From Spatial Data Mining in Precision Agriculture to Environmental Data Mining

Georg Ruß

Abstract. In the first part of this article, the main results from applying data mining methods and algorithms to spatial precision agriculture data sets will be outlined. In particular, the task of yield prediction will be handled as a spatial regression problem. To account for the spatial nature of the data sets, a few modeling pitfalls resulting from spatial autocorrelation will be tackled. Based on a cross-validation approach, the yield prediction setting will be used to determine spatial variable importance. Another task called management zone delineation will be briefly outlined. A novel hierarchical spatially constrained clustering algorithm will be presented which aims to provide a tradeoff between spatial contiguity of the resulting clusters and cluster similarity. These two tasks are a summary of [26]. In the second part of this article, the emerging field of *environmental data mining* will be briefly laid out.

1 Introduction

While the (spatial) data sets around us grow rapidly, the tools and algorithms to match those data sets are struggling to keep up. While geographical information systems and location-based services are rapidly expanding, the agricultural sector is currently experiencing an influx of information technology, mostly based on the global positioning system and technological advances in sensors and data aggregation. However, even *precision agriculture* is still in its infancy and requires novel data mining tools and algorithms adapted for the special spatial data sets.

Agricultural companies nowadays harvests not only crops but also growing amounts of data. These data are site specific – which is essentially why the combination of GPS, agriculture and data has been termed *site-specific crop management*. A large amount of information about the soil and crop properties enabling a higher operational efficiency is often contained in these spatial data sets – appropriate

Georg Ruß
TecData AG, 9240 Uzwil, Switzerland
e-mail: georg.russ@buhlergroup.com

C. Moewes et al. (Eds.): Computational Intelligence in Intelligent Data Analysis, SCI 445, pp. 263–273.
springerlink.com

techniques should therefore be applied to find this information. This is a rather common problem for which the term *data mining* has been coined. Data mining techniques aim at finding those patterns in the data that are both valuable and interesting for crop management. This article primarily summarizes the author's two main lines of research. Furthermore, it extends the work on *data mining in precision agriculture* towards a broader scope on *environmental data mining*. The first two parts shortly recapitulate existing work based mainly on [27] and [25].

2 Data Description

The data available in this work were collected during the growing season of 2007 on three sites south of Köthen, Germany. The data for the sites, called *F440*, *F611* and *F631*, respectively, were interpolated using kriging [30] to a grid with 10 by 10 meters grid cell sizes. Each grid cell represents a record with all available information. The fields grew winter wheat. Nitrogen fertilizer (N) was applied three times during the growing season. Overall, for each field there are six input attributes, accompanied by the respective current year's yield (2007) as the target attribute. In total, there are 6446 (F440), 4970 (F611) and 7875 records (F631).

Yield is measured in metric tons per hectare ($\frac{t}{ha}$), along the harvesting lanes (spaced 8 m apart), roughly every ten meters. Apparent electrical soil conductivity (EC25) as a measure for a number of soil properties is acquired. Satellite or aerial image processing provides a measure of vegetation called the red edge inflection point (REIP) value, at two points into the growing season (REIP32, REIP49), according to the growing stage defined in [17]. The REIP value may also be used directly for guiding fertilizations [10]. A simplified assumption is that a higher REIP value means more vegetation. Three nitrogen fertilizer dressings are applied (N1, N2, N3, in $\frac{kg}{ha}$). In the available data, due to the fields being experimental agriculture sites, the nitrogen dressings were not temporally autocorrelated. However, this phenomenon may be considered in production sites. EC, REIP and N are measured in 10-m-intervals along the lanes which are spaced 24 meters apart.

3 Spatial Cross-Validation and Regression

According to [9], *spatial autocorrelation* is the correlation among values of a single variable strictly attributable to the proximity of those values in geographic space, introducing a deviation from the *independent observations* assumption of classical statistics. Given a spatial data set, spatial autocorrelation can be determined using Moran's I ([18]) or semivariograms. For the data sets used in this article, each of the attributes exhibits spatial autocorrelation. In practice, it is usually also known from the data origin whether spatial autocorrelation exists. For further information it is referred to, e.g., [3].

In previous articles using the above data, such as [28, 24], the main focus was on finding a suitable regression model to predict the current year's yield sufficiently well. However, the used regression models, such as neural networks [28, 29] or support vector regression [24], among others, generally assume statistical independence of the data records. However, with the given geo-tagged data records at hand, this is clearly not the case, due to (natural) spatial autocorrelation. Therefore, the spatial relationships between data records have to be taken into account.

Due to the shortcomings in classical regression and cross-validation learning approaches when using them on spatial data, this section will present a novel regression model for data sets which exhibit spatial autocorrelation. In non-spatial regression models, data records which appear in the training set are not supposed to appear in the test set during a cross-validation learning setup. Classical sampling methods do not take spatial neighborhoods of data records into account. Therefore, the above assumption may be rendered invalid when using non-spatial models on spatial data. This inevitably leads to overfitting and underestimates the true prediction error of the regression model (compare [1, 2] for similar observations in a classification context). Therefore, the main issue is to avoid having neighboring or the same samples in training and testing data subsets during a cross-validation learning approach. The basic idea therefore is to apply changes to the resampling method and keep the regression modeling techniques as-is. The resulting procedure can be seen as spatial cross-validation technique.

Traditionally, cross-validation for regression randomly subdivides a given data set into two or three parts: a training set, (optionally a validation set) and a test set. A 10- to 20-fold cross-validation is usually considered appropriate to remove bias [14]. The regression model is trained on the training set until the prediction error on the validation set starts to rise. Once this happens, the training process is stopped and the error on the test set is reported for this fold. This procedure is repeated r times to remove a possible sampling bias. In our case, r has been empirically determined as 100. Instead of sampling randomly from the data set to generate the training and test sets, a clustering step is inserted. A simple k-Means clustering on the data records' x/y-coordinates yields a spatial tessellation of the site under study. The sub-areas of the site are roughly the same size on typical sites. Once the tessellation exists, the cross-validation samples randomly from the sub-areas rather than from the whole data set. The regression is then performed on the data records within the sampled sub-areas. Since k-Means is sensitive to initialization, the clustering is repeated r times.

The spatial clustering procedure may thus be considered as a broader definition of the standard cross-validation setup. This can be seen as follows: when refining the clustering further, the spatial zones on the field become smaller. The border case is reached when the field is subdivided into as many clusters as there are data records, i.e. each data record describes its own cluster. In this special case, the advantages of spatial clustering are lost since no spatial neighborhoods are taken into account in this approach. Therefore, the number of clusters should be seen as a tradeoff between predictive precision and statistical validity of the model. The parameter k for the size of the tessellation has to be determined heuristically.

In previous work ([24, 28]), numerous regression modeling techniques have been compared on similar data sets to determine which of those modeling techniques works best. Support vector regression has been determined as the best modeling technique. It has furthermore recently been shown to work rather successfully in spatial classification tasks, albeit without spatial cross-validation, as in [21]. Hence, in this work support vector regression will serve as a benchmark technique against which further models will have to compete. The techniques and the respective R packages used here are *support vector regression* (e1071), *regression trees* (rpart), *random forests* (randomForest), *bagging* with trees (ipred). The models' performance is measured via the root mean squared error (RMSE) between actual value and predicted value.

The results in Table 1 confirm that the spatial autocorrelation inherent in the data set leads classical, non-spatial regression modeling setups to a substantial underestimation of the prediction error. This outcome is consistent throughout the results, regardless of the used technique and regardless of the parameters. Furthermore, it can be seen that Random Forests seem to yield better performance in terms of lower prediction error, regardless of the setup used. Moreover, the spatial setup can be easily set to emulate the non-spatial setup: set *k* to be the number of data records in the data set. Therefore the larger the parameter *k* is set, the smaller the difference between the spatial and the non-spatial setup should be. This assumption also holds true for almost all of the obtained results.

Table 1 Results of running different setups on the data sets F440 and F611; comparison of spatial vs. non-spatial treatment of data sets; root mean squared error in t/ha is shown, averaged over clusters/folds; *k* is either the number of clusters in the spatial setup or the number of folds in the non-spatial setup. The average yield is around 8-10 t/ha.

		F440		F611	
	k	spatial	non-spatial	spatial	non-spatial
Support Vector Regression	10	1.06	0.54	0.73	0.40
	20	1.00	0.54	0.71	0.40
	50	0.91	0.53	0.67	0.38
Regression Tree	10	1.09	0.56	0.69	0.40
	20	0.99	0.56	0.68	0.42
	50	0.91	0.55	0.66	0.40
Random Forest	10	0.99	0.50	0.65	0.41
	20	0.92	0.50	0.64	0.41
	50	0.85	0.48	0.63	0.39
Bagging	10	1.09	0.59	0.66	0.42
	20	1.01	0.59	0.66	0.42
	50	0.94	0.58	0.65	0.41

4 Management Zone Delineation with HACC-Spatial

The second task in precision agriculture which is summarized in this article is *management zone delineation*. In brief, it aims to generate a subdivision of the site under study into homogeneous zones which are, to a certain degree, contiguous in space. Further details can be acquired from [25]. From a data mining point of view, this task is essentially a clustering challenge where a specific constraint (spatial contiguity) must be taken into account accordingly.

The underlying idea of HACC-SPATIAL is to adapt hierarchical agglomerative clustering (HAC) for spatial data sets. In HAC, the clustering of arbitrary objects starts with each object in a single cluster. Consecutively, two clusters are merged into one new cluster: the decision which clusters to merge is often done based on cluster similarity or distance, using an appropriate distance measure. Furthermore, constraints have been introduced into HAC, leading to hierarchical agglomerative constrained clustering (HACC): the decision which clusters to merge is not only done based on the similarity, but also according to constraints which can have two types. The first is a *must-link* constraint, which means that two clusters belong into one cluster. The second is of the opposite *cannot-link* constraint, which determines that two clusters must not be merged. The idea of HACC-SPATIAL is now to use a *cannot-link* constraint to enforce spatial contiguity of the resulting clusters. Furthermore, in the beginning of the clustering the algorithm strictly enforces spatial contiguity due to the constraint, while the constraint may be relaxed after a certain threshold between adjacent and non-adjacent clusters is reached.

The cluster distance is determined in feature space, while the constraint ensures spatial contiguity in geographic space. For lower-dimensional feature spaces, Euclidean distance is used, while for higher dimensions, due to the curse of dimensionality, the Cosine distance may be used. The details of HACC-SPATIAL can be obtained from [25].

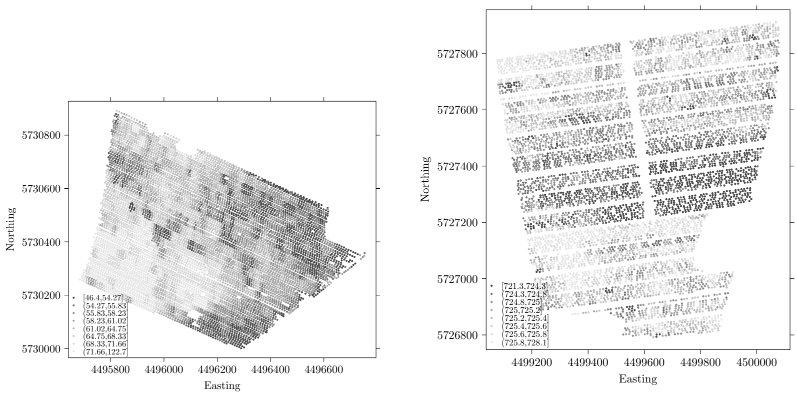
To exploit spatial autocorrelation (which is typically present in precision agriculture data sets) and reduce the computational burden, HACC-SPATIAL does not start directly with each data object in a single cluster. Instead, it can safely be assumed that a few spatially adjacent data objects are similar enough to be put into an initial cluster. To achieve this initial clustering, a round of k-Means clustering is applied initially to the spatial coordinates of the data objects. Depending on the heterogeneity of the site, the number of initial clusters in the tessellation which is generated by k-Means should be set in a range of around 100 to N, where N is the number of data objects to be clustered.

4.1 Results on Different Precision Agriculture Data Sets

The two variables from two actual sites which HACC-SPATIAL will be applied to are depicted in Figure 1. While the REIP value alone has no practical use in this zone delineation task, it certainly is of major importance in other YIELD-related tasks. The experiments are designed such that the algorithm's results can be easily visually compared with the actual variable under study. Practically, zone delineation is often done using the EC variable.

4.1.1 F631

A result demonstrating the different settings of the contiguity parameter is presented in Figure 2, where the variable EC25 of the F631 field is used for management zone



(a) F631: EC25

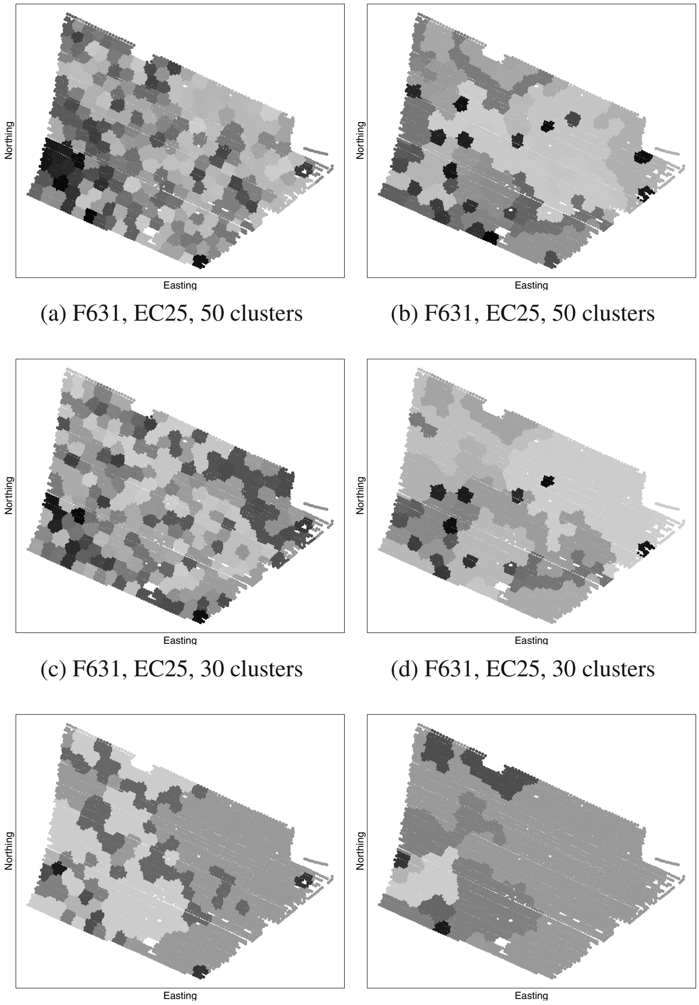
(b) F440: REIP32

Fig. 1 F631: EC25, F440: REIP32. The spatial distribution of the EC25 and the REIP32 variables clearly exhibits spatial autocorrelation. In the bottom figure, strips of data are missing, while the underlying distribution can still be identified by a human. An appropriate clustering algorithm has to be developed to generate management zones from this variable.

delineation. The field is initially tessellated into 250 clusters and the clustering is run with low and high contiguity settings to compare the results. Clustering with low spatial contiguity yields mostly non-contiguous clusters (as expected) until spatially contiguous clusters start emerging towards the very end of the clustering (Figure 2e). On the other hand, clustering with high spatial contiguity starts showing emergent clusters after around 200 merging steps (Figure 2b) and subsequent clusters clearly correspond to the actual variable value (Figure 1a). The clusters are not limited to convex shapes and account for the irregular shape of the field (missing data, irregular borders, "holes"). If the clustering in Figure 2f is deemed too coarse, the hierarchically structured clustering easily allows for subdividing single clusters by traversing the dendrogram.

4.1.2 F440

In the preceding sections, the main purpose was to show the effect of enforcing or neglecting spatial contiguity throughout the clustering. This was achieved by setting the contiguity ratio threshold accordingly. A direct comparison of the results of HACC-SPATIAL when applied to the same input data is provided here. Figure 3 shows the REIP32 variable on the F440 field, clustered by HACC-SPATIAL, showing the stage at which 15 clusters are left. While Figure 3a shows almost no visible spatial contiguity, this changes gradually towards Figure 3d where the clusters are clearly spatially contiguous.



(a) F631, EC25, 50 clusters

(b) F631, EC25, 50 clusters

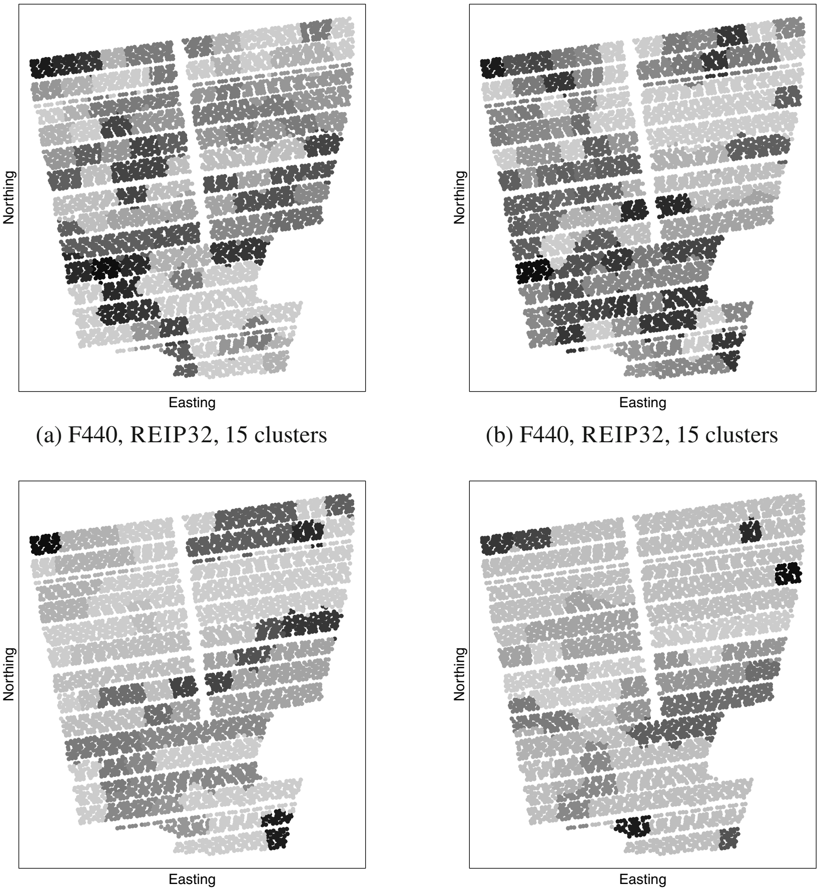
(c) F631, EC25, 30 clusters

(d) F631, EC25, 30 clusters

(e) F631, EC25, 8 clusters

(f) F631, EC25, 8 clusters

Fig. 2 HACC-SPATIAL on F631, using the variable EC25 (cp. Figure 1a on Page 268), starting with 250 clusters. Clustering with low (left figures) and high (right figures) spatial contiguity shows considerable differences in the spatial structure of the resulting clusters. At low spatial contiguity the algorithm starts producing visible spatially contiguous clusters only towards the end of the clustering (e), while spatially contiguous clusters start emerging much earlier when clustering with high spatial contiguity (b).



(a) F440, REIP32, 15 clusters

(b) F440, REIP32, 15 clusters

(c) F440, REIP32, 15 clusters

(d) F440, REIP32, 15 clusters

Fig. 3 HACC-SPATIAL on F440, 120 initial clusters, using the REIP32 variable and demonstrating the effect of different spatial contiguity settings. While (a) shows spatially rather scattered clusters, the change in the designed contiguity ratio threshold varies the spatial contiguity of the clusters until spatial contiguity is strictly enforced in Figure (d).

4.2 Clustering Summary

Based on both the practical and the theoretical need for an efficient and understandable algorithm for management zone delineation in precision agriculture, a novel algorithm HACC-SPATIAL has been devised. It is able to exploit spatial autocorrelation in the precision agriculture data and successfully extends hierarchical agglomerative constrained clustering towards spatial data sets. An algorithmic description and results on one-dimensional spatial data sets have been presented. The main parameter *contiguity threshold* has been experimentally validated and shown to be successful in three practical data sets.

5 Environmental Data Mining

The original definition of data mining within the process of knowledge discovery in databases by [6] described it as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. The data collected in environmental sciences such as ecology, geology, remote sensing and agriculture are by their very nature spatial and/or temporal which are important additional properties when it comes to data mining. Hence, it is proposed to extend the above definition towards environmental data mining as follows:

> *Environmental Data Mining* is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in *spatial* and *temporal* data *from environmental sciences.*

Many of the developed techniques in data mining, though not particularly adapted for the specifics of environmental data sets, are rather flexible. They can often be tailored to fit environmental data, such as the regression and clustering problems presented in this article. Introductions to this increasingly active field can be found in [8], [13], [12] and [7].

Given the classicals tasks of classification and regression, especially the work in ecology has started around the year 2000, ranging from neural networks [15] over bayesian statistics and belief networks [16] to bagging and random forests [22],[1], [2]. The related task of clustering in environmental data sets has a history that dates back to 1990 [5], with numerous further applications of fuzzy clustering, such as in agriculture [19] and remote sensing [20]. A third frequent task in classical data mining is association analysis, which has also been introduced into ecology [31], remote sensing [11] and agriculture [4], among others.

With those prerequisites, the term *environmental data mining* encompasses most of the existing work under a common umbrella term, while distinctively combining the fields of environmental sciences and data mining.

Acknowledgements. The implementation is carried out in R [23]. The R scripts are available on request from the author of this article. The data in this work have been acquired on the experimental farm Görzig in the federal state Sachsen-Anhalt, in Germany. The data were obtained from Martin Schneider and Peter Wagner from Martin-Luther-Universität Halle-Wittenberg, Germany, Lehrstuhl für landwirtschaftliche Betriebslehre.

References

[1] Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Science 5(6), 853–862 (2005)

[2] Brenning, A., Lausen, B.: Estimating error rates in the classification of paired organs. Statistics in Medicine 27(22), 4515–4531 (2008)

[3] Cressie, N.A.C.: Statistics for Spatial Data. Wiley, New York (1993)

[4] El-Beltagy, S.R., Rafea, A., Mabrouk, S.: Agrimine: A tool for mining agricultural problems and their solutions. In: Proc. of Int. Computer Engineering Conference (ICENCO), ICENCO, pp. 81–85 (2010)

[5] Equihua, M.: Fuzzy clustering of ecological data. Journal of Ecology 78(2), 519–534 (1990)

[6] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Commun. ACM 39, 27–34 (1996)

[7] Fielding, A.: An introduction to machine learning methods. In: Fielding, A. (ed.) Machine Learning Methods for Ecological Applications, pp. 1–35. Kluwer Academic Publishers, Dordrecht (1999)

[8] Gibert, K., Spate, J., Snchez-Marr, M., Athanasiadis, I.N., Comas, J.: Data mining for environmental systems. In: Jakeman, A.J., Voinov, A.A., Rizzoli, A.E., Chen, S. (eds.) Environmental Modelling, Software and Decision Support, Developments in Integrated Environmental Assessment, vol. 3, pp. 205–228. Elsevier (2008)

[9] Griffith, D.A.: Spatial Autocorrelation and Spatial Filtering. Advances in Spatial Science. Springer, New York (2003)

[10] Heege, H., Reusch, S., Thiessen, E.: Prospects and results for optical systems for site-specific on-the-go control of nitrogen-top-dressing in germany. Precision Agriculture 9(3), 115–131 (2008)

[11] Iisaka, J., Sakurai-Amano, T.: Spatial association analysis for radar image interpretation. In: International Geoscience and Remote Sensing Symposium, IGARSS, pp. 1200–1203 (1993)

[12] Kanevski, M. (ed.): Advanced Mapping of Environmental Data: Geostatistics, Machine Learning and Bayesian Maximum Entropy. ISTE, London (2010)

[13] Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V., Canu, S.: Environmental data mining and modeling based on machine learning algorithms and geostatistics. Environmental Modelling and Software 19(9), 845–855 (2004)

[14] Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of International Joint Conference on Artificial Intelligence (1995)

[15] Lek, S., Guegan, J.: Application of artificial neural networks in ecological modelling. Ecological Modelling 120(2-3) (1999)

[16] Marcot, B.G., Holthausen, R.S., Raphael, M.G., Rowland, M.M., Wisdom, M.J.: Using bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. Forest Ecology and Management 153, 29–42 (2001)

[17] Meier, U.: Entwicklungsstadien mono- und dikotyler Pflanzen. In: Biologische Bundesanstalt für Land- und Forstwirtschaft, Braunschweig, Germany (2001)

[18] Moran, P.A.P.: Notes on continuous stochastic phenomena. Biometrika 37, 17–33 (1950)

[19] Papajorgji, P.J., Pardalos, P.M. (eds.): Advances in Modeling Agricultural Systems. Springer Optimization and Its Applications, vol 25. Springer (2009)

[20] Petcu, D., Zaharie, D., Panica, S., Hussein, A.S., Sayed, A., El-Shishiny, H.: Fuzzy clustering of large satellite images using high performance computing. In: Proc. of SPIE. SPIE, vol. 8183 (2011)

[21] Pozdnoukhov, A., Foresti, L., Kanevski, M.: Data-driven topo-climatic mapping with machine learning methods. Natural Hazards 50(3), 497–518 (2009)

[22] Prasad, A.M., Iverson, L.R., Liaw, A.: Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 9, 181–199 (2006)

[23] R Development Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009), `http://www.R-project.org`, ISBN 3-900051-07-0

[24] Ruß, G.: Data Mining of Agricultural Yield Data: A Comparison of Regression Models. In: Perner, P. (ed.) ICDM 2009. LNCS (LNAI), vol. 5633, pp. 24–37. Springer, Heidelberg (2009)

[25] Ruß, G.: Hacc-spatial: Hierarchical agglomerative spatially constrained clustering. In: Bichindaritz, I., Perner, P., Ruß, G. (eds.) 11th ICDM Conference, New York, USA, Workshop Proceedings. IBaI Publishing, Leipzig (2011)

[26] Ruß, G.: Spatial Data Mining in Precision Agriculture. PhD thesis, Otto-von-Guericke-Universität Magdeburg (2012)

[27] Ruß, G., Brenning, A.: Data Mining in Precision Agriculture: Management of Spatial Information. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010. LNCS, vol. 6178, pp. 350–359. Springer, Heidelberg (2010)

[28] Ruß, G., Kruse, R., Schneider, M., Wagner, P.: Estimation of neural network parameters for wheat yield prediction. In: Artificial Intelligence in Theory and Practice II. IFIP, vol. 276, pp. 109–118. Springer, Boston (2008)

[29] Ruß, G., Kruse, R., Schneider, M., Wagner, P.: Optimizing wheat yield prediction using different topologies of neural networks. In: Verdegay, J.L., Ojeda-Aciego, M., Magdalena, L. (eds.) Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2008), pp. 576–582. University of Málaga (2008)

[30] Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics. Springer (1999)

[31] Su, F., Zhou, C., Lyne, V., Du, Y., Shi, W.: A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. Ecological Modelling 174, 421–431 (2004)