Hierarchical Spatial Clustering for Management Zone Delineation in Precision Agriculture

Georg Ruß[1], Martin Schneider[2], Rudolf Kruse[1]

[1] Otto-von-Guericke-Universität Magdeburg, Germany
[2] Agri Con GmbH, Jahna, Germany

Abstract. Precision agriculture is concerned with the application of current technology in the field of agriculture. Huge data sets are nowadays collected during standard farming operations. These data are fine-scale and available in high resolution, usually reflecting the heterogeneity of any agricultural field. The data may result from a multitude of sensors and electronic equipment and can be used for a number of purposes which affect the efficiency and effectiveness in farming operations.
An important task in classic agriculture is *base fertilization*. This term is commonly used to describe the process to make minerals like potassium (K), phosphor (P) and magnesium (Mg) available for the planted crops. Since the field is usually heterogeneous, the question arises which part of the field should be treated and how it should be treated. This is usually associated with the concept of *management zone delineation*. There have been quite a number of approaches towards using fine-scale data for subdividing the field into a small number of zones. These approaches often require multi-year data sets and are based on high-resolution sampling methods for data acquisition. Existing research into the semi-automatic generation of management zones usually applies a variety of clustering methods to subdivide the field into homogeneous parts. Most of the existing approaches use only parts of the available data for the clustering process.
In this paper, we propose a novel approach which tackles the above issues. We base our work on a site-year of high-resolution spatial data from selected wheat fields in Germany. Our approach may best be described as *divide-and-conquer*: we split the field into homogeneous subfields and consecutively merge these subfields using a hierarchical agglomerative clustering approach with a spatial constraint.
Keywords: Spatial Clustering, Hierarchical Clustering, Management Zone Delineation, Precision Agriculture

1 Introduction

In recent years, the area of *precision agriculture* has seen an enormous increase in research activity. This has partly been triggered by a decline in technology cost (such as GPS) and the successful integration of these new technologies into agricultural equipment. These technologies essentially enable to measure physical properties of an agricultural field with

a high resolution. Some examples are crop growth sensors, fertilizer usage sensors and high-resolution satellite or aerial imaging. These sensors generate spatial data sets, i.e. each data record in these data sets has a geographical location and therefore natural neighbors. Therefore, methods that take these special properties into account have to be developed to cope with the tasks encountered in precision agriculture.

One of these tasks is the delineation of management zones. This process is usually required before the growing season when the availability of Potassium (K), Phosphorus (P) and Magnesium (Mg) is measured. It is, however, prohibitively expensive to sample the soil at the spatial resolution required for precise management. Therefore, more cost- and time-efficient data acquisition sensors are used. This gives the drawback that the availability of these minerals can not directly be determined but should be somehow inferred from the available sensor data. Furthermore, a lot of research effort has been devoted to automating the process of finding appropriate management zones when high-resolution soil-sensing data are available. Nevertheless, a literature review laid out below will point out the importants studies and their drawbacks from the agricultural point of view, clarifying the need for a novel approach.

On the other hand, the management zone delineation approach might be tackled as a data mining problem, i.e. an exploratory clustering problem: partition the field into a fixed number of possibly contiguous zones and evaluate those in practice, using expert knowledge. However, due to the spatial nature of the precision agriculture data sets, this narrows down the choice of algorithms from the large foundation of clustering algorithms available. A literature review using the major subtypes of clustering algorithms will demonstrate that none of those is capable of dealing with the data sets encountered here.

Keeping the limitations of the agricultural approaches and the existing clustering algorithms in mind, a novel clustering algorithm will be developed which uses parts of the existing work. In previous work, a novel tessellation of the field has been presented, based on simple k-means clustering applied to the geographical coordinates of the data records [12]. Since the tessellation has empirically shown to be practical and more reliable than a grid-based tessellation, it is going to be used in this article to provide the first stage of a two-stage management zone delineation algorithm. The underlying idea of this algorithm is rather simple: while the first stage divides the field into small subfields, the second stage works on those subfields and consecutively merges the most similar and neighboring clusters into a new, larger cluster. Due to the nature of the algorithm, we term it *hierarchical agglomerative clustering with spatial constraint*.

The structure of the article is as follows: while Section 2 reviews the agricultural approaches towards management zone delineation, Section 3 does so with a short overview about existing clustering algorithms. Both sections will point out the need for a novel management zone delineation algorithm tailored to the spatial, high-resolution data sets encountered here. This novel algorithm will be explained in detail in Section 4. It will be evaluated shortly in Section 5, followed by the finishing conclusions in Section 6.

2 Literature Review on Management Zones

The delineation of management zones has been used as a method of subdividing fields into parts with different properties for a long time. However, this has usually been done using expert and long-term knowledge of the respective field. The advent of modern GPS technology in the early 1990s has not had a significant impact until 2000, when the *selective availability* restriction was decommissioned, allowing for higher precision of GPS data. Due to further recent advances in technology, the delineation of management zones has turned into a data-driven approach for subdividing the fields. Therefore, the approaches below represent the cutting edge in this topic, spawned by the public availability of GPS from 2000 onwards. This provides the agricultural basics for a novel approach towards management zone delineation. The following literature review on management zone delineation is presented in chronological order.

In the study of [9], spatially coherent regions are defined by fuzzy classification, smoothing the fuzzy memberships, then defuzzifying the smoothed memberships to allocate each individual data record to one of the classes. Each of these three consecutive steps is well reasoned for by the author. The overall goal is to find spatially coherent regions by classification of multi-variate data. The author uses three years' worth of yield maps from a 6-ha field in Bedfordshire, East England, UK. Without going into too much detail, it is worth noting that the spatial coherence of the regions is achieved in a two-step process, by applying fuzzy classification and then smoothing the fuzzy memberships. It should be investigated whether this approach can be performed in an integrated step.

The approaches of [2, 3] represent a basic data analysis approach to management zone delineation. In [2] different data attributes are evaluated towards their potential for zone delineation. First, single attributes are used and their correlation coefficients with base nitrate results are determined. Second, combinations of attributes and their correlation coefficients are determined. The six attributes are topography, yield, order 1 soil survey aerial photography, satellite imagery and EC_a. A subset of the variables consisting of topography, satellite imagery and yield mapping is determined as yielding the highest and most consistent correlation. Using all the available attributes produced worse results than using just a subset of those. In subsequent work, those results are confirmed [3]. The management zones are determined via a simple weighted summation method, resulting in the desired four different management zones.

In [6] a multi-year approach for determining management zones on a field in Iowa, USA, is presented. Sampled data from six years are used, with a total of seven variables in addition to the yield variable. The data are sampled along eight transects with 28 sampling points each, resulting in 224 data points for this 16 ha field. Based on the available attributes, the authors propose a three-step process of *partitioning, interpretation and profiling*. A k-means cluster analysis is performed, mainly for the purpose of data reduction to a manageable number of types. Once the clusters are obtained, a canonical multiple discriminant analysis is performed to reveal which field attributes contributed significantly towards classifying yield plots into clusters.

The authors of [13] follow the idea of assessing spatial and temporal variability on irrigated fields with management zones. First, a principal components analysis is performed. Out of the five principal components, the first two explain 85% of the total variability and are therefore retained for the management zone delineation approach. The authors do not elaborate in more detail how the delineation is performed, except for the fact that an unsupervised classification is applied. This is likely to be a clustering approach. The resulting four management zones are quite distinct in their soil properties and also in their yield variabilities.

[8] proposes an idea for semi-automatic management zone delineation using fuzzy c-means clustering. They use yield, elevation and EC_a data from ten consecutive years on claypan soil fields in north-central Missouri, USA, growing corn and soybeans. They split the data into two sets, one containing the yield information and the other one carrying the EC_a and elevation information. Both data sets are then clustered using fuzzy c-means clustering as built into the management zone analyst software [4].

[7] investigates the relationship between electromagnetic sensing (EC_a) and yield mapping for delineation of management zones. On these data, fuzzy clustering is applied to determine management zones, where the optimal number of zones is determined by using the normalized classification entropy value, resulting in *four* management zones. The zones based on the yield maps seem to be closely related to some of the soil properties, which is determined by regression modeling.

[1] investigate the delineation of management zones for a corn-soybean rotation in east-central Indiana, USA. Special attention is paid to evaluate the usage of fuzzy c-means clustering for delineating zones and to compare different management zones for different crops with each other. The data sets consist of multi-year precise yield data, collected at 1 s-intervals using yield monitoring systems, coupled with a differential GPS. After preprocessing, these data are available to the management zone delineation approach. Both objectives laid out above are achieved by running a fuzzy clustering algorithm on the yield data and post-processing the results according to the fuzzy performance index and the normalized classification entropy.

[10] present an approach that uses remote sensing and sampled soil data to delineate management zones in coastal saline land in northern China. The available data for this field are the normalized differenced vegetation index (NDVI) from satellite imagery, 139 soil samples on an almost regular grid and EC_a measurements. The soil samples are processed into the following seven variables: OM, TN, AN, AP, AK and CEC. Furthermore, the end-of-current-season yield is available. Following a conventional statistical analysis, a principal components analysis is performed, resulting in two principal components which together explain 88% of the total variability in the data set. Based on these principal components maps, a fuzzy c-means algorithm is used to delineate three management zones. The number of management zones is determined via the fuzzy performance index and the normalized classification entropy, both yielding an optimal number of three components.

In [11], a cluster-based approach to management zone delineation is compared with two additional ad-hoc approaches fulfilling the same purpose. The data are sampled from 13 fields in Central Chile with an average density for 8.1 and 11.9 soil samples per hectar, for the 2002/2003 and 2003/2004 corn-growing seasons, respectively. The soil samples which are taken are analyzed for six chemical properties: pH, EC_a, OM (organic matter), N (available nitrogen), P (extractable phosphorus) and K (potassium). First, a k-means clustering is computed for the field, based on the standardized variable values. A fixed number of four management zones is chosen because this is the maximum number of management zones that can be handled with conventional equipment. Second, based on the available six variables plus the dry-matter yield variable, a SI (soil index) variable is computed which is then mapped and classified into management zones.
[15] present a further, very recent approach to the delineation of management zones. Their data are based on manually taken soil samples on a 100 m grid, resulting in 81 points for this 87 ha field.
Fuzzy clustering is applied on the first and on the combination of the first and second principal components (PCs) of the data set. The first PC explained roughly 50% of the total variance and the second an additional 21%. The optimal number of clusters is determined via the FPI (fuzzy performance index) and the NCE (normalized classification entropy).
[14] undertake an approach of using data layers of soil sampling, yield and remote sensing data. Fuzzy c-means clustering is used to create management zone maps. The resulting maps are compared via the coefficient of variation of the different variables.

3 Towards a Novel Clustering Approach for Management Zone Delineation

Most of the aforementioned research focuses on partitioning the field via a clustering approach. This is a natural path to follow, since clustering approaches have been explicitly developed for this purpose. However, none of the authors consider the issue of spatial autocorrelation since the available data records in each article are, without exception, considered to be independent. This often misleads the clustering approaches into generating non-contiguous management zones. Often, a three- or two-step process as depicted in Table 1 is performed.
Based on the special properties of the available agriculture data in this work, standard clustering approaches can not be used directly here. First of all, the available data sets consist of vectors which have a *data part* (results from sensors) and a *geographical location part* which is essentially the point's northing, easting and elevation. Furthermore, the data records are aligned on a regular grid (with some irregularities), and they exhibit spatial autocorrelation. These properties prevent the usual clustering approaches from succeeding: density-based clustering approaches can not be used since the data point density is rather homogeneous on the field. Grid-based approaches may have issues with the grid irregularities. Simple partitional approaches such as k-means or fuzzy c-means are likely to produce non-contiguous field areas.

phase	traditional approaches	improved clustering approach
1	tessellation of field via grid	k-means tessellation
2	(fuzzy) clustering of data records	clustering with spatial constraint
3	Smoothing / Filtering	—

Table 1: Comparison of frequently used management zone delineation approaches vs. hierarchical agglomerative clustering with spatial constraint

However, with slight modifications, hierarchical clustering approaches with a spatial constraint may be used in a straightforward manner. This will be described in the following section.

4 Hierarchical Agglomerative Clustering with Spatial Constraint

Hierarchical agglomerative clustering (HAC) is a bottom-up technique to generate a tree-like structure of clusters, a dendrogram. In each level of the dendrogram, a full clustering of the underlying data is depicted. HAC usually starts at the level of single data records and consecutively merges records and/or clusters, thereby creating the dendrogram. In the agriculture data, hierarchical clustering may be applied, given that two specialties are taken into account. First, due to spatial autocorrelation, the lowest level from which HAC starts may be replaced by small contiguous zones which can be generated using a spatial tessellation. Second, HAC may proceed as usual but should consider a spatial constraint: since the resulting management zones are suggested to be contiguous, only spatially neighboring zones are to be merged.

The first step of spatially *partitioning* the data points may be achieved by overlaying a grid. Due to the irregularities in the field shape and the gaps and holes in natural data density, running a k-means algorithm on the coordinates of the points in the data set provides a more flexible solution to the initial tessellation. An upper bound for the parameter k is given by the size of the resulting smallest zone – zones below a threshold provided by the precision of the used farming equipment can not be managed. A lower bound for the k parameter is set by the granularity of the final management zones and by the amount of heterogeneity on the field – for more heterogeneity, a higher k value should be used.

The second step of repetitively merging two zones has two constraints: first, zones which are to be merged must be similar in their attributes; second, they must be direct neighbors in geographical space (spatial constraint). While the first condition ensures that the resulting zones are rather homogeneous, the second condition ensures that the resulting zones are contiguous. In future work, the second condition (spatial constraint) may be weighted to allow for non-contiguous areas to form one zone if they are otherwise similar enough.

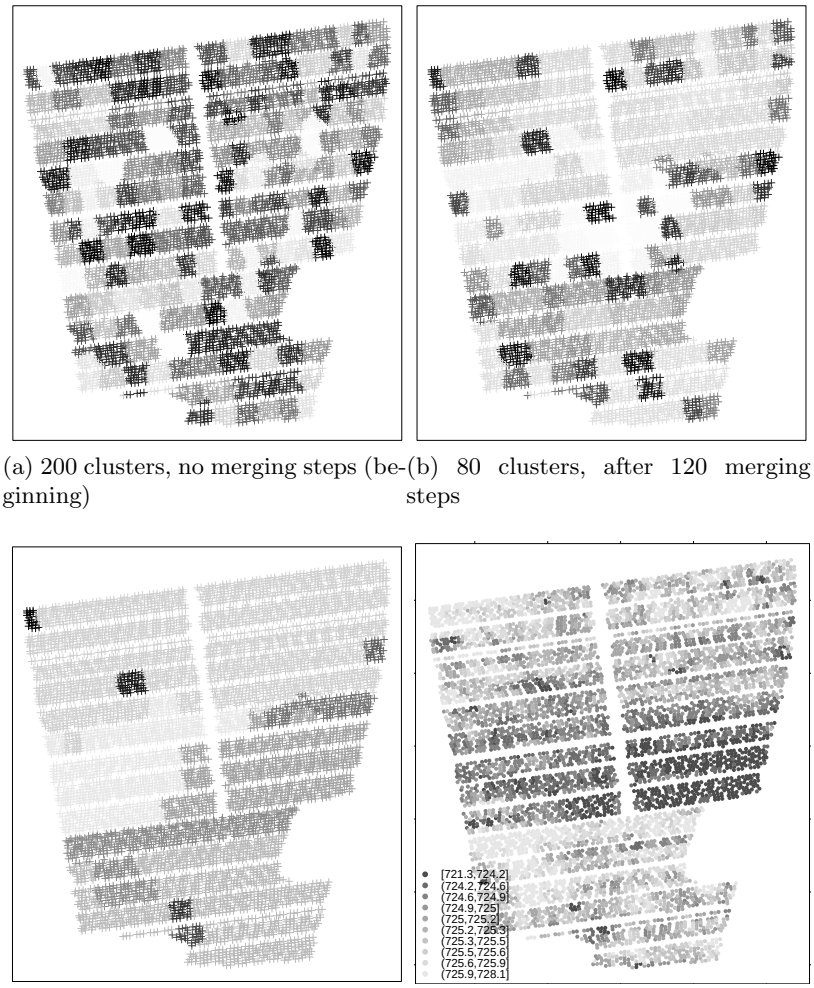
The spatial constraint can easily be fulfilled by generating a list of neighbors for each cluster and updating this list accordingly on cluster merge steps. The similarity measure which decides which of the spatially neighboring clusters are to be merged is usually one of *single-linkage*, *complete-linkage* or *average-linkage* (cp. [5] for an overview). Single-linkage would determine the data records which are most similar in two candidate clusters and merge the clusters containing these records. Due to spatial autocorrelation in these data, neighboring clusters share a common border at which neighboring data records are expected to be very similar. Therefore, this criterion is prone to choose clusters which are similar at the common border, but not otherwise. Complete-linkage would determine those data records which are most dissimilar in candidate clusters and merge those clusters. Again, due to spatial autocorrelation and especially once the clusters grow, the most dissimilar data records are not expected to characterize the neighboring clusters sufficiently well. Average-linkage would determine the average vector of the data records in one cluster and would then merge those clusters which are closest according to the (usually Euclidean or Cosine) attribute distance. It is assumed that an average vector characterizes a (spatial) cluster sufficiently well.

Since the agricultural necessity of zones is not targeted towards an automatic process, but rather an exploratory one, it should be left to the expert user to investigate the resulting management zones.

5 Results

The aforementioned hierarchical agglomerative clustering approach with spatial constraint has been evaluated on two data sets from precision agriculture sites. These consist of multiple data layers which have been recorded on two fields in Northern Germany. The data consist of measurements of biophysical variables such as apparent electrical conductivity (EC25) and vegetation indices (red edge inflection point – REIP), as well as controllable variables such as nitrogen fertilizer and yield. These data are available in a 10-by-10-meter spatial resolution, after kriging the original measurements. Since the usability of the clustering method has not yet been evaluated in practice, the method is presented here in a simplified version for demonstrative purposes. The REIP32 value is shown in Figure 1d. The initial tessellation of the field into spatially contiguous small clusters is shown in Figure 1a, with a setting of 200 initial clusters. The process outcome is depicted after 120 merging steps (Figure 1b) and further 70 merging steps (Figure 1c).

The zones in Figure 1c correlate well with the actual REIP32 measurements shown. The zones are necessarily contiguous since this has been enforced by the design of the clustering algorithm. The zones may be non-convex which can not usually be achieved using classical clustering algorithms. The small outlier clusters can easily be merged into the neighboring zones when comparing them with the REIP32 map.



(a) 200 clusters, no merging steps (beginning)

(b) 80 clusters, after 120 merging steps

(c) 10 clusters, after 190 merging steps

(d) REIP32 value

Fig. 1: Three stages of the clustering process, compared with the original REIP32 map

6 Conclusions and Future Work

This article elaborated upon the algorithmic side of an important task in modern precision agriculture: management zone delineation. Based on the existing research work in agriculture in conjunction with the specialties of the available geographically located data sets, a novel clustering algorithm has been developed which fulfils the requirements of management zone delineation approaches. The resulting zones are necessarily contiguous and reflect the heterogeneity of the map well. Since the approach is meant to be used in an exploratory data analysis manner, the parameter for the initial tessellation size (k in k-means) and the stopping criterion for the merging process are to be determined on-the-fly. In future work, the second condition (spatial constraint) of the algorithm may be weighted to allow for non-contiguous areas to form one zone if they are otherwise similar enough.

Note that the data sets which the algorithm was evaluated upon are not directly suitable for management zone delineation when considering the agricultural task of base fertilization. However, the measurements of electrical conductivity and the vegetation indices exhibit spatial autocorrelation and are expected to be characteristic for the (mineral) measurements of soil sampling. Therefore, the algorithm will be evaluated once these soil sampling data become available.

The algorithm will also be evaluated using multiple variables as input, e.g. such that the resulting zones are based on REIP and EC instead of either of these. For the sake of algorithmic clarity, this article focused on presenting the novel algorithm and illustrating its purpose rather than yielding direct results.

Acknowledgements

The data sets were obtained from Peter Wagner and Martin Schneider from the experimental precision agriculture sites of the Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für landwirtschaftliche Betriebslehre, Germany.

References

1. A. Brock, S. M. Brouder, G. Blumhoff, and B. S. Hofmann. Defining yield-based management zones for corn-soybean rotations. *Agronomy Journal*, 97(4):1115–1128, 2005.
2. David W. Franzen and T. Nanna. Comparison of nitrogen management zone delineation methods. In *Proceedings of the North Central Extension-Industry Soil Fertility Conference*, volume 19, Des Moines, IA, 2003.
3. David W. Franzen and T. Nanna. Use of data layering to address changes in nitrogen management zone delineation. In *USDA Forest Service Proceedings*, 2006.

4. Jon J. Fridgen, Newell R. Kitchen, Kenneth A. Sudduth, Scott T. Drummond, William J. Wiebold, and Clyde W. Fraisse. Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation. *Agronomy Journal*, 96(1):100–108, 2004.
5. A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.
6. D.B. Jaynes, T.C. Kasper, T.S. Colvin, and D.E. James. Cluster analysis of spatiotemporal corn yield patterns in an iowa field. *Agronomy Journal*, 95(3):574–586, 2003.
7. J. A. King, P. M. R. Dampney, R. M. Lark, H. C. Wheeler, R. I. Bradley, and T. R. Mayr. Mapping potential crop management zones within fields: Use of yield-map series and patterns of soil physical properties identified by electromagnetic induction sensing. *Precision Agriculture*, 6:167–181, 2005.
8. N.R. Kitchen, K.A. Sudduth, D.B. Myers, S.T. Drummond, and S.Y. Hong. Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. *Computers and Electronics in Agriculture*, 46(1-3):285 – 308, 2005. Applications of Apparent Soil Electrical Conductivity in Precision Agriculture.
9. R. M. Lark. Forming spatially coherent regions by classification of multi-variate data: an example from the analysis of maps of crop yield. *International Journal of Geographical Information Science*, 12(1):83–98, 1998.
10. Yan Li, Zhou Shi, Feng Li, and Hong-Yi Li. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. *Comput. Electron. Agric.*, 56(2):174–186, 2007.
11. Rodrigo A. Ortega and Oscar A. Santibez. Determination of management zones in corn (zea mays l.) based on soil fertility. *Computers and Electronics in Agriculture*, 58(1):49 – 59, 2007. Precision Agriculture in Latin America.
12. Georg Ruß and Alexander Brenning. Data mining in precision agriculture: Management of spatial information. In *Proceedings of IPMU'2010*, pages –. Springer, July 2010. accepted for publication.
13. Aaron R. Schepers, John. F. Shanahan, Mark A. Liebig, James S. Schepers, Sven H. Johnson, and Ariovaldo Luchiari. Appropriateness of Management Zones for Characterizing Spatial Variability of Soil Properties and Irrigated Corn Yields across Years. *Agronomy Journal*, 96(1):195–203, 2004.
14. Xiaoyu Song, Jihua Wang, Wenjiang Huang, Liangyun Liu, Guangjian Yan, and Ruiliang Pu. The delineation of agricultural management zones with high resolution remotely sensed data. *Precision Agriculture*, 1:not available, 2009.
15. Wang Xin-Zhong, Liu Guo-Shun, Hu Hong-Chao, Wang Zhen-Hai, Liu Qing-Hua, Liu Xu-Feng, Hao Wei-Hong, and Li Yan-Tao. Determination of management zones for a tobacco field based on soil fertility. *Computers and Electronics in Agriculture*, 65(2):168 – 175, 2009.