) 583-603. https://doi.org/10.1016/j.jspi.2004.04.017