

Optimization algorithms for multivariate sampling reduction using spatial-temporal data

Tamara Cantú Maltauro^{(1)*} , **Luciana Pagliosa Carvalho Guedes⁽¹⁾**  and **Miguel Angel Uribe-Opazo⁽¹⁾** 

⁽¹⁾ Universidade Estadual do Oeste do Paraná, Centro de Ciências Exatas e Tecnológicas, Programa de Pós-Graduação em Engenharia Agrícola, Cascavel, Paraná, Brasil.

ABSTRACT: Knowing and defining the spatial and temporal variability of soil chemical properties becomes important for soil management. The definition of application zones in agricultural areas consists of dividing the area into homogeneous subareas, thus allowing the development of localized management. These zones can be defined by cluster methods and one of their advantages is to direct the determination of a future soil sampling, with a possible sample reduction. This study aimed to propose a methodology that integrates multivariate and space-time analyses to obtain an optimized sample configuration, using spatio-temporal data and application zones. First, three spatial multivariate clustering methods were evaluated to obtain the application zone (fuzzy C-means, K-means and Ward). Clusters were obtained by considering the analysis of spatio-temporal dependence through the empirical orthogonal function. Afterward, 50, 60 and 70 % of the sampling points of the initial sampling configuration were selected in each application zone, generating an optimized sample configuration. This choice was made by an optimization process, whose efficiency was evaluated based on spatial prediction, using the sum of the Overall Accuracy Index, of all soil properties. The results indicated the division of the agricultural area into two application zones, considering the K-means method. For most soil properties, when comparing the original and reduced sampling configurations, a similarity was observed in descriptive statistics. Regarding the estimates of the accuracy indexes, considering all the optimized sample configurations, the best estimates were observed comparing the initial sample configurations and the optimized ones in 70 % for most soil chemical properties.

Keywords: application zones, empirical orthogonal function, genetic algorithm, geostatistics.

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

Received: May 21, 2024

Approved: January 24, 2025

How to cite: Maltauro TC, Guedes LPC, Uribe-Opazo MA. Optimization algorithms for multivariate sampling reduction using spatial-temporal data. Rev Bras Cienc Solo. 2025;49:e0240097.

<https://doi.org/10.36783/18069657rbc20240097>

Editors: José Miguel Reichert  and Quirijn de Jong van Lier .

Copyright: This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.



INTRODUCTION

Soils are not evenly distributed on the land surface, and they are more homogeneous in some parts than others (Noetzold et al., 2018). Even those considered homogeneous have some spatial and temporal variability in their chemical, physical and biological properties. Spatial and temporal variability in any crop can be affected by several factors, such as climate, genetics, soil physical and chemical properties, topography, management practices, pests, diseases and the dynamic interaction among all the above-mentioned factors (Ortega and Santibáñez, 2007). Thus, knowing and defining the spatial and temporal variability of the soil chemical and physical properties becomes an important factor for soil management, as these properties have an important influence on crop productivity (Amaro Filho et al., 2007; Souza et al., 2010; Barbosa et al., 2019).

Geostatistical techniques allow the study of the spatial variability of continuous georeferenced variables. These techniques determine the degree of spatial dependence between the sample elements in the region and display the spatial dependence structure of the georeferenced variable throughout the area, enabling the creation of thematic maps (Cressie, 2015; Uribe-Opazo et al., 2021, 2023).

Thus, understanding the spatial distribution of soil chemical properties is important to establish appropriate management practices, especially with the application of precision agriculture (PA) techniques, which allows the optimization of agricultural production and the minimization of possible environmental damage, because the adequate soil management, which includes localized management of soil chemical properties, allows soil fertility maintenance and the production system sustainability (Noetzold et al., 2018).

In PA, the definition of management zones (MZ) or application zones (AZ) in agricultural areas is fundamental for the localized management, since it allows to delimit the area into subareas with similar characteristics (Ikenaga and Inamura, 2008). The difference between AZs and MZs is related to the available variables used for the generation of zones, as well as the intention of zones generation. The MZs use stable variables and these zones are for long-term use, while AZs use unstable variables, and these zones are only used for future fertilizer application.

When samples present spatial dependence these management or application zones are outlined through a study of spatial variability, considering a period of several crop years, they aggregate greater information about the soil, such as chemical and physical properties, pH, as well as about crop development and agricultural yield (productivity, biomass, vegetation indices) (Perez-Quezada et al., 2003; Ikenaga and Inamura, 2008).

In addition, ZAs can portray indicators for future soil samplings, and thus reduce the number of samples that need to be collected to perform soil and crop analysis (Gavioli et al., 2019; Maltauro et al., 2023a). In this context, Maltauro et al. (2023a) used an integration of multivariate statistical techniques (principal components (PC) and clustering analysis), optimization processes and spatial statistics to determine a reduction in the number of samples initially collected in the area, thus defining the application zones. However, the methodology used by these authors was performed for each crop year individually.

The gap that this article seeks to fill is to propose a methodology that incorporates spatio-temporal information in the definition of these AZs, to enable the determination of sample configurations with reduced size. A form of description of an extensive spatio-temporal data set is through Empirical Orthogonal Functions (EOFs), a statistical method that decomposes a data set in terms of orthogonal basis functions determined from the data. A literature review carried out by the authors found several applications in the climatological area (Vilela et al., 2018; Ma et al., 2021; Hannachi et al., 2023; Benestad et al., 2023) and in agricultural data, such as soil moisture and electrical conductivity, linked

to irrigation management (Perry and Niemann, 2007; Gibson and Franz, 2018; Finkenbiner et al., 2019; Zhao et al., 2020). The general idea is that the EOF/PCA technique seeks to find a new set of variables that explains most of the variance observed using linear combinations of the original variables, allowing a reduction of dimensionality (Hannachi et al., 2007). However, EOF considers that the vector “traces” of the multivariate data are spatially indexed, and the samples are taken over time (Wikle et al., 2019). This analysis is considered an exploratory spatio-temporal analysis, which, by reducing the data dimensionality, allows a visualization of spatial and temporal variability trends, revealing the spatial structure in spatio-temporal data. This can be useful for several scientific areas, e.g., atmospheric sciences (Finkenstädt et al., 2007; Hannachi et al., 2007). This dimensionality reduction is made considering the matrix of spatial empirical cross-covariance and is mathematically based on the analysis of main components (Wikle et al., 2019). However, the use of EOFs in the study of the spatio-temporal variability of soil chemical properties, in the generation of Application Zones, and in the definition of spatial sampling was not found in the literature.

This study proposed a methodology that integrates multivariate and space-time analyses to obtain a reduced sample configuration, considering spatio-temporal agricultural data and application zones, using optimization processes, cluster analysis and empirical orthogonal functions. Sample reduction is extremely important, as it reduces the costs of chemical soil analysis. However, this reduction will be more efficient when associated with optimization and statistical techniques that maintain the efficiency in analyzing the spatial variability of soil properties.

MATERIALS AND METHODS

Area under study, soil data and geostatistical analysis

The agricultural area used is a grain production area, located at Agassiz Farm in Cascavel, PR, latitude 24.95° S, longitude 53.37° W and an average elevation of 650 m, with 167.35 ha. The soil in the area is classified as Latossolo Vermelho Distroférrico típico, according to the Brazilian Soil Classification System (Santos et al., 2018), which corresponds to an Oxisol, according to the Soil Taxonomy (USDA Soil Survey Staff, 1999), with a clayey texture. Climate of the region is classified as mesothermal temperate and superhumid, type Cfa (Köppen classification system) (Aparecido et al., 2016).

The area has 102 sampling points, arranged by lattice plus close pairs sampling (Chipeta et al., 2017; Maltauro et al., 2021, 2023a). This sampling design has a minimum distance between the points of the regular grid equal to 141 m, and in some places, randomly chosen, the sampling was performed with smaller distances between pairs of points (75 and 50 m) (Figure 1a). The 141 m sampling grid was defined by studies initially carried out in the agricultural area and considered adequate. However, some points at 50 and 75 m were considered to estimate the nugget effect better. The combination of these distances followed the combination of samples from different sampling grids (Uribe-Opazo et al., 2007; Guedes et al., 2011). All samples were localized and georeferenced through a Global Navigation Satellite System (GNSS) receiver (GeoExplorer, Trimble Navigation Limited, Sunnyvale, CA, USA), in a Datum WGS84 coordinate reference system, Universal Transverse Mercator (UTM) projection.

Soil chemical properties that showed spatial dependence were used: Carbon (C) [g dm^{-3}], Calcium (Ca^{2+}) [$\text{cmol}_c \text{ dm}^{-3}$], Manganese (Mn) [mg dm^{-3}] and Zinc (Zn) [mg dm^{-3}]. For the generation of AZs (Figure 1b), since the properties used are not considered stable, presenting variations over the years-harvest, it was decided to work with four crop years, being 2012-2013, 2013-2014, 2014-2015 and 2015-2016 (in each crop year, described by “year1-year2”, where soybean was planted in October of the first year and harvested in February of the second year).

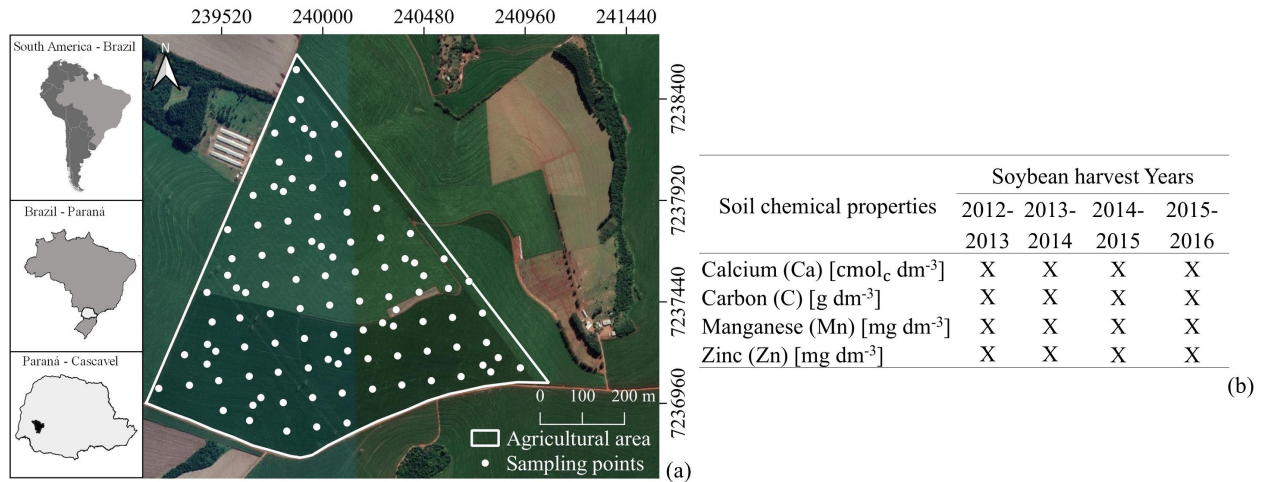


Figure 1. Agricultural area with the location of the sampling points (a) and the soil chemical properties used in the research (b).

Considering all the harvest years and soil chemical properties, the descriptive and geostatistical analyses of each soil chemical property were initially performed (Figure 2). The directional trend was evaluated by Pearson's linear correlation coefficient (Callegari-Jacques, 2003). While anisotropy was evaluated through the analysis of directional semi-variograms (Guedes et al., 2018) and the Maity and Sherman test (non-parametric) (Maity and Sherman, 2012), considering 5 % significance. Parameters of geostatistical models were estimated: gaussian, exponential, and Matérn family with a smoothing parameter $k = 1.5$; 2 and 2.5 by means of the maximum likelihood method (Uribe-Opazo et al., 2012), using the criteria of cross-validation and Akaike (AIC) for the choice of the best adjusted model (Faraco et al., 2008). Spatial dependence was evaluated by classification of relative nugget effect (RNE) (Cambardella et al., 1994). Subsequently, considering the spatial prediction of each property, in non-sampled locations in the agricultural area, using ordinary kriging, thematic maps of each soil chemical property were created (Cressie, 2015).

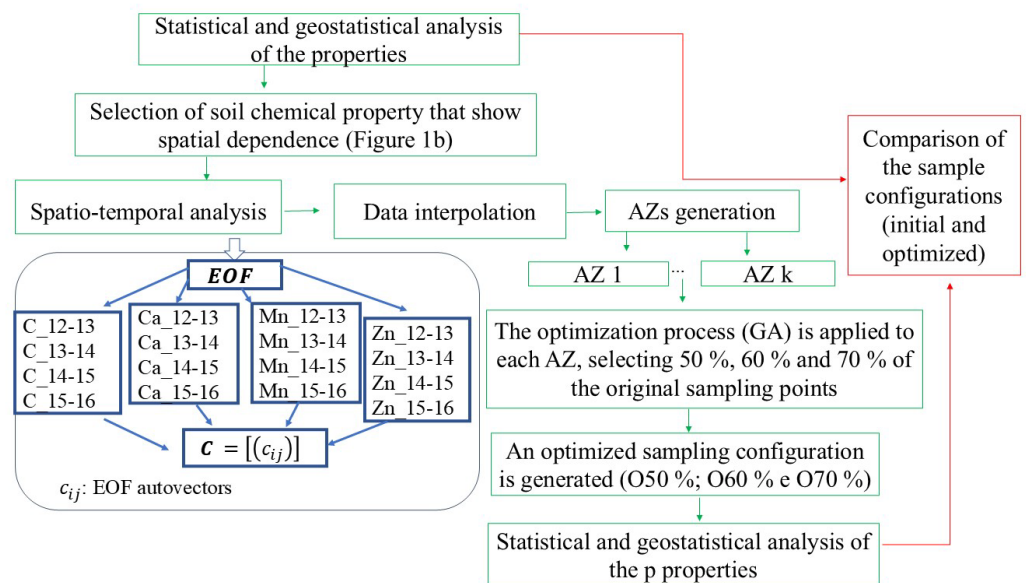


Figure 2. Methodology to obtain the optimized sample configuration.

Spatio-temporal analysis

In each crop year, considering the selected soil chemical properties, the aim was to analyze the spatial-temporal dependence structure, using the Empirical Orthogonal Function (EOF), which has a relationship with the principal component analysis (PCA) (Cressie and Wikle, 2011). In addition to being calculated from the decomposition of the empirical spatial covariance matrix with temporal lag equal to 0 (lag-0), the EOFs can be obtained through a singular value decomposition (SVD), providing computational benefits in some cases. To see the equivalence, we first show how to calculate the EOFs based on empirical covariance.

Be \mathbf{Z} the spatio-temporal, $T \times m$ dimension data matrix, whose j -th line is given by the vector $\mathbf{Z}_{t_j} = (Z(\mathbf{s}_1; t_j), \dots, Z(\mathbf{s}_m; t_j))$, in which $Z(\mathbf{s}_i; t_j)$, the value of the georeferenced variable, is observed in the \mathbf{s} location and in the j -th time (t_j), with $i=1, \dots, m$ and $j=1, \dots, T$; in which m is the number of locations observed and T is the number of moments observed in time. The matrix \mathbf{Z} must be a spatio-temporal matrix "no trend" and in scale, obtained by the transformation given by equation 1 (Wikle et al., 2019):

$$\tilde{\mathbf{Z}} \equiv \frac{1}{\sqrt{T-1}} (\mathbf{Z} - \mathbf{1}_T \hat{\boldsymbol{\mu}}'_{\mathbf{Z}, \mathbf{s}}) \quad \text{Eq. 1}$$

in which: $\mathbf{1}_T$ is a T -dimensional vector of ones and $\hat{\boldsymbol{\mu}}'_{\mathbf{Z}, \mathbf{s}}$ is the vector of averages of the georeferenced variable \mathbf{Z} ($m \times 1$) in each vector location $\mathbf{s} = (\mathbf{s}_1, \dots, \mathbf{s}_m)$, considering all times. Temporal lag consists of the time interval, and the temporal lag equal to zero represents each moment analyzed. Thus, considering temporal lag equal to zero, the empirical spatial covariance matrix of \mathbf{Z} , with dimension $m \times m$ and given by $\mathbf{C}_Z^{(0)} = \tilde{\mathbf{Z}}' \tilde{\mathbf{Z}}$, can be decomposed according to equation 2.

$$\mathbf{C}_Z^{(0)} = \boldsymbol{\Psi} \boldsymbol{\Lambda} \boldsymbol{\Psi}' \quad \text{Eq. 2}$$

in which: $\boldsymbol{\Lambda} \equiv \text{diag}(\lambda_1, \dots, \lambda_m)$ is a diagonal matrix, $m \times m$, of eigenvalues of $\mathbf{C}_Z^{(0)} \mid \lambda_k > 0$, with $k = 1, \dots, m$, ordinated incrementally, and $\boldsymbol{\Psi} \equiv (\psi_1, \dots, \psi_m)'$ is a matrix $m \times m$ of eigenvectors of $\mathbf{C}_Z^{(0)}$, indexed spatially and corresponding to the eigenvalues of $\boldsymbol{\Lambda}$, whose k -th eigenvector, $\boldsymbol{\Psi}_k = (\psi_k(\mathbf{s}_1), \dots, \psi_k(\mathbf{s}_m))'$, is called k -th EOF, for $k = 1, \dots, m$. (Hannachi et al., 2007). Likely PCA, the EOFs have two good properties, namely: (A) the EOFs make a discrete orthonormal basis (that is, $\boldsymbol{\Psi}' \boldsymbol{\Psi} = \boldsymbol{\Psi} \boldsymbol{\Psi}' = \mathbf{I}$), and (b) the first EOFs contain the majority of the data information, considering the total of spatial empirical variance with zero time lag, which corresponds to the sum of the diagonal of $\mathbf{C}_Z^{(0)}$. In addition, each EOF can be represented graphically through spatial maps, once each value of the same is associated with a known spatial location. These maps allow to obtain some understanding on important spatial standards of variability in a sequence of spatio-temporal data.

The k -th column of $\mathbf{A} = (\sqrt{T-1}) \tilde{\mathbf{Z}} \boldsymbol{\Psi}$ generates the time series of k -th PC obtained in the analysis of EOF, in which $k = 1, \dots, m$, for each element of this time series is then calculated for time t_j , with $j = 1, \dots, T$, given by $a_k(t_j) = \tilde{\mathbf{Z}}_{t_j} \boldsymbol{\Psi}_k$. The time series of the PCs can be normalized by $\mathbf{A}_{\text{norm}} = \mathbf{A} \boldsymbol{\Lambda}^{-1/2}$, which are just the time series of PCs divided by their standard deviation, so that the temporal variance of the normalized time series is equal to one, allowing the time series of each EOF to be plotted on the same scale, letting their relative importance be captured by their corresponding eigenvalues.

On the other hand, considering the SVD of the data matrix without trend and sized $\tilde{\mathbf{Z}}$, $t \times m$, we have the equation 3.

$$\tilde{\mathbf{Z}} = \mathbf{U} \times \mathbf{D} \times \mathbf{V}' \quad \text{Eq. 3}$$

in which: \mathbf{U} is the singular vectors to the left, $T \times T$; \mathbf{D} is a matrix containing singular values in the main diagonal, $T \times m$; and \mathbf{V} is a matrix, $m \times m$, containing the singular values to the right, where \mathbf{U} and \mathbf{V} are orthonormal matrices. Thus, it is possible to see that the EOFs are given by $\Psi = \mathbf{V}$, and $\Lambda = \mathbf{D}'\mathbf{D}$.

In addition, the $\mathbf{A} = (\sqrt{T} - 1)\mathbf{U}\mathbf{D}$ and the first m columns of $(\sqrt{T} - 1)\mathbf{U}$ correspond to the normalized PC time series, denominated by \mathbf{A}_{norm} . Thus, the advantages of the SVD calculation approach are: (1) It is not necessary to calculate the empirical spatial covariance matrix; (2) the normalized PC time series and EOFs are obtained simultaneously; and (3) the procedure still works when $T < m$. The case of $T < m$ can be problematic in the covariance context, as $\mathbf{C}_z^{(0)}$ is not positively defined, although, as shown in Cressie and Wikle (2011), in this case, it is still possible to calculate the time series EOFs and PC. Considering every crop year, the EOF was applied for each soil chemical property to reveal the spatial structure in spatio-temporal data, besides reducing the data dimensionality.

Most of the papers do not explain a method to select the number of EOFs that are really significant. Wikle et al. (2019) suggest that perhaps the simplest explanation for this is to consider the number of EOFs that represent some desired proportion of the total variance, or that one can use the PCA criteria to select the number of EOFs. Thus, it was used the criterion that considers the first EOFs (PCs) that explain more than 70 % of the total variability of the original variables (Furtado, 1996). The selected EOFs (eigenvectors) were distributed in columns forming an array \mathbf{C} . In this matrix \mathbf{C} , a geostatistical model was adjusted, and the data were interpolated using Kriging, with pixels representing areas of 10×10 m. These data generated by interpolation were used as input in the application to divide the study area into AZs (Gavioli et al., 2016; Maltauro et al., 2021).

Generation of the Application Zone

For the generation of AZ, three clusters methods were evaluated, namely: Fuzzy C-means, K-means and Ward (Ward Jr, 1963; Macqueen, 1967; Kaufman and Rousseeuw, 1990). To select the best clustering method, five evaluation criteria were used 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 ideal number of groups for the set of data, it was used, in each cluster method, the scatter plots of the Sum of Square of Errors Method (SSE) versus the number of groups (knee chart), and silhouette scatter versus the number of groups (Tan et al., 2009; Yi et al., 2013).

Reduced sample configuration

After the definition of AZ, obtaining the optimized sample configurations in 50, 60 and 70 % of the initial sample size were considered an optimization problem (Maltauro et al., 2019), aiming to select sampling points within each AZ. A methodology developed by Maltauro et al. (2021, 2023a,b), similar to that developed for the genetic algorithm (GA), was used to obtain the optimized sample configuration, using 1500 crossings among the individuals.

As the aim is to generate an optimized sample configuration for all the crop years and soil chemical properties, a multi-objective optimization was performed using the weighted sum method (WS), which consists of the sum of the objective function of each property, weighted by a weight (Branke et al., 2008; Pantuza Jr., 2016).

The efficiency was evaluated based on spatial prediction, considering the objective function Overall Accuracy (OA), a similarity measure between the initial and optimized sampling configurations of each soil chemical property (Maltauro et al., 2021). Thus, in the WS method, the multi-objective function is given by equation 4.

$$\min F(x_i) = \sum_{k=1}^p [1 - OA_k(x_i)] \times w_m, \quad \text{Eq. 4}$$

in which: x_i is the i -th sample configuration to the problem, with sample sizes analyzed in this study of 50, 60 and 70 % of the sample points; $w_m = 1/p$ and the weight for each function objective $f_k(x_i) = 1 - OA_k(x_i)$ of the k -th soil property, with $k = 1, \dots, p$, in which p is the number of soil properties, such that $w_k \in [0, 1]$, $\sum_{k=1}^p w_k = 1$ and $OA_k(x_i) \in [0, 1]$. Thus, upon minimizing $F(x_i)$, which is a linear combination of the k objective functions, the lowest values of $f_k(x_i)$ will be obtained, which correspond to higher values of $OA_k(x_i)$.

Considering the optimized sample configuration, descriptive and geostatistical data analyses were performed again. Finally, the initial and optimized sample configurations were purchased through metrics that express the similarity of the thematic maps obtained by kriging, namely: OA (Anderson et al., 2001) and Kappa concordance indexes (Kp) and Tau (T) (Krippendorff, 2013) (Figure 2).

Computational resources

The routines of the calculation of spatio-temporal analysis, sample configuration, optimization and other statistical and geostatistical analyses were performed in the R (R Development Core Team, 2023) software, considering the geoR packages (Ribeiro Jr and Diggle, 2016), gstat (Pebesma and Graeler, 2022), and spacetime (Pebesma et al., 2023).

RESULTS AND DISCUSSION

Descriptive statistics and initial geostatistics

There was a variation in the average values of each soil chemical property over the years. The average values of C content were 32.20, 31.28, 29.42, and 32.01 g dm⁻³, the average values of the Ca²⁺ content were 6.50, 6.22, 5.38, and 5.50 cmol_c dm⁻³, while the average values of the Mn content were 74.33, 60.96, 76.54, and 86.41 mg dm⁻³. Finally, the average values of the Zn content were 3.97, 2.93, 2.81 and, 4.97 mg dm⁻³, respectively for the harvest years 2012-2013, 2013-2014, 2014-2015, and 2015-2016 (Table 1).

In all the harvest years studied, soil chemical properties presented heterogeneity of their values (Zn with a coefficient of variation (CV) >30) (Pimentel-Gomes and Garcia, 2002), or homogeneous data (CV ≤30) (Pimentel-Gomes and Garcia, 2002) (Table 1). Also, the contents of Ca²⁺, C, Mn showed average values considered medium, high or very high for soil use (Oliveira, 2007; Pavinato et al., 2017). The mean values of Zn can be classified as low or medium (Pavinato et al., 2017). The soil chemical properties are fundamental for the plants' growth and development, some in greater quantities (Ca), while others require lesser quantities (Mn and Zn), and excess or lack of it can harm productivity (Mendes, 2007; Oliveira, 2007).

As for the directional trend, only the soil chemical property Zn for the harvest-year 2014-2015 showed a moderate linear association of their respective values with the coordinates of the X-axis (Pearson's linear coefficient value greater than 0.30) (Callegari-Jacques, 2003). Regarding spatial dependence, soil chemical properties presented moderate or strong spatial dependence, and this makes thematic maps more accurate than those generated with weak spatial dependence (Cambardella et al., 1994).

Table 1. Descriptive statistics and estimated values of the geostatistical model parameters for the soil chemical properties, referring to each harvest year and considering the initial sampling configuration

Harvest years	Property	Descriptive statistics				Estimation of the attributes by the geostatistical model					
		Mean	CV	Coef X	Coef Y	Model	$\hat{\mu}$	$\hat{\psi}_1$	$\hat{\psi}_2$	\hat{a}	RNE
2012-2013	C (g dm ⁻³)	32.20	11.12	0.23	0.03	M k = 2.5	31.58	5.50	7.15	400.56	43.47
	Ca ²⁺ (cmol _c dm ⁻³)	6.50	24.93	0.13	0.29	Gaus.	6.50	1.900	0.70	377.62	73.18
	Mn (mg dm ⁻³)	74.33	25.72	-0.05	0.36	M k = 2.5	-58280; 0.008	65.66	281.40	403.20	18.92
	Zn (mg dm ⁻³)	3.97	58.76	0.25	0.19	Exp.	3.94	2.31	2.99	268.23	43.62
2013-2014	C (g dm ⁻³)	31.28	12.40	0.07	-0.12	M k = 2.5	31.25	7.50	7.38	112.95	50.38
	Ca ²⁺ (cmol _c dm ⁻³)	6.22	22.46	-0.08	-0.05	Gaus.	6.19	1.08	0.87	179.00	55.27
	Mn (mg dm ⁻³)	60.96	33.69	-0.11	0.12	Gaus.	60.31	171.59	225.55	203.98	43.20
	Zn (mg dm ⁻³)	2.93	90.85	-0.03	0.01	Gaus.	2.89	1.21	5.90	173.05	16.97
2014-2015	C (g dm ⁻³)	29.42	12.67	0.23	-0.12	Exp.	29.53	8.60	5.41	511.34	61.39
	Ca ²⁺ (cmol _c dm ⁻³)	5.38	25.11	0.22	0.03	Exp.	5.40	1.05	0.75	231.56	58.49
	Mn (mg dm ⁻³)	76.54	27.43	0.08	-0.03	Gaus.	77.30	233.70	209.80	453.07	52.69
	Zn (mg dm ⁻³)	2.81	61.61	0.31	0.01	Gaus.	-326.73; 0.001	0.54	2.28	162.73	19.12
2015-2016	C (g dm ⁻³)	32.01	10.58	0.11	-0.23	Exp.	31.80	5.97	5.37	576.28	52.65
	Ca ²⁺ (cmol _c dm ⁻³)	5.50	24.12	-0.01	0.05	Gaus.	5.53	1.29	0.48	284.08	72.78
	Mn (mg dm ⁻³)	86.41	25.66	-0.11	0.09	Gaus.	86.78	268.79	226.14	367.29	54.30
	Zn (mg dm ⁻³)	4.97	40.92	0.21	0.23	Gaus.	5.10	1.59	3.04	367.65	34.30

CV: coefficient of variation; Coef X (Y): Pearson's linear correlation coefficient for each coordinate (X and Y) with each of the soil chemical properties; $\hat{\mu} = \beta_0$: estimated mean; $\hat{\psi}_1$: estimated nugget effect; $\hat{\psi}_2$: estimated contribution; \hat{a} : estimated practical range; RNE : estimated relative nugget effect ($RNE = \hat{\psi}_1 / (\hat{\psi}_1 + \hat{\psi}_2)$ (%); for properties that showed a directional trend $\hat{\mu} = \beta_0 + \beta_1 Y_1$, in which β_0 (first value of the $\hat{\mu}$ column), β_1 (second value of the $\hat{\mu}$ 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 k = 2.5: Matérn model with smoothness parameter k = 2.5.

Spatio-temporal analysis

As for the spatial average of the soil chemical properties, considering all the studied harvest years, it is noticed that there was a variation of the values of these properties in the sampling points of the area under study (Figure 3).

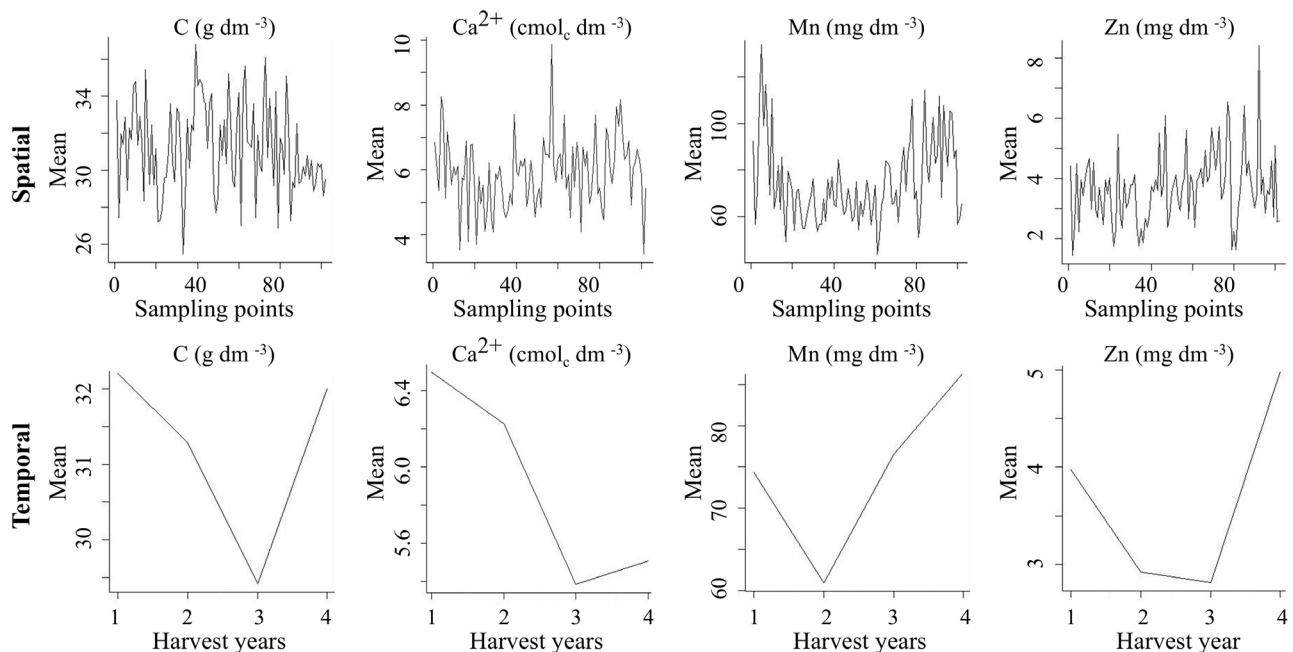


Figure 3. Temporal mean and spatial mean of soil chemical properties C, Ca²⁺, Mn and Zn, considering the four harvest years 2012-2013 (1), 2013-2014 (2), 2014-2015 (3), and 2015-2016 (4).

Temporal means of the soil chemical properties, considering all spatial locations, show the nature of the soil chemical property and the variations of these properties over the four harvest years (Figure 3). A decrease in the concentration of the chemical properties is observed in the harvest year 2014-2015 for most soil chemical properties, and this may have been caused by leaching (Figure 3). This figure shows the temporal means of the concentration of soil chemical properties at each location, allowing for interpretation of property concentrations over the years (Wikle et al., 2019).

For each soil chemical property, the EOFs were obtained, as well as the percentages of variance explained by the EOFs. Some studies presented in the literature use the amount of EOFs that explain a high variance value (greater than 85 %) (Zhao et al., 2020; Ma et al., 2021). However, in this study, the first two EOFs were selected for all the soil chemical properties, because they explain more than 70 % of the total variability of the original variables (Furtado, 1996). The EOF1 and EOF2 explain, respectively, 45.82 and 33.61 % of the total variability for the C content, 45.17 and 35.27 % of the total variability for the Ca^{2+} content, 66.61 and 20.85 % of the total variability for the Mn content, and 43.36 and 34.90 % of the total variability for the Zn content.

One advantage of using EOFs is the reduction of dimension in a spatial or spatial-temporal representation of random effects (Wikle et al., 2019). In addition, the method presents several advantages, such as consistency and mathematical elegance, and is a tool that highlights the interrelations among the variables. For each soil chemical property, the first two EOFs are presented, considering the four harvest years (Equations 5 to 12).

$$\text{EOF1}_C = 0.4474X_1 + 0.5610X_2 + 0.4623X_3 + 0.5209X_4 \quad \text{Eq. 5}$$

$$\text{EOF2}_C = -0.1612X_1 + 0.7069X_2 - 0.6865X_3 - 0.0120X_4 \quad \text{Eq. 6}$$

$$\text{EOF1}_{\text{Ca}^{2+}} = 0.6097X_1 + 0.4131X_2 + 0.4857X_3 + 0.4708X_4 \quad \text{Eq. 7}$$

$$\text{EOF2}_{\text{Ca}^{2+}} = -0.6068X_1 + 0.7892X_2 + 0.0939X_3 - 0.0036X_4 \quad \text{Eq. 8}$$

$$\text{EOF1}_{\text{Mn}} = -0.4845X_1 - 0.4587X_2 - 0.5138X_3 - 0.5391X_4 \quad \text{Eq. 9}$$

$$\text{EOF2}_{\text{Mn}} = 0.1137X_1 + 0.7906X_2 - 0.2292X_3 - 0.5562X_4 \quad \text{Eq. 10}$$

$$\text{EOF1}_{\text{Zn}} = -0.0388X_1 - 0.9253X_2 + 0.0796X_3 - 0.3688X_4 \quad \text{Eq. 11}$$

$$\text{EOF2}_{\text{Zn}} = 0.9608X_1 - 0.1293X_2 + 0.0633X_3 + 0.2370X_4 \quad \text{Eq. 12}$$

in which: X_1 is the harvest year 2012-2013; X_2 is the harvest year 2013-2014; X_3 is the harvest year 2014-2015; and X_4 is the harvest year 2015-2016.

For the C content in the soil, EOF1 represents a weighted mean of all harvest years, with more influence of 2013-2014 and 2015-2016 directly. From the map of this EOF1 (Figure 4), considering all the harvest years, the highest values are distributed more to the North and South of the study area. Thus, the C content in the soil had higher values in these sub-regions, mainly in the 2013-2014 and 2015-2016 harvest years. On the other hand, the lowest values of EOF1 and consequently of the C content in the soil

are distributed in the central region. Meanwhile, the EOF2 represents a contrast mainly between the harvest years 2014-2015 (with inverse influence) and 2013-2014 (with direct influence). Through the EOF2 map (Figure 4), the north and south regions highlight exhibiting the highest values of EOF2, and consequently, when comparing the harvest years, the highest values of C content in the soil in the harvest year 2013-2014 and the lowest values in 2014-2015.

Regarding the Ca^{2+} content in the soil, EOF1 represents a weighted average across all harvest years, with a stronger direct influence from the 2012-2013 period. The spatial distribution of EOF1 (Figure 4) reveals values dispersed throughout the entire study area when considering all harvest years. In addition, it is highlighted that the Ca^{2+} content in the soil obtained the highest values in the South sub-regions in all the harvest years, mainly in 2012-2013. On the other hand, the lowest values of EOF1 and consequently of Ca^{2+} content in the soil are distributed in the North region in all the harvest years, mainly in 2012-2013.

Whereas the EOF2 of Ca^{2+} content in the soil represents a contrast mainly between the harvest years 2012-2013 (with inverse influence) and 2013-2014 (with direct influence). Through the EOF2 map (Figure 4), the north and west regions highlight exhibiting the highest values of EOF2, and consequently, when comparing the harvest years, the highest values of Ca^{2+} content in the soil in the harvest year 2013-2014 and the lowest values in 2014-2015 and 2015-2016.

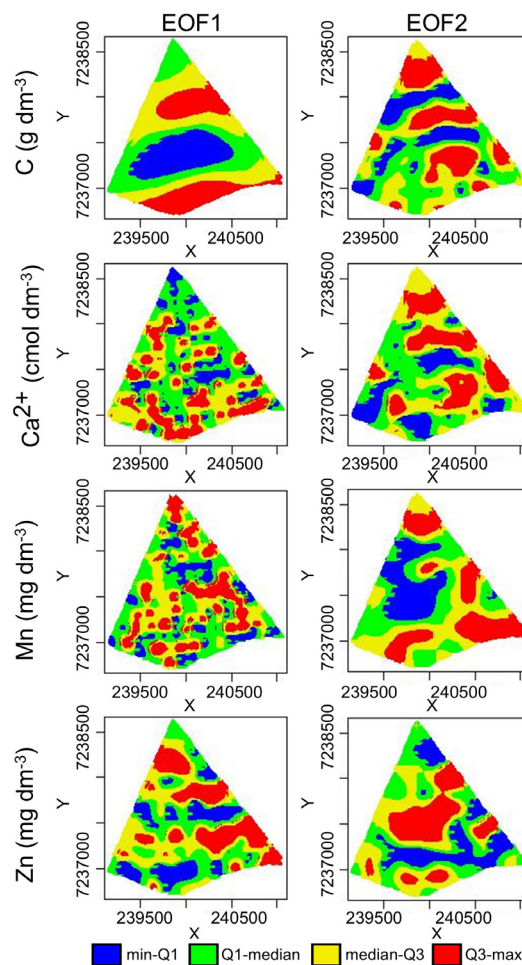


Figure 4. Map of the first two EOFs for each soil chemical property, considering all harvest years 2012-2013, 2013-2014, 2014-2015 and 2015-2016.

For the Mn content in the soil, the EOF1 represents a weighted mean of all the harvest years, with more influence of 2014-2015 and 2015-2016 inversely. Whereas EOF2 represents a contrast mainly between the harvest years 2014-2015 (with direct influence) and 2015-2016 (with inverse influence). The EOF1 map (Figure 4) highlights the region north and west, with higher sub-regions having higher values of EOF1. What represents lower values in these sub-regions of Mn content in the soil, in all the harvest years, mainly in 2014-2015 and 2015-2016.

For the map of EOF2 (Figure 4) of Mn content in the soil, considering all harvest years, the values of EOF2 are more divided in the whole study area, highlighting a higher formation of the lowest EOF2 values in the central and central-west region of the study area and higher values in the north and south regions. When comparing harvest years, the 2013-2014 period exhibited the lowest soil Mn content values in the central and central-west regions, while the highest values were observed in the north and south regions. In contrast, the 2015-2016 harvest year showed an inverse trend across these same regions.

Finally, for the Zn content in the soil, the EOF1 represents a weighted mean of all harvest years, with more influence of 2013-2014 and 2015-2016 indirectly. Through the EOF1 map (Figure 4), it is observed that higher values are in the regions north and southeast of the study area, representing lower values of Zn content in the soil in these same regions. Meanwhile, in EOF2, a higher influence is highlighted directly in the harvest year 2012-2013. The EOF2 map (Figure 4) revealed that the southern region exhibited the lowest EOF2 values, whereas the central region displayed the highest values. This pattern indicates that the Zn content in the soil during the 2012-2013 harvest year was highest in the southern region and lowest in the central region.

In general, the EOF maps indicated distinct variability patterns for each soil chemical property, presenting important patterns of variability. The maps of the first EOFs explain the largest and main variations in soil chemical properties, while the second EOFs highlight secondary characteristics (Wikle et al., 2019). Furthermore, since soil chemical properties are necessary for plant growth and development, it is essential to identify their availability in the soil (Mendes, 2007; Oliveira, 2007). Plants require high amounts of Ca^{2+} , while Mn and Zn levels are required in smaller amounts, being absorbed in the form of cations (Mendes, 2007; Oliveira, 2007). Carbon plays several roles in biomass formation and plant metabolism (Lopes, 1998; Ferreira et al., 2014; Assad et al., 2019). It was observed that two crop years stood out in the EOF analyses, 2013-2014 and 2015-2016. According to Gasparin et al. (2024), when developing agroclimatic regionalization in Paraná, the west of Paraná was characterized in 2015-2016 as rainy, with the highest average precipitation values between 352 and 833 mm and low values of water balance, that is, the counting of water inflows and outflows, with the aim of establishing the variation in storage and, as a consequence, the availability of water in the soil. In contrast, 2013-2014 had a precipitation close to the average of the historical series evaluated, varying between 1073 to 1842 mm, and a positive water balance. Also, these authors identified high values of minimum, average, and maximum temperatures during the crop cycle for both harvest years, varying from 14 to 33 °C for 2013/2014, and from 18 to 32 °C for 2015/2016.

Moreover, comparing these two crop years (2013/2014 and 2015/2016), there was a 4.74 % increase in soybean productivity in Paraná. However, 2015/2016 was considered a year with lower productivity. There was a 6.2 % drop in soybean productivity compared to the 2014/2015 crop year and an 8 % loss in productivity when comparing the estimated potential production with the real production obtained (Hirakuri, 2016). The difference among the crop years, indicated by EOFs, may have been influenced by climatic factors, because the temporal variability of soil chemical properties, soybean productivity, and the interaction between soil and plant is influenced by the climate (Liu et al., 2016; Cordovil et al., 2024).

Generation of the Application Zones

It is important to consider spatial and temporal variability when making management decisions (Luchiari Junior et al., 2011). Furthermore, the generation of AZs with more than one year-crop aggregates a lot of information, considering the variability of properties over time. Since the used agricultural area already considers precision agriculture, it is possible to show that the spatial variability decreases over the years, indicating a stability of these properties over the harvest years.

Considering the scatter plots of the number of groups *versus* the SSE and the silhouette, the best number of groups for all the clustering methods was two groups (Figures 5a and 5b), being that the K-means method was the best cluster method for the generation of AZ, according to the values obtained from the following indices: Davies Bouldin (DB), C, SD, and Variance reduction (VR) (Table 2).

Borge et al. (2022) also used the scatter plots of the number of groups *versus* the SSE and the silhouette, along with the gap statistics method, to find the ideal number of groups, concluding that the optimal number of groups for NO₂, PM10 and PM2.5 was equal to 5, while for O₃ was equal to 4. Syakur et al. (2018) used the scatter plot of the number of groups *versus* the SSE (knee chart) for the selection of the number of groups in the customers' mapping, concluding that the chart shows a sharp drop in the sum of squared errors values with three groups, and this was the ideal number of groups.

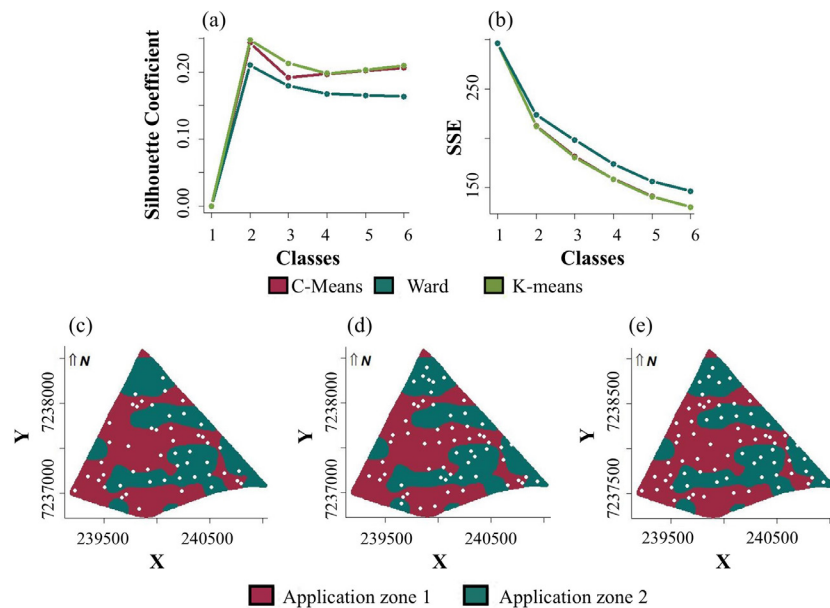


Figure 5. Silhouette graph (a) and Knee graph (SSE - sum of squared errors) (b); and thematic maps with the best number of application zones and clustering method chosen for the sample configuration reduced by 50 % (a), 60 % (b), and 70 % (c).

Table 2. Evaluation measures according to the clustering method used to generate application zones

Indices	C-means	K-means	Ward
D	0.0045	0.0076	0.0084
DB	1.4826	1.4593	1.6362
C	0.2494	0.2468	0.2843
SD	13.9458	13.7557	15.0600
VR	28.2434	28.3243	24.4748

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.

Jipkate and Gohokar (2012) also concluded that the K-means method produces better computing times and results compared to the C-means method. Still, Rodrigues Junior et al. (2011) and Alves et al. (2013) showed that the K-means method was also efficient for the delimitation of the MZs from the interpolated variability maps, and for the coffee farming, based on determinations carried out with a chlorophyll sensor and through foliar analysis. Crispim et al. (2019) showed that the K-means and Ward methods allowed the decrease of the number of municipalities for four groups, composed of municipalities that possibly have similar characteristics within each group and different from the others. Maltauro et al. (2023a,b) also concluded that the K-means method was the best cluster method to generate application zones, considering chemical properties of the three-year-crop soil.

Therefore, with the best number of groups and cluster method selected, the map of ZA was generated. Considering that the highest AZ (AZ1, red color; Figures 5c, 5d and 5e) occupied 102.63 ha (61.33 % of the area), and it was composed of 58 initial sample points (57 % of the total sample points of the initial configuration), while the lowest AZ (AZ2, dark green; Figures 5c, 5d and 5e) occupied 64.72 ha (38.67 % of the area), covering 44 initial sample points (43 % of the total sample points of the initial sample configuration) (Figures 5c, 5d and 5e).

The main advantage for the farmer in this reduction in the number of samples will be the reduction in costs with laboratory analyses and the time involved in collecting soil to carry out these laboratory analyses. This reduction is only possible because, due to spatial dependence, there are similar sampling units within a spatial radius and, therefore, have redundant information. Therefore, for a new sample to represent regions with more similar pairs of points, there will be a need for a smaller number of sample points that summarize the information of the original sample (Dal' Canton et al., 2021).

In addition, the AZs represent most of the spatial trends in the studied harvest years described in figure 5, highlighting AZ2 that exhibits the highest values of all the chemical properties of the soil considered, in the harvest year of 2013-2014 and the lowest values of these properties in the harvest year 2015-2016. Whereas the AZ1 presents the central region with lower C and Mn content values in the soil in the 2013-2014 and 2015-2016 harvest years, the highest values of Zn content in the soil were in 2012-2013.

Optimized sample configuration

Considering the generated AZs, it has been observed that in the optimized sample configuration in 50 % (O50) of the initial sampling points, 51 sampling points were obtained (Figure 5c); in the optimized sampling configuration that removed 40 % (O60) of the initial sampling points, 61 sampling points were obtained (Figure 5d); while the optimized sampling configuration, reducing 30 % (O70) of the initial sampling points, comprised 71 sampling points distributed in the agricultural area (Figure 5e).

Taking into account all the optimized sample configurations (O50, O60, O70), the AZ1 covered 29, 35 and 41 sampling points (57 % of the total sampling points of the reduced sample configuration), while AZ2 included 22, 26 and 30 sampling points (43 % of the total sampling points of the reduced sample configuration), indicating that the same proportion of sample points in each AZ, comparing with the initial sample configuration.

A similarity was observed in the descriptive statistics of the initial and optimized sample configurations (Tables 1, 3, 4 and 5), thus indicating that the optimized sample configurations are representative due to the similarity found (Maltauro et al., 2023a). These results have also been found by Maltauro et al. (2019; 2021; 2023a,b), finding similar sample sizes even when working with different methodologies to obtain a sample reduction. Furthermore, these studies were developed in the same agricultural area, considering different soil chemical properties.

Table 3. Descriptive statistics and estimated values of the geostatistical model parameters for the soil chemical properties, referring to each harvest year and considering the sample configuration optimized by 50 %

Harvest years	Property	Descriptive statistics				Estimation of the properties by the geostatistical model					
		Mean	CV	Coef X	Coef Y	Model	$\hat{\mu}$	$\hat{\phi}_1$	$\hat{\phi}_2$	\hat{a}	RNE
2012-2013	C (g dm ⁻³)	32.05	11.07	0.38	-0.08	M k = 2.5	-807.01; 0.0035	3.07	7.22	357.02	29.82
	Ca ²⁺ (cmol _c dm ⁻³)	6.62	22.95	0.08	0.43	Gaus.	-12850; 0.0018	1.84	0.01	190.28	99.98
	Mn (mg dm ⁻³)	74.55	25.06	-0.01	0.35	Gaus.	-0,0008; 0.0116	115.90	197.00	314.68	37.03
	Zn (mg dm ⁻³)	3.94	60.23	0.27	0.20	Exp.	3.94	0.00	5.57	189.08	0.00
2013-2014	C (g dm ⁻³)	31.34	12.61	0.09	0.05	Gaus.	31.33	15.30	0.01	153.66	99.95
	Ca ²⁺ (cmol _c dm ⁻³)	6.28	24.50	-0.17	0.13	Gaus.	6.21	1.06	1.26	220.68	45.66
	Mn (mg dm ⁻³)	60.35	36.98	-0.10	0.16	Gaus.	59.70	217.00	258.40	305.25	45.66
	Zn (mg dm ⁻³)	2.96	97.55	0.08	0.23	Gaus.	2.85	1.73	6.50	163.40	20.99
2014-2015	C (g dm ⁻³)	28.60	12.20	0.37	-0.07	Gaus.	-781.37; 0.0034	9.27	1.09	555.60	89.43
	Ca ²⁺ (cmol _c dm ⁻³)	5.22	23.95	0.20	0.12	Exp.	5.20	0.99	0.53	186.25	65.32
	Mn (mg dm ⁻³)	74.84	26.02	0.14	-0.11	Exp.	73.89	0.00	357.82	272.50	0.00
	Zn (mg dm ⁻³)	2.71	69.10	0.23	0.04	Exp.	2.82	0.00	3.71	364.51	0.00
2015-2016	C (g dm ⁻³)	31.41	11.58	0.13	-0.21	Exp.	31.56	4.87	8.07	276.03	37.62
	Ca ²⁺ (cmol _c dm ⁻³)	5.47	21.08	-0.16	0.26	Gaus.	5.48	0.88	0.38	350.26	69.94
	Mn (mg dm ⁻³)	85.19	27.34	-0.18	0.20	Gaus.	85.19	310.92	205.48	435.57	60.21
	Zn (mg dm ⁻³)	4.73	45.95	0.18	0.29	Gaus.	4.95	1.22	3.78	354.03	24.47

CV: coefficient of variation; Coef X (Y): Pearson's linear correlation coefficient for each coordinate (X and Y) with each of the soil chemical properties; $\hat{\mu} = \beta_0$: estimated mean; $\hat{\phi}_1$: estimated nugget effect; $\hat{\phi}_2$: estimated contribution; \hat{a} : estimated practical range; RNE : estimated relative nugget effect ($RNE = \hat{\phi}_1 / (\hat{\phi}_1 + \hat{\phi}_2)$ (%); for properties that showed a directional trend $\hat{\mu} = \beta_0 + \beta_1 Y_1$, in which β_0 (first value of the $\hat{\mu}$ column), β_1 (second value of the $\hat{\mu}$ 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 k = 2.5: Matérn model with smoothness parameter k = 2.5.

Considering the sample configuration reduced by 50 %, the C content in the soil showed a directional trend in the X direction (East-West) for the harvest years 2012-2013 and 2014-2015, while the Ca²⁺ and Mn contents in the soil indicated a directional trend in the Y direction (North-South) for the 2012-2013 harvest year (Table 3).

In the sample configuration optimized in 60 %, only the Mn content in the soil showed a directional trend towards Y (North-South) to the harvest year 2012-2013 (Table 4). Whereas in the sample configuration optimized in 70 %, the C content in the soil showed a directional trend in the X direction (East-West), and the chemical property of the soil Mn in the direction Y (North-South) for the harvest year 2012-2013 (Table 5).

Considering spatial dependence, most of the soil chemical properties presented moderate or strong spatial dependence, as in the initial sample configuration, except the C content in the soil, which presented weak spatial dependence for the harvest years 2013-2014 and 2014-2015, considering the sample configuration optimized in 50 %. Meanwhile, in the sample configuration optimized at 60 %, only the Ca content in the soil indicated weak spatial dependence for the harvest years 2014-2015 and 2015-2016 (Tables 4 and 5). However, for all the sample configurations optimized, some soil chemical properties presented pure nugget, namely Zn content in the soil for the harvest years 2012-2013 and 2014-2015, and the Mn content in the soil, in the harvest year 2014-2015 (Tables 3, 4 and 5).

A comparison of the thematic maps for soil chemical properties between the initial and optimized sampling configurations revealed that only Mn content in the 2012-2013 harvest year achieved an estimated overall accuracy (OA) exceeding 85 % under the

70 % optimized sampling configuration. This high OA value (>85 %; Anderson et al., 2001) indicates strong similarity between the maps, suggesting consistent spatial distributions of the studied properties across both sampling designs. Half of the soil chemical properties in all harvest years present high values of OA, above 80 %, and this measure indicates that 80 % of the pixels were equally classified in the same ranges of variation of the soil chemical properties values, when comparing the elaborated thematic maps with the original and optimized sample configurations. However, according to the classification of Anderson et al. (2001), these maps are considered non-similar, as the OA values are lower than 85 % (Table 6; Figures 6, 7, and 8).

For all the soil chemical properties, the spatio-temporal trends observed in the maps generated by the EOFs values of each property (Figure 4) were the same in the maps elaborated considering the initial sample configurations (Figures 6, 7, and 8), highlighting similarities in the distribution of these properties by the study area in both maps.

A similarity is observed in the results obtained by the estimated values of the Kp and T concordance indices. The comparison of the initial sample configuration with the optimized ones showed low or average accuracy, with values between 0.01 and 79 % (low accuracy if Kp; T <67 %; mean accuracy if 67 % ≤ Kp; T ≤80 %) (Krippendorff, 2013), for most soil chemical properties (Table 6; Figure 6, 7, and 8). Only the chemical properties of the soil Mn and Zn, considering respectively the configuration optimized in 70 % and harvest year 2012-2013, and the sample configuration optimized in 60 % and harvest year 2015-2016, presented high accuracy, with Kappa or Tau values higher than 80 % (Krippendorff, 2013) (Table 6; Figures 6, 7, and 8).

Table 4. Descriptive statistics and estimated values of the geostatistical model parameters for the soil chemical properties, referring to each harvest year and considering the sample configuration optimized by 60 %

Harvest year	Property	Descriptive statistics				Estimation of the properties by the geostatistical model					
		Mean	CV	Coef X	Coef Y	Model	$\hat{\mu}$	$\hat{\Phi}_1$	$\hat{\Phi}_2$	\hat{a}	\widehat{RNE}
2012-2013	C (g dm ⁻³)	32.35	12.16	0.12	0.07	Exp.	31.73	2.76	12.56	406.28	18.01
	Ca ²⁺ (cmol _c dm ⁻³)	6.77	28.05	0.03	0.29	Gaus.	6.75	2.45	1.17	351.79	67.74
	Mn (mg dm ⁻³)	76.11	28.40	-0.09	0.36	Gaus.	-0,0009; 0.013	89.99	338.70	295.76	20.99
	Zn (mg dm ⁻³)	4.09	65.54	0.19	0.28	Exp.	3.95	4.82	2.21	418.81	68.54
2013-2014	C (g dm ⁻³)	31.03	13.30	0.13	-0.14	M k = 1.5	30.94	4.54	12.19	159.46	27.14
	Ca ²⁺ (cmol _c dm ⁻³)	6.45	21.97	-0.08	-0.04	Gaus.	6.46	1.42	0.57	302.99	71.41
	Mn (mg dm ⁻³)	61.79	36.64	-0.21	0.05	Exp.	60.69	63.07	415.50	232.10	13.18
	Zn (mg dm ⁻³)	2.97	100.93	0.03	0.09	Gaus.	2.93	1.55	7.56	172.39	17.02
2014-2015	C (g dm ⁻³)	29.03	10.89	0.24	-0.23	M k = 1.5	29.15	6.86	3.32	641.09	67.38
	Ca ²⁺ (cmol _c dm ⁻³)	5.54	22.44	0.27	0.05	Gaus.	5.54	1.34	0.18	335.26	87.94
	Mn (mg dm ⁻³)	78.17	28.98	0.04	-0.01	Exp.	76.80	0.00	479.64	251.12	0.00
	Zn (mg dm ⁻³)	2.70	58.74	0.24	0.13	Gaus.	2.69	1.58	0.99	349.69	61.58
2015-2016	C (g dm ⁻³)	31.43	11.93	0.25	-0.28	Gaus.	31.47	8.90	5.02	437.24	63.96
	Ca ²⁺ (cmol _c dm ⁻³)	5.65	25.98	0.01	0.01	Exp.	5.67	0.00	2.12	118.08	0.00
	Mn (mg dm ⁻³)	87.17	26.83	-0.11	0.07	M K = 1.5	86.42	280.19	255.00	398.76	52.35
	Zn (mg dm ⁻³)	5.02	48.32	0.24	0.22	Gaus.	5.14	2.18	4.15	344.94	34.44

CV: coefficient of variation; Coef X (Y): Pearson's linear correlation coefficient for each coordinate (X and Y) with each of the soil chemical properties; $\hat{\mu} = \beta_0$: estimated mean; $\hat{\Phi}_1$: estimated nugget effect; $\hat{\Phi}_2$: estimated contribution; \hat{a} : estimated practical range; \widehat{RNE} : estimated relative nugget effect ($\widehat{RNE} = \hat{\Phi}_1 / (\hat{\Phi}_1 + \hat{\Phi}_2)$ (%); for properties that showed a directional trend $\hat{\mu} = \beta_0 + \beta_1 Y_1$, in which β_0 (first value of the $\hat{\mu}$ column), β_1 (second value of the $\hat{\mu}$ 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 k = 1.5: Matérn model with smoothness parameter k = 1.5.

Table 5. Descriptive statistics and estimated values of the geostatistical model parameters for the soil chemical properties, referring to each harvest year and considering the sample configuration optimized by 70 %

Harvest year	Property	Descriptive statistics				Estimation of the properties by the geostatistical model					
		Mean	CV	Coef X	Coef Y	Model	$\hat{\mu}$	$\hat{\phi}_1$	$\hat{\phi}_2$	\hat{a}	RNE
2012-2013	C (g dm ⁻³)	32.11	11.25	0.33	0.01	M k = 1.5	-590.68; 0.0026	5.43	6.45	455.82	45.73
	Ca ²⁺ (cmol _c dm ⁻³)	6.63	26.19	0.17	0.29	Exp.	6.59	2.21	0.75	573.73	74.57
	Mn (mg dm ⁻³)	76.04	25.24	-0.02	0.35	Gaus.	-58730; 0.008	73.00	257.10	301.54	22.11
	Zn (mg dm ⁻³)	4.28	61.23	0.29	0.19	Gaus.	4.23	4.42	2.31	252.48	65.66
2013-2014	C (g dm ⁻³)	31.17	12.29	0.13	-0.15	Exp.	31.05	3.95	10.53	162.42	27.26
	Ca ²⁺ (cmol _c dm ⁻³)	6.23	24.17	-0.13	-0.02	M k = 1.5	6.19	0.87	1.36	149.90	39.16
	Mn (mg dm ⁻³)	62.38	34.39	-0.08	0.08	Gaus.	61.37	186.19	253.23	224.74	42.37
	Zn (mg dm ⁻³)	2.99	94.91	-0.01	0.03	Gaus.	2.93	2.87	5.11	195.09	35.96
2014-2015	C (g dm ⁻³)	29.01	12.91	0.28	-0.03	M k = 1.5	29.08	4.93	9.01	198.61	35.36
	Ca ²⁺ (cmol _c dm ⁻³)	5.41	26.67	0.20	0.14	Exp.	5.41	1.49	0.57	527.06	72.44
	Mn (mg dm ⁻³)	76.52	28.43	0.10	-0.06	M k = 2.5	78.20	270.30	242.40	650.41	52.71
	Zn (mg dm ⁻³)	2.80	67.31	0.27	0.05	Exp.	2.88	0.00	3.64	237.55	0.00
2015-2016	C (g dm ⁻³)	31.52	10.80	0.19	-0.26	Exp.	31.60	7.76	3.72	407.88	67.62
	Ca ²⁺ (cmol _c dm ⁻³)	5.55	26.15	-0.01	0.08	Gaus.	5.56	1.35	0.72	234.96	65.22
	Mn (mg dm ⁻³)	86.28	28.59	-0.13	0.10	Exp.	87.13	202.89	393.03	452.40	34.05
	Zn (mg dm ⁻³)	4.98	45.11	0.30	0.21	Gaus.	5.20	1.99	3.22	356.74	38.16

CV: coefficient of variation; Coef X (Y): Pearson's linear correlation coefficient for each coordinate (X and Y) with each of the soil chemical properties; $\hat{\mu} = \beta_0$: estimated mean; $\hat{\phi}_1$: estimated nugget effect; $\hat{\phi}_2$: estimated contribution; \hat{a} : estimated practical range; RNE : estimated relative nugget effect ($RNE = \hat{\phi}_1 / (\hat{\phi}_1 + \hat{\phi}_2)$ (%); for properties that showed a directional trend $\hat{\mu} = \beta_0 + \beta_1 Y_1$, in which β_0 (first value of the $\hat{\mu}$ column), β_1 (second value of the $\hat{\mu}$ 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 k = 1.5; 2.5: Matérn model with smoothness parameter k = 1.5.

Table 6. Estimated values of similarity measures Overall Accuracy (OA), Kappa (Kp) and Tau (T) for comparison between the initial sampling configuration and the optimized configurations at 50, 60 and 70 %, for the 2012-2013 harvest years, 2013-2014, 2014-2015 and 2015-2016

Harvest Year	Property	50 %			60 %			70 %		
		OA	Kp	T	OA	Kp	T	OA	Kp	T
2012-2013	C (g dm ⁻³)	72.05	58.85	65.07	77.48	65.46	71.86	78.26	75.05	72.79
	Ca ²⁺ (cmol _c dm ⁻³)	25.89	0.01	7.36	56.36	39.04	45.44	60.40	44.26	50.50
	Mn (mg dm ⁻³)	71.89	56.12	64.86	83.81	75.41	79.76	85.07	77.55	81.34
	Zn (mg dm ⁻³)	76.11	48.59	70.14	77.14	57.45	71.42	75.27	57.95	69.09
2013-2014	C (g dm ⁻³)	66.35	0.01	57.93	70.63	32.60	63.29	81.01	52.11	76.26
	Ca ²⁺ (cmol _c dm ⁻³)	57.82	36.66	47.28	51.84	28.55	39.80	77.17	61.35	71.46
	Mn (mg dm ⁻³)	66.29	46.40	57.86	82.76	68.86	78.45	81.64	68.76	77.05
	Zn (mg dm ⁻³)	61.34	38.78	51.68	74.01	58.96	67.51	74.18	58.78	67.73
2014-2015	C (g dm ⁻³)	46.66	21.10	33.32	30.24	0.69	12.80	74.76	45.85	68.45
	Ca ²⁺ (cmol _c dm ⁻³)	60.37	24.97	50.46	56.40	22.66	45.50	67.01	48.20	58.76
	Mn (mg dm ⁻³)	65.25	38.07	56.56	66.66	40.97	58.33	81.97	74.62	77.46
	Zn (mg dm ⁻³)	71.44	48.17	64.30	59.91	26.68	49.88	81.19	65.85	76.50
2015-2016	C (g dm ⁻³)	61.27	39.05	51.59	64.80	52.50	56.01	61.70	42.62	52.12
	Ca ²⁺ (cmol _c dm ⁻³)	61.85	41.11	52.31	75.97	50.82	69.96	82.16	70.05	77.69
	Mn (mg dm ⁻³)	56.87	39.30	46.08	68.88	55.04	61.10	81.27	70.75	76.59
	Zn (mg dm ⁻³)	73.48	56.33	66.85	84.05	75.53	80.06	83.62	74.02	79.52

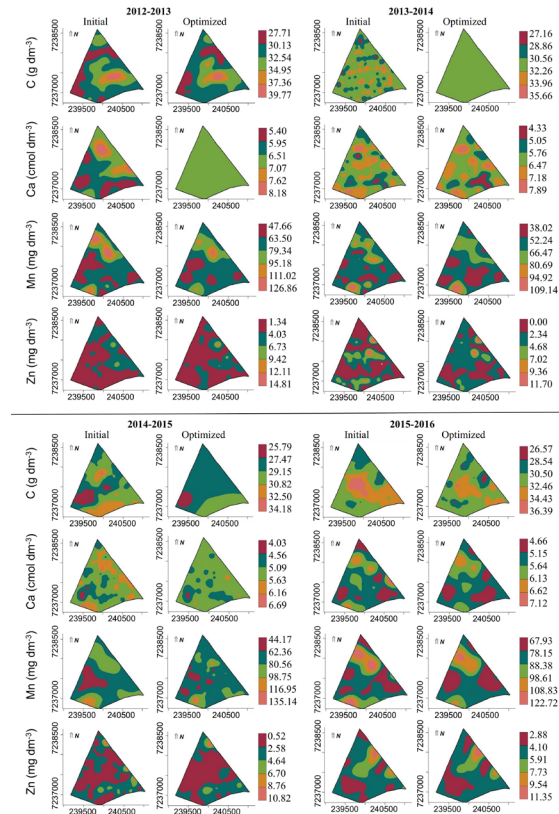


Figure 6. Thematic maps of the soil chemical properties considering the initial and optimized sample configuration by 50 % for the harvest years 2012-2013, 2013-2014, 2014-2015, and 2015-2016.

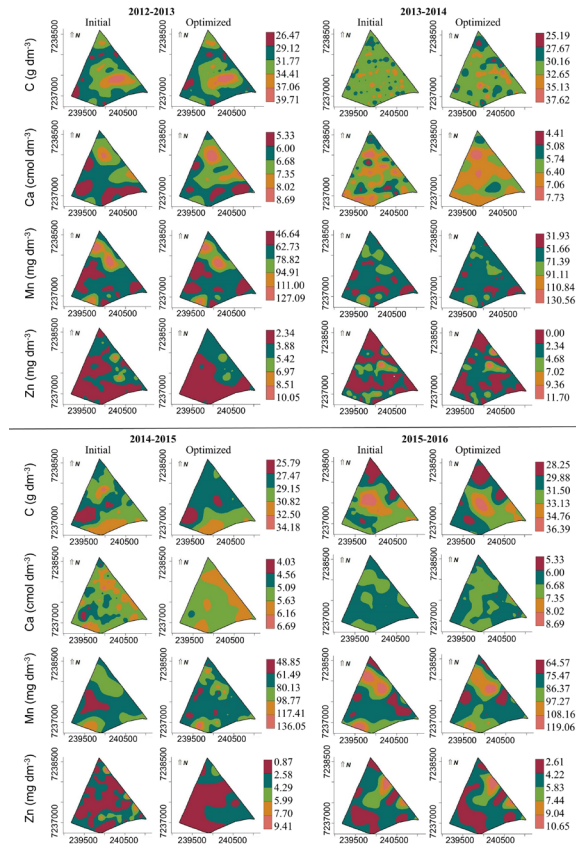


Figure 7. Thematic maps of the soil chemical properties considering the initial and optimized sample configuration by 60 % for the harvest years 2012-2013, 2013-2014, 2014-2015, and 2015-2016.

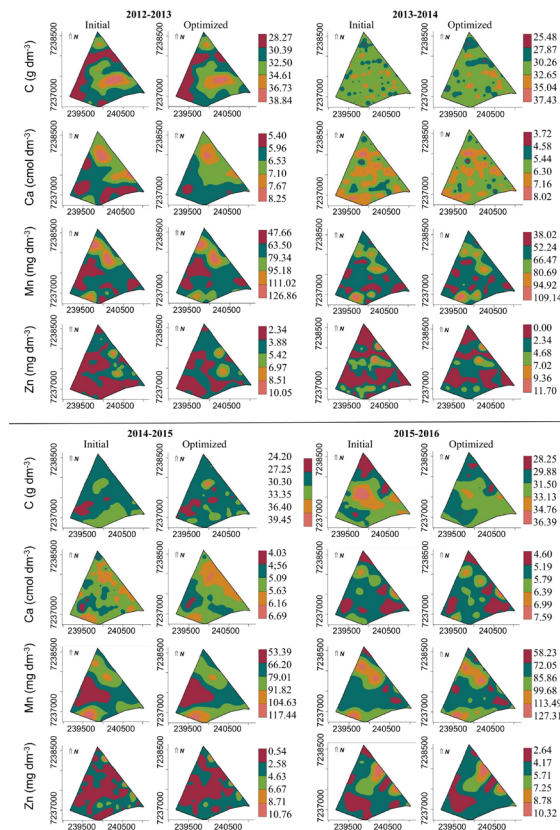


Figure 8. Thematic maps of the soil chemical properties considering the initial and optimized sample configuration by 70 % for the harvest years 2012-2013, 2013-2014, 2014-2015, and 2015-2016.

Considering all the optimized sample configurations, the best estimate for the accuracy index was observed when comparing the initial sample configurations and those optimized at 70 % (Table 6) for most of the soil chemical properties. This result was already expected because it contains more sample points than the sample reduction of 50 and 40 %. Thus, there was an influence of the sample reduction of the spatial variability.

CONCLUSION

Clustering methods were efficient for the application zones (AZ) definition. For all clustering methods, the ideal number of groups was two. Furthermore, considering the evaluation criteria, K-means was the best clustering method. Sample configurations could be optimized by 50, 60 and 70 % with genetic algorithm (GA). Nonetheless, for most cases, when compared with the initial sample configuration, the optimized sample configurations presented low or medium accuracy in the estimates of the accuracy index. The best estimates, for most of the soil chemical properties, were observed comparing the initial sample configurations and the optimized ones in 70 %. Finally, the cluster methods integration, EOFs for the description of the spatial-temporal data, and optimization method were efficient in generating the optimized sample configurations that conserved the spatio-temporal trends observed in the original set data.

DATA AVAILABILITY

Not applicable. The datasets presented in this article are not readily available because the data belong to a group of researchers at the University and are currently part of ongoing studies by researchers in the area of spatiotemporal statistics.


FUNDING

The authors gratefully acknowledge the funding received from the Coordination for the Improvement of Higher Education Personnel (CAPES), Financing Code 001, from the National Council for Scientific and Technological Development (CNPq) and Fundação Araucária of the State of Paraná.


ACKNOWLEDGEMENTS




The authors would also like to thank the Spatial Statistics Laboratory - LEE/UNIOESTE and Applied Statistics Laboratory -LEA/UNIOESTE - Cascavel, PR, Brazil.


AUTHOR CONTRIBUTIONS

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

Data curation:  Tamara Cantú Maltauro (lead).




Formal analysis:  Tamara Cantú Maltauro (lead).

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

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

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

Writing - original draft:  Tamara Cantú Maltauro (lead).

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

REFERENCES

- 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.
- Amaro Filho J, Negreiros RFDD, Assis Júnior RND, Mota JCA. Amostragem e variabilidade espacial de atributos físicos de um Latossolo Vermelho em Mossoró, RN. *Rev Bras Cienc Solo*. 2007;31:415-22. <https://doi.org/10.1590/S0100-06832007000300001>
- Anderson JF, Hardy EE, Roach JT, Witmer RE. A land use and land cover classification system for use with remote sensor data. 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>
- 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>

- Benestad RE, Mezghani A, Lutz J, Dobler A, Parding KM, Landgren OA. Various ways of using empirical orthogonal functions for climate model evaluation. *Geosci Model Dev.* 2023;16:2899-913. <https://doi.org/10.5194/gmd-16-2899-2023>
- Borge R, Jung D, Lejarraga I, de la Paz D, Cordero JM. Assessment of the Madrid region air quality zoning based on mesoscale modelling and k-means clustering. *Atmos Environ.* 2022;287:119258. <https://doi.org/10.1016/j.atmosenv.2022.119258>
- Branke J, Deb K, Miettinen K, Slowiński R. Multiobjective optimization: Interactive and evolutionary approaches. Berlin Heidelberg: Springer; 2008.
- 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>
- 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>
- Cordovil HPL, Lima HC, Oliveira JJLC, Silva AA. Impacto das mudanças climáticas na qualidade do solo e na produção agrícola. *Observ Econ Latino-Americ.* 2024;22:e6017. <https://doi.org/10.55905/oelv22n6-292>
- Cressie NAC. Statistics for spatial data. rev. ed. New York: John Wiley & Sons; 2015.
- Cressie NAC, Wikle CK. Statistics for Spatio-Temporal Data. Hoboken, NJ: Wiley; 2011.
- Crispim DL, Fernandes LL, Albuquerque RLO. Aplicação de técnica estatística multivariada em indicadores de sustentabilidade nos municípios do Marajó-PA. *Rev Principia.* 2019;46:145-54. <https://doi.org/10.18265/1517-03062015v1n46p145-154>
- 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>
- 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.
- Finkenbiner CE, Franz TE, Gibson J, Heeren DM, Luck J. Integration of hydrogeophysical datasets and empirical orthogonal functions for improved irrigation water management. *Precis Agric.* 2019;20:78-100. <https://doi.org/10.1007/s11119-018-9582-5>
- Finkenstadt B, Held L, Isham V. Statistical methods for spatio-temporal systems. New York: Chapman and Hall/CRC; 2007.
- Furtado DF. Análise Multivariada. Lavras-MG: Ministério da Educação e do Desporto, Universidade Federal de Lavras, Departamento de Ciências Exatas; 1996.
- Gasparin PP, Silva EM, Becker WR, Paludo A, Guedes LPC, Johann JA. Agroclimatic and spectral regionalization for soybean in different agricultural settings in the state of Paraná, Brazil. *J Agric Sci.* 2024;162:291-306. <https://doi.org/10.1017/S0021859624000340>
- 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>
- Gibson J, Franz TE. Spatial prediction of near surface soil water retention functions using hydrogeophysics and empirical orthogonal functions. *J Hydrol.* 2018;561:372-83. <https://doi.org/10.1016/j.jhydrol.2018.03.046>
- 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>
- Halkidi M, Vazirgiannis M, Batistakis Y. Quality scheme assessment in the clustering process. In: Zighed DA, Komorowski J, Żytkow J, editors. *Proceedings 4th European Conference, PKDD 2000, Lyon, France, September 13-16, 2000: Principles of data mining and knowledge discovery.* Berlin Heidelberg: Springer; 2000. v. 1910. p. 265-76. https://doi.org/10.1007/3-540-45372-5_26
- Hannachi A, Jolliffe IT, Stephenson DB. Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int J Climatol.* 2007;27:1119-52. <https://doi.org/10.1002/joc.1499>
- Hannachi A, Finke K, Trendafilov N. Common EOFs: A tool for multi-model comparison and evaluation. *Clim Dyn.* 2023;60:1689-703. <https://doi.org/10.1007/s00382-022-06409-8>
- Hirakuri MH. Impactos econômicos de estresses na produção de soja da safra 2015/16. Londrina: Embrapa Soja; 2016. (Circular técnica, 125).
- 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>
- Ikenaga S, Inamura T. Evaluation of site-specific management zones on a farm with 124 contiguous small paddy fields in a multiple-cropping system. *Precis Agric.* 2008;9:147-59. <https://doi.org/10.1007/s11119-008-9062-4>
- 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.
- Liu XL, Fu XQ, Li Y, Shen JL, Wang Y, Zou GH. Spatio-temporal variability in N₂O emissions from a tea-planted soil in subtropical central China. *Geosci Model Dev Discuss [Preprint];* 2016. p. 1-45. <https://doi.org/10.5194/gmd-2015-251>
- Lopes AS. *Manual internacional de fertilidade do solo.* 2. ed rev amp. Piracicaba: Potafos; 1998.
- Luchiari Junior A, Borghi E, Avanzi JC, Freitas AA, Bortolon L, Bortolon ESO, Ummus ME, Inamasu RY. Zonas de Manejo: teoria e prática. In: Inamasu RY, Naime JM, Resende AV, Bassoi LH, Bernardi AC, editors. *Agricultura de precisão: Um novo olhar.* São Carlos: Embrapa Instrumentação; 2011. p. 60-4.
- Ma Y, Liu H, Xu G, Lu Z. Empirical orthogonal function analysis and modeling of global tropospheric delay spherical harmonic coefficients. *Remote Sens.* 2021;13:4385. <https://doi.org/10.3390/rs13214385>
- MacQueen J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th 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. 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>
- 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, Canton LED. Spatial multivariate optimization for a sampling redesign with a reduced sample size of soil chemical properties. *Rev Bras Cienc Solo*. 2023a;47:e0220072. <https://doi.org/10.36783/18069657rbcs20220072>
- Maltauro TC, Guedes LPC, Uribe-Opazo MA, Canton LED. Multivariate spatial sample reduction of soil chemical attributes by means of application zones. *Span J Agric Res*. 2023b;21:e0205. <https://doi.org/10.5424/sjar/2023212-19521>
- 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.
- Noetzold R, Silva LM, Schoninger EL, Tomé PCDT, Alves MC. Variabilidade espacial e temporal de atributos químicos do solo durante cinco safras. *Rev Bras Geom*. 2018;6:328-45. <https://doi.org/10.3895/rbgeo.v6n4.8102>
- Oliveira EF. Treinamento: Fertilidade do solo e nutrição das plantas. Cascavel: Coodetec - Cooperativa Central de Pesquisa Agrícola; 2007.
- 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>
- Pantuzza 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.
- Pebesma E, Graeler B. gstat: Spatial and spatio-temporal geostatistical modelling, prediction and simulation. CRAN – Package; 2022. Available from: <https://cran.r-project.org/web/packages/gstat/gstat.pdf>.
- Pebesma E, Graeler B, Gottfried T, Hijmans RJ. Spacetime: Classes and methods for spatio-temporal data. CRAN – Package; 2023. Available from: <https://cran.r-project.org/web/packages/spacetime/spacetime.pdf>.
- Perez-Quezada JF, Pettygrove GS, Plant RE. Spatial-temporal analysis of yield and soil factors in two four-crop-rotation fields in the Sacramento Valley, California. *Agron J*. 2003;95:676-87. <https://doi.org/10.2134/agronj2003.6760>
- Perry MA, Niemann JD. Analysis and estimation of soil moisture at the catchment scale using EOFs. *J Hydrol*. 2007;334:388-404. <https://doi.org/10.1016/j.jhydrol.2006.10.014>
- Pimentel-Gomes F, Garcia CH. Estatística aplicada a experimentos agrônômicos e florestais. Piracicaba: Fealq; 2002.
- 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/>.
- Ribeiro Jr PJ, Diggle PJ. geoR: A package for geostatistical analysis. *R News*. 2016;1:15-8.
- Rodrigues Junior 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 HG, Jacomine PKT, Anjos LHC, Oliveira VA, Lumberreras 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.

Soil Survey Staff. Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys. 2nd ed. Washington, DC: United States Department of Agriculture, Natural Resources Conservation Service; 1999. (Agricultural Handbook, 436).

Souza ZMD, Cerri DG, Magalhães PS, Siqueira DS. Spatial variability of soil attributes and sugarcane yield in relation to topographic location. *Rev Bras Eng Agric Ambient*. 2010;14:1250-6. <https://doi.org/10.1590/S1415-43662010001200001>

Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conf Ser: Mater Sci Eng*. 2018;336:012017. <https://doi.org/10.1088/1757-899X/336/1/012017>

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

Uribe-Opazo MA, Johann JA, Boas MAV, Lunkes C, Borssoi JA. Métodos de ajuste à semivariogramas experimentais utilizando diferentes grades amostrais na produtividade da soja. *Rev Eng Agric*. 2007;15:319-30.

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>

Uribe-Opazo MA, Dalposso GH, Galea M, Johann JA, De Bastiani F, Moyano ENC, Grzegozewski D. Spatial variability of wheat yield using the Gaussian spatial linear model. *Aust J Crop Sci*. 2023;17:179-89. <https://doi.org/10.21475/ajcs.23.17.02.p3742>

Vilela I, Araujo M, Tyaquicã P, Velela D. Empirical orthogonal function analysis of satellite-derived currents in the Tropical Atlantic. *Trop Oceanogr*. 2018;46:1-24. <https://doi.org/10.5914/tropocean.v46i2.239346>

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

Wikle CK, Zammit-Mangion A, Cressie N. Spatio-temporal statistics with R. New York: Chapman and Hall/CRC; 2019.

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>

Zhao Y, Li F, Yao R, Jiao W, Hill RL. An empirical orthogonal function-based approach for spatially-and temporally-extensive soil moisture data combination. *Water*. 2020;12:2919. <https://doi.org/10.3390/w12102919>