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):

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

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.