Spatial Variable Importance Assessment for Yield Prediction in Precision Agriculture

Georg Ruß[1] and Alexander Brenning[2]

[1] Otto-von-Guericke-Universität Magdeburg, Germany
[2] University of Waterloo, Canada

Abstract. *Precision Agriculture* applies state-of-the-art GPS technology in connection with site-specific, sensor-based crop management. It can also be described as a data-driven approach to agriculture, which is strongly connected with a number of data mining problems. One of those is also an inherently important task in agriculture: yield prediction. Given a yield prediction model, which of the predictor variables are the important ones?

In the past, a number of approaches have been proposed towards this problem. For yield prediction, a broad variety of regression models for non-spatial data can be adapted for spatial data using a novel spatial cross-validation technique. Since this procedure is at the core of variable importance assessment, it will be briefly introduced here. Given this spatial yield prediction model, a novel approach towards assessing a variable's importance will be presented. It essentially consists of picking each of the predictor variables, one at a time, permutating its values in the test set and observing the deviation of the model's RMSE. This article uses two real-world data sets from precision agriculture and evaluates the above procedure.

Keywords: Precision Agriculture, Spatial Data Mining, Regression, Spatial Cross-Validation, Variable Importance.

1 Introduction

Data-driven technology has become part of our everyday lives, while data-driven management techniques have become necessary and common in industry and services. Improvements in efficiency can be made in almost any part of businesses. This is especially true for agriculture, due to the modernization and better affordability of state-of-the-art GPS technology. Agricultural companies nowadays harvest not only crops but also a growing amount of data. Site-specific crop management (SSM) therefore heavily depends on knowledge discovery from large amounts of site-specific, possibly noisy geodata. This led to the term *precision agriculture* (PA) being coined. PA is an agricultural concept based on the assumption of in-field heterogeneity. The abovementioned GPS, as well as ground-based, aerial or satellite sensors and image acquisition in connection with geographic information systems (GIS) allow to assess and understand variations. These data are nowadays routinely collected and available to farm operators. It can be expected that a large amount of information is contained, yet hidden, in these agricultural field data. This is usually information about the soil and crop properties enabling a

P.R. Cohen, N.M. Adams, and M.R. Berthold (Eds.): IDA 2010, LNCS 6065, pp. 184–195, 2010.

higher operational efficiency – appropriate data processing 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. In PA, it must additionally be taken into account that the data are spatial: each data record has a specific location on the field and is accompanied by natural neighboring data points. Therefore, it must not be considered independent of its neighbors.

Yield prediction is a specific agricultural problem commonly occurring. As early as possible, a farmer would like to know 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 is assumed to be available in the data collected during normal farming operations throughout the season(s). A multitude of sensor data are nowadays collected, measuring a field's heterogeneity. These data are fine-scale, often highly correlated and carry spatial information which must not be neglected. Furthermore, it is of interest to know which of these sensor data are the most relevant for a yield prediction setup. Given novel sensors, we would like to assess whether they contribute novel information or whether this information is not already contained in more traditional data sources.

The problem of yield prediction can be treated as a problem of data mining and, specifically, regression. However, it should be noted that a regression problem on spatial data must be treated differently from regression on non-spatial data, as described in [7]. This article will serve as a continuation of [19]: in the previous article, the spatial data were treated with regression models which do not take the spatial relationships into account. This lead to serious underestimation of the true prediction error, which is shown in [20] where we compared the results on non-spatial data with those obtained on spatial data. This article builds on the previously established suitable spatial cross-validation framework (summarised in Section 3.1) and presents an approach to assess the importance of certain variables in the abovementioned precision agriculture data sets.

1.1 Research Target

The main research target of this work is to build upon yield prediction approaches to establish a novel approach towards assessing a variable's importance. The regression work presented in [19,23] will be used as a baseline for this work. The spatial regression model with spatial cross-validation has recently been described in [20] (a short summary is given below) and will be used as a core of the proposed variable importance assessment. Our approach will be described in detail and results from its application on two precision agriculture data sets will be detailed below.

1.2 Article Structure

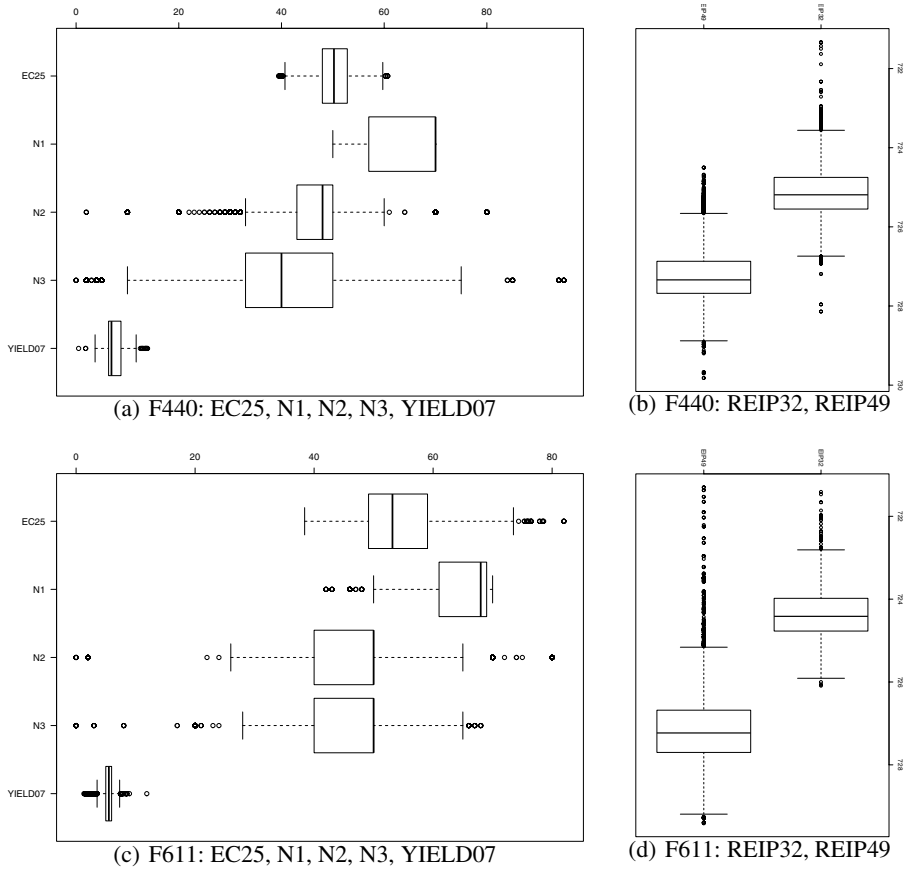
This article will start with a brief introduction into the area of precision agriculture and a more detailed description of the available data in Section 2. This will be followed by an outline of the key techniques, a short summary of our novel spatial sampling

technique described previously in [20], as well as our proposed approach towards variable importance assessment in Section 3. The results obtained from 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

This section will present a summary on the available data sets.

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 [24] to a grid with a resolution of 10 by 10 meters. 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. These available attributes



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

(b) F440: REIP32, REIP49

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

(d) F611: REIP32, REIP49

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

Table 1. Statistical summary of data sets, see also Figure 1

	F440				F611			
	min	mean	median	max	min	mean	median	max
EC25	39.47	50.13	50.22	60.69	38.41	54.44	53.17	81.98
N1	50.00	63.57	70.00	70.00	42.00	65.09	68.00	70.00
N2	2.00	47.60	48.00	80.00	0.00	47.89	50.00	80.00
N3	0.00	37.98	40.00	95.00	0.00	45.61	50.00	68.00
REIP32	721.33	725.11	725.19	728.14	721.41	724.37	724.41	726.09
REIP49	724.50	727.20	727.34	729.82	721.30	727.12	727.23	729.41
YIELD07	0.49	7.37	6.89	13.92	1.32	5.42	5.51	11.88

will be described in the following. In total, for the F440 field there are 6446 records, for F611 there are 4970 records. Descriptive statistics are displayed in Table 1. For further information on the available data attributes we refer to [19].

The response variable YIELD07 (wheat yield) is measured in metric tons per hectare ($\frac{t}{ha}$). The soil's apparent electrical conductivity EC25 is measured by non-invasive geophysical instruments and represents a number of physical soil properties. The *red edge inflection point* (REIP32, REIP49) values are obtained through image processing of high-resolution imagery of the field. The plants' chlorophyll content can be measured by calculating the REIP value and allows to deduce the plants' state of nutrition and thus the previous crop growth. The 32 and 49 numbers in the indicators refer to the growing stage of winter wheat. The amount of fertilizer applied to each subfield can be measured, resulting in three attributes N1, N2, N3. N1 on the F440 field had only four levels from {50,57,60,70}.

3 Spatial Data Mining

Spatial autocorrelation is the correlation among values of a single variable attributable to the proximity of those values in geographic space, introducing a deviation from the independent observations assumption of classical statistics. Spatial autocorrelation can be assessed using the semivariogram, and its presence is often known beforehand because of the nature of the data at hand [9].

In previous articles using the above data, such as [21,19], the main focus was on finding a suitable regression model to predict the current year's yield sufficiently well. However, it should be noted that the used regression models, such as neural networks [21,22] or support vector regression [19], among others, generally assume statistical independence of the data records. However, with the given geolocated data records at hand, this is clearly not the case, due to (natural) spatial autocorrelation (cp. [7]). Therefore, the spatial relationships between data records have to be taken into account.

3.1 Spatial vs. Non-spatial Data Treatment

To account for these spatial relationships, a novel and model-independent spatial cross-validation technique has been presented in [20] and is summarised below. It has partly

been adapted from existing approaches [4,5] towards the context of crop yield prediction and spatial regression more generally.

The approach consists of subdividing the agriculture site into spatially contiguous regions. This may be performed by overlaying the site with a regular grid. However, there are drawbacks with this approach: due to the fields being different, the grid would have to be manually adapted to each field being processed. Furthermore, agriculture fields are rarely of regular shape, hence at the field borders the grid cells are likely to be irregular as well, leading to further necessary processing. A third issue is to which of the surrounding grid cells a data record on the cell borders should be assigned. To overcome these issues, a clustering-based approach has been developed. This simple, yet effective method employs k-means clustering on the data records' coordinates (longitude, latitude) to partition the site into k contiguous parts of roughly equal size. This is depicted for the F440 site in Figure 2.

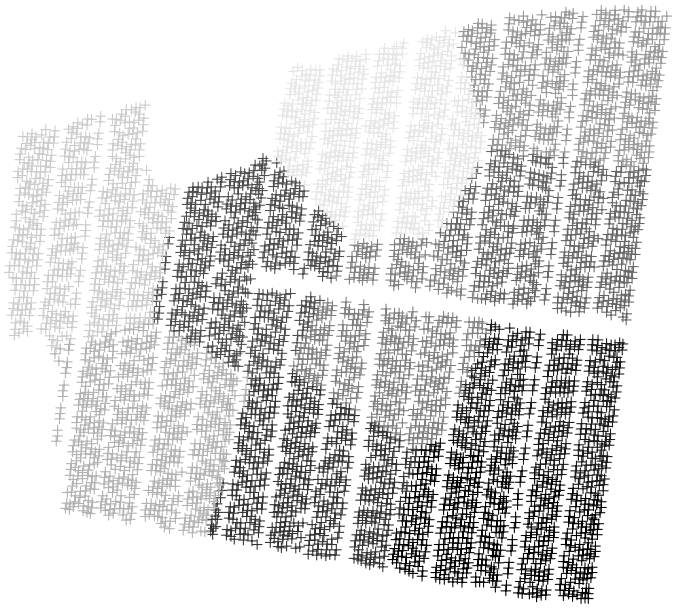


Fig. 2. k-means clustering on F440, $k = 10$

Having partitioned the field into k subfields, cross-validation at the level of spatial partitions may be carried out in order to assess the predictive performance of different regression models in the spatial domain. The only difference is that the subdivision of the whole data set into training and test subsets has to be done according to the spatial clusters determined in the previous step. This avoids the issue of having the same or similar samples in training and test sets, which non-spatial techniques mostly neglect. In a non-spatial setup, this leads to model overfitting and an underestimation of the prediction error (compare [4,6] for similar observations in a classification context). With the above procedure, the estimated prediction error is much less influenced by

model overfitting. Therefore, the prediction error of the spatial procedure is usually much higher than the one in a non-spatial setup. This result could be confirmed in [20] using tree-based regression techniques as well as a support vector machine.

3.2 Regression Techniques and Error Estimation

In previous work ([19,21]), numerous regression modeling techniques have been compared on similar data sets. It was determined that among those models which represent non-linear relationships between variables, SVR has constantly shown favorable RMSE values. It has furthermore recently been shown to work rather successfully in spatial classification tasks, albeit without spatial cross-validation, as in [17]. Resampling-based estimation methods (such as cross-validation and the bootstrap) for dependent data in general have been investigated recently in the context of time series data [8] and paired data [6].

In this work, SVR will be compared against standard linear regression modeling and tree-based models, all of those in the aforementioned spatial cross-validation setup. Experiments are conducted in R [18]. It is assumed that the reader is mostly familiar with the regression techniques below. Therefore, the techniques used are described in short. References to further details are given, where appropriate.

Support Vector Regression. Support Vector Machines (SVMs) are a supervised learning method discovered by [1]. They were originally described for the use in classification, but can also be applied to regression tasks, where optimization of a cost function is achieved. The model produced by support vector regression depends only on a subset of the training data – which are essentially the support vectors. Further details can be found in [19]. In the current experiments, the *svm* implementation from the *e1071* R package has been used. We set a radial kernel and `cost=50`. These settings were determined empirically by `best.svm`.

Regression Trees. Regression trees have seen some usage in agriculture [10,12,16]. Essentially, they are a special case of decision trees where the outcome (in the tree leaves) is a continuous function instead of a discrete classification. The *rpart* R package has been used, with the settings for `rpart.control` of `minsplit=30` and the complexity parameter `cp=0.001`, `pruning` and `maxdepth` were left at the defaults. These settings were determined experimentally.

Bootstrap Aggregating. Bootstrap aggregating (or bagging) [2] 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. The bagging implementation in the R *ipred* package has been used here. We set `nbagg=250`, while leaving the parameters in `rpart.control` at their defaults.

Random Forests. According to [3], 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. Random forests are

basically a variant of bagging where regression trees are used as the internal predictor. In addition to resampling the observations, it also resamples the variables, which prevents overfitting, especially in high-dimensional problems. The *randomForest* R package has been used. The setting for `ntree` was kept at the default of 500 trees to grow.

Linear Regression. Linear regression is the most widely used regression technique in an agricultural context. It is therefore used as a benchmark model for accuracy assessment. We use the implementation of ordinary-least-squares linear regression *lm* from the *stats* package.

The performance of the models will be determined using the root mean squared error (RMSE). The RMSE is based on the difference between an observed target value y_a and the model prediction y.

3.3 Variable Importance

The interdependencies between variables in the data sets rule out standard feature selection approaches such as *forward selection* [14] or *backward elimination* [11]. A relatively new and intuitive computational approach for assessing variable importance is based on measuring the increase in prediction error associated with permuting a predictor variable [26]. We adapt this approach to spatial prediction problems by assessing RMSE increase on spatial cross-validation partitions rather than non-spatial test sets.

Given any model fitted on a cross-validation training sample, we choose one variable at a time and permute its values randomly in the test set. The remaining variables are left unchanged. We quantify the increase in RMSE caused by the permutation. No specific model is needed, each of the regression models is suitable for assessing the variable importance in this way. To obtain a sufficient number of replications, the permutation is repeated 200 times for each variable. The permutation is embedded in a 10-fold leave-one-out cross-validation setup, after having partitioned the agriculture sites into $k = 10$ sub-parts. The partitioning is rather stable due to the characteristics of the used clustering technique, therefore the k-means procedure is repeated only 10 times. The setting of parameter $k = 10$ was determined empirically and is currently under investigation.

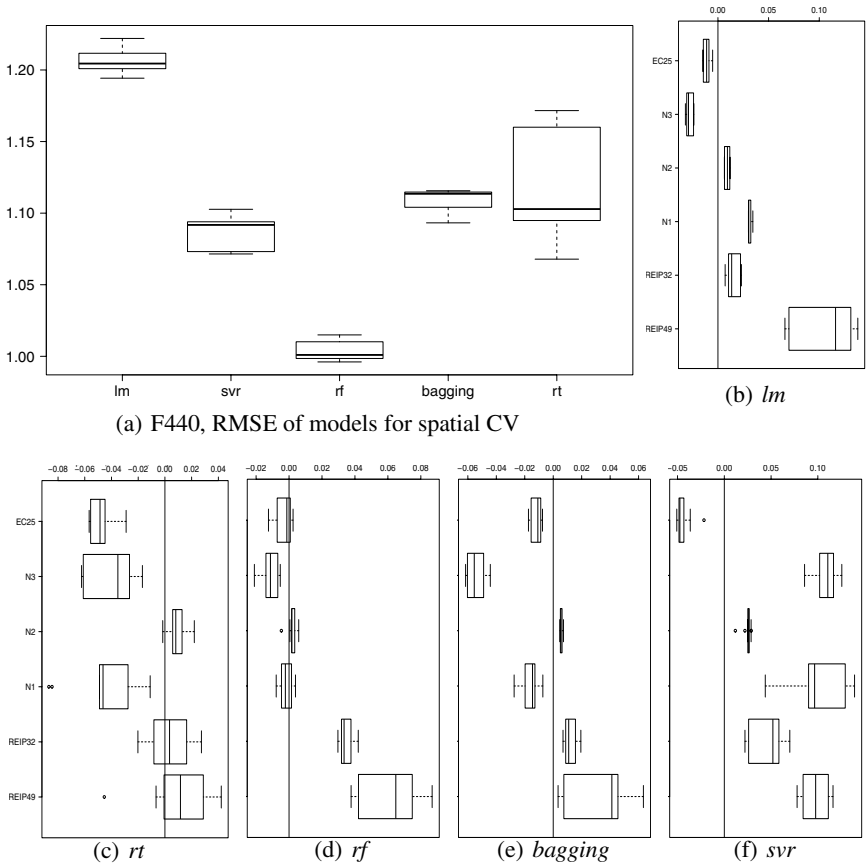
4 Results

The main research target of this article is to assess the importance of variables in a spatial cross-validation setup. The results for the two available data sets described in Section 2 are presented below.

4.1 Results for F440

The overall RMSE of different models for the F440 field can be seen in Figure 3(a). *lm* returns the highest RMSE at 1.2 dt/ha, *rf* the lowest at around 1 dt/ha. This range has to be kept in mind when considering the RMSE increase after variable permutation.

The relative order of the variable importance for the tree-based models *bagging*, *rf* and *rt* is quite similar, which is due to the underlying tree construction. In each case,



(a) F440, RMSE of models for spatial CV

(b) *lm*

(c) *rt*

(d) *rf*

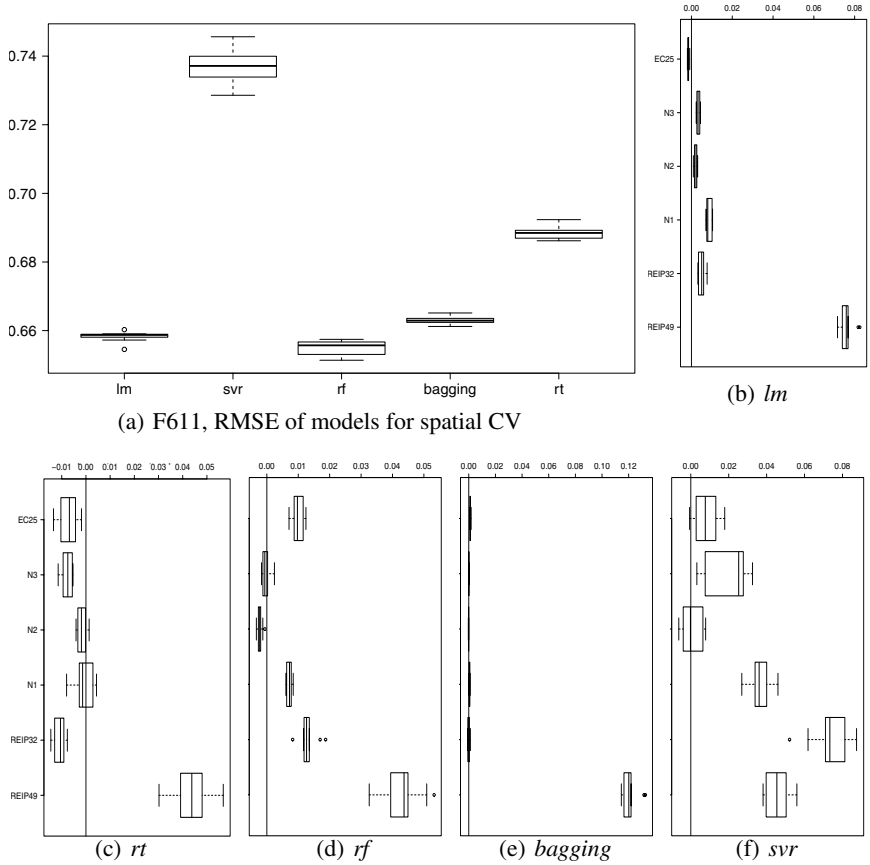
(e) *bagging*

(f) *svr*

Fig. 3. F440: 3(a) shows each model's RMSE for a spatial cross-validation setup with 10 repetitions, 10 clusters; remaining figures show the increase in RMSE after permuting the respective variable (200 permutations): 3(b) *lm*, 3(c) *rt*, 3(d) *rf*, 3(e) *bagging*, 3(f) *svr*; the predictors are EC25, N3, N2, N1, REIP32, REIP49 (top to bottom)

the best predictor is the REIP49 variable with an increase of 0.06 at an average model RMSE of 1.0 (rf). This was expected, since the REIP49 value is closest to harvest and represents the amount of biomass on the field. For *bagging* and *rf*, the remaining variables increase the RMSE much less when permuted. For *rf*, the second most important predictor is the REIP32 variable – this is also the expected behaviour since the application of N2 and N3 tries to equalize the vegetation growth.

The remaining variables, especially the fertilizer dressings, are rather insignificant in the tree-based models. Since the N1 variable has only four different values in F440, it is clear that it will be rarely built into the trees since these are usually biased towards variables with a high number of values. Similarly, this also holds true for N2, N3: these have 45 and 50 different levels, whereas REIP32 and REIP49 have 367 and 397 different levels, respectively. Nevertheless, EC25 has 851 different levels and still has close to no importance for yield prediction. Another reason for the insignificance of the variables



(a) F611, RMSE of models for spatial CV

(b) *lm*

(c) *rt*

(d) *rf*

(e) *bagging*

(f) *svr*

Fig. 4. F611: 4(a) shows each model's RMSE for a spatial cross-validation setup with 10 repetitions, 10 clusters; remaining figures show the increase in RMSE after permuting the respective variable (200 permutations): 4(b) *lm*, 4(c) *rt*, 4(d) *rf*, 4(e) *bagging*, 4(f) *svr*; the predictors are EC25, N3, N2, N1, REIP32, REIP49 (top to bottom)

with a low number of distinct levels is that it is much more likely that the variable's value in the test set is the same before and after the permutation.

Different results can be reported for *svr*. While the EC25 variable is still insignificant, most importance is given to N1, N3 and REIP49. This different ordering is likely to be due to *svr* measuring smooth non-linear variable importance with interaction, while tree-based models measure non-linear variable importance with higher-order interactions and discontinuities.

4.2 Results for F611

The overall RMSE for each of the models is quite different from the F440 site. *lm* returns an RMSE of 0.66 dt/ha and is very close to being the best model. *rf* and *bagging* seem are rather close around 0.66 dt/ha, while *rt* and *svr* are rather far from this mark

at 0.69 and 0.73 dt/ha. This might hint to linear interactions between the variables. The rather bad performance of *svr* might as well be attributed to simple linear interactions.

The results for the variable importance are similar to the ones on the F440 site. Except for the *svr* model, the REIP49 value is by far the most important predictor. It influences the *lm* model's RMSE by 0.08 dt/ha at an average RMSE of 0.66 dt/ha. Similarly, the *bagging* model's RMSE is raised by 0.12 dt/ha at an average RMSE of roughly 0.66 dt/ha. There may yet be non-linear, high-order interactions, seen in the *svr*, where REIP32 is the most important variable, followed by REIP49 and N1. However, since the *svr*'s overall level of RMSE is higher at 0.74 dt/ha and the RMSE increase of the *REIP* values is around 0.07 and 0.04 dt/ha, this result should be taken with care.

4.3 Remarks

Assessing the importance of individual variables in multiple-variable models is difficult since the importance of any particular variable is influenced by other, correlated or interacting, variable in the data set. In our data sets, due to agricultural management decisions, the N2 and N3 variable are determined partly based on the REIP32 and REIP49 values and are therefore not fully independent of those. Therefore, e.g. if one of REIP32,REIP49 has been included in the model, the remaining variable N2,N3 is less likely to be chosen. If the other variable is included nevertheless, it is likely to have a smaller influence on the prediction than if included without the presence of the other variable.

5 Conclusions and Future Work

This article focused on a central data analysis task: regression. A procedure for measuring spatial variable importance has been developed and tested using two real-world data sets. This was achieved by adopting a variable importance approach and combining it with a clustering-based spatial cross-validation approach developed earlier [20]. The resulting spatial variable importance assessment was tested using two precision agriculture data sets. The key question which of the available data attributes is informative regarding yield prediction was answered.

Based on practical experiences, the hypothesis that the vegetation indicators REIP49 and REIP32 should be most important for yield prediction could be confirmed. Although the different modeling techniques are partly biased (as explained), it is highly likely that the N1 and N3 dressings have a significant impact on yield, whereas the EC25 and N2 variables are rather unimportant. It can also be confirmed that the results of variable importance are quite similar for two different fields, growing the same crop in the same season and in the same climatic region.

The developed procedure may be applied to further areas where similar spatial data need to be analysed, such as land cover classification [5], landslide susceptibility modeling [4], species habitat analysis [13,15].

5.1 Future Work

Despite having improved and validated the yield prediction task, the data sets carry further information. From a theoretical point of view, our spatial cross-validation setup is

a spatially constrained case of a standard cross-validation setup. Therefore, it would be interesting to see whether the spatial case converges towards the non-spatial case when the number of clusters is raised. Furthermore, additional predictor variables should be considered in future yield prediction setups.

Conditional permutation-based variable importance measures as proposed by [25] provide a framework for investigating the importance of variables conditional on other variables, which provides additional insights when interactions are present, in the context of the F440/F611 data sets, e.g. between N2/N3 and REIP32/REIP49, as noted above.

Acknowledgements

The F440 and F611 data sets were obtained from Martin Schneider and Peter Wagner, Professur für Landwirtschaftliche Betriebslehre, Martin-Luther-Universität Halle-Wittenberg, Germany.

References

1. 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)
2. Breiman, L.: Bagging predictors. Technical report, Department of Statistics, Univ. of California, Berkeley (1994)
3. Breiman, L.: Random forests. Machine Learning, 45(1):5–32 (2001)
4. Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Science 5(6), 853–862 (2005)
5. Brenning, A., Itzerott, S.: Comparing classifiers for crop identification based on multitemporal landsat tm/etm data. In: Proceedings of the 2nd workshop of the EARSeL Special Interest Group Remote Sensing of Land Use and Land Cover, September 2006, pp. 64–71 (2006)
6. Brenning, A., Lausen, B.: Estimating error rates in the classification of paired organs. Statistics in Medicine 27(22), 4515–4531 (2008)
7. Brenning, A., Piotraschke, H., Leithold, P.: Geostatistical analysis of on-farm trials in precision agriculture. In: Ortiz, J.M., Emery, X. (eds.) GEOSTATS 2008, Proceedings of the Eighth International Geostatistics Congress, December 12, vol. 2, pp. 1131–1136 (2008)
8. Bühlmann, P.: Bootstraps for time series. Statistical Science 17, 52–72 (2002)
9. Cressie, N.A.C.: Statistics for Spatial Data. Wiley, New York (1993)
10. Crone, S.F., Lessmann, S., Pietsch, S.: Forecasting with computational intelligence - an evaluation of support vector regression and artificial neural networks for time series prediction. In: International Joint Conference on Neural Networks, IJCNN 2006, pp. 3159–3166 (2006)
11. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis 1, 131–156 (1997)
12. Huang, C., Yang, L., Wylie, B., Homer, C.: A strategy for estimating tree canopy density using landsat 7 etm+ and high resolution images over large areas. In: Proceedings of the Third International Conference on Geospatial Information in Agriculture and Forestry (2001)
13. Knudby, A., Brenning, A., LeDrew, E.: New approaches to modelling fish-habitat relationships. Ecological Modelling 221, 503–511 (2010)
14. Langley, P.: Selection of relevant features in machine learning. In: Proceedings of the AAAI Fall symposium on relevance, pp. 140–144. AAAI Press, Menlo Park (1994)

15. Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T., Taylor, P.: Variation in demersal fish species richness in the oceans surrounding new zealand: an analysis using boosted regression trees. Marine Ecology Progress 321, 267–281 (2006)
16. Lobell, D.B., Ortiz-Monasterio, J.I., Asner, G.P., Naylor, R.L., Falcon, W.P.: Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. Agronomy Journal 97, 241–249 (2005)
17. Pozdnoukhov, A., Foresti, L., Kanevski, M.: Data-driven topo-climatic mapping with machine learning methods. Natural Hazards 50(3), 497–518 (2009)
18. 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
19. 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)
20. Ruß, G., Brenning, A.: Data mining in precision agriculture: Management of spatial information. In: Proceedings of IPMU 2010. Springer, Heidelberg (submitted for review 2010)
21. Ruß, G., Kruse, R., Schneider, M., Wagner, P.: Estimation of neural network parameters for wheat yield prediction. In: Bramer, M. (ed.) Proceedings of AI in Theory and Practice II, IFIP 2008, July 2008, vol. 276, pp. 109–118. Springer, Heidelberg (2008)
22. 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)
23. 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)
24. Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging, June 1999. Springer Series in Statistics. Springer, Heidelberg (1999)
25. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC Bioinformatics 9(1), 307 (2008)
26. Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8(1), 25 (2007)