Visualization of Agriculture Data Using Self-Organizing Maps

Georg Ruß, Rudolf Kruse, Martin Schneider, Peter Wagner

Abstract The importance of carrying out effective and sustainable agriculture is getting more and more obvious. In the past, additional fallow ground could be tilled to raise production. Nevertheless, even in industrialized countries agriculture can still improve on its overall yield. Modern technology, such as GPS-based tractors and sensor-aided fertilization, enables farmers to optimize their use of resources, economically and ecologically. However, these modern technologies create heaps of data that are not as easy to grasp and to evaluate as they have once been. Therefore, techniques or methods are required which use those data to their full capacity – clearly being a data mining task. This paper presents some experimental results on real agriculture data that aid in the first part of the data mining process: understanding and visualizing the data. We present interesting conclusions concerning fertilization strategies which result from data mining.

Key words: Precision Farming, Data Mining, Self-Organizing Maps, Neural Networks

1 Introduction

Recent worldwide economic development shows that agriculture will play a crucial role in sustaining economic growth, both in industrialized as well as in developing countries. In the latter countries agricultural development is still in its early stages and production improvements can easily be achieved by simple means like

Georg Ruß, Rudolf Kruse
Otto-von-Guericke-Univ. Magdeburg, Germany e-mail: {russ,kruse}@iws.cs.uni-magdeburg.de

Martin Schneider
Agri Con GmbH, Germany

Peter Wagner
Martin-Luther-University Halle-Wittenberg, Germany e-mail: peter.wagner@landw.uni-halle.de

introduction of fertilization. In industrialized countries, on the other hand, even the agricultural sector is mostly quite industrialized itself, therefore improvements are harder to achieve. Nevertheless, due to the adoption of modern GPS technology and the use of ever more different sensors on the field, the term *precision farming* has been coined. According to [16], precision farming is the sampling, mapping, analysis and management of production areas that recognises the spatial variability of the cropland.

In artificial intelligence terms, the area of precision farming (PF) is quite an interesting one as it involves methods and algorithms from numerous areas that the artificial intelligence community is familiar with. When analyzing the data flow that results from using PF, one is quickly reminded of *data mining*: an agriculturist collects data from his cropland (e.g., when fertilizing or harvesting) and would like to extract information from those data and use this information to his (economic) advantage. A simplified data flow model can be seen in Figure 1. Therefore, it is clearly worthwile to consider using AI techniques in the light of precision farming.

1.1 Research Target

With this contribution we aim at finding suitable methods to visualize agricultural data with a high degree of precision and generality. We present different data sets which shall be visualized. We present experimental results on real and recent agricultural data. Our work helps in visualizing and understanding the available data, which is an important step in data mining.

1.2 Article Structure

This article concentrates on the third and fourth step of the data flow model from Figure 1, namely building and evaluating different models. Here, the modeling will clearly be aimed at visualizing the data. Nevertheless, details which are necessary for the understanding and judgment of the modeling stage will not be omitted. This article starts with a description of the data and (partly) how they have been acquired in Section 2. After the data have briefly been shown, the existing modeling approach and the basics of self-organizing maps will be shown in Section 3. Section 4 is at the core of this article: the different data sets will be visualized and conclusions will be drawn from the visualisations – and compared with farmers' experience. Section 5 presents a short conclusion and lays out our future work.

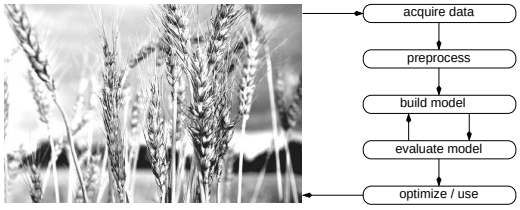


Fig. 1: Data mining for agriculture data

2 Data Description

The data available in this work have been obtained in the year 2006 on a field near Köthen, north of Halle, Germany[1] All information available for these 72- and 32-hectare fields[2] was interpolated using kriging [12] to a grid with 10 by 10 meters grid cell sizes. Each grid cell represents a record with all available information. During the growing season of 2006, the field was subdivided into different strips, where various fertilization strategies were carried out. For an example of various managing strategies, see e.g. [11], which also shows the economic potential of PA technologies quite clearly. The field grew winter wheat, where nitrogen fertilizer was distributed over three application times during the growing season.

Overall, there are seven input attributes – accompanied by the yield in 2006 as the target attribute. Those attributes will be described in the following. In total, for the smaller field (F131) there are 2278 records, for the larger field (F330) there are 4578 records, thereof none with missing values and none with outliers.

2.1 Nitrogen Fertilizer – N1, N2, N3

The amount of fertilizer applied to each subfield can be easily measured. It is applied at three points in time into the vegetation period.

2.2 Vegetation – REIP32, REIP49

The *red edge inflection point* (REIP) is a first 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

[1] GPS: Latitude N 51 40.430, Longitude E 11 58.110

[2] We will call them *F330* and *F131*, respectively

nitrogen availability (see, e.g. [6]), 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 [15] or [5]. 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. For the range of REIP values encountered in the available data, see Tables 1 and 2. The numbers in the REIP32 and REIP49 names refer to the growing stage of winter wheat.

2.3 Electric Conductivity – EM38

A non-invasive method to discover and map a field's heterogeneity is to measure the soil's conductivity. Commercial sensors such as the EM-38[3] 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 the range of EM values encountered in the available data, see Tables 1 and 2.

2.4 YIELD *2005/2006*

Here, yield is measured in metric tons per hectare ($\frac{t}{ha}$), where one metric ton equals roughly 2204 pounds and one hectare roughly equals 2.47 acres. For the yield ranges for the respective years and sites, see Tables 1 and 2. It should be noted that for both data sets the yield was reduced significantly due to bad weather conditions (lack of rain) during the growing season 2006.

2.5 Data Overview

In this work, we evaluate data sets from two different fields. A brief summary of the available data attributes for both data sets is given in Tables 1 and 2. On each field, different fertilization strategies have been used as described in Section 2.6. For each field, one data set will contain all records, thus containing all the different fertilization strategies. Another data set for each field will be a subset of the first that only contains those data records where the MLP has been used, respectively. Table 3 serves as a short overview about the resulting four different data sets.

[3] trademark of Geonics Ltd, Ontario, Canada

Table 1: Data overview, F131

F131	*min*	*max*	*mean*	*std*	*Description*
YIELD05	1.69	10.68	5.69	0.93	yield in 2005
EM38	51.58	84.08	62.21	8.60	electrical conductivity of soil
N1	47.70	70	64.32	6.02	amount of nitrogen fertilizer applied at the first date
N2	14.80	100	51.71	15.67	amount of nitrogen fertilizer applied at the second date
N3	0	70	39.65	13.73	amount of nitrogen fertilizer applied at the third date
REIP32	719.6	724.4	722.6	0.69	red edge inflection point vegetation index
REIP49	722.3	727.9	725.8	0.95	red edge inflection point vegetation index
YIELD06	1.54	8.83	5.21	0.88	yield in 2006

Table 2: Data overview, F330

F330	*min*	*max*	*mean*	*std*	*Description*
YIELD05	4.64	14.12	10.62	0.97	yield in 2005
EM38	25.08	49.48	33.69	2.94	electrical conductivity of soil
N1	24.0	70	59.48	14.42	amount of nitrogen fertilizer applied at the first date
N2	3.0	100	56.38	13.35	amount of nitrogen fertilizer applied at the second date
N3	0.3	91.6	50.05	12.12	amount of nitrogen fertilizer applied at the third date
REIP32	719.2	724.4	721.5	1.03	red edge inflection point vegetation index
REIP49	723.0	728.5	726.9	0.82	red edge inflection point vegetation index
YIELD06	1.84	8.27	5.90	0.54	yield in 2006

Table 3: Overview on available data sets for specific fertilization strategies for different fields

F131-all	YIELD05, EM38, N1, REIP32, N2, REIP49, N3, YIELD06, *fert. strategy*
F131-net	subset of F131-all where fertilization strategy is *neural network*
F330-all	YIELD05, EM38, N1, REIP32, N2, REIP49, N3, YIELD06, *fert. strategy*
F330-net	subset of F330-all where fertilization strategy is *neural network*

2.6 Fertilization Strategies

There were three different strategies that have been used to guide the nitrogen fertilization of the fields. F131 contains data resulting from two strategies (F, N) and F330 contains data from three strategies (F, N, S). The three strategies are as follows:

F – uniform distribution of fertilizer according to long-term experience of the farmer

N – fertilizer distribution was guided by an economic optimization with a multilayer perceptron model; the model was trained using the above data with the current year's yield as target variable that is to be predicted

S – based on a special nitrogen sensor – the sensor's measurements are used to determine the amount of nitrogen fertilizer that is to be applied.

2.7 Points of Interest

From the agricultural perspective, it is interesting to see how much the influencable factor "fertilization" really influences the yield in the current site-year. Furthermore, there may be additional location factors that correlate directly or indirectly with yield and which can not be discovered using regression or correlation analysis techniques like principal component analysis. Self-organizing maps (SOMs), on the other hand, provide a relatively self-explanatory way to analyse those yield data visually, find correlations and, eventually, make predictions for current year's yield from past data. The overall research target is to find those indicators of a field's heterogeneity which are optimal for prediction. In this paper we will present advances in visualizing the available data with SOMs which helps in understanding and will ultimately lead to new heterogeneity indicators. The following section will briefly summarize an appropriate technique to model the data that we have presented in earlier work. Afterwards, SOMs will be outlined briefly, with the main focus on data visualization.

3 Using Multi-Layer Perceptrons and Self-organizing Maps Approach

This section deals with the basic techniques that we used to model and visualize the agricultural yield data. For modeling, we have used Multi-Layer Perceptrons, as discussed in [10]. To visualize the data we will use Self-Organizing Maps (SOMs). Therefore, SOMs will comprise the main part of this section.

3.1 Multi-Layer Perceptrons for Modeling

In recent years, we have modeled the available data using a multi-layer perceptron (MLP). To gain more insights into what the MLP has learned, in this paper we will use self-organizing maps to try to better understand the data and the modeling process that underlies MLPs. In [7], neural networks have been used for optimization of fertilizer usage for wheat, in [13] the process has been carried out for corn. In [8] we could show that MLPs can be used for predicting current year's yield. For a detailed discussion of the used MLP structure and parameters, we refer to [9]. We basically used a feedforward-backpropagation multi-layer perceptron with two hidden layers. The network parameters such as the hidden layer sizes were determined experimentally. A prediction accuracy of between 0.45 and 0.55 metric tons per hectare (100×100 metres) at an average yield of $9.14 \frac{t}{ha}$ could be achieved by using this modeling technique.

3.2 Self-Organizing Maps for Visualization

Our approach of using SOMs is motivated by the need to better understand the available yield data and extract knowledge from those data. SOMs have been shown to be a practical tool for data visualization [1]. Moreover, SOMs can be used for prediction and correlation analysis, again, mostly visually [3]. As such, the main focus in explaining Self-Organizing Maps in the following will be on the visual analysis of the resulting maps.

Self-Organizing Maps have been invented in the 1990s by Teuvo Kohonen [4]. They are based on unsupervised competitive learning, which causes the training to be entirely data-driven and the neurons on the map to compete with each other. Supervised algorithms like MLPs or Support Vector Machines require the target attribute's values for each data vector to be known in advance whereas SOMs do not have this limitation.

Grid and Neigborhood: An important feature of SOMs that distinguishes them from Vector Quantisation techniques is that the neurons are organized on a regular grid. During training, not only the Best-Matching Neuron, but also its topological neighbors are updated. With those prerequisites, SOMs can be seen as a scaling method which projects data from a high-dimensional input space onto a typically two-dimensional map, preserving similarities between input vectors in the projection.

Structure: A SOM is formed of neurons located on a usually two-dimensional grid having a rectangular or hexagonal topology. Each neuron of the map is represented by a weight vector $m_i = [m_{i1}, \cdots, m_{in}]^T$, where n is equal to the respective dimension of the input vectors. The map's neurons are connected to adjacent neurons by a neighborhood relationship, superimposing the structure of the map. The number of neurons on the map determines the granularity of the resulting mapping, which, in turn, influences the accuracy and generalization capabilities of the SOM.

Training: After an initialization phase, the training phase begins. One sample vector **x** from the input data set is chosen and the similarity between the sample and each of the neurons on the map is calculated. The Best-Matching Unit (BMU) is determined: its weight vector is most similar to **x**. The weight vector of the BMU and its topological neighbors are updated, i.e. moved closer to the input vector. The training is usually carried out in two phases: the first phase has relatively large learning rate and neighborhood radius values to help the map adapt towards new data. The second phase features smaller values for the learning rate and the radius to fine-tune the map.

Visualization: The reference vectors of the SOM can be visualized via a component plane visualization. The trained SOM can be seen as multi-tiered with the components of the vectors describing horizontal layers themselves and the reference vectors being orthogonal to these layers. From the component planes the distribution of the component values and possible correlations between components can be obtained easily. The visualization of the component planes is the main feature of the SOMs that will be utilized in the following section.

In this work, we have used the Matlab SOM toolbox authored by [14] with the default presets and heuristics for determining map sizes and learning parameters.

4 Experimental Results

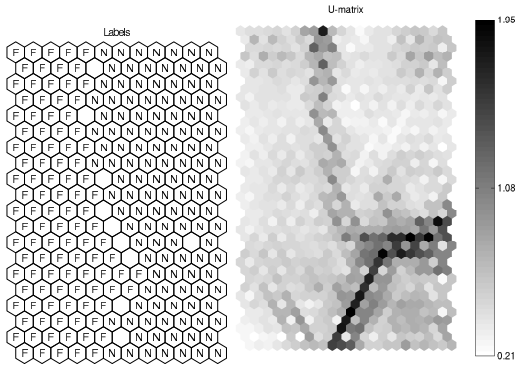
This section will present some of the experimental results that we have obtained using SOMs on agricultural data. The first two parts will deal with the analysis of the maps generated from the complete data set (containing different fertilization strategies). The subsequent two parts will deal with those subsets of the data where a MLP has been used for yield prediction and optimization. The data sets have been described in Section 2, an overview has been given in Table 3.

4.1 Results for F131-all

The full F131-all dataset consists of the **F** and **N** fertilization strategies where each data record is labeled accordingly. After training the SOM using the preset heuristics from the toolbox [14], the labeled map that results is shown in Figure 2a. The corresponding U-Matrix that confirms the clear separability of the two fertilization strategies is shown in Figure 2b. In Figures 3a to 3c the amount of fertilizer for the three different fertilization times is projected onto the same SOM. On those three maps it can also be seen that the different strategies are clearly separated on the maps. Another result can be seen in Figures 3d and 4b. As should be expected, the REIP49 value (which is an indicator of current vegetation on the field) correlates with the YIELD06 attribute. This hypothesis that we obtained from simple visual inspection of the SOM's component planes can be substantiated by the corresponding scatter plot in Figure 4c.

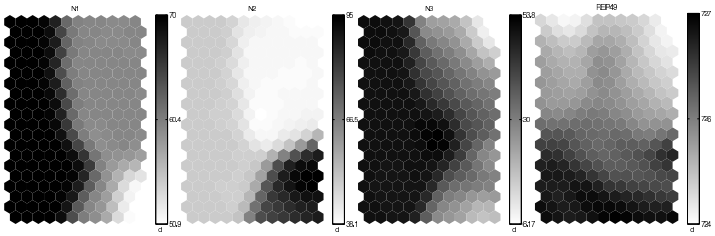
4.2 Results for F330-all

In contrast to the F131 dataset, F330 contains three different fertilization strategies. The "farm" strategy (labeled *F*), the "neural network (MLP)" strategy (labeled *N*) and the "sensor" strategy (labeled *S*) In Figure 5a it can be seen that, as in the preceding section, the *N* strategy is separable from the other two variants. However, the *F* and *S* strategies are not clearly separable. The U-matrix in Figure 5b also represents this behaviour. When looking at the projected values of N1, N2 and N3 in the component planes in Figures 6a to 6c, the differences between the *N* and *F* or *S* strategies are again clearly visible. There is, however, no such clear connection between the REIP49 (Figure 6d) and YIELD06 (Figure 7b) parameters as in the preceding section. This can also be seen from the scatter plot in Figure 7c. This



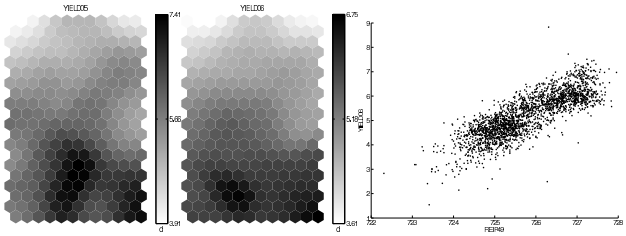
(a) Labels (b) U-Matrix

Fig. 2: F131-all, Labels and U-Matrix



(a) N1 (b) N2 (c) N3 (d) REIP49

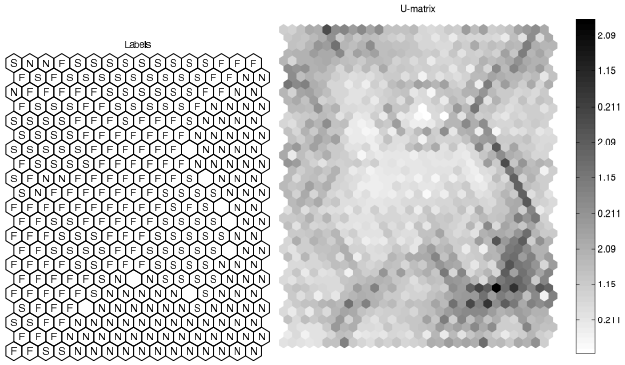
Fig. 3: F131-all: N1, N2, N3, REIP49



(a) YIELD05 (b) YIELD06 (c) Scatter plot for F131-all, REIP49 vs. YIELD06

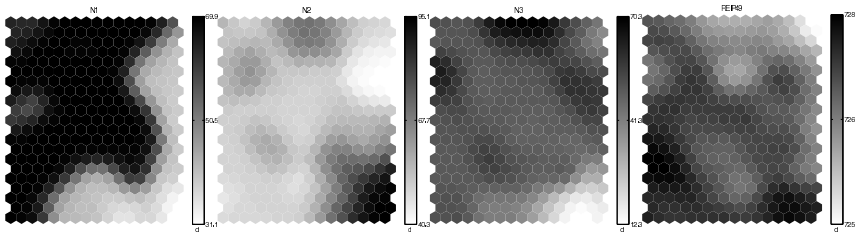
Fig. 4: F131-all: YIELD05, YIELD06, scatter plot

might be due to the fact that the overall yield was significantly reduced by bad weather conditions in 2006. Nevertheless, there is a certain similarity between the relative yields that can be easily obtained by comparing YIELD05 to YIELD06 in Figures 7a and 7b.



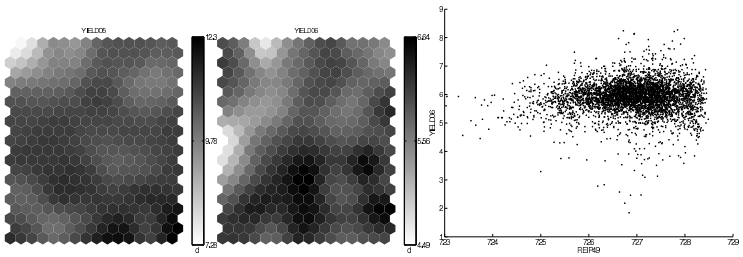
(a) Labels (b) U-Matrix

Fig. 5: F330-all, Labels and U-Matrix



(a) N1 (b) N2 (c) N3 (d) REIP49

Fig. 6: F330-all: N1, N2, N3, REIP49



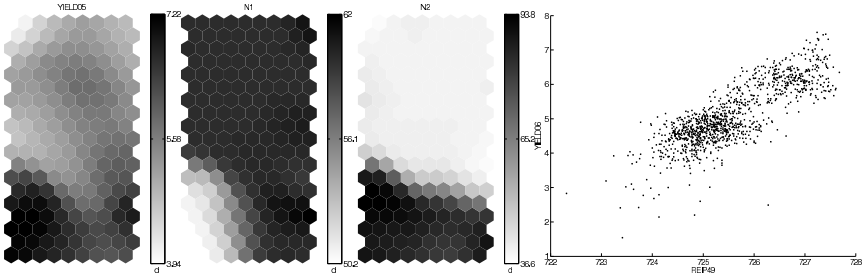
(a) YIELD05 (b) YIELD06 (c) Scatter plot for F330-all, REIP49 vs. YIELD06

Fig. 7: F330-all: YIELD05, YIELD06, scatter plot

4.3 Results for F131-net

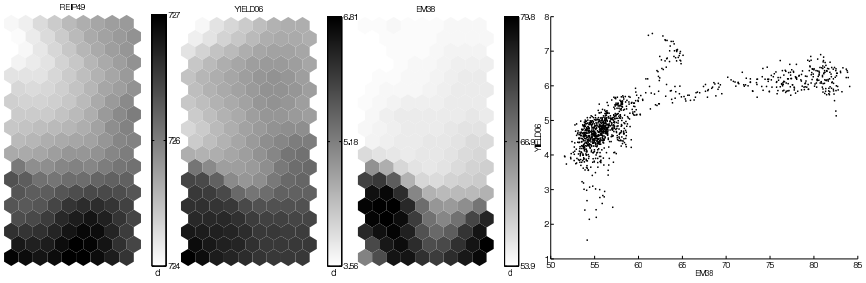
F131-net represents a subset of F131-all: it contains those data records from F131-all that were labeled *N*, i.e. in those field parts the neural network predictor was used for fertilizer optimization. Figures 8a and 8b seem to convey a connection: the MLP has learned that where YIELD05 was high (lower left of map), there is less need of N1 fertilizer whereas the rest of the field needs a high amount. For N2, another network is trained with more input, now N2 and YIELD05 seem to correlate (Figures 8a and 8c).

Furthermore, it is expected that REIP49 and YIELD06 correlate, as can be seen from Figures 9a and 9b. Furthermore, even the EM38 value for electromagnetic conductivity correlates with the said attributes, see Figure 9c. Additionally, the corresponding scatter plot in Figure 9d shows a separation between clusters of low EM38/YIELD06 values and high EM38/YIELD06 values.



(a) YIELD05 (b) N1 (c) N2 (d) Scatter plot for F131-net, REIP49 vs. YIELD06

Fig. 8: F131-net: YIELD05, N1, N2, scatter plot



(a) REIP49 (b) YIELD06 (c) EM38 (d) Scatter plot for F131-net, EM38 vs. YIELD06

Fig. 9: F131-net: REIP49, YIELD06, EM38, scatter plot

4.4 Results for F330-net

As in the preceding section, F330-net represents a subset of F330-all: it contains those data records from F330-all that were labeled *N*, i.e. in those field parts the MLP predictor was used for fertilizer optimization. Again, Figures 10a and 10b seem to convey a connection: the MLP has learned that where YIELD05 was high (lower left of map), there is less need of N1 fertilizer whereas the rest of the field needs a high amount. For N2, another network is trained with more input, now N2 and YIELD05 seem to correlate (Figures 10a and 10c), although the correlation is not as clear as with the F131-net dataset. Furthermore, it is expected that REIP49 and YIELD06 correlate, as can be seen from Figures 9a,9b and 8d. Furthermore, even the EM38 value for electromagnetic conductivity correlates with the said attributes, see Figure 9c. Additionally, the corresponding scatter plot in Figure 9d shows a separation between clusters of low EM38/YIELD06 values and high EM38/YIELD06 values.

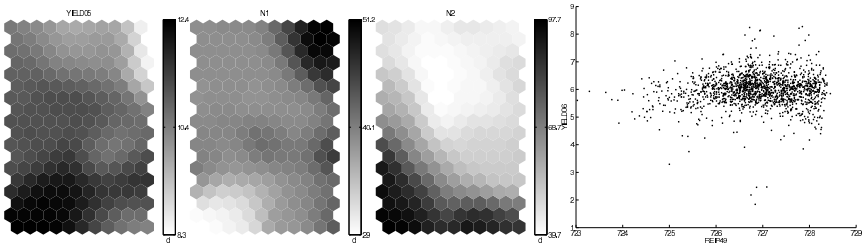
From the agricultural point of view, the F330 field is quite different from the one where the F131 data set was obtained, they are located 5.7km away from each other. This difference can be clearly shown on the SOMs. So, even though the fields are quite close, it is definitely necessary to have different small-scale and fine-granular fertilization and farming strategies.

5 Conclusion

In this paper we have presented a novel application of self-organizing maps by using them on agricultural yield data. After a thorough description and statistical analysis of the available data sets, we briefly outlined the advantages of self-organizing maps in data visualization. A hypothesis on the differences between two fields could clearly be confirmed by using SOMs. We presented further results, which are very promising and show that correlations and interdependencies in the data sets can easily be assessed by visual inspection of the resulting component planes of the self-organizing map. Those results are of immediate practical usefulness and demonstrate the advantage of using data mining techniques in agriculture.

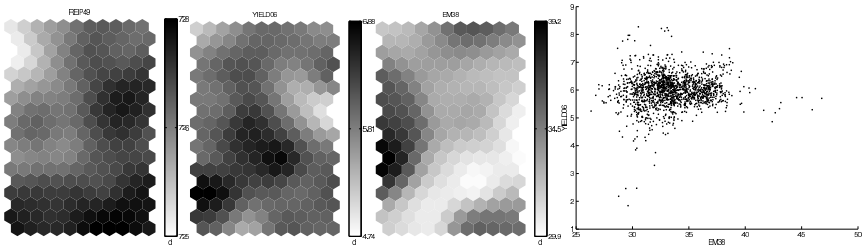
5.1 Future Work

The presented work is part of a larger data mining process. In earlier work, we have presented modeling ideas to represent the agriculture data and use them for prediction and optimization [10]. This work presented ideas on using advanced visualization techniques with the available, real data. Future work will certainly cover further optimization of the prediction capabilities and evaluating different modeling techniques as well as working with additional data such as low-altitude flight



(a) YIELD05 (b) N1 (c) N2 (d) Scatter plot for F330-net, REIP49 vs. YIELD06

Fig. 10: F330-net: YIELD05, N1, N2, scatter plot



(a) REIP49 (b) YIELD06 (c) EM38 (d) Scatter plot for F330-net, EM38 vs. YIELD06

Fig. 11: F330-net: REIP49, YIELD06, EM38, scatter plot

sensors [2]. As of now, those additional sensor data are becoming available for data mining – this will eventually lead to better heterogeneity indicators by refining the available models.

Acknowledgements The figures in this work were generated using Matlab R2007b with the SOM toolbox downloadable from `http://www.cis.hut.fi/projects/somtoolbox/`. The Matlab script that generated the figures can be obtained from the first author on request.

References

1. Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. WEBSOM—self-organizing maps of document collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6*, pages 310–315. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.
2. T. Jensen, A. Apan, F. Young, and L. Zeller. Detecting the attributes of a wheat crop using digital imagery acquired from a low-altitude platform. *Comput. Electron. Agric.*, 59(1-2):66–77, 2007.

3. T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *Neural Networks, IEEE Transactions on*, 11(3):574–585, 2000.
4. Teuvo Kohonen. *Self-Organizing Maps*. Springer, December 2000.
5. J. Liu, J. R. Miller, D. Haboudane, and E. Pattey. Exploring the relationship between red edge parameters and crop variables for precision agriculture. In *2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 2, pages 1276–1279 vol.2, 2004.
6. E. M. Middleton, P. K. E. Campbell, J. E. Mcmurtrey, L. A. Corp, L. M. Butcher, and E. W. Chappelle. "Red edge" optical properties of corn leaves from different nitrogen regimes. In *2002 IEEE International Geoscience and Remote Sensing Symposium*, volume 4, pages 2208–2210 vol.4, 2002.
7. D. Pokrajac and Z. Obradovic. Neural network-based software for fertilizer optimization in precision farming. In *Int. Joint Conf. on Neural Networks 2001*, volume 3, pages 2110–2115, 2001.
8. Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Estimation of neural network parameters for wheat yield prediction. In *Proceedings of the WCC 2008*, Science and Business Media. Springer, 2008. (to appear).
9. Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Optimizing wheat yield prediction using different topologies of neural networks. In José Luis Verdegay, Manuel Ojeda-Aciego, and Luis Magdalena, editors, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-08)*, pages 576–582. University of Málaga, June 2008.
10. Georg Ruß, Rudolf Kruse, Peter Wagner, and Martin Schneider. Data mining with neural networks for wheat yield prediction. In *Advances in Data Mining*. Springer Verlag, 2008. (to appear).
11. M. Schneider and P. Wagner. Prerequisites for the adoption of new technologies - the example of precision agriculture. In *Agricultural Engineering for a Better World*, Düsseldorf, 2006. VDI Verlag GmbH.
12. Michael L. Stein. *Interpolation of Spatial Data : Some Theory for Kriging (Springer Series in Statistics)*. Springer, June 1999.
13. Y. Uno, S. O. Prasher, R. Lacroix, P. K. Goel, Y. Karimi, A. Viau, and R. M. Patel. Artificial neural networks to predict corn yield from compact airborne spectrographic imager data. *Computers and Electronics in Agriculture*, 47(2):149–161, May 2005.
14. J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. Self-organizing map in matlab: the SOM toolbox. In *Proceedings of the Matlab DSP Conference*, pages 35–40, Espoo, Finland, November 1999.
15. Georg Weigert. *Data Mining und Wissensentdeckung im Precision Farming - Entwicklung von ökonomisch optimierten Entscheidungsregeln zur kleinräumigen Stickstoff-Ausbringung*. PhD thesis, TU München, 2006.
16. Michael D. Weiss. Precision farming and spatial economic analysis: Research challenges and opportunities. *American Journal of Agricultural Economics*, 78(5):1275–1280, 1996.