Spanish Journal of Agricultural Research
21 (2), e0205, 13 pages (2023)
eISSN: 2171-9292
https://doi.org/10.5424/sjar/2023212-19521
INIA-CSIC

RESEARCH ARTICLE

OPEN ACCESS

Multivariate spatial sample reduction of soil chemical attributes by means of application zones

Tamara C. MALTAURO*, Luciana P. C. GUEDES, Miguel A. URIBE-OPAZO, and Letícia E. D. CANTON

Western Paraná State University, 2069 Universitária Street, 85819-110, Cascavel, PR, Brazil.

*Correspondence should be addressed to Tamara C. Maltauro: tamara_ma02@hotmail.com

Abstract

Aim of study: In precision agriculture, the definition of Application Zones (AZs) in agricultural areas consists in delimiting the area in subareas with similar characteristics, using soil chemical attributes. To such end, the use of clustering methods is common. Therefore, the AZs make up a database that can be used to target future soil sampling, thus seeking a possible sample reduction. The objective of this paper is to assess the acquisition of sample configurations, with reduced sample size, contained in application zones generated by spatial multivariate clustering. The sampling protocol proposed in this work evaluated five clustering methods (C-means, Fanny, K-means, Mcquitty, and Ward) for the creation of AZs, and, through these AZs, to obtain reduced sample configurations with 50% and 75% of the initial sampling points.

Area of study: Commercial agricultural area, Cascavel, Brazil.

Material and methods: Data of the soil chemical attributes from a commercial agricultural area were used, referring to three soybean harvest years (2013-2014; 2014-2015; and 2015-2016). The clustering methods considered a dissimilarity matrix that aggregates the information about the Euclidean distance between the sample elements and the spatial dependence structure of the attributes.

Main results: The results indicated division of the agricultural area into two or three AZs for the aforementioned harvest years, considering the K-means method. Comparing all the reduced sample configurations with the initial one, it was observed that the one proportionally reduced by 25% was the most effective to obtain a reduced sample configuration.

Research highlights: The sampling protocol using AZs showed that it is possible to reduce the sample size.

Additional key words: clustering; dissimilarity matrix; precision agriculture; sampling design.

Abbreviation used: AZ (Application Zone); GPS (global positioning system); Kp (Kappa); MZ (Management Zone); OA (overall accuracy); PA (Precision Agriculture); PCA (principal component analysis); PC1 (first principal component); PR (proportional random); PR50, PR25 (proportional random sample configurations reduced by 50% and 25%); PS (proportional systematic); PS50, PS25 (proportional systematic sample configurations reduced by 50% and 25%); R (random); R50, R25 (random sample configurations reduced by 50% and 25%); S (systematic); S50, S25 (systematic sample configurations reduced by 50% and 25%) SSE (sum of squares of errors); T (Tau); UTM (Universal Transverse Mercator).

Citation: Maltauro, TC; Guedes, LPC; Uribe-Opazo, MA; Canton, LED (2023). Multivariate spatial sample reduction of soil chemical attributes by means of application zones. Spanish Journal of Agricultural Research, Volume 21, Issue 2, e0205. https://doi.org/10.5424/sjar/2023212-19521

Supplementary material (Tables S1-S4; Figs. S1-S3) accompanies the paper on SJAR's website.

Received: 27 Apr 2022. **Accepted:** 03 May 2023.

Funding agencies/institutions
Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil. Financing Code 001
Post-Graduate Program in Agricultural Engineering (PGEAGRI), Brazil
National Council for Scientific and Technological Development (CNPq)
Spatial Statistics Laboratory-LEE/UNIOESTE, Brazil and Applied Statistics Laboratory –LEA/UNIOESTE, Brazil

Competing interests: The authors have declared that no competing interests exist.

Introduction

The proposal of precision agriculture (PA) aimed at localized soil management involves agronomic, technological and, consequently, financial changes, which hinders adoption of the PA techniques, especially by small producers (Srinivasan, 2006). PA allows studying and analysing the spatial variability of crop yield and the chemical and physical attributes of the area under study. The study of this spatial variability of spatially georeferenced continuous variables can be done using geostatistical techniques (Cressie, 2015), which is crucial for the feasibility of differentiated and localized management practices, and allowing to measure the degree of spatial dependence between the sample elements in a given region and to describe the spatial dependence structure of the georeferenced variable throughout the area, thus generating thematic maps (Cressie, 2015).

An important part of the PA is to obtain data for the elaboration of maps of attributes that influence crop productivity (Bernardi et al., 2014). There are commercially available equipment and sensors that can be highlighted for obtaining these data, as well as for collecting soil samples (Gonçalves et al., 2020).

Agricultural sensors are instruments that can be attached to implements or agricultural machinery, allowing, for example, to analyze soil attributes or characteristics in real time. Scherer et al. (2018) proposed the development of a sensor to determine the electrical conductivity of the soil at low cost. Prudente et al. (2021) compared the NDVI (Normalized Difference Vegetation Index) spectro-temporal profiles obtained by active (GreenSeeker 505 Handheld) and passive (FieldSpec4 Standard-Res model) proximal sensors to monitor soybeans and beans. By means of multispectral sensors onboard the AT120 Remotely Piloted Aircraft Systems, Facco & Pegoraro (2019) captured aerial images, and thus built digital surface models, orthoimages, contour lines, generated vegetation indices and performed the detection of planting failures. Thinking about agility and quality, Resende et al. (2020) used a remotely piloted aircraft equipped with an RGB camera and a MAPPIR 3 camera to capture images of a corn crop, to estimate the leaf area index of a plot infested by *Spodoptera frugiperda.* Furthermore, the wireless sensor network can be used in underground and terrestrial environments to detect soil and climate conditions, as well as detect pests and insects (Bayrakdar, 2020a,b).

Another important procedure in PA is soil sampling, which represents a critical step in the assessment of soil fertility, and may be responsible for 98% of the errors made in the inappropriate recommendation of fertilizers (Mendes et al., 2001). Still, the choice and influence of the location of the sampled points, as well as the distance that separates them, are essential for the success of the sampling design (Carvalho & Nicollela, 2002).

In this context, it is necessary to determine a soil sampling scheme that is as efficient as possible, with the smallest sample size to minimize operating costs and maximize the quality of the results of the spatial variability analysis (Marchant & Lark, 2010).

For the choice of this sampling scheme, there are traditional spatial sampling designs, such as simple random sampling and the systematic design, which are classified as design-based, assuming that the values of the georeferenced variable are considered fixed. Traditional spatial sampling designs are advantageous in terms of the simplicity of choice of sampling points in the area under study, either because of the occurrence of different spacing between pairs of sampling points (in simple random sampling) or because of coverage uniformity of the area under study (in systematic sampling) (Haining, 2015).

However, the spatial heterogeneity and spatial autocorrelation perform a particular impact on the variability analysis accuracy, in the context of the sampling design. The spatial distribution of an attribute is considered heterogeneous when the attribute varies in the study area, thus requiring division of the area into homogeneous subareas. On the other hand, spatial autocorrelation indicates that the georeferenced samples nearby have similar values.

Characteristics of spatial variability, such as anisotropy, can guide a choice of the most intense number of sampling points in the direction of greater spatial variability (Delmelle, 2009). Or the inclusion of more sampling points with a smaller distance radius in the systematic sampling design, to minimize the nugget effect problem, as proposed by Diggle & Lophaven (2006), in the "lattice plus close pairs" and "lattice plus infill" designs.

The spatial sampling designs that consider the sample as a realization of a probability process are classified as model-based design or geostatistical design (Diggle & Lophaven, 2006). The spatial sampling design that considers the spatial correlation makes it possible to reduce the sample size by restricting redundant spatial information, in relation to traditional sampling (Haining, 2015; Dal'Canton et al., 2021).

Several methodologies for spatial re-planning of the sample, with sample size reduction, are developed in the literature, such as calculation of the effective sample size (Dal'Canton et al., 2021) or the optimized sample re-planning (Maltauro et al., 2019; 2021). Moreover, prior knowledge of the spatial autocorrelation of the variable can be used in the spatial stratified sampling, which provides a lesser loss of precision of the estimate than spatial simple random and systematic sampling. The spatial stratified sampling can use relevant previous historical data in the same area, such as prior knowledge. Associated with the multivariate clustering analysis, this information allows generating zones, which will represent the stratified space (Wang et al., 2013).

In the PA context, the identification of homogeneous Management Zones (MZs) within the cultivation areas becomes a more technically and economically viable strategy for the implementation of localized soil management,

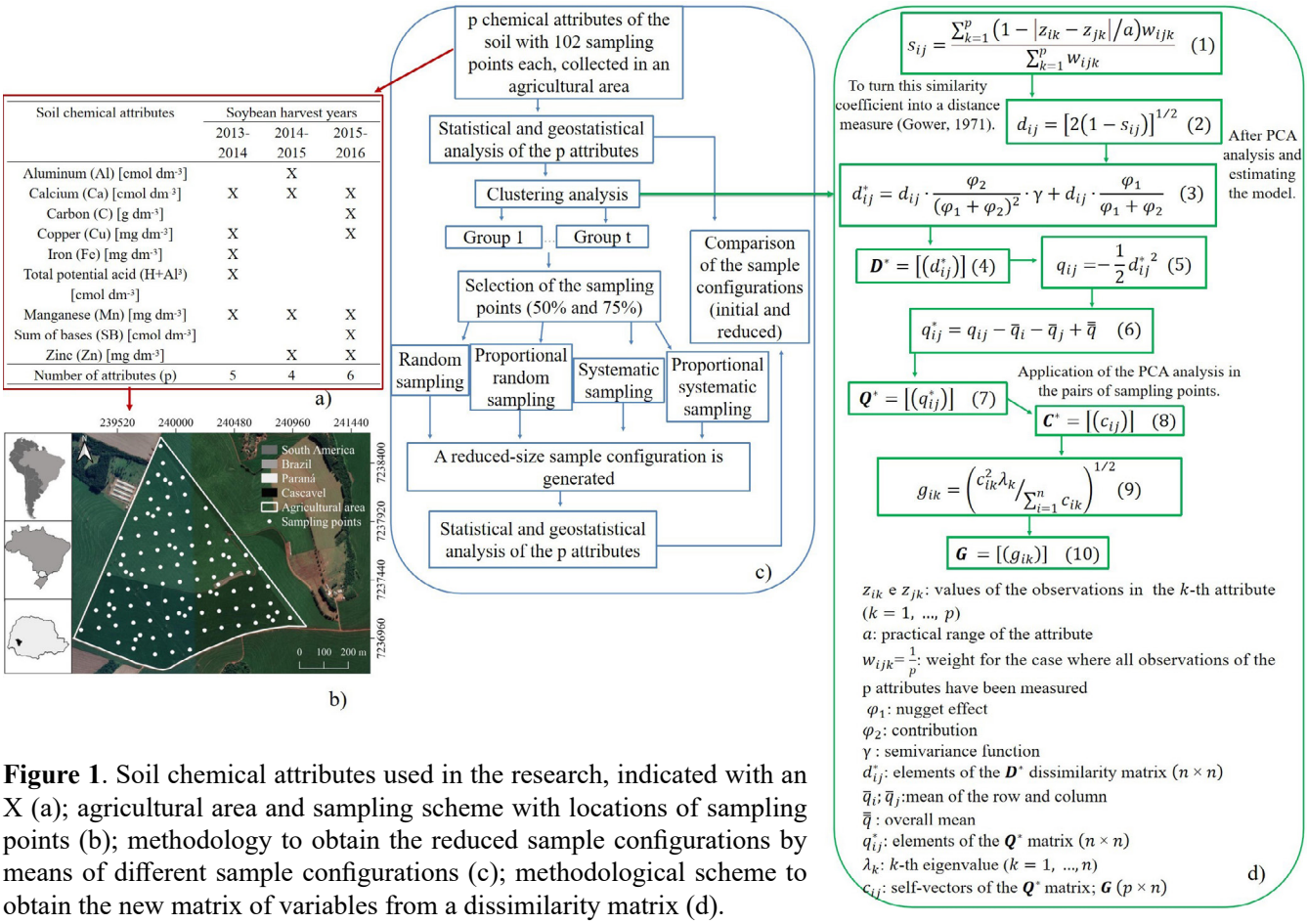

Figure 1. Soil chemical attributes used in the research, indicated with an X (a); agricultural area and sampling scheme with locations of sampling points (b); methodology to obtain the reduced sample configurations by means of different sample configurations (c); methodological scheme to obtain the new matrix of variables from a dissimilarity matrix (d).

allowing a long-term that all area treatments to be carried out uniformly within each zone (Molin, 2006; Srinivasan, 2006; Bernardi et al., 2014). Thus, to produce more stable MZs, it is recommended to use variables that do not vary significantly over time, such as topographic data and physical data (Aikes et al., 2021).

On the other hand, the chemical attributes are important and useful to generate the zones for recommendations of variable rate fertilizer applications (Aikes et al., 2021). These zones are called application zones (AZs) (Molin, 2006). Therefore, the difference between MZs and AZs is related to the variables available (stables or not) and with the intention of generating the zones (long-term use or only for a future application of fertilizers). Moreover, the spatial statistics and the multivariate cluster analysis are methods commonly used by MZ and AZ.

These MZs or AZs make up a database that can be used to direct, in the future, the reduction of the sample size in areas more homogeneous (Rodrigues Jr. et al., 2011). In this context, there are several ways to select samples, which depend on the purpose and the resources available for the survey (Haining, 2015). When we do not have information about the spatial variability, the sampling points can be selected using traditional sampling criteria (Benedetti et al., 2015), or multi-phase adaptive sampling schemes (Marchant & Lark, 2010).

Determination of the sample configurations with reduced size by the use of AZs as a spatial stratified sampling provide the following advantages: AZs generate more homogeneous regions from the point of view of spatial similarity, allowing the researcher to choose a smaller number of sampling points in each AZ; the methodology considers the spatial information redundant; there is a design-based and model-based combination; use of less computational time than in optimized sampling; and greater simplicity of execution.

Given the above, the objective of the paper was to assess the acquisition of reduced sample configurations contained in application zones generated by spatial multivariate clustering. The sampling protocol proposed in this work is intended to guide the definition of new intervention cycles in the study area, with a reduced sample size, but maintained the refined localized management in this agricultural area.

Material and methods

Agricultural area or experimental field

The data set regarding the soil chemical attributes in the three soybean harvest years (2013-2014, 2014-2015, and 2015-2016) used in this research was observed in a 167.35

ha commercial grain production area located at the Three Girls Farm in the city of Cascavel-PR, Brazil, located approx. 24.95º S 53.37º W with a mean altitude of 650 m above sea level. The soil is classified as Dystroferric Red Latosol, with a clayey texture. The region's climate is classified as mesothermal and super-humid temperate, climate type Cfa (Köeppen), and the mean annual temperature is 21ºC (Embrapa, 2013). The area under study had 102 sampling points arranged in the lattice plus close pairs design (Chipeta et al., 2017), with a minimum distance between the points of the regular grid of 141 meters and, in some places, randomly chosen, the sampling points were arranged at smaller distances (75 and 50 m between point pairs (Fig. 1b).

The 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 at each point indicated (Fig. 1b). At these points, four soil sub-samples were collected, from 0.0 to 0.2 m deep, mixed and placed in plastic bags, with samples of approx. 500 g comprising the representative sample of the plot.

Descriptive and geostatistical analyses

Each harvest year, descriptive and geostatistical analyses were performed for each of the soil chemical attributes, in order to verify the presence of directional trend and anisotropy. Anisotropy was assessed through the analysis of the directional semivariograms (Guedes et al., 2018) and the non-parametric Maity & Sherman´s (2012) test, considering 5% significance. Spatial dependence was assessed by the nugget-to-sill radio classification (Cambardella et al., 1994), which is a good and consolidated method to assess the intensity of spatial dependence, being used in several works in the field of Soil Science (Siqueira et al., 2010; Guedes et al., 2018; Maltauro et al., 2019, 2021; Dal'Canton et al., 2021). The soil chemical attributes that presented spatial dependence were studied (Cambardella et al., 1994), referring to each harvest years (Fig. 1a).

The following models were estimated by the maximum likelihood method: exponential, Gaussian, and Matérn family with shape parameter =2.5 (Cressie, 2015). The best model was chosen by means of the cross-validation method (Faraco et al., 2008). Subsequently, the spatial prediction was carried out in non-sampled locations in the agricultural area under study, by kriging, and thematic maps of each attribute were prepared (Landim, 2006).

Acquisition of the spatial and multivariate dissimilarity matrix

Subsequently, all the locations were compared in pairs. For this, in each pair of i and j locations ($i, j=1,...,n$) in which the p attributes had already been measured (Fig. 1a), the similarity coefficient proposed by Gower (1971) was calculated and, for quantitative attributes, the practical range is a form of standardizing the attributes (Eq. 1; Fig. 1d). The dissimilarity matrix was obtained based on Oliver & Webster (1989).

In the principal component analysis (PCA) of the original data (Eq. 2; Fig. 1d), the first principal component (pc1) was selected, as this explains most of the data variation. Considering the pc1 scores, the geostatistical models were estimated in a way analogous to the methodology used for the soil chemical attributes. The dissimilarity matrix was obtained with the estimation of the parameters of the pc1 scores' geostatistical model (Eq. 3; Fig. 1d). In this way, the matrix adds information about the Euclidean distance between the sample elements, as well as the spatial dependence structure of the attributes.

The columns of matrix (Eq. 10; Fig. 1d) are the new variables. Consequently, the number of columns corresponding to the number of original attributes was selected. Subsequently, a geostatistical model was estimated and data interpolation by kriging was made. The interpolated data were used to obtain the AZs (Gavioli et al., 2016).

Clustering and choice of the number of clusters and criteria to evaluate the clusterings

Initially for each harvest year, the best clustering method was chosen among the following: Fanny, Fuzzy C-means, McQuity, Ward, and K-means. Details about the clustering methods evaluated are described in Ward Jr. (1963), McQuitty (1966), MacQueen (1967), Bezdek (1981), and Kaufman & Rousseeuw (2009). This choice was made by means of the following indices: Dunn, Davies Bouldin, C, SD, and variance reduction (Dunn, 1974; Hubert & Levin, 1976; Davies & Bouldin, 1979; Halkidi et al., 2000; Gavioli et al., 2016, respectively). To define the number of clusters, the scatter plot of the sum of squares of errors (SSE) versus the number of clusters (knee graph) was used, as well as the silhouette scatter plot versus the number of clusters (Yi et al., 2013).

Sample configuration

With the best AZ chosen for each harvest year, a sample reduction was performed in each AZ. Sample configuration was reduced considering 75% and 50% of the sampling points and using different sample configurations (Fig. 1c). Greater reductions were not possible, as the number of sampling points would not meet the geostatistical analysis criteria (at least 30 pairs for calculation of the semivariances; Journel & Huijbregts, 1976).

First, sample reduction was carried out considering random (R) and proportional random (PR) sampling, selecting, respectively, the sampling points randomly within each AZ

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

Harvest year	Attribute[1]	Descriptive statistics[2]				Estimated parameters of the geostatistical model[3]				
		Mean	CV	Coef.X	Coef.Y	Model	$\hat{\mu}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{a}$
2013-2014	Ca	6.22	22.46	-0.08	-0.04	Gaus.	6.19	1.08	0.87	179.00
	Cu	1.21	60.18	0.05	-0.29	Gaus.	1.26	0.28	0.25	707.86
	Fe	37.10	22.41	-0.07	-0.07	Gaus.	37.37	35.93	33.06	217.86
	H+Al[3]	8.60	22.55	0.15	0.02	Exp.	8.62	2.62	2.18	157.70
	Mn	60.96	33.69	-0.11	-0.11	Gaus.	60.31	171.59	225.55	203.98
2014-2015	Al	0.28	126.28	-0.15	0.09	M. κ=2.5	0.28	0.02	0.10	128.61
	Ca	5.38	25.11	0.22	0.03	Exp.	5.40	1.05	0.75	231.56
	Mn	76.54	27.43	0.07	-0.02	Gaus.	77.30	233.70	209.80	453.07
	Zn	2.81	61.61	0.30	0.01	Gaus.	-326.7; 0.001	0.54	2.28	162.73
2015-2016	C	32.01	10.58	0.11	-0.23	Exp.	31.80	5.97	5.37	576.28
	Ca	5.50	24.12	-0.01	0.05	Gaus.	5.53	1.29	0.48	284.08
	Cu	3.82	23.78	-0.10	0.24	Exp.	4.02	0.33	0.52	855.10
	Mn	86.41	25.66	-0.11	0.09	Gaus.	86.78	268.79	226.14	367.29
	SB	7.93	25.20	-0.01	-0.04	Exp.	7.93	2.73	1.22	149.73
	Zn	4.97	40.92	0.21	0.23	Gaus.	5.10	1.59	3.04	367.65

[1] H+Al[3]: total potential acid. [2] CV: coefficient of variation; Coef.X, Coef.Y: Pearson's linear correlation coefficient (r) for each coordinate (X and Y) with each of the soil's chemical attributes. [3] Gaus.: Gaussian; Exp.: exponential; M. κ=2.5: Matérn with κ=2.5; $\hat{\mu}$, $\hat{\varphi}_1$, $\hat{\varphi}_2$, $\hat{a}$: estimated values of the mean, nugget effect, partial sill, and practical range (meters) parameters, respectively.

or proportionally to the number of hectares within each AZ. In the sample reduction using systematic (S) and proportional systematic (PS) sampling, the sampling points were obtained by selecting of points from the regular sampling grid, to obtain 50% and 75% of the sampling points. The points of the agricultural area's lattice plus close pairs were selected until the number of required points was completed.

With each reduced sample configuration, the exploratory and geostatistical data analyses were repeated. Finally, the initial and the reduced sample configurations were compared using the overall accuracy (Anderson et al., 2001) and the Kappa and Tau agreement indices (Krippendorff, 2013).

Computational resources

All statistical and geostatistical analyses were developed using the R software (R Development Core Team, 2021), or its geoR package (Ribeiro Jr. & Diggle, 2001).

Results

Descriptive and geostatistical statistics

Regarding all the soil's chemical attributes for the 2013-2014 and 2014-2015 harvest years, there was a wide variation in the coefficient of variation (CV), from 22.41 to 60.18 and from 25.11 to 126.28 respectively, what indicates high dispersion ($20 < CV \leq 30$) or heterogeneity of the soil chemical attributes ($CV > 30$). The 2015-2016 harvest year presented a CV between 10.58 and 40.92, ranging from average ($10 \leq CV \leq 20$) to high ($20 < CV \leq 30$) dispersion of the soil chemical attributes (Table 1).

Soil chemical attributes Ca, C, Cu, Fe, H+Al[3], Mn, and SB had mean values considered average or high, whereas the mean value of Zn can be classified as low or average, and that of Al can be classified as low regarding the need for the soil (Table 1). The Zn content of the 2014-2015 harvest year presented a moderate linear association of its respective values with the X-axis coordinates, that is, showing a mean linear trend of the deterministic term in relation to the East-West direction (Table 1) ($r \geq 0.30$).

For all the soil's chemical attributes of the 2013-2014 harvest year, an estimated value for the spatial dependence radius (range) from 157.70 m to 707.86 m was observed (Table 1). For the 2014-2015 harvest year, the practical range varied from 128.61 m to 453.07 m. The 2015-2016 harvest year presented greater variation in the practical range, between 149.73 m to 855.10 m. Regarding spatial dependence intensity, and according to the criteria used, it is observed that, for all the harvest years, the soil chemical attributes presented moderate or strong spatial dependence (Table 1).

Table 2. Estimated values of the overall accuracy (OA), Kappa (Kp) and Tau (T) similarity measures for comparison between the initial sample configuration and the reduced configurations for the 2013-2014, 2014-2015 and 2015-2016 harvest years.

	Attributes	Indexes	R50	R25	PR50	PR25	S50	S25	PS50	PS25
2013-2014	Ca	OA	58.61	70.30	61.63	59.02	62.40	70.59	71.06	80.14
		K_p	31.56	53.98	39.46	35.40	37.92	44.57	33.85	60.78
		T	48.26	72.87	52.04	48.77	53.00	63.24	63.82	75.18
	Cu	OA	47.38	81.47	50.19	75.77	66.86	74.85	68.45	82.44
		K_p	26.59	76.02	20.29	69.31	51.26	67.77	59.38	77.44
		T	34.22	76.83	37.73	69.71	58.57	68.56	60.56	78.06
	Fe	OA	53.58	79.71	71.90	53.46	52.66	78.00	62.94	57.91
		K_p	30.61	62.05	50.86	31.21	28.65	64.35	22.79	33.63
		T	41.98	74.64	64.88	41.82	40.83	72.50	53.67	47.38
	$H+Al^3$	OA	58.45	78.72	46.12	56.08	66.93	46.12	46.51	58.95
		K_p	30.61	60.42	0.00	24.77	39.43	0.00	0.69	20.95
		T	48.07	73.39	32.65	45.09	58.66	32.65	33.13	48.69
	Mn	OA	66.75	77.28	61.98	74.66	72.57	89.79	59.19	84.37
		K_p	34.13	58.11	22.16	55.00	45.66	82.03	0.00	70.52
		T	41.98	71.60	52.47	68.32	65.71	87.24	49.98	80.47
2014-2015	Al	OA	84.94	85.81	78.86	87.44	81.99	81.97	85.94	86.45
		K_p	24.11	60.43	36.77	63.62	0.99	0.00	50.73	54.07
		T	81.18	82.27	73.57	84.29	77.48	77.47	82.43	83.06
	Ca	OA	59.24	61.42	61.05	80.19	54.82	45.23	64.28	85.13
		K_p	23.92	38.50	23.83	64.23	1.02	18.05	29.97	71.47
		T	49.05	51.78	51.32	75.23	43.53	31.54	55.36	81.41
	Mn	OA	67.35	79.04	71.08	88.34	59.34	80.47	48.87	65.93
		K_p	53.30	70.42	50.06	82.97	28.09	72.35	24.09	47.18
		T	59.19	73.80	63.84	85.43	49.18	75.59	36.09	57.41
	Zn	OA	62.27	66.36	67.76	79.82	68.69	74.12	70.68	76.56
		K_p	31.88	40.70	41.23	60.44	43.63	55.39	47.15	59.40
		T	52.83	57.95	59.69	74.77	60.86	67.65	63.36	70.70
2015-2016	C	OA	70.82	65.40	56.69	73.46	47.51	63.15	48.22	74.19
		K_p	59.08	40.87	35.53	63.82	15.39	47.34	20.62	63.87
		T	63.53	56.75	45.86	66.82	34.38	53.94	35.27	67.74
	Ca	OA	51.37	81.83	67.73	66.59	73.73	51.37	81.23	76.55
		K_p	0.00	58.12	37.15	41.23	35.17	0.56	44.27	50.87
		T	39.21	77.29	59.67	58.24	67.16	39.21	76.54	70.69
	Cu	OA	73.79	76.69	64.95	79.59	50.76	73.79	52.90	60.48
		K_p	58.65	65.26	44.18	68.51	15.26	60.06	19.41	35.57
		T	67.24	70.87	56.18	74.48	38.46	67.24	41.13	50.60
	Mn	OA	61.49	72.57	59.96	63.45	52.20	70.29	55.03	72.50
		K_p	44.74	62.67	42.04	49.24	32.73	57.14	36.63	61.94
		T	51.87	65.71	49.95	54.31	40.25	62.87	43.79	65.63
	SB	OA	64.32	68.50	66.66	53.40	59.03	64.93	64.93	64.93
		K_p	0.00	39.21	27.51	20.72	23.58	0.00	0.00	7.67
		T	55.40	60.62	58.32	41.75	48.79	56.16	56.16	56.16
	Zn	OA	43.13	75.19	69.93	85.36	70.97	83.03	71.01	85.83
		K_p	14.14	57.77	48.91	74.52	52.14	73.21	52.44	77.72
		T	28.92	68.99	62.41	81.70	63.72	78.79	63.76	82.29

R50 (R25): sample configuration randomly reduced by 50% (25%). PR50 (PR25): sample configuration randomly proportionally reduced by 50% (25%). S50 (S25): sample configuration systematically reduced by 50% (25%). PS50 (PS25): sample configuration systematically proportionally reduced by 50% (25%).

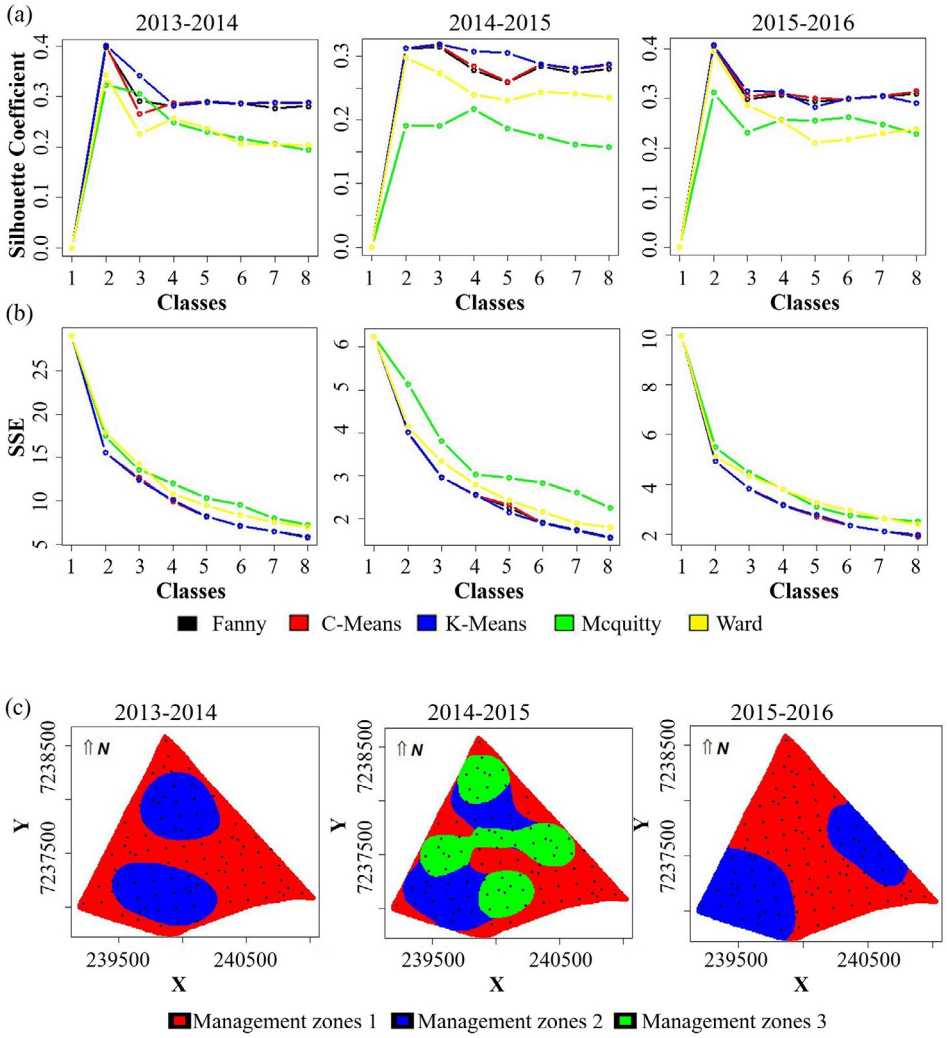

Figure 2. Silhouette graph (a), knee graph (b) and thematic maps (c) with the best number of application zones (AZs) and the best clustering method.

Clustering

For the 2013-2014 and 2015-2016 harvest years, the scatter plots of the number of clusters versus the SSE and the silhouette (Fig. 2a,b) showed that the best number of clusters for all the clustering methods was 2. For the 2014-2015 harvest year, and for most of the clustering methods, the ideal number of clusters was 3 (Fig. 2a,b). Moreover, K-means was the best clustering method for all the harvest years (Table S1 [suppl]). With the best number of clusters and the best clustering method selected, the AZ map was generated, for all the harvest years (Fig. 2c).

It is observed that the AZs present differences in the seasons, as they are generated with the chemical attributes of the soil, to be used for a single application recommendation. It is noted that a larger AZ was created in all harvest years (red colour, Fig. 2c). In addition to that, it was observed that there is at least one AZ in the Southwest region in all the harvest years (Fig. 2c). And, except for the 2015-2016 harvest year, it was also possible to find at least one AZ in the North region (Fig. 2c).

In the harvest years that presented two AZs, the largest (red) occupied 106.85 ha (64% of the plot) and 99.61 ha (60% of the plot) for the 2013-2014 and 2015-2016 harvest years, respectively (Fig. 2c). The 2014-2015 harvest year, which featured three AZs, had 69.64 and 54.03 ha in the two largest AZs, corresponding to 42% and 32% of the total area, respectively (Fig. 2c).

Sample configuration

For the 2013-2014, 2014-2015, and 2015-2016 harvest years, AZ 1 covered 60, 32, and 65 sampling points, respectively, which correspond to 59%, 31%, and 64% of the total points in the study area (Fig. 2c). AZ 2 comprises 42, 26 and 37 sampling points, corresponding, respectively, to 41%, 26%, and 36% of the total points in the study area (Fig. 2c). In addition, in the 2014-2015 harvest year, the third AZ included 44 sampling points (43% of the total points in the study area). Sample configurations reduced by 50% obtained 51 sampling points; on the other hand,

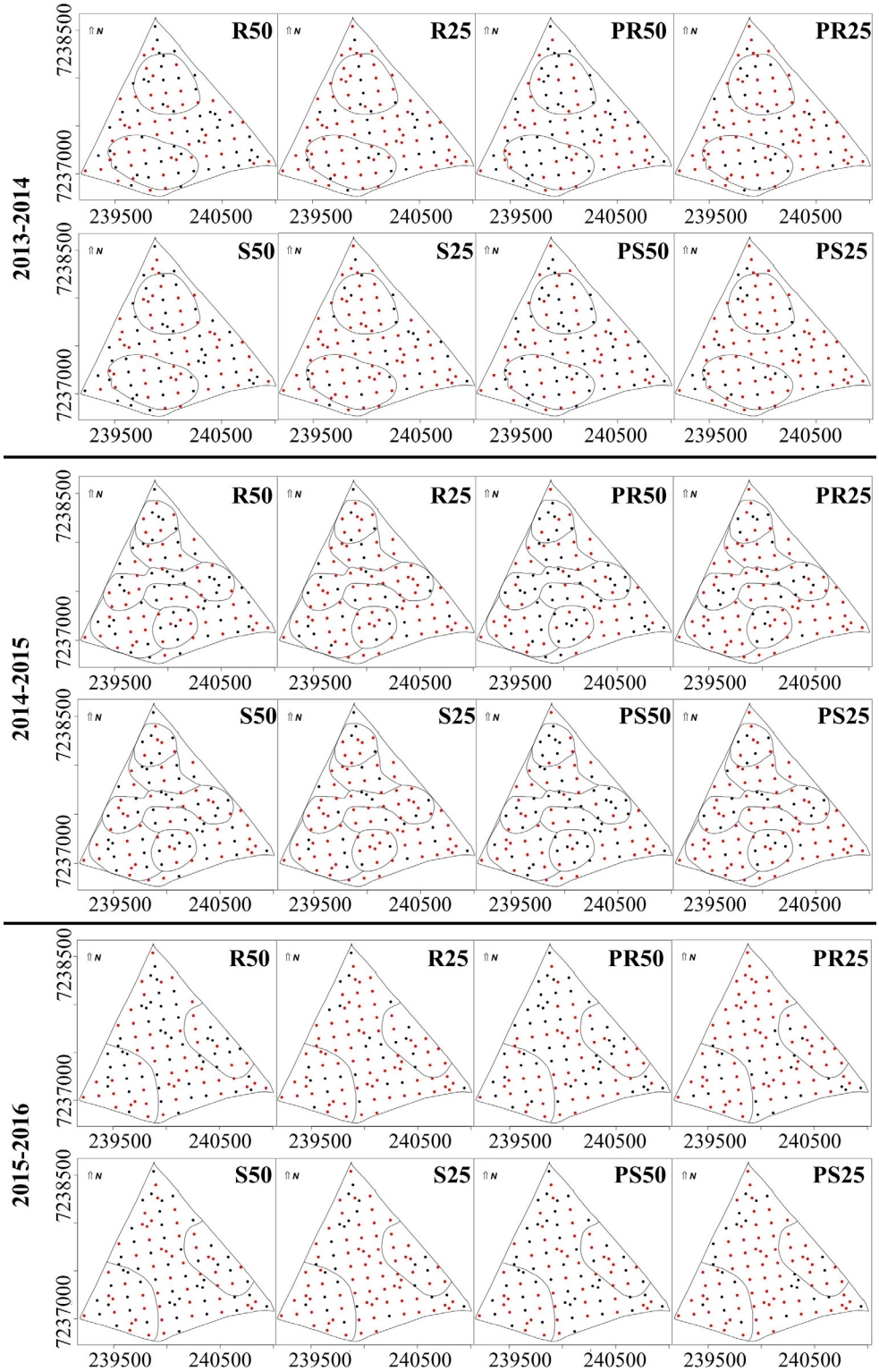

Figure 3. Initial and reduced sample configurations (points in red •) for the 2013-2014, 2014-2015 and 2015-2016 harvest years (random [R], proportional random [PR], systematic [S], and proportional systematic [PS] sample configurations reduced by 50% and 25%).

sample configurations reduced by 25% had 76 sampling points distributed in the agricultural area (Fig. 3).

For all harvest years, similarity in descriptive statistics was observed when comparing all sample configurations reduced by 50% and 25% with the initial one (Table 1 and Fig. 4). For all harvest years, most of the attributes did not present any directional trend for the initial sample configurations, except for soil chemical attribute Zn for the

2014/2015 harvest year, which showed a directional trend in the Y direction (North-South). As for the reduced sample configurations, soil chemical attributes Cu, $H+Al^3$, Mn, and Zn presented a directional trend in the X (East-West) or Y (North-South) direction for at least one sample, with Pearson's linear correlation coefficient (r) values greater than 0.30 in a module.

Regarding the classification of spatial dependence intensity of the soil chemical attributes, for attributes $H+Al^3$ and Mn in the 2013-2014 harvest year, there was a change in the classification, from moderate (Table 1) to weak (Table S2) spatial dependence intensity, in the PR50, PR25, S25, PS50 and PS25 sample configurations. The presence of pure nugget effect was found in soil chemical attributes Ca, Cu and Fe, mainly in the systematic sample configurations (Table S2). For the 2014-2015 harvest year, soil chemical attributes Al and Ca presented weak spatial dependence in the R50 and S25 sample configurations (Table S3).

The proportional and random sample configurations had a greater number of attributes with pure nugget effect, namely: Al, Mn, and Zn (Table S3). For the 2015-2016 harvest year, soil chemical attributes C, Ca, Cu, and SB also presented weak spatial dependence in at least one of the systematic sample configurations (Table S4).

Disregarding the cases that presented low spatial dependence and pure nugget effect, the spatial dependence radius of all the reduced sample configurations were compared with the initial one, showing variations (of more or less) of: 4.94 m to 106.75 m for Al; 27.55 m to 381.36 m for C; 1.38 m to 209.41 m for Ca; 26.64 m to 541.63 m for Cu; 5.76 m to 574.31 m for Fe; 9.42 m to 45.48 m for $H+Al^3$; 1.00 m to 288.82 m for Mn; 65.14 m to 168.61 m for SB; and 75.44 m to 213.18 m for Zn (Table 1 and Table S2 to S4), regardless of the harvest year.

When comparing the thematic maps of the chemical attributes generated considering the initial and the reduced configurations in all harvest years, most of the soil chemical attributes presented low or average accuracy by the estimated values of the Kappa and Tau agreement indices; with values between 0.00% and 78.79% (low accuracy if Kappa; Tau < 67%, average accuracy if 67% ≤ Kappa; Tau < 80%) (Table 2).

For the 2013-2014 harvest year, it was observed that only the Mn content in the soil in the S25 sample presented an estimated overall accuracy (OA) value greater than 85%, which indicates that the maps of both configurations are similar regarding distribution of the content of this soil attribute in the study area (OA ≥ 85%) (Table 2; Fig. S1). For the 2014-2015 harvest year, the Al attribute with samples PR25, PS50 and PS50; the Ca attribute in SP75; and the Mn attribute in PR25 also presented estimated OA values above 85% (Table 2; Fig. S2). Attribute Al exhibited high accuracy with values between 81.18% and 84.29% for most of the sample configurations. However, one of the main reasons is the fact the pixels fall into only one class. Finally, for the 2015-2016 harvest year, only the Zn attribute presented high overall accuracy in sample configurations PR25 and PS25, with estimated OA values above 85% and Kappa; Tau ≥ 80% (Table 2; Fig. S3).

Discussion

Regardless of considering the same variables year to year to generate zones, Schenatto et al. (2016), working with different soil attributes, also found different amounts of the zones for an agricultural area considering different harvest years. The sampling points obtained by the reduced sampling configurations (points in red in Fig. 3) within each AZ (Fig. 3), showed that sampling points selected throughout the study area, seeking greater concentration of points in regions where there is greater variability, as well as a reduction in sample density in more uniform locations (Rodrigues Jr. et al., 2011).

The similarity in descriptive statistics when comparing the reduced sample configurations with the initial sample configurations was also found by Maltauro et al. (2019); and Dal'Canton et al. (2021) working with the sample reduction in the same agricultural area with grain cultivation, considering the chemical attributes of the soil.

The results regarding spatial dependence are explained by the bilateral relationship of sample size in the spatial scale (Kerry et al., 2010). A reduction in sample size can generate a reduction in spatial dependence intensity, as evidenced by these authors, who presented examples with sample size reduction that generated an increase in the nugget effect. These authors also mention the importance of defining an adequate sample size, to remove the effect of the "noise" that may have been generated exclusively by the choice of sample size.

In general, most of the soil's chemical attributes showed moderate spatial dependence; this fact makes the thematic maps more accurate than those generated considering weak spatial dependence (Cambardella et al., 1994). In addition, a high nugget effect indicates low spatial dependence and leads to estimates around the sample mean (Alencar et al., 2019), as the nugget effect is associated with sampling or analysis errors, which indicates that two observations close to each other have very different values. Regarding the range, Dal'Canton et al. (2021) concluded that the larger the spatial dependence radius, the smaller the sample size, estimated by the effective size of the univariate sample and the greater the area of homogeneity between sampling points.

The best estimates for the accuracy indices were observed comparing the initial sample configurations with those reduced by 25% (Table 2). This result was already expected, for containing more sampling points than the 50% sample reduction. However, it was possible to observe that the spatial variability pattern was maintained in most classes of the thematic map of the soil chemical

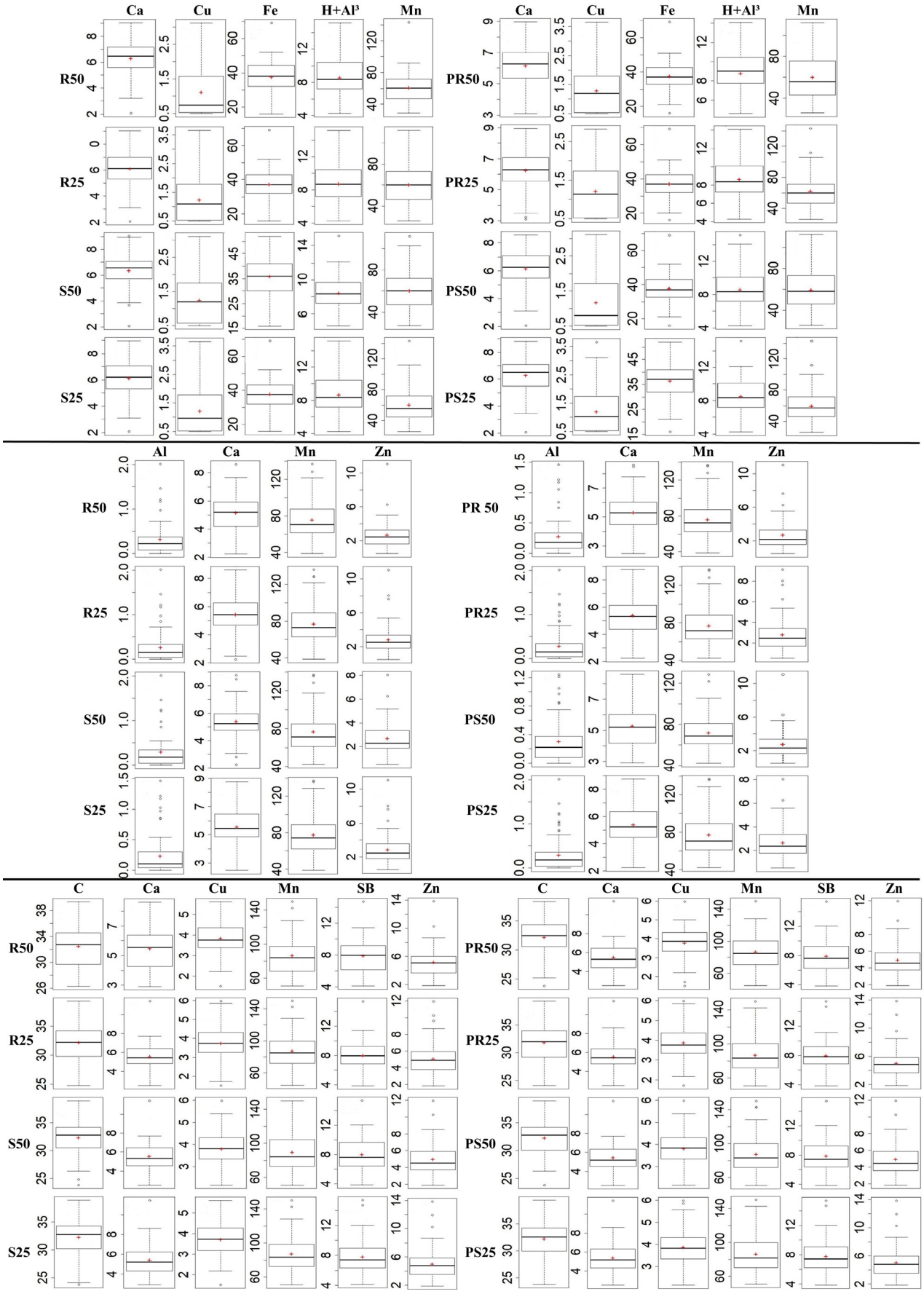

Figure 4. Boxplots of the reduced sample configurations for the 2013-2014, 2014-2015 and 2015-2016 harvest years, with the respective soil chemical attributes (random [R], proportional random [PR], systematic [S], and proportional systematic [PS] sample configurations reduced by 50% and 25%).

attributes, even considering the largest sample reduction. This same trend was observed in Maltauro et al. (2019) and by Dal'Canton et al. (2021), even the authors working with different methods for sample reduction in an area with grain cultivation.

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

Among the configurations analysed, the sample configuration proportionally reduced by 25%, when compared to the initial sample configuration, presented the best estimates for the values of the spatial dependence radius and the highest values for the accuracy indices; while the worst estimates for the accuracy indices were shown when comparing the initial sample configuration and those randomly and systematically reduced by 50%.

Overall, the AZs allowed dividing the agricultural area into more homogeneous sub-regions, as well as to select a smaller number of points within each AZ. Thus, the definition of AZs is efficient in obtaining a reduced sample configuration and in defining future soil samplings in the study area, aiming to minimize, in the long term, the spatial variability of the soil chemical attributes soil in this agricultural area, also allowing the producer to reduce costs when carrying out the soil analysis. This methodology used presents a design-based and model-based combination, greater simplicity in execution and brief computational times.

As future works, we seek to apply the methodology developed in spatio-temporal data.

Authors' contributions

Conceptualization: T. C. Maltauro, L. P. C. Guedes.
Data curation: Not applicable.
Formal analysis: T. C. Maltauro, L. E. D. Canton.
Funding acquisition: T. C. Maltauro, L. P. C. Guedes.
Investigation: T. C. Maltauro.
Methodology: T. C. Maltauro, L. P. C. Guedes.
Project administration: L. P. C. Guedes.
Resources: Not applicable.
Software: T. C. Maltauro, L. E. D. Canton, L. P. C. Guedes.
Supervision: L. P. C. Guedes, M. A. Uribe-Opazo.
Validation: T. C. Maltauro, L. E. D. Canton, L. P. C. Guedes.
Visualization: T. C. Maltauro, L. E. D. Canton, L. P. C. Guedes.
Writing – original draft: T. C. Maltauro, L. E. D. Canton, L. P. C. Guedes, M. A. Uribe-Opazo.
Writing – review & editing: T. C. Maltauro, L. P. C. Guedes, M. A. Uribe-Opazo.

References

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

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

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

Bayrakdar ME, 2020a. Employing sensor network based opportunistic spectrum utilization for agricultural monitoring. Sust Comput: Inform Syst 27: 100404. https://doi.org/10.1016/j.suscom.2020.100404

Bayrakdar ME, 2020b. Enhancing sensor network sustainability with fuzzy logic based node placement approach for agricultural monitoring. Comput Electron Agric 174: 105461. https://doi.org/10.1016/j.compag.2020.105461

Benedetti R, Piersimoni F, Postiglione P, 2015. Sampling spatial units for agricultural surveys. Springer, Berlin. https://doi.org/10.1007/978-3-662-46008-5

Bernardi ACC, Naime JM, Resende AV, Bassoi LH, Inamasu RY, 2014. Agricultura de precisão: Resultados de um novo olhar. Embrapa, São Paulo.

Bezdek JC, 1981. Pattern recognition with fuzzy objective function algorithms. Springer, Boston. https://doi.org/10.1007/978-1-4757-0450-1

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

Carvalho JRP, Nicollela G, 2002. Uso de geoestatística na definição de plano de amostragem em levantamento de parâmetros do solo-uma proposta. Embrapa Informática Agropecuária, Campinas.

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

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

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

Davies DL, Bouldin DW, 1979. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1: 224-227. https://doi.org/10.1109/TPAMI.1979.4766909

Delmelle E, 2009. Spatial sampling. In: The SAGE handbook of spatial analysis, pp: 183-206. https://doi.org/10.4135/9780857020130.n10

Diggle P, Lophaven S, 2006. Bayesian geostatistical design. Scand J Stat 33(1): 53-64. https://doi.org/10.1111/j.1467-9469.2005.00469.x

Dunn JC, 1974. Well-separated clusters and optimal fuzzy partitions. J Cybern 4(1): 95-104. https://doi.org/10.1080/01969727408546059

Embrapa, 2013. Sistema Brasileiro de Classificação de Solos, 3ed. Centro Nacional de Pesquisa de Solos, Empresa Brasileira de Pesquisa Agropecuária, Brasília. 306 pp.

Facco BFL, Pegoraro A, 2019. Utilização de sistemas de aeronaves remotamente pilotadas na agricultura de precisão. Rev Geonorte 10(34): 129-152. https://doi.org/10.21170/geonorte.2019.V.10.N.34.129.152

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

Gavioli A, Souza EG, Bazzi CL, Guedes LPC, Schenatto K, 2016. Optimization of management zone delineation by using spatial principal components. Comput Electron Agric 127: 302-310. https://doi.org/10.1016/j.compag.2016.06.029

Gonçalves JRM, Ferraz GA, Reynaldo ÉF, Marin DB, Ferraz PF, 2020. Comparative economic analysis of soil sampling methods used in precision agriculture. An Acad Bras Cienc 92(suppl. 1): e20190277. https://doi.org/10.1590/0001-3765202020190277

Gower JC, 1971. A general coefficient of similarity and some of its properties. Biometrics 27(4): 857-871. https://doi.org/10.2307/2528823

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

Haining R, 2015. Spatial sampling. International Encyclopedia of the Social & Behavioral Sciences 23: 185-190. https://doi.org/10.1016/B978-0-08-097086-8.72065-4

Halkidi M, Vazirgiannis M, Batistakis Y, 2000. Quality scheme assessment in the clustering process. Eur Conf Principles of Data Mining and Knowledge Discovery, Lyon, France. Proc. PKDD, pp. 265-276. https://doi.org/10.1007/3-540-45372-5_26

Hubert LJ, Levin JR, 1976. A general statistical framework for assessing categorical clustering in free recall. Psychol Bull 83(6): 1072-1080. https://doi.org/10.1037/0033-2909.83.6.1072

Journel AG, Huijbregts CJ, 1976. Mining geostatistics. Academic Press, London.

Kaufman L, Rousseeuw PJ, 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New Jersey.

Kerry R, Oliver MA, Frogbrook ZL, 2010. Sampling in precision agriculture. In: Geostatistical applications for precision agriculture, pp. 35-63. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-9133-8_2

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

Landim PMB, 2006. Sobre geoestatística e mapas. Terrae Didatica 2(1): 19-33. https://doi.org/10.20396/td.v2i1.8637463

MacQueen J, 1967. Some methods for classification and analysis of multivariate observations. Proc 5th Symp Mathematical Statistics and Probability, Berkeley. pp. 281-297.

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

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

Maltauro TC, Guedes LPC, Uribe-Opazo MA, Canton LED, 2021. A genetic algorithm for resizing and sampling reduction of non-stationary soil chemical attributes optimizing spatial prediction. Span J Agric Res 19(4): e0210. https://doi.org/10.5424/sjar/2021194-17877

Marchant B, Lark R, 2010. Sampling in precision agriculture, optimal designs from uncertain models. In: Geostatistical applications for precision agriculture, pp. 65-87. Springer. https://doi.org/10.1007/978-90-481-9133-8_3

McQuitty LL, 1966. Similarity analysis by reciprocal pairs for discrete and continuous data. Educ Psychol Measur 26(4): 825-831. https://doi.org/10.1177/001316446602600402

Mendes AM, Locatelli M, Quisen RC, Vieira AH, 2001. Amostragem de solo para as culturas de pupunha, freijó e castanha-do-brasil. Embrapa 202 (01): 1-2.

Molin JP, 2006. Agricultura de precisão aprimora o gerenciamento. Visão Agrícola 5: 115-118.

Oliver MA, Webster R, 1989. A geostatistical basis for spatial weighting in multivariate classification. Math Geol 21(1): 15-35. https://doi.org/10.1007/BF00897238

Prudente VHR, Mercante E, Johann JA, de Souza CHW, Oldoni LV, Almeida L, et al., 2021. Comparison between vegetation index obtained by active and passive proximal sensors. J Agric Stud 9(2): 391-405. https://doi.org/10.5296/jas.v9i2.18462

R Development Core Team, 2020. R: A language and environment for statistical computing. Version 4.0.0. R Foundation for Statistical Computing, Vienna, Austria.

Resende DB, de Abreu Jr CAM, Martins GD, José O, Marques LCMX, 2020. Uso de imagens tomadas por aeronaves remotamente pilotadas para detecção da cultura do milho infestada por Spodoptera frugiperda. Rev Bras de Geogr Fís 13(1): 156-166. https://doi.org/10.26848/rbgf.v13.1.p156-166

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

Rodrigues Jr FA, Vieira LB, Queiroz DMD, Santos NT, 2011. Geração de zonas de manejo para cafeicultura empregando-se sensor SPAD e análise foliar. Rev Bras Eng Agríc

Ambient 15(8): 778-787. https://doi.org/10.1590/S1415-43662011000800003

Schenatto K, Souza EG, Bazzi CL, Bier VA, Betzek NM, Gavioli A, 2016. Data interpolation in the definition of management zones. Acta Sci Technol 38(1): 31-40. https://doi.org/10.4025/actascitechnol.v38i1.27745

Scherer FS, Maldaner S, dos Santos Mello MV, Miranda PB, da Cunha Lima A, Luz AR, 2018. Construção de um sensor de condutividade elétrica do solo: uma proposta multidisciplinar. Ciência e Natura 40 (Edição especial: II mostra de projetos da UFSM): 107-110. https://doi.org/10.5902/2179460X35507

Srinivasan A, 2006. Handbook of precision agriculture: Principles and applications. CRC Press, NY. https://doi.org/10.1201/9781482277968

Siqueira DS, Marques Jr J, Pereira GT, 2010. The use of landforms to predict the variability of soil and orange attributes. Geoderma 155(1-2): 55-66. https://doi.org/10.1016/j.geoderma.2009.11.024

Wang JF, Jiang CS, Hu MG, Cao ZD, Guo YS, Li LF, et al., 2013. Design-based spatial sampling: Theory and implementation. Environ Model Softw 40: 280-288. https://doi.org/10.1016/j.envsoft.2012.09.015

Ward Jr JH, 1963. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301): 236-244. https://doi.org/10.1080/01621459.1963.10500845

Yi J, Du Y, Wang X, He Z, Zhou C, 2013. A clustering analysis of eddies' spatial distribution in the South China Sea. Ocean Sci 9(1): 171-182. https://doi.org/10.5194/os-9-171-2013