Regression Models for Spatial Data: An Example from Precision Agriculture

Georg Ruß and Rudolf Kruse

Otto-von-Guericke-Universität Magdeburg
georg.russ@ieee.org

Abstract. The term *precision agriculture* refers to the application of state-of-the-art GPS technology in connection with small-scale, sensor-based treatment of the crop. This data-driven approach to agriculture poses a number of data mining problems. One of those is also an obviously important task in agriculture: yield prediction. Given a precise, geographically annotated data set for a certain field, can a season's yield be predicted?

Numerous approaches have been proposed to solving this problem. In the past, classical regression models for non-spatial data have been used, like regression trees, neural networks and support vector machines. However, in a cross-validation learning approach, issues with the assumption of statistical independence of the data records appear. Therefore, the geographical location of data records should clearly be considered while employing a regression model. This paper gives a short overview about the available data, points out the issues with the classical learning approaches and presents a novel spatial cross-validation technique to overcome the problems and solve the aforementioned yield prediction task.

Keywords: Precision Agriculture, Data Mining, Regression, Modeling.

1 Introduction

In recent years, information technology (IT) has become more and more part of our everyday lives. With data-driven approaches applied in industry and services, improvements in efficiency can be made in almost any part of nowadays' society. This is especially true for agriculture, due to the modernization and better affordability of state-of-the-art GPS technology. A farmer nowadays harvests not only crops but also growing amounts of data. These data are precise and small-scale – which is essentially why the combination of GPS, agriculture and data has been termed *precision agriculture*.

In those agriculture (field) data, often a large amount of information is contained, yet hidden. This is usually information about the soil and crop properties enabling a higher operational efficiency – appropriate techniques should therefore be applied to find this information. This is a common problem for which the term *data mining* has been coined. Data mining techniques aim at finding those patterns or information in the data that are both valuable and interesting to the farmer.

A specific problem commonly occurring is *yield prediction*. As early into the growing season as possible, a farmer is interested in knowing how much yield he is about to expect. The ability to predict yield used to rely on farmers' long-term knowledge of particular fields, crops and climate conditions. However, this knowledge can be expected

P. Perner (Ed.): ICDM 2010, LNAI 6171, pp. 450–463, 2010.

to be available in the data collected during normal farming operations throughout the season. A multitude of sensor data are nowadays collected, measuring a field's heterogeneity. These data are precise, often highly correlated and carry spatial information which must not be neglected.

Hence, the problem of yield prediction encountered can be treated as a problem of data mining and, specifically, multi-dimensional regression. This article will serve as a reference of how to treat a regression problem on spatial data with a combination of classical regression techniques and a number of novel ideas. This article will furthermore serve as a continuation of [17]: in the previous article, the spatial data were treated with regression models which do not take the spatial relationships into account. The current work aims to check the validity of the *statistical independence* assumption inherent in classical regression models in conjunction with spatial data. Based upon the findings, spatial regression will be carried out using a novel clustering idea during a cross-validation procedure. The results will be compared to those obtained while neglecting the spatial relationships inherent in the data sets.

1.1 Research Target

The main research target of this work is to improve and further substantiate the validity of *yield prediction* approaches using multi-dimensional regression modeling techniques. Previous work, mainly the regression work presented in [17,21], will be used as a baseline for this work. Some of the issues of the previous approach will be clearly pointed out in this article. Nevertheless, this work aims to improve upon existing yield prediction models and, furthermore, incorporates a generic, yet novel spatial clustering idea into the process. Therefore, different types of regression techniques will be incorporated into a novel spatial cross-validation framework. A comparison of using spatial vs. non-spatial data sets shall be presented.

1.2 Article Structure

This article will start with a brief introduction into the area of precision agriculture and a more detailed description of the available data in Section 2. This will be followed by an outline of the key techniques used in this work, embedded into a data mining workflow presented in Section 3. The results obtained during the modeling phase will be presented in Section 4. The article will be completed with a short conclusion in Section 5, which will also point out further lines of research.

2 Data Description

With the recent advances in technology, ever larger amounts of data are nowadays collected in agriculture during standard farming operations. This section first gives a short categorization of the data into four classes. Afterwards, the actual available data are presented. The differences between spatial and non-spatial data are pointed out.

2.1 Data Categorization

A commonality among data collected in agriculture is that every data record has a spatial location on the field, usually determined via (differential) GPS with a high degree of precision. These data can roughly be divided into four classes as follows:

Yield Mapping has been a standard approach for many years. Based on maps of previous years' yields, recommendations of farming operations for the current season are determined.

Topography is often considered a valuable feature for data mining in agriculture. The spatial location of data points (longitude, latitude) is a standard variable to be used in spatial modeling. Furthermore, variables like elevation, slope and derivatives of those values can be obtained easily.

Soil Sampling is a highly invasive means of acquiring data about a field. Furthermore, it is labour-intensive and therefore rather expensive. Obtaining a high resolution of soil sampling data therefore requires lots of effort. From soil sampling, variables like organic matter, available minerals, water content etc. can be derived.

Remote Sensing recently has become a rather cheap and high-resolution data source for data-driven agricultural operations. It usually consists of aerial or satellite imaging using multiple spectral bands at different times into the vegetation period. From those images, vegetation indices are derived and used for assessing the crop status.

2.2 Available Data

The data available in this work were collected during the growing season of 2007 on two fields north of Köthen, Germany. The data for the two fields, called *F440* and *F611*, respectively, were interpolated using kriging [23] 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, where nitrogen fertilizer was distributed over three application times during the growing season.

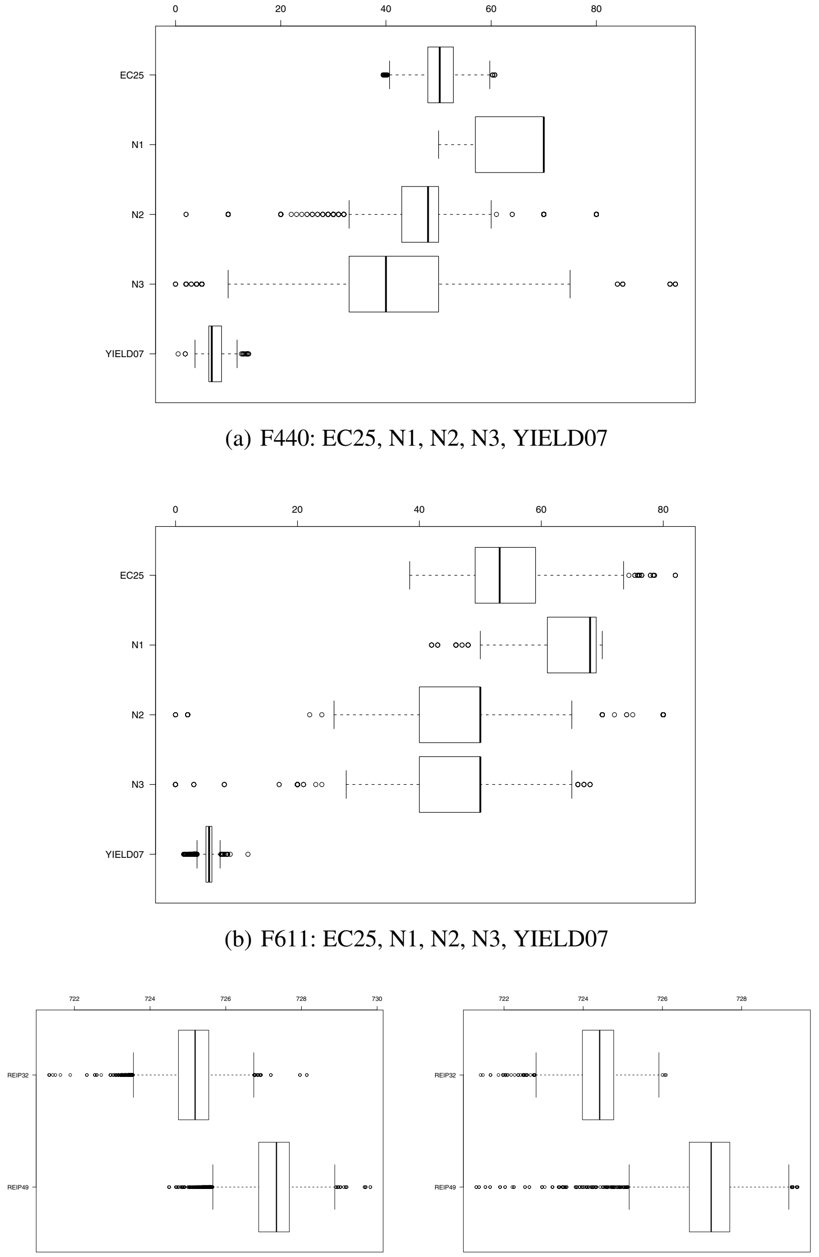
Overall, for each field there are six input attributes – accompanied by the respective current year's yield (2007) as the target attribute. Those attributes will be described in the following. In total, for the F440 field there are 6446 records, for F611 there are 4970 records, thereof none with missing values and none with outliers. A short statistical summary of the fields and variables can be found in Figure 1. In the following sections, further details about the individual attributes is provided.

2.3 YIELD07

Here, yield is measured in metric tons per hectare ($\frac{t}{ha}$). For the yield ranges for the respective years and sites, see Figures 1(a) and 1(b).

2.4 Apparent Electric Conductivity – EC25

A non-invasive method to discover and map a field's heterogeneity is to measure the soil's apparent electrical conductivity. It is assumed that the EC25 readings are closely related to soil properties which would otherwise have to be sampled in a time-consuming



(a) F440: EC25, N1, N2, N3, YIELD07

(b) F611: EC25, N1, N2, N3, YIELD07

(c) F440: REIP32, REIP49

(d) F611: REIP32, REIP49

Fig. 1. Statistical Summary for the two available data sets (F440, F611)

and expensive manner. Commercial sensors such as the EM-38[1] are designed for agricultural use and can measure small-scale conductivity to a depth of about 1.5 metres. There is no possibility of interpreting these sensor data directly in terms of its meaningfulness as yield-influencing factor. But in connection with other site-specific data, as explained in the rest of this section, there could be coherences. For a more detailed analysis of this particular sensor, see, e.g. [5]. For the range of EC25 values encountered in the available data, see Figures 1(a) and 1(b).

2.5 Vegetation – REIP32, REIP49

The *red edge inflection point* (REIP) is a second derivative value calculated along the red edge region of the spectrum, which is situated from 680 to 750nm. Dedicated REIP sensors are used in-season to measure the plants' reflection in this spectral band. Since the plants' chlorophyll content is assumed to highly correlate with the nitrogen availability (see, e.g. [13]), the REIP value allows for deducing the plants' state of nutrition and thus, the previous crop growth. For further information on certain types of sensors and a more detailed introduction, see [9] or [24]. Plants that have less chlorophyll will show a lower REIP value as the red edge moves toward the blue part of the spectrum. On the other hand, plants with more chlorophyll will have higher REIP values as the red edge moves toward the higher wavelengths. Obviously, later into the growing season the plants are expected to have a higher chlorophyll content, which can easily be assessed by visually comparing the REIP values in Figures 1(c) and 1(d). The numbers in the REIP32 and REIP49 names refer to the growing stage of winter wheat, as defined in [11].

2.6 Nitrogen Fertilizer – N1, N2, N3

The amount of fertilizer applied to each subfield can be measured easily. Since it is a variable that can and should be influenced by the farmer, it does not appear in the preceding categorization. Fertilizer is applied at three points in time into the vegetation period, which is the standard strategy for most of Northwestern Europe [15]. The ranges in the data sets can be obtained from Figures 1(a) and 1(b).

2.7 Spatial vs. Non-spatial Data Treatment

According to [7], *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 ([14]) or semivariograms. Spatial autocorrelation appears in such diverse areas as econometrics [1], geostatistics [6] and social sciences [10], among others. In practice, it is usually also known from the data configuration whether spatial autocorrelation is existent. For further information it is referred to, e.g., [6].

In previous articles using the above data, such as [17,21], the main focus was on finding a suitable regression model to predict the current year's yield sufficiently well.

[1] Trademark of Geonics Ltd, Ontario, Canada.

However, it should be noted that the used regression models, such as neural networks [18,19] or support vector regression [17], among others, usually 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. The following section further elaborates upon this topic in detail.

3 Regression Techniques on Spatial Data

Based on the findings at the end of the preceding section, this section will present a novel regression model for data sets which exhibit spatial autocorrelation. In classical regression models, data records which appear in the training set must not appear in the test set during a cross-validation learning setup. Due to classical sampling methods which do not take spatial neighborhoods of data records into account, this assumption may be rendered invalid when using non-spatial models on spatial data. This leads to overfitting (overlearning) and underestimates the true prediction error of the regression model. Therefore, the core issue is to avoid having neighboring or the same samples in training and testing data subsets during a cross-validation approach.

As should be expected, the data sets F440 and F611 exhibit spatial autocorrelation. Therefore, classical regression models must either be swapped against different ones which take spatial relationships into account or may be adapted to accommodate spatial data. In order to keep standard regression modeling techniques such as neural networks, support vector regression, bagging, regression trees or random forests as-is, a meta-approach will be presented in the following. In a nutshell, it replaces the standard sampling approach of the cross-validation process with an approach that is aware of spatial relationships.

3.1 From Classical to Spatial Cross-Validation

Traditionally, k-fold cross-validation for regression randomly subdivides a given data set into two (without validation set) or three parts: a training set, a validation set and a test set. A ratio of 6:2:2 for these sets is usually assumed appropriate. 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 k times, with the root mean squared error (RMSE) often used as a performance measure.

The issue with spatial data is that, due to spatial autocorrelation, almost identical data records may end up in training and test set, such that the model overfits the data and underestimates the error. Therefore, one possible solution might be to ensure that only a very small number (if any) of neighboring and therefore similar samples end up in training and test subsets. This may be achieved by adapting the sampling procedure for spatial data. Once this issue has been accommodated, the cross-validation procedure may continue as-is. A rather straightforward approach using the geo-tagged data is described in the following.

3.2 Employing Spatial Clustering for Data Sampling

Given the data sets F440 and F611, a spatial clustering procedure can be employed to subdivide the fields into spatially disjunct clusters or zones. The clustering algorithm can easily be run on the data map, using the data records' longitude and latitude. Depending on the clustering algorithm parameters, this results in a tesselation map which does not consider any of the attributes, but only the spatial neighborhood between data records. A depiction of this clustering process can be found in Figures 2(a) and 2(b). Standard k-means clustering was used with a setting of $k = 20$ clusters per field for demonstration purposes. In analogy to the non-spatial regression treatment of these data records, now a spatially-aware cross-validation regression problem can be handled using the k zones of the clustering algorithm as an input for k-fold cross-validation. Standard models, as described below, can be used straightforwardly, without requiring changes to the models themselves. The experimental setup and the results are presented in the following section.

It should be noted that this spatial clustering procedure is 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 precision and statistical validity of the model.

3.3 Regression Techniques

In previous work ([17,21]), numerous regression modeling techniques have been compared on similar data sets to determine which of those modeling techniques works best. Although those models were run in a non-spatial regression setup, it is assumed that the relative differences between these models will also hold in a spatial cross-validation regression setup. In the aforementioned previous work, support vector regression has been determined as the best modeling technique when comparing the models' root mean squared prediction error. Hence, in this work support vector regression will serve as a benchmark technique against which further models will have to compete. Experiments are conducted in R [16], a link to the respective scripts is provided in Section 5.

Support Vector Regression. Support Vector Machines (SVMs) are a supervised learning method discovered by [2]. However, the task here is regression, so the focus is on support vector regression (SVR) in the following. A more in-depth discussion can be found in [8]. Given the training set, the goal of SVR is to approximate a linear function $f(x) = \langle w, x \rangle + b$ with $w \in \mathbb{R}^N$ and $b \in \mathbb{R}$. This function minimizes an empirical risk function defined as

$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} L_\varepsilon(\hat{y} - f(x)), \tag{1}$$

where $L_\varepsilon(\hat{y} - f(x)) = \max((|\xi| - \varepsilon), 0)$. $|\xi|$ is the so-called slack variable, which has mainly been introduced to deal with otherwise infeasible constraints of the optimization problem, as has been mentioned in [22]. By using this variable, errors are basically ignored as long as they are smaller than a properly selected ε. The function here is called ε-insensitive loss function. Other kinds of functions can be used, some of which are presented in chapter 5 of [8].

To estimate $f(x)$, a quadratic problem must be solved, of which the dual form, according to [12] is as follows:

$$max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j)-\varepsilon\sum_{i=j}^{N}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{N}y_i(\alpha_i-\alpha_i^*) \quad (2)$$

with the constraint that $\sum_{j=1}^{N}(\alpha_i-\alpha_i^*)=0, \alpha_i, \alpha_i^* \in [0,C]$. The regularization parameter $C > 0$ determines the tradeoff between the flatness of $f(x)$ and the allowed number of points with deviations larger than ε. As mentioned in [8], the value of ε is inversely proportional to the number of support vectors. An adequate setting of C and ε is necessary for a suitable solution to the regression problem.

Furthermore, $K(x_i,x_j)$ is known as a kernel function which allows to project the original data into a higher-dimensional feature space where it is much more likely to be linearly separable. Some of the most popular kernels are radial basis functions (equation 3) and a polynomial kernel (equation 4):

$$K(x,x_i) = e^{-\frac{||x-x_i||^2}{2\sigma^2}} \quad (3)$$

$$K(x,x_i) = (\langle x,x_i\rangle+1)^{\rho} \quad (4)$$

The parameters σ and ρ have to be determined appropriately for the SVM to generalize well. This is usually done experimentally. Once the solution for the above optimization problem in equation 2 is obtained, the support vectors can be used to construct the regression function:

$$f(x) = \sum_{i=1}^{N}(\alpha_i-\alpha_i^*)K(x,x_i)+b \quad (5)$$

In the current experiments, the *svm* implementation from the *e1071* R package has been used.

Random Forests and Bagging. In previous work ([17]), one of the presented regression techniques were regression trees. They were shown to be rather successful, albeit in a non-spatial regression setup. Therefore, this article considers an extension of regression trees: random forests. According to [4], random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In the version used here, the random forest is used as a regression technique. Basically, a random forest is an ensemble method that consists of many regression trees and outputs a combined result of those trees as a prediction for the target variable. Usually, the generalization error for forests converges to a limit as the number of trees in the forest becomes large.

Let the number of training cases be N and the number of variables in the regression task be M. Then, each tree is constructed using the following steps:

1. A subset with size m of input variables is generated. This subset is used to determine the decision at a node of the tree; $m \ll M$.
2. Take a bootstrap sample for this tree: choose N times with replacement from all N available training cases. Use the remaining cases to estimate the tree's regression error.
3. Randomly choose m variables from which to derive the regression decision at that node; repeat this for each node of the tree. Calculate the best tree split based on these m variables from the training set.

It should be noted that each tree is fully grown and not pruned. This is a difference from normal regression tree construction. Random forests mainly implement the key ideas from bagging, which is therefore explained in the following.

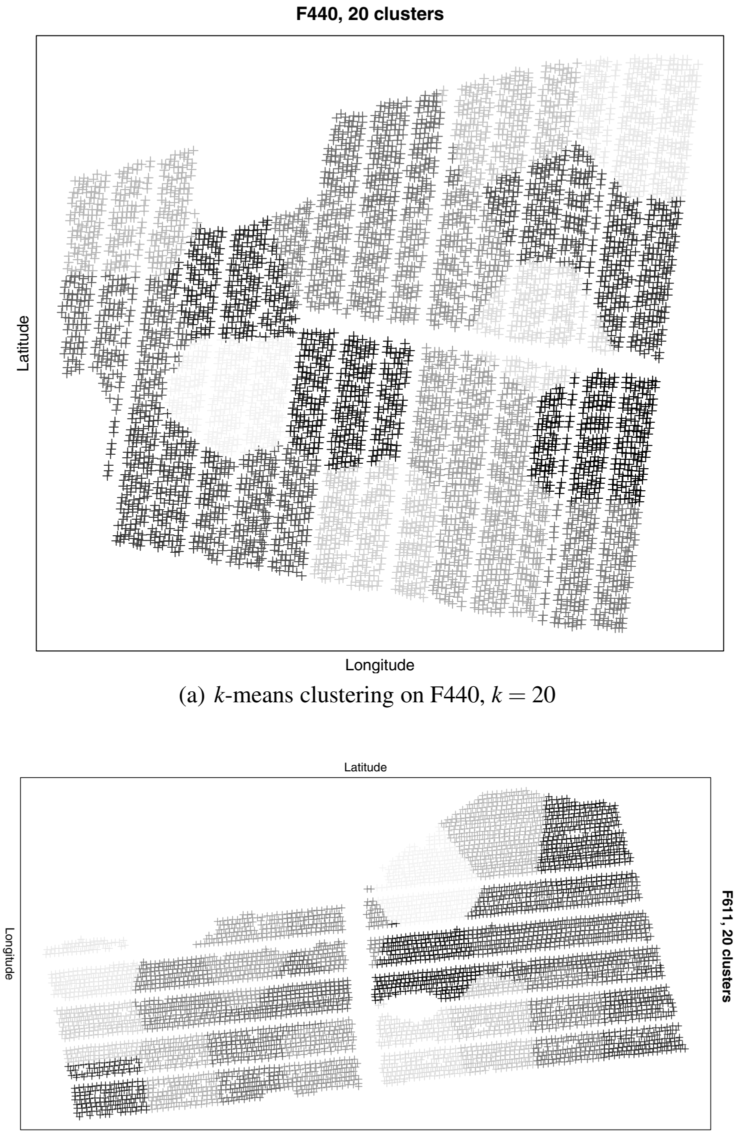
Bootstrap aggregating (or bagging) has first been described in [3]. It is generally described as a method for generating multiple versions of a predictor and using these for obtaining an aggregate predictor. In the regression case, the prediction outcomes are averaged. Multiple versions of the predictor are constructed by taking bootstrap samples of the learning set and using these as new learning sets. Bagging is generally considered useful in regression setups where small changes in the training data set can cause large perturbations in the predicted target variables. Since random forests are a special case of bagging where regression trees are used as the internal predictor, both random forests and bagging should deliver similar results. Both techniques are available in the R packages *randomForest* and *ipred*. Running them on the available data sets should therefore deliver similar results, since the bagging implementation in the R *ipred* package internally uses regression trees for prediction as well. Therefore, the main difference between random forests and bagging in this article is that both techniques are implicitly run and reported with different parameters.

Performance Measurement. The performance of the models will be determined using the root mean squared error (RMSE). For the RMSE, first the difference between an actual target value y_a and the model output value y is computed. This difference is squared and averaged over all training examples before the root of the mean value is taken, see equation 6.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=j}^{n}(y_i - y_{a,i})^2} \tag{6}$$

4 Results

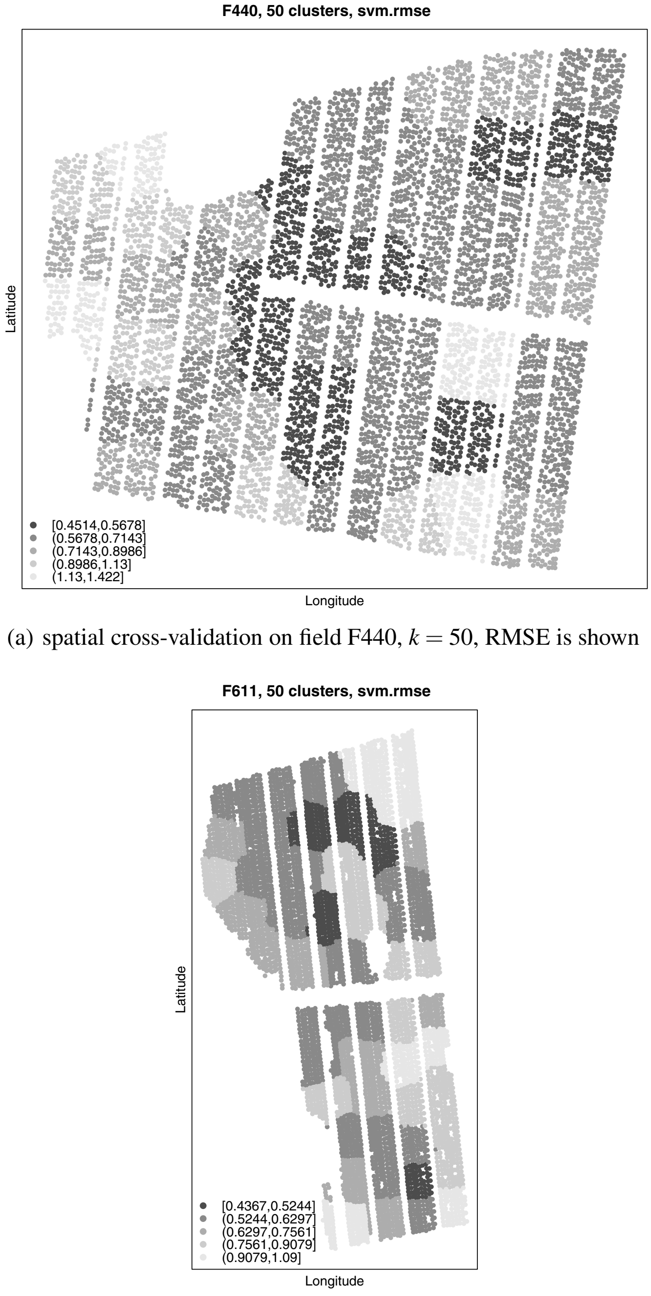
As laid out in the preceding sections, the main research target of this article is to assess whether existing spatial autocorrelation in the data sets may fail to be captured in standard, non-spatial regression modeling setups. The approach consists of a simple comparison between a spatial and a non-spatial setup.



(a) k-means clustering on F440, $k = 20$

(b) k-means clustering on F611, $k = 20$

Fig. 2. k-means clustering on F440 and F611 (the bottom figure has been rotated by 90 degrees)



(a) spatial cross-validation on field F440, $k = 50$, RMSE is shown

(b) spatial cross-validation on field F611, $k = 50$, RMSE is shown

Fig. 3. Results for spatial cross-validation on F440/F611 fields, using 50 clusters and support vector regression

non-spatial setup. The non-spatial setup is similar to the one presented in [17], although different data sets are used. A standard cross-validation procedure is performed, where *k* is the number of folds. Support vector machines, random forests and bagging are trained on the training set. The squared errors on the test set are averaged and the square root is taken. The resulting value is reported in Table 1.

spatial setup. Since the amount of research effort into spatial data sets is rather sparse when compared to this special setup, a simple, yet effective generic approach has been developed. The spatial data set is clustered into *k* clusters using the *k*-means algorithm (see Figure 2). This non-overlapping partitioning of the data set is then used in a spatial cross-validation setup in a straightforward way. This ensures that the number of neighboring data points (which are very similar due to spatial autocorrelation) in training and test sets remains small. The root mean squared error is computed similarly to the non-spatial setup above and may optionally be displayed (see Figure 3).

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 could be shown that for these particular data sets, random forests or bagging yield more precise predictions than support vector regression. However, the standard settings of the respective R toolboxes were used in both the spatial and the non-spatial setup, therefore the difference between these setups will remain similar regardless of parameter changes. Nevertheless, changes to model parameters might slightly change the outcome of the prediction accuracy and the ranking of the models in terms of root mean squared error. The drawback is that parameter tuning via grid search easily extends computation times by orders of magnitude.

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

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 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

		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
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

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.

5 Conclusions and Future Work

This article presented a central data mining task: regression. Based on two data sets from precision agriculture, a continuation and improvement over previous work ([17,21]) could be achieved. The difference between spatial data and non-spatial data was pointed out. The implications of spatial autocorrelation in these data sets were mentioned. From a statistical and machine learning point of view, neighboring data records in a spatially autocorrelated data sets should not end up in training and test sets since this leads to a considerable underestimation of the prediction error, possibly regardless of the used regression model.

It can be concluded that it is indeed important to closely consider spatial relationships inherent in the data sets. As a suggestion, the following steps should be taken: for those data, the spatial autocorrelation should be determined. If spatial autocorrelation exists, standard regression models must be adapted to the spatial case. A straightforward and illustrative approach using simple *k*-means clustering has been described in this article.

5.1 Future Work

Despite having improved and validated upon the yield prediction task, the data sets carry further information. Two rather interesting task are *variable importance* and *management zones*.

The first refers to the question which of the variables is actually contributing most to the yield prediction task. This has practical implications for the farmers and sensor-producing companies. A first non-spatial approach has been presented in [20] as a standard feature selection approach, which should accommodate the spatial relationships in future implementations. The bagging approach presented in this article might be considered.

The second refers to discovering interesting zones on the (heterogeneous) field which should be managed differently from each other. This is a classical data mining question where the *k*-means approach used in this article is likely to be considered.

Further material, including the R scripts for creating the figures in this article and computing the results, can be found at `http://research.georgruss.de/?cat=24`.

References

1. Anselin, L.: Spatial Econometrics, pp. 310–330. Basil Blackwell, Oxford (2001)
2. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152. ACM Press, New York (1992)
3. Breiman, L.: Bagging predictors. Technical report, Department of Statistics, Univ. of California, Berkeley (1994)
4. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)

5. Corwin, D.L., Lesch, S.M.: Application of soil electrical conductivity to precision agriculture: Theory, principles, and guidelines. Agron J. 95(3), 455–471 (2003)
6. Cressie, N.A.C.: Statistics for Spatial Data. Wiley, New York (1993)
7. Griffith, D.A.: Spatial Autocorrelation and Spatial Filtering. In: Advances in Spatial Science, Springer, New York (2003)
8. Gunn, S.R.: Support vector machines for classification and regression. Technical Report, School of Electronics and Computer Science, University of Southampton, Southampton, U.K (1998)
9. Liu, J., Miller, J.R., Haboudane, D., Pattey, E.: Exploring the relationship between red edge parameters and crop variables for precision agriculture. In: 2004 IEEE International Geoscience and Remote Sensing Symposium, vol. 2, pp. 1276–1279 (2004)
10. Goodchild, M., Anselin, L., Appelbaum, R., Harthorn, B.: Toward spatially integrated social science. International Regional Science Review 23, 139–159 (2000)
11. Meier, U.: Entwicklungsstadien mono- und dikotyler Pflanzen. Biologische Bundesanstalt für Land- und Forstwirtschaft, Braunschweig, Germany (2001)
12. Mejía-Guevara, I., Kuri-Morales, Á.: Evolutionary feature and parameter selection in support vector regression. In: Gelbukh, A., Kuri Morales, Á.F. (eds.) MICAI 2007. LNCS (LNAI), vol. 4827, pp. 399–408. Springer, Heidelberg (2007)
13. Middleton, E.M., Campbell, P.K.E., Mcmurtrey, J.E., Corp, L.A., Butcher, L.M., Chappelle, E.W.: "Red edge" optical properties of corn leaves from different nitrogen regimes. In: 2002 IEEE International Geoscience and Remote Sensing Symposium, vol. 4, pp. 2208–2210 (2002)
14. Moran, P.A.P.: Notes on continuous stochastic phenomena. Biometrika 37, 17–33 (1950)
15. Neeteson, J.J.: Nitrogen Management for Intensively Grown Arable Crops and Field Vegetables, ch. 7, pp. 295–326. CRC Press, Haren (1995)
16. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009) ISBN 3-900051-07-0
17. Ruß, G.: Data mining of agricultural yield data: A comparison of regression models. In: Perner, P. (ed.) Advances in Data Mining. Applications and Theoretical Aspects. LNCS, vol. 5633, pp. 24–37. Springer, Heidelberg (2009)
18. Ruß, G., Kruse, R., Schneider, M., Wagner, P.: Estimation of neural network parameters for wheat yield prediction. In: Bramer, M. (ed.) AI in Theory and Practice II, July 2008. Proceedings of IFIP-2008, vol. 276, pp. 109–118. Springer, Heidelberg (2008)
19. Ruß, G., Kruse, R., Schneider, M., Wagner, P.: Optimizing wheat yield prediction using different topologies of neural networks. In: Verdegay, J., Ojeda-Aciego, M., Magdalena, L. (eds.) Proceedings of IPMU 2008, June 2008, pp. 576–582. University of Málaga (2008)
20. Ruß, G., Kruse, R., Schneider, M., Wagner, P.: Visualization of agriculture data using self-organizing maps. In: Allen, T., Ellis, R., Petridis, M. (eds.) Applications and Innovations in Intelligent Systems, January 2009. Proceedings of AI-2008, vol. 16, pp. 47–60, BCS SGAI. Springer, Heidelberg (2009)
21. Ruß, G., Kruse, R., Wagner, P., Schneider, M.: Data mining with neural networks for wheat yield prediction. In: Perner, P. (ed.) ICDM 2008. LNCS (LNAI), vol. 5077, pp. 47–56. Springer, Heidelberg (2008)
22. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Technical report, Statistics and Computing (1998)
23. Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging (Springer Series in Statistics). Springer, Heidelberg (June 1999)
24. Weigert, G.: Data Mining und Wissensentdeckung im Precision Farming - Entwicklung von ökonomisch optimierten Entscheidungsregeln zur kleinräumigen Stickstoff-Ausbringung. PhD thesis, TU München (2006)