Brazilian Journal of Cartography (2017), Nº 69/5, Special Issue GEOINFO 2017: 923-935
Brazilian Society of Cartography, Geodesy, Photgrammetry and Remote Sense
ISSN: 1808-0936

GEOINFO
Brazilian Symposium on GeoInformatics
November 27th to 30th, Campos do Jordão, SP, Brazil

USING ENSEMBLES WITH SPATIAL CLUSTERING APPROACHES APPLIED IN THE DELINEATION OF MANAGEMENT CLASSES IN PRECISION AGRICULTURE

Utilizando Ensembles com Abordagens de Agrupamento Espacial para o Delineamento de Classes de Manejo em Agricultura de Precisão

Eduardo Antonio Speranza[1] & Ricardo Rodrigues Ciferri[2]

[1]Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA
Embrapa Informática Agropecuária
Av. André Tosello, 209, Campus da Unicamp, Barão Geraldo, CEP 13083-886, Campinas, SP, Brasil
eduardo.speranza@embrapa.br

[2]Universidade Federal de São Carlos – UFSCar
Departamento de Computação
Rod. Washington Luís, km 235, Caixa Postal 676, CEP 13565-905, São Carlos, SP, Brasil
ricardo@dc.ufscar.br

Received on February 21, 2017/ Accepted on Maio 16, 2017
Recebido em 21 de Fevereiro, 2017/ Aceito em 16 de May, 2017

ABSTRACT

This paper describes experiments performed using different approaches for spatial data clustering, aiming to assist the delineation of management classes in Precision Agriculture (PA). These approaches were established from the partitional clustering algorithm *Fuzzy c-Means* (*FCM*), traditionally used in PA, and from the hierarchical clustering algorithm *HACC-Spatial*, especially designed for PA. We also performed experiments using different clustering ensembles approaches, evaluating their behavior to achieve consensus solutions from individual clusterings obtained from attribute splitting or using exclusively *FCM* or *HACC-Spatial*. The achieved results exhibited some differences between *FCM* and *HACC-Spatial*, mainly for the visualization of management classes in the form of maps. The *HACC-Spatial* algorithm achieved, in general, better results when compared to *FCM* and ensembles approaches. Regarding the consensus clusterings provided by ensembles, we can point out the attempt to achieve agreement results which most closely matches the original clusterings, decreasing or increasing the stratification of the management classes maps.

Keywords: Precision Agriculture, Management Classes, Spatial Data Clustering, Ensembles.

RESUMO

Este artigo descreve experimentos realizados utilizando diferentes abordagens para agrupamento de dados espaciais, com o objetivo de auxiliar no delineamento de classes de manejo em Agricultura de Precisão (AP). Essas abordagens foram estabelecidas a partir do algoritmo de agrupamento particional *Fuzzy c-Means* (*FCM*), tradicionalmente utilizado em AP, e do algoritmo de agrupamento hierárquico *HACC-Spatial*, especialmente desenvolvido para AP. Também foram realizados experimentos utilizando diferentes abordagens de *ensembles* para agrupamentos disponíveis na literatura, avaliando o seu funcionamento para obter soluções de consenso para agrupamentos individuais obtidos a partir do

particionamento do conjunto de atributos ou da utilização exclusiva do *FCM* ou do *HACC-Spatial*. Os resultados obtidos mostraram algumas diferenças entre o *FCM* e o *HACC-Spatial*, principalmente com relação a visualização das classes de manejo em forma de mapas. O algoritmo *HACC-Spatial*, alcançou, de uma maneira geral, melhores resultados quando comparado ao *FCM* e as abordagens de *ensembles*. Levando-se em consideração os agrupamentos consensuais obtidos pelas abordagens de *ensembles*, ficou evidente a tentativa de se obter resultados concordantes que se aproximam das soluções fornecidas pelos agrupamentos originais, proporcionando o aumento ou a diminuição da estratificação dos mapas de classes de manejo.

Palavras-chave: Agricultura de Precisão, Classes de Manejo, Agrupamento de Dados Espaciais, Ensembles.

1. INTRODUCTION

Precision Agriculture (PA) is an agricultural management system driven by spatio-temporal variability of soil and culture attributes of a crop. These parameters may be obtained from particular procedures and techniques based on information technology, remote sensing and Global Positioning System (GPS) (MOLIN, 2003; VENDRUSCULO & KALEITA, 2011). Unlike conventional agriculture, where agricultural inputs and correctives are evenly applied across the cultivation area, PA enables its users to manage them in a site-specific way, allowing farmers to fit crop needs and supply of inputs (SCHWALBERT *et al.*, 2014). Therefore, the main aim of PA is to increase yield in a sustainable way, reducing the environmental impacts with the site-specific use of agricultural inputs and, consequently, increasing the profit (BERNARDI *et al.*, 2014).

Because of its highly dependency of the spatio-temporal variability built-in data collected on the field, the adoption of decision-making processes based on PA suggests data collection at high spatial resolutions. However, this usually is not possible for most farmers, because several factors such as the high cost of acquiring satellite images and gathering data on the field, beyond the need to acquire services and automated machinery able to perform variable rate interventions. In these cases, the delineation of subfields spatially internal to the crop area, which the internal spatial variability is so negligible as to allow for evenly distributed internal interventions, is a way to disseminate the adoption of PA even using accurate spatial resolutions (e.g. between 10 and 30 meters). These subfields, known as management classes, may be composed by one or many spatially contiguous areas in the coordinate space, known as management zones (TAYLOR *et al.*, 2007).

Taking into account these concepts, it is really intuitive to relate the delineation of management classes with traditional clustering algorithms, such as Fuzzy c-Means (FCM) (BEZDEK *et al.*, 1984). However, PA tasks produce complex and non-conventional data, composed by two distinct spaces: attribute space, regarding the events occurring in the crop; and coordinate space, regarding the spatial location where these events took place. Thereby, because of its complexity, the coordinate space must to be handled in different ways by clustering algorithms. With the purpose of solving this challenge, Ruß and Kruse (2011) developed an agglomerative hierarchical clustering algorithm, known as HACC-Spatial. The HACC-Spatial enables the delineation of management classes preserving the spatial contiguity as much as possible, in order to facilitate easy visual interpretation of the user while maintain the coherence of the clustering obtained by events related to soil and plants.

Using algorithms composed by different attributes and parameters, such as FCM and HACC-Spatial, to solve the delineation of management classes in PA, may generate different results and hence questions regarding which of them is the best solution. In order to clarify such questions, several approaches enabling consensual and more robust clusterings have emerged in the literature. These clusterings, known as ensembles, must be obtained from different ways, such as individual clusterings using different kinds of algorithms, parameters configurations or subsets of attributes at the same data set (GHOS & ACHARYA, 2011).

In a preliminary version of this work, clustering ensembles approaches based on graph and hypergraph partitioning (STREHL &

GHOSH, 2002) were evaluated in their ability to provide more robust management classes maps regarding both the individual clusterings from FCM and HACC-Spatial algorithms using all available attribute space, and the same clustering algorithm spliting the attribute space (SPERANZA *et al.*, 2016). Here, we extend this evaluation performing new experiments, using a more recent and simple clustering ensemble approach, based on evidence accumulation obtained across individual clusterings (FRED; JAIN, 2005). Therefore, it was possible to achieve a more complete evaluation on which situations each approach should be used, either individually or using clustering ensembles.

The remainder of this paper is structured as follows. In section 2, we describe the FCM and HACC-Spatial algorithms and approaches commonly used to delineate management classes in PA, beyond the ensemble approaches used in this work and in its preliminary version. In section 3, we present the methodology used for the experiments. In section 4, we present results for experiments using real data. Finally, in section 5, we present our conclusions and provide suggestions for future work proposals.

This paper is based on Speranza *et al.* (2016), previously presented in XVII Brazilian Symposium on Geoinformatics (http://www.geoinfo.info/).

2. BACKGROUND AND RELATED WORK

Some clustering approaches have been used to assist the delineation of management classes in PA. Nevertheless, most of the approaches available in the literature use the FCM algorithm as a basis for this task. Based on the standard clustering algorithm k-means (MACQUEEN *et al.*, 1967), the FCM (BEZDEK *et al.*, 1984) calculates, at each iteration, the membership degree (ω) of each data sample i with respect to each cluster j, for j varying from 1 to K, where K should be defined by the end user (Equation (1)).

$$\omega_{(i,j)} = \frac{1}{\sum_{k=1}^{K}\left[\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}\right]} \quad (1)$$

In Equation 1, m is a fuzzification parameter, defined by the user with default value 2, K is the number of desired clusters, c_j is the centroid of cluster j and c_k is the centroid of the cluster K, also for K varying from 1 to K. At the end of each iteration, the centroids of each cluster j are recalculated, taking into account all N dataset samples and their membership values for the respective cluster (Equation 2).

$$c_j = \frac{\sum_{i=1}^{N}\omega(i,j)^m * x_i}{\sum_{i=1}^{N}\omega(i,j)^m} \quad (2)$$

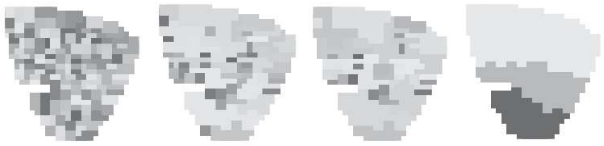
Instead of k-means, FCM convergence results not only assign each sample to a unique cluster (non-overlapping clustering), but also in a membership matrix with 0 to 1 values for each sample with respect to each cluster, known as fuzzy partition matrix (overlapping clustering). This matrix is one of the FCM advantages regarding non-overlapping clustering algorithms, by providing better results for situations having difficult separation and overlapping datasets. However, like k-means, in the original version of FCM the centroids are randomly initialized. While this feature helps to reduce the computational cost of running FCM, it can also make the results susceptible to a local minima. Consequently, FCM may provide different results to the end user for different runs using the same parameters, which allows us to classify it as a non-deterministic algorithm.

The main reason for using FCM in the context of the delineation of management classes in PA is linked with the fact that abrupt changes do not occurs in soil and plant attributes in small enough parcels of the crop, causing input data and the obtained clusters to consider a membership degree. Over the years, several approaches in the literature using FCM and considering different types of these attributes have been developed. Brock *et al.* (2005) used FCM to delineate management classes considering historical yield data from corn-soybean rotation crops, indentifying the spatial association of the obtained maps with soil maps. Kitchen *et al.* (2005) used FCM to

delineate management classes considering ratios of soil electrical conductivity (EC) in different depths (bulk of EC) and relief data, comparing them with yield classes obtained from historical yield data. As a result, it was found that the bulk of EC combined with relevant data are strong indications of management classes. Similar conclusions were obtained by Morari *et al.* (2009), including measures of soil and electrical resistivity data. The work of Li *et al.* (2007) used, in addition to abovementioned attributes, features indicating rates of organic matter and biomass. In this case, due to the large number of attributes, an intermediate phase of principal component analysis was performed before getting the management classes by FCM. High-resolution satellite images also appears as inputs to obtain management classes using the FCM, such as in works by Song *et al.* (2009) and Zhang *et al.* (2009). Milne *et al.* (2012) used FCM to find management classes from smoothed spatial data obtained from three different methods. The results were compared with crop responses regarding the application of different nitrogen rates. The work of Scudiero *et al.* (2013), using FCM to obtain management classes, shows that combined bare-soil and EC data can contribute to find spatial variability of a crop. The KM-sPC approach (CORDOBA *et al.*, 2013) allowed to show the importance of principal component analysis considering the coordinate space to reduce the stratification provided by FCM when management classes are displayed in the form of maps. This approach were used again in a pratical nitrogen management of wheat (PERALTA *et al.*, 2015). The study by Chang *et al.* (2014) compared management classes generated by FCM using reflectance data regarding the soil properties and productivity, showing that it is feasible to use an active canopy sensor for this PA application.

Despite the widespread use of FCM for this task, the coordinate space of PA datasets composed by spatial coordinates variables (e.g., latitude and longitude) have been used only to show the management classes provided by clustering in the form of maps. This fact does not block the use of these maps by automated machinery for variable rate interventions, but can reduce the spatial contiguity, causing stratification of management classes in too many management zones which can confuse visual analysis by experts. The approach proposed by Cordoba *et al.* (2013) attempts to reduce the effect of the contiguity loss by treating the coordinate space during the preprocessing of the data. However, although it is possible to achieve better results using this approach rather than traditional FCM, it is still very difficult for the end user to differentiate management classes in a visual way.

In order to solve this kind of problem, Ruß and Kruse (2011) proposed the HACC-Spatial hierarchical clustering algorithm. This approach takes into account spatial restrictions for clustering samples, and considers a preprocessing step to perform an initial tessellation of them in small spatial clusters, using the k-means algorithm at the coordinate space. Such subdivision aims to reduce computational costs by decreasing the number of steps of the construction of the hierarchical tree (or dendrogram) produced by the algorithm, considering the geostatistics principle that very spatially close samples tend to have close enough values in attribute space (MATHERON, 1963). As a result, a structure similar to a Voronoi diagram should be obtained in the preprocessing step (Figure 1a). From this moment, each dendrogram step merges the most similar clusters, according to the feature space. First, only spatially adjacent clusters can be merged, providing the maintenance of spatial contiguity (Figure 1b). However, when a user-defined contiguity threshold cp is reached, this restriction is switched off and from this point on non-adjacent clusters can also be merged (Figure 1c). This threshold is associated with the ratio of the average distances between the samples belonging to adjacent clusters and the average distances between samples belonging to non-adjacent clusters. At the end of this run, it is expected that HACC-Spatial will provide maps of management classes as contiguous as possible, regarding the parameters values provided by the end user (Figure 1d).



(a) (b) (c) (d)

Fig. 1 - Clusterings obtained running HACC-Spatial, represented in the form of management classes maps, in sequencial dendrogram steps: (a) initial tesselation; (b) 60 clusters before reaching the *cp* threshold; (c) 30 clusters after reaching the *cp* threshold; (d) 3 clusters, representing useful management classes in practice.

Because of the distinct nature of FCM and HACC-Spatial (partitional and hierarchical, respectively) and the spatial restrictions used by the second algorithm, distinct clustering results for the same dataset are expected, making it difficult for the user to choose the best approach. A feasible solution to solve this question can be achieved by using ensembles. Ensembles are able to combine multiple sample clusterings in a unique and consolidated one, known as consensus solution. These kind of approach can be used to meet several requirements, such as: increase the quality of the solution, by providing more robust clusterings; select models; reuse knowledge; find consensus between clusterings obtained from subsets of features or subsamples; among others (GHOS & ACHARYA, 2011).

The main aim of a clustering ensemble is to find a consensus solution composed by an unique clustering to share as much information as possible derived from original clusterings. This sharing can be measured by the average of normalized mutual information (ANMI), where the desired optimal value is ANMI equal to 1 (STREHL & GHOSH, 2002). The main goal of the three ensembles algorithms developed by Strehl and Ghosh (2012) is to build general approaches to obtain consensus from individual clusterings aiming at maximizing the ANMI value. These algorithms were evaluated by the authors in scenarios where individual clusterings were composed by distinct features, distinct subsamples or distinct clustering algorithms. The Cluster-based Similarity Partitioning Algorithm (CSPA) is the simplest and with the most obvious heuristic. It is based on the fact that two samples have a similarity of 1 if they are in the same cluster and 0 otherwise. Thus, a n x n binary matrix, where n is the number of samples, is created for each original clustering. To recluster these samples, a similarity-based clustering algorithm based on graph partitioning is used (KARYPI & KUMAR, 1998). The HyperGraph Partitioning Algorithm (HGPA) addresses the clustering ensemble as a hypergraph partitioning problem, where hyperedges represent the original clusters as indication of strong bonds. To recluster the samples, a partitional hypergraph algorithm, cutting a minimal number of hyperedges is used (HAN *et al.*, 1997). In this case, while CPSA only considers pairwise relationships, HGPA includes original clustering relationships. Finally, the Meta-Clustering Algorithm (MCLA) represents each cluster by a hyperedge, and then group and collapse related hyperedges (or clusters), attaching each sample to the collapsed hyperedge in which it belongs more actively. At the end, a graph-based clustering of hyperedges is performed, indentifying consolidated clusters of clusters. According to Strehl and Ghosh (2002), the MCLA tends to provide better ANMI values when the consensus solution were obtained from individual clusterings with low noise rates and diversity; and HGPA and CSPA are usually better were obtained from individual clusterings with high noise rates and diversity.

Despite their effectiveness of obtain robust clusterings, the CPSA, HGPA and MCLA approaches are dependent of graph and hypergraph partitioning complex algorithms run by external software. In this way, and with the aim of extending the experiments and analyzes performed in our previous work (SPERANZA *et al.*, 2016), here a simplified concept of clustering ensembles was used, based on evidence accumulation. This concept treats each original clustering as an independent evidence of data organization. Fred and Jain (2005) developed an approach based on this concept, where the original clusterings are combined by a voting mechanism, building a new similarity matrix known as co-association matrix. Equation 3 defines the co-association matrix calculation for n samples grouped by N different original clusterings.

$$M(i,j) = \frac{s_{ij}}{N} \quad (3)$$

In Equation 3, sij represents the number of times the pair of samples (i,j), for i different from j and i and j less than or equal to n, is associated with the same cluster, considering the N original clusterings. Next, this matrix is used as input for an hierarchical clustering algorithm which regroups the samples in order to obtain a more robust result.

According to experiments performed by Fred and Jain (2005), the evidence accumulation approach presents better performance, in general, than the approaches based on partitioning of graphs and hypergraphs, mainly regarding the ability of this approach to recognize clusters with arbitrary forms in the attribute space, making it possible to use it in data sets with well-correlated attributes.

From the concepts described in this section, we extend the experiments developed in our previous work (SPERANZA *et al.*, 2016), now comparing the use of the clustering ensemble approaches based on graph and hypergraph partitioning with the evidence accumulation approach in generating more robust clusterings for the delineation of management classes in PA.

3. METHODOLOGY

The methodology used in our experiments follows the concepts of Knowledge Discovery in Databases (KDD). According to Fayyad *et al.* (1996) and Weiss and Indurkhya (1998), at least three main steps of KDD process should be taken into account when it will be used: preprocessing, data mining (or pattern extraction) and post processing. The planned activities for each one of these steps, in the context of management classes in PA, are described below.

3.1 Preprocessing

The preprocessing step comprises the changes that should be made in a raw dataset when it will be used by a KDD process, preparing it to the next steps. Regarding spatial data, in addition to very common preprocessing activities, such as standardization, cleaning and feature selection, the spatial interpolation must be performed in order to accommodate data samples in a single and regular spatial grid (VIEIRA, 2000). This activity is required, because PA datasets are caught using different kinds of sensors and samples densities, usually at distinct spatial spots in the same area. Other important activities in this step are: verifying data distribution using probabilistic density functions, as a preassessment of possible distortions that can occur in clustering algorithms when using non-Gaussians distributed features; verifying features correlations, using methods such as Pearson's Coefficient Correlation (BENESTY *et al.*, 2009); and data standardization, reducing the bias caused by features with highly predominant scales relative to the others.

3.2 Data Mining

The data mining step can be viewed as an iterative process, where different solutions should be used to improve the accuracy of the results. In the context of our work, due to the fact that datasets had no previous classification, clusterings tasks need to be considered. Therefore, the approaches to be used are classified as non-supervised machine learning algorithms (MITCHELL, 1997). In this step, we used the HACC-Spatial and FCM algorithms in the traditional way and also combining results with ensembles. HACC-Spatial was run using non-spatial attributes of the whole dataset to calculate dissimilarity values at each step of dendrogram, and spatial attributes to build the initial tessellation and to support adjacency treatments at each step of dendrogram (Approach I). On other hand, FCM was run in its traditional way, i.e., using only non-spatial attributes (Approach II). Regarding ensembles, new approaches were created to found consensus clusterings from individual results provided by Approach I and Approach II, using both the graph and hypergraph partitioning algorithms (Approach III-A) and the evidence accumulation algorithm (Approach III-B). In the same way, other four approaches were created to fiund consensus clustering from individual results provided by attributes subsets of soil, altimetry and yield using HACC-Spatial (Approaches IV-A and IV-B) and FCM (Approaches V-A and V-B). Regarding the ensembles approaches run using the graph and hypergraph partitioning algorithms (Approaches III-A, IV-A and V-A), the chosen the result was the one which achieved the best values of ANMI considering different runs using CSPA, HGPA and MCLA.

According to expert domain users, at least 2 and at most 5 management classes should be considered for a crop (MOLIN *et al.*, 2015). Thereby, the eight abovementioned approaches were run using k=2 to 5 clusters for the experiments, when using FCM (partitional), and the same values for dendrogram cuts, when using HACC-Spatial (hierarchical). Regarding to dissimilarity measures, the Euclidean distance were used for all approaches. In relation to other parameters and customizations, for approaches using FCM, the standard fuzzification value m=2 was fixed, and samples were associated with the cluster where a higher membership degree was achieved. HACC-Spatial parameters, initial tessellation number of clusters (k) and cp, were defined during the experiments.

3.3 Post Processing

Finally, in the post processing step, we used two internal validation criteria: the SD criteria and the silhouette width criteria. These criteria allow comparing and evaluating the effectiveness of the eight approaches when they are run at the same number of clusters. The SD criteria (HALKIDI *et al.* 2000; HALKIDI; VAZIRGIANNIS, 2001) allows to verify, for each obtained clustering, how cohesive and well separated are the clusters, from average values of intra-cluster variance and distances between clusters centroids. In this case, optimal values should be closer to 0. The silhouette width criteria (ROUSSEEUW, 1987) follows the same principles of SD, but using dissimilarity values of a sample regarding its associated cluster and the nearest neighbor cluster. In this case, values closer to 1 indicates that the sample has been allocated to the correct cluster; and values closer to -1 indicates that the sample should have been better allocated to the nearest neighbor cluster. According to Vendramim et. al (2010), the silhouette width criteria, in comparison to other internal criteria in the literature, can provide in general, a more effective assessments about the internal structure of the clusters.

4. EXPERIMENTS

In this section, we present the results obtained from experiments using real data, following the methodology described in section 3 and extending the results achieved in Speranza *et al.* (2016). These data are composed by samples collected on an experimental crop field of sugarcane culture. This field has an area around 17 hectares belonging to Fazenda Aparecida, located in Mogi-Mirim, São Paulo state, Brazil, with central coordinates 7505136N (latitude) and 299621E (longitude), in the spatial reference system UTM Zone 23S.

The raw datasets used in our work comprises measures of soil electrical conductivity (EC) in milisiemens per meter; altimetry quota, in meters; and historical yield, in tons per hectare or culms per square meter. The samples were collected at different times and by different sensors or processes, providing us six conventional attributes associated with spatial coordinates: soil electrical conductivity at 30 and 90 cm depth in 2010 (EC30 and EC90); altimetry quota (Quota); and historical yield in 2010 (Yield2010), 2012 (Yield2012) and 2013 (Yield2013). It is worth mentioning the need for historical yield data, because they could be susceptible to climatic factors over the years. In addition, the rainfall data of the whole farm in the agricultural years should be considered to support some analysis: 1601 mm in 2010 (July 2009 to June 2010), 1538 mm in 2012 (July 2011 to June 2012) and 1599 mm in 2013 (July 2012 to June 2013). The probabilistic density distribution of EC30, EC90 and Yield2010 attributes could be described by Gaussians, with most values around the mean. On the other hand, the distributions of Yield2012 and Yield2013 indicates, respectively, predominance of higher and lower yield values, probably affected by the climatic factors. A special case occurs with the Quota attribute, where average values are the minority because the experimental area has a slight slope and is narrow in the central region. These distributions are shown in Figure 2.

Applying the Pearson's Coefficient Correlation between pairs of attributes, it was verified that EC30 and EC90 hold the most positive correlation of the dataset. In general, the Quote attribute was well correlated with all other attributes, and negatively (oppositely) correlated with Yield2010. Regarding yield data, Yield2012 and Yield2013 attributes are highly correlated, and negatively correlated with Yeld2010 attribute. The negative correlation of Yield2010 with other yield years could be the influence again of climatic factors.

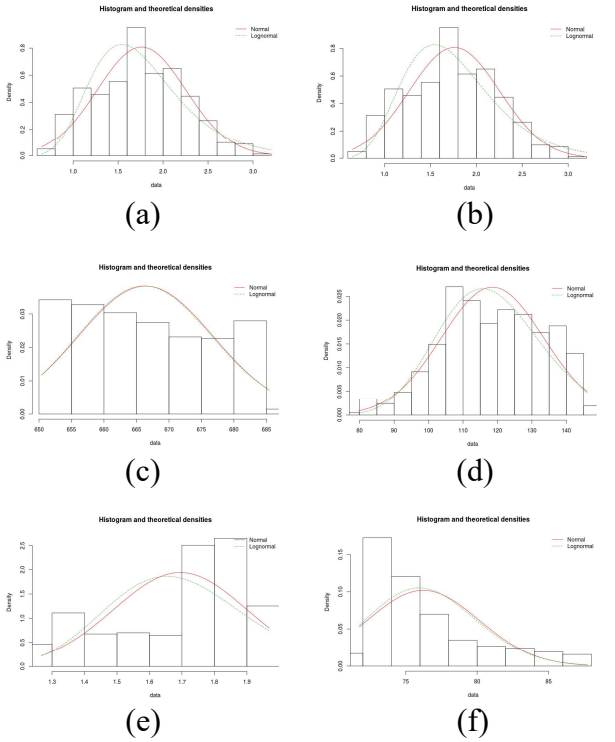


Fig. 2 - Probabilistic density distributions of dataset attributes: (a) EC30; (b) EC90; (c) Quote; (d) Yield2010; (e) Yield2012; e (f) Yield2013.

Using the concepts of preprocessing described above, the dataset features were interpolated in a single regular spatial grid with spatial resolution of 20 meters. This value was calculated using the average coordinates spacing between samples for each one of the six features of the original data set. Simple algorithms, such as the average of k nearest neighbors (ALTMAN, 1992), were used to interpolate attributes with higher sample densities. On the other hand, more sophisticated algorithms, such as kriging (MATHERON, 1969), were used to interpolate attributes with smaller sample densities. After applying this process, each dataset feature was distributed in 415 samples spatially represented by points with latitude and longitude coordinates. Figure 3 shows raw samples of soil electrical conductivity (high density) and yield (medium density) and their respective interpolated samples in the same regular spatial grid. Lower values are represented by lighter shades of gray, while higher values are represented by darker shades of gray.

Especially for the HACC-Spatial algorithm, when it was run in the context of approaches I, III-A and IV-A, the cp parameter was set to 0.5, according to the best results obtained by Ruß and Kruse (2011). Initial tessellation (k) was set to 200, after checking a significant increase in internal variance of the clusters for the following levels of the dendrogram.

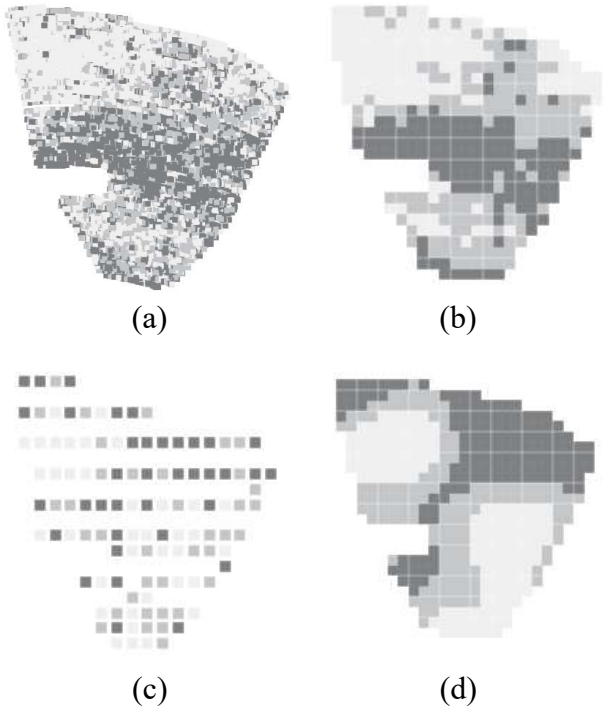
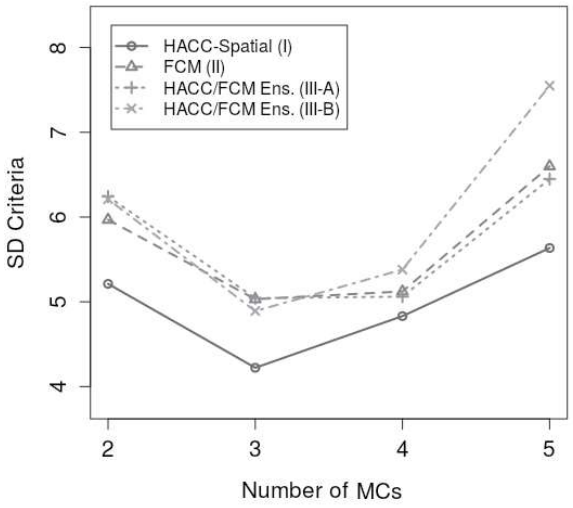


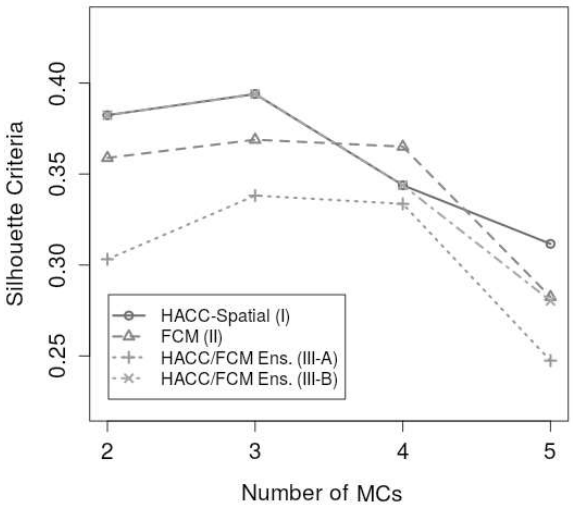
Fig. 3 - Example of raw and interpolated data in 3 classified intervals: (a) EC30 raw data (9046 samples); (b) EC30 interpolated data (415 samples); (c) Yield2010 raw data (111 samples); (d) Yield2010 interpolated data (415 samples).

First, the results achieved with the two approaches of clusterings ensembles used in this paper were qualitatively compared to each other and in relation to the results achieved by the original clusterings using the internal validation criteria SD and silhouette width. Figure 4 shows charts containing the indices of SD and silhouette width criteria achieved by the approaches I, II, III-A and III-B.

As verified in previous experiments, charts from Figure 4 show that ensemble approaches (III-A and III-B) attempt to find consensus solutions for the results obtained by the original clustering approaches. Therefore, it can be observed the trend maintenance in the value of the criteria for clusterings regarding 2 to 5 management classes, with slight variations in values favoring one or other approach. In addition, these results show that HACC-Spatial (Approach I) achieved better performance in general when compared with FCM (Approach II) and ensembles approaches (III-A and III-B).



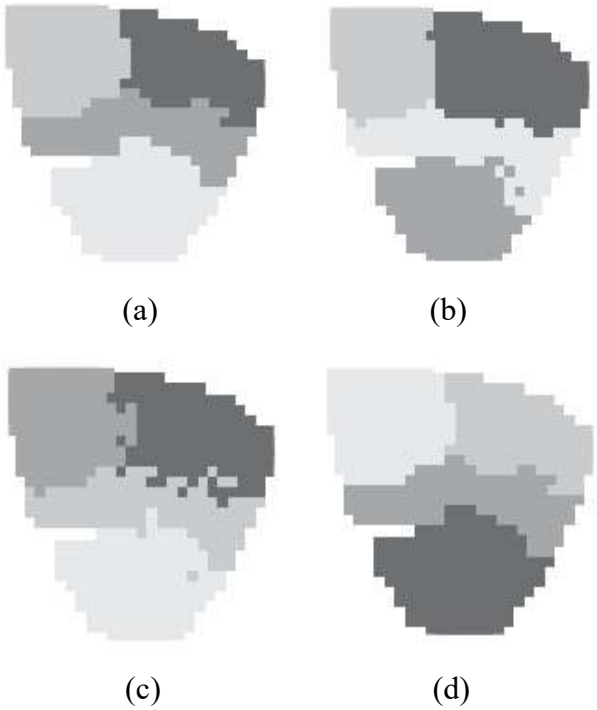
(a)



(b)

Fig. 4 - Indices of SD (a) and silhouette width (b) internal validation criteria achived by clusterings generated using approaches I, II, III-A and III-B.

Completing this first analysis, Figure 5 shows maps with the four management classes generated by these approaches. In this figure, very similar maps obtained by approaches I (Figure 5a) and II (Figure 5b) can be observed, suggesting the use of an ensemble approach. Regarding the ensembles approaches, its easy to observe the influence of FCM on the final result of Approach III-A and the influence of HACC-Spatial on the final result of Approach III-B. Therefore, the Approach III-A expanded the stratification of the map obtained by Approach II, and the Approach III-B took advantage of the best spatial structure of the map from Approach I to generate a spatially better-distributed map. Due to the fact that the validation criteria used in our experiments do not have the ability to evaluate the results regarding both attribute and coordinate space, spatially-stratified maps could obtain better indices than spatially well-distributed maps, as occurred in this case.



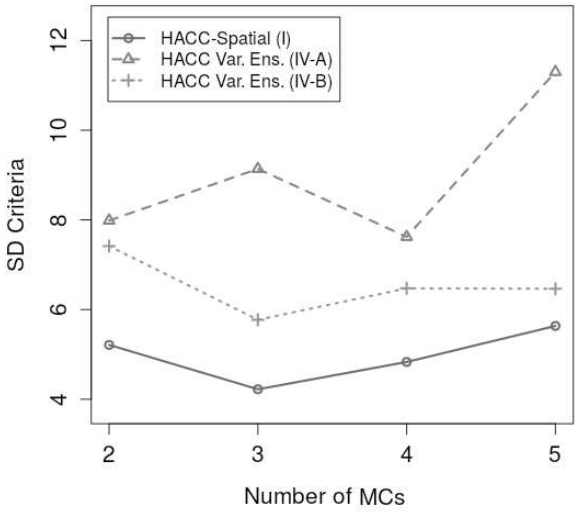
(a) (b)

(c) (d)

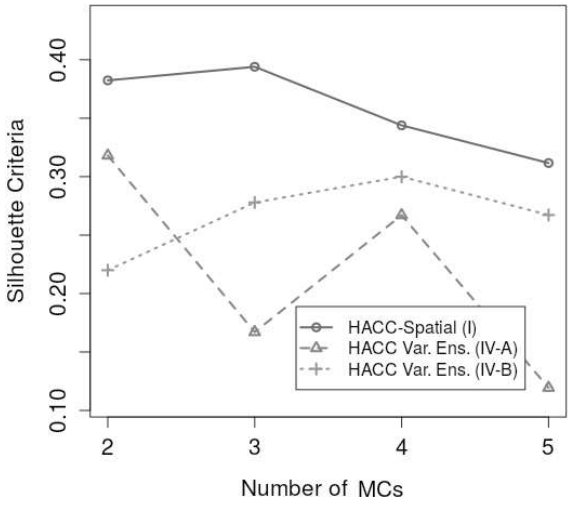
Fig. 5 - 4 Management classes maps generated using approaches I (a), II (b), III-A (c) and III-B (d).

Another experiment was performed using approaches IV-A and IV-B, where original clusterings were obtained using splits of the attribute space and the HACC-Spatial algorithm. Figure 6 shows charts containg the SD and silhouette width indices achieved by these approaches, in comparison with the indices achieved by Approach I.

According to Figure 6, the ensemble approach using evidence accumulation (IV-B) achieved better performance in this experiment than the approach using graph and subgraph partitioning (IV-A), confirming its better ability to take advantage of the better spatial distribution of the results provided by the HACC-Spatial algorithm. Nevertheless, the results achieved by the Approach IV-A can not be considered more robust than those achieved by Approach I, since in the same way as in our previous work, they may cause spatial contiguity loss, harming the analysis of management classes maps by the end users (Figure 7).

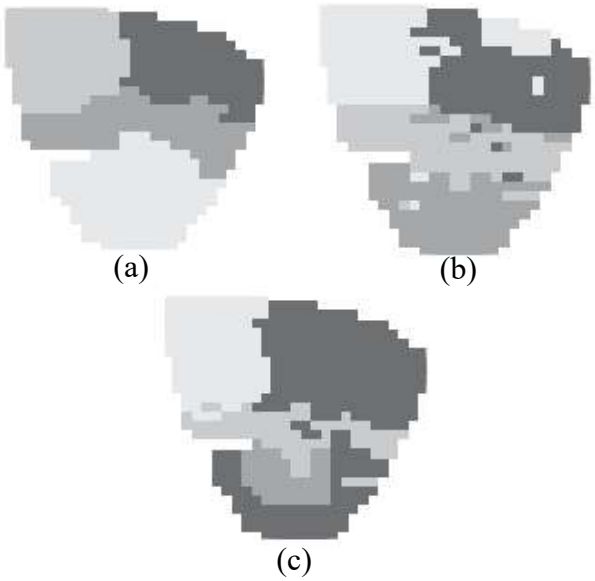


(a)



(b)

Fig. 6 - Indices of (a) and silhouette width (b) internal validation criteria achieved by clusterings generated using approaches I, IV-A and IV-B.



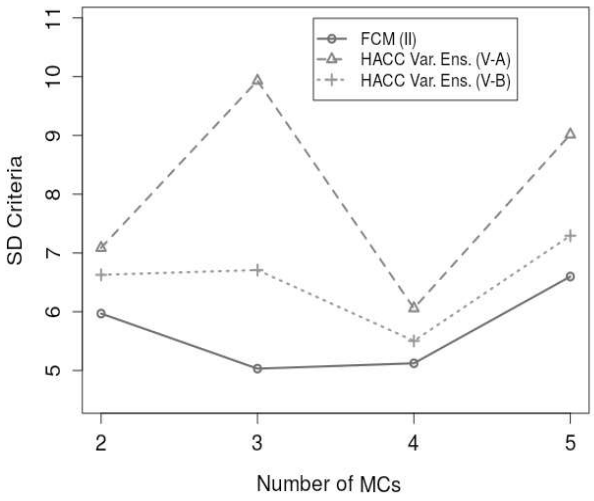
(a) (b)

(c)

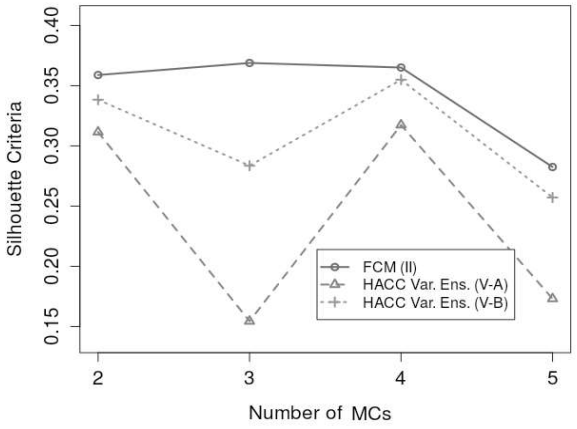
Fig. 7 - 4 Management classes maps generated using approaches I (a), IV-A (b) and IV-B (c).

Finally, an experiment similar to the previous one was performed, now regarding the approaches V-A and V-B, where the original clusterings were obtained using splits of the attribute space and the FCM algorithm. Figure 8 shows charts containing the SD and silhouette width indices achieved by these approaches, in comparison with the indices achieved by Approach II.

In the same way, the ensemble approach using evidence accumulation (V-B) achieved better performance in this experiment than the approach using graph and subgraph partitioning (V-A). From the resulting management class maps shown in Figure 9, the better ability of the Approach V-B to dealing with the stratification issue, in comparison to the Approach V-A, can be noted. In addition, for this case, the Approach V-B was able to obtain management classes maps quite compatible with the Approach II, without generating additional stratification.



(a)



(b)

Fig. 8 - Indices of SD (a) and silhouette width (b) internal validation criteria achieved by clusterings generated using approaches II, V-A and V-B.

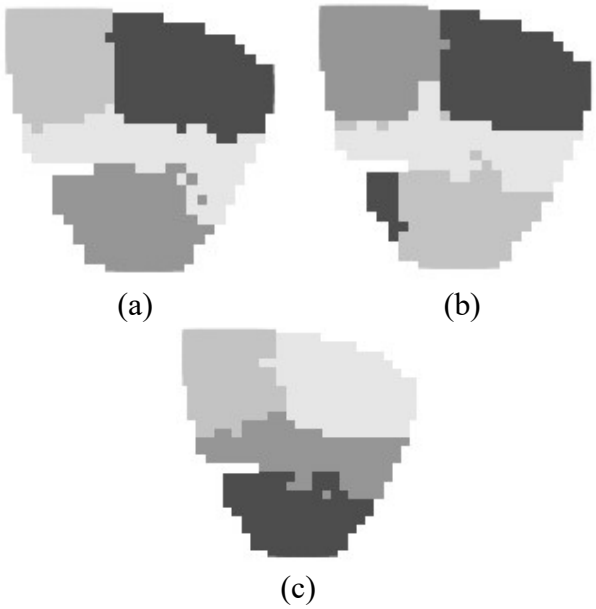


Fig. 9 - 4 Management classes maps generated using approaches II (a), V-A (b) and V-B (c).

5. CONCLUSIONS AND FUTURE WORK

In this paper, new experiments and analysis were performed using clustering ensembles approaches applied in the delineation of management classes in PA. Here, we used an approach based on evidence accumulation and compare the results of the new experiments with the results performed by an approach based on partitioning of graphs and hypergraphs algorithms, already evaluated in a preliminary work.

If we consider maps delineated by traditional clustering approaches using all attributes (approaches I and II), according to both the internal validation criteria and visual analysis, the HACC-Spatial algorithm achieved in general better results compared to FCM algorithm. This finding is different from the preliminary work, and it can be explained by the non-determinism also verified in running HACC-Spatial, which just like FCM, uses random initiation of centroids in the initial tessellation step that may provide different results to the end user using the same values of parameters and attributes.

Regarding the use of ensembles, the results obtained by approaches III-A and III-B show an attempt to obtain consensus clusterings extracting the main features from both algorithms used in obtaining individual clusterings (HACC-Spatial and FCM). In this case, the ensemble approach based on graphs and hypergraphs partitioning (III-A) provided solutions closer to those obtained by the FCM algorithm, causing an increase in stratification. In the other hand, the ensemble approach based on accumulation of evidence (III-B) favored the best spatial arrangement generated by the HACC-Spatial algorithm, providing results with no stratification and easier interpretation by the end user. However, the indices for internal validation criteria achieved by the Approach III-B have not always been better in comparison to the indices achieved by other approaches, evidencing the fact that the internal validation criteria used in this experinent only address issues related to attribute space.

Finally, although the ensembles approaches using subdivision of the attribute space have obtained results worse than those achieved by the traditional algorithms using all attributes, the performed experiments were useful to confirm the skill of the Approach V-B in better dealing with stratification issues than the Approach V-A.

The clustering ensembles approaches used in this work is rather general and try to find consensus clusterings using only final clusterings obtained from splitting of features or from different algorithms. In a future work, new clustering ensembles approaches could be proposed, which allow extracting the main features of each algorithm, such as the membership values provided by FCM, or the contiguity threshold provided by HACC-Spatial. In addition, new internal validation criteria can be proposed to evaluate clusterings regarding the attribute and coordinate space, in order to improve the analysis performed by the end user.

Acknowledgement

The authors thanks Fazenda Aparecida for the availability of the area for data collection that allowed the this study, and to Embrapa researchers, Celia Grego and Luiz Vicente, responsible for collecting the data. We also thank the following Brazilian research agencies: CNPq, CAPES, FAPESP and FINEP. The second author has been supported by the grant 311868/2015-0 from CNPq.

REFERENCES

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician**, v. 46, n. 3, p. 175–185, 1992.

BENESTY, J.; CHEN, J.; HUANG, Y.; COHEN, I. Pearson Correlation Coefficient. In: **Noise Reduction in Speech Processing SE - 5**. Springer Berlin Heidelberg, 2009. v.2 of Springer Topics in Signal Processing, p. 1–4.

BERNARDI, A. C. de C. ; NAIME, J. de M.; RESENDE, A.V. ; INAMASU, R.Y. ; BASSOI, L.H. **Agricultura de Precisão - Resultados de um Novo Olhar**. 1. ed.Brasília: Empresa Brasileira de Pesquisa Agropecuaria, 2014. 596 p.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: The Fuzzy c-Means Clustering Algorithm. **Computer & Geosciences**, v. 10, n. 2-3, p. 191–203, 1984.

BROCK, A.; BROUDER, S. M.; BLUMHOFF, G.; HOFMANN, B. S. Defining Yield-Based Management Zones for Corn-Soybean Rotations. **Agronomy Journal**, v. 97, n. 4, p. 1115–1128, July 2005.

CHANG, D.; ZHANG, J.; ZHU, L.; GE, S. H.; LI, P. Y.; LIU, G. S. Delineation of management zones using an active canopy sensor for a tobacco field. **Computers and Electronics in Agriculture**, v. 109, p. 172–178, 2014.

CÓRDOBA, M.; BRUNO, C.; COSTA, J.; BALZARINI, M. Subfield management class delineation using cluster analysis from spatial principal components of soil variables. **Computers and Electronics in Agriculture**, v. 97, p. 6–14, 2013.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

FRED, A. N. L.; JAIN, A. K. Combining multiple clusterings using evidence accumulation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 6, p. 835–850, 2005.

GHOSH, J.; ACHARYA, A. Cluster ensembles. Wiley Interdisciplinary Reviews: **Data Mining and Knowledge Discovery**, v. 1, n. 4, p. 305–315, 2011.

HALKIDI, M.; VAZIRGIANNIS, M.; BATISTAKIS, Y. Quality scheme assessment in the clustering process. In: ZIGHED, D.; KOMOROWSKI, J.; ZYTKOW, J. (Eds.) . **Principles of Data Mining and Knowledge Discovery**. Springer Berlin Heidelberg, 2000. v. 1910 of Lecture Notes in Computer Science, p. 265–276.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment: finding the optimal partitioning of a data set. In: **Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on**. IEEE, 2001. p. 187-194.

HAN, E.-H.; KARYPIS, G.; KUMAR, V.; MOBASHER, B. Clustering based on association rule hypergraphs. In: **DKMD**, 1997. p. 9-13.

KARYPIS, G.; KUMAR, V. Multilevelk-way partitioning scheme for irregular graphs. **Journal of Parallel and Distributed Computing**, v. 48, n. 1, p. 96–129, 1998.

KITCHEN, N.; SUDDUTH, K.; MYERS, D.; DRUMMOND, S.; HONG, S. Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. **Computers and Electronics in Agriculture**, v. 46, n. 1-3, p. 285–308, Mar. 2005.

LI, Y.; SHI, Z.; LI, F.; LI, H.-Y. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. **Computers and Electronics in Agriculture**, v. 56, n. 2, p. 174–186, Apr. 2007.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, v. 1, p. 281–297.

MATHERON, G. Principles of geostatistics. **Economic geology**, v. 58, n. 8, p. 1246–1266,1963.

MATHERON, G. **Le krigeage universel**. Paris, France, 1969. 84 p.

MILNE, A. E.; WEBSTER, R.; GINSBURG, D.; KINDRED, D. Spatial multivariate classification of an arable field into compact management zones based on past crop yields. **Computers and Electronics in Agriculture**, v. 80, p. 17–30, 2012.

MITCHELL, T. M. **Machine Learning**. New York; London: McGraw-Hill, 1997. 414 p.

MOLIN, J. P. Agricultura de Precisão: Situação atual e perspectivas. In: FANCELLI, A. L.; NETO, D. D. (Eds.) **Milho: Estratégias de**

Manejo para Alta Produtividade. Piracicaba: ESALQ/USP/LPV, 2003. p. 89–98.

MOLIN, J. P.; DO AMARAL, L. R.; COLAÇO, A. **Agricultura de precisão**. Oficina de Textos, 2015. 224 p.

MORARI, F.; CASTRIGNANÒ, A.; PAGLIARIN, C. Application of multivariate geostatistics in delineating management zones within a gravelly vineyard using geo-electrical sensors. **Computers and Electronics in Agriculture**, v. 68, n. 1, p. 97–107, Aug. 2009.

PERALTA, N. R.; COSTA, J. L.; BALZARINI, M.; Castro Franco, M.; CORDOBA, M.; BULLOCK, D. Delineation of management zones to improve nitrogen management of wheat. **Computers and Electronics in Agriculture**, v. 110, p. 103–113, 2015.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, n. 0, p. 53–65, 1987.

RUSS G.; KRUSE, R. Exploratory hierarchical clustering for management zone delineation in precision agriculture. In: PERNER, P. (Ed.) **Advances in Data Mining. Applications and Theoretical Aspects**. Springer Berlin Heidelberg, 2011. v. 6870 of Lecture Notes in Computer Science, p. 161–173.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis**. Journal of Computational and Applied Mathematics**, v. 20, n. 0, p. 53–65, 1987.

SCHWALBERT, R. A.; AMADO, T. J. C.; GEBERT, F. H.; SANTI, A. L.; TABALDI, F. Zonas de manejo: atributos de solo e planta visando a sua delimitação e aplicações na agricultura de precisão. **Revista Plantio Direto**, p. 21–32, 2014.

SCUDIERO, E.; TEATINI, P.; CORWIN, D. L.; DEIANA, R.; BERTI, A.; MORARI, F. Delineation of site-specific management units in a saline region at the Venice Lagoon margin, Italy, using soil reflectance and apparent electrical conductivity. **Computers and Electronics in Agriculture**, v. 99, p. 54–64, 2013.

SONG, X.; WANG, J.; HUANG, W.; LIU, L.; YAN, G.; PU, R. The delineation of agricultural management zones with high resolution remotely sensed data. **Precision Agriculture**, v. 10, n. 6, p. 471–487, 2009.

SPERANZA, E. A.; CIFERRI, R. R.; CIFERRI, C. D. A. Clustering Approaches and Ensembles Applied in the Delineation of Management Classes in Precision Agriculture. In: **XVII Brazilian Symposium on GeoInformatics**. Campos do Jordão: MCTIC/INPE, 2016. p. 152–165.

STREHL, A.; GHOSH, J. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. **Journal of Machine Learning Research**, v. 3, p. 583–617, 2002.

TAYLOR, J. A.; MCBRATNEY, A. B.; WHELAN, B. M. Establishing Management Classes for Broadacre Agricultural Production. **Agronomy Journal**, v. 99, n. 5, p. 1366–1376, Sept. 2007.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. **Statistical Analysis and Data Mining**, v.3, n. 4, p. 209–235, 2010.

VENDRUSCULO, L. G.; KALEITA, A. L. Modeling zone management in precision agriculture through Fuzzy C-Means technique at spatial database. In: . **St. Joseph, MI: American Society of Agricultural and Biological Engineers**, 2011. 17 p.

VIEIRA, S. R. Geoestatistica em Estudos de Variabilidade Espacial do Solo. In: NOVAIS, R. F.; ALVAREZ, V H, S. G. R. (Eds.) **Tópicos em ciência do solo**. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 1. ed., 2000. p. 1–54.

WEISS, S. M.; INDURKHYA, N. **Predictive Data Mining: A Practical Guide**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. 228 p.

ZHANG, X.; SHI, L.; JIA, X.; SEIELSTAD, G.; HELGASON, C. Zone mapping application for precision-farming: a decision support tool for variable rate application. **Precision Agriculture**, v. 11, n. 2, p. 103–114, 2010.