Application of Supervised Self Organising Models for Wheat Yield Prediction

Xanthoula Eirini Pantazi, Dimitrios Moshou, Abdul Mounem Mouazen, Boyan Kuang, Thomas Alexandridis

APPLICATION OF SUPERVISED SELF ORGANISING MODELS FOR WHEAT YIELD PREDICTION

X.E. Pantazi[1], D.Moshou[1], A.M. Mouazen[2], B.Kuang[2], T. Alexandridis[1]

[1]Aristotle University of Thessaloniki, School of Agriculture, Department of Hydraulics, Soil Science and Agriculture Engineering, Laboratory of Agricultural Engineering, P.0 275 54 124, Thessaloniki, Greece
[2]Environmental Technology and Science Department, Cranfield University, Bedfordshire MK43 0AL, United Kingdom.

renepantazi@gmail.com,dmoshou@agro.auth.gr

Abstract. The management of wheat yield behavior in agricultural areas is a Fvery important task because it influences and specifies the wheat yield production. An efficient knowledge-based approach utilizing an efficient Machine Learning algorithm for characterizing wheat yield behavior is presented in this research work. The novelty of the method is based on the use of Supervised Self Organizing Maps to handle existent sensor information by using a supervised learning algorithm so as to assess measurement data and update initial knowledge. The advent of precision farming generates data which, because of their type and complexity, are not efficiently analyzed by traditional methods. The Supervised Self Organizing Maps have been proved from the literature efficient and flexible to analyze sensor information and by using the appropriate learning algorithms can update the initial knowledge. The Self Organizing models that are developed consisted of input nodes representing the main factors in wheat crop production such as biomass indicators, Organic Carbon (OC), pH, Mg, Total N, Ca, Cation Exchange Capacity (CEC), Moisture Content (MC) and the output weights represented the class labels corresponding to the predicted wheat yield.

Keywords: neural networks, data mining, precision agriculture, machine learning

1 Introduction

A large number of approaches, models, algorithms, and statistical tools have been proposed and used for assessing the yield prediction in agriculture. Many authors used simple linear correlations of yield with soil properties but the results varying from

field to field and year to year (Drummond, et al., 1995; Khakural et al., 1999). Many other studies, contain complex linear methods like multiple linear regression, were accomplished with similar results (Drummond et al., 1995; Khakural et al., 1999; Kravchenko & Bullock, 2000). Some authors proposed non-linear statistical methods to investigate the yield response (Adams, et al., 1999; Wendroth, et al,. 1999). Expert systems and artificial intelligent algorithms are a relatively new subset of nonlinear techniques. They have been proposed in agriculture for decision making and decision support tasks. More specifically, expert systems (Plant & Stave, 1991; Rao, 1992) have been developed and applied in different fields in agriculture to give advices and make management decisions. In this context many studies have been reported using artificial intelligence techniques and a few of them focused on the spatial analysis of produced data in precision agriculture. The most of them use artificial neural networks (ANNs) and machine learning algorithms for setting target yields which is one of the problems in PA (Canteri et al., 2002; Liu, et al., 2001; Miao et al., 2006). Schultz, et al. (2000) summarized the advantages of applying neural networks in agro ecological modeling, including the ability of ANN to handle both quantitative and qualitative data, merge information and combine both linear and non-linear responses. Neural networks have been proposed for identifying important factors influencing corn yield and grain quality variability (Miao et al., 2006), for data analysis (Irmak et al., 2006), for prediction crop yield based on soil properties (Drummond et al., 2003), for setting target corn yields (Liu et al., 2001). Shearer et al. (1999) studied a large number of variables, including fertility, satellite imagery, and soil conductivity, for a relatively small number of observations in one site-year of data.

Self-Organizing Maps (SOMs) are one of the most well-known among the several Artificial Neural Networks architectures proposed in literature (Kohonen, 1988). Their applications have increased during the last decade and they have been applied in several different fields and nowadays they are considered as one of the foremost machine learning tools and an important tool for multivariate statistics (Marini, 2009) Self-Organizing Maps (SOMs) are self-organizing systems able to solve problems in an unsupervised way, without needing target data. In order to cover certain needs, unsupervised models have been extended in order to be able to work in a supervised framework. To this end, methods like counterpropagation Artificial Neural Networks (CP-ANNs), which are very similar to SOMs, since an output layer is added to the SOM layer (Zupan et al., 1995), have been introduced. When dealing with classification issues, CP-ANNs are generally efficient methods for achieving class separation in non-linear boundaries. Recent modifications to CP-ANNs have led to the introduction of new supervised neural network architectures and relevant learning algorithms such as Supervised Kohonen Networks (SKNs) and XY-fused Networks (XY-Fs) (Melssen, 2006).

The aim of the work reported is to present a methodology that can determine wheat yield behavior in precision farming, based on Machine Learning techniques and particularly based on aspects related to cluster visualization. In the current paper several Self Organizing Map models using supervised learning approach and algorithm are used to classify precision agriculture data in order to predict the yield productivity. To achieve this, soil physical and chemical parameters have been fused together with biomass indicators.

2 Materials and Methods

2.1 Crop Parameters affecting yield

For the calculation of crop cover the Normalized Different Vegetation Index (NDVI) was used. The NDVI is (NIR-RED)/ (NIR+RED) where NIR is the Near Infrared Radiation (0,725 to 1μm) and RED is the Red Radiation (0, 58 to 0, 68 μm). The NDVI was calculated based on satellite images that were collected two times on the 2nd May and 3rd June in the spring of 2013 (Fig. 1). These images were provided by DMCii (http://dmcii.tumblr.com/) for the TF1 in the Duck End farm in the UK.
Below is the processing workflow chain for crop NDVI, based on satellite imagery:

1. We received L1R (geo-rectified from DMCii) or L1T (ortho-rectified imagery). With the L1R we had to complete the ortho-rectification using software called 'keystone workstation', otherwise the L1T ortho product should be purchased directly from DMCii (the preferred method),
2. We performed in-band reflectance calibration using ArcGIS,
3. A map of NDVI was created with ArcGIS,
4. The NDVI was calculated for 5mX5m grid resulting 16500 values.

The yield was calculated for 5mX5m grid by a combine harvester that able to measure automatically the yield during harvest resulting in 16500 values.
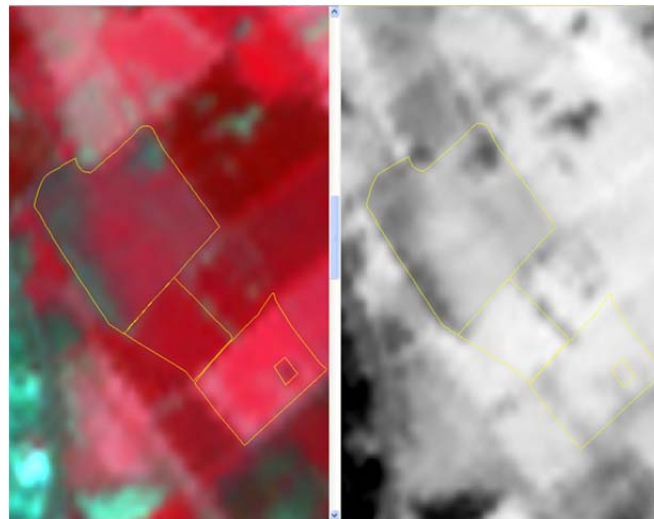
Fig. 1. False colour infrared (left) and NDVI (right) of 3rd Jun 2013 obtained from DMCii UK-DMC2 satellite imagery for TF1 (e.g. Hawnes End field) in the Duck End farm in the UK

2.2 Soil Parameters affecting yield

Precision farming requires development of on-line sensors for real-time measurement of soil properties, because these sensors can lead to reducing labor and time cost of soil sampling and analysis. With the emergence of commercial NIR spectrophotometers and multivariate calibration software packages, the vis–NIR spectroscopy has been adopted much widely for soil analysis. Numerous researchers have extended the vis–NIR spectroscopy applications from the measurement of key soil properties (Moisture Content, pH, Soil Organic Matter, Total N, and soil Organic Carbon) with high accuracy to almost all other micro and macro elements with less accuracy. The analysis of soil with this technique was also extended to soil biological, physical, and engineering properties. Multivariate calibration techniques allowed for simultaneous measurements of several soil properties under consideration. Shibusawa et al. (2001) developed an on-line vis–NIR (400– 1700 nm) sensor to predict Moisture Content, pH, Soil Organic Matter, and NO_3-N. A simpler design to the one of Shibusawa et al. (2001) without sapphire window optical configuration was developed by Mouazen et al. (2005). The system was successfully calibrated for Moisture Content, Soil Organic Matter, Total N, pH, and available P in different soils in Belgium and northern France (Mouazen et al., 2005, 2007, 2009).

On-line measurements of soil were carried out in the selected field of Horn's End. The measurement was carried out by Cranfield University (CU). The raw soil spectra was recorded and stored for time on analysis. Soil samples collected from the field were sent to laboratory analysis with standard methods.

The following soil parameters were estimated from the spectra in a 5mX 5m grid: soil Organic Carbon (OC), pH, Mg, Total N, Ca, Cation Exchange Capacity (CEC) and Moisture Content (MC). This resulted in 16500 vectors of soil parameters for the whole field matching the number of values obtained from the NDVI calculation.

2.3 Counterpropagation Artificial Neural Networks

Counterpropagation Artificial Neural Networks (CP-ANNs) are modeling methods which combine features from both supervised and unsupervised learning (Zupan et al., 1995) CP-ANNs consist of two layers, a Kohonen layer and an output layer, whose neurons have as many weights as the number of classes to be modelled .The class vector is used to define a matrix C, with I rows and G columns, where I is the number of samples and G the total number of classes; each entry cig of C represents the membership of the i-th sample to the g-th class expressed with a binary code (0 or 1). When the sequential training is adopted, the weights of the rth neuron in the output layer (y_r) are updated in a supervised manner on the basis of the winning neuron selected in the Kohonen layer. Considering the class of each sample i, the update is calculated as follows:

$$\Delta y_r = \eta \left(1 - \frac{d_n}{d_{max} + 1} \right) \left(c_i - y_r^{old} \right)$$

(1)

where d_{ri} is the topological distance between the considered neuron r and the winning neuron selected in the Kohonen layer; c_i is the ith row of the unfolded class matrix C,

that is, a G-dimensional binary vector representing the class membership of the ith sample. At the end of the network training, each neuron of the Kohonen layer can be assigned to a class on the basis of the output weights and all the samples placed in that neuron are automatically assigned to the corresponding class.

2.4. XY-fused Networks

XY-fused Networks (XY-Fs) (Melssen et al., 2006) are supervised neural networks for building classification models derived from Self- Organizing Maps (SOMs). In XY-fused Networks, the winning neuron is selected by calculating Euclidean distances between a) sample (xi) and weights of the Kohonen layer, b) class membership vector (ci) and weights of the output layer. These two Euclidean distances are then combined together to form a fused similarity, that is used to find the winning neuron. The influence of distances calculated on the Kohonen layer decreases linearly during the training epochs, while the influence of distances calculated on the output layer increases.

2.5 Supervised Kohonen Networks (SKNs)

As in the case for CP-ANNs and XY-Fs, Supervised Kohonen Networks (SKNs) (Melssen et al., 2006) are supervised neural networks derived from Self-Organizing Maps (SOMs) and used to calculate classification models. In Supervised Kohonen Networks, Kohonen and output layers are glued together to give a combined layer that is updated according to the training scheme of Self-Organizing Maps. Each sample (xi) and its corresponding class vector (ci) are combined together and act as input for the network. In order to achieve classification models with good predictive performances, xi and ci must be scaled properly. Therefore, a scaling coefficient for ci is introduced for tuning the influence of class vector in the model calculation.

3   Results and Discussion.

The values of the eight soil parameters were concatenated with the NDVI values so as to form 16500 feature vectors which correspond to fusion of both soil and crop parameters. Aiming to avoid bias during clustering the fusion vectors were preprocessed so that they had zero mean and start a deviation equal to unity. In order to predict the yield fusion vectors were used as inputs while the yield values were divided in three classes with equal number of samples containing 5500 each in ascending order, thus corresponding to low medium and high yield.
The Supervised Map models Xyf, Skn and Cpann were trained with the fusion vectors as input and the yield classes as output. In order to be able to test the generalization capability of the neural networks cross validation was applied by leaving one out of ten samples randomly so that after training on nine samples the prediction was tested on the tenth. The results of the cross validation are shown in Tables 1, 2, 3 for Xyf, SKn and Cpann. The best overall result is obtained from the Skn network. The Skn gives better results than the other two methods because in the case of Skn the clustering of input layers and output yield is performed using one combined vector and this

reduces the possibility of deviating values of yield from affecting the result of classification. In the other two architectures there are values of yield that can affect the clustering in a negative way since it is possible that other non-measured parameters are capable of affecting the yield. In all cases the best prediction is obtained for the low category of yield which is advantageous since the low yield spots in the field require additional fertilization. The SOM clusters of the components of the training vectors are shown in Figures 2, 3, 4. The first subplot corresponds to NDVI component while the second to ninth correspond to Ca, CEC, MC, Mg, OC, P, pH, TN. The last two correspond to target yield and normalized target yield. From this can be seen that the NDVI is the first component and shows partial correlation to the yield map which is shown in the tenth subplot. The other components correspond to soil parameters and provide complementary information to the yield. The Component Maps are useful in interpreting the correlations between the soil factors and NDVI as related to the yield. In the case where tendency of the soil factors is similar to the yield this factor is important for higher yield while in the opposite case, a high value in a soil factor could limit the yield. Generally the NDVI has a positive correlation to the yield.

**Table 1.** Results of cross validation for Xyf Network

Real Network Estimation	Network Prediction (%)		
	low	medium	high
low	95.62	3.89	0.49
medium	5.45	89.95	4.6
high	5.35	11.29	83.36

Table 2. Results of cross validation for Skn Network

Real Network Estimation	Network Prediction (%)		
	low	medium	high
low	93.11	5.69	1.20
medium	4.09	89.05	6.85
high	2.33	6.64	91.04

Table 3. Results of cross validation for Cpann Network

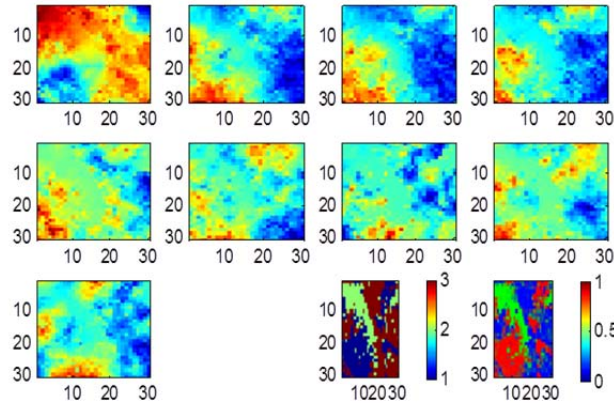| Real Network Estimation	Network Prediction (%)		
	low	medium	high
low	92.40	4.36	3.24
medium	8.07	81.22	10.71
high	4.18	6.02	89.80
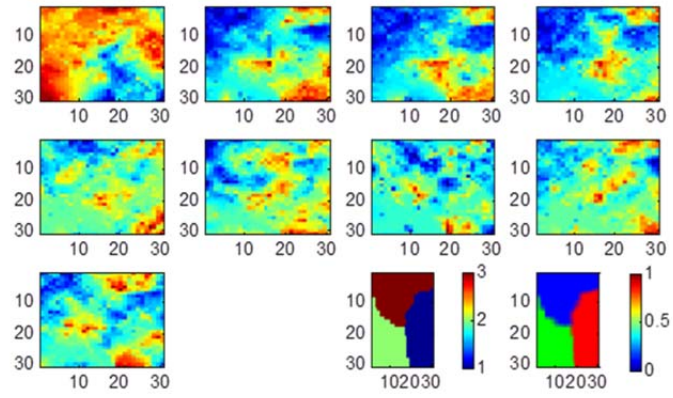
**Fig. 2.** Skn Component Maps for NDVI and soil parameters
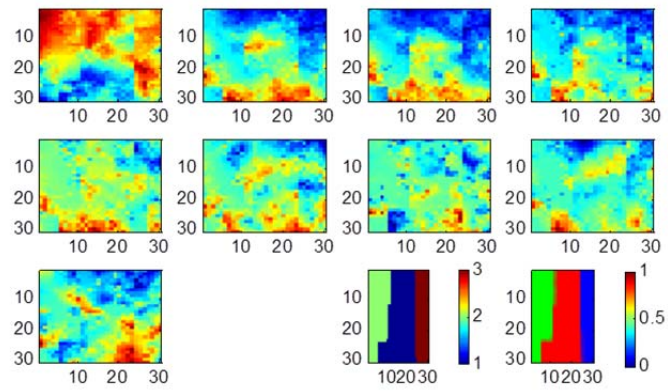
**Fig. 3.** Xyf Component Maps for NDVI and soil parameters

Fig. 4. Cpann Component Maps for NDVI and Soil parameters

The target yield is shown in Fig.5. in three classes labeled as red for high yield, blue for low yield and yellow for medium yield. By looking at Fig.6 one can see that there is a high correlation between the target yield and the predicted yield. The NDVI Shown in Fig.7 demonstrates partial correlation to the yield in the left part of the field, thus explaining the cluster distribution of the Neural Network models. This discrepancy in the right part of the field can be explained by the distribution of the other nutrients which counterbalance the effect on yield. In Figure 8, the spatial variation of total N is shown divided in three classes corresponding to different levels of total N.
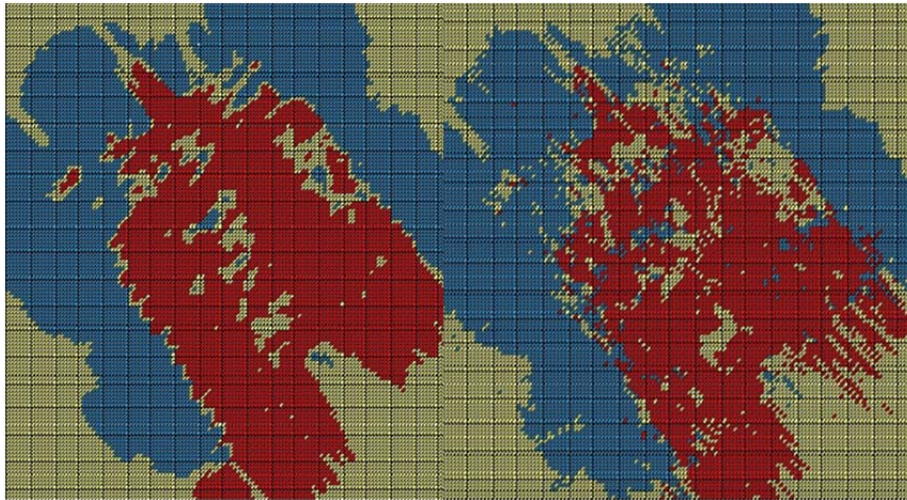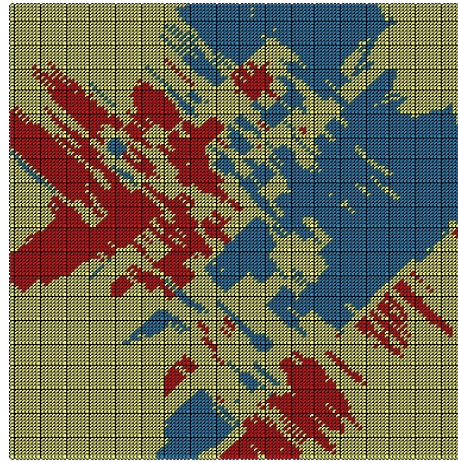

Fig. 5. Target Yield in three classes produced by ArcGIS

Fig. 6. Prediction in three classes produced by ArcGIS
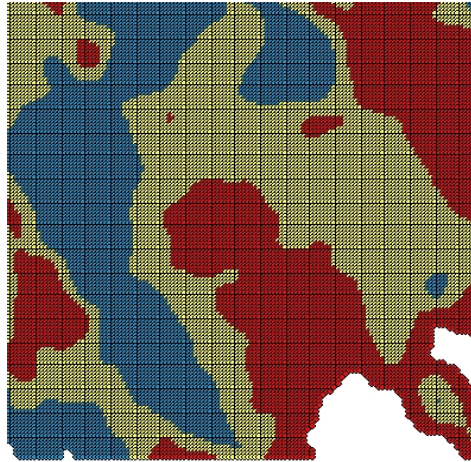
Fig. 7. NDVI Satellite based in three classes produced by ArcGIS

Fig. 8. Total Nitrogen in three classes produced by ArcGIS

Conclusions.

In this paper, three Self Organizing models were developed that consisted of input nodes representing the main factors in wheat crop production such as biomass indicators, organic carbon (OC), pH, Mg, total N, Ca, Cation Exchange Capacity (CEC), moisture content (MC) and the output weights represented the class labels corresponding to the predicted wheat yield. The results indicate that yield prediction is possible with a very high accuracy reaching 93% and can be used in order to improve fertilizer administration by using the yield prediction models that were presented. Visualization of component maps could reveal extra information that is useful in order to interpret the relations between the soil and crop parameters and the yield.

Acknowledgement

References

1. Adams, M. L., Cook, S. E., Caccetta, P. A., & Pringle, M. J. (1999). Machine learning methods in site-specific management research: An Australian case study. In Proceedings of the fourth international conference on precision agriculture, Madison, USA, pp. 1321–1333.
2. Canteri, M., Avila, B. C., Dos Santos, E. L., Sanches, M. K., Kovaleschyn, D., Molin, J. P., et al. (2002). Application of data mining in automatic description of yield behavior in

agricultural areas. Proceedings of the World Congress of Computers in Agriculture and Natural Resources, 183–189.

3. Drummond, S. T., Sudduth, K. A., & Birrell, S. J. (1995). Analysis and correlation methods for spatial data. ASAE Paper No. 95–1335, St. Joseph, Michigan, ASAE.

4. Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J., & Kitchen, N. R. (2003).Statistical and neural methods for site-specific yield prediction. Transactions of the ASAE, 46(1), 5–14.

5. Irmak, A., Jones, J. W., Batchelor, W. D., Irmak, S., Boote, K. J., & Paz, J. O. (2006). Artificial neural network model as a data analysis tool in precision farming. Transactions of the ASABE, 49(6), 2027–2037.

6. Khakural, B. R., Robert, P. C., & Huggins, D. R. (1999). Variability of corn/soybean yield and soil/landscape properties across a southwestern Minnesota landscape. In Proceedings of the fourth international conference on precision agriculture, pp. 573–579.

7. Kohonen T. (1988). Self-Organization and Associative Memory, Springer Verlag, Berlin.

8. Kravchenko, A. N., & Bullock, D. G. (2000). Correlation of corn and soybean grain yield with topography and soil properties. Agronomy Journal, 92(1), 75–83.

9. Liu, J., Goering, C. E., & Tian, L. (2001). A neural network for setting target corn yields. Transactions of the ASAE, 44(3), 705–713.

10. Marini F. (2009). Artificial neural networks in food analysis: trends and perspectives Analytica Chimica Acta 635, 121–131.

11. Miao, Y., & Liu, Z. (2000). On causal inference in fuzzy cognitive maps. IEEE 1034 Transactions on Fuzzy Systems, 8, 107–119

12. Melssen W., Wehrens R., Buydens L. (2006) .Supervised Kohonen networks for classification problems Chemometrics and Intelligent Laboratory Systems 83, 99–113.

13. Mouazen, A. M., De Baerdemaeker, J., and Ramon, H. (2005). Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. Soil Till. Res. 80, 171–183.

14. Mouazen, A. M., Maleki, M. R., De Baerdemaeker, J., and Ramon, H. (2007). On-line measurement of some selected soil properties using a VIS-NIR sensor. Soil Till. Res. 93, 13–27.

15. Mouazen, A. M. (2009). The future of for on-line measurement of soil properties with sensor fusion. Landwards. Nelson Publishing, Natural Resources Department, Cranfield University, MK43 OAL, United Kingdom, 14–16.

16. Shearer, S. A., Thomasson, J. A., Mueller, T. G., Fulton, J. P., Higgins, S. F., & Samson, S. (1999). Yield prediction using a neural network classifier trained using soil landscape features and soil fertility data. ASAE Paper No. 993042. St. Joseph, Michigan, USA.

17. Plant, R., & Stave, N. (1991). Knowledge based systems in agriculture. USA: McGraw-Hill.

18. Rao,J.P.(1992).Expert systems in agriculture,<http://www.manage.gov.in/managelib/faculty/PanduRanga.htm>.

19. Shibusawa, S., Made Anom, S. W., Sato, H. P., and Sasao, A. (2001). Soil mapping using the real-time soil spectrometer. In "ECPA 2001" (G. Gerenier and S. Blackmore, Eds.), agro Montpellier, 2, pp. 485–490. Montpellier, France.

20. Zupan J., Novic M., Gasteiger J. (1995). Neural networks with counter-propagation learning strategy used for modelling Chemometrics and Intelligent Laboratory Systems 27, 175–187.

21. Wendroth, O., Jurschik, P., & Nielsen, D. R. (1999). Spatial crop yield prediction from soil and land surface state variables using an autoregressive state- space approach. Precision Agriculture'99 (pp. 419–428). Sheffield, U.K: J.V.Stafford, ed.