Clustering Approaches and Ensembles Applied in the Delineation of Management Classes in Precision Agriculture

Eduardo A. Speranza[1], **Ricardo R. Ciferri**[2], **Cristina D. A. Ciferri**[3]

[1]Embrapa Agricultural Informatics – Brazilian Research Agricultural Corporation
13083-886 – Campinas – SP – Brazil

eduardo.speranza@embrapa.br

[2]Departament of Computer Science – Federal University of São Carlos
13.565-905 – São Carlos – SP – Brazil

ricardo@dc.ufscar.br

[3]Department of Computer Science – University of São Paulo at São Carlos
13.560-970 – São Carlos – SP – Brazil

cdac@icmc.usp.br

Abstract. *This paper describes an experiment 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 this context, and from the hierarchical clustering algorithm HACC-Spatial, especially designed for this PA task. We also performed experiments using traditional ensembles approaches from the literature, evaluating their behavior to achieve consensus solutions from individual clusterings obtained from features splitting or running one of the abovementioned algorithms. Results showed some differences between FCM and HACC-Spatial, mainly for the visualization of management classes in the form of maps. Considering the consensus clusterings provided by ensembles, it became clear the attempt to achieve an agreement result that most closely matches the original clusterings, showing us some details that may go undetected when we analyse only the individual clusterings.*

1. Introduction

Precision Agriculture (PA) is an agricultural management system driven by spatio-temporal variability of soil and culture features 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 and 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, aiming the maximization of profit cutting of yield limiting factors. Moreover, this system allows farmers to fit crop needs and supply of inputs, helping to reduce the environmental damage [Schwalbert 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 coordinates 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: features, regarding the events occurring in the crop; and coordinates, regarding the spatial location where these events took place. Thereby, because of its complexity, the coordinates 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 features and parameters, such as FCM and HACC-Spatial, to solve clustering problems present in any domain, can 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 been 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 features at the same data set [Ghosh and Acharya 2011]. Our work described in this paper were aimed to evaluate, from internal clustering validation measures, the accuracy of clusterings representing management classes that were obtained individually using the FCM and HACC-Spatial algorithms, as well as using more robust and consensual clustering ensembles to consolidate individual results and feature space partitioning.

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

2. Background and Related Work

Some clustering algorithms have been used to assist the delineation of management zones in PA. Nevertheless, most of the approaches available in the literature use the Fuzzy c-Means algorithm (FCM) as a basis for this task. Based on the standard clustering algorithm k-means [MacQueen et al. 1967], the Fuzzy c-Means algorithm (FCM) [Bezdek et al. 1984] calculates, at each iteration, the membership (ω_k) of each data sample with respect to each one of the desired clusters. This calculation takes into account the distance (d) from any particular data sample to each cluster centroid and a fuzzification parameter (*m*), defined by the user with default value of 2. At the end of each iteration, clusters centroids are recalculated taking into account all dataset samples and their membership values to each cluster. Instead of k-means, FCM convergence results not only to assign each sample to a unique cluster (hard clustering), but in a membership matrix with 0 to 1 values for each sample with respect to each cluster, known as fuzzy partition matrix (soft clustering). This matrix is one of the FCM advantages regarding hard clustering algorithms, providing better results for situations that have a difficult separation and overlapping datasets. However, like k-means, FCM centroids are randomly initialized, making the results susceptible to a local minima.

The main reason for using FCM in the context of this application 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 zones considering historical yield data from corn-soybean rotation crops, indentifying the spatial association of the obtained maps with soil maps. Already Kitchen et al. 2005 used FCM to delineate management zones considering ratios of soil electrical conductivity (EC) in different depths (bulk of EC) and relief data, comparing them with yield zones obtained from historical yield data. As a result, it was found that the bulk of EC combined with relevant data are strong indications for management zones. 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 with 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 before getting the management zones by FCM was performed. High-resolution satellite images also appears as inputs to obtain management zones using the FCM, as in works of Song et al. 2009 and Zhang et al. 2010. More recently, Milne et al. 2012 used FCM to find management zones 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 shows, using FCM to obtain management zones, that combined bare-soil and EC data can contribute to find spatial variability of a crop. The KM-sPC approach [Córdoba et al. 2013] allowed to show the importance of a principal component analysis considering the coordinate space to reduce the stratification provided by FCM when management zones are displayed in form of maps. This approach were used again in a pratical nitrogen management of wheat [Peralta et al. 2015]. The study of Chang et al. 2014 compared management zones generated by FCM using reflectance data regarding the soil properties and productivity, showing that it is feasible the use of an active canopy sensor for this PA application.

Despite the widespread use of FCM for this task, the coordinates space of PA datasets, composed by spatial coordinates variables (e.g., latitude and longitude), have been used only in preprocessing steps or to show the management classes provided by clustering in the form of maps. This fact does not prevent the use of these maps by automated machinery for variable rate interventions, but the reduction of spatial contiguity, causing stratification of management classes in too many areas, can confuse visual analysis by experts. In order to solve this problem, the HACC-Spatial hierarchical clustering algorithm were developed by Ruß and Kruse 2011. 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 coordinates 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, regarding the geostatistics principle claiming that spatially very close samples tends to have close enough values in the features space [Matheron 1963]. As a result, a structure similar to a Voronoi diagram should be obtained by the preprocessing step. 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. However, when a user-defined contiguity threshold *cp* is reached, this restriction is switched off. This threshold is associated to the ratio of the average distances between the samples belonging to adjacent clusters and the average distances between samples belonging to non-adjacent clusters.

Because of the differing nature of FCM and HACC-Spatial (partitional and hierarchical, respectively) and the spatial restrictions used for one of them, are expected distinct clustering results for the same dataset, making it difficult for the user to choose the best approach. A feasible solution to solve this question can be achieved 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, providing more robust clusterings; select models; reuse knowledge; find consensus between clusterings obtained from subsets of features or subsamples, among others [Ghosh and 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 measure by the average of normalized mutual information (ANMI), where the desired optimal value is ANMI equal to 1 [Strehl and Ghosh 2002]. The main goal of the three ensembles algorithms developed by Strehl and Ghosh 2002 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 most obvious heuristic. It is based on the fact that two objects 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 [Karypis and Kumar 1998]. The computational and storage complexity of this algorithm are both quadratic in n. The HyperGraph Partitioning Algorithm (HPGA) addresses the clustering ensemble as a hypergraph partitioning problem, where hyperedges represent the original given clusters as indications 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, HPGA includes original clustering relationships. Finally, the Meta-Clustering Algorithm (MCLA) represent 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 activelly. At the end, a graph-based clustering of hyperedges is performed, indentifying consolidated ”clusters of clusters”. In contrast to CPSA, HPGA e MCLA have linear computational and storage complexity. Still 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 HPGA and CSPA are usually better were obtained from individual clusterings with high noise rates and diversity.

From the abovementioned algorithms, it were possible for us to prepare some experiments, described in section 3, combining distinct approaches that can be applied in the delineation of management classes in PA. Results of these experiments are presented in section 4.

3. Methodology

The methodology used in ours 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 to 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. Another 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 should be used different solutions 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 by ensembles. HACC-Spatial was run using non-spatial features of the whole dataset to calculate dissimilarity values at each step of dendrogram, and spatial features to build the initial tessellation and to support adjacency treatments at each step of dendrogram (Approach I). In the other hand, FCM was run in its traditional way, i.e., using only non-spatial features (Approach II). Regarding ensembles, it was created an approach to found consensus clusterings from individual results provided by Approach I and Approach II (Approach III); and another two approaches to found consensus clustering from individual results provided by non-spatial features subsets of soil, altimetry and yield using HACC-Spatial (Approach IV) and FCM (Approach V). The ensembles approaches was run using CSPA, HPGA and MCLA algorithms described in section 2, and the results with best values of ANMI were chosen as the best solution for each approach.

According to domain expert users, at least 2 and at most 5 management classes should be considered for a crop [Molin et al. 2015]. Thereby, the five 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 were achieved a higher membership degree. For approaches using HACC-Spatial, were used a binding criteria similar to average-linkage algorithm [Sokal 1958], because of its ability to handle data sets with presence of outliers. Other HACC-Spatial parameters, like 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 five approaches when they are run at the same number of clusters. The SD criteria [Halkidi et al. 2000, Halkidi and 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 could have been better allocated to the nearest neighbor cluster. According to Vendramin et al. 2010, the silhouette width criteria, in comparison to other internal criteria in the literature, can provide, in general, 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. 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), given the spatial reference system UTM Zone 23S. Figure 1 shows the contour shape and a cropped image of the experimental field.

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 features associated with spatial coordinates: soil electrical conductivity at 30 e 90 cm deep in 2010

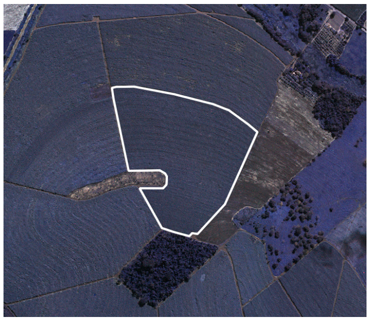
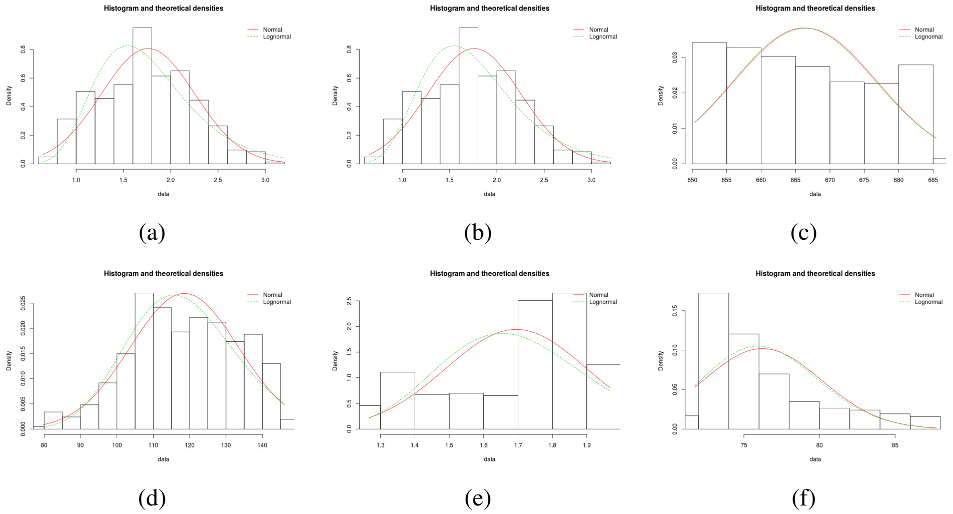


Figure 1. Experimental crop field of sugarcane (white contour) with a cropped image in the background provided by the World View 2 satellite (April 30, 2011).

(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 considered susceptible to anthropic and 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 features 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 abovementioned factors. A special case occurs with the Quota feature, where average values are the minority because the experimental area has a slight slope and narrow in the central region. These distributions are shown in Figure 2.



(a) (b) (c)

(d) (e) (f)

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

Applying the Pearson's Coefficient Correlation between pairs of features, were verified that EC30 and EC90 hold the most positive correlation of the dataset. In general, the Quote feature was well correlated with all other features, and negatively (oppositely) correlated with Yield2010. Regarding to yield data, Yield2012 and Yield2013 features are highly correlated, and negatively correlated with Yeld2010 feature. The negative correlation of Yield2010 with other yield years could be influenced again by the anthropic and climatological factors.

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, like the average of *k* nearest neighbors [Altman 1992], were used to interpolate features with higher sample densities. On the other hand, more sophisticated algorithms, like kriging [Matheron 1969], were used to interpolate features with smaller sample densities. After applying this process, each dataset feature were 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 colors, while higher values are represented by darker colors.

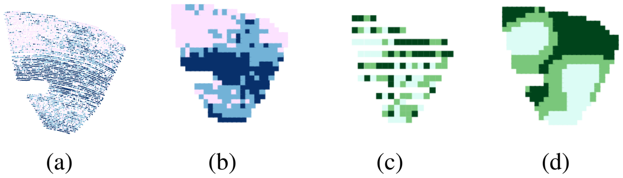


Figure 3. Example of raw and interpolated data in 3 classified intervals: (a) EC30 raw data (9046 samples); (b) EC30 interpolated data (415 samples); (c) Yield2010 raw data (111 samples); (d) Yield2010 interpolated data (415 samples).

Especially for the HACC-Spatial algorithm, when it was run in the context of approaches I, III and IV, 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.

Figures 4 and 5 show, respectively, linear charts containing values achieved by both SD and silhouette width criteria for the five proposed approaches, regarding *k* values between 2 and 5. Through these charts, we can observe better results for *k*=3, where we can found, in general, smaller values of SD and larger values of silhouette width.

By analyzing the results using ensembles, the charts of figures 4 and 5 show us that the approach IV, in the most of cases, achieved poor results regarding both the internal criteria. Therefore, we can conclude that the heuristic of HACC-Spatial algorithm, considering spatial relationships during the construction of the hierarchy, tends to be more consistent when using all features (approach I) than when using individual clusterings by features split to obtain a subsequent consensus by ensembles (approach IV). On the other hand, approach V achieved better results than approach IV, showing that in some cases consensus solutions from individual FCM clusterings by features division can be used to replace solutions provided by approach II. Finally, the approach III results shown, for all *k*

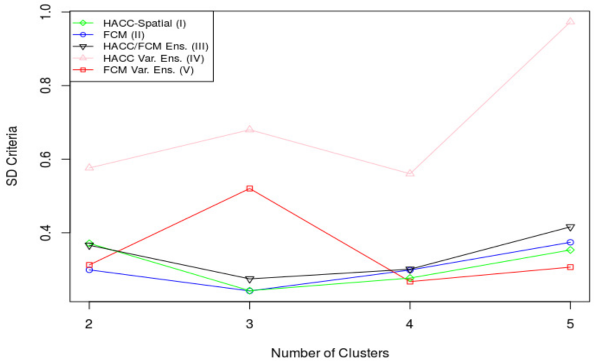


Figure 4. SD criteria values. Each line corresponds to values of SD achieved by the respective approach, considering *k*=2 to 5 clusters.

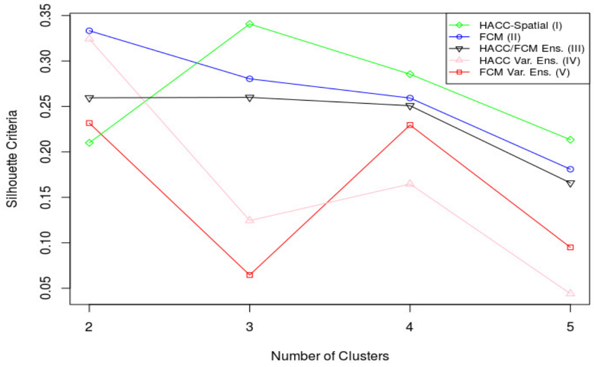


Figure 5. Silhouette width criteria values. Each line corresponds to values of SD achieved by the respective approach, considering *k*=2 to 5 clusters.

values, attempts to find consensus from clusterings obtained by approaches I and II, with slight variations in both the internal criteria values.

Beyond the analysis using internal criteria, we used the visualization of management classes in the form of maps to perform some observations. These analysis were performed from *k*=5 to 2 clusters, in order to observe some effects of agglomerative hierarchy provided by HACC-Spatial approaches. For *k*=5, management classes exhibited generally pronounced stratification, hindering the analysis and an accurate understanding by expert users. For *k*=4, were used, for comparsion with the clustering results, the interpolated dataset from each feature, classified in 4 classes of equal intervals (Figure 6). For each feature, lighter colors represent samples with higher values, while darker colors represent samples with lower values.

Figure 7 shows the results obtained by approaches I, II and III for *k*=4. As can be seen, the results are quite similar for management classes identified with the same color. We can observe the shaping of an isolated area on the top left of the map (blue),

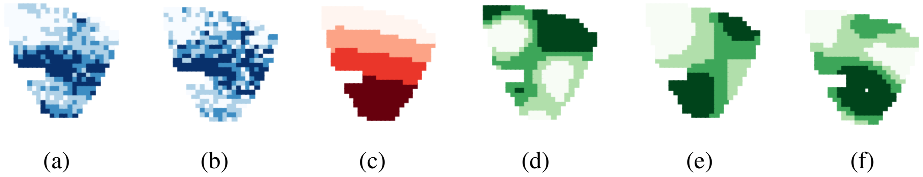


Figure 6. Interpolated data classified in 4 equal intervals: (a) EC30; (b) EC90; (c) Quote; (d) Yield2010; (e) Yield2012; e (f) Yield2013.

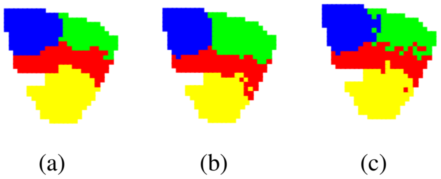
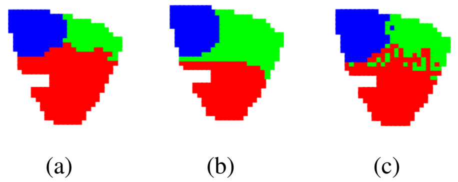


Figure 7. Results for *k*=4: (a) HACC-Spatial (I); (b) FCM (II); and (c) HACC/FCM Ensemble(III).

representing a low elevation region with lower rates of soil EC and historical yield. At the green area, also located in a low elevation region, can be observed medium values of yield and soil EC. Already at the red area, corresponding to a middle elevation region, can be observed a strong influence of extreme values of soil EC for its formation. Finally, the yellow area, located at a high elevation region, shows higher rates of yield.

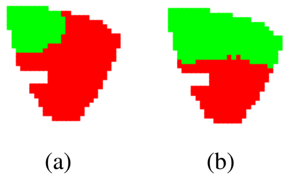
Figure 8 shows the results obtained for *k*=3, regarding the approaches I, II and III, where were achieved the lowest value of SD criteria (approach II) and the highest value of silhouette width criteria (approach I) of the whole experiment. From this figure, can be observed that approaches I and II achieved very similar results. In both cases, the green and blue areas obtained for *k*=4 were practically kept. The main difference between these results is focused at the subdivision between green and red areas. While approach I is forced to merge two clusters because of the hierarchical caracteristics of HACC-Spatial, making the red area be composed by the most similar areas in *k*=4 (red and yellow), the approach II recalculates again which are the clusters where all samples should be assigned, promoting a greater amount of change. Neverthless, the differences observed between both approaches are quite small, which may still be noticed a strong influence of the low frequency of medium values of Quota in approach II, contributing for the user to clearly note the red region with higher values and blue and green regions with lower values of this feature. Regarding to ensemble approaches, Figure 8 (c) further reinforces that approach III, in turn, tried to find a consensus for these subdivision differences, turning the final map quite stratified.

Finally, for *k*=2 (Figure 9), we can verify many differences between approach I, that achieved the worst value of silhouette width criteria, and approach II, that achieved the best values for both internal criteria. While approach I strongly took into account low levels of historical yield in order to identify an isolated area at the top left region (green), merging clusters representing green and red classes for *k*=3, the approach II was affected again by the low frequency of average altitude values, clearly separating a low (green) from a high elevation region (red).



(a) (b) (c)

Figure 8. Results for *k*=3: (a) HACC-Spatial (I); (b) FCM (II); and (c) HACC/FCM Ens.(III).



(a) (b)

Figure 9. Results for *k*=2: (a) HACC-Spatial (I); e (b) FCM (II).

5. Conclusions and Future Work

If we take into account visual analysis and measures of cohesion and separation provided by SD criteria, approaches purely based on FCM (II e V) achieved, in general, better results in comparison to the approaches using the HACC-Spatial (I, III e IV) . Due to the fact that FCM is based in *k-means* algorithm, its bias is always performed to achieve the minimization of intracluster variance and maximization of intercluster dissimilarity. Because of this, internal criteria based on these measures, like SD and silhouette width, tends to provide suitable results for clusterings obtained by this algorithm. On the other hand, in some cases the visual perception of the expert user, of major importance in PA tasks, may be harmed. However, for the silhouette width criteria, these approaches achieved, in general, worst results in relation to those obtained by approach I, except for *k*=2. These results were likely influenced by intrinsic FCM fuzzy features, which can generate doubts if a sample was properly associated with a particular cluster or whether it will be better allocated to the nearest neighbor cluster.

Regarding to the use of ensembles, splitting of features (approaches IV and V) was important for clarifying some details that can get unnoticed in clusterings obtained using all features. However, the high stratification rates generated in the final maps can be very harmful to the users analysis. In the consensus approach between different kinds of algorithms (III), we can observe an increased stratification, causing damage to the visual user analysis. On the other hand, were observed slight variations in SD and silhouette width criteria for different values of *k*, indicating that this approach can be used as solution in some specific cases.

The ensembles approach used in this work is rather general and try to find consensus using only final clusterings obtained from splitting of features or from different algorithms. In an future work, could be used ensembles approaches that allow extracting the main features of each algorithm, making useful data like the membership values provided by FCM, might be used to obtain a better consensus solution.

6. Acknowledgement

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

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing SE - 5*, volume 2 of *Springer Topics in Signal Processing*, pages 1–4. Springer Berlin Heidelberg.

Bezdek, J. C., Ehrlich, R., and Full, W. (1984). FCM : The Fuzzy c-Means Clustering Algorithm. *Computer & Geosciences*, 10(2-3):191–203.

Brock, A., Brouder, S. M., Blumhoff, G., and Hofmann, B. S. (2005). Defining Yield-Based Management Zones for Corn-Soybean Rotations. *Agronomy Journal*, 97(4):1115–1128.

Chang, D., Zhang, J., Zhu, L., Ge, S. H., Li, P. Y., and Liu, G. S. (2014). Delineation of management zones using an active canopy sensor for a tobacco field. *Computers and Electronics in Agriculture*, 109:172–178.

Córdoba, M., Bruno, C., Costa, J., and Balzarini, M. (2013). Subfield management class delineation using cluster analysis from spatial principal components of soil variables. *Computers and Electronics in Agriculture*, 97:6–14.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM*, 39(11):27–34.

Ghosh, J. and Acharya, A. (2011). Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):305–315.

Halkidi, M. and Vazirgiannis, M. (2001). Clustering validity assessment: finding the optimal partitioning of a data set. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 187–194.

Halkidi, M., Vazirgiannis, M., and Batistakis, Y. (2000). Quality scheme assessment in the clustering process. In Zighed, D., Komorowski, J., and Zytkow, J., editors, *Principles of Data Mining and Knowledge Discovery*, volume 1910 of *Lecture Notes in Computer Science*, pages 265–276. Springer Berlin Heidelberg.

Han, E.-H., Karypis, G., Kumar, V., and Mobasher, B. (1997). Clustering based on association rule hypergraphs. In *DMKD*, page 0.

Karypis, G. and Kumar, V. (1998). Multilevelk-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed computing*, 48(1):96–129.

Kitchen, N., Sudduth, K., Myers, D., Drummond, S., and Hong, S. (2005). Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. *Computers and Electronics in Agriculture*, 46(1-3):285–308.

Li, Y., Shi, Z., Li, F., and Li, H.-Y. (2007). Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. *Computers and Electronics in Agriculture*, 56(2):174–186.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58(8):1246–1266.

Matheron, G. (1969). Le krigeage universel.

Milne, A. E., Webster, R., Ginsburg, D., and Kindred, D. (2012). Spatial multivariate classification of an arable field into compact management zones based on past crop yields. *Computers and Electronics in Agriculture*, 80:17–30.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York; London.

Molin, J. P. (2003). Agricultura de Precisão: Situação atual e perspectivas. In Fancelli, A. L. and Neto, D. D., editors, *Milho: Estratégias de Manejo para Alta Produtividade*, pages 89–98. ESALQ/USP/LPV, Piracicaba.

Molin, J. P., do Amaral, L. R., and Colaço, A. (2015). *Agricultura de precisão*. Oficina de Textos.

Morari, F., Castrignanò, a., and Pagliarin, C. (2009). Application of multivariate geostatistics in delineating management zones within a gravelly vineyard using geo-electrical sensors. *Computers and Electronics in Agriculture*, 68(1):97–107.

Peralta, N. R., Costa, J. L., Balzarini, M., Castro Franco, M., C??rdoba, M., and Bullock, D. (2015). Delineation of management zones to improve nitrogen management of wheat. *Computers and Electronics in Agriculture*, 110:103–113.

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

Ruß G. and Kruse, R. (2011). Exploratory hierarchical clustering for management zone delineation in precision agriculture. In Perner, P., editor, *Advances in Data Mining. Applications and Theoretical Aspects*, volume 6870 of *Lecture Notes in Computer Science*, pages 161–173. Springer Berlin Heidelberg.

Schwalbert, R. A., Amado, T. J. C., Gebert, F. H., Santi, A. L., and Tabaldi, F. (2014). Zonas de manejo: atributos de solo e planta visando a sua delimitação e aplicações na agricultura de precisão. *Revista Plantio Direto*, pages 21–32.

Scudiero, E., Teatini, P., Corwin, D. L., Deiana, R., Berti, A., and Morari, F. (2013). 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*, 99:54–64.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–1438.

Song, X., Wang, J., Huang, W., Liu, L., Yan, G., and Pu, R. (2009). The delineation of agricultural management zones with high resolution remotely sensed data. *Precision Agriculture*, 10(6):471–487.

Strehl, A. and Ghosh, J. (2002). Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3:583–617.

Taylor, J. A., McBratney, A. B., and Whelan, B. M. (2007). Establishing Management Classes for Broadacre Agricultural Production. *Agronomy Journal*, 99(5):1366–1376.

Vendramin, L., Campello, R. J. G. B., and Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235.

Vendrusculo, L. G. and Kaleita, A. L. (2011). Modeling zone management in precision agriculture through Fuzzy C-Means technique at spatial database. In *2011 Louisville, Kentucky, August 7 - August 10, 2011*, St. Joseph, MI. American Society of Agricultural and Biological Engineers.

Vieira, S. R. (2000). Geoestatistica em Estudos de Variabilidade Espacial do Solo. In Novais, R. F. and Alvarez, V H, S. G. R., editors, *Tópicos em ciência do solo*, pages 1–54. Sociedade Brasileira de Ciência do Solo, Viçosa, MG, 1 edition.

Weiss, S. M. and Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Zhang, X., Shi, L., Jia, X., Seielstad, G., and Helgason, C. (2010). Zone mapping application for precision-farming: a decision support tool for variable rate application. *Precision Agriculture*, 11(2):103–114.