2950 Niles Road, St. Joseph, MI 49085-9659, USA
269.429.0300 fax 269.429.3852 hq@asabe.org www.asabe.org

An ASABE Meeting Presentation

Paper Number: 1111774

Modeling zone management in precision agriculture through Fuzzy C-Means technique at spatial database

Laurimar G. Vendrusculo, PhD. Candidate and Researcher

Iowa State University - Department of Agricultural and Biosystems Engineering

100 Davidson Hall, Ames, Iowa 50011 - USA

Embrapa Informatics - Av. André Tosello, 209 - Barão Geraldo - Caixa Postal 6041- Campinas, Sao Paulo, 13083-886 – Brazil

Amy F. Kaleita, Associate Professor

Iowa State University - Department of Agricultural and Biosystems Engineering

100 Davidson Hall, Ames, Iowa 50011 - USA

Written for presentation at the
2011 ASABE Annual International Meeting
Sponsored by ASABE
Gault House
Louisville, Kentucky
August 7 – 10, 2011

Abstract. *Predict the optimal number of zones to manage tasks evolved in precision agriculture applications is challenging issue in classification tasks. Important decisions in the farm required maps of yield classes which contain relative large, similar and spatially contiguous partitions and sometimes without a priori knowledge of the field. The main goal of this study was to apply Fuzzy C-means (FCM), an unsupervised classification technique, in a geo-referenced yield and grain moisture dataset in order to find optimal number for homogeneous zones. Those data were produced by Long-Term Ecological Research in a Biological Station (KBS-LTER), Michigan, during growing season at 2008. The best results presented by this algorithm ranged from 8 to 10 zones which were validated using the indexes Partition Coefficient (PC), Classification Entropy (CE) and Dunn's Index (DI). Even though, only two attributes were collected in the dataset, the Fuzzy C-means has shown promissing results for zone mapping.*

Keywords. *soft computing, fuzzy classification, optimal zone number, precision agriculture.*

(The ASABE disclaimer is on a footer on this page, and will show in Print Preview or Page Layout view.)

Introduction

The Precision Agriculture approach has been established under the pressure on produce more and health food considering the negative impact in the environment and also the fast development of the many mechanical and electronic devices. The precision agriculture is defined, mostly, as a combination of techniques to detect, analyze and deal with many sources of variability in the field. All this techniques are based on information technology, remote sensing or GPS technology that address applications of in-field inputs such as seed, fertilizer, pesticides or water, according to soil or crop needs. (Whelan et al., 1996; Auernhammer, 2001).

During the past two decades, satellite guidance system for tractors and rate variable technology allowed the farmer a precise manner to manage within-field variability. One valuable and practical tool to deal with changes in the field is grouping areas with similar numerical characteristics, for instance, subfields with the same range of yield. This tool, called zone mapping, split the field in a reasonable number of regions in order to allow site specific operations. According Fridgen et al. (2004), to develop a zone map, usually, is necessary a procedure to group or classify the data. Thus, one optimal number of zones is obtained as outcome of the previous step. A prior knowledge of the field is also essential. Many studies have carried out to identify areas that are similar based on, crop yield and soil properties (Boydell and MacBratney, 1999; Fraisse et al., 2001; Goktepe et al. ,2005)

Traditionally, the analysis of crop yield forecasting has been done, mainly, through regression techniques from empirical statistical models or simulations models. The data mining approach offers an effective form of knowledge discovery in order to find out unexpected patterns of information in large database. Additionally, soft computing methodologies applied in the data mining step (neural networks, support vector machines, genetic algorithm, and fuzzy logic) provide a robust and low cost solution with a tolerance of imprecision and uncertainty for many fields. Soft computing techniques used in agricultural and biological engineering are widely discussed in Huang et al. (2010).

Due to uncertain behavior of some agricultural systems as soil properties or continuous variability in natural phenomena (Burrough, 1989) the fuzzy set theory or fuzzy logic may be suited to take in account such uncertainties.

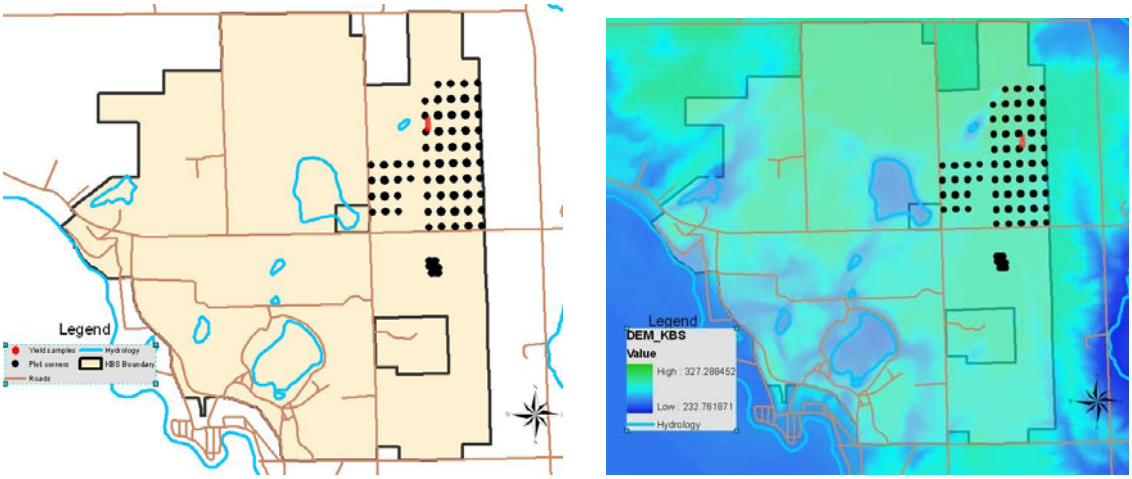
For this reason, the goals of this study are: (i) apply the fuzzy C-means in a yield and grain moisture dataset in order to identify management zones (ii) test cluster validation techniques to determine the optimal number of class groups from unsupervised fuzzy clustering. The formal definition fuzzy c-means and the validity indices pattern are given in Section 2. An experimental evaluation of our approach is presented in Section 3. Finally, in Section 4 some conclusions are drawn from this paper.

Material and Methods

Site description

This study has used one subset of the database provided by Kellogg Biological Station in Kalamazoo County, southwestern of Michigan (85^{o} 24' W, 42^{o} 24' N). In that area has held Long-

Term Ecological Research (LTER) Program studies of intensive field crop ecosystems. This station was established in 1989 as part of a national network of LTER sites created by the National Science Foundation. The red spot shown in Fig. 1 (a) represents the study area and also few layers such as boundary, hydrology, and roads are present. The digital elevation model (Fig. 1b) shows the variation of altitude in the study area.



(a) (b)

Figure 1 – Kellog Biological Station Long Term Ecological Research Site with emphasis at (a) boundary, roads, hydrology resources, corner of plots and study area (b) Digital Elevation Model.

According Smith et al. (2007) the Soils at the KBS LTER site are a combination of Kalamazoo (fine-loamy, mixed, semiactive, mesic Typic Hapludalfs and Oshtemo (coarse-loamy, mixed, active, mesic Typic Hapludalfs) sandy loams. The Mean annual temperature is 9.4^{o}C and the annual precipitation (30-yr mean) is 860 mm.

Data Collection

The corn (*Zea Mays L.)* field was harvested with a combine equipped with yield monitor system (grain mass flow and moisture sensors). Yield geo-referenced data was exported from the John Deere Ag software package in order to perform temporal and spatial analysis. The database[1] comprises 509 records of grain yield (bushels/acre) and grain moisture (grams/hectograms) for 2008 growing season.

The data was standardized in order to simply rescale de values for the same magnitude. This process has no impact on the shape of the distribution curve.

[1] Available at - http://lter.kbs.msu.edu/datatables/185

Mapping data

In order to visualize and analyze precision agriculture data, interpolations of grain yield and grain moisture were plotted in ArcGIS software (Fig. 1) using, initially, 4 classes. Through the kriging interpolation technique, the variables yield and grain moisture were mapped. Principles of the kriging interpolation method are broadly explained in Isaaks and Srivastava (1989) and Cressie (1993).

Delineation of management zones

By the nature of the phenomena, some responses for soil conditions, nutrient and weather in crop systems, for instance, higher classes of yield are characterized by transitional zones, therefore the change is gradual between the classes. The hard partition among those classes became subjective if one supervised method is applied.

In order to overcome the issues with blurred boundaries among classes many studies have been used the fuzzy membership model as an approach to treat the uncertainty in classifying data. Especially, fuzzy c-means technique deals with grouping similar instances with minimum variance.

Initially proposed by Bezdek et al. (1984) for geostatistical data studies, fuzzy c-means (FCM) has been used in many studies (Lark and Stafford, 1997; Odeh et al., 1992) and have shown effectives results for zone mapping. Zhang et al. (2010) has developed a web-based tool (ZoneMap) for Zone Mapping Application for Precision Farming. In that study, FCM is applied in combined data from precision agriculture and remote sensing.

The zone mapping procedures comprise three mains steps: (1) Data preprocessing: Generally, one of the most time-consuming activity that allows trimmer the outliers points through mapping data, for example. Overlapped views of variables from GIS also provide some idea about the spatial patterns present at the field (2) Application of the fuzzy c-means at yield dataset (3) Validation of results by four validity indices.

The fuzzy c-means algorithm

The method FCM for clustering allows points to belong to more than one cluster. Thus, this method is frequently used in pattern recognition and it is based on minimization of the objective function, given by equation 1:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - c_j \right\|^{\alpha}. \qquad 1 \leq m < \infty \tag{1}$$

Where N is the number of the specified clusters and m is fuzziness exponent and can be any real number greater than 1, u_{ij} the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

The fuzzy partitioning process is conducted through an iterative optimization of the objective function, as indicated in equation 2. For each interaction the membership is updated. So new values of membership, u_{ij} and the cluster center c_j are calculated by equation 2:

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{1=1}^{n} u_{ij}^{m} x_i}{\sum_{1=1}^{n} u_{ij}^{m}}$$

(2)

This iteration will finish when $\max_1(i,j) = \{u_1(ij)^T(m(k+1)) - u_1(ij)^T(m(k))\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

The fuzzy Clustering Toolbox for MatLab Environment, developed by Abonyi et al. (2003) and also the library E1071 of R package were employed in the clustering tasks.

Validation

The step of validation is related to the procedure to verify if a given fuzzy zone fits as best it can to the whole database. Commonly, the cluster validity indexes calculated in this step measure statistical properties of clustering results, mostly the distance within cluster or among clusters. This fitting step includes also other aspects as a fixed number of the cluster and the shapes of the clusters found. This study has validated the dataset through four validity indices described as following.

(a) Partition Coefficient (PC): measures the amount of "overlapping" between clusters. It is defined by Bezdek et al. (1984) as following equation.

$$PC(c) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^2 \qquad (3)$$

Where μ_{ij} is the membership of data joint j in cluster i.

(b) Classification Entropy (CE): Based on the study of Cheng et al. (1998), the CE measures the fuzziness of the cluster partition only, which is similar to calculate in the previous coefficient.

$$CE(c) = -\frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij} \log(\mu_{ij}) \qquad (4)$$

(c) Dunn's Index (DI): The function of this measure originally was used to identify how compact and well separated were the clusters. Under this condition the final result of the clustering must to be recalculated as it was a hard partition algorithm.

$$DI(C) = \min_{i \in c} \left\{ \min_{j \in c, i \neq j} \left\{ \frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_{k \in c} \left\{ \max_{x,y \in C^d(x,y)} \right\}} \right\} \right\} \quad (5)$$

As C and N, in equation 5, increase the calculation of Dunn's index becomes computationally expensive.

(d) Alternative Dunn Index (ADI): The goal of this index is speed up the Dunn calculus through its simplification (Trauwaert, 1988). This principle is used when dissimilarity function between two clusters ($\min_{i}(x \in C_i(i,j) \in C_i j)\, d(x,y)$) is rated in value by the triangle-non equal.

$d(x,y) \geq |d(y,v_j) - d(x,v_j)|$ Where v_j is the cluster center of the j-th cluster.

$$ADI(C) = \min_{i \in c} \left\{ \min_{j \in c, i \neq j} \left\{ \frac{\min_{x \neq C_i, y \neq C_j} |d(y,v_j) - d(x,v_j)|}{\max_{k \in c} \left\{ \max_{x,y \in C^d(x,y)} \right\}} \right\} \right\} \quad (6)$$

Results and Discussion

Summary statistics of crop yield data

Fig. 2 depicts the summary statistics and frequency distribution of yield and grain moisture at corn fields in a Long-Term Ecological Research in a Biological Station, MI, during growing season at 2008. Generally, the frequency distributions of yield are close to normal. This finding suggests that the yield variable has fewer tendency to produce unusually extreme values; compared to grain moisture. For the grain moisture the histogram is slightly negatively skewed. The mean and median in both variables are quite similar. But, the higher value of yield standard deviation may be an indicator of intense variability of secondary attributes such as soil properties.

Max: 106.2

Median: 55.0

Max: 22.3

Median: 17.9

Min: 16.9

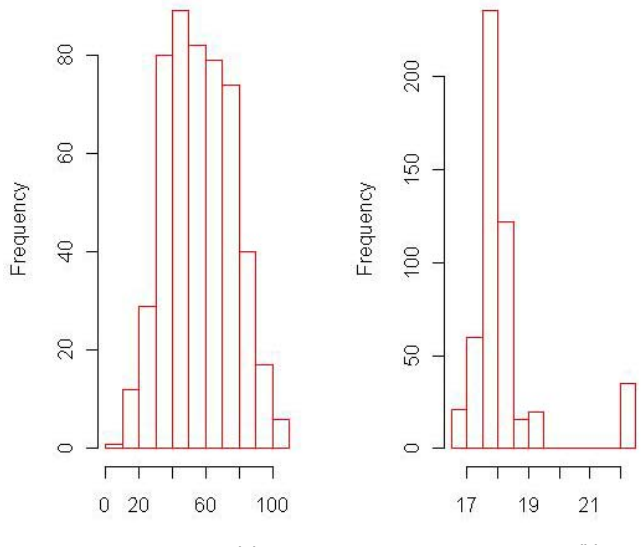


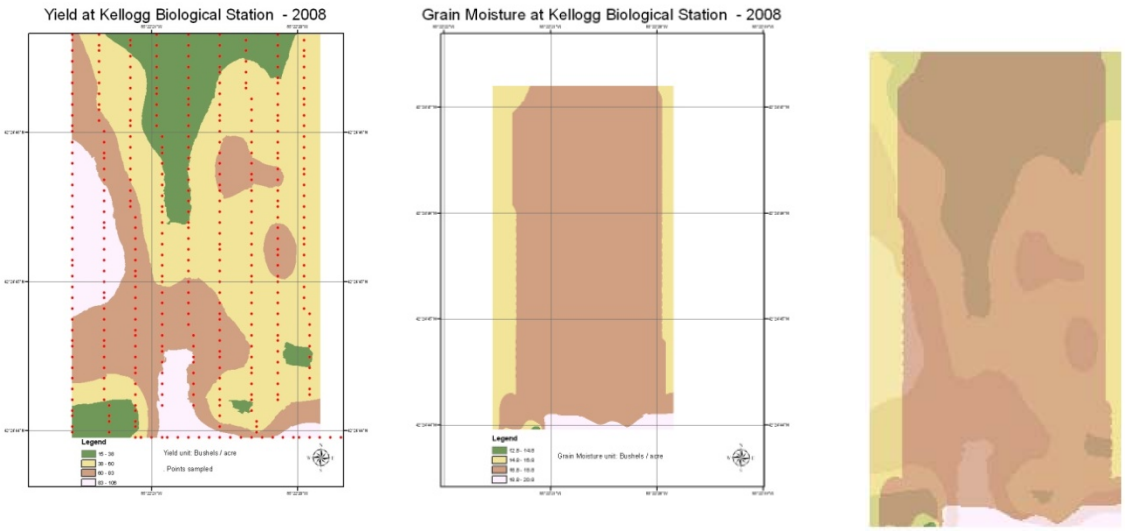
Figure 2 – Frequency distribution of (a) yield and (b) grain moisture and summary statistics.

Spatial Distribution of yield and grain moisture

The visual analysis of corn yield in Fig. 3(a) shows that the direction of the lower classes of yield corresponds from north to south. In the north portion of field the altitudes range from 291 m to approximately 283 m at south. Furthermore, the higher values of yield appear only in small areas in the west and south of the field. The sampling task was carried out in the dots presented in the Fig. 3a. The values most frequently found vary from 36 to 60 bushels/acres. According the USDA 2008 Crop Production Report (USDA, 2009), the average U.S. grain yield for corn was estimated at 153.8 bushels per acre. Thus, the higher yield class considered in this study (80 to 106 bushel/acre) is only for scientific purpose.

Therefore, analyzing grain moisture shown in the Fig. 2b, the most frequent values for grain moisture found are 16.8 to 18.8 grams/hectograms. These values cross the whole field in the same direction. However, small stripes on the east and west sides of the field are found and demonstrate that there is more humidity in the center of the area than the borders. Only an insignificant area of higher rate of grain moisture was found on the south side of the field.

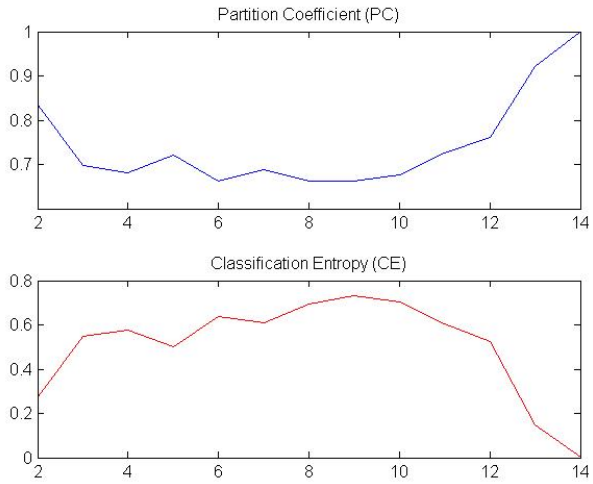
The overlapped view of yield and grain moisture is shown in Fig. 3(c). In this case, it is highly improbable that those attributes have spatial correlation because only one class of grain moisture comprises diverse ranges of yield. This finding suggests that more variables related to soil and weather might need to be analyzed to explain the spatial variability of the yield. Some variation on the nutrient levels might be also another indication of the yield variability.



(a) (b) (c)

Figure 3 – Map for four classes interpolated corn yield map and the respective sampling points (a), grain moisture (b) and two-layers view (c) of the study area in 2008 at KBS LTER.

The validity measures are calculated and shown in Figure 4.



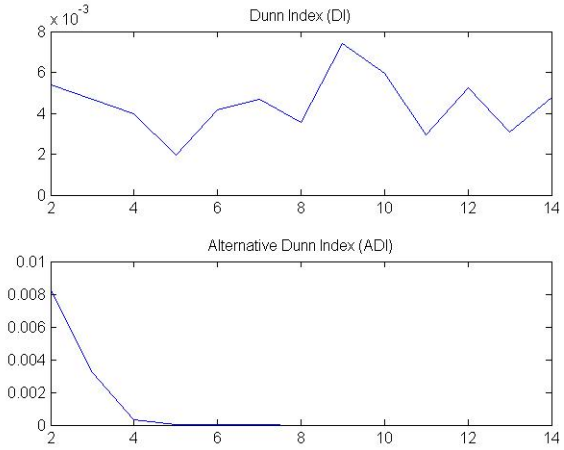
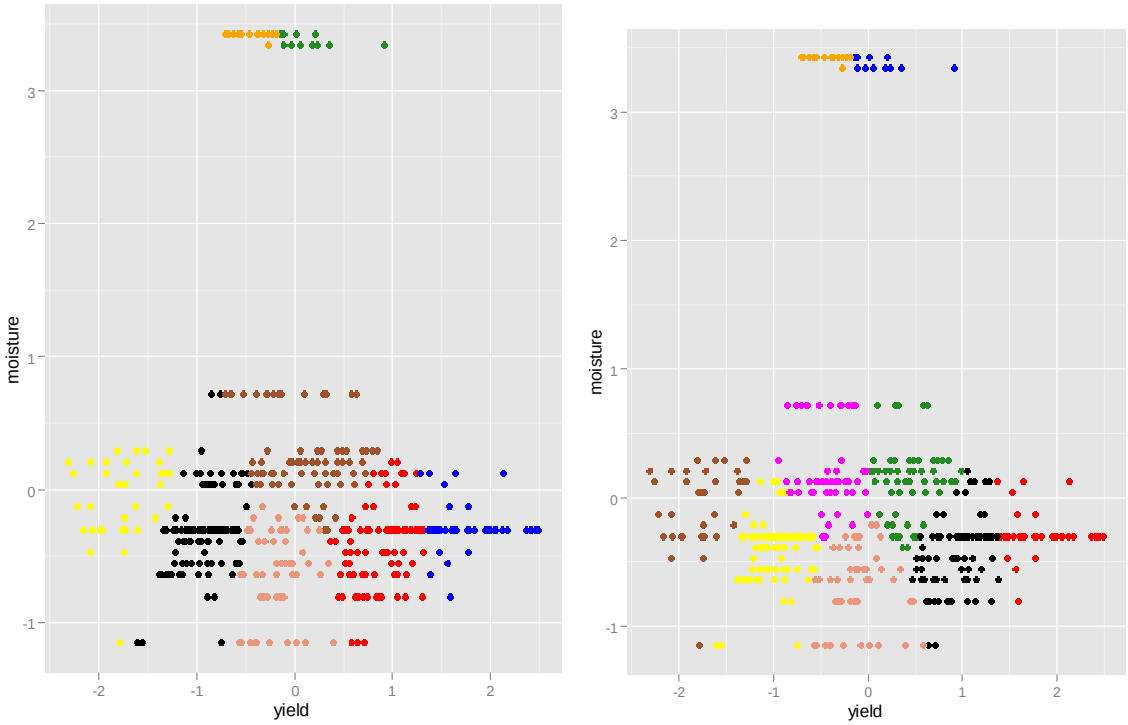


Figure 4 – Measures of validity applied to yield dataset considering the maximum of 14 clusters.

The behavior of the PC index is similar to CE, however in opposite way. This trend makes sense since the formulas are quite close. The values close to 1 in PC and to 0 in CE measure represent the desired choices for the optimal number of cluster. But, those indices have some limitations, once they are not connected directly to some attributes of the dataset.

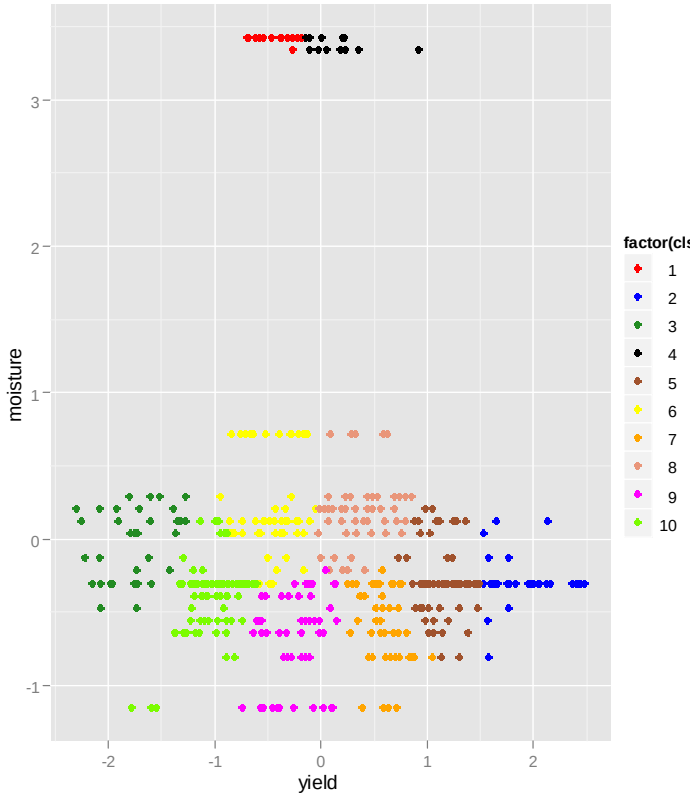
Lower values in PC and higher in CE are eligible indices to balance the choice of optimal number of cluster. For this study, the range from 8 to 10 clusters is likely probably to be the better ones and 9 clusters should be the best one. Despite of the indication of the indices (8 to 10 clusters) we decide explore a broad number of clusters in order to analyze the possible shapes of the clusters and its accuracy in create plausible partitions to the dataset. This range is supported for high values in Dunn's index; however the saturation of ADI is noticed when the number of cluster reaches 4.

The fuzzy clustering technique was performed on the dataset of the study area for 8,9,10 clusters. In order to reproduce the same results was chosen a seed for the random step of FCM. The images illustrated in Figure 5 show the FCM results through the plots of yield (axis X) versus grain moisture (axis Y). The data was standardized. It can be noticed that this approach tries to enclose the clusters through a circle shape.



(a)

(b)



(c)

Figure 5 – Lower dimension of 8(a), 9(b), 10(c) clusters employing FCM approach.

Fig. 6 shows variation of the sum of square distances within the clusters from 2 to 15 groupings. Observing only the natural distribution of points in Figure 5, there are at least 4 well-separated groups. One of those groups contains most of the observations. This fact explains the higher rate of variation of the weighted standard deviation shown in figure 6 for the number of cluster from 2 to 6. The changes of sum of square distances are smooth from 8 groups. Thus, the potential management zones start from 8 clusters when the weighted standard variations within the groups vary less.

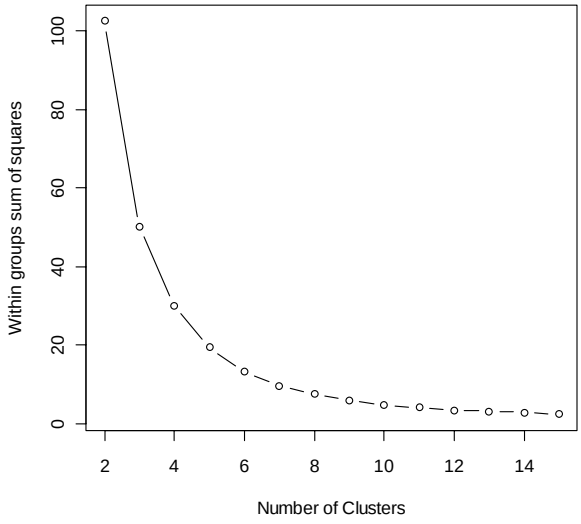
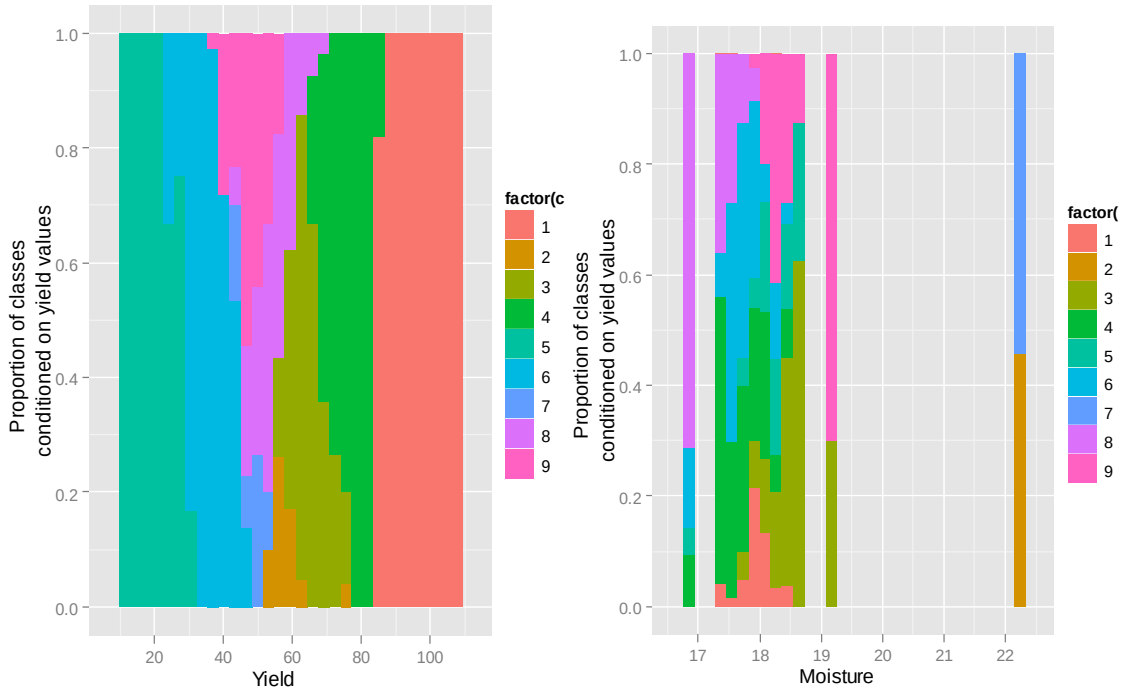


Figure 6 – Sum of square distance within the clusters.

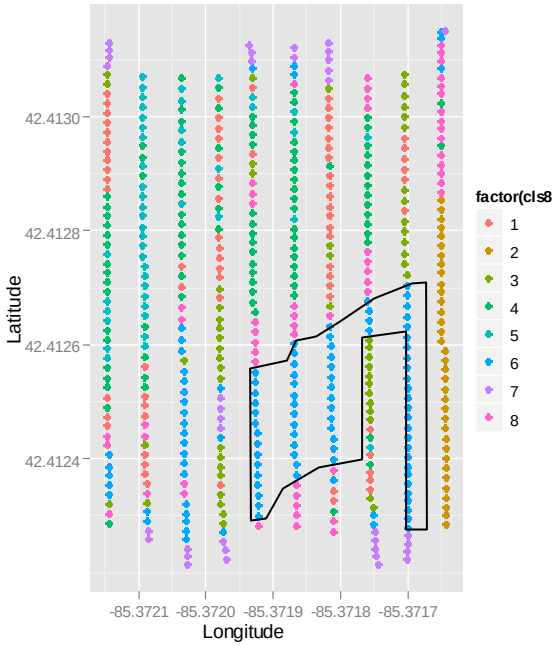
Besides the clustering process was based on employing only two variables, the study decides to explore the distribution of cluster according original values of yield and grain moisture. According Fig. 7a, the classes 5 and 1 represent the lower and higher intervals of yield, respectively. Class 4 can be truly described from 78 to 82 bushels/acre. Nevertheless, the remained classes have participation in values from 30 to 70 bushels/acre. The Fig. 7b, demonstrates that grain moisture attribute is cumbersome to extract a pattern based on range because the classes are overlapped.



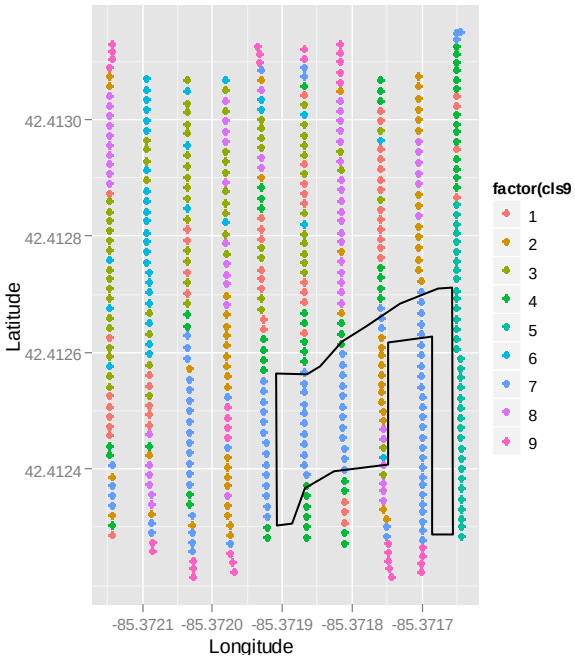
(a) (b)

Figure 7 - Distribution of clusters according yield and moisture values.

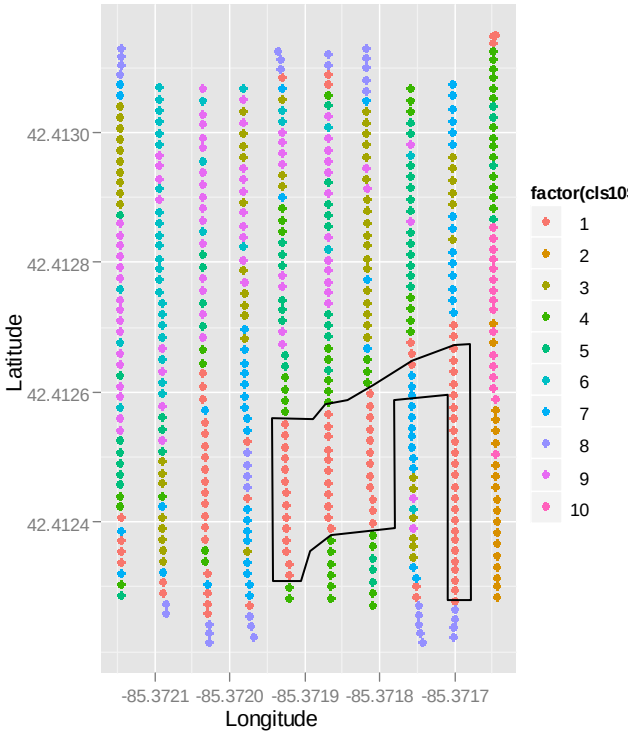
After find the potential configurations for 8, 9 and 10 clusters, the last step was plotting each point geo referenced and its correspondently number of cluster. The figure 8 presents those configurations. The class 3 in field divided by 8 clusters (Fig. 8a) corresponds to the edges in top and bottom of field. With the exception of class 7 in cluster number 8, all the dual-top clusters in Figure 5 are located in bottom right portion of Figure 8. However, the FCM was not able to detect continuous clusters in the top areas of the Figure 7. The reason is because those points come from close classes such as classes 1, 4, 5 in Figure 5a and theirs respectively plotting in Figure 8a. The lack of clear boundaries in the circled clusters in Fig. 5 is likely caused by the physical discontinuity of yield shown in Figure 8.



(a)



(b)



(c)

Figure 8 - Field representation of FCM partitioning for 8 (a), 9(b) and 10 (c) clusters

Conclusions

In this study, the fuzzy c-means approach was employed to find out optimal number for clusters for zone mapping application from precision agriculture data. According the partition coefficient, classification entropy and Dunn Index, 9 is the potential optimal number of management zones for this data. However, we decided to explore the range from 8 to 10 clusters for this study.

It is important to observe the limitations of this study. First, only one season was considered for the FCM approach. Boydell and MacBratney (1999) recommend at least five years of yield data in order to find a stable number of management zones. Secondly, the number of samples is not quite big enough to establish a pattern for the yield in general. However, we have no reason to believe that our conclusions affected the number of management zones found in the field.

The work carried out in this study suggests that the use of fuzzy k-means on yield and grain moisture provide sensible groupings. However, the plotting of clustered points, especially in the top of study area presents many small similar areas that may difficult a practical rate variable operation.

The underlying structure of the dataset used in this study shows that the majority of points have significant membership in multiples classes. Thus, the partitioning of these data becomes cumbersome. Although the dataset used was composed only for two attributes (yield and grain moisture), the FCM technique has shown promising outcomes for zone mapping.

Better delineation of these areas could be achieved either by sampling chemical and physical attributes of the soil or obtaining detailed information of the climate at the previous grown seasons.

As a further research, the comparison with another soft computing technique such as Model based or hierarchical will provide the level of accuracy of the fuzzy c-means technique against those unsupervised methodologies. We also intend to use the full version of the database that comprises a time series and calculate other topographical features such as slope, aspect, flow accumulation and Wetness Index. This approach will lead us for a more complex version of this model.

Acknowledgments

We would like to thanks Dr Suzanne Sippel who allows using the KBS LTER data table KBS037-002. The Support for this research was also provided by the NSF Long-Term Ecological Research Program at the Kellogg Biological Station and by the Michigan Agricultural Experiment Station.

References

Abonyi, J., B, Feil; Nemeth, S., Arva, P. 2003. Fuzzy Clustering based Time Series Segmentation*, Lecture Notes in Computer Science.* 2810, 75-84.

Auernhammer, H. 2001. Precision farming - the environmental challenge. *Computers and Electronics in Agriculture*, 30: 41-33.

Bezdek, J., R. Ehlich, W. Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences.* 10(2), 191-203.

Boydell, B., McBratney, A.B. 1999. Identifying potential within-field management zones from cotton yield estimates. P. 331-341. In J. V. Stafford (ed) Precision Agriculture '99. Proc. European Conf. on Precision Agric., 2nd, Odense Congress Cent., Denmark. 11-15 July 1999. SCI, London.

Burrough, P. A. 1989. Fuzzy mathematical methods for soil survey and land evaluation. J. Soil Sci. 40:477-492

Cheng, H.D., Chen, J.R., Li, J. 1998. Threshold selection based on fuzzy c-partition entropy approach. *Pattern Recognition*, 31 (7), 857–870.

Cressie, Noel A.C. 1993. *Statistics for Spatial Data* Revised Edition. New York: John Wiley & Sons, Inc. 900 p.

Fraisse, C.W.**,** K. A. Sudduth, N. R. Kitchen. 2001. Delineation of site-specific management zones by unsupervised classification of topographic attributes and soil electric conductivity. *Transactions of the ASAE* 44(1): 155-166.

Fridgen, J. J., Kitchen,N. R., Sudduth, K. A. , Drummond, S. T., Wiebold, W. J., Fraisse, C. W., 2004, Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation *Agronomy Journal* 96:100–108.

Goktepe, A.B, Altun, S., Sezer, A. 2005. Soil Clustering by fuzzy c-means algorithm. *Advances in Engineering Software* 36: 691-698.

Huang, Yanbo, Lan, Yubin, Thomson, Steven J., Fang, Alex, Hoffmann, Wesley C., Lacey, Ronald E. 2010. Development of soft computing and applications in agricultural and biological engineering. *Computers and Electronics in Agriculture*, 71:107–127.

Isaaks, E. H., Srivastava, R.M. 1989. *An Introduction to applied geostatistics*, New York: Oxford University Press, 561p.

Lark, R. M., Stafford, J. V. 1997. Classification as a first step in the interpretation of temporal and spatial variation of crop yield. *Annals of Applied Biology*, 130: 111–121.

Trauwaert, E. 1988. On the meaning of Dunn's partition coefficient for fuzzy clusters. *Fuzzy sets and systems*, 25(2), 2127-242.

Odeh, I. O. A, Chittleborough, D. J. & McBratney, A. B. 1992. Soil pattern recognition with fuzzy c-means: Application to classification and soil-landform interrelationships. *Soil Science Society of America Journal*, 56: 505–516.

Smith, R. G., F. D. Menalled, and G. P. Robertson. 2007. Temporal yield variability under conventional and alternative management systems. *Agronomy Journal* 99: 1629-1634.

USDA. 2008 Crop Production. 2008 . National Agricultural Statistics Database. Washington, D.C.: USDA National Agricultural Statistics Service. Available at: < http://usda.mannlib.cornell.edu/usda/nass/CropProd//2000s/2008/CropProd-12-11-2008.pdf> Accessed : 20 April 2011

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

Whelan, B.M., McBratney, A.B., Viscarra Rossel, R.A. 1996. Spatial prediction for Precision Agriculture. In P.C. Robert, R.H. Rust & W.E. Larson (ed.) Precision Agriculture: *Proceedings of the 3rd International Conference on Precision Agriculture, ASA/CSSA/SSS*, Madison, p. 331-342.