Computers and Electronics in Agriculture 121 (2016) 57–65

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Wheat yield prediction using machine learning and advanced sensing techniques

X.E. Pantazi [a,*], D. Moshou [a], T. Alexandridis [b], R.L. Whetton [c], A.M. Mouazen [c]

[a] *Laboratory of Agricultural Engineering, Aristotle University of Thessaloniki, Faculty of Agriculture, Univ. Box 275, Thessaloniki 54124, Greece*
[b] *Laboratory of Remote Sensing and GIS, Aristotle University of Thessaloniki, Faculty of Agriculture, Univ. Box 259, Thessaloniki 54124, Greece*
[c] *Cranfield Soil and AgriFood Institute, Cranfield University, Bedfordshire MK43 0AL, United Kingdom*

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form 1 October 2015
Accepted 9 November 2015
Available online 21 December 2015

Keywords:
Artificial neural networks
Data fusion
Satellite imagery
Soil spectroscopy
NDVI
Precision farming

ABSTRACT

Understanding yield limiting factors requires high resolution multi-layer information about factors affecting crop growth and yield. Therefore, on-line proximal soil sensing for estimation of soil properties is required, due to the ability of these sensors to collect high resolution data (>1500 sample per ha), and subsequently reducing labor and time cost of soil sampling and analysis. The aim of this paper is to predict within field variation in wheat yield, based on on-line multi-layer soil data, and satellite imagery crop growth characteristics. Supervised self-organizing maps capable of handling existent information from different soil and crop sensors by utilizing an unsupervised learning algorithm were used. The performance of counter-propagation artificial neural networks (CP-ANNs), XY-fused Networks (XY-Fs) and Supervised Kohonen Networks (SKNs) for predicting wheat yield in a 22 ha field in Bedfordshire, UK were compared for a single cropping season. The self organizing models consisted of input nodes corresponded to feature vectors formed from normalized values of on-line predicted soil parameters and the satellite normalized difference vegetation index (NDVI). The output nodes consisted of yield isofrequency classes, which were predicted from the three trained networks. Results showed that cross validation based yield prediction of the SKN model for the low yield class exceeded 91% which can be considered as highly accurate given the complex relationship between limiting factors and the yield. The medium and high yield class reached 70% and 83% respectively. The average overall accuracy for SKN was 81.65%, for CP-ANN 78.3% and for XY-F 80.92%, showing that the SKN model had the best overall performance.

1. Introduction

Yield prediction in precision farming, is considered of high importance for the improvement of crop management and fruit marketing planning. Once the yield is site-specifically predicted, the farm inputs such as fertilizers could be applied variably according to the expected crop and soil needs. A variety of approaches, models and algorithms, have been presented and used to enable yield prediction in agriculture. Simple linear correlations of yield with soil properties have been proposed based on limited number of soil samples. However, the results of prediction were variable spatially and temporally (Drummond et al., 1995; Khakural et al., 1999). Ayoubi et al. (2009) used factor analysis to quantify the relationship of several soil properties with grain yield. Numerous other studies using complicated linear methods such as multiple linear regression analyses showed similar outcomes (Kravchenko and Bullock, 2000). Computational intelligence and expert systems are considered as quite new subdivision of nonlinear algorithms. They have been recommended in agriculture to aid decision support.

In particular, expert systems (Rao, 1992) have been established and applied for several agricultural purposes related to advisory and management services. In this field, many researches have introduced the use of computational intelligence algorithms. Schultz et al. (2000) presented the benefits of neural networks application in agro ecological case studies to manage simultaneously quantitative and qualitative data, combine information and handle both linear and non-linear responses. Besalatpour et al. (2012) used Artificial Neural Networks (ANN) trained on several soil physical properties in order to predict soil shear strength. Some researchers have focused on the spatial behavior of the data within precision agriculture context. Literature showed that the majority of research has utilized the ANNs and other computational intelligence techniques for predicting target yields which

[*] Corresponding author.
E-mail addresses: renepantazi@gmail.com (X.E. Pantazi), dmoshou@agro.auth.gr (D. Moshou).

http://dx.doi.org/10.1016/j.compag.2015.11.018
0168-1699/© 2015 Elsevier B.V. All rights reserved.

constitutes a major issue in precision agriculture (Miao et al., 2006). ANNs, as non-linear modeling techniques, have also been applied to understand and quantify yield response to soil variables (Effendi et al., 2010; Fortin et al., 2010). In ANNs, the observed dataset of the selected variables is fitted aiming at providing a better picture of the problem as follows: the weights of linkages connecting input and output variables are adjusted and are also utilized as multivariate non-linear tools for further analysis. Neural networks have been suggested for finding important factors that are considered responsible for corn yield and grain quality variation (Miao et al., 2006), for data mining (Irmak et al., 2006), for crop yield prediction by using soil properties (Drummond et al., 2003), and for determining target corn yields (Liu et al., 2001). Ayoubi and Sahrawat (2011) and Norouzi et al. (2010) used ANN to predict grain yield as a function of soil properties which were collected/analyzed with traditional lab methods. Zolfaghari et al. (2015) developed ANN models that could explain a majority of the variability (62–94%) in Atterberg limits and indices. Shearer et al. (1999) investigated a vast number of variables such as soil and crop parameters based on satellite imagery for a few observations during the year from one site. In their paper, soil fertility, elevation and electrical conductivity were used together with spectral satellite image features in order to predict corn yield. This model failed to predict the spatial variation in the yield. Only fertility and conductivity features were closely correlated to yield. The ANNs can be combined with other artificial intelligence methods or other statistical techniques so as to guarantee the advantages of ANN modeling, and to avoid some of the limitations they are often imposed, such as the necessity of large amounts of data that are needed to be trained. Taking into account the various ANNs architectures provided by the literature, self-organizing maps (SOMs) are regarded as one of the most eminent, whose implementations in various fields experienced a large increase during the last decade (Kohonen, 1988). They belong to machine learning tools of high importance, especially the most suitable for solving multivariate statistic problems (Marini, 2009). They are also capable of providing solutions by operating in a non-supervised mode, which is based on data clustering.

So far the use of some of the non-linear methods described above for yield prediction was limited and based on traditional soil sampling (e.g. 1 sample per ha) and laboratory analyses that is tedious, time consuming and expensive. No attempt was taken to utilize high sampling resolution data collected with an on-line soil sensor (e.g. Mouazen, 2006), although the sensor was proved to successfully measure key soil properties affecting crop yield with different degree of accuracy. These include total nitrogen (TN), organic carbon (OC), moisture content (MC), phosphorous (P), pH, calcium (Ca), magnesium (Mg), clay (CC), cation exchange capacity (CEC), soil organic matter content (SOMC) and plasticity index (PI) (Mouazen et al., 2009; Marin-González et al., 2013; Kuang and Mouazen, 2013; Mouazen et al., 2014). Furthermore, none of the previous work attempted to fuse high resolution data on key soil properties with crop growth indicated as NDVI to predict crop yield in arable crops. The main gap in knowledge concerns the absence of a unification framework under which individual factors affecting yield to a certain extent convey complementary information so the combination of all these factors in an integrated model can provide more accurate prediction. This unification framework is data fusion of the various layers of information. This need calls for a flexible modeling technique for data fusion, which can model the non-lineal relationship between soil parameters, biomass and yield. Previous work has been hindered by the low accuracy offered by the linear models and the lack of fusion of high resolution data on key soil properties with remotely sensed crop growth to predict crop yield in arable crops. The aim of the present work is to overcome the limitations of the above mentioned non-linear modeling approaches for the prediction of crop yield by introducing an original algorithm, based on an extension of SOMs with supervised learning. The proposed modeling techniques allow the integration of high sampling resolution of multi-layer data on soil and crop by establishing a data fusion model capable of predicting the spatial distribution of wheat yield, with high accuracy compared to the current techniques. These techniques provide an innovative way of visualizing correlations between soil, crop parameters and yield.

2. Materials and methods

In the current research, three Self Organizing Map models, namely, Counter-propagation Artificial Neural Network (CPANN), Supervised Kohonen Network (SKN) and XY-fusion network (XYF), based on Supervised Learning to associate precision agriculture data with isofrequency classes of yield productivity, were utilized. For the implementation of this approach, physicochemical soil parameters were gathered by means of an on-line visible and near infrared (vis–NIR) spectroscopy sensor, which was subsequently combined with biomass indicators, following a sensor fusion approach.

2.1. Experimental site

The study site was a 22 ha Horn End Field at Duck End Farm, Wilstead, Bedfordshire, U.K. (Latitude 52°05′51″N, Longitude 0°27′19″W) (Fig. 1). The soil type was defined as “Haplic Luvisols” according to the FAO soil classification system. The textures of selected soil samples according to the United State Department of Agriculture (USDA) indicated the presence of clay, clay loam, sandy clay loam and loam textures. The terrain has a gentle slope of 2% with an elevation that varies between 30 and 38 m, determined by differential global positioning system (DGPS) (EZ-Guide 250, Trimble, USA). The study took place over 2013 cropping season with winter wheat crop.

2.2. Crop parameters affecting yield

In order to estimate crop performance characteristics, crop and yield parameters were utilized. The NDVI was calculated from satellite data acquired by the UK-DMC-2 of the Disaster Monitoring Constellation for International Imaging (DMC_{ii}) on May 2nd and June 3rd, 2013. The second NDVI measurement was collected due to low quality of the first measurement. The images acquired by UK-DMC-2 are multispectral (green, red, near-infrared bands) at 22 m spatial resolution, and 14 bit radiometric resolution.

The image pre-processing and analysis involved ortho-rectification, in-band reflectance calibration, and NDVI calculation using the following formula (Rouse et al., 1974):

$$\text{NDVI} = (\text{NIR} - \text{R})/(\text{NIR} + \text{R}) \tag{1}$$

where NIR and R is the is reflectance in the near-infrared and red bands, respectively. The NDVI layer was resampled to match the 5 m × 5 m grid of the other data layers (e.g. soil layers discussed above) using bilinear interpolation, consequently resulting in 8798 values.

Yield data were gathered with a New Holland CX8070 combine harvester, which was equipped with a yield sensor. The data collection was performed during August of 2013. A harvesting methodology for the field was devised which maximized the accuracy of the yield measurements. The aim was to (I) record wheat yield when the machine header of a width of 7.35 m was full for the full length of the study area, (II) avoid the bare soil in the tramlines.

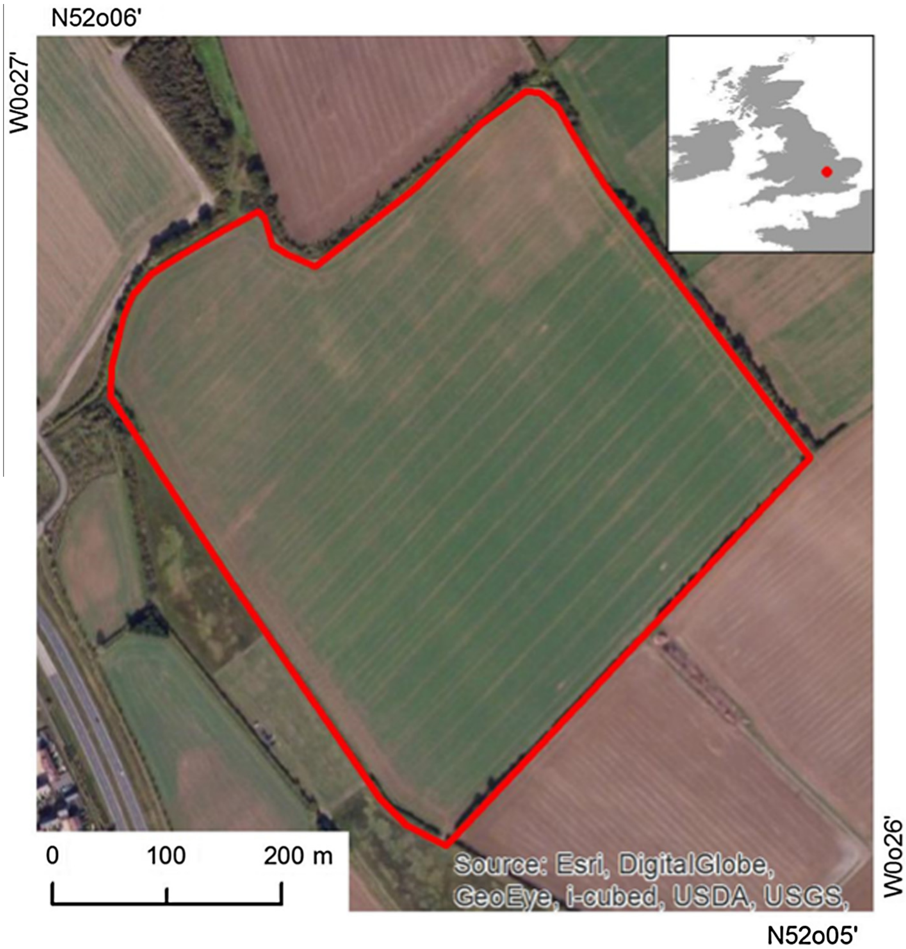

Fig. 1. Horns End field outlined on a web mapping service.

This enabled accurate calculations of yield per harvested area. The yield was interpolated at the same 5 × 5 m grid as the NDVI, resulting in 8798 values.

2.3. Soil parameters affecting yield

Precision farming needs development of on-line sensors so as to become capable of measuring various soil properties, due to the fact that these sensors are able of reducing the cost of soil sampling and analysis in terms of time and labor. In addition, they provide high sample resolution of multi-soil properties. Taking into account the development of vis–NIR spectrophotometers that are launched into the market and chemometrics software packages, the application of vis–NIR spectroscopy has been espoused much broadly for soil analysis. Several researchers have widened the vis–NIR spectroscopy applications so as to enable the key soil property measurements (MC, pH, soil organic matter content (SOMC), TN, and OC) with higher accuracy than those reported for micro elements (Viscarra Rossela et al., 2005; Mouazen et al., 2007). Several calibration practices enabled parallel measurements of various soil properties under consideration. An on-line vis–NIR (400–1700 nm) sensor has been developed for the prediction of MC, pH, SOMC, and NO_3-N (Shibusawa et al., 2001). A more plain design to the one that is above presented was developed by Mouazen (2006), which consisted of no sapphire window optical configuration. The sensor consists of a subsoiler, able to penetrate the soil to a specific depth forming a trench. The bottom of the trench is smoothened by the vertical forces which act on the subsoiler. The optical probe is installed in a steel lens holder which is mounted on the back of the subsoiler chisel so as to register soil spectral reflectance data acquired from the trench smooth bottom.

The subsoiler included a frame in which the optical unit was retrofitted. This system was effectively tested successfully for the measurement of pH, MC, TN, TC, Mg, Ca, cation exchange capacity (CEC) and available P in different soils in Europe (Mouazen et al., 2009, 2007; Kuang and Mouazen, 2013; Marin-González et al., 2013). Online measurements of soil were conducted in the Horn's End field after crop harvest in summer 2013 (27th of August 2013). As some soil properties are dynamic (e.g. MC), their absolute values would change with time. However, the spatial distribution of their variability is expected to remain unchanged, which justifies the inclusion of MC in the analysis.

Measurement was carried out in parallel transects at an average speed of 1.5–2 km h^{-1}. A constant gap of 20 m was kept between neighboring transects. AgroSpec fiber type, mobile, vis–NIR spectrophotometer (Tec5 Technology for Spectroscopy, Germany) with a measurement range of 305–2200 nm was utilized for soil spectra measurement in diffuse reflectance mode. Further information about the sensor configuration can be obtained from Kuang and Mouazen (2013). A total of 400–500 measurements were registered per ha at a depth of 15 cm. The raw soil spectra data was recorded and kept for time analysis. During the online measurement soil samples were collected for the evaluation of the measurement accuracy of selected soil properties. A total of 60 soil samples were collected from the bottom of the survey trench opened by the subsoiler at a depth of 15 cm. Soil samples obtained from the field were dispatched to laboratory so as to be subjected to standard methods of laboratory analysis. Predicted values of TN, OC, MC, P, pH, CEC, Ca and Mg were obtained using partial least squares regression (PLSR), as described in more details by Kuang and Mouazen (2013). The Root Mean Square Error of Cross Validation (RMSECV) values were for TN 0.026 (%), for OC 0.26 (%), for MC 1.92 (%), for P 0.55 (mg 100^{-1} ml^{-1}), for pH 0.4, for CEC (cmolc kg^{-1}) 1.77, for Ca (cmolc kg^{-1}) 22.05 and for Mg (cmolc kg^{-1}) 0.38.

The acquisition points from the on-line soil sensor needed a method of interpolation, to give a continuous data set for all the locations. Krigging was opted because it is a non-biased approach for predicting the values of parameters between the sample points. The predicted values, based on 5 m by 5 m grid were extracted, leading to the same number of points of 8798 as those of the NDVI and crop yield data points.

2.4. Hierarchical self-organizing artificial neural networks

For the purpose of covering specific needs, unsupervised models have been extended so as to be capable of working in a supervised manner. Methods like CP-ANNs have been introduced, which are very similar to SOMs due to the fact that an output layer is appended to the SOMs input layer (Zupan et al., 1995). In the case of classification tasks, CP-ANNs are regarded as efficient methods for attaining non-linear class limits separation. Modifications that have driven to CP-ANNs recently, have introduced novel supervised neural network models and associated learning algorithms, namely, Supervised Kohonen Networks (SKNs) and XY-fused Networks (XY-Fs) (illustrated in Melssen et al., 2006). These three networks were implemented in the current work using the Matlab software platform (MathWorks, Natick, Massachusetts). A detailed description of the networks is provided below.

2.4.1. Counter-propagation artificial neural networks

CP-ANNs are regarded as modeling methods, capable of combining features from not only supervised but also unsupervised learning techniques (Zupan et al., 1995). CP-ANNs comprise of two layers, namely, a Kohonen and an output layer. Every neuron of both layers consists of an equal number of weights to the number of classes that have to be modeled. The class vector is utilized so as to define a matrix C, which comprises of I rows and G columns, where I represents the amount of samples and G the total amount of classes. The membership of the ith sample to the gth class is expressed with a binary code (0 or 1) and is represented by each entry c_{ig} of C. At the time the sequential training is employed, the rth neuron weights in the output layer (y_r) are updated in a supervised way based on the winning neuron in the Kohonen layer. Taking into account the class of each sample i, the increment. Δy_r is estimated as follows (Melssen et al., 2006):

$$\Delta y_r = \eta\left(1 - \frac{d_{ri}}{d_{\max} + 1}\right)(c_i - y_r^{old}) \tag{2}$$

where d_{ri} represents the topological distance separating neuron r and the best matching neuron selected in the Kohonen layer; c_i represents the ith row of the unfolded matrix C formed from classes, which is, a binary G-dimensional vector, representing class membership corresponding to the ith sample. At the time the network is about to be trained completely, each neuron, belonging to the Kohonen layer can be allocated to a class on the basis of the output weights while all the samples which are placed in that neuron are allocated to the corresponding class in an automatic way.

2.4.2. XY-fused networks (XY-F)

XY-fused Networks (XY-Fs) (Melssen et al., 2006) are regarded as supervised neural networks capable of forming classification models that are resulting from SOMs. In these networks, the winning neuron is determined by estimating the Euclidean distances between (a) the data sample (x_i) and the weights of the Kohonen layer, (b) the class membership vector (c_i) and the weights of the output layer. Then, these two Euclidean distances are joined together to create a fused similarity, which is utilized for the winning neuron indication.

2.4.3. Supervised Kohonen networks (SKNs)

Similarly to CP-ANNs and XY-Fs, the Supervised Kohonen Networks (SKNs) models (Melssen et al., 2006) are considered as supervised neural networks, derived from SOMs and are utilized for classification models estimation. In the case of SKNs, the Kohonen and output layers are joined together to bring up a combined layer that is updated taking into account the training regime of SOMs. Every sample (x_i) and its related class vector (c_i) are associated and form an input for the network. The x_i and c_i must be scaled properly for constructing classification models with good predictive performances. Hence, a scaling coefficient for c_i is proposed aiming at tuning the class vector influence in the model calculation.

2.5. Prediction of crop yield

The values of the eight soil parameters collected in 2013 with the on-line soil sensor (pH, MC, TN, TC, Mg, Ca, CEC and available P) were concatenated with the 2013 satellite imagery calculated NDVI values and historic yield data from two years (2011–2012) in order to form 8798 feature vectors of dimension equal to 11 to predict crop yield classes. This procedure concerned a single cropping season. In order to avoid bias during training, the fusion vectors were subjected to preprocessing so that their mean is equal to zero and standard deviation equal to unity. The preprocessing was achieved by subtracting the mean vector and by dividing by the standard deviation of all the samples that belong to the training set. For the yield prediction, fusion vectors were utilized as input for the three ANNs. The yield values were divided in three classes with equal number of samples containing 2933 each in ascending order, thus corresponding to low (0.65–3.86 t/ha), medium (3.86–5.63 t/ha) and high yield (5.63–8.68 t/ha).

3. Results and discussion

3.1. Accuracy of yield prediction with supervised models

The supervised map models XYF, SKN and CP-ANN were trained with the 8798 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 25% of all samples randomly so that after training on the training set (75% of samples), the prediction was tested on this prediction set (25% of samples). Independent validation was performed by leaving 1000 samples for testing, and training with the remaining 7798 samples. The statistics of the calibration and validation datasets that were used for the independent validation are shown in Table 1. Results of the cross validation are shown in Table 2 for XYF, SKN and CP-ANN, respectively, whereas results of the independent validation are shown in Table 3. The difference between cross-validation and independent validation is that the cross-validation assesses overall model capability by rotating training and testing sets but does not result in a usable final model, while independent validation actually produces a usable model that can be used for future

Table 1
Sample statistics of the parameter vectors of the calibration and independent validation sets for yield prediction in the whole field.

Parameter vectors	Calibration set (7798 vectors)					Validation set (1000 vectors)				
	Min	Max	Mean	SD	CV	Min	Max	Mean	SD	CV
NDVI	0.413	0.721	0.620	0.0702	0.113	0.423	0.720	0.619	0.072	0.116
Ca	6.757	57.854	26.317	10.106	0.384	7.690	55.282	26.324	10.286	0.391
CEC	9.514	20.893	13.702	2.229	0.163	9.497	20.642	13.716	2.307	0.168
MC	11.669	24.611	16.617	2.269	0.137	11.695	24.556	16.615	2.285	0.138
Mg	0.195	2.047	1.062	0.270	0.254	0.242	2.078	1.059	0.279	0.263
OC	1.250	2.284	1.753	0.170	0.097	1.323	2.214	1.750	0.169	0.097
P	1.170	2.860	1.951	0.207	0.106	1.341	2.729	1.954	0.207	0.106
pH	4.304	8.159	6.031	0.567	0.094	4.604	7.939	6.027	0.539	0.089
TN	0.158	0.364	0.2450	0.034	0.139	0.168	0.344	0.248	0.034	0.137
Y2011	3.611	7.650	5.437	0.746	0.137	3.748	7.562	5.425	0.725	0.134
Y2012	0.224	9.410	.4.346	1.084	0.249	0.565	8.626	4.353	1.113	0.256

Table 2
Results of cross validation for supervised Kohonen networks (SKN), counter-propagation artificial networks (CP-ANN) and XY-fusion (XY-F) networks having 30 by 30 neurons for the prediction of wheat yield, based on normalized difference vegetation index (NDVI), on-line measured soil variables and historic yield data. Validation was made on four alternating prediction sets (25% of samples), whereas the network has been trained on the training set (75% of samples). The numbers shown are numbers of samples of each predicted yield category, normalized as a percentage of the number of samples in the actual yield maps.

Actual yield isofrequency class	Network prediction (%)		
	Low	Medium	High
SKN			
Low	91.3	7.23	1.4
Medium	7.84	70.54	21.61
High	1.26	15.62	83.12
CP-ANN			
Low	91.48	7.43	1.09
Medium	10.19	68.56	21.24
High	1.87	23.26	74.86
XY-F			
Low	92.15	7.09	0.75
Medium	8.9	72.48	18.62
High	1.29	20.56	78.14

Table 3
Results of independent validation for supervised Kohonen networks (SKN), counter-propagation artificial networks (CP-ANN) and XY-fusion (XY-F) Networks having 30 by 30 neurons for the prediction of wheat yield, based on normalized difference vegetation index (NDVI), on-line measured soil variables and historic yield data. Validation was made on 1000 samples, whereas the network has been trained on 7798 samples. The numbers shown are numbers of samples of each predicted yield category, normalized as a percentage of the number of samples in the actual yield maps.

Actual yield Isofrequency Class	Network Prediction (%)		
	Low	Medium	High
SKN			
Low	91.3	6.96	1.74
Medium	10.87	64.35	24.78
High	1.54	16.98	81.48
CP-ANN			
Low	90.09	9.29	0.62
Medium	9.57	69.86	20.58
High	2.11	24.40	73.49
XY-F			
Low	87.91	11.21	0.89
Medium	5.76	85.15	9.09
High	2.11	38.67	59.21

prediction. In Tables 2 and 3 the network prediction capability in % expresses the number of samples that the network predicts as correctly belonging to the actual yield isofrequency class. In the first column low–medium–high refers to the actual yield isofrequency class. So, correct predictions are situated in the diagonal while off-diagonal elements refer to misclassifications as to belong to another class.

The best overall results for the prediction of wheat yield in cross validation and independent validation were obtained from the SKN networks for the prediction of the low yield category. The accuracy of prediction reached 91.3% for both cross validation and independent validation (Tables 2 and 3). The average overall accuracy of cross-validation for SKN was 81.65%, for CP-ANN 78.3% and for XY-F 80.92%, showing that the SKN model had the best overall performance. In the case of cross-validation CP-ANN obtained 91.48% for the low yield category while it obtained 90.09% for the independent validation. Regarding the medium yield class, the prediction accuracy was lower than that of the low yield category, varying between 70.54%, obtained with SKN networks in cross validation to 85.15% obtained with XY-F in independent validation. The yield prediction of the high yield category was best obtained with SKN networks, with classification accuracy of 83.12% in cross validation and 81.48% in independent validation. Among all modeling cases, the best prediction is obtained for the low category of yield. The presented results agree with the work of other researchers like Ayoubi and Sahrawat (2011) where ANN models could explain 93% and 89% of the total variability in barley biomass and grain yield, respectively by using 14 soil variables. In Norouzi et al. (2010) results indicated that the ANN models could explain 89–95% of the total variability in wheat biomass, grain yield, and grain protein content. More specifically the yield could be predicted with an R^2 of 0.93. Also, the predictability of wheat yield and quality could be further improved by considering management practices followed during the growing season. The accurate yield as reported above can serve as a tool for establishing fertiliser management decisions. Working in the same experimental field as that of the current study (e.g. Horns End field), Halcro et al. (2013) suggested adding the largest amount of nitrogen fertilizer in the poor fertility zones of the field. However, it was shown by these authors that it is not necessarily correct to assume that poor zones of the field will result in the lowest yield, as soil physical properties, particularly, high water content levels and poor drainage play an important role in crop establishment, crop growth and accordingly yield. Therefore, the nutrient rich parts of a field with water logging problems may not result in plausible crop growth and yield.

3.2. Yield maps

The comparison between measured and predicted yield maps is shown in Fig. 2 where SKN has been used for prediction since it provided the best accuracy in prediction as explained in Section 3.1. The predicted yield is classified into three classes labeled as red for high yield (5.63–8.68 t/ha), blue for low yield (0.65–3.86 t/ha) and yellow for medium yield (3.86–5.63 t/ha) based

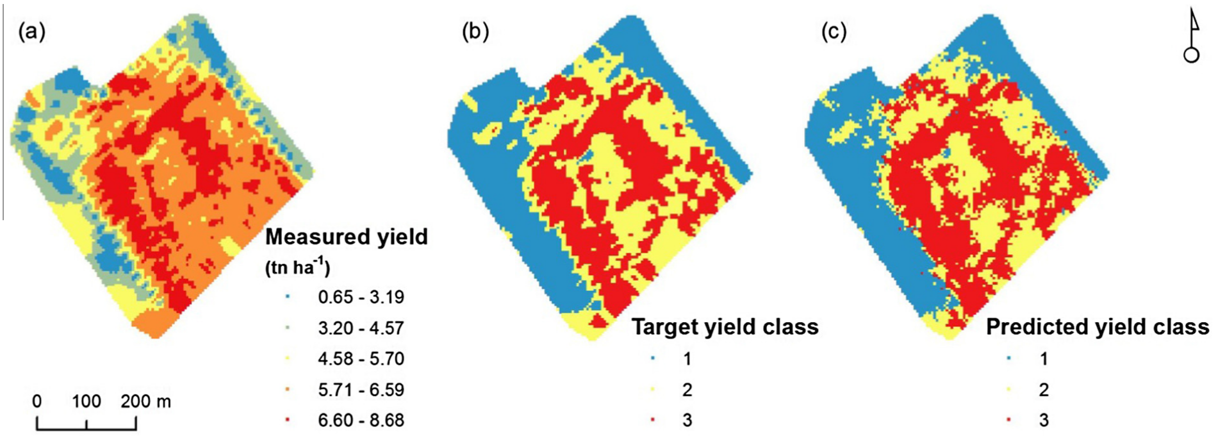

Fig. 2. Measured wheat yield map of 2013 (a) in Horns End field and the target yield map developed by dividing the measured yield data into three iso-frequency classes (b), as compared to the predicted yield map in three classes, resulted from the supervised Kohonen networks (SKN) (c).

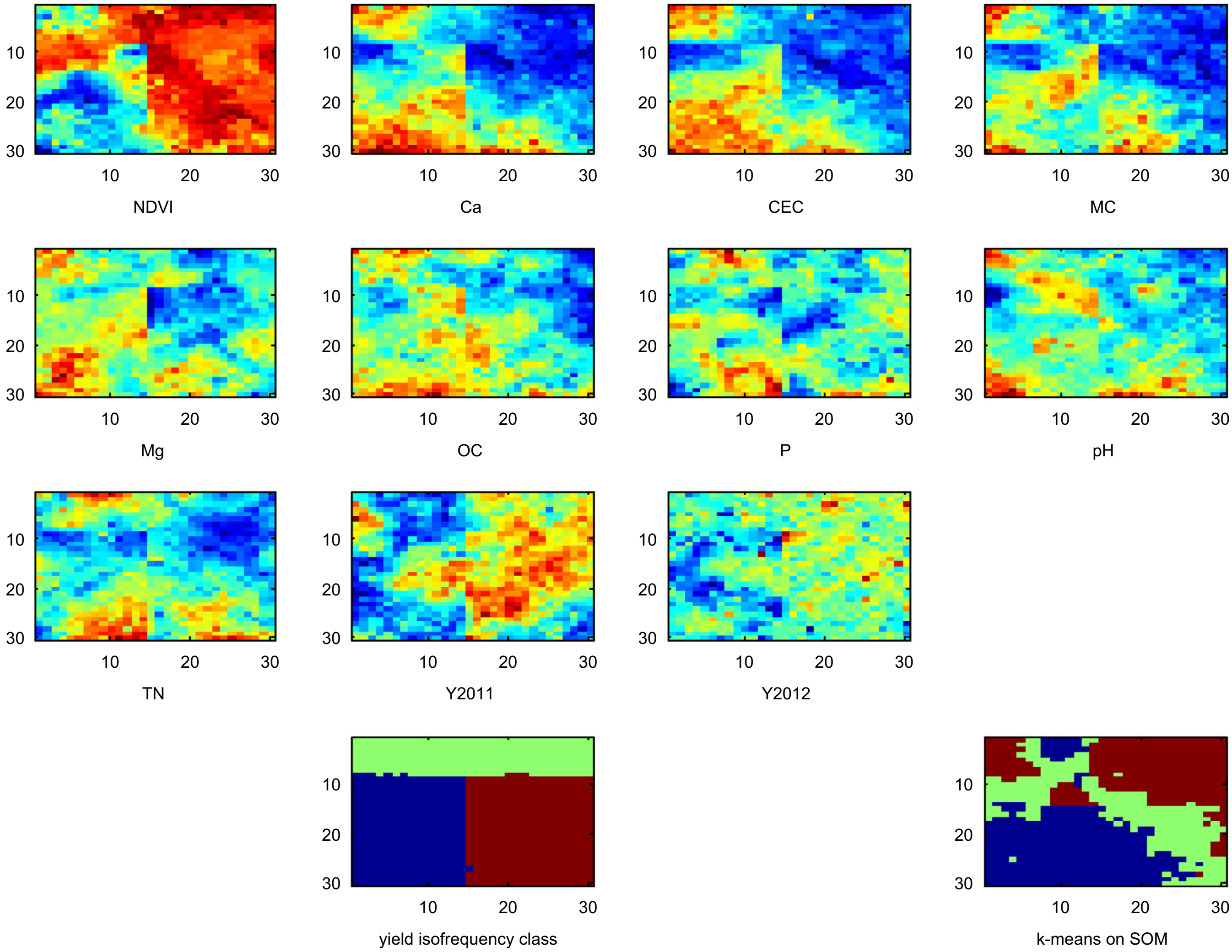

Fig. 3. Supervised Kohonen networks (SKN) predicted maps of normalized difference vegetation index (NDVI), Soil parameters, historic yields for 2011–2012, target yield classes and SOM predicted yield classes. The lower values are depicted in blue and the high values are depicted in red, while other values obtain intermediate colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

on three equal class division of the yield samples. Examination of Fig. 2 reveals high spatial similarity between the target yield Fig. 2b and the predicted yield in Fig. 2c, as has been manifested in the contingency Table 3 (SKN). Based on the values from Table 3, the kappa hat coefficient was calculated as an indicator of reliability of the agreement between the two maps (Congalton, 1991). The value of kappa hat coefficient is equal to 0.8386 which shows that the high agreement of the predicted classes in map 2c with respect to the real classes shown in map 2b cannot be achieved by chance.

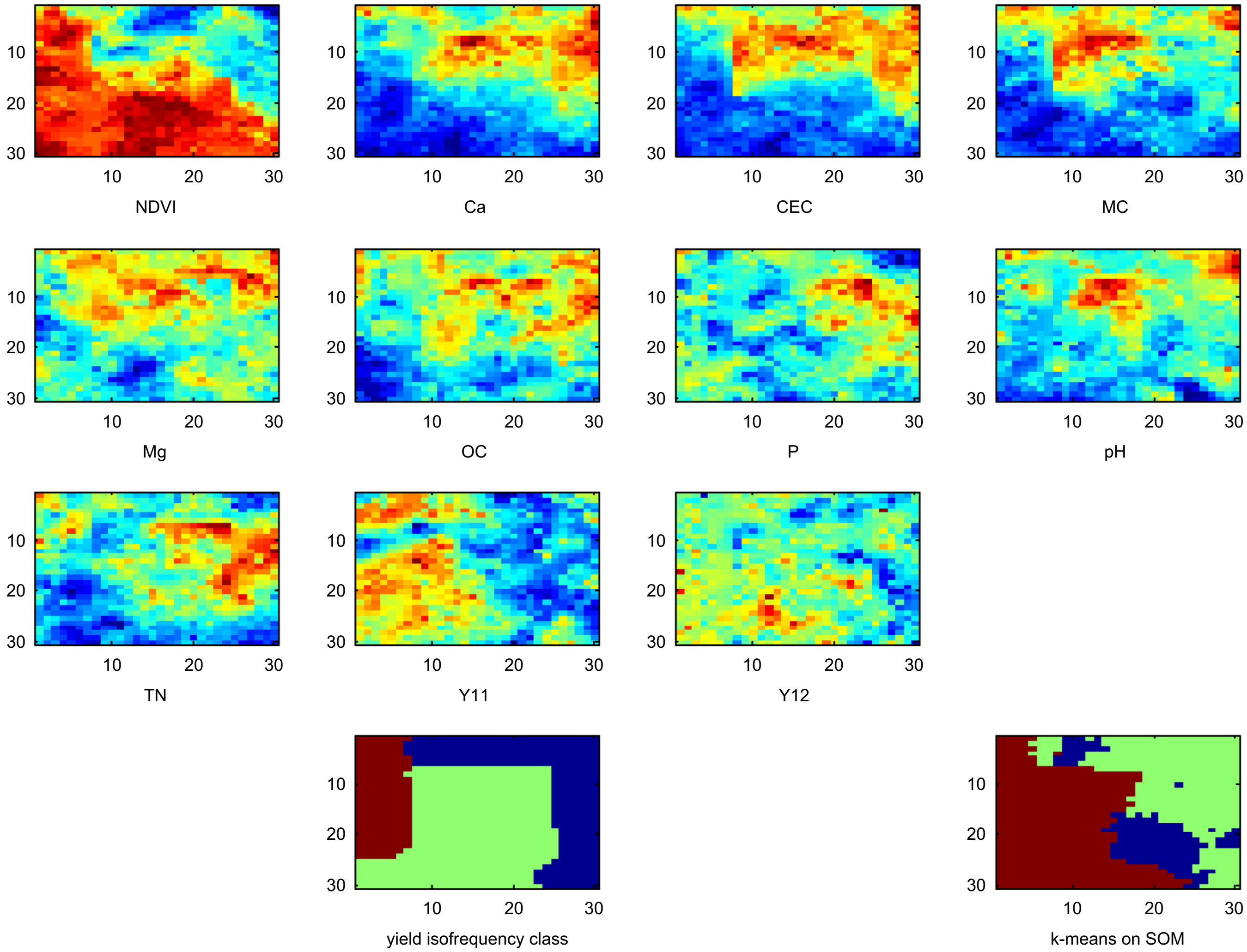

Fig. 4. XY-fusion (XY-F) predicted maps of normalized difference vegetation index (NDVI), soil parameters, historic yields for 2011–2012, target yield classes and SOM predicted yield classes. The lower values are depicted in blue and the high values are depicted in red, while other values obtain intermediate colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The training vectors that have been used for calibration of the hierarchical SOM models create a specific topological structure, which exploits the similarity between the vectors to create clusters. The most persistent phenology in the behavior of the training data imposes a particular structure which can be used to infer visual relationships between components of the training vectors that are capable of revealing correlations between different components. In the particular problem of yield prediction of the current work, the topological structure enables to infer the relationships between limiting factors affecting the yield and also inter relationship between these factors. Practically, this means that variations in soil parameters and NDVI will appear in the same local neurons of the SOM as correlated to certain yield class thus revealing an automatic extraction of a relationship between individual tendency of a specific parameter and the yield. In practical terms, the two dimensional SOM grid allows the visualization of SOM component vectors in a color coded matrix where the lower values are depicted in blue and the high values are depicted in red, while other values obtain intermediate colors. The SKN, XY-F and CP-ANN SOM clusters of the components of the training vectors are shown in Figs. 3–5 respectively. The first subplot corresponds to NDVI component while the second to ninth correspond, respectively, to Ca, CEC, MC, Mg, OC, P, pH, TN, measured with the on-line soil sensor (Mouazen, 2006). The tenth and eleventh subplot correspond to historic yield collected in 2011–2012. The last two subplots correspond to target yield classes and SOM predicted yield classes and classified into three clusters based on k-means algorithm (MacQueen, 1967). In the case of SKN network, every sample (x_i) and corresponding class vector (c_i) are associated and form an input for the network, which explains the vertical line in the component planes, as the input and output components co-develop during clustering.

By examining Figs. 3–5, it is evident that NDVI spatial distribution shows the highest similarity with predicted yield of three isofrequency classes. Most soil parameters demonstrate spatial similarity to the lower class of yield, although high concentrations of certain soil nutrients and MC were observed in the corresponding low yield area of the field. The CEC shows the inverse behavior with the spectrum of NDVI (Figs. 3–5). Out of the properties that influence crop yield MC is the most variable because there is an optimum for soil moisture (depending on the crop growth stage) but would become a limiting factor to crop yield after reaching a threshold. Discussing this issue with the farmer, and by field examination, it was revealed that although the low yield zones was rich with nutrients, a water logging problem was responsible for the low class yield in the affected areas. Due to the above mentioned water logging problem associated to the poor drainage system in the north-west part of the field, MC has minor influence on the

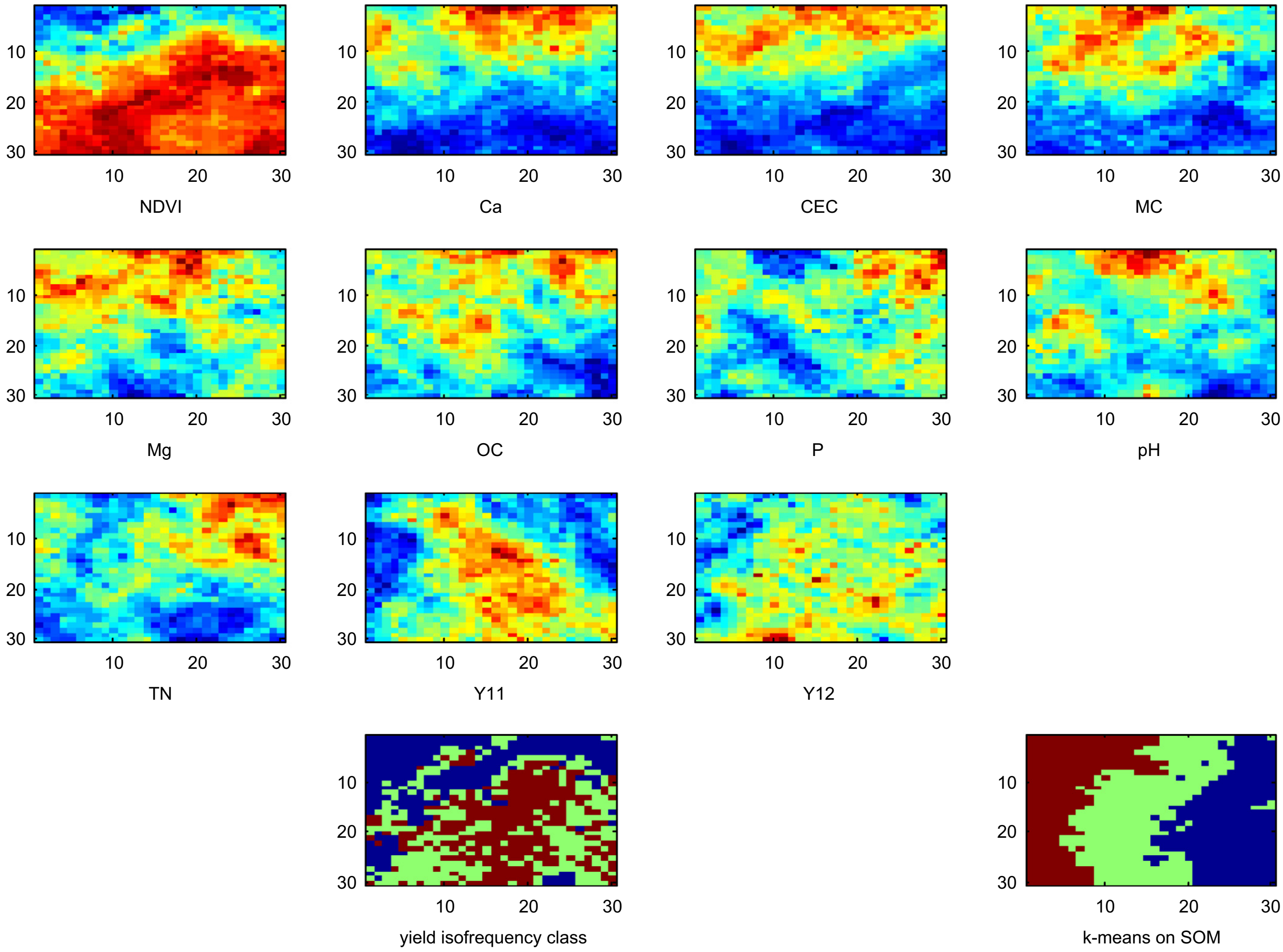

Fig. 5. Counter-propagation artificial networks (CP-ANN) predicted maps of normalized difference vegetation index (NDVI), soil parameters, historic yields for 2011–2012, target yield classes and SOM predicted yield classes. The lower values are depicted in blue and the high values are depicted in red, while other values obtain intermediate colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

crop yield. The affected areas are high in MC and nutrient concentrations but low in yield. Water logging causes respiration problems in the crop roots. Water logging has been reported as a factor reducing grain yield in UK, causing oxygen depletion in the root zone, leading to reduced photosynthesis and plant growth (Cannell et al., 1980). Additionally, it can result in significant loss in grain (Condon and Giunta, 2003).

Among the three networks used in this study, each network has a different cluster structure, which is due to different initialization and learning dynamics. However the correspondence in different components indicates the same tendencies. Considering the fact, that each vector belongs to different yield topology, the SOM maps can demonstrate in a more successful way, how clusters from various points of the field can be quite compact while on the other hand, they can be much dispersed spatially in the field. Therefore, the advantage of the SOM maps lays on their ability to provide information about the factors affecting the yield productivity in a more precise way due to the clusters tendency, which has been described above.

The results obtained in the current study for the prediction of wheat yield and classification of field into three different yield categories based on combination of high sampling resolution data and modeling techniques are superior to those presented in literature. Previous researches developed yield prediction models based on only limited soil parameters (Papageorgiou et al., 2013), or collected with traditional soil sampling of few samples per ha (Norouzi et al., 2010), as compared to the approach proposed in the current study of multi-hundred sample per ha. Uno et al. (2005) have utilized solely crop parameters in the form of vegetation indices to predict yield. The combination of soil and crop characteristics as they correlate with yield historical data seems to provide more science based prediction of crop yield. However, it should be noted that it is not necessary correct to assume that high fertile zones should link with high crop performance and yield. Therefore, a more holistic approach should be considered for correct predictions, where field observations and practical discussions with farmers need to be done before recommendations for fertilizer applications can be implemented.

The presented hierarchical map models appear as effective methods for solving classification problems, due to their capability to model classes which are separated with non-linear boundaries. The advantage of these models lies in the local updates compared with the global updates that are needed for training classical ANN architectures such as Multi-Layer Perceptron (MLP).

A limitation of the current work, concerns the inability to model continuous output relations. Thus, they are more appropriate for classification tasks. This could be overcome by enhancing the proposed architectures with smooth interpolating kernels.

The presented approach is generic in the sense that it can be used for modeling arbitrary classification functions connecting agronomic parameters and yield or quality indices.

4. Conclusions

In the current research, three SOM based models, namely, supervised Kohonen networks (SKN), counter-propagation artificial networks (CP-ANN) and XY-fusion (XY-F), which use Supervised Learning to associate high resolution data on soil and crop with isofrequency classes of wheat yield productivity were utilized. For the implementation of this approach, physicochemical soil parameters were gathered by utilizing an on-line visible and near infrared (vis–NIR) spectroscopy sensor, which were subsequently combined with crop growth indicators measured with satellite imagery by means of a sensor fusion approach The best overall results were obtained from the SKN network for the prediction of the low category of wheat yield with a correct classification reached 91.3% for both cross validation and independent validation. The hierarchical SOM maps provided visual information about the factors affecting the yield productivity in a more precise way, as compared to existing approaches. The similarity of maps of soil and crop properties with crop yield map revealed that the NDVI is more strongly correlated with yield compared to soil properties. It was also worth to note that high fertile zones in the field associated with poor yield, which was explained by water logging problems in these nutrients rich zones. It was concluded that the SKN model can be used to predict wheat yield and to classify field area into different yield potential zones.

Acknowledgements

This research was sponsored by the ICT-AGRI2 under the European Commission's ERA-NET scheme under the 7th Framework Programme. The research received funding for FarmFUSE project entitled "Fusion of multi-source and multi-sensor information on soil and crop for optimised crop production system".

References

Ayoubi, S., Sahrawat, K.L., 2011. Comparing multivariate regression and artificial neural network to predict barley production from soil characteristics in northern Iran. Arch. Agron. Soil Sci. 57 (5), 549–565.

Ayoubi, S., Khormali, F., Sahrawat, K.L., 2009. Relationships of barley biomass and grain yields to soil properties within a field in the arid region: use of factor analysis. Acta Agric. Scand. Sect. B – Soil Plant Sci. 59 (2), 107–117.

Besalatpour, A., Hajabbasi, M.A., Ayoubi, S., Afyuni, M., Jalalian, A., Schulin, R., 2012. Soil shear strength prediction using intelligent systems: artificial neural networks and an adaptive neuro-fuzzy inference system. Soil Sci. Plant Nutr. 58 (2), 149–160.

Cannell, R.Q., Belford, R.K., Gales, K., Dennis, C.W., Prew, R.D., 1980. Effects of waterlogging at different stages of development on the growth and yield of winter wheat. J. Sci. Food Agric. 31 (2), 117–132.

Condon, A.G., Giunta, F., 2003. Yield response of restricted-tillering wheat to transient waterlogging on duplex soils. Aust. J. Agric. Res. 54 (10), 957–967.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37 (1), 35–46.

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.

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

Effendi, Z., Ramli, R., Ghani, J.A., 2010. A back propagation neural networks for grading *Jatropha curcas* fruits maturity. Am. J. Appl. Sci. 7 (3), 390–394.

Fortin, J.G., Anctil, F., Parent, L.É., Bolinder, M.A., 2010. A neural network experiment on the site-specific simulation of potato tuber growth in Eastern Canada. Comput. Electr. Agric. 73 (2), 126–132.

Halcro, G., Corstanje, R., Mouazen, A.M., 2013. Site-specific land management of cereal crops based on management zone delineation by proximal soil sensing. In: Stafford, J. (Ed.), Precision Agriculture (2013). Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 475–481.

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. Trans. ASABE 49 (6), 2027–2037.

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.

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

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

Kuang, B., Mouazen, A.M., 2013. Non-biased prediction of soil organic carbon and total nitrogen with vis–NIR spectroscopy, as affected by soil moisture content and texture. Biosyst. Eng. 114 (3), 249–258.

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

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

Marin-González, O., Kuang, B., Quraishi, M.Z., Munoz-Garcia, M.A., Mouazen, A.M., 2013. On-line measurement of soil properties without direct spectral response in near infrared spectral range. Soil Tillage Res. 132, 21–29.

Marini, F., 2009. Artificial neural networks in food analysis: trends and perspectives. Anal. Chim. Acta 635, 121–131.

Melssen, W., Wehrens, R., Buydens, L., 2006. Supervised Kohonen networks for classification problems. Chemometr. Intell. Lab. Syst. 83, 99–113.

Miao, Y., Mulla, D.J., Robert, P.C., 2006. Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. Precis. Agric. 7 (117), 135.

Mouazen, A.M., 2006. Soil Survey Device. International Publication Published Under the Patent Cooperation Treaty (PCT). World Intellectual Property Organization, International Bureau. International Publication Number: WO2006/015463; PCT/BE2005/000129; IPC: G01N21/00; G01N21/00.

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

Mouazen, A.M., Maleki, M.R., Cockx, L., Van Meirvenne, M., Van Holm, L.H.J., Merckx, R., De Baerdemaeker, J., Ramon, H., 2009. Optimum three-point linkage set up for improving the quality of soil spectra and the accuracy of soil phosphorous measured using an on-line visible and near infrared sensor. Soil Tillage Res. 103 (1), 144–152.

Mouazen, A.M., Alhwaimel, S.A., Kuang, B., Waine, T., 2014. Multiple on-line soil sensors and data fusion approach for delineation of water holding capacity zones for site specific irrigation. Soil Tillage Res. 143, 95–105.

Norouzi, M., Ayoubi, S., Jalalian, A., Khademi, H., Dehghani, A.A., 2010. Predicting rainfed wheat quality and quantity by artificial neural network using terrain and soil characteristics. Acta Agric. Scand. Sect. B – Soil Plant Sci. 60 (4), 341–352.

Papageorgiou, E.I., Aggelopoulou, K.D., Gemtos, T.A., Nanos, G.D., 2013. Yield prediction in apples using Fuzzy Cognitive Map learning approach. Comput. Electr. Agric. 91, 19–29.

Rao, J.P., 1992. Expert Systems in Agriculture, <http://www.manage.gov.in/managelib/faculty/PanduRanga.htm>.

Rouse Jr., J., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring Vegetation Systems in the Great Plains with ERTS. NASA Special Publication, 351, p. 309.

Schultz, A., Wieland, R., Lutze, G., 2000. Neural networks in agroecological modelling – stylish application or helpful tool? Comput. Electr. Agric. 29, 73–97.

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.

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

Uno, Y., Prasher, S.O., Lacroix, R., Goel, P.K., Karimi, Y., Viau, A., Patel R.M., 2005. Artificial neural networks to predict corn yield from Compact Airborne Spectrographic Imager data. Comput. Electr. Agric., 47(2), 149–161.

Viscarra Rossela, R.A., Walvoortb, D.J.J., McBratneya, A.B., Janikc, L.J., Skjemstadc, J. O., 2005. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1–2), 59–75.

Zolfaghari, Z., Mosaddeghi, M.R., Ayoubi, S., 2015. ANN-based pedotransfer and soil spatial prediction functions for predicting Atterberg consistency limits and indices from easily available properties at the watershed scale in western Iran. Soil Use Manage. 31 (1), 142–154.

Zupan, J., Novic, M., Gasteiger, J., 1995. Neural networks with counter-propagation learning strategy used for modelling. Chemometr. Intell. Lab. Syst. 27, 175–187.