Revista Brasileira de Ciência do Solo

Division - Soil in Space and Time | Commission – Soil Survey and Classification

Spatial multivariate optimization for a sampling redesign with a reduced sample size of soil chemical properties

Tamara Cantú Maltauro[(1)*], **Luciana Pagliosa Carvalho Guedes**[(1)], **Miguel Angel Uribe-Opazo**[(1)] **and Letícia Ellen Dal Canton**[(1)]

[(1)] Universidade Estadual do Oeste do Paraná, Centro de Ciências Exatas e Tecnológicas, Cascavel, Paraná, Brasil.

*** Corresponding author:**
E-mail: tamara_ma02@hotmail.com

Received: July 04, 2022

Approved: December 07, 2022

How to cite: Maltauro TC, Guedes LPC, Uribe-Opazo MA, Canton LED. Spatial multivariate optimization for a sampling redesign with a reduced sample size of soil chemical properties. Rev Bras Cienc Solo. 2023;47:e0220072.
https://doi.org/10.36783/18069657rbcs20220072

Editors: José Miguel Reichert and Yuri Jacques Agra Bezerra da Silva.

ABSTRACT: Precision agriculture can improve the decision-making process in agricultural production, as it gathers, processes and analyzes spatial data, allowing, for example, specific fertilizer application in each location. One of the proposals to deal with spatial heterogeneity of the soil or the distribution of chemical properties is to define application zones (homogeneous subareas). These zones allow reducing both spatial variability of the yield of the crop under study and of the environmental impacts. Considering the soil data, application zones can also represent strata or indicators to direct future soil sampling, thus seeking sample size reduction, for example. This study aimed to obtain an optimized sampling redesign using application zones generated from the assessment of five clustering methods (Fuzzy C-means, Fanny, K-means, McQuitty and Ward). Soil samples were collected in an agricultural area located in the city of Cascavel-Paraná-Brazil, and analyzed in the laboratory to determine the soil chemical properties, referring to four soybean harvest years (2013-2014, 2014-2015, 2015-2016 and 2016-2017). The application zones were obtained through a dissimilarity matrix that aggregates information about the Euclidean distance between the sample elements and the spatial dependence structure of the properties. Subsequently, an optimized sampling redesign, with reduction of the initial sample points, was obtained in these application zones. For the harvest years under study, the K-means and Ward clustering methods efficiently defined the application zones, dividing the study area into two or three application zones. Among the reduced sample configurations obtained by the optimization process, when comparing the initial sample configuration, the one optimized by 25 % (selecting 75 % of the initial configuration points, which corresponds to 76 sample points) was the most effective in terms of the accuracy indices (overall accuracy, Kappa, Tau). This fact indicates greater similarity between the thematic maps of these sample configurations. In this way, the reduced sample configurations could be used to generate the application zones and reduce the costs regarding the laboratory analyses involved in the study.

Keywords: clustering, genetic algorithm, multivariate spatial dissimilarity matrix, sample configuration.

INTRODUCTION

Precision Agriculture (PA) can improve the decision-making process in agricultural production. Differently from traditional agriculture, PA allows specific fertilizer application, irrigation or amendments in each location, that is, at a variable rate. Consequently, its use can contribute to improving yield efficiency and reduce environmental impacts (Ortega and Santibáñez, 2007; Bottega et al., 2017).

In addition to enabling a reduction of contaminants and maximization of agricultural productivity, proper soil handling is directly related to knowledge of the soil attributes' spatial variability (Barbosa et al., 2019). This spatial variability of georeferenced variables can be studied by means of Geostatistics techniques, which also make it possible to determine the degree of spatial dependence between the sample elements in the region and describe the spatial dependence structure of the georeferenced variable in the entire area, thus elaborating the thematic maps (Cressie, 2015; Uribe-Opazo et al., 2021).

One of the proposed ways to deal with the soil spatial heterogeneity and soil chemical properties of the agricultural area is defining management zones (MZs), which is nothing more than delimiting the study area into subareas with similar characteristics, i.e., spatially homogeneous subareas according to certain variables/attributes. With this, it is possible to manage each subarea uniformly and with a similar amount of fertilizers, enabling a more viable strategy for localized management (Rodrigues Jr et al., 2011; Galambošová et al., 2014). As the MZs are generated to be used in several harvest years, it is recommended to use soil variables that do not vary significantly over time, such as topographic data (elevation and slope) and soil physical data (bulk density, soil texture, soil penetration resistance – SPR) (Aikes Jr et al., 2021). When the farmer only has data on soil chemical properties, application zones (AZs) for variable rate fertilizer application recommendations can be generated (Molin, 2006; Aikes Jr et al., 2021).

The difference between MZs and AZs is related to the available variables (stable or unstable) and to the intended use time of the zones (long-term use or merely for future fertilizer application). On the other hand, spatial statistics and multivariate cluster analysis are methods that can be used by both zones. In addition, these zones can represent indicators or strata for soil samplings, reducing the number of samples that need to be collected to perform the soil and crop analyses (Gavioli et al., 2019). Another methodology that allows reducing the number of sample points in a spatial variability study, i.e., spatially characterizing a property by studying its distribution and dispersion in the agricultural area and, consequently, raising the crop productivity level (Landim and Yamamoto, 2013), is the one related to the optimization processes, such as the Genetic Algorithm (GA).

The GA can be used to solve optimization problems found in the real world; it consists of an optimization technique based on the evolution and adaptation process of individuals in a population and intends that only those fit remain in the population, constituting a solution to the problem, i.e., it consists of an iterative process, starting with a population of individuals, which are the possible solutions to the problem. The individuals are evaluated to select the fittest, according to an objective function to be maximized or minimized. The selected individuals are recombined based on the genetic operators and, thus, a new population is generated. This process is carried out until finding the optimal solution or until reaching a stop criterion pre-established by the researcher; more details can be seen in Guedes et al. (2011) and Maltauro et al. (2019).

In the context of obtaining a reduced sample configuration, with a size previously fixed by the researcher and considering univariate optimization processes, the hybrid GA showed that a 50 % sample size reduction produces effective results for the classification of the potassium fertilizer in the area (Guedes et al., 2011). Optimizing the efficiency of the geostatistical model estimation and based on Fisher's information matrix (objective function) as well as the estimation of the values predicted by kriging, maximizing Overall Accuracy (OA, objective function) measure, Maltauro et al. (2019; 2021), respectively obtained that the GA was efficient in sample size reduction, determining a sample size with 29.41 to 39.22 % of the initial sampling points for the soil chemical properties.

In this study, we seek a joint definition of AZs and to determine an optimized sampling redesign (a new reduced configuration) with a reduced size using the GA. Thus, AZs allow to collect more samples in areas with greater variability (heterogeneous areas) and reduce such numbers in more homogeneous areas (Rodrigues Junior et al., 2011). This study aimed to obtain a spatial multivariate optimized sampling redesign with reduced sample size of an agricultural area using application zones as a way to stratify the agricultural area, as well as the optimized process called GA.

MATERIALS AND METHODS

Study area and soil data

Soil samples were collected in an agricultural area located in the city of Cascavel, Paraná State, and the following chemical properties were determined: Al, Base Saturation [V], Ca^{2+}, C, Cu, Fe, K^+, H+Al, Mg^{2+}, Mn, organic matter (OM), P, pH, sum of bases (SB), and Zn; and the Shoemaker, Mac lean and Pratt (SMP) index was calculated (Table 1). The agricultural area has 167.35 ha and is a commercial grain production area located at *Fazenda Três Meninas* at approximately 24.95° South and 53.37° West and with a mean altitude of 650 m above sea level. The soil is classified as *Latossolo Vermelho Distroférrico típico* (Santos et al., 2018) ou Oxisols (Soil Taxonomy), with a clayey texture. The region's climate is classified as mesothermal and super-humid temperate, climate type Cfa (Köeppen classification system), and the mean annual temperature is 21 °C (Aparecido et al., 2016).

Table 1. Soil chemical properties and SMP index used in the study, indicated with an X

Soil chemical properties	Soybean harvest Years			
	2013-2014	2014-2015	2015-2016	2016-2017
Al^{3+} [$cmol_c$ dm^{-3}]		X		
Ca^{2+} [$cmol_c$ dm^{-3}]	X	X	X	X
C [g dm^{-3}]			X	
Cu [mg dm^{-3}]	X		X	
Fe [mg dm^{-3}]	X			
P [g dm^{-3}]				X
H+Al [$cmol_c$ dm^{-3}]	X			X
Shoemaker, Mac lean and Pratt (SMP index)				X
Mg^{2+} [$cmol_c$ dm^{-3}]				X
Mn [mg dm^{-3}]	X	X	X	
OM [g dm^{-3}]				X
K^+ [$cmol_c$ dm^{-3}]				X
pH($CaCl_2$) 1:2.5 (w/v)				X
SB [$cmol_c$ dm^{-3}]			X	X
V [%]				X
Zn [mg dm^{-3}]		X	X	
Number of properties (*p*)	5	4	6	10

The sample configuration or sampling design (arrangement of sampling points) used in this area is lattice plus close pairs (Chipeta et al., 2017), with 102 sampling points. This sampling was chosen because regular sampling allows for a uniform distribution of sampling points throughout the study area. This sampling design consists of a regular grid with a minimum distance of 141 m between the points. In some randomly chosen places, the sampling points were arranged at smaller distances (75 and 50 m between point pairs) (Figure 1). Adding nearby points minimizes estimation errors at smaller scales. Samples were located and georeferenced using a GNSS receiver (GeoExplorer, Trimble Navigation Limited, Sunnyvale, CA, USA) in a Datum WGS84 coordinate reference system, UTM (Universal Transverse Mercator) projection.

Soil sampling was performed in each point indicated (Figure 1). According to the recommendations found in the literature, four soil subsamples were collected at these points, from 0.0 to 0.2 m depth in the vicinity of the points (Arruda et al., 2014), mixed and placed in plastic bags with samples of approximately 500 g, thus comprising the representative sample of the plot. The chemical analyses were performed using the Walkley-Black method (Walkley and Black, 1934).

Considering the database from the Laboratory of Applied and Spatial Statistics (LEA and LEE) at UNIOESTE, the last consecutive harvest years used in this research (2013-2014, 2014-2015, 2015-2016, and 2016-2017) for which soil samples were collected and analyzed in the laboratory to determine the chemical properties. Only the soil chemical properties were used because the database does not have soil physical properties for all years.

Initial analysis

Considering all harvest years, descriptive and geostatistical analyses were performed for each soil chemical property to verify the existence of directional trends, spatial dependence and anisotropy (Figure 2a). Directional trends represent a linear association between the respective values of the soil chemical properties with the coordinates of the X or Y axis, and were assessed by Pearson’s linear correlation coefficient (r), in which values above 0.30 in a module indicate a directional trend (Callegari-Jacques,

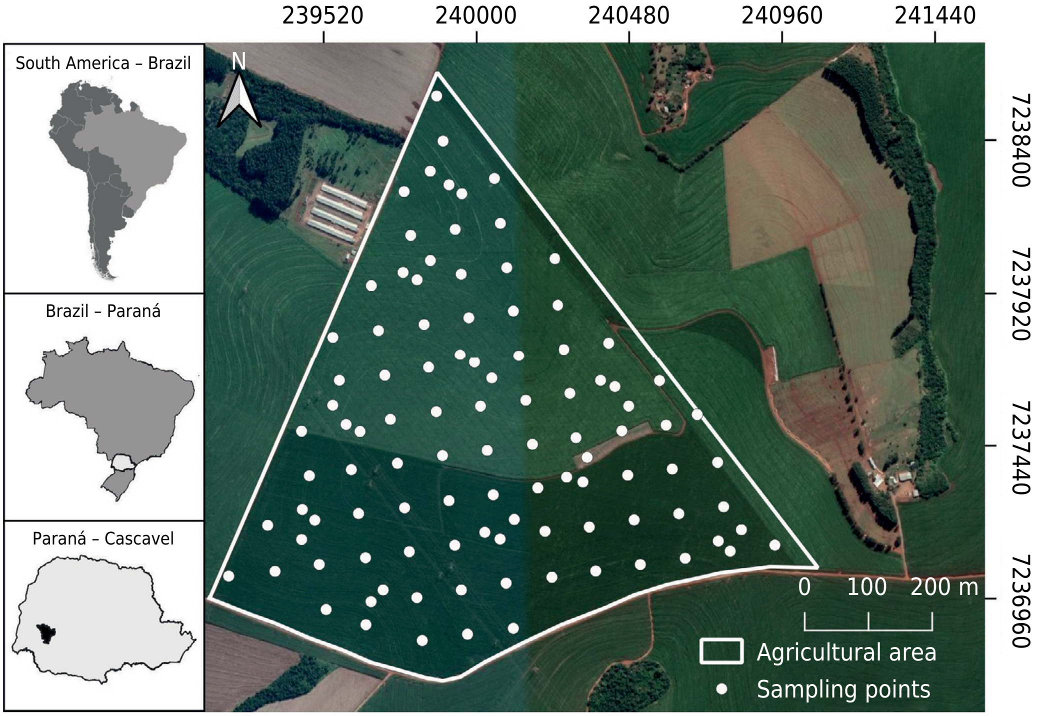

Figure 1. Agricultural area and sampling points.

2003). Anisotropy was assessed by analyzing the directional semivariograms (Guedes et al., 2018) and the non-parametric Maity and Sherman (2012), considering 5 % significance.

To understand and describe spatial dependence, the parameters of the geostatistical models were estimated: exponential, Gaussian and Matérn family with shape parameter k_m = 2.5 using the maximum likelihood method (Uribe-Opazo et al., 2012), with choice of the best-adjusted model performed following the cross-validation method (Faraco et al., 2008), as it is widespread and widely used in the literature and a technique of estimation errors that allows making comparisons between estimated and sampled values. The Relative Nugget Effect (RNE) coefficient was calculated with the best-estimated model. Consequently, only the soil chemical properties that presented spatial dependence were used in this study, disregarding those that were not spatially dependent (strong spatial dependence: 0.25≤ RNE; moderate spatial dependence: 0.25< RNE <0.75; and weak spatial dependence: RNE ≥0.75) (Cambardella et al., 1994) (Table 1).

Subsequently, thematic maps corresponding to each soil chemical property were prepared considering the spatial prediction of each property at locations not sampled in the agricultural area under study, through ordinary kriging (considered an optimal interpolator, due to the way in which the weights are distributed, so that the estimator cannot be biased and must have minimum variance) and with pixels representing 10 × 10 m areas (maximum number of interpolated points that made it possible to implement cluster analyses) (Cressie, 2015).

Acquisition of the multivariate spatial dissimilarity matrix

Considering the soil chemical properties selected in each harvest year (Table 1), a dissimilarity matrix was used to generate the AZs, which aggregates information on the Euclidean distance between the sample elements and the spatial dependence structure of the properties. Consequently, all the locations were compared in pairs to obtain the spatial and multivariate dissimilarity matrix. For this, in each pair of *i* and *j* locations (*i*,*j* = 1, …, *n*) in which the *p* properties had already been measured (Table 1), the similarity coefficient proposed by Gower (1971) was calculated (Equation 1; Figure 2b).

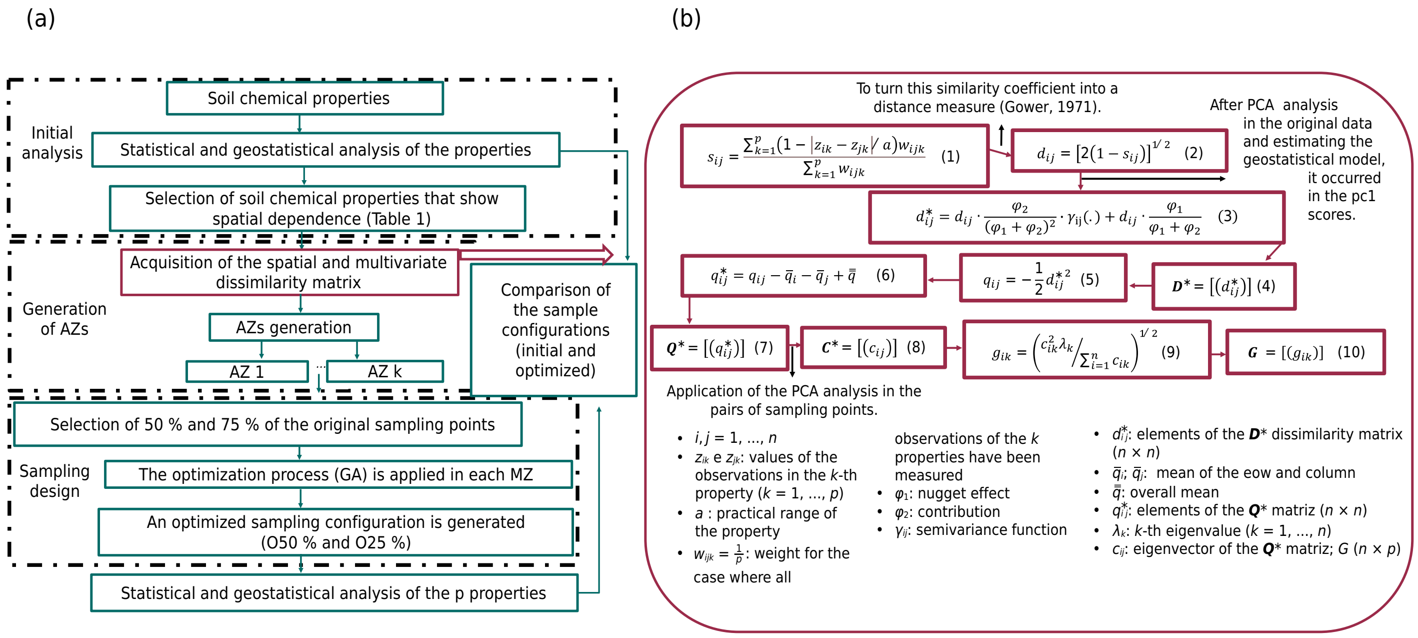
(a)
Initial analysis
Soil chemical properties
Statistical and geostatistical analysis of the properties
Selection of soil chemical properties that show spatial dependence (Table 1)
Generation of AZs
Acquisition of the spatial and multivariate dissimilarity matrix
AZs generation
AZ 1
...
AZ k
Comparison of the sample configurations (initial and optimized)
Sampling design
Selection of 50 % and 75 % of the original sampling points
The optimization process (GA) is applied in each MZ
An optimized sampling configuration is generated (O50 % and O25 %)
Statistical and geostatistical analysis of the p properties
(b)
To turn this similarity coefficient into a distance measure (Gower, 1971).
After PCA analysis in the original data and estimating the geostatistical model, it occurred in the pc1 scores.
$s_{ij} = \frac{\sum_{k=1}^{p}(1-|z_{ik}-z_{jk}|/a)w_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$ (1)
$d_{ij} = [2(1-s_{ij})]^{1/2}$ (2)
$d_{ij}^{*} = d_{ij} \cdot \frac{\varphi_2}{(\varphi_1+\varphi_2)^2} \cdot \gamma_{ij}(.) + d_{ij} \cdot \frac{\varphi_1}{\varphi_1+\varphi_2}$ (3)
$\boldsymbol{D}^{*} = [(d_{ij}^{*})]$ (4)
$q_{ij} = -\frac{1}{2} d_{ij}^{*2}$ (5)
$q_{ij}^{*} = q_{ij} - \bar{q}_i - \bar{q}_j + \bar{\bar{q}}$ (6)
$\boldsymbol{Q}^{*} = [(q_{ij}^{*})]$ (7)
$\boldsymbol{C}^{*} = [(c_{ij})]$ (8)
$g_{ik} = \left(c_{ik}^{2}\lambda_k / \sum_{i=1}^{n} c_{ik}\right)^{1/2}$ (9)
$\boldsymbol{G} = [(g_{ik})]$ (10)
Application of the PCA analysis in the pairs of sampling points.
• $i,j = 1, \ldots, n$
• z_{ik} e z_{jk}: values of the observations in the k-th property ($k = 1, \ldots, p$)
• a : practical range of the property
• $w_{ijk} = \frac{1}{p}$: weight for the case where all observations of the k properties have been measured
• φ_1: nugget effect
• φ_2: contribution
• γ_{ij}: semivariance function
• d_{ij}^{*}: elements of the $\boldsymbol{D}^{*}$ dissimilarity matrix ($n \times n$)
• $\bar{q}_i$; $\bar{q}_j$: mean of the eow and column
• $\bar{\bar{q}}$: overall mean
• q_{ij}^{*}: elements of the $\boldsymbol{Q}^{*}$ matriz ($n \times n$)
• λ_k: k-th eigenvalue ($k = 1, \ldots, n$)
• c_{ij}: eigenvector of the $\boldsymbol{Q}^{*}$ matriz; G ($n \times p$)

Figure 2. Methodology to obtain the optimized sample configurations (a) and the new matrix of properties from a dissimilarity matrix (b).

The dissimilarity matrix was obtained based on the methodological sequence described by Oliver and Webster (1989) (Figure 2b).

To modify equation 2 of this methodology (Figure 2b), in order to consider the geographic distances between the observations sampled and the spatial variability of the properties, a Principal Component Analysis (PCA) was performed in the original data to reduce dimensionality without losing the information contained in all properties. In this PCA, the first principal component (PC1) was selected, as it explains most of the data variation. Considering the PC1 scores, the geostatistical models were adjusted analogously to the methodology used for the soil chemical properties in order to obtain an estimation of the parameters of the geostatistical model for the PC1 scores. With these parameters, the dissimilarity matrix $\boldsymbol{D}^*$ was obtained (Equation 3; Figure 2b). In this way, the matrix adds information about the Euclidean distance between the sample elements (Landim and Yamamoto, 2013), as well as the spatial dependence structure of the properties (Uribe-Opazo et al., 2012).

Based on the calculations (Equations 5 to 9; Figure 2b) performed in the $\boldsymbol{D}^*$ matrix elements, the $\boldsymbol{G}$ matrix columns were obtained (Equation 10; Figure 2b), which are the new variables (chemical properties of the soil after using the dissimilarity matrix). In this way, one selects the number of ρ columns corresponding to the number of original attributes. Subsequently, a geostatistical model was adjusted and data interpolation of the values of the soil chemical properties was performed through kriging, with pixels representing 10 × 10 m areas. The interpolated values of the soil chemical properties were used to obtain the AZs in the agricultural area (Gavioli et al., 2016).

Spatial clustering of the agricultural area

Considering the multivariate spatial dissimilarity matrix and the most cited clustering methods in the literature, five methods were evaluated for the agricultural area's clustering, three of them hierarchical and two partitioned, namely: Fuzzy analysis clustering (Fanny), Fuzzy C-means, K-means, McQuitty and Ward (Ward Jr, 1963; McQuitty, 1966; MacQueen, 1967; Bezdek, 1981; Kaufman and Rousseeuw, 1990), respectively. Five evaluation criteria were used to select the method that provided the best data clustering, namely: Dunn Index (D), Davies Bouldin Index (DB), C Index, SD Index and Variance Reduction Index (VR) (Dunn, 1974; Hubert and Levin, 1976; Davies and Bouldin, 1979; Halkidi et al., 2000; Gavioli et al., 2016).

To define the adequate number of clusters for each harvest year, the scatter plots of the Sum of Squares of Errors (SSE) against the number of clusters (knee graph) were used in each clustering method, as well as the silhouette scatter plot against the number of clusters. These methods were used for their stabilization and for presenting satisfactory results in the literature (Shi and Zeng, 2013; Yi et al., 2013; Martarelli and Nagano, 2016). In the SSE graph, the mean distance decreases as the number of clusters increases. To find the optimal number of clusters, it is necessary to find the cluster with a sharp drop; therefore, this will be the sweet spot of the clusters. For the silhouette graph, the cluster that presents the highest value or peak of the graph is observed (Tan et al., 2009; Yi et al., 2013). Thus, with the optimal number of clusters and clustering method selected, AZ maps were generated for all harvest years considering the soil chemical properties that showed spatial dependence.

Optimized sampling redesign (sampling points selected by the GA with sampling reduction)

After dividing the agricultural area into AZs, a sample reduction process was carried out, selecting sampling points within each AZ for all the harvest years. Obtaining a new reduced sample configuration was considered an optimization problem (Guedes et al., 2014; Maltauro et al., 2019), as the objective of optimization is to seek the best solutions to achieve the objective of the problem; in the proposal of this article, the research was to obtain the best sample configurations with reduced sizes. The intention was to reduce the initial sample configuration by 25 and 50 % (Figure 2b), selecting the sampling points within each AZ.

The methodology developed by the GA to obtain the optimized sample configuration was similar to the one developed by Maltauro et al. (2021), only changing the objective function. To obtain an optimized sample configuration for each harvest years, considering all soil properties, it was decided to work with multi-objective optimization, in which it is possible to find viable solutions that simultaneously optimize several objective functions (Deb and Kalyanmoy, 2001). To such end, the Sum of Weights (SW) method was used, which consists of the sum of the objective function corresponding to each property, adjusted by a weight (Branke et al., 2008; Pantuza Junior, 2016).

In this study, optimization efficiency was evaluated based on spatial prediction. Subsequently, we considered a multi-objective function to be minimized using the SW method (Equation 11), based on a measure of similarity between the initial and optimized sample configurations of each of the soil chemical properties, called OA (Guedes et al., 2014; Maltauro et al., 2021) methodology.

$$\begin{cases} min\ \ F(x_i) = \sum_{k=1}^{p} [1 - \mathrm{OA_k}\ (x_i)] * w_m \\ 0 \leq x_i \leq 1 \qquad\qquad i = 1, 2, \ldots, n,) \end{cases} \qquad \text{Eq. 11}$$

in which: x_i is a possible sample configuration for the problem, with sample size i (i=1,2,⋯,n), in which n is the number of sampling points; w_m = 1/p is the weight for each objective function the k-th soil properties (Equation 12)

$$f_k\ (x_i) = 1 - OA_{\mathrm{k}}\ (x_i) \qquad \text{Eq. 12}$$

with k = 1,..., p, in which p is the number of soil attributes, so that $w_{\mathrm{k}} \in [0, 1]$, $\sum_{\mathrm{k}=1}^{p} w_{\mathrm{k}} = 1$ and $OA_k\ (x) \in [0,1]$. Consequently, when minimizing $F(x_i)$, which is the linear combination of the k objective functions, lower $f_k\ (x_i)$ values will be obtained, which corresponds to getting a higher OAk (x_i) values.

With each optimized sample configuration, the descriptive and geostatistical data analyses were performed again. Finally, the initial and optimized sample configurations were compared using metrics that express the similarity of the thematic maps obtained through kriging, namely: OA (Anderson et al., 2001) and the Kappa (Kp) and Tau (T) agreement indices (Krippendorff, 2013) (Figure 2b).

Computational resources

The routines for calculating the spatial and multivariate dissimilarity matrix, clustering, sample configuration, optimization and other statistical and geostatistical analyses were developed in the R software (R Development Core Team, 2022), considering the ClassInt, cluster, clusterCrit, data.table, e1071, fastcluster, geoR, psych, Splancs and vegan packages.

RESULTS

Initial sample configuration

For all the harvest years, the soil chemical properties presented dispersion of their values around the mean, or even homogeneous data. The Ca^{2+}, C, Cu, Fe, H+Al, Mn, K^+, Mg^{2+}, OM and P contents, as well as pH and SB, had mean values considered average, high or very high for land-use. In turn, the Zn mean value can be classified as low or average, and the Al and V values were classified as very low or low.

For the directional trend, only the Zn content in the 2014-2015 harvest year presented a moderate linear association of its respective values with the X axis coordinates, with a Pearson's linear correlation coefficient value greater than 0.30 in a module. For all the harvest years, the soil chemical properties presented moderate (0.25< RNE <0.75) or strong (0.25≤ RNE) spatial dependence (Table 2).

Regarding the estimated value for the spatial dependence radius (practical range), the 2013-2014 and 2015-2016 harvest years presented greater variation, from 157.70 to 707.86 m and from 149.73 to 855.10 m, respectively, whereas the 2014-2015 and 2016-2017 harvest years exhibited less variation in the practical range, from 128.61 to 453.07 m and from 126.32 to 385.62 m, respectively. This variation in the practical ranges can be influenced by the chosen model and the sample size reduction (Table 2).

Spatial clustering of the agricultural area

For the 2013-2014, 2015-2016 and 2016-2017 harvest years, considering the scatter plots of the number of clusters versus the SSE and silhouette ones, the best number of

Table 2. Descriptive statistics and estimated values of the geostatistical model parameters for the soil chemical properties and SMP index, referring to each harvest year and considering the initial sample configuration

Year	Properties	Descriptive statistics		Estimation of the properties by the geostatistical model					
		Mean	CV	Models	$\hat{\mu}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{a}$	$\widehat{RNE}$
2013-2014	Ca^{2+}	6.22	22.46	Gaus.	6.19	1.08	0.87	179.00	55.27
	Cu	1.21	60.18	Gaus.	1.26	0.28	0.25	707.86	52.68
	Fe	37.10	22.41	Gaus.	37.37	35.93	33.06	217.86	51.67
	H+Al	8.60	22.55	Exp.	8.62	2.62	2.18	157.70	54.54
	Mn	60.96	33.69	Gaus.	60.31	171.59	225.55	203.98	43.20
2014-2015	Al	0.28	126.28	M. k_m=2.5	0.28	0.02	0.10	128.61	15.65
	Ca^{2+}	5.38	25.11	Exp.	5.40	1.05	0.75	231.56	58.49
	Mn	76.54	27.43	Gaus.	77.30	233.70	209.80	453.07	52.69
	Zn	2.81	61.61	Gaus.	-326.73; 0.001	0.54	2.28	162.73	19.12
2015-2016	C	32.01	10.58	Exp.	31.80	5.97	5.37	576.28	52.65
	Ca^{2+}	5.50	24.12	Gaus.	5.53	1.29	0.48	284.08	72.28
	Cu	3.82	23.78	Exp.	4.02	0.33	0.52	855.10	39.41
	Mn	86.41	25.66	Gaus.	86.78	268.79	226.14	367.29	54.30
	SB	7.93	25.20	Exp.	7.93	2.73	1.22	149.73	69.03
	Zn	4.97	40.92	Gaus.	5.10	1.59	3.04	367.65	34.30
2016-2017	Ca^{2+}	4.05	20.93	Gaus.	4.07	0.35	0.34	223.41	50.82
	H+Al	6.30	20.10	M. k_m=2.5	6.31	0.72	0.91	293.01	44.21
	K^+	0.29	38.69	Gaus.	0.29	0.003	0.009	126.32	27.50
	Mg^{2+}	1.72	41.45	M. k_m=2.5	1.79	0.06	0.44	239.11	13.17
	Mo	41.86	24.82	Gaus.	43.04	52.09	58.72	385.62	47.00
	P	19.38	54.16	M. k_m=2.5	19.15	32.27	74.79	236.45	30.14
	pH	4.53	6.63	M. k_m=2.5	4.53	0.03	0.05	200.62	38.71
	SB	6.07	22.38	Exp.	6.16	0.02	1.79	236.59	1.40
	SMP	5.71	4.35	M. k_m=2.5	5.72	0.04	0.02	332.18	67.76
	V	48.94	18.91	Exp.	49.16	17.52	66.94	227.56	20.74

CV: coefficient of variation; $\hat{\mu} = \beta_0$: estimated mean; $\hat{\varphi}_1$: estimated nugget effect; $\hat{\varphi}_2$: estimated contribution; $\hat{a}$: estimated practical range; $\widehat{RNE}$: estimated relative nugget effect ($\widehat{RNE} = \hat{\varphi}_1/\hat{\varphi}_1 + \hat{\varphi}_2$) (%) for properties that showed a directional trend $\hat{\mu} = \beta_0 + \beta_1 Y_1$, in which $\hat{\beta}_0$ (first value of the mean column), $\hat{\beta}_1$ (second value of the mean column): estimated values of the parameters of the regression model and Y_1 represents the directional trend identified; Exp.: exponential model; Gaus.: Gaussian model; M. km=2.5: Matérn model with smoothness parameter km = 2.5; the units of measure of soil chemical properties are found in table 1.

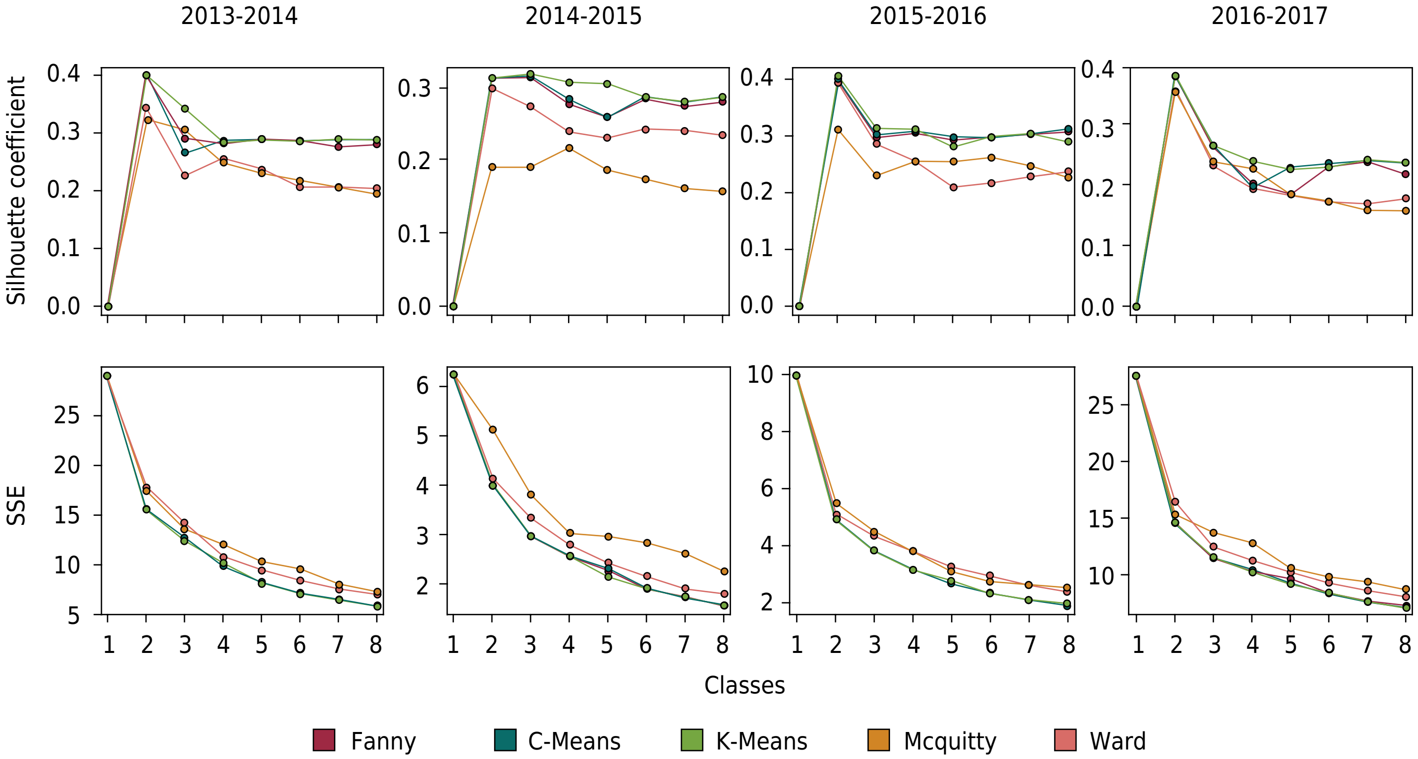

Figure 3. Silhouette coefficient and knee graphs (SSE).

clusters for all the clustering methods was $k_c = 2$ since, with this number of clusters, the highest value of the Silhouette coefficient and the lowest SSE value were obtained. In turn, for the 2014-2015 harvest year and for most of the clustering methods, the ideal number of clusters was $k_c = 3$ (Figure 3). As for the interpretation of the chemical properties available in the soil within each AZ, it was noticed that all the soil chemical properties presented average, high or very high values for the soil in the state of Paraná, except for the Al and pH soil chemical properties, which were classified as low or very low.

Considering the evaluation criteria, K-means was the best clustering method for the 2013-2014, 2014-2015 and 2015-2016 harvest years since, in this method, the lowest values of the DB, C and SD indices were obtained, as well as the highest values of the D and VR indices (Table 3). Regarding the 2016-2017 harvest year, a tie was observed between the Ward and Fuzzy C-means clustering methods (Table 3). In addition to that, certain similarity was verified in the maps of the clusterings in relation to the definition of the AZs. Consequently, the Ward clustering method was selected for the 2016-2017 harvest year, as its execution is simpler and requires less computational time.

Differences are observed when comparing the AZs generated for each harvest year, which can be explained by the fact that different soil chemical properties are used in each harvest year. However, it is noted that a larger AZ was created in all harvest years (red color, Figure 4). In addition to that, it was observed that there is at least one AZ in the Southwest region in all the harvest years (Figure 4). And, except for the 2015-2016 harvest year, it was also possible to find at least one AZ in the North region (Figure 4).

In harvest years that had two AZs, the largest (red color) occupied 106.85 ha (64 % of the area), 99.61 ha (60 % of the area) and 120.38 ha (74 % of the area) for the 2013-2014, 2015-2016 and 2016-2017 harvest years, respectively (Figure 4). The 2014-2015 harvest year, which featured three AZs, had 69.64 and 54.03 hectares in the two largest AZs, corresponding to 42 and 32 % of the total area, respectively (Figure 4). Then, within the harvest years studied, two or three AZs were selected in the study area.

Table 3. Evaluation measures according to the clustering method used to generate application zones

Harvest year	Indices	Fanny	Fuzzy C-means	K-means	Mcquitty	Ward
2013-2014	D	0.0013	0.0049	**0.0053**	0.0038	0.0040
	DB	1.0136	1.0007	**1.0004**	1.1620	1.1519
	C	0.1264	0.1230	**0.1228**	0.2051	0.1834
	SD	34.3238	33.9659	**33.9549**	38.2531	37.0772
	VR	46.1936	46.2690	**46.2692**	39.4736	38.3279
2014-2015	D	0.0031	0.0044	0.0034	0.0047	**0.0083**
	DB	1.1280	1.1115	**1.1081**	1.4141	1.1822
	C	0.1391	0.1368	**0.1367**	0.2464	0.1759
	SD	79.1748	77.4998	**76.5687**	99.0186	82.8300
	VR	52.3989	**52.5292**	51.1738	38.8561	46.4901
2015-2016	D	0.0096	0.0076	**0.0106**	0.0099	0.0095
	DB	0.9552	0.9471	**0.9467**	0.9787	0.9564
	C	0.1032	0.1016	**0.1015**	0.1457	0.1159
	SD	57.0478	56.6668	56.6466	57.1677	**56.6015**
	VR	50.4092	50.4698	**50.4701**	44.7660	48.6003
2016-2017	D	0.0078	**0.0084**	0.0083	0.0069	0.0063
	DB	1.0282	1.0180	1.0184	1.0726	**0.9981**
	C	0.1085	**0.1049**	0.1050	0.1291	0.1568
	SD	35.3339	35.0227	35.0346	36.1047	**34.2426**
	VR	46.8857	46.9487	**46.9490**	44.3445	40.0984

D: Dunn Index; DB: Davies Bouldin Index; C: C Index; SD: SD Index; and VR: Variance Reduction Index. The best results of the indices are highlighted in bold type.

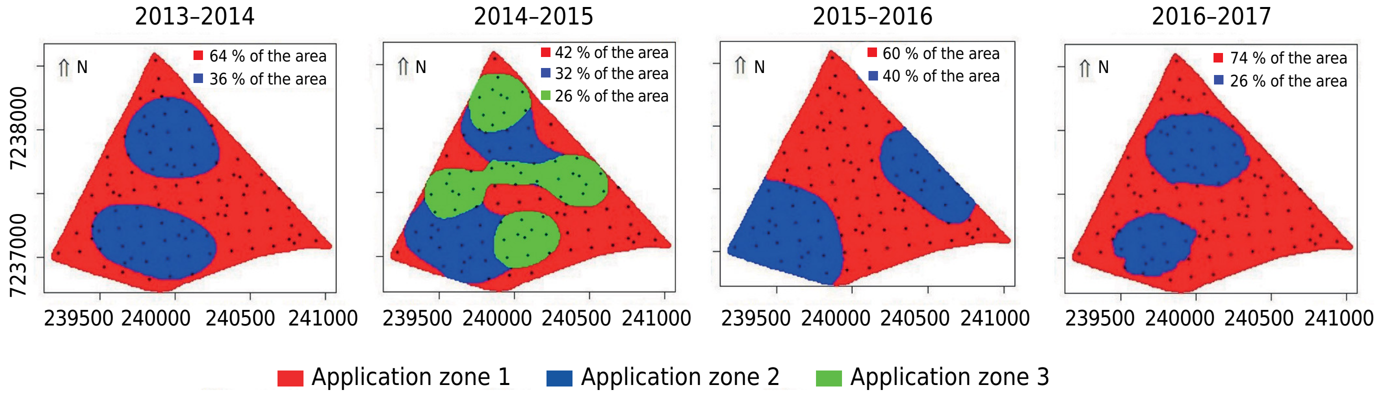

Figure 4. Thematic maps with the best number of application zones and clustering method chosen for each harvest year.

Optimized sample configuration

Considering all harvest years, the sampling points were distributed across all AZs, with the largest zones presenting the highest number of sampling points, thus collecting a higher number of samples in areas with greater spatial variability, as well as reducing this number in more homogeneous areas. This result was also simultaneously influenced by the size of the AZs and by the uniform distribution of sample points over the agricultural area based on the original design (lattice plus close pairs). Thus, AZ1 covered 60, 32, 65 and 73 sampling points, which correspond to 59, 31, 64 and 72 % of the total points in the study area, respectively, for the 2013-2014, 2014-2015, 2015-2016 and 2016-2017 harvest years. In turn, AZ2 comprised 42 sampling points (41 % of the total points of the study area), 26 sampling points (26 % of the total points of the study area), 37 sampling points (36 % of the total points of the study area) and 29 sampling points (28 % of the total points in the study area) (Figure 5). In addition, in the 2014-2015 harvest year, AZ3 included 44 sampling points (43 % of the total points in the study area). The optimized sample configurations that removed 50 % of the initial sampling points (O50) obtained 51 sampling points, and those that removed 25 % (O25) of the initial sampling points had 76 sampling points distributed in the agricultural area (Figure 5). Greater reductions were not possible, as the number of sampling points would not meet the geostatistical analysis criteria, that is, having at least 30 pairs for the calculation of semivariances.

For all the harvest years, similarity in the descriptive statistics was observed between the O50 and O25 sample configurations and the initial sample configurations (Tables 2, 4, and 5). For the 2015-2016 harvest year, both optimized sample configurations to obtain the Cu content presented a directional trend in the Y direction (North-South) (r >30). For the 2014-2015 harvest year, the Zn content presented a directional trend in the X direction (East-West) (r >30) for the O50 and O25 sample configurations, unlike the initial sample configuration, which presented a directional trend in the Y direction (North-South) (r >30).

For the 2013-2014 harvest year, only the H+Al and Mn contents presented a change in the classification of the spatial dependence intensity, from moderate (0.25< RNE <0.75) to weak (RNE ≥0.75) or to strong (0.25≤ RNE) spatial dependence for the optimized sample configurations (Table 4). Regarding the 2014-2015 harvest year, the Al and Zn soil chemical properties presented a nugget effect for at least one optimized sample configuration; and the Ca content had strong spatial dependence (0.25≤ RNE) in the O50 sample configuration (Table 4).

For the 2015-2016 harvest year, all soil chemical properties presented moderate spatial dependence (0.25< RNE <0.75) in the initial sample configuration and in the O25 sample configuration, whereas in the O50 sample configurations, the Ca and Mn contents showed a pure nugget effect, SB indicated weak spatial dependence (RNE ≥0.75) and Zn evidenced strong spatial dependence (0.25≤ RNE) (Table 5). Finally, for the

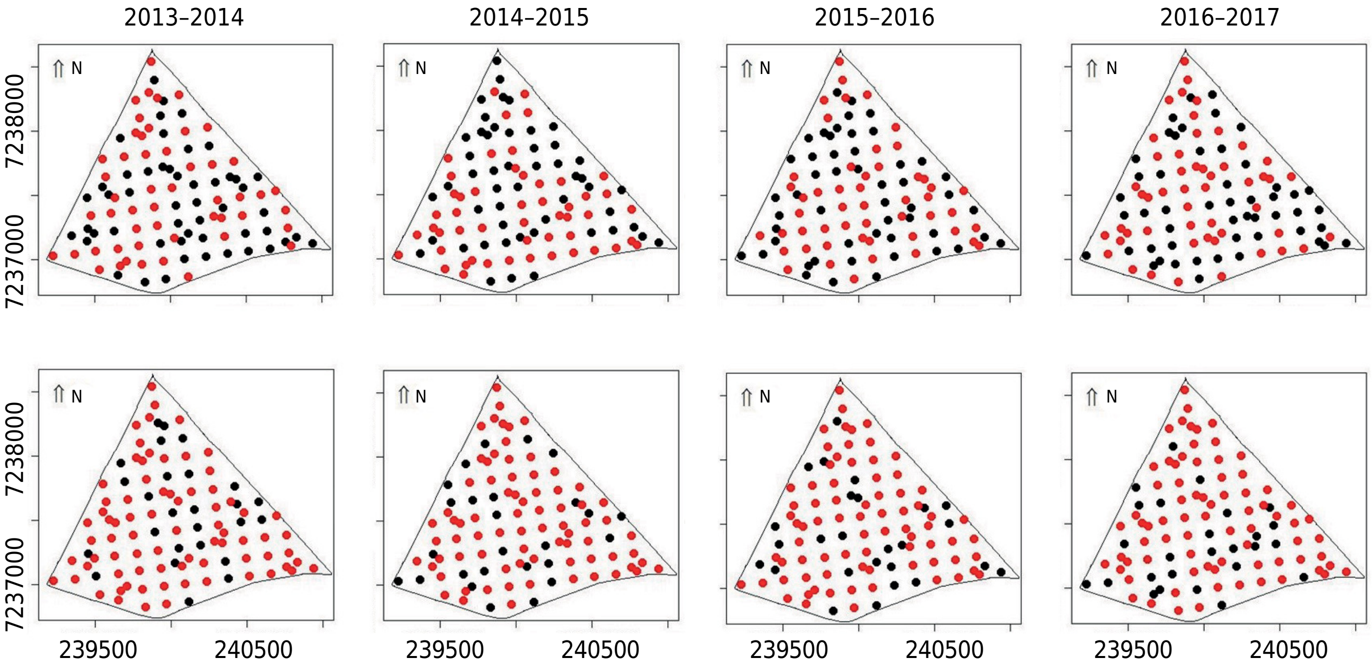

Figure 5. Initial (•) and optimized (•) sample configurations for the 2013-2014, 2014-2015, 2015-2016 and 2016-2017 harvest years.

Table 4. Descriptive statistics and estimated values of the adjusted geostatistical model parameters for the soil chemical properties and with the sample configurations optimized by 50 and 25 %, referring to the 2013-2014 and 2014-2015 harvest years

Year	Method	Properties	Descriptive statistics		Estimation of the parameters by the geostatistical model					
			Mean	CV	Models	$\hat{\mu}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{a}$	$\widehat{RNE}$
2013-2014	O50	Ca^{2+}	6.30	24.80	M. k_m=2.5	6.30	1.00	1.38	178.55	42.04
		Cu	1.17	61.12	Gaus.	1.27	0.35	0.24	987.02	59.47
		Fe	36.69	26.55	Gaus.	36.65	71.46	26.93	673.32	72.63
		H+Al	8.57	26.70	Gaus.	8.60	1.26	3.91	139.37	24.38
		Mn	62.69	33.42	Exp.	61.37	2.18	407.07	176.08	0.53
	O25	Ca^{2+}	6.28	24.09	M. k_m=2.5	6.25	1.34	0.92	193.98	59.27
		Cu	1.25	59.13	Gaus.	1.29	0.35	0.20	770.36	63.37
		Fe	37.21	24.69	Gaus.	37.31	54.72	29.14	252.47	65.25
		H+Al	8.53	27.64	Gaus.	8.53	4.93	0.56	195.46	89.81
		Mn	61.64	36.12	Exp.	60.93	80.43	382.05	268.36	17.39
2014-2015	O50	Al^{3+}	0.34	110.69	Exp.	0.34	0.00	0.14	104.92	0.00
		Ca^{2+}	5.20	27.45	Exp.	5.19	0.28	1.71	164.84	14.10
		Mn	77.91	29.41	Gaus.	80.48	181.55	348.17	494.94	34.27
		Zn	2.92	72.40	Exp.	-430.57; 0.002	0.00	3.93	189.67	0.00
	O25	Al^{3+}	0.28	130.83	M. k_m=2.5	0.29	0.04	0.10	123.67	25.58
		Ca^{2+}	5.38	25.92	Exp.	5.39	1.23	0.69	373.31	63.95
		Mn	77.38	29.74	Gaus.	77.84	283.48	247.03	423.07	53.44
		Zn	2.82	68.25	Exp.	-410.97; 0.002	0.00	3.36	195.25	0.00

CV: coefficient of variation; $\hat{\mu} = \beta_0$: estimated mean; $\hat{\varphi}_1$: estimated nugget effect; $\hat{\varphi}_2$: estimated contribution; $\hat{a}$: estimated practical range; $\widehat{RNE}$: estimated relative nugget effect ($\widehat{RNE} = \hat{\varphi}_1/\hat{\varphi}_1 + \hat{\varphi}_2$) (%) for properties that showed a directional trend $\hat{\mu} = \beta_0 + \beta_1 Y_1$; in which $\hat{\beta}_0$ (first value of the mean column), $\hat{\beta}_1$ (second value of the mean column): estimated values of the parameters of the regression model and Y_1 represents the directional trend identified; Exp.: exponential model; Gaus.: Gaussian model; M. km=2.5: Matérn model with smoothness parameter km=2.5; the units of measured soil chemical properties are found in Table 1.

2016-2017 harvest year, in the O50 sample configuration, the Ca^{2+} and V contents had strong (0.25≤ RNE) and moderate (0.25< RNE <0.75) spatial dependence, respectively; in turn, the SB presented moderate spatial dependence (0.25< RNE <0.75) in the O25 sample configuration, and the pH had strong spatial dependence (0.25≤ RNE) in both optimized sample configurations (Table 5).

Comparing the thematic maps of the Ca^{2+}, Mn, MO and Zn contents and that of V generated considering the initial and the O25 configurations, estimated OA values above 85 % were found, which indicates that the maps are similar; in other words, the maps prepared considering both configurations are similar in terms of distribution of the properties contents in the study area (OA ≥85 %) (Figures 6 to 9). Therefore, the O25 sample configuration could also be used to delimit the AZs, similarly to those generated with the initial sample configuration.

According to the estimated values of the Kp and T agreement indices, most of the soil chemical properties presented low or average accuracy, with values between 0.001 and 79.18 % (low accuracy if Kp and T <67 %, average accuracy if 67 % ≤ Kp and T <80 %); whereas the Ca^{2+}, Cu, Mn, MO and Zn contents and pH, SB and V presented high accuracy, with values between 80.01 and 91.56 %, mainly for the T index (high accuracy if Kp and T ≥80 %) with the initial and the O25 configurations. In turn, the Al presented high accuracy when comparing O25 and O50 to the initial configuration (Figures 6 to 9). This shows no relevant differences in the spatial prediction of these properties in the area under study, described by the thematic maps.

Table 5. Descriptive statistics and estimated values of the adjusted geostatistical model parameters, for the soil chemical properties and SMP index with the sample configurations optimized by 50 and 25 %, referring to the 2015-2016 and 2016-2017 harvest years

Year	Method	Property	Descriptive statistics		Estimated of the parameters of the geostatistical model					
			Mean	CV	Models	$\hat{\mu}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{a}$	$\widehat{RNE}$
2015-2016	O50	C	31.95	11.56	Exp.	31.61	6.80	6.50	518.93	51.13
		Ca^{2+}	5.46	27.58	Exp.	5.49	0.00	2.23	111.79	0.00
		Cu	3.82	26.29	Exp.	-6254.8; 0.001	0.43	0.47	504.57	47.38
		Mn	83.57	28.44	Gaus.	85.35	303.88	276.93	326.97	52.32
		SB	7.86	29.69	Gaus.	7.86	5.09	0.26	16.26	95.22
		Zn	5.12	46.99	Exp.	5.33	0.18	5.68	464.94	3.04
	O25	C	31.93	11.31	Exp.	31.83	5.35	7.63	531.75	41.18
		Ca^{2+}	5.48	25.76	M. k_m=2.5	5.48	0.55	1.49	201.81	26.86
		Cu	3.87	24.20	Exp.	-6045.7; 0.001	0.28	0.50	526.79	35.58
		Mn	86.47	28.34	Gaus.	86.59	329.87	267.00	367.19	55.27
		SB	7.87	27.16	M. k_m=2.5	7.87	3.38	1.14	170.88	74.81
		Zn	4.96	44.11	Gaus.	5.03	2.32	2.78	399.51	45.45
2016-2017	O50	Ca^{2+}	4.02	22.16	M. k_m=2.5	4.08	0.08	0.68	203.56	10.77
		H+Al	6.33	22.19	Gaus.	6.35	0.78	1.19	193.37	39.58
		K^+	0.30	39.27	Gaus.	0.30	0.01	0.01	401.61	51.49
		Mg	1.66	45.77	Gaus.	1.77	0.01	0.54	200.87	0.87
		MO	42.34	27.25	Gaus.	43.52	43.85	95.99	260.06	31.36
		P	20.25	53.52	Exp.	19.07	57.34	54.91	455.86	51.09
		pH	4.54	6.97	Gaus.	4.55	0.02	0.09	167.13	15.18
		SB	5.97	24.39	M. k_m=2.5	6.15	0.15	1.82	22.74	7.60
		SMP	5.72	4.75	Gaus.	5.71	0.04	0.04	175.61	49.98
		V	48.39	21.19	M. k_m=2.5	48.96	25.40	74.84	202.54	25.33
	O25	Ca^{2+}	4.07	22.78	Gaus.	4.10	0.41	0.42	236.38	49.78
		H+Al	6.32	21.20	M. k_m=2.5	6.28	0.81	1.03	311.34	44.19
		K	0.29	39.27	Gaus.	0.29	0.01	0.01	132.08	43.55
		Mg	1.69	45.04	M. k_m=2.5	1.77	0.07	0.50	255.07	12.17
		MO	42.43	25.46	Gaus.	43.10	51.05	62.05	347.40	43.97
		P	19.18	56.88	Gaus.	19.35	36.38	78.00	200.10	31.80
		pH	4.53	7.21	Exp.	4.54	0.01	0.10	201.97	7.88
		SB	6.06	24.61	Gaus.	6.17	0.83	1.34	256.04	38.22
		SMP	5.72	4.64	Exp.	5.73	0.03	0.04	371.75	44.33
		V	48.72	20.77	Exp.	49.21	22.56	77.65	270.87	22.51

CV: coefficient of variation; $\hat{\mu} = \beta_0$: estimated mean, $\hat{\varphi}_1$: estimated nugget effect; $\hat{\varphi}_2$: estimated contribution; $\hat{a}$: estimated practical range; $\widehat{RNE}$: estimated relative nugget effect ($\widehat{RNE} = \hat{\varphi}_1/\hat{\varphi}_1 + \hat{\varphi}_2$) (%); for attributes that showed a directional trend $\hat{\mu} = \beta_0 + \beta_1 Y_1$, where $\hat{\beta}_0$ (first value of the mean column), $\hat{\beta}_1$ (second value of the mean column): estimated values of the parameters of the regression model and Y_1 represents the directional trend identified; Exp.: Exponential model; Gaus.: Gaussian model; M. km=2.5: Matérn model with smoothness parameter km=2.5; the units of measure of soil chemical properties are found in Table 1.

DISCUSSION

Clustering of the agricultural area

When comparing the clustering methods, it was observed that the Fanny method was not selected in any of the criteria, and requires more computational time compared to other partitioned methods (Rajkumar et al., 2019). Among the hierarchical methods, the only one that was chosen for a given harvest year was Ward's, which is in line with the

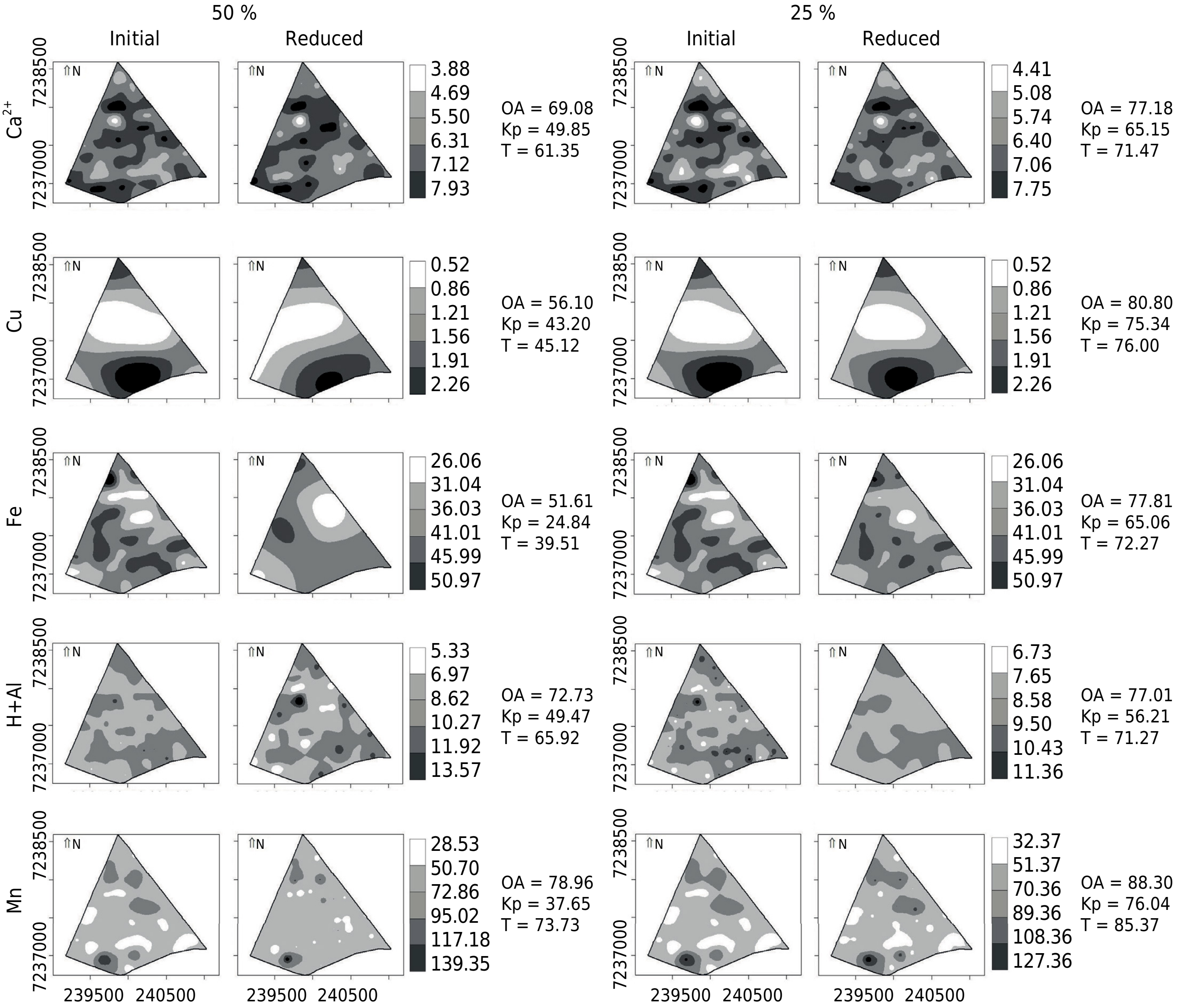

Figure 6. Thematic maps of the soil chemical properties considering the initial and optimized sample configurations and estimated values of the OA, Kp, and T similarity measures for the 2013-2014 harvest year.

result obtained by Ossani et al. (2020); by analyzing clusterings in a coffee plantation, the authors verified that regardless of the distance considered, this method stood out among the other hierarchical methods. Dobermann et al. (2003) showed that the Ward method is one of the algorithms that provides the best results, analyzing various configurations of the input data. In addition, Freitas et al. (2014) and Santos et al. (2015) showed that the Ward method was also efficient in verifying similarities or differences based on the chemical and physical properties of the soil; in addition, integrated with the characterization of the soil properties' spatial variability, this method was effective in defining MZs.

Comparing the K-means and Fuzzy C-means methods to analyze the performance of various segmentation techniques for color images, Jipkate and Gohokar (2012) concluded that K-means clustering produces better results and computational times. The K-means method was also efficient for delimiting MZs from the interpolated variability maps and for coffee, based on determinations carried out with a chlorophyll sensor and by leaf analysis (Rodrigues Jr et al., 2011; Alves et al., 2013).

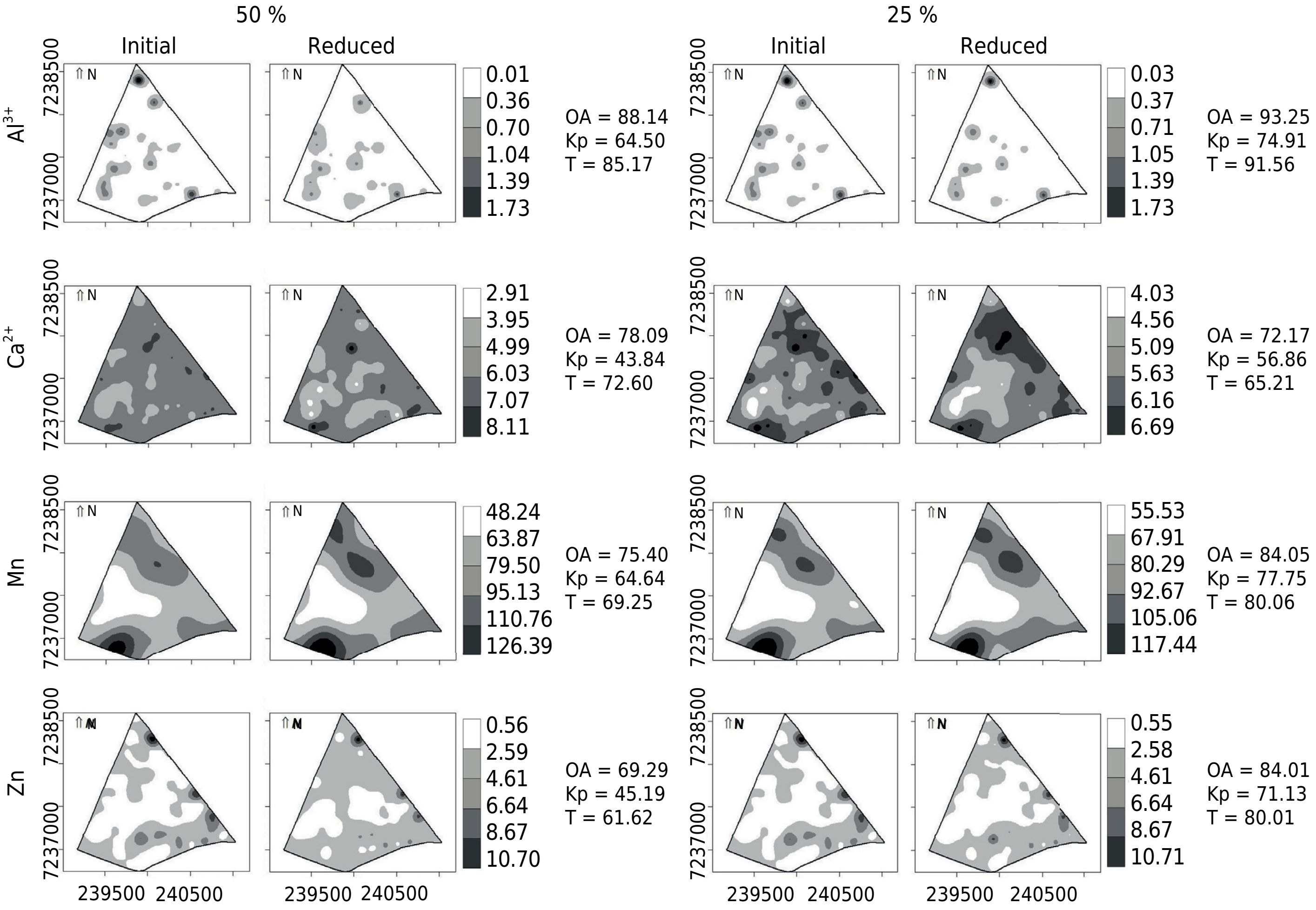

Figure 7. Thematic maps of the soil chemical properties considering the initial and optimized sample configurations and estimated values of the OA, Kp, and T similarity measures for the 2014-2015 harvest year.

The study only considered chemical properties for the generation of zones. Ortega and Santibáñez (2007) also used soil chemical properties to evaluate three zoning methods and quantitatively determine the relationships between the methods evaluated. The same result regarding the generation of two or three zones was also obtained by Barbosa et al. (2019) and by Breunig et al. (2020) when analyzing different numbers of AZs for various harvest years in grain production areas. In the definition of AZs, obtaining a small number of zones renders the application of localized management practices more economically viable, mainly due to greater simplicity in the subdivision of the production field (Carvalho et al., 2016). Then, as the number of AZs increases, they end up becoming increasingly irregular, which leads to difficulties managing them due to technical and economic limitations, as small zones can become unmanageable.

Optimized sample configuration

Similarity in the descriptive statistics was observed between the O50 and O25 sample configurations and the initial sample configuration; this result was also found by Maltauro et al. (2019; 2021) and by Dal'Canton et al. (2021), obtaining similar sample reductions even when working with different methodologies to obtain a sample reduction; in addition, these research studies were developed in the same agricultural area, considering different soil chemical properties. Thus, it can be stated that the reduced sample configurations are representative due to the similarity obtained.

The few cases in the optimized configurations that showed weak spatial dependence and a pure nugget effect might be influenced by the sample size reduction, as low concentrations

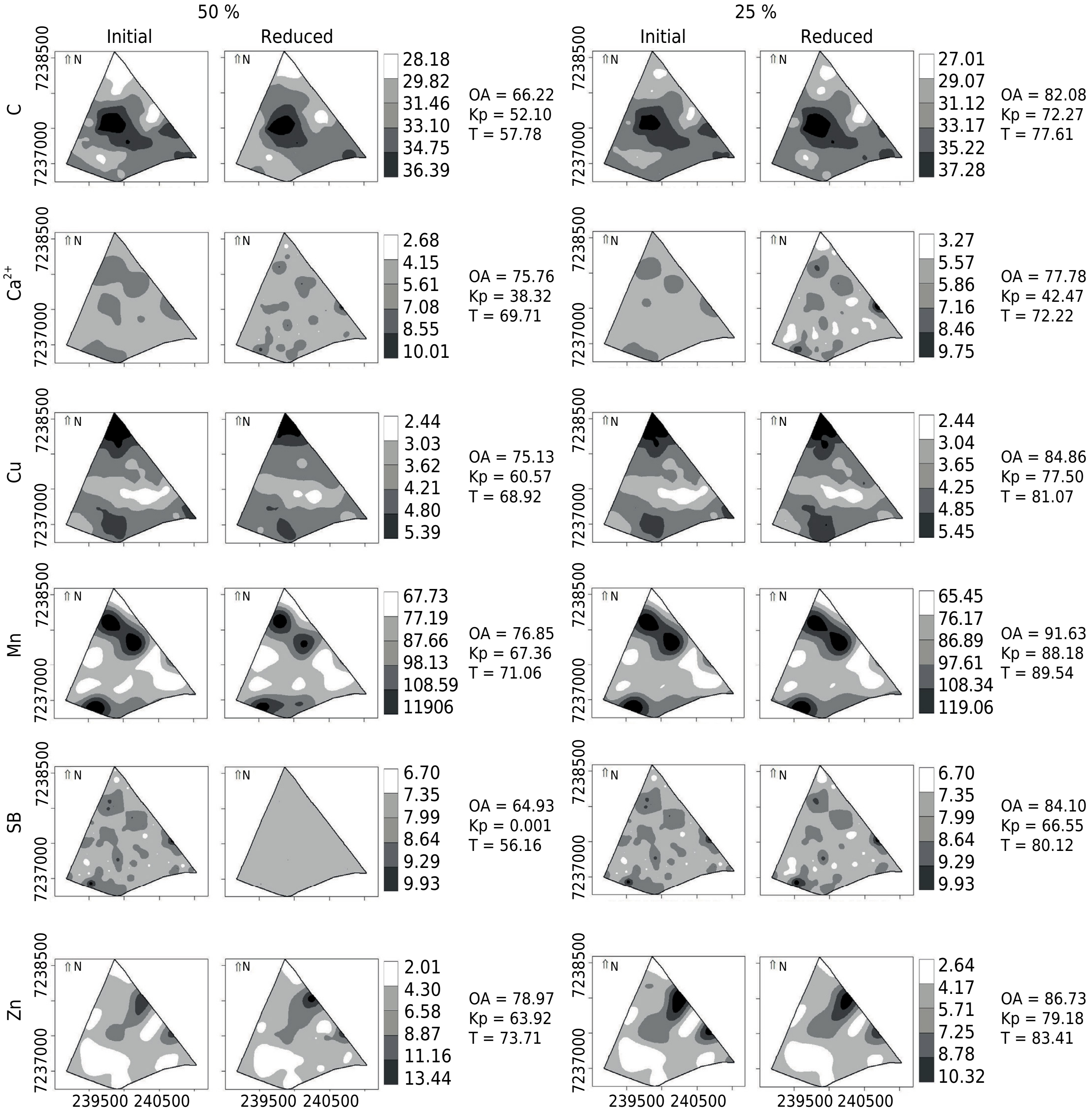

Figure 8. Thematic maps of the soil chemical properties considering the initial and optimized sample configurations and estimated values of the OA, Kp, and T similarity measures for the 2015-2016 harvest year.

in samples might lead to an overestimation of the nugget effect (Hofmann et al., 2010). Most of the optimized configurations remained spatially dependent. One of the factors that may have contributed to this is the maintenance of the close pairs of points that lattice plus close pairs sampling provides an optimized sampling. In fact, these pairs of close points make it possible to more accurately estimate the nugget effect and minimize sampling error on a small scale (Hofmann et al., 2010).

Disregarding the soil chemical properties that presented weak spatial dependence and/or a pure nugget effect, the spatial dependence radium of the initial and optimized

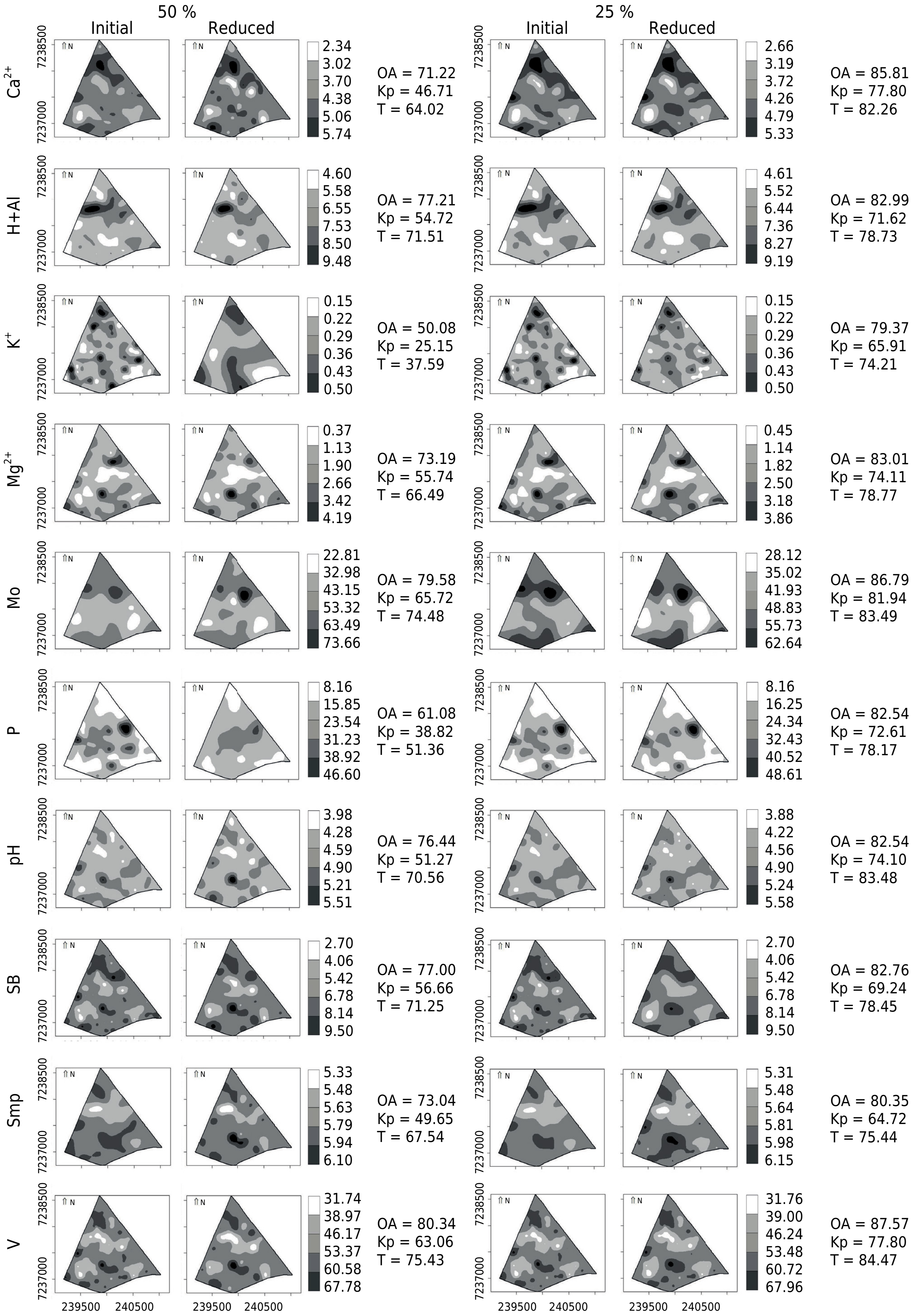

Figure 9. Thematic maps of the soil chemical properties and SMP index considering the initial and optimized sample configurations and estimated values of the OA, Kp, and T similarity measures for the 2016-2017 harvest year.

sample configurations was compared. A smaller variation in the practical ranges (upwards or downwards) considering the initial and O25 sample configurations (Tables 4 and 5) was observed, with this variation in the practical ranges influenced by the model chosen (Diggle and Ribeiro Jr, 2007). Such being the case, the O25 sample configuration presented the best estimate for the spatial dependence radium values when compared to the initial sample configuration (Tables 4 and 5).

Furthermore, for most of the soil chemical properties, when compared to the initial sample configurations, O25 presented higher values for the accuracy indices (Figures 6 to 9). This fact was already expected, as this optimized sampling contains more sampling points compared to the one with a 50 % sample reduction. In the practice, there is a certain similarity between the thematic maps generated with the initial and optimized sample configurations; therefore, the initial sample configuration or the optimized sample configuration could be used for the localized application of inputs. Furthermore, in both optimized sample configurations, it was possible to observe that the spatial variability pattern is maintained in most classes of the thematic map of the soil chemical properties, a fact also observed by Maltauro et al. (2019) and by Dal' Canton et al. (2021), working with sample reductions in an area of grain plantations.

As the soil chemical properties (macronutrients and micronutrients) are important for the development and growth of plants, it is necessary to know their availability in the soil; however, macronutrients (Ca, K, Mg, and P) are elements that plants need in high amounts, while micronutrients (Cu, Fe, Mn, and Zn) are the ones they need in smaller amounts and are absorbed in the form of cations (Mendes, 2007; Oliveira, 2007). Carbon plays numerous roles in the formation of biomass and plant metabolism, being necessary for plant growth as well as OM, which makes the soil richer in nutrients (Lopes, 1998; Ferreira et al., 2014; Assad et al., 2019). As for V, dystrophic soils (V <50 %) apparently have a lower ability to yield nutrients to plants when compared to eutrophic soils (V ≥50 %) (Mendes, 2007).

As for the macronutrients, it is estimated that the P utilization rate by the plant is from 20 to 40 %, with 80 to 95 % being fixed to the soil (Oliveira, 2007). It is absorbed in anionic forms, presenting a strong covalent bond with the O atom, maintained even after incorporation into plant tissues (Mendes, 2007). As for the K attribute, it is estimated that its utilization rate by the plant varies from 50 to 70 %, with K losses occurring in the soil due to leaching and water erosion (Oliveira, 2007). It is absorbed by plants in the ionic form of K^+, and absorption depends on the diffusion of the element through the soil solution and mass flow (Villar, 2007). In the same way as P, during assimilation, its redox state does not change, remaining in the same ionic form in which it was absorbed (Mendes, 2007). Calcium and Mg are absorbed by plants as Ca^{2+} and Mg^{2+} and are found at high levels in the soil solution, as root interception attends to a considerable absorption percentage, while the mass flow supplies the rest. The Cu addresses the same process. All three processes supply the Fe and Zn: root interception, mass flow and diffusion (Lopes, 1998; Villar, 2007).

Therefore, by interpreting the chemical properties available in the soil within each AZ described in this paper, it was possible to observe that almost all the soil chemical properties presented high or very high average values for the soil in the state of Paraná. However, the Al and pH were classified as low or very low (Oliveira 2007; Pavinato et al., 2017). The Al content in the soil exerts a beneficial effect on the plant when it is supplied in low concentrations (Mendes, 2007). Soil pH is one of the most important factors influencing the availability of nutrients to plants; however, low pH values in the soil indicate greater soil acidity, which affects plant growth (Lopes, 1998; Oliveira, 2007). One solution to reduce soil acidity is to apply lime in the study area (Lopes, 1998). Therefore, the SMP index is used to correct the soil with liming recommendations: the method consists in adding a volume of buffer solution to the soil sample, with the pH reading in the suspension of the sample representing the SMP index (Shoemaker et al., 1961; Lopes, 1998; Villar, 2007). For most soil chemical properties, these results agree with those obtained for the original sample configurations.

Therefore, at these sampling points, localized application of inputs can be carried out according to the need for each soil chemical property. As the amounts of macronutrients and micronutrients contained in the soil and their availability depend on several factors, and mainly on the interaction between them, it is known that the nutrients present in the soil are not always available (or easy to be absorbed - due to strong chemical bonds between nutrients and soil) to the plant. Thus, the fertilizer is in a format of easy availability of nutrients to the plant. In this way, fertilization is also carried out in a minimal amount (for nutrients that are in excess), only applying the amounts that the plant needs to absorb to develop (Santos and Silva, 2010). The methodology presented in this study can only be used to redefine a sample configuration that already exists in the study area and, therefore, cannot be used for new samples.

CONCLUSION

For all the harvest years, the clustering methods were efficient for defining the application zones (AZs) and, for the 2013-2014, 2015-2016 and 2016-2017 harvest years, the best number of clusters for all the clustering methods was $k_c = 2$. For the 2014-2015 harvest year and for most of the clustering methods, the ideal number of clusters was $k_c = 3$. Considering the evaluation criteria, K-means and Ward were the best clustering methods. Therefore, from a practical point of view, it is concluded that the AZs allow for localized application of inputs in the agricultural area.

Optimized sample configurations can be obtained by 50 and 25 % with the Genetic Algorithm (GA). However, among the sample configurations, when compared to the initial sample configuration, the O25 optimized sample configurations presented the best estimates for the spatial dependence radius values and the highest values for the accuracy indices, indicating greater similarity between the thematic maps of these sample configurations. The AZs effectively obtained an optimized sample configuration with 75 % of the sampling points of the agricultural area.

Therefore, this study showed that the reduced sample configurations allowed reducing the number of soil samples required and, consequently, the costs inherent to carrying out laboratory analyses, in addition to achieving efficiency in the analysis of the special variability of properties and yields.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the funding from the Coordination for the Improvement of Higher Education Personnel (CAPES), Financing Code 001, and the National Council for Scientific and Technological Development (CNPq). The authors would also like to thank the Spatial Statistics Laboratory - LEE/UNIOESTE - Cascavel, PR, Brazil.

AUTHOR CONTRIBUTIONS

Conceptualization: Luciana Pagliosa Carvalho Guedes (equal), Miguel Angel Uribe-Opazo (equal) and Tamara Cantú Maltauro (equal).

Data curation: Letícia Ellen Dal Canton (equal) and Tamara Cantú Maltauro (equal).

Formal analysis: Letícia Ellen Dal Canton (equal) and Tamara Cantú Maltauro (equal).

Methodology: Letícia Ellen Dal Canton (equal), Luciana Pagliosa Carvalho Guedes (equal), Miguel Angel Uribe-Opazo (equal) and Tamara Cantú Maltauro (equal).

Software: Letícia Ellen Dal Canton (equal), Luciana Pagliosa Carvalho Guedes (equal) and Tamara Cantú Maltauro (equal).

Validation: Luciana Pagliosa Carvalho Guedes (equal) and Tamara Cantú Maltauro (equal).

Writing – original draft: Letícia Ellen Dal Canton (equal) and Tamara Cantú Maltauro (equal).

Writing – review & editing: Luciana Pagliosa Carvalho Guedes (equal), Miguel Angel Uribe-Opazo (equal) and Tamara Cantú Maltauro (equal).

REFERENCES

Aikes Jr J, Souza EG, Bazzi CL, Sobjak R. Thematic maps and management zones for precision agriculture: Systematic literature study, protocols, and practical cases. Curitiba: Poncã; 2021.

Alves SMDF, Alcântara GR, Reis EFD, Queiroz DMD, Valente DSM. Definição de zonas de manejo a partir de mapas de condutividade elétrica e matéria orgânica. Biosci J. 2013;29:104-14. https://seer.ufu.br/index.php/biosciencejournal/article/view/13687

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

Aparecido LEO, Rolim GS, Richetti J, Souza PS, Johann JA. Köppen, Thornthwaite and Camargo climate classifications for climatic zoning in the State of Paraná, Brazil. Cienc Agrotec. 2016;40:405-17. https://doi.org/10.1590/1413-70542016404003916

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

Assad ED, Martins SC, Cordeiro LAM, Evangelista BA. Sequestro de carbono e mitigação de emissões de gases de efeito estufa pela adoção de sistemas integrados. In: Almeida RG, Bungenstab DJ, Ferreira AD, Balbino LC, Laura VA, editors. ILPF: Inovação com integração de lavoura, pecuária e floresta. Brasília, DF: Embrapa; 2019. p. 153-67.

Barbosa DP, Bottega EL, Valente DSM, Santos NT, Guimarães WD, Ferreira MDP. Influence geometric anisotropy in management zones delineation. Rev Cienc Agron. 2019;50:543-51. https://doi.org/10.5935/1806-6690.20190064

Bezdek JC. Pattern recognition with fuzzy objective function algorithms. Boston: Springer; 1981.

Bottega EL, Queiroz DM, Pinto FAC, Souza CMA, Valente DSM. Precision agriculture applied to soybean: Part I - Delineation of management zones. Aust J Crop Sci. 2017;11:573-9. https://doi.org/10.21475/ajcs.17.11.05.p381

Branke J, Deb K, Miettinen K, Slowiński R. Multiobjective optimization: Interactive and evolutionary approaches. Berlin Heidelberg: Springer; 2008.

Breunig FM, Galvão LS, Dalagnol R, Dauve CE, Parraga A, Santi AL, Flora DPD, Chen S. Delineation of management zones in agricultural fields using cover–crop biomass estimates from PlanetScope data. Int J Appl Earth Obs Geoinf. 2020;85:102004. https://doi.org/10.1016/j.jag.2019.102004

Callegari-Jacques SM. Bioestatística: Princípios e aplicações. Porto Alegre: Artmed; 2003.

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

Carvalho PSMD, Franco LB, Silva SDA, Sodré GA, Queiroz DMD, Lima JSDS. Cacao crop management zones determination based on soil properties and crop yield. Rev Bras Cienc Solo. 2016;40:e0150520. https://doi.org/10.1590/18069657rbcs20150520

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

Cressie NAC. Statistics for Spatial Data. rev. ed. New York: John Wiley & Sons; 2015.

Dal' Canton LE, Guedes LPC, Uribe-Opazo MA. Reduction of sample size in the soil physical-chemical attributes using the multivariate effective sample size. J Agr Stud. 2021;9:357-76. https://doi.org/10.5296/jas.v9i1.17473

Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979;1:224-7. https://doi.org/10.1109/TPAMI.1979.4766909

Deb K, Kalyanmoy D. Multi-objective optimization using evolutionary algorithms. New York: John Wiley & Sons; 2001.

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

Dobermann A, Ping JL, Adamchuk VI, Simbahan GC, Ferguson RB. Classification of crop yield variability in irrigated production fields. Agron J. 2003;95:1105-20. https://doi.org/10.2134/agronj2003.1105

Dunn JC. Well-separated clusters and optimal fuzzy partitions. J Cybern. 1974;4:95-104. https://doi.org/10.1080/01969727408546059

Faraco MA, Uribe-Opazo MA, Silva EAA, Johann JA, Borssoi J. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. Rev Bras Cienc Solo. 2008;32:463-76. https://doi.org/10.1590/S0100-06832008000200001

Ferreira JT, Ferreira E, Silva W, Rocha I. Atributos químicos e físicos do solo sob diferentes manejos na microrregião serrana dos quilombos - Alagoas. Rev Agr Acad. 2014;1:89-101.

Freitas L, Casagrande JC, Oliveira IA, Souza Jr PR, Campos MC. Análises multivariadas de atributos químicos do solo para caracterização de ambientes. Rev Agro@mbiente On-line. 2014;8:155-64. https://doi.org/10.18227/1982-8470ragro.v8i2.1684

Galambošová J, Rataj V, Prokeinová R, Prešinská J. Determining the management zones with hierarchic and non-hierarchic clustering methods. Res Agr Eng. 2014;60:44-51. https://doi.org/10.17221/34/2013-RAE

Gavioli A, Souza EG, Bazzi CL, Schenatto K, Betzek NM. Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. Biosyst Eng. 2019;181:86-102. https://doi.org/10.1016/j.biosystemseng.2019.02.019

Gavioli A, Souza EG, Bazzi CL, Guedes LPC, Schenatto K. Optimization of management zone delineation by using spatial principal components. Comput Electron Agric. 2016;127:302-10. https://doi.org/10.1016/j.compag.2016.06.029

Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971;27:857-71. https://doi.org/10.2307/2528823

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

Guedes LPC, Uribe-Opazo MA, Ribeiro Jr PJ, Dalposso GH. Relationship between sample design and geometric anisotropy in the preparation of thematic maps of chemical soil attributes. Eng Agr. 2018;38:260-9. https://doi.org/10.1590/1809-4430-Eng.Agric.v38n2p260-269/2018

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

Halkidi M, Vazirgiannis M, Batistakis Y. Quality scheme assessment in the clustering process. In: European conference on principles of data mining and knowledge discovery. Berlin Heidelberg: Springer; 2000. p. 265-76.

Hofmann T, Darsow A, Schafmeister MT. Importance of the nugget effect in variography on modeling zinc leaching from a contaminated site using simulated annealing. J Hydrol. 2010;389:78-89. https://doi.org/10.1016/j.jhydrol.2010.05.024

Hubert LJ, Levin JR. A general statistical framework for assessing categorical clustering in free recall. Psychol Bull. 1976;83:1072-80. https://doi.org/10.1037/0033-2909.83.6.1072

Jipkate BR, Gohokar VV. A comparative analysis of Fuzzy C-Means clustering and K-Means clustering algorithms. Int J Comput Eng Sci. 2012;2:737-9. https://doi.org/10.14569/IJACSA.2013.040406

Kaufman L, Rousseeuw PJ. Finding groups in data: An introduction to cluster analysis. Hoboken, New Jersey: John Wiley & Sons; 1990.

Krippendorff K. Content analysis: An introduction to its methodology. 2nd ed. California: Sage Publications Ltda; 2013.

Landim PMB, Yamamoto JK. Geoestatística: Conceitos e aplicações. São Paulo: Oficina de Textos; 2013.

Lopes AS. Manual internacional de fertilidade do solo. 2. ed rev amp. Piracicaba: Potafos; 1998.

MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley, Califórnia: University of California Press; 1967. p. 281-97.

Maity A, Sherman M. Testing for spatial isotropy under general designs. J Stat Plan Inference. 2012;142:1081-91. https://doi.org/10.1016/j.jspi.2011.11.013

Maltauro TC, Guedes LPC, Uribe-Opazo MA, Canton LED. A genetic algorithm for resizing and sampling reduction of non-stationary soil chemical attributes optimizing spatial prediction. Span J Agric Res. 2021;19:e0210. https://doi.org/10.5424/sjar/2021194-17877

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

Martarelli NJ, Nagano MS. Socioeconomic class of Brazilian cities for health, education and employment & income IFDM: A clustering data analysis. IEEE Latin America Trans. 2016;14:1513-8. https://doi.org/10.1109/TLA.2016.7459643

McQuitty LL. Similarity analysis by reciprocal pairs for discrete and continuous data. Educ Psychol Meas. 1966;26:825-31. https://doi.org/10.1177/001316446602600402

Mendes AMS. Introdução a fertilidade do solo. In: Curso de manejo e conservação do solo e da água; 2007; Barreiras. Barreiras: MAPA; Superintendência Federal de Agricultura, Pecuária e Abastecimento do Estado da Bahia; Embrapa Semi-Árido; Recife: Embrapa Solos - UEP; 2007. [CD-ROM].

Molin JP. Agricultura de precisão aprimora o gerenciamento. Visão Agrícola. 2006;3:115-8.

Oliveira EF. Treinamento: Fertilidade do solo e nutrição das plantas. Cascavel: Coodetec - Cooperativa Central de Pesquisa Agrícola; 2007.

Oliver MA, Webster R. A geostatistical basis for spatial weighting in multivariate classification. Math Geology. 1989;21:15-35. https://doi.org/10.1007/BF00897238

Ortega RA, Santibanez OA. Determination of management zones in corn (*Zea mays* L.) based on soil fertility. Comput Electron Agric. 2007;58:49-59. https://doi.org/10.1016/j.compag.2006.12.011

Ossani PC, Rossoni DF, Cirillo MA, Borém FM. Unsupervised classification of specialty coffees in Homogeneous sensory attributes through machine learning. Coffee Sci. 2020;15:e151780. https://doi.org/10.25186/cs.v15i.1780

Pantuza Jr G. Uma abordagem multiobjetivo para o problema de sequenciamento e alocação de trabalhadores. Gest Prod. 2016;23:132-45. https://doi.org/10.1590/0104-530X1432-14

Pavinato PS, Pauletti V, Motta ACV, Moreira A. Manual de adubação e calagem para o estado do Paraná. 2. ed. Curitiba: Núcleo Estadual Paraná da Sociedade Brasileira de Ciência do Solo - NEPAR - SBCS; 2017.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: http://www.R-project.org/.

Rajkumar KV, Yesubabu A, Subrahmanyam K. Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset. Int J Electr Comput Eng. 2019;9:2760-70. https://doi.org/10.11591/ijece.v9i4.pp2760-2770

Rodrigues Jr FA, Vieira LB, Queiroz DM, Santos NT. Geração de zonas de manejo para cafeicultura empregando-se sensor SPAD e análise foliar. Rev Bras Eng Agric Ambient. 2011;15:778-87. https://doi.org/10.1590/S1415-43662011000800003

Santos DRD, Silva LSD. Fertilidade do solo e nutrição de plantas. Santa Maria: UFSM, NTE, UAB; 2010.

Santos EODJ, Pinto FB, Barbosa MDA, Gontijo I. Delineamento de zonas de manejo para macronutrientes em lavoura de café conilon consorciada com seringueira. Coffee Sci. 2015;10:309-19. http://www.sbicafe.ufv.br:80/handle/123456789/8132

Santos HG, Jacomine PKT, Anjos LHC, Oliveira VA, Lumbreras JF, Coelho MR, Almeida JA, Araújo Filho JC, Oliveira JB, Cunha TJF. Sistema brasileiro de classificação de solos. 5. ed. rev. ampl. Brasília, DF: Embrapa; 2018.

Shi W, Zeng W. Genetic k-means clustering approach for mapping human vulnerability to chemical hazards in the industrialized city: A case study of Shanghai, China. Int J Environ Res Public Health. 2013;10:2578-95. https://doi.org/10.3390/ijerph10062578

Shoemaker HE, McLean EO, Pratt PF. Buffer methods for determining lime requirement of soils with appreciable amounts of extractable aluminum. Soil Sci Soc Am J. 1961;25:274-7. https://doi.org/10.2136/sssaj1961.03615995002500040014x

Tan PN, Steinbach M, Kumar V. Introdução ao Data Mining: Mineração de dados. Rio de Janeiro: Ciência Moderna; 2009.

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

Uribe-Opazo MA, De Bastiani F, Galea M, Schemmer RC, Assumpção RAB. Influence diagnostics on a reparameterized t-Student spatial linear model. Spat Stat. 2021;41:100481. https://doi.org/10.1016/j.spasta.2020.100481

Villar MLP. Manual de interpretação de análise de plantas e solos e recomendação de adubação. Mato Grosso: Empresa Mato-grossense de Pesquisa, Assistência e Extensão Rural - EMPAER-MT; 2007.

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

Ward Jr JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58:236-44. https://doi.org/10.2307/2282967

Yi J, Du Y, Wang X, He Z, Zhou C. A clustering analysis of eddies' spatial distribution in the South China Sea. Ocean Sci. 2013;9:171-82. https://doi.org/10.5194/os-9-171-2013