Australian Journal of Crop Science AJCS

AJCS 17(7):570-580 (2023) ISSN:1835-2707
doi: 10.21475/ajcs.23.17.07.p3895

Regionalization of an agricultural area by means of multivariate data and their relationship with soybean productivity

Rodrigo Lorbieski*, Luciana Pagliosa Carvalho Guedes, Miguel Angel Uribe-Opazo, Franciele Buss Frescki Kestring

State University of Western Paraná (UNIOESTE), Cascavel, PR, Brazil

***Corresponding author: rodrigo.lorbieski@hotmail.com**

Abstract

Regionalization of an agricultural area by dividing it into different clusters is an important strategy in the precision agriculture scope. Multivariate and spatial data are common in the design of these divisions. This paper sought to characterize regional differences in the area under study through different subsets of variables formed by soil physical-chemical variables and vegetative indices, in an agricultural area for four soybean harvest years in the period from 2013/2014 to 2016/2017. To such end, three subsets were generated comprised by these variables, which presented spatial dependence and were grouped according to their characteristics. By means of decision trees, it was identified which of these variables exerted the most influence on subdivision of the area. The multivariate and non-parametric spatial clustering technique was used to generate the clusters. Finally, by means of maps and boxplots, the spatial relationships between these variables and soybean productivity were evaluated. There was variation across the harvest years in relation to the subset of variables that determined the best design of the different clusters. The regional differences determined by the different variables used in the study showed no relationship with soybean productivity, which presented spatial homogeneity in its data for the harvest years evaluated. This approach is recommended when there is high spatial variability of factors that exert impacts on productivity, advising on using both soil physical-chemical variables and the vegetative indices to explain the causes of soybean productivity spatial variability.

Keywords: cluster; decision tree; non-parametric spatial statistics; core-estimator function; vegetative indices; soil physical-chemical variables.

Abbreviations: ALL_ subset that contains all the variables; Al_ aluminum; ARVI_ atmospherically resistant vegetation index; ASC_ average silhouette coefficient; CCC_ cophenetic correlation coefficient; CFA_ humid subtropical climate; C_ carbon; Ca_ calcium; Cindex_ C index; Cu_ copper; CV_ coefficient of variation; DB_ Davies-Bouldin index; DUNN_ Dunn index; EVI2_ enhanced vegetation index 2; Fe_ iron; K_ potassium; Mg_ magnesium; Mn_ manganese; NDVI_ normalized difference vegetation index; OSAVI_ optimized soil adjusted vegetation index; P_ phosphorus; PC_ subset that contains the soil physical and chemical variables; PE_ plant emergence; R4_ full pod stage; R6_ full seed stage; R7_ grain maturation stage; RNE_ relative nugget effect; SAVI_ soil adjusted vegetation index; SD_ SD index; SPR_ soil penetration resistance; VI_ subset that contains the vegetative indices; WDRI_ wide dynamic range vegetation index; Zn_ zinc.

Introduction

Farmers face the challenge of increasing crop yields without expanding the planted area and, to such end, they seek technological advances to improve what they know about each crop, allowing for efficient use of inputs (Deiss et al., 2020). A number of research studies highlight the importance of investigating the relationship between the soil variables and crop yields in order to improve management of the planted area (Malvezi et al., 2019; Deus et al., 2020). The soil is a dynamic and complex system that undergoes the influence of various physical and chemical processes (Marinković et al., 2018) and, consequently, its variability is affected by factors such as its own characteristics or the management and use practices to which it is subjected (Gülser et al., 2016; Santos Jr. et al., 2021). Some studies show that soil variables such as texture, structure, nutrient contents and pH, among others, can present considerable variations within the same rural property (Rosemary et al., 2017; Behera et al., 2018; Metwally et al., 2019; Mwendwa et al., 2022). This variability can become a challenge for agricultural production, as it can affect distribution of the nutrients, water retention and oxygen availability, exerting a direct influence on plant growth and, consequently, on productivity (Sanchez et al., 2011; Nyéki et al., 2022). In addition to the analysis of the spatial patterns of the soil physical-chemical variables, analyzing the vegetative indices also allows us to monitor the changes in the growth environments of a crop and, thus, assess the vegetation response pattern to the various factors that affect the crop, assisting in management of this area (Rodriguez et al., 2006; Cordeiro et al., 2017). Therefore, it is crucial to understand this variability to maximize productivity and ensure sustainability of the agricultural production system (Oliveira et al., 2018; Vian et al., 2016). Precision agriculture is a management practice that takes into account spatial variability to improve the efficacy of agricultural production (Cherubin et al., 2022). Applying techniques associated with precision agriculture allows acquiring more detailed information, which assists in more efficient decision-making in relation to crop management (Dalchiavon et al., 2017). Among the techniques used in precision agriculture are those related to machine learning, namely: clustering techniques and classification techniques, such as decision trees (Liakos et al., 2018). The clustering techniques allow for a more precise identification of areas with similar characteristics, assisting in understanding patterns that may not be easily seen (Priya and Venkateswari, 2018). This tool groups the areas into different clusters, allowing for a detailed analysis of the relationship between the different variables being worked on, which assists in identifying areas with specific problems (Gavioli et al., 2019). In turn, the classification techniques, such as decision trees, allow identifying the most important variables within a set, which can be useful to reveal spatial homogeneity patterns and to identify spatial relationships between the different variables (Zheng et al., 2009; Burdett and Wellen, 2022).

Consequently, this paper sought to delimit subregions in the area under study that presented regional differences through different data subsets and, thus, to identify the variables that most contributed to regionalization of this area and whether this regionalization was in accordance with soybean productivity spatial distribution. This knowledge allows us to better understand the relationships of spatial distributions between the different variables analyzed and that can be related to plant development and final productivity.

Results

Analysis and choice of the best cluster

Tables 1 and 2 present the fit metrics of the clusters for each subset tested in each harvest year. Thus, for the 2013/2014 and 2014/2015 harvest years and, according to the following indices (ASC, DUNN, DB and CCC) and (ASC, SD and DB) (Table 1), respectively, the subset that presented the best result was the one considering all the variables (ALL). The results of the (Cindex, DUNN and CCC) and (SD, DB and CCC) indices (Table 2) indicate that, for the 2015/2016 and 2016/2017 harvest years, respectively, the subset chosen in both years was the one comprised only by the soil physical-chemical (PC) variables.

By means of these same metrics, it is noticed that the optimum number of clusters in the study area was two in all the harvest years evaluated. For the 2013/2014 and 2014/2015 harvest years, this number of clusters is indicated by the ASC, SD and DB indices evaluated for the subset comprised by all the variables (ALL), respectively. For the 2015/2016 harvest year, the indices that indicate this number of divisions as the best for the area under study are ASC, C, SD, DUNN and DB, evaluated for the subset consisting of the soil physical-chemical (PC) variables. For the 2016/2017 harvest year, the indices that indicate this number of clusters are ASC, SD, DUNN and DB, also evaluated for the subset comprised only by the soil physical-chemical (PC) variables (Tables 1 and 2).

The subgroups chosen according to the indices proposed were formed by all the variables (ALL) for the first two harvest years (2013/2014 and 2014/2015) and by the physical-chemical variables for the last two (2015/2016 and 2016/2017). Thus, in general, it is noticed that there was not a specific set of variables that stood out in terms of defining the clusters; in other words, the subset that best grouped the data varied according to the harvest year.

For each harvest year, the sampling maps described with their respective clusters (Figure 1) are divided into CLUSTER1 and CLUSTER2. In these maps, it is observed that, for the 2013/2014, 2014/2015 and 2016/2017 harvest years, a smaller region (CLUSTER2) was formed in the Southwest region of the map, as well as another larger region (CLUSTER1), occupying the rest of the area. For the 2015/2016 harvest year, partition of the area was different, with the map divided between North (CLUSTER2) and South (CLUSTER1).

Profile of the different clusters by harvest year

Considering the decision tree (Figure 2) prepared for the 2013/2014 harvest year, using the (ALL) subset that generated the clusters, it is generally observed that CLUSTER1 presented the highest values for the ARVI vegetative index for 09/13/2013, 12/18/2013 and 01/19/2014, which respectively correspond to the plant emergence (PE), full seed (R6) and beginning of grain maturation (R7) stages; in other words, this cluster presented higher values in all three vegetative periods evaluated, which correspond to the initial and final phases of the soybean vegetative cycle (Figure 2). In turn, CLUSTER2 presented slightly higher values for moisture in the layer from 0 to 10 cm deep and in copper (Cu) content in the soil (Figure 2).

In relation to the 2014/2015 harvest year, the clusters were generated considering the subset in which all the study variables were included. According to the decision tree (Figure 3), CLUSTER1 presents higher values for the SAVI index calculated for 12/05/2014 as its main characteristic, corresponding to the full pod stage (R4). Also according to the decision tree, there are places within this cluster where high values were observed for the SAVI index in the R4 stage and low values for the NDVI index in the R6 stage (full seed), with this region located at the Northwest of the area according to the maps (Figure 3). In turn, CLUSTER2 is characterized by slightly higher values for SPR in the layers between 20 and 30 cm deep and for calcium (Ca) content in the soil.

Also for the 2014/2015 harvest year, although not identified by the decision tree, the plant emergence (PE) stage also presented a relevant difference between the clusters, according to the analysis of the boxplots and maps for this variable, where CLUSTER1 presented higher values than CLUSTER2 (Figure 4). In general, the indices presented similar results; however, the maps generated by NDVI and SAVI showed more evident differences between the clusters.

The 2015/2016 harvest year was the one that most differentiated itself from the others analyzed, in relation to the formation of clusters. The subgroup that best divided the study area was comprised by the soil physical-chemical variables. Thus, when analyzing the decision tree generated for this harvest year, it is observed that the variables that exerted the most influence on differentiation of the clusters were carbon (C) and copper (Cu) content (Figure 5). As its main characteristic, CLUSTER1 presents carbon content values greater than or equal to $32\ \mathrm{g\ dm^{-3}}$, representing the highest values for this variable in the entire plantation (Figure 5). In the regions within this cluster, where the carbon (C) content value is below $32\ \mathrm{g\ dm^{-3}}$, regions with low values for copper (Cu) and high values for iron (Fe) and zinc (Zn) content are also found (Figure 5). In turn, CLUSTER2 is characterized for having slightly lower values for carbon (C) content and higher ones for copper (Cu) content, with the highest concentration of this latter variable in the plantation (Figure 5).

Also for the 2015/2016 harvest year, among all the soil physical variables, moisture and density in the layers between 0 and 10 cm and between 10 and 20 cm deep, respectively, were the ones that presented the highest differentiation between the clusters formed (Figure 6), with the region comprised by CLUSTER1 showing low values for both variables. Even so, these differences are not so large, with most of the maps presenting intermediate values (Figure 6).

In the clusters generated for the 2016/2017 harvest year, the soil physical-chemical variables were also used. As its main characteristic, CLUSTER1 has lower SPR values in the layers from 0 to 10 cm deep, with values below 2.437 kPa (Figure 7). It is also possible to notice that, in the places within this same cluster, where the value of this variable is above 2.437 kPa, those for phosphorus (P) content in the soil were above $19\ \mathrm{mg\ dm^{-3}}$, while carbon (C) content in the soil was above $36\ \mathrm{g\ dm^{-3}}$ (Figure 7). CLUSTER2 is characterized by having higher values for SPR in the layers between 0 and 10 cm of soil depth and lower ones for phosphorus (P) content and density in the layer between 11 and 20 cm deep (Figure 7).

Soybean productivity analysis by harvest year

In this study, it was observed that the regional differences generated by the different data subsets in the area under study exerted little influence on productivity during the 2013/2014, 2014/2015 and 2016/2017 harvest years. In other words, when analyzing the boxplots and maps for these variables (Figure 8), no significant difference was observed in productivity in relation to the clusters formed. In all three cases, the different clusters presented intermediate values in relation to productivity. The 2015/2016 harvest year was the one that presented the highest difference among the clusters in relation to yield, according to the boxplot and the thematic map. It is observed that the productivity values in CLUSTER2 are lower than in CLUSTER1, which characterizes CLUSTER2 as a less productive region in relation to the total area.

In the thematic maps it is also possible to notice certain homogeneity in distribution of the values for all four harvest years under study. This fact is also made evident when the descriptive statistics for each harvest year is verified (Table 3). According to this table, in all harvest years, the coefficients of variation were in the ranges considered low (< 10%) or average (10% < CV < 20%) according to the criteria proposed by Pimentel-Gomes (2009); in other words, these results also point to the little variation of the values of this variable in the field, also representing an indication of this homogeneity.

Discussion

The results obtained showed that there is variation across the harvest years regarding the definition of which subset presented the best adjustments to group the locations and, consequently, to define the different clusters. Therefore, there is no specific subset that stands out from the others to generate these clusters. This result was already expected due to the fact that many of the variables used in the paper, such as the soil chemical variables, are not considered temporally stable (Gavioli et al., 2016). Therefore, it is to be expected that the subset that best divides the area will also vary according to the years.

Analyzed together with the thematic maps and the boxplots of the values corresponding to the variables in each group, the decision tree was able to identify the existence of differences in the study area, which is important to identify different behavior patterns in the plants that may come to influence productivity.

In relation to the clusters generated in this study, it is verified that the vegetative indices and the soil physical variables were the variables with the greatest contribution to differentiating the area and generating clusters for the 2013/2014 and 2014/2015 harvest years. According to Alvino et al. (2020), it is possible to establish relationships between the vegetative indices and the characteristics of the crops observed in the field and, thus, to interpret vegetative vigor of the crops and guide management decisions. In general, CLUSTER1 presented greater vegetative vigor in almost all phenological stages evaluated, and the plat emergence (PE) stage (Figures 2, 3 and 4) was the one that most stood out in terms of differentiation of the clusters generated for both harvest years, indicating a smaller vegetation cover area in this region during these periods or late development of the CLUSTER2 plants in relation to CLUSTER1.

The main characteristics that distinguish both regions delimited for the 2013/2014 harvest year are essentially given by slightly higher values for soil moisture in the layer between 0 and 10 cm deep and lower vegetative vigor, mainly in the plant emergence (PE) stage for CLUSTER2. In other words, at first sight, there seems to be an association between moisture and soybean development. According to Collares et al. (2006), moisture controls soil aeration, temperature and mechanical strength, which in turn regulate root growth and functionality, reflecting in growth and productivity. High soil moisture values can cause a decrease in aeration, which is not considered beneficial for plant development and may exert effects on productivity (Grable and Siemer, 1968).

The 2014/2015 harvest year is characterized by having higher SPR values and lower vegetative indices in the early stages of CLUSTER2 soybean development. The values for SPR may have influenced initial soybean development since, according to Freddi et al. (2006), SPR exerts a significant influence on plant development, mainly affecting the roots in their initial period, when they are very susceptible to compacted soil layers.

However, productivity for these two harvest years presented a small difference between the clusters generated (Figure 8). In other words, the differences found between the delimited regions did not influence productivity in general.

In turn, the soil physical-chemical variables were the most important for the generation of clusters in the last two harvest years studied (2015/2016 and 2016/2017). However, there was variation across the harvest years in relation to which soil physical-chemical variables were more relevant in differentiating the study area.

For the 2015/2016 harvest year, the most important soil chemical elements for the differentiation of clusters were carbon (C) and copper (Cu) contents. Distribution of these elements on the map (Figure 8) is similar to the one corresponding to productivity (Figure 11) when compared in

Table 1. Evaluation of group formation using cluster fit metrics. Values in bold represent the best results for each index, relative to the various subsets under study.

		CROP YEAR 2013/2014			CROP YEAR 2014/2015		
		SUBSET			SUBSET		
INDEX	CLUSTER	ALL	VI	PC	ALL	VI	PC
	4	0.139	0.132	0.095	0.14	0.13	0.03
ASC	3	0.158	0.156	0.088	0.13	0.15	0.04
	2	**0.249**	0.205	0.079	**0.25**	0.20	0.09
	4	0.714	0.726	0.640	**0.22**	0.60	0.50
Cindex	3	0.710	0.714	0.643	0.48	0.59	0.52
	2	0.639	**0.596**	0.659	0.46	0.51	0.52
	4	9.466	7.161	**2.155**	3.40	8.77	12.5
SD	3	4.582	3.805	2.678	3.15	5.14	10.9
	2	4.190	3.400	3.426	**2.30**	3.82	7.47
	4	**0.898**	0.884	0.709	0.63	**0.91**	0.87
DUNN	3	0.886	0.861	0.705	0.62	0.88	0.87
	2	0.773	0.741	0.696	0.62	0.78	0.85
	4	1.861	1.604	1.266	1.02	1.90	3.05
DB	3	1.026	0.948	1.484	1.12	1.19	2.79
	2	**0.704**	0.915	1.917	**0.75**	0.91	1.69
CCC	-	**0.86**	0.78	0.77	0.81	**0.83**	0.73

ASC: Average Silhouette Coefficient; Cindex: C index; SD: Standard Deviation index; CCC: Cophenetic Correlation Coefficient; DUNN: Dunn index; DB: Davies-Bouldin index; VI: vegetative index; PC: soil physicochemical variables; Nº CLUSTER: number of clusters formed within the study area.

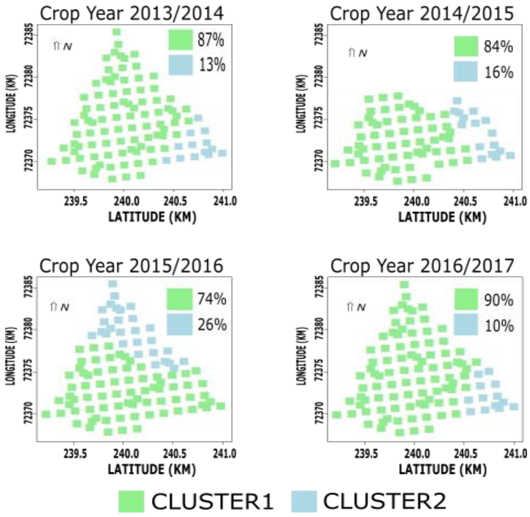


Fig 1. Maps with their sampling points according to their respective cluster and percentage of occupation of each cluster for the 2013/2014, 2014/2015, 2015/2016 and 2016/2017 crop years, CLUSTER1: cluster 1; CLUSTER2: cluster 2.

Table 2. Evaluation of group formation using cluster fit metrics. Values in bold represent the best results for each index, relative to the various subsets under study.

		CROP YEAR 2015/2016			CROP YEAR 2016/2017		
		SUBSET			SUBSET		
INDEX	CLUSTER	ALL	VI	PC	ALL	VI	PC
	4	0.05	0.07	0.06	0.21	0.06	0.13
ASC	3	0.06	0.10	0.05	0.24	0.07	0.13
	2	**0.25**	0.15	0.09	0.28	**0.39**	0.27
	4	0.53	0.54	**0.39**	0.51	0.56	0.52
Cindex	3	0.52	0.55	0.56	0.50	0.58	0.47
	2	0.42	0.43	0.55	0.41	**0.40**	0.49
	4	5.64	4.92	4.87	3.66	4.92	3.83
SD	3	4.54	3.33	6.22	2.50	4.26	3.28
	2	3.45	**3.11**	4.41	2.37	2.53	**2.33**
	4	0.37	0.55	0.80	0.65	0.40	0.77
DUNN	3	0.35	0.53	0.78	0.62	0.37	0.72
	2	0.52	0.42	**0.82**	0.49	0.66	**0.81**
	4	1.50	1.34	1.75	0.97	1.30	1.10
DB	3	1.43	**0.87**	2.10	0.78	1.21	0.98
	2	0.89	0.92	1.66	0.83	0.73	**0.67**
CCC	-	0.61	0.51	**0.78**	0.74	0.65	**0.88**

ASC: Average Silhouette Coefficient; Cindex: C index; SD: Standard Deviation index; CCC: Cophenetic Correlation Coefficient; DUNN: Dunn index; DB: Davies-Bouldin index; VI: vegetative index; PC: soil physicochemical variables; Nº CLUSTER: number of clusters formed within the study area.

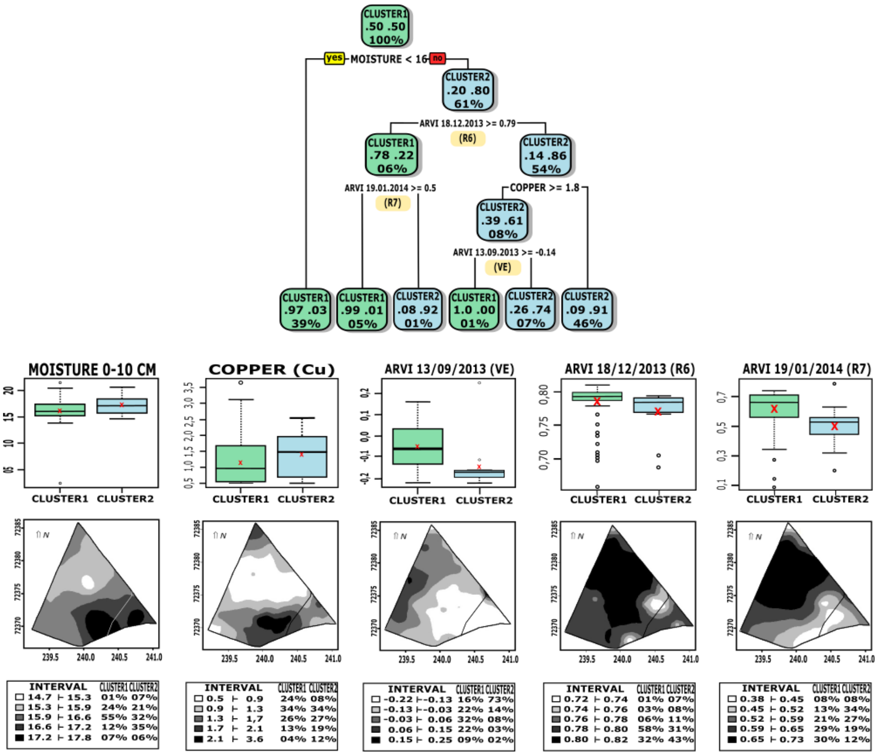


Fig 2. Decision tree for the subset formed by all variables (crop year 2013/2014), Boxplot and thematic maps with the attributes indicated by the tree as those that most contributed to the generation of clusters. The line on the map indicates the division between the clusters.

Table 3. Descriptive statistics of productivity data

	2013/2014	2014/2015	2015/2016	2016/2017
Minimum	2.90	1.87	2.14	1.58
Maximum	5.76	3.18	2.82	4.21
Average	4.22	2.37	2.44	3.12
Standard deviation	0.58	0.27	0.18	0.53
CV	13.74%	11.55%	7.57%	17%

CV: Coefficient of variation

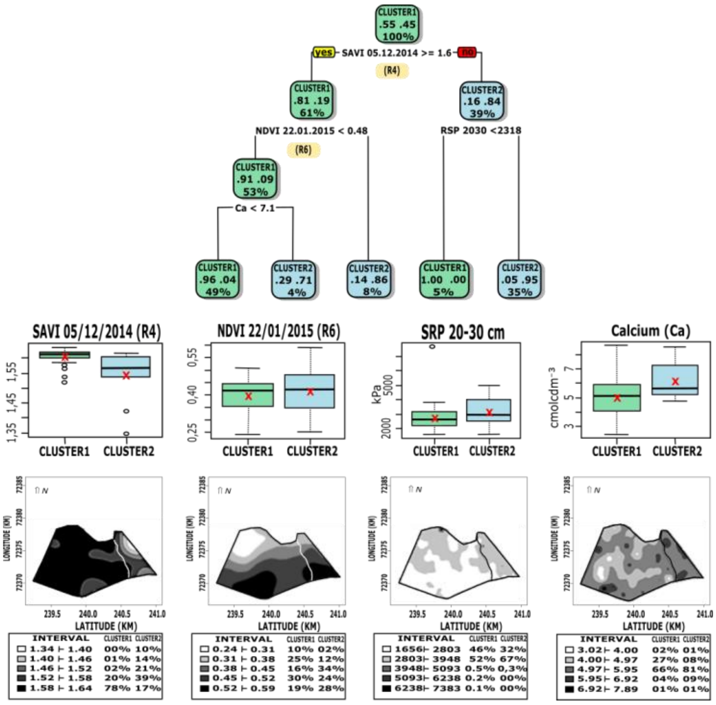


Fig 3. Decision tree for the subset formed by all attributes (crop year 2014/2015), Boxplot and thematic maps with the attributes indicated by the tree as the that most contributed to the generation of clusters. The line on the map indicates the division between the clusters.

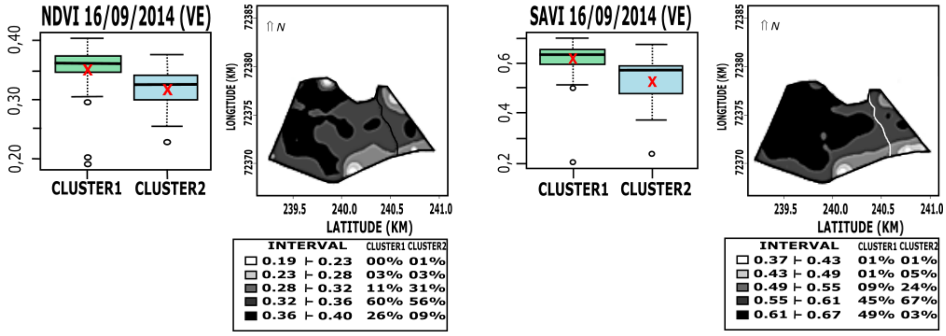


Fig 4. Boxplot and thematic maps generated for the NDVI and SAVI index for the plant emergence stage (VE) (crop year 2014/2015). The line on the map indicates the division between the clusters.

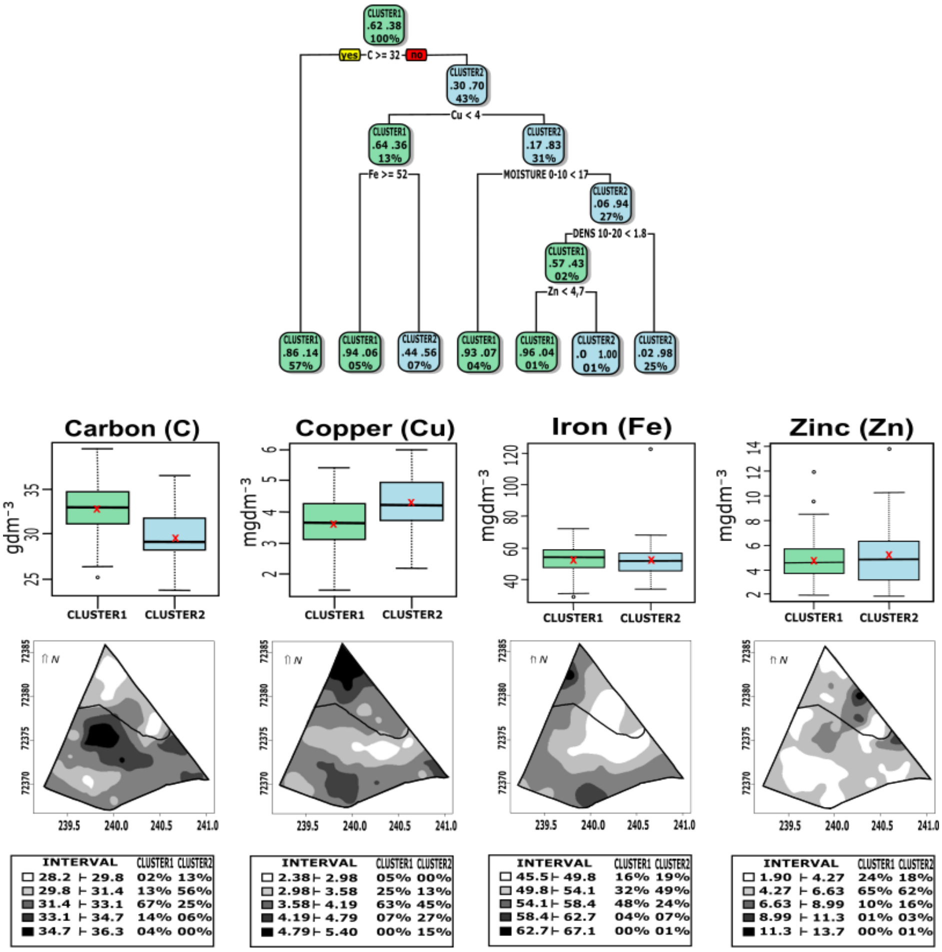


Fig 5. Decision tree for the physicochemical attributes (crop year 2015/2016), Boxplot and thematic maps with the chemical attributes indicated by the tree as those that most contributed to the generation of clusters. The line on the map indicates the division between the clusters.

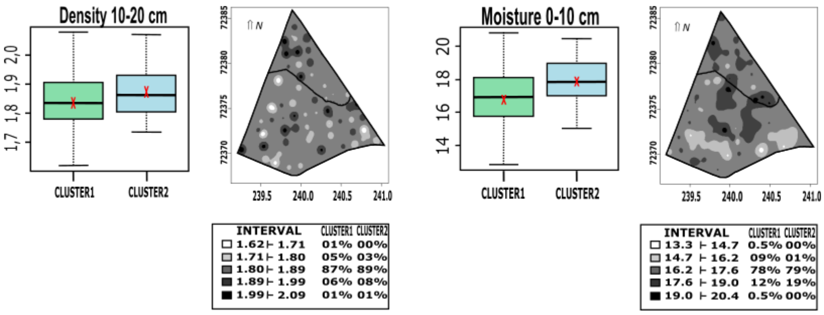


Fig 6. Boxplot and thematic maps with the physical attributes pointed out by the tree as those that most contributed to the generation of clusters. The line on the map indicates the division between the clusters

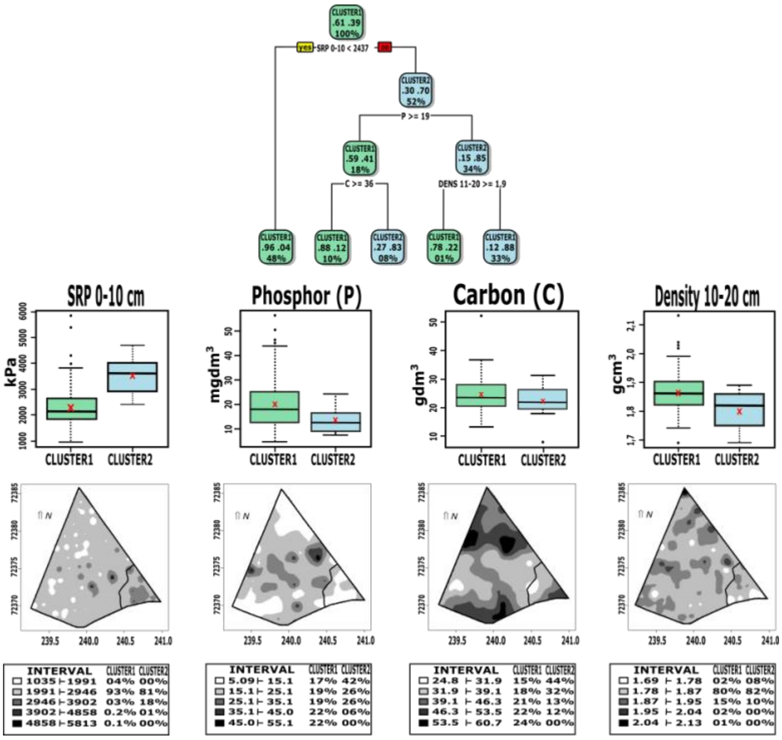


Fig 7. Decision tree for the physicochemical attributes (crop year 2016/2017), boxplot and thematic maps with the attributes indicated by the tree as the most contributed to the generation of clusters. The line on the map indicates the division between the clusters.

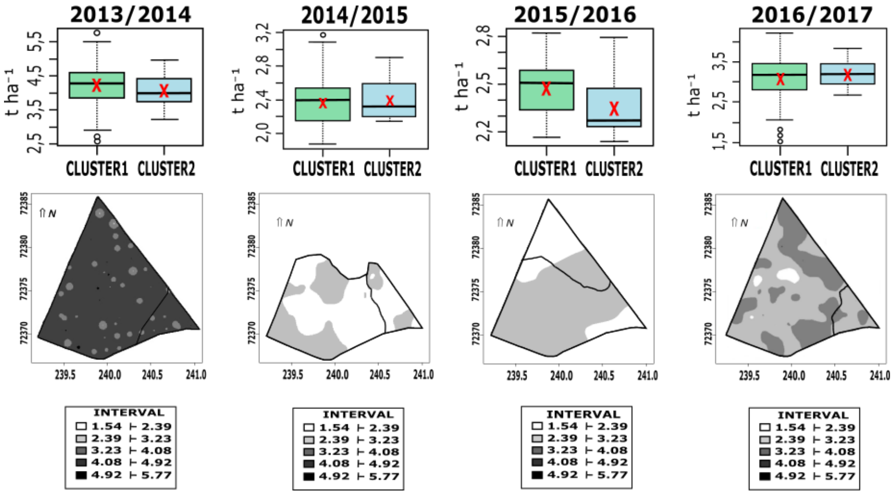
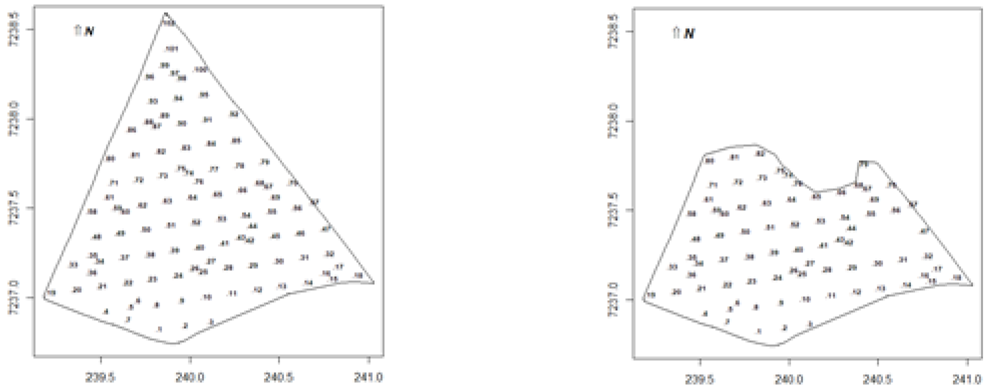


Fig 8. Comparative productivity map for all crop years analyzed in the study. The line on the map indicates the division between the clusters.



(a) (b)

Fig 9. Representation of the plot with the respective numbers of sampling points for the crop years (a) 2013/2014, 2015/2016 and 2016/2017 and (b) 2014/2015.

relation to the clusters formed; in other words, there seems to be a relationship between these elements and productivity, which was not so evident in the other harvest years.

According to the classification proposed by SEAB (1989), carbon (C) content in the soil has concentrations considered high (from $20\,\mathrm{g\,dm^{-3}}$ to $35\,\mathrm{g\,dm^{-3}}$) in both clusters, although there are areas within CLUSTER1 with concentrations considered very high ($> 35\,\mathrm{g\,dm^{-3}}$) for that variable. In turn, copper (Cu) content in the soil presented average concentrations ($> 1\,\mathrm{mg\,dm^{-3}}$) for both clusters.

Carbon (C) content in the soil plays an important role in the terrestrial ecosystem, being directly associated with soil fertility and, therefore, with productive capacity (Guo et al., 2015). In this study, high availability of this variable was verified in both clusters, according to the soybean needs (SEAB, 1989). In other words, even with different concentrations of this element in the field, this variation takes place within the range considered high for soybean, so that it is available for full development of this crop and, thus, not limiting it.

In turn, copper (Cu) content in the soil plays a fundamental role in the biochemistry and physiology of plants, so that lack or low concentration of this element causes a decrease in productivity (Malavolta, 2006). As its concentration is considered average for both clusters (SEAB, 1989), this element is also not a limiting factor for productivity in CLUSTER2; in other words, likewise to carbon, the variation between the clusters takes place within the limits necessary for plant development. Thus, it is verified that the soybean productivity variability between both clusters formed for this harvest year can be related to other factors not analyzed in this study, such as climatic factors, for example, which may also be similarly affecting the spatial distribution of carbon and copper content. For the 2016/2017 harvest year, the main variables responsible for dividing the area into two clusters were SPR (0-10 cm) and phosphorus (P) content in the soil. According to the results presented, SPR (0-10 cm) has higher soil penetration resistance in the layer from 0 to 10 cm deep in CLUSTER 2 when compared to the rest of the area. Changes in soil structure exert effects on SPR, which in turn influence root and seedling growth (Botta et al., 2006). The SPR increase makes the energy required for root development to be higher, in addition to reducing root elongation and growth (Lipiec and Hatano, 2003). However, although there are differences in the distribution of the values corresponding to this variable between the clusters, it did not prove to be limiting in relation to productivity, as it did not show relevant differences between both clusters generated (Figure 8) for this harvest year. This result is in agreement with Girardello et al. (2014), which establish that, in adequate environmental situations, the relationship between SPR and productivity has been low.

The results obtained show that, according to the classification by SEAB (1989), the concentration for the phosphorus (P) content variable is considered very high ($> 9\,\mathrm{mg\,dm^{-3}}$) for both clusters. Phosphorus (P) content is an essential element for the development of plants, which cannot reach their maximum potential without adequate nutritional supply of this chemical element (Marschner, 1995). However, in our case, according to the soybean needs, high availability of this element was verified in both clusters and even with variations between the different regions delimited in this study, which takes place within this availability range, therefore not causing any effect on productivity.

It is noted that the regional differences found in the field derived from the different sets of variables had little agreement with the productivity data, especially for the 2013/2014, 2014/2015 and 2016/2017 harvest years. This low agreement can be associated with the little relationship between the variation of the different variables used in this paper and the variation in productivity. This absence of relationship between different variables and productivity can also be seen in some other papers, such as the one presented by Stafford et al. (1996), who analyzed the relationship between the variation of soil nutrient contents and the variation in productivity. In another study, when evaluating four case studies related to problems associated with the analysis of agricultural data, Wendroth et al. (2001) observed that there was no spatial association between the productivity maps and the vegetation index and soil index maps (which include several soil variables). The homogeneity seen in the productivity maps also evidences that the variability of the different variables present in the different clusters exerted little influence on productivity spatial variation. This homogeneity can be related to management and handling of the area where the study was carried out, as it is a commercial agricultural area with technical monitoring for several years. According to Kayad et al. (2021), field management and environmental factors are the main causes of productivity spatial and time variability and, according to Freddi et al. (2006), soil management is the main factor for its variability.

Even though there was no major effect on soybean productivity in the harvest years evaluated, it was possible to identify regional differences in the study area, mainly characterized by vegetative indices and some soil physical-chemical variables. Understanding these regional differences is important to establish appropriate management practices, not only in relation to productivity optimization, but also to minimize possible environmental damage (Alves et al., 2013).

Materials and Methods

Study area

The study was carried out with data collected in a grain production commercial agricultural area with 167.35 ha (Figure 9a) for the 2013/2014, 2015/2016 and 2016/2017 harvest years and 124.22 ha (Figure 99b) for the 2014/2015 harvest year, located in the municipality of Cascavel, state of Paraná - Brazil. The mean altitude in this area is 650 m. This region has its soil classified as typical dystroferric red oxisol (Santos et al., 2018) and presents super-humid mesothermic temperate climate, Cfa climate type (Köppen), with a mean annual temperature of 21ºC. For 2014/2015, part of the study area was used (Figure 9b), as the region to the North lacked values for soybean productivity and was disregarded from the paper.

Data acquisition

In order to better understand the methodology proposed, the stages that make up this paper were organized according to the flowchart described in Supplementary Figure 1. Information for the 2013/2014, 2014/2015, 2015/2016 and 2016/2017 harvest years obtained through field collection and remote sensing was used. The sampling grid consisted of 102 points and followed a sampling plan called "Lattice plus close pairs" (Chipeta et al., 2017) (Figure 9a) with the exception of the 2014/2015 harvest year, in which 77 points were used (Figure 9b) because the missing points were planted with corn that year. From these sampling grids (a) and (b) in Figure 9, the following sets of soil variables were obtained through field collection: soil penetration resistance (SPR), density, soil moisture, soil chemical variables - Carbon (C) (g dm^{-3}), Phosphorus (P) (mg dm^{-3}), Potassium (K) (cmolc dm^{-3}), potential hydrogen (pH), Aluminum (Al) (cmolc dm^{-3}), potential acidity (H+Al) (cmolc dm^{-3}), Calcium (Ca) (cmolc dm^{-3}), Copper (Cu) (cmolc dm^{-3}), Iron (Fe) (cmolc dm^{-3}), Magnesium (Mg) (cmolc dm^{-3}), Manganese (Mn) (cmolc dm^{-3}), Zinc (Zn) (cmolc dm^{-3}) and soybean productivity (t ha^{-1}). In addition to that, vegetation indices (VIs) were calculated through sensors using the images obtained from the Landsat-8 satellite OLI sensor. The indices used in this study were the following: ARVI, NDVI, EVI2, SAVI, OSAVI and WDRVI (Huete, 1988; Steven, 1988; Gitelson, 2004; Jiang et al., 2008).

Geostatistical analysis and selection of variables

For each harvest year, a geostatistical analysis was carried out on each variable, in order to observe the spatial dependence behavior and subsequently generate the thematic map through ordinary kriging. The parameters of the geostatistical models were estimated by the maximum likelihood method. The Matérn family theoretical models were estimated with model order parameter k equal to 0.5 (exponential model), 1 and 2 and with $\mathrm{k} \rightarrow \infty$ (Gaussian model) (Uribe-Opazo et al., 2012). Choice of the best fitted model was through cross-validation and Akaike's criterion (Faraco et al., 2008). The estimate of an index (RNE) that evaluates the spatial dependence degree was also calculated (Cambardella et al., 1994). Once the spatial dependence degree was known, the variables that presented RNE only classified as strong ($\mathrm{RNE} > 0.75$) or average ($0.25 \leq \mathrm{RNE} \leq 0.75$) spatial dependence were selected.

For a more detailed study of the area, the variables with spatial dependence (average or strong) were grouped into three subsets according to their characteristics: (1st subset) containing all the variables (ALL), (2nd subset) vegetative indices (VIs); and (3rd subset) soil physical-chemical (PC) variables; with the objective of finding out if any of these subsets presented better performance in characterizing the study area. Subsequently, for each of these subsets, their dimensionality was reduced using the MULTISPATI-PCA technique (Dray et al., 2008), with which each subset of variables was transformed into synthetic variables called spatial principal components (SPCs), from which the score values for each sampling point were obtained. The number of SPCs that represented at least 70% of the total variability of the variables was used (Gavioli et al., 2016).

Generation of the dissimilarity matrix and data clustering

Subsequently, a dissimilarity matrix used to perform the clustering was generated. For this, a methodology created by Fouedjio (2016) (Supplementary Figure 2) was employed, in which a spatial dissimilarity measure was elaborated through calculation of non-parametric, univariate and crossed experimental semivariances, considering a non-parametric core-estimator function. With the set of estimated values corresponding to the direct and cross semivariances, a dissimilarity measure between two locations, $\mathbf{s}_\mathrm{k}$ and $\mathbf{s}_\mathrm{l}$, denoted by $\mathrm{d}_\lambda(\mathbf{s}_\mathrm{k}, \mathbf{s}_\mathrm{l})$, was obtained according to Supplementary Figure 2 (Theodoridis and Koutroumbas, 2009). From this dissimilarity measure, the $\mathbf{D}$, nxn symmetric dissimilarity matrix was generated for all sampled locations (Fouedjio, 2016). After obtaining dissimilarity matrix $\mathbf{D}$, a cumulative full link algorithm was used to perform the clusters, from which interpolated maps with 2, 3 and 4 clusters were generated. For each harvest year, choice of the subset that presented the best clustering results (which best defined the different partitions), as well as choice of the optimal number of clusters in the area under study, was performed by calculating the following measures: cophenetic correlation coefficient (CCC), average silhouette coefficient (ASC), Cindex, SD index, Dunn index (DUNN) and DB index (Xiao et al., 2017; Mota et al., 2018; Halkidi et al., 2000; Rousseeuw, 1987).

Statistical data analysis and classification

The analysis of the profile corresponding to the different clusters and the analysis of the relationship between the variables belonging to the subset selected and the clusters chosen for each harvest year, as well as the evaluation between the variables that most contributed to regionalization of the area and productivity, were carried out through boxplots, thematic maps (generated by kriging) and decision trees (Witten et al., 2011).

In the decision tree, the clusters generated by the subset that best clustered the data were used as dependent variable; whereas the explanatory variables corresponded to those that were part of the subset chosen to generate these divisions in the field. Due to the small number of sample points to be used in the decision tree, the sample set considered in such tree corresponded to a grid with interpolated points (8,244 points for the 2013/2014, 2015/2016, 2016/2017 harvest years and 5,218 points for 2014/2015) through ordinary kriging.

Subsequently, this set was randomly divided into 70%, which corresponded to a sample for training, and the other 30% for testing. There was also balancing of the clusters formed and pruning in the decision trees to generate smaller and easy-to-interpret trees. The classification was evaluated by analyzing the confusion matrix between the elements classified and the test and training data (Congalton and Green, 1999).

Conclusion

The methodology applied was able to identify regional differences caused by the variability of the different variables used in the research; however, this variation took place within a range not limiting productivity which, in turn, showed homogeneity throughout the study area and, therefore, few differences between the clusters formed. Even so, the fact that local characteristics were identified was important to better understand the influence exerted by the various factors studied on the field.

Acknowledgments

This paper was carried out with the support of the Coordination for the Improvement of Higher Education Personnel (CAPES), Funding Code 001; National Council for Scientific and Technological Development (CNPq), Laboratory of Applied Statistics (LEA-UNIOESTE) and Laboratory of Spatial Statistics (LEE-UNIOESTE).

References

Alves SM, Alcântara GR, Reis EF, Queiroz DM, Valente DSM (2013) Definição de zonas de manejo a partir de mapas de condutividade elétrica e matéria orgânica. Bioscience Journal. 29(1): 104-114.

Alvino FCG, Aleman CC, Filgueiras R, Althoff D, Da Cunha FF (2020) Vegetation indices for irrigated corn monitoring. Engenharia Agrícola. 40: 322–333. doi:10.1590/1809-4430-Eng.Agric.v40n3p322-333/2020.

Behera SK, Mathur RK, Shukla AK, Suresh K, Prakash C (2018) Spatial variability of soil properties and delineation of soil management zones of oil palm plantations grown in a hot and humid tropical region of southern India. Catena. 165: 251-259. doi: 10.1016/j.catena.2018.02.008.

Botta GF, Jorajuria D, Balbuena R, Ressia M, Ferrero C, Rosatto H, Tourn M (2006) Deep tillage and traffic effects on subsoil compaction and sunflower (Helianthus annus L.) yields. Soil and Tillage Research. 91:164-170. doi: 10.1016/j.still.2005.12.011.

Burdett H, Wellen C (2022) Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precision Agriculture. 23(5): 1553-1574. doi: 10.1007/s11119-022-09897-0.

Cambardella CA, Moorman TB, Novack JM, Parkin TB, Karlen DL, Turco RF, Knopka AE (1994) Field-scale variability of soil proprieties in central Iowa soils. Soil Science Society America Journal. 58: 1240-1248. doi: 10.2136/sssaj1994.03615995005800050033x.

Cherubin MR, Damian JM, Tavares TR, Trevisan RG, Colaço AF, Eitelwein MT, Molin JP (2022) Precision Agriculture in Brazil: The Trajectory of 25 Years of Scientific Research. Agriculture. 12(11): 1882. doi: 10.3390/agriculture12111882.

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

Collares GL, Reinert DJ, Reichert JM, Kaiser DR (2006) Qualidade física do solo na produtividade da cultura do feijoeiro num Argissolo. Pesquisa Agropecuária Brasileira, 41:1663-1674. doi: 10.1590/S0100-204X2006001100013.

Congalton RG, Green K (2019) Assessing the accuracy of remotely sensed data: principles and practices. 3rd ed. New York: Lewis Publisher.

Cordeiro APA, Berlato MA, Fontana DC, Melo RW, Shimabukuro YE, Fior CS (2017) Regiões homogêneas de vegetação utilizando a variabilidade do NDVI. Ciência Florestal, 27: 883-896. doi: 10.5902/1980509828638.

Dalchiavon FC, Rodrigues AR, De Lima ES, Lovera LH, Montanari R (2017) Variabilidade espacial de atributos químicos do solo cultivado com soja sob plantio direto. Revista de Ciências Agroveterinárias. 16(2): 144-154. doi: 10.5965/223811711622017144.

Deiss L, Kleina GB, Moraes A, Franzluebbers AJ, Motta AC, Dieckow J, Sandini IE, Anghinoni I, Carvalho PC (2020) Soil chemical properties under no-tillage as affected by agricultural trophic complexity. European Journal of Soil Science. 71(6):1090-1105.

Deus ACF, Büll LT, Guppy CN, Santos SDMC, Moreira LLQ (2020) Effects of lime and steel slag application on soil fertility and soybean yield under a no till-system. Soil Tillage Res. 196:104422.

Dray S, Saïd S, Débias F (2008) Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. Journal of Vegetation Science. 19: 45-56. doi: 10.3170/2007-8-18312.

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

Fouedjio FA (2016) hierarchical clustering method for multivariate geostatistical data. Spatial Statistics. 18: 333-351. doi: 10.1016/j.spasta.2016.07.003.

Freddi OS, Carvalho MP, Veronesi Júnior V, Carvalho GJ (2006) Produtividade do milho relacionada com a resistência mecânica à penetração do solo sob preparo convencional. Engenharia Agrícola. 26:113-121. doi:10.1590/S0100-69162006000100013.

Gavioli A, Souza EG, Bazzi CL, Guedes LPC, Schenatto K (2016) Optimization of management zone delineation by using spatial principal Componentes. Computers and Electronics in Agriculture. 127: 302-310. doi: 10.1016/j.compag.2016.06.029.

Gavioli A, de Souza EG, Bazzi CL, Schenatto K, Betzek NM (2019) Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. Biosystems engineering. 181: 86-102. doi: 10.1016/j.biosystemseng.2019.02.019.

Girardello VC, Amado TJC, Santi AL, Cherubin MR, Kunz J, Teixeira TG (2014) Resistência à penetração, eficiência de escarificadores mecânicos e produtividade da soja em Latossolo Argiloso manejado sob plantio direto de longa duração. Revista Brasileira de Ciência do Solo. 38: 1234-1244. doi: 10.1590/S0100-06832014000400020.

Gitelson, AA (2004) Wide Dynamic Range Vegetation Index for remote quantification of biophysical characteristics of vegetation. Journal of Plant Physiology. 161: 165-173. doi: 10.1078/0176-1617-01176.

Grable AR, Siemer EG (1968). Effects of bulk density, aggregate size, and soil water suction on oxygen diffusion, redox potential and elongation of corns roots. Soil Science Society of America Proceedings. 32:180-186. doi: 10.2136/sssaj1968.03615995003200020011x.

Gülser C, Ekberli I, Candemir F (2016) Spatial variability of soil physical properties in a cultivated field. Eurasian Journal of Soil Science. 5(3): 192-200. doi: 10.18393/ejss.2016.3.192-200.

Guo PT, Li MF, Luo W, Tang QF, Liu ZW, Lin ZM (2015) Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma. 237-238: 49-59. doi: 10.1016/J.GEODERMA.2006.07.002.

Halkidi M, Vazirgiannis M, Batistakis Y (2000) Quality Scheme Assessment in the Clustering Process. In: Zighed DA, Komorowski J, Żytkow J (eds) Principles of Data Mining and Knowledge Discovery. PKDD 2000. Lecture Notes in Computer Science. Berlin: Springer.

Huete ARA (1988) soil-adjusted vegetation index (SAVI). Remote Sensing Of Environment**.** 25: 295-309. doi:10.1016/0034-4257(88)90106-X.

Jiang Z, Huete AR, Didan K, Miura T (2008) Development of a two-band enhanced vegetation index without a blue band. Remote Sensing of Environment. 112: 3833-3845. doi: 10.1016/j.rse.2008.06.006.

Kayad A, Sozzi M, Gatto S, Whelan B, Sartori L, Marinello F (2021) Ten years of corn yield dynamics at field scale under digital agriculture solutions: a case study from North

Italy. Computers and Electronics in Agriculture. 185. doi: 10.1016/j.compag.2021.106126.

Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D (2018) Machine learning in agriculture: A review. Sensors. 18(8): 2674. doi:10.3390/s18082674.

Lipiec J, Hatano R (2003) Quantification of compaction effects on soil physical properties and crop growth. Geoderma. 116:107- 136. doi: 10.1016/S0016-7061(03)00097-1.

Malavolta E (2006). Manual de nutrição mineral de plantas. Agronômica. 1ed. São Paulo: Ceres.

Malvezi KED, Zanão Jr. LA, Guimaraes EC, Vieira SR, Pereira N (2019) Soil chemical attributes variability under tillage and no-tillage in a long-term experiment in southern Brazil. Bioscience Journal. 35(2):467-476.

Marinković J, Bjelić D, Ninkov J, Vasin J, Tintor B, Živanov M (2018) Effect of different soil usage on microbial properties in soils of Central Serbia. Ratarstvo i povrtarstvo/Field and Vegetable Crops Research. 55(2): 58-64. doi: 10.5937/ratpov55-15307**.**

Marschner H (1995) Mineral nutrition of higher plants. 2ed. London: Academic Press.

Metwally MS, Shaddad SM, Liu M, Yao RJ, Abdo AI, Li P, Chen X (2019) Soil properties spatial variability and delineation of site-specific management zones based on soil fertility using fuzzy clustering in a hilly field in Jianyang, Sichuan, China. Sustainability. 11(24): 7084. doi: 10.3390/su11247084.

Mota VC, Damasceno FA, Leite DF (2018) Fuzzy clustering and fuzzy validity measures for knowledge discovery and decision making in agricultural engineering. Computers and Electronics in Agriculture. 150: 118–124. doi: 10.1016/j.compag.2018.04.011.

Mwendwa SM, Mbuvi JP, Kironchi G, Gachene CK (2022). Assessing spatial variability of selected soil properties in Upper Kabete Campus coffee farm, University of Nairobi, Kenya. Heliyon. 8. doi: 10.1016/j.heliyon.2022.e10190

Nyéki A, Daróczy B, Kerepesi C, Neményi M, Kovács AJ (2022) Spatial variability of soil Properties and its effect on maize yields within field – a case study in Hungary. Agronomy. 12: 395. doi: 10.3390/agronomy12020395.

Oliveira DG, Reis EF, Medeiros JC, Martins MPO, Silva AU (2018) Correlação espacial de atributos físicos do solo e produtividade de tomate industrial. Revista Agro@mbiente on-line. 12(1): 1-10. doi: 10.18227/1982-8470ragro.v12i1.4211.

Pimentel-Gomes F (2009) Curso de estatística experimental. 15ed. Piracicaba: Fealq.

Priya KCB, Venkateswari S (2018). Delineation of management zones in precision agriculture using different clustering algorithms. International Journal of Applied Engineering Research. 13(22):15951-15955.

Rodriguez D, Fitzgerald GJ, Belford R, Christensen L (2006) Detection of nitrogen deficiency in wheat from spectral reflectance indices and basic crop ecobiophysiological concepts. Australian Journal of Agricultural Research. 57: 781-89.doi: http://dx.doi.org/10.1071/ AR05361.

Rosemary F, Indraratne SP, Weerasooriya R, Mishra U (2017) Exploring the spatial variability of soil properties in an Alfisol soil catena. Catena. 150: 53-61. doi: 10.1016/j.catena.2016.10.017.

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal Computational Applied Mathematics. 20: 53–65. doi: 10.1016/0377-0427(87)90125-7.

Sánchez TJD, Ligarreto MGA, Leiva FR (2011) Spatial variability of soil chemical properties and its effect on crop yields: a case study in maize (Zea mays L.) on the Bogota Plateau. Agronomía Colombiana. 29(2): 456-466.

Santos HG, Jacomine PT, Anjos LHC, Oliveira VÁ, Lumbreras JF, Coelho MR, Araujo Filho JO, Oliveira JB, Cunha TJF (2018) Brazilian soil classification system. 5th ed. Brasilia: Embrapa.

Santos Júnior AB, Silveira Júnior O, Lima ÍCS, Nunes ME, Santos AC, Faria AFG (2021) Variabilidade espacial dos atributos químicos do solo sob diferentes usos agrícolas no ecótono cerrado - amazônia. Agri-environmental sciences. 7(1):16. https://doi.org/10.36725/agries.v7i1.5224.

Seab (1989) Manual Técnico do Subprograma de Manejo e Conservação do Solo. Secretaria de Estado da Agricultura e do Abastecimento do Paraná. 1ed. Curitiba.

Stafford JV, Ambler B, Lark RM, Catt J (1996) mapping and interpreting yield variation in cereal crops. Computer and Electronics in Agriculture. 14: 101-119. doi: 10.1016/0168-1699(95)00042-9.

Steven MD (1998) The sensitivity of the OSAVI vegetation index to observational parameters. Remote Sensing of Environmental. 63: 49-60. doi: 10.1016/S0034-4257(97)00114-4.

Theodoridis S, Koutroumbas K (2009) Pattern Recognition. 4ed. London: Academic Press.

Uribe-Opazo MA, Borssoi J, Galea M (2012) Influence diagnostics in Gaussian spatial linear models. Journal of Applied Statistics. 39(3): 615–630. doi:10.1080/02664763.2011.607802.

Vian AL, Santi AL, Amado TJC, Cherubin MR, Simon DH, Damian JM, Bredemeier C (2016) Variabilidade espacial da produtividade de milho irrigado e sua correlação com variáveis explicativas de planta. Ciência Rural. 46: 464-471. doi: 10.1590/0103-8478cr20150539.

Wendroth O, Jürchiik P, Kersebaun KC, Reuter H, Van Kessel C, Nielsen DR (2001) Identifying, understanding, and describing spatial processes in agricultural landscapes - four case studies. Soil and Tillage Research. 58: 113-27. doi:10.1016/S0167-1987(00)00162-8.

Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. 3rd ed. Califórnia: Elsevier Science.

Xiao J, Lu J, Li X (2017). Davies Bouldin Index based hierarchical initialization K-means. Intelligent Data Analysis. 21: 1327-1338. doi: 10.3233/IDA-163129.

Zheng H, Chen L, Han X, Zhao X, Ma Y (2009). Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. Agriculture, Ecosystems and Environment. 132(1-2): 98-105. doi: 10.1016/j.agee.2009.03.004.