Machine Learning Methods for Spatial Clustering on Precision Agriculture Data

Georg RUSS [a,1] Rudolf KRUSE [a]

[a] *Otto-von-Guericke-Universität Magdeburg, Germany*

Abstract.
Machine learning techniques typically result from the need for intelligent solutions to practical tasks. Nowadays, large data volumes are usually involved and machine learning techniques are focused on particular tasks like classification, regression or clustering. For the latter task, clustering, quite a few algorithms have been proposed, typically tailored to particular application domains and their data sets. Recently, georeferenced (or *spatial*) data sets keep emerging in lots of disciplines. Therefore, algorithms which are able to handle these spatial data sets should be developed. This article shortly describes a particular application area, precision agriculture, and the spatial data sets which exist there. A particular task from this area, management zone delineation, is outlined and existing spatial clustering algorithms are evaluated for this task. Based on the experiences with those algorithms and a few requirements, HACC-SPATIAL is developed. The algorithm is based on hierarchical agglomerative clustering with a spatial constraint and it is demonstrated to produce practically advantageous results on precision agriculture data sets.

Keywords. Spatial Clustering, Machine Learning, Management Zone Delineation

1. Motivation and Application Area

Figure 1 shows a part of the data sets which are of interest here. Data such as these are nowadays collected regularly. They are obtained in precision agriculture operations, which would best be described as a data-driven and GPS-based approach to agriculture. A number of ground-based sensors, aerial imagery and other variables are regularly collected and, due to (differential) GPS, georeferenced at a high spatial resolution up to a few centimeters. This article deals with these data sets and a specific task for spatial clustering.

1.1. Problem Description

Given data sets from precision agriculture, the classical question of *management zone delineation* should be answered: are there parts or zones on the field which can be treated

[1]Corresponding Author: Georg Ruß, Otto-von-Guericke-Universität Magdeburg, Germany E-mail: georg.russ@ovgu.de

similarly, for example during fertilization operations? In machine learning, this is treated as a clustering task: given a data set, which of the points in the data set are similar and thus fall into one cluster? However, since the data in precision agriculture are of spatial nature, any clustering algorithm to be used must account for this spatial nature.

In this article, a real precision agriculture data set will be used, consisting of 1079 spatial data records for the following variables: pH value, potassium (K) content, magnesium (Mg) content, phosphorus (P) content, apparent electrical soil conductivity (EC25) and yield. These data records result from precision agriculture operations and are preprocessed to a regular grid – each grid cell is represented by a data vector containing the previously mentioned variable values at this particular cell. The data can also be thought of as having the above variable layers.

Personal experience shows that currently for management zone delineation only the EC25 variable is used. It is clearly stated that it would indeed be worthwile to use further available data variables, but that there is a lack of readily usable methods tailored to this particular task. Therefore, this article aims to develop a spatial clustering algorithm fit for these particular data sets.

2. Task and Algorithm Requirements

A definition of a spatial cluster with regard to geographical information systems is provided in [9]:

> A spatial cluster might then be defined as an excess of [. . .] values (for field-based data, such as a grouping of excessively high concentrations of cadmium in soils) in geographic space. [. . .] For now, it is useful to think of a "cluster" as a spatial pattern that differs in important respects from the geographic variation expected in the absence of the spatial processes that are being investigated.

With the aforementioned data sets at hand, a few further requirements for the outcome of any spatial clustering algorithm for management zone delineation are specified in the following.

exploratory nature The process of management zone delineation is typically user-dependent and requires user intervention. Therefore, a developed clustering algorithm is not supposed to result in an *optimal* outcome where the algorithm performance can be measured and confirmed. It is rather supposed to be supportive in providing the user with suggestions for the zone delineation.

understandability End users of a geographical information system are much more likely to apply an algorithm if it can be easily understood. Therefore, an algorithm which is unlike a "black-box" algorithm is suggested.

spatial contiguity It would certainly be desirable to have spatially contigous management zones. It can, however, not be ensured due to the heterogeneity of a field. Therefore, this contiguity should be seen as a soft constraint which may be violated up to a certain degree.

spatial autocorrelation It may safely be assumed that the available data sets are spatially autocorrelated: data records which are close to each other are very likely to have similar variable values in contrast to data records which are rather distant from each other. A spatial clustering algorithm should exploit this data property accordingly.

3. Literature Review on Spatial Clustering

Given the data set presented in the preceding section, the task is to establish an algorithm which is able to delineate the field into spatially (mostly) contiguous clusters, so-called management zones. From a machine learning point of view, the task is the following: given a set of georeferenced data records consisting of a certain number of variables, find a spatial tessellation of these data records which is appropriate for management zone delineation. Since it is as of now unclear which of the available variables contribute to the physical and biological underpinnings of management zones [11], the above broad task should be narrowed to the following: develop an algorithm for the above type of data sets which returns a spatially (mostly) contiguous tessellation and which can be easily parameterized by a human expert.

In precision agriculture, there are a number of approaches using standard clustering algorithms such as fuzzy c-means clustering [12,13,15]. However, these rely solely on the data records' variables and totally neglect the spatial structure of the data records. This results in zones which are non-contiguous and spread over the whole field, as well as small islands of outliers and insignificant records which must be smoothed out manually after the clustering. A similar approach is undertaken by fuzzy classification of the data records, which exhibits the same problems [14]. In addition, there is no clear guidance available as to which input variables enable a successful management zone delineation [4,17]. It seems, however, clear that management zones must rely on more than just yield data [11].

In the area of computer science, there are, to the best of the authors' knowledge, no clustering algorithms which would allow tackling the above task on the given type of data sets. Density-based algorithms like DBSCAN [5], CLIQUE [1] or STING [22] usually rely on a non-uniform distribution of the data records (density differences) to find clusters. With our data sets, the records are spatially uniformly distributed on a grid, which renders the aforementioned algorithms useless. An extension to include geographic information, as presented in MOSAIC [3], would be appropriate, but MOSAIC does not distinguish between geographic space and feature space. Algorithms like SKATER [2] and REDCAP [7] are different in that they explicitly incorporate spatial contiguity constraints into the clustering process. However, these algorithms may fail to report adjacent clusters correctly (SKATER) or are too strict in terms of management zone contiguity (REDCAP). In addition, they both rely on the fact that data records are spatially non-uniformly distributed, which is not the case here. This last assumption is also used by ICEAGE [8], which is therefore not applicable either. CLARANS [16] is a further algorithm designed for clustering spatial data sets but is based on the assumption that the structure to be discovered is hidden exclusively in the spatial part of the data, which is not the case here. Finally, AMOEBA [6] works on two-dimensional spatial data by building a hierarchy of spatial clusters using a Delaunay triangulation, but lacks the extension to non-spatial variables and also assumes that the 2D points are non-uniformly distributed in space.

4. Recommendations towards a Novel Approach

One of the more common approaches to spatial clustering is a hierarchical agglomerative one: start with each point in a single cluster and subsequently merge clusters according to some criterion or constraint. Further research into constraints-based clustering [20] reveals that it may in principle be applied here. The author of [20] explicitly describes the "spatial contiguity" constraint for spatial data as a type of global clustering constraint using neighborhood information, albeit for image segmentation. The constraints are presented as "hard" or "soft", meaning that the final clustering outcome "must" or "can" consider these constraints. The task encountered in this article, namely generating *mostly contiguous* clusters, could therefore be tackled by using a soft spatial contiguity constraint. An additional feature of constrained clustering algorithms is the existence of "must-link" and "cannot-link" pairwise constraints for data records. Although an algorithm can usually be constructed this way or the other, it seems more appropriate to model the spatial contiguity requirement as a "cannot-link" (soft) constraint for spatially non-adjacent data records or clusters. In addition, the work of [21] encounters a similar agricultural problem to the one in this article, but the focus is slightly shifted to yield prediction on a county scale with low-resolution data, rather than using high-resolution data for management zone delineation. Since the focus in this work is more on exploratory data mining in an unsupervised setup we postpone the performance question.

Additionally, hierarchical agglomerative clustering seems like a rather natural approach since the solution ultimately has to be presented to domain experts who typically prefer easy-to-understand solutions over black-box models. Therefore, our focus will be on developing a hierarchical agglomerative algorithm for zone delineation which takes the special properties of the data sets into account. Our data sets are different from the ones in existing work since the data records are located on a uniformly spaced hexagonal grid and exhibit spatial autocorrelation. This autocorrelation will be used explicitly in our approach.

5. HACC-spatial

This section presents an extended and refined version of the hierarchical, divide-and-conquer approach to delineating spatially mostly contiguous management zones based on precision agriculture data presented in [18,19]. Our approach can best be described as *hierarchical agglomerative clustering with a spatial contiguity constraint* (HACC-SPATIAL) and an additional (optional) initialization step which exploits the spatial autocorrelation in the data. It consists of two phases, in a divide-and-conquer manner. First, the field is tessellated into a fixed number of (spatial) clusters. Second, these clusters are merged iteratively, using a similarity measure and adhering to a spatial contiguity constraint, which shifts from being a hard constraint to a soft constraint throughout the algorithm. The algorithm is given in pseudo-code in Algorithm 1 (syntax close to R).

Phase 1 of HACC-SPATIAL is intended to create small initial clusters or single objects for the second phase of the algorithm. For hierarchical agglomerative clustering on single objects no further action needs to be taken. However, due to spatial autocorrelation, spatially neighboring data records are likely to be very similar in their variables. Therefore, by tessellating the field into a fixed number of spatial clusters $n \leq N$, the

Algorithm 1. HACC-SPATIAL

\# input:
\# V . . . set of i georeferenced data vectors
\# k – tessellation resolution, $k \leq i$
\# cp – contiguity constraint parameter
\# output: a dendrogram of the hierarchical clustering

\# split phase, run k-means clustering on spatial locations of data vectors
C $\leftarrow$ k-means(V, k)
return spatial clustering C

\# merging phase, iteratively merge clusters according to cp
spatialconstraint $\leftarrow$ “hard”
repeat
\# determine and store cluster distances
for each spatially adjacent cluster pair (c_i,c_j) $\in C$ **do**
dist$_a$[i,j] $\leftarrow$ dist(c_i,c_j)
end for
for each spatially non-adjacent cluster pair (c_i,c_j) $\in C$ **do**
dist$_{\overline{a}}$[i,j] $\leftarrow$ dist(c_i,c_j)
end for
\# determine minimum/median distances and contiguity
mindist$_a$ $\leftarrow$ minimum(dist$_a$), meddist$_a$ $\leftarrow$ median(dist$_a$)
mindist$_{\overline{a}}$ $\leftarrow$ minimum(dist$_{\overline{a}}$), meddist$_{\overline{a}}$ $\leftarrow$ median(dist$_{\overline{a}}$)
contiguity $\leftarrow -\frac{\text{meddist}_{\overline{a}}}{\text{meddist}_a}$
\# switch from hard to soft constraint when cp is reached
if contiguity $\geq$ cp **and** spatialconstraint = “hard” **then**
spatialconstraint $\leftarrow$ “soft”
end if
if spatialconstraint == “hard” **then**
clusterpair $\leftarrow$ which(dist$_a$ == mindist$_a$, arr.ind=TRUE)
else
if mindist$_a$ $\leq$ mindist$_{\overline{a}}$ **then**
clusterpair $\leftarrow$ which(dist$_a$ == mindist$_a$, arr.ind=TRUE)
else
clusterpair $\leftarrow$ which(dist$_{\overline{a}}$ == mindist$_{\overline{a}}$, arr.ind=TRUE)
end if
end if
i $\leftarrow$ clusterpair[1], j $\leftarrow$ clusterpair[2]
$C \leftarrow C \setminus (c_i, c_j)$ # remove most similar cluster pair
$C \leftarrow C \cup (c_i \cup c_j)$ # add newly merged cluster
until number of clusters $= 1$
return dendrogram of management zones C

clusters are still very likely to contain similar (adjacent) data records while some of the ensuing computational effort of the merging step can be saved. With the above prerequisites, the simplest tessellation approach fulfilling the requirements is to perform a k-means clustering on the data records' spatial coordinates. This explicitly assumes that, due to spatial autocorrelation, the resulting spatial clusters contain similar data records. This phase may be omitted, such that the second phase would then begin with each point in a single cluster.

Phase 2 of HACC-SPATIAL starts either with the small contiguous clusters or from single data records as clusters. The idea is to merge these clusters consecutively into larger clusters. In addition to the standard similarity or distance measure, a spatial constraint is taken into account. Since the final result of the clustering is assumed to be a set of spatially mostly contiguous clusters, only those clusters should be merged which are a) similar (with regard to their variables' values) and b) spatial neighbors (adjacent).

In classical hierarchical clustering, the standard measures for cluster similarity are single linkage, complete linkage and average linkage [10]. However, when considering the spatial data encountered here, these three criteria merit some explanation. *Single linkage* determines cluster similarity based on the smallest distance between objects from adjacent clusters. Due to spatial autocorrelation, it is likely that there are always some points at the borders of the clusters which are very similar, for each neighbor. Therefore, single linkage will not provide us with a good measure for which neighbor to choose. *Complete linkage* determines the similarity of neighboring clusters based on the distance of those objects which are farthest away from each other. Since we are considering spatially adjacent clusters, this would lead to very dissimilar clusters being merged. Due to spatial autocorrelation, these objects would also be spatially rather far away from each other, which leads to a chaining effect and less meaningful clusters. *Average linkage* determines the similarity of adjacent clusters based on the average of the (Euclidean or other) distances between all objects in the clusters. A combination of the aforementioned arguments for single and complete linkage may be applied here: points in adjacent clusters which are spatially close/far apart are likely to also be very similar/dissimilar. Therefore, an appropriate distance for adjacent clusters may be determined by *average group linkage*: we compute an average vector for each cluster and determine the distance between these vectors.

It is not required that one zone is strictly contiguous, i.e. consists of just one spatially contiguous area on the field. It is a valid result if one zone comprises those data records which are similar but is made up of two or more larger areas on the field. This *"mostly contiguous"* description should be seen as a soft constraint in the final merging steps. To prevent the algorithm from producing too many scattered zones, we propose to set it as a hard constraint during the beginning of the merging phase. As long as adjacent clusters are similar enough, these are merged. If this is not the case, clusters which are not direct neighbors of each other may be merged if they are similar enough. The spatial constraint is changed from a hard to a soft constraint as soon as the contiguity ratio (see Algorithm 1) is exceeded. In the results for Figure 2, the algorithm performs well with the hard constraint in the beginning and would switch to a soft constraint only after the bottom plot, which has 28 clusters left, with the contiguity threshold set to -2.

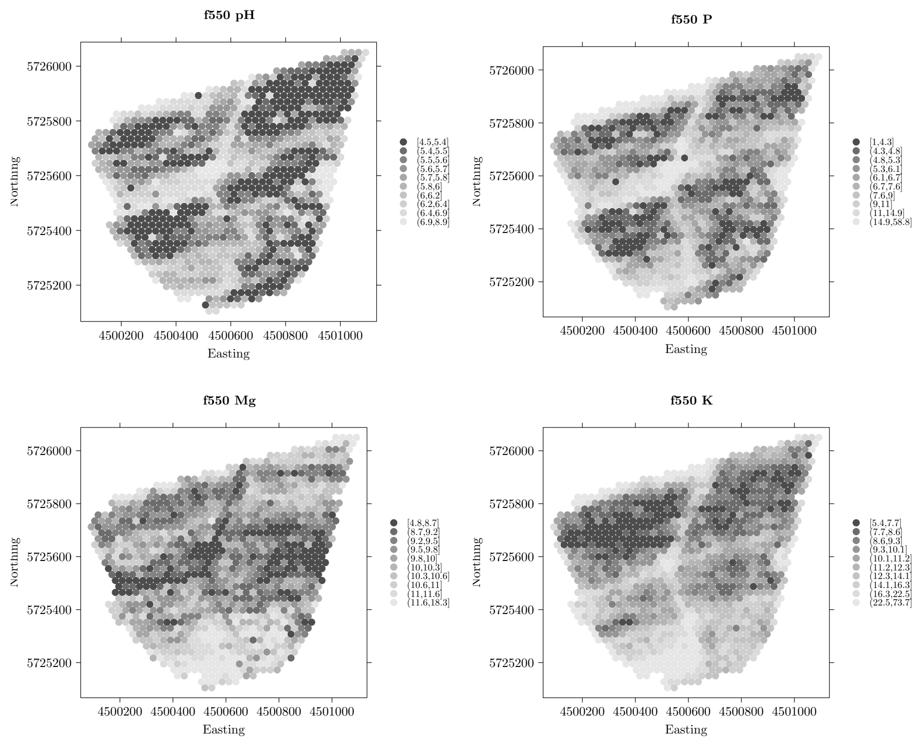

Figure 1. Four chosen variables for which the management zone delineation is applied: pH value, P, Mg, K concentration

6. Results and Discussion

We now demonstrate the algorithm on multi-variate data and start with a correlated subset of the original data set: we choose the four soil sampling variables (pH-value, P, K, Mg content). From the four plots in Figure 1 it can be seen that a certain spatial structure is emergent, with four to six visible areas, separated by another cross-shaped area in the middle. This structure is the one we would like HACC-SPATIAL to discover.

The data set has 1079 spatial data records. As mentioned in the algorithm description, a hierarchical agglomerative clustering procedure may start with each of the data records forming one cluster. However, due to spatial autocorrelation, spatially adjacent data records are likely to be similar and are therefore grouped by using a k-means clustering on the spatial part only. This is depicted in the top left figure of Figure 2: we choose k to be 350, such that on average three neighboring data records are in one cluster initially. The algorithm then proceeds to consecutively merge adjacent, similar clusters. This is depicted in Figure 2, top right and bottom left plot, with 250 and 150 clusters left, respectively. The final plot in Figure 2 shows the outcome with 28 clusters left. We can roughly see six zones. Those at the borders are, of course, not (yet) zones in the sense of the algorithm, but they are easily visually distinguishable. For an exploratory data mining task, this result is what the algorithm is supposed to deliver.

Upon further examination of the resulting six zones, it turns out that these are actually just three zones. Comparing the clustering result with the original data set yields

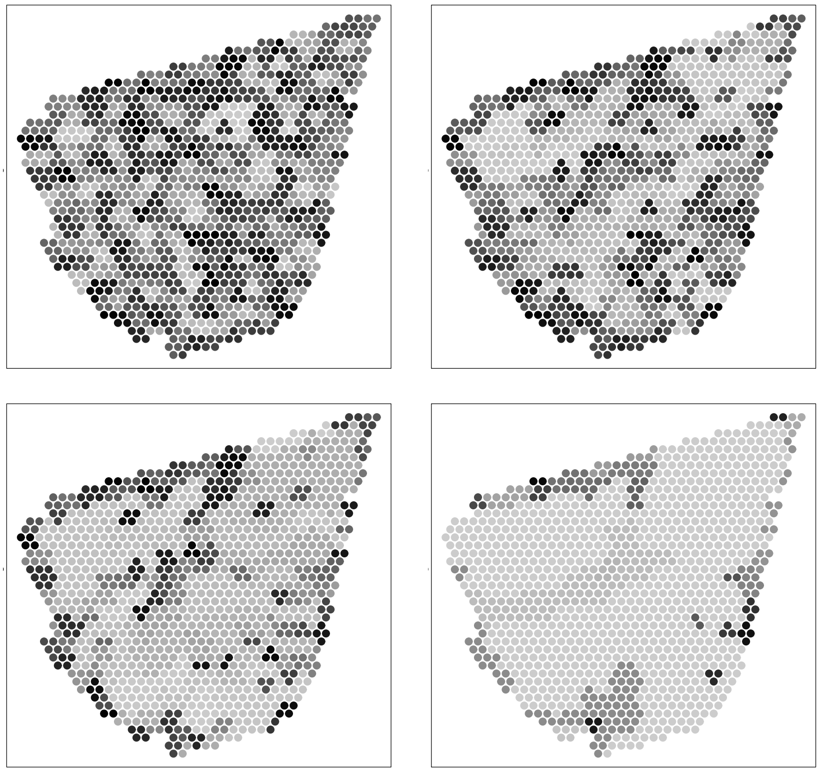

Figure 2. Clustering on the four variables from Figure 1, beginning of clustering (350 clusters), after 100/200 merging steps, with 28 clusters left (left to right, top to bottom)

the following: the largest zone which covers roughly 80% of the field could be described with *low pH, low P, medium/low Mg, low K*. The border zones on the top left, the left and the bottom left of the field can be described with *high pH, high P, high Mg, high K*. The small zone at the right field border and the one extending from the left border mostly horizontally into the middle would be *high pH, high P, low Mg, high K*. For practical purposes of basic fertilization this simple characterization of a field's principal zones is very convenient.

Setting the parameter k for the k-means tessellation depends on the data set. For rather homogeneous fields, this can be set to a lower value such as $\frac{N}{10}$, where N is the number of available data records. For rather heterogeneous data sets such as the one encountered here, we may set it to as high as $\frac{N}{3}$, thereby combining roughly three adjacent data records into one initial cluster. If the number k of initial clusters is set to N, we obtain a setting which may be used for data where no spatial autocorrelation exists.

Setting the contiguity ratio threshold is rather straightforward: a high value (in the implementation here: ≥ -1) leads to a later switch from a hard to a soft constraint – therefore, the spatial contiguity is higher. A value smaller than, but closer to -2 further weakens this hard constraint. A value ≤ -3 favors the merging of non-adjacent clusters early in the algorithm, probably resulting in rather scattered zones. For other data sets, this parameter setting may vary, and depending on the implementation, the scale may change. The *average-linkage* similarity computation using Euclidean distance may be replaced by a different distance measure. For higher numbers of variables, the Cosine distance measure may be employed.

7. Summary and Future Work

This article presented a hierarchical agglomerative clustering approach with a spatial constraint for the task of management zone delineation in precision agriculture. Based on the specifics of the data sets from precision agriculture, namely the uniform spatial distribution of the data records on a hexagonal grid and the existence of spatial autocorrelation, we established and recognized the shortcomings (or the lack) of existing approaches. Henceforth, we specified the requirements of a novel approach: the spatial contiguity of the resulting zones and the explicit assumption of spatial autocorrelation.

This research lead to a two-phase divide-and-conquer approach. In the first phase we tessellated the field using k-means on the data records' 2D coordinates. In the second phase, we iteratively merged those spatially adjacent clusters that are similar. This was done in two sub-phases: in the first sub-phase, the spatial contiguity was a hard constraint, meaning that only adjacent clusters may be merged. In the second sub-phase, this was relaxed to a soft constraint. Switching from the hard to the soft constraint can be user-influenced by a contiguity factor cf. Proceeding like this provided us with a hierarchical structure which can then be examined by a human expert for guidance on the management zone delineation. Our focus was on providing an exploratory and easy-to-understand approach rather than a fixed, black-box solution. Our approach worked successfully for spatially autocorrelated precision agriculture data sets. The parameter setting for k (initial tessellation) was explained. An additional parameter cf was suggested for further analysis on the spatial contiguity of the resulting clusters.

Once the clustering algorithm finishes, a certain clustering should usually be examined further. The clusters may easily be examined using frequent itemset mining. Numerical variables can be converted to a three- or five-value categorical scale and the resulting frequent sets could be generated as we did manually for the bottom plot of Figure 2. Although the *average linkage* similarity calculation turns out to work rather well in practice, it may be further researched whether different linkage criteria in combination with other similarity measures could be more appropriate. A drawback of our work is the lack of reference data sets from precision agriculture and similar domains in conjunction with a similar task. We are currently investigating the possibility of making our data sets publicly available for this purpose.

This work has been researched and implemented during a research stay in the database group of Kjetil Nørvåg and Heri Ramampiaro at NTNU Trondheim, IDI, August–October 2010.

References

[1] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD '98: Proc. of the 1998 ACM SIGMOD Int. Conf. on Management of Data*, pages 94–105, New York, NY, USA, 1998. ACM.

[2] R. M. Assuncao, M. C. Neves, G. Camara, and C. Da Costa Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006.

[3] Jiyeon Choo, Rachsuda Jiamthapthaksin, Chun-sheng Chen, Oner Celepcikay, Christian Giusti, and Christoph Eick. Mosaic: A proximity graph approach for agglomerative clustering. In *Data Warehousing and Knowledge Discovery*, pages 231–240, 2007.

[4] Horst Domsch, Michael Heisig, and Katrin Witzke. Estimation of yield zones using aerial images and yield data from a few tracks of a combine harvester. *Precision Agriculture*, 9:321–337, 2008.

[5] Martin Ester, Hans-Peter Kriegel, Joerg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *2nd Int. Conf. on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[6] Vladimir Estivill-Castro and Ickjai Lee. Multi-level clustering and its visualization for exploratory spatial analysis. *GeoInformatica*, 6(2):123–152, June 2002.

[7] D. Guo. Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). *International Journal of Geographical Information Science*, 22(7):801–823, 2008.

[8] Diansheng Guo, Donna J. Peuquet, and Mark Gahegan. ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. *Geoinformatica*, 7(3):229–253, 2003.

[9] Geoffrey M. Jacquez. Spatial cluster analysis. In Stewart Fotheringham and J. Wilson, editors, *The Handbook of Geographic Information Science*, pages 395–416. Blackwell Publishing, 2008.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Survey*, 31(3):264–323, 1999.

[11] R. Khosla, D. Inman, D. G. Westfall, R. M. Reich, M. Frasier, M. Mzuku, B. Koch, and A. Hornung. A synthesis of multi-disciplinary research in precision agriculture: site-specific management zones in the semi-arid western Great Plains of the USA. *Precision Agriculture*, 9:85–100, 2008.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Raymond T. Ng and Jiawei Han. Clarans: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5):1003–1016, 2002.

[17] Rodrigo A. Ortega and Oscar A. Santibáñez. 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.

[18] Georg Ruß, Rudolf Kruse, and Martin Schneider. A clustering approach for management zone delineation in precision agriculture. In Rajiv Khosla, editor, *Proceedings of the Int. Conf. on Precision Agriculture 2010*, July 2010.

[19] Georg Ruß, Martin Schneider, and Rudolf Kruse. Hierarchical spatial clustering for management zone delineation in precision agriculture. In Isabelle Bichindaritz, Petra Perner, and Georg Ruß, editors, *Advances in Data Mining*, pages 95–104, Leipzig, Germany, July 2010. IBaI Publishing.

[20] Kiri Lou Wagstaff. *Intelligent Clustering with Instance-Level Constraints*. PhD thesis, Cornell University, 2002.

[21] Kiri Lou Wagstaff, Dominic Mazzoni, and Stephan Sain. HARVIST: A system for agricultural and weather studies using advanced statistical models. In *Proceedings of the Earth-Sun Systems Technology Conference*, 2005.

[22] Wei Wang, Jiong Yang, and Richard Muntz. Sting: A statistical information grid approach to spatial data mining. In *Proceedings of the 23rd VLBD conference, Athens, Greece, 1997*, pages 186–195. Morgan Kaufmann Publishers Inc., 1997.