Soil Research, Vol. 56, Issue 1, 2017, pp. 19-27
DOI:10.1071/SR16264

Quantifying individual and collective influences of soil properties on crop yield

Rebecca Whetton[a], Yifan Zhao[b,*], Abdul M Mouazen[a,c]

[a] *Cranfield Soil and AgriFood Institute, Cranfield University, Bedfordshire MK43 0AL, UK.*

[b] *Through-life Engineering Services Centre, Cranfield University, Bedfordshire MK43 0AL, UK*

[c] *Department of Soil Management, Ghent University, Coupure 653, 9000 Gent, Belgium*

* *E-mail of corresponding author: Abdul.Mouazen@UGent.be*

Abstract

Quantifying the agronomic influences of soil properties, collected at high sampling resolution, on crop yield is essential for site specific soil management. This study implements a novel Volterra Non-linear Regressive with eXogenous inputs (VNRX-LN) model, to quantify causal factors to explain yield using high resolution data on key soil properties affecting wheat yield in a 22 ha field with waterlogging problem in Bedfordshire, UK. A total of eight soil properties including total nitrogen (TN), organic carbon (OC), pH, available phosphorous (P), magnesium (Mg), calcium (Ca), moisture content (MC), and cation exchange capacity (CEC) were collected with an on-line (tractor mounted) visible and near infrared spectroscopy (vis-NIR) sensor and used as multiple-input to the VNRX-LN model, while crop yield represented the single-output in the system.

Results showed that the largest contributors to wheat yield were CEC, Mg and TN, with error reduction ratio contribution (ERRC) values of 14.6%, 4.69% and 1%, respectively. The overall contribution (SEER) of the soil properties considered in this study totals a value of 23.21%. This was attributed to a large area of the studied field having been

waterlogged, which masked the actual effect of soil properties on crop yield. It is recommended to validate the introduced concept on a larger number of fields, where other crop yield affecting parameters e.g., crop disease, pests, drainage, topography and microclimate conditions should be taken into account.

Keywords: Yield limiting factors; proximal soil sensing; VNRX; nonlinear parametric modelling.

Table of abbreviation

AFOLS	Adaptive-forward-orthogonal least squares
Ca	Calcium
CAA	Circle-based average approximation
CEC	Cation exchange capacity
DGPS	Differential global positioning system
ERR	Error reduction ratio
ERRC	Error reduction ratio contribution
LAI	Leaf area index
MC	Moisture content
Mg	Magnesium
NDVI	Normalised difference vegetation index
NFIR	nonlinear finite impulse response
OC	Organic carbon
OLS	Orthogonal least squares
P	Phosphorus
SDA	Shortest distance approximation
SERR	Sum of error reduction ratio
TN	Total nitrogen
vis-NIRS	Visible and near infrared
UTM	Universal transverse Mercator
VNRX	Volterra non-linear regressive with eXogenous
VNRX-LN	Volterra non-linear regressive with eXogenous, accounting for both linear and non-linear variability

1. Introduction

The world's population is expected to rise to 9 billion by 2050 and, based on the current land available, an increase in crop yield of 60% will be required. Precision management of farm resources (e.g., fertilisers, seeds, water, etc.) is one potential way to increase crop yield. The spatial variability in agricultural fields exists at different spatial scales (Raun 1998; Dhillon *et al.* 1994), which requires precise management with the aim to increase yield at reduced input cost and related environmental impacts. This is hardly achievable by conventional agriculture that relies on homogeneous applications of external inputs. For example, current fertiliser applications are made based on a bulked composite soil sample collected per field or 1-3 ha in the best scenario, which ignore within field variability. This may result in over-application in rich zones, and under-applications in poor zones in the field. In this context, recent years have seen a surge of variable-rate application technologies where external farm inputs are applied in response to input data from normalised difference vegetation index (NDVI), leaf area index (LAI), high resolution soil properties or a combination of these (Lowenberg-DeBoer and Aghib 1999; Maleki *et al.* 2008; Mouazen *et al.* 2009; Halcro *et al.* 2013; Mouazen and Kuang 2016). Although variable rate fertilisation is a strategy to increase crop yield, understanding and quantifying the yield limiting factors is still a crucial research question to be answered, before variable rate applications can be optimised.

Since spatial variability in the majority of agricultural fields exist, proximal sensor technologies are invaluable to measure this variability accurately. This will require robust and reliable sensing platforms of crop and soil. Proximal (e.g., Crop Circle ACS 470, Holland Scientific, Lincoln, NE USA) and remote sensing (e.g., satellite imagery, unmanned aerial vehicles or aircrafts) both can provide high resolution data on crop canopy characteristics indicated e.g., as NDVI or LAI (Mulla 2013; Kipp *et al.* 2014) and they are commercially available. However, remote sensing methods based on spectral reflectance provide data on the top millimetres of soil and require a bare soil surface. Furthermore, due to the complex nature and vast variability of agricultural soils, the majority of proximal soil sensors are still premature to fulfil this requirement. Kuang *et al.* (2012) concluded in an extensive review that the most promising proximal sensing technologies for quantifying soil properties are electrochemical technique and optical visible and near infrared (vis-NIR) spectroscopy. Although they are limited to particular research groups worldwide, on-line (tractor mounted) vis-NIR sensors (Shibusawa *et al.* 2001; Mouazen *et al.* 2006; Christy 2008) enable the collection of high sampling resolution (e.g., >500 samples per ha) of key soil properties (Kuang *et al.* 2012; Kuang and Mouazen 2013; Marin-González *et al.* 2013; Kodaira and Shibusawa 2013; Kweon *et al.* 2013), which are valuable sources of information to manage the within field spatial variability.

Nonlinear parametric modelling approaches offer novel tools for the quantification and better understanding of the influences of soil related yield limiting factors, collected at high sampling resolution with on-line soil sensors, which cannot be obtained with the traditional soil sampling and laboratory analytical methods. One of these parametric methods is Volterra Non-linear Regressive with eXogenous inputs (VNRX-LN) model, which was broadly used in the engineering sector, but not common in agriculture.

The aim of this work was to use the VNRX-LN model to quantify causal factors to explain yield using high resolution data on key soil properties affecting wheat yield in a 22 ha field with waterlogging problem in Bedfordshire, UK.

2. Materials and Methods

2.1 Study site

The study site was one field designated as Horns End, and located at a commercial farm, called Duck end farm, in Wilstead, Bedfordshire UK (52°5’52.087”W latitude and 0°27’19.76”N longitude). The field is about 22 ha area, with an average annual rainfall of 598 mm. According to the UK meteorology Office (http://www.metoffice.gov.uk/climate/uk/summaries), May was particularly wet in 2013, and spring was cooler than average, whilst summer was the driest for the UK since 2003. Nevertheless, there were some notably wet days, particularly in July and August. The farm has a crop rotation of barley (Hordeum vulgare), wheat (Triticum aestivum) and oil seed rape (Brassica napus). The soil texture over the field down to 0.20 m is non-homogeneous, including three textures of sandy loam, loam, and sandy clay loam according to the United State Department of Agriculture (USDA) texture classification system. Wheat was cultivated during the experiment in 2013.

2.2 On-line collected data

The on-line vis-NIR sensor (Mouazen 2006) was used (Figure 1) to carry out the field measurement. It consists of a subsoiler that penetrates the soil to the required depth, making a trench, whose bottom is smoothened due to the downwards forces acting on the subsoiler (Mouazen *et al.* 2005).

[Figure 1]

The optical probe, housed in a steel lens holder, was attached to the rear of the subsoiler chisel to acquire soil spectra in reflectance mode from the smooth bottom of the trench. The subsoiler, retrofitted with the optical unit, was attached to a frame that was mounted onto the three point hitch of the tractor. An AgroSpec mobile, fibre type, vis-NIR spectrophotometer (tec5 Technology for Spectroscopy, Germany) with a measurement range of 305-2200 nm was used to measure soil spectra in diffuse reflectance mode. A differential global positioning system (DGPS) (EZ-Guide 250, Trimble, USA) was used to record the position of the on-line measured spectra with sub-metre accuracy. On-line soil measurement occurred in summer 2012 after the harvest of the previous crop, at parallel transects of 15 m space, with an average forward speed of the tractor of 2 km h^{-1} and the measurement depth set at 150 mm. A few on-line collected vis-NIRS spectra are shown in Figure 2, as an example.

[Figure 2]

During on-line measurement, two or three soil samples per line were collected from the bottom of a trench and the sampling positions were carefully recorded with the DGPS. These samples were analysed for calcium (Ca), magnesium (Mg), cation exchange capacity (CEC), phosphorous (P), pH, moisture content (MC), organic carbon (OC) and total nitrogen (TN), using the following laboratory analytical methods:

- Exchangeable Ca and Mg were determined by Agilent 240 FS AA atomic absorption spectrophotometry (Agilent Technologies, Inc. USA).
- CEC was determined using a Flame Photometer (Chapman 1965).

- Available P concentration was determined by an ascorbic acid method (Olsen *et al.* 1954).
- pH was measured potentiometrically on a suspension of soil to water ratio (1:2.5) (DEFRA 1986).
- MC was determined by oven drying of samples at 105º for 24 h.
- OC was determined using a combustion method (British Standard BS 7755 Section 3.8 1995).
- TN was determined by the Dumas method, where soil samples are heated to 900 ºC in the presence of oxygen gas (British Standard BS EN 13654-2:2001).

The selection of these eight soil properties was attributed to the fact that these properties are considered important in explaining crop yield response and can be measured with the on-line vis-NIRS sensor with appreciable accuracy (Kuang and Mouazen 2013; Marin-González *et al.* 2013).

Partial least squares regression (PLSR) based calibration models, developed with Unscrambler V9.8 software (Camo Software, Norway) were used to predict all eight soil properties using the on-line collected soil spectra (>500 samples per ha). The on-line prediction accuracy of properties with direct spectral responses (i.e., MC, OC and TN) indicated as residual prediction deviation (the ration of standard deviation divided by root mean square error of prediction (RMSEP) ranged between 1.96 and 3.06 (good to excellent predictions). For the soil properties with indirect spectral responses (i.e., Ca, Mg, CEC, P and pH), RPD ranged between 1.30 and 2.14 (moderate to good predictions). More details about the on-line vis-NIR sensor and accuracy of measurement can be found in Kuang and Mouazen (2013) and Marin-González *et al.* (2013).

Wheat yield data was collected in August, 2013 by the on-board yield sensor and GPS system of the farmer's combine harvester (New Holland, CX8070 model), with a header width of 7.25 m commonly used for barley and wheat harvest. In addition, the harvest was optimised to: I) record wheat yield when the machine header was full for the full length of the study area, and II) avoid the bare soil in the tramlines. Total yield was calculated from the mean yield (tonnes per hectare) of an area, multiplied by the size of the area (m^2), which was derived using ArcGIS (Esri, USA).

2.3 Data processing

Features in the environment, are the product of many interacting processes, including physical, chemical and biological. They are determined with exceedingly complex interactions, which along with incomplete understanding can make the occurrence seem random. Due to this, a way of overcoming the prediction of distribution is to treat the variation as if it is random (Matheron 1963). The measurement points from the on-line soil sensor and yield sensor required a method of interpolation, to provide a continuous data set across the locations. Kriging was selected as a non-biased approach to predict the values between the sample points, where semi-variograms were first produced and then applied in Kriging predictions. The interpolated data were then converted into a common 5 m raster grid in ArcGIS (Esri, USA) in order to assist data fusion (Frogbrook and Oliver 2007). The raster squares of the layers were converted into this common grid of points by extracting the value at the midpoint of each raster square. A smaller resolution has no practical implementation, due to the limitations of the size and response time of the precision farming equipment. The 5 m grid size provided a balance between adequately characterising the spatial variation and practical farm management. These steps ensured that all layers consisted of a common set of 5 m grid point-values, to allow the application of parametric modelling to be carried out. This method allowed data from a diverse range of soil and crop property surveys, measured at different resolutions, to be merged (Khosla *et al.* 2008). The different soil and crop layers of a 5 by 5 m grid were subjected to the VNRX-LN detailed in the following section.

2.4 Volterra Non-linear Regressive with eXogenous Model

In this study, the simplified VNRX-LN model, also known as NFIR model, was implemented, which represent a multi-inputs and single-output system:

$$y = f(u_1, u_2, \ldots, u_R) + \varepsilon \tag{1}$$

where R is the number of the system inputs, f is some unknown linear or non-linear mapping, which links the system output y to the system inputs $u_1, u_2, \ldots, u_R$; ε denotes the model residual.

The on-line measured soil properties (i.e., TN, OC, pH, P, Mg, Ca, MC, and CEC) were normalised and used as inputs ($R = 8$) to the VNRX-LN model, whereas the model output was wheat yield. The analysis also included the interaction between pairs of soil properties and their contribution to crop yield. The aim was to investigate the contribution of each soil property and their pairwise interaction on crop yield.

Parameters are estimated based on the observations, and these are determined by the structure, using the orthogonal least squares (OLS) estimation procedures. Adaptive-forward-orthogonal least squares (AFOLS) was employed not only to determine the model structure but also to estimate the unknown parameters. More detailed description of this method can be found in Zhao *et al.* (2012).

Performance of VNRX-LN model output was evaluated by considering the value of error reduction ratio (ERR) for each parameter to the prediction of yield (system outputs). Values of ERR always range from 0% to 100%. The larger the ERR is, the higher the dependence is between this term and the output. It is, therefore, a useful index to indicate the contribution of each term to the output. To calculate the contribution of each input variable to the output, the sum of ERR values (SERR) of all selected terms is used to describe the percentage explained by the identified model to the system output. If the considered inputs can fully explain the variation of system output, the value of SERR is equal to 100%. It is an indicator of model performance and uncertainty. The contribution of the i^{th} input variable to the variation of the system output, denoted as $ERRC_i$, is defined as the sum of ERR values of the terms that include this input variable. The value of $ERRC_i$ should be always between 0% and 100%.

2.5 Significance Test

To determine the statistical significance of the contribution from each input to the system output, a threshold τ_i, representing the level of contribution, above which value had less than a 5% probability of occurring by chance, requires being determined. The conventional 95% confidence interval is not suitable for this study because the distribution of ERRC value is unknown. For this purpose, the following surrogate data technique was used.

Assuming the signal Y is a function of the signal X, this sort of dependence is destroyed when Y is ordered randomly in some way while X keeps the same order. For this purpose, the order of the data in Y was randomised by a shuffle procedure that saves the distribution properties of the Y signal, but destroys the spatial relationship between X and Y. This procedure was repeated 100 times and then the 95% quantile was determined as the threshold. A significance threshold for each term is firstly calculated, and then the significance threshold for each input can then be derived by the same way to calculate $ERRC_i$.

2.6 Optimal spatial resolution of soil properties versus yield

Since the spatial sampling resolutions of soil properties and crop yield are different, before applying the proposed VNRX-LN modelling method, the data must be re-sampled to establish the correspondence between the inputs and the output. Two re-sampling techniques have been used in this study. In the first technique, for each crop yield data $y(e_i, w_i)$ on a location (e_i, w_i), the corresponding soil properties were approximated by the properties on the location that has the shortest distance to (e_i, w_i), which must be smaller than a radius r. It is possible that some crop yield data cannot find corresponding soil properties if r is too small, for which scenario this yield data will be discarded. In the second technique, for each crop yield data $y(e_i, w_i)$, each corresponding soil property was approximated by the averaging value of all values of this soil property inside a circle with a radius r. A small value of r refers to more accurate correspondence between yield and soil properties, but a lower number of samples included in the analysis. The former method of re-sampling is designated here as ‘shortest distance approximation (SDA)’, whereas the latter method is designated as ‘circle-based average approximation (CAA)’.

3. Results and discussion

3.1 Pearson correlations

Pearson coefficient (r) values between pairs of soil properties suggest collective (positive) linear relationships to exist between Ca and CEC, MC, Mg, OC, pH and TN (r = 0.519 - 0.747) and between CEC and Ca, Mg, MC and pH (r = 0.590 - 0.748). This may indicate that although Ca has no direct spectral response in the NIR range, it is measured with vis-NIR spectroscopy through covariation with MC and OC, both having direct spectral response (Stenberg *et al.* 2010; Kuang *et al.* 2012). However, CEC is measured through covariation with MC only. As expected, TN correlated with OC, which is a similar result to that reported elsewhere (Carlyle 1993; Kuang and Mouazen 2011).

Examining r values between the eight on-line measured soil properties and yield, reveals negligible (negative) relationships (Table 1) between laboratory measured soil properties and yield. The highest linear correlation is calculated between CEC and yield (r = -0.349). This again proves the complexity of the system and necessitates the need for more advanced modelling techniques that account for both linear and nonlinear interactions.

[Table 1]

3.2 Model output

The detailed correspondence between inputs variables and soil properties are described in Table 2. The initial full model, based on quadratic terms, was chosen in this paper, which can be written as follows:

$$y = \theta_0 + \sum_{i=1}^{8} \theta_i u_i + \sum_{i=1}^{8} \sum_{j=i}^{8} \theta_{ij} u_i u_j + \varepsilon \qquad (2)$$

This model has 45 terms. All inputs and output were normalised by removing the mean. The proposed method was then applied to calculate the ERRC of each term. Table 3 lists the first 10 terms selected using the SDA re-sampling technique with a 3 m radius. From this calculation it was observed that the contribution of CEC to the wheat yield variability was the largest (e.g. ERRC = 15.68%) among the 45 terms, including all soil properties and their interactions. This was followed successively by Mg (ERRC = 3.57%) and Ca * CEC (ERRC = 1.13%) terms. This is explained by the fact that although CEC is not a nutrient, it is a widely accepted measure to assess the fertility of the soil. In fact, CEC represents the soil ability to hold positively charged ions e.g., exchangeable cations, which is directly linked to nutrients, hence, it is an important indicator of soil fertility (Hazelton and Murphy 2007). Its significant contribution to crop yield could be due to the quantity of nutrients in the field being variable through the field. Furthermore, CEC is an important indicator influencing soil structure stability, nutrient availability, soil pH and the soil's reaction to fertilisers and other ameliorants (Hazelton and Murphy, 2007), which as a result will have a positive influence of crop growth and yield. Furthermore, CEC is also related to potassium content and clay particles, which affect available water content (Bergaya and Vayer 1997), hence, influencing crop growth and development.

[Table 2]

[Table 3]

By comparing the contribution of each soil property to the wheat yield with the corresponding significance threshold, the soil properties having significant contribution to the crop yield can then be highlighted as shown in Table 4. Amongst the eight studied soil properties, CEC, Mg, TN, Ca, OC and MC all have significant influence on the crop yield, with declining order. However, the largest influence is attributed to CEC, followed successively by Mg and TN. It is worth noting that pH is normally associated with soil fertility and CEC (Hazelton and Murphy 2007) has the lowest influence on yield. But, pH level directly affects nutrient availability and crop nutrient uptake (HGCA 2014). With acidic soils (soil pH is smaller than 5), the pH would have negative influence on nutrient uptake. It is commonly stated in farmer's guides that the optimum pH for soils under continuous arable cropping of cereal crops is between 6 and 7 with 6.5 being the ideal. However, in the Horns End experimental field, the pH value of the majority of the field area ranged between 5.6 and 8, which may explain the low contribution of pH to yield prediction (Bruulsema 2015). Similar observation can be made for P. Although P is a key nutrient for crop growth and development, no significant contribution to wheat yield was observed. One explanation could be that P is not a limiting property in Horns End field, as manure is being frequently applied (Mouazen and Kuang 2016). Another reason might be the fact that a part of the field i.e., the north-west part experienced a waterlogging problem associated with a poor drainage system for many years. This is also reflected on the poor yield harvested in 2013, as shown in Figure 3, where low harvest can be observed particularly on the northern and south western parts of the field, coinciding well with areas with the waterlogging problem.

[Figure 3]

[Table 4]

A multiple linear regression analyses with least square estimation conducted by Kravchenko and Bullock (2000) found OC as the main and most consistent, positively correlated parameter with corn and soybean yield. Interestingly, they found that the contribution from K, CEC and P was mostly negligible, and this was attributed to K and P being ample in abundance in the soils. This finding is in line with those of the current work regarding P only. However, Kravchenko and Bullock (2000) stated that the performance of crop prediction models varies from field to field across different cropping seasons.

After CEC and Mg, TN ranked as the third largest contributor to wheat yield, a result which is supported by previous research suggesting that nitrogen supply is a large limiting factor of crop yield (Agegnehu *et al.* 2016) and is strongly linked with soil TN content before planting and uptake rate by plants during the growing season. Surprisingly, Mg has the second largest contribution to the wheat yield variance. Mg is an essential plant nutrient for plant growth, as it has well-known roles in photosynthesis process and chlorophyll building (Mengel and Kirkby 1987). Deficiency in Mg by leaching may take place in highly acidic sandy soils. However, this is not the case of the current experimental field, where pH varies between 5.6 and 8 in a mixture of medium soil texture classes of sandy loam, loam, and sandy clay loam. This could explain the high contribution of magnesium distribution to crop yield variation.

Due to the waterlogging problem associated with the poor drainage system in the north-west part of the field (Figure 4), MC had only a minor influence on crop yield as it is ranked sixth among the eight soil properties included in the analysis. There is an optimum for soil moisture (varying with crop growth stage) being beneficial to crop yield. As MC increases it may become a hindrance to crop yield after reaching a threshold. The waterlogged areas are of high MC and nutrient concentrations but low in yield due to the water stress, which affects crop establishment, growth and yield. Waterlogging causes the crop roots to be unable to respire and when there is too little oxygen in the soil pores, the demand for oxygen varies with crop and crop growth stage (Boyer 1982). Waterlogging at grain filling stages can cause a significant loss in grain yield (Condon and Giunta 2003).

[Figure 4]

3.3 Model sensitivity to sampling technique

All results discussed above are based on the SDA re-sampling technique with a 3 m radius. To evaluate the sensitivity of the results to the selection of re-sampling technique and the size of radius, more tests were performed, whose results are shown in Table 5, in which only the top 3 significant soil properties are presented. Inspection of Table 5 reveals that the top two soil properties (e.g., CEC and Mg) showed exactly same response for all tests, appearing at first and second factors affecting yield, respectively, whereas TN appears three times and Ca appears once in the third position. Additionally, the CAA re-sampling technique consistently had a larger total contribution (SERR = 22.97% for 3 m radius) to wheat yield than that of the SDA re-sampling technique (SERR = 20.29% for 3 m radius), which indicates the CAA technique may be more suitable for the high resolution soil and yield data, because the identified model explains more of the system output. Also, the total contribution decreases following the increase of the sample number, which is expected because more samples indicate more spatial variations of the underlying rule (Billings 2013). This is also true for the radius, because with a larger radius, larger samples are included in the analysis.

[Table 5]

Results showed that the overall contribution of the eight soil properties to wheat yield is 23.21%. One would expect that the contribution of soil properties to yield should be larger than the overall calculated contribution in this study. However, the results obtained confirmed this to be a significant contribution, but also shows that there is variability still at play, influencing the crop yield (e.g., crop disease, pests, topography, micro-climatic conditions etc.). For example, whilst TN and OC should have significant effects, and both are required by the crop for healthy growth and grain production, they can also increase and prolong the leaf area index of the crop, which in turn increases humidity, making the plant more susceptible to disease, hence, crop yield is negatively affected (Bryson *et al.* 1997). Therefore, there is a need for a future work to expand on the current data mining approach to quantify yield limiting factors, under larger number of fields with different crops and different agricultural systems. The study should also account for the other affecting factors of crop yield including crop disease, pests, topography, micro-climatic conditions etc.

4. Conclusions

A volterra non-linear regressive with eXogenous inputs (VNRX) model accounting for the linear and non-linear variability (VNRX-LN) was used to quantify yield limiting factors of wheat in one field in Bedfordshire, the UK. The input data were eight soil properties (e.g. OC, TN, CEC, Mg, MC, Ca, pH and P), collected at a high sampling resolution rate (>500 sample per ha), with an on-line visible and near infrared spectroscopy (vis-NIRS) sensor, whereas crop yield represented the single-output in the system. Based on the results obtained the following conclusions can be drawn:

1. The VNRX-LN model can be successfully used to quantify the influence of multi-soil properties, collected at high sampling resolution with an on-line soil sensor, on crop yield.
2. The effect of soil properties on crop yield varied with soil property, with the largest contribution observed for CEC, Mg and TN, with error reduction ratio contribution (ERRC) values of 14.6%, 4.69% and 1%, respectively.

3. The overall contribution of the eight soil properties sums up to an ERRC value of 23.21%. This value was found to be surprisingly low, but was explained by the fact that a large part of the studied field suffers of a drainage problem, which masked the actual effect of soil properties on crop yield.

It was recommended to validate the concept introduced in this study on a larger number of fields, where other affecting parameters (e.g. crop diseases, pests, topography, microclimate conditions) of crop growth and yield should be taken into account.

Acknowledgements

We acknowledge the funding received for FarmFUSE project from the ICT-AGRI under the European Commission's ERA-NET scheme under the 7th Framework Programme, and the UK Department of Environment, Food and Rural Affairs (contract no: IF0208).

References

Agegnehu G, Nelson P, Bird M (2016) Crop yield, plant nutrient uptake and soil physicochemical properties under organic soil amendments and nitrogen fertilization on Nitisols. *Soil & Tillage Research* **160**, 1–13.

Bergaya F, Vayer M (1997) CEC of clays:measurement by absorbtion of a copper ethylenediamine comples. *Applied clay science* **12(3)**, 275-280.

Billings SA (2013) 'Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains.' (London: John Wiley & Sons)

Boyer J (1982) Plant productivity and environment. *Science* **218**, 443–448.

Bruulsema T (2015) 'Plant Nutrition TODAY.' (Georgia USA: International Plant Nutrition Institute (IPNI))

Bryson RJ, Paveley ND, Clark WS, Sylvester-Bradley R, Scott RK (1997) Use of in-field measurements of green leaf area and incident radiation to estimate the effects of yellow rust epidemics on the yield of winter wheat. *European Journal of Agronomy* **7**, 53-62.

Carlyle J (1993) Carbon in forested sandy soils: properties, processes, and the impact of forest management. *New Zealand Journal of Forestry Science* **23(3)**, 390-402.

Christy C (2008) Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. *Computers and Electronics in Agriculture* **61**, 10-19.

Condon A, Giunta F (2003) Yield response of restricted-tillering wheat to transient waterlogging on duplex soils. *Australian Journal of Agricultural Research* **54(10)**, 957-967.

Corwin DL, Lesch S, Shouse PJ, Soppe R, Ayars JE (2003) Identifying soil properties that influence cotton yield using soil sampling directed by apparent soil electrical conductivity. *Agronomy Journal* **95(2)**, 352-364.

Dhillon N, Samra J, Sadana U, Nielson D (1994) Spatial variability of soil test values in a typic Ustochrept. *Soil Technology* **7**, 163–171.

Frogbrook ZL, Oliver MA (2007) Identifying management zones in agricultural fields using spatially constrained classification of soil and ancillary data. *Soil Use and Management* **23(1)**, 40-51.

Halcro G, Corstanje R, Mouazen A (2013) Site-specific land management of cereal crops based on management zone delineation by proximal soil sensing. Precision agriculture '13. (Wageningen, Neitherlands, Wageningen Academic Publishers) pp. 475-481.

Hazelton PA and Murphy BW (2007) 'Interpreting soil test results: what do all the numbers mean Australia.' (Melbourne, Australia.: CSIRO publisher)

HGCA (2014). Oilseed rape guide, s.l.: Agriculture and Horticulture Development Board.

Khosla R, Inman D, Westfall, DG, Reich, RM, Frasier M, Mzuku M, Koch B, Hornung A (2008) A synthesis of multi-disciplinary research in precision agriculture: site-specific management zones in the semi-arid Western Great Plains of the USA. *Precision Agriculture* **9**, 85-100.

Kipp S, Mistele B, Schmidhalter U (2014) The performance of active spectral reflectance sensors as influenced by measuring distance, device temperature and light intensity. *Computers and Electronics in Agriculture* **100**, 24–33.

Kodaira M, Shibusawa S (2013) Using a mobile real-time soil visible-near infrared sensor for high resolution soil property mapping. *Geoderma* **199**, 64-79.

Kravchenko A, Bullock D (2000) Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal* **92(1)**, 75-83.

Kuang B, Mahmood HS, Quraishi Z, Hoogmoed WB, Mouazen AM, Henten E (2012) Sensing soil properties in the laboratory, in situ, and on-line: a review. *Advances in Agronomy* **114**, 155-223.

Kuang B, Mouazen A (2011) Calibration of a visible and near infrared spectroscopy for soil analysis at field scales across three European farms. *European Journal of Soil Science* **62(4)**, 629- 636.

Kuang B, Mouazen AM (2013). Effect of spiking strategy and ratio on calibration of on-line visible and near infrared soil sensor for measurement in European farms. *Soil & Tillage Research* **128**, 125-136.

Kweon G, Lund E, Maxton C (2013). Soil organic matter and cation-exchange capacity sensing with on-the-go electrical conductivity and optical sensors. *Geoderma* **199**, 80-89.

Lowenberg-DeBoer J, Aghib A (1999). Average return and risk characteristics of site specific P and K management: eastern Corn Belt on-farm trial results. *Journal of Production Agriculture* **12**, 276–282.

Maleki MR, Mouazena AM, Ketelaerea BD, Ramona H, Baerdemaekera, JD (2008). On-the-go variable-rate phosphorus fertilisation based on a visible and near infrared soil sensor. *Biosystems Engineering* **99(1)**, 35-46.

Marin-González O, Kuang B, Quraishi MZ, Munóz-García MA, Mouazen AM (2013). On-line measurement of soil properties without direct spectral response in near infrared spectral range. *Soil & Tillage Research* **132**, 21-29.

Matheron G (1963). Principles of geostatistics. *Economic Geology* **58**, 1246–1266.

Mouazen A (2006). Soil survey device. BE, Patent No. WO/2006/015463.

Mouazen AM, De Baerdemaeker J, Ramon H (2005) Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil & Tillage Research* **80(1-2)**, 171-183.

Mouazen AM, De Baerdemaeker J, Ramon H (2006) Effect of wavelength range on the measurement accuracy of some selected soil constituents using visual-near infrared spectroscopy. *Journal of Near Infrared Spectroscopy* **14**, 189–199.

Mouazen AM, Kuang B (2016) On-line visible and near infrared spectroscopy for in-field phosphorous management. *Soil & Tillage Research* **156**, 471-477.

Mouazen AM, Maleki MR, Cockx L, Van Meirvenned M, Van Holm LHJ, 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 Research* **103(1)**, 144-152.

Mulla D (2013) Twenty Five Years of Remote Sensing in Precision Agriculture: Key Advances and Remaining Knowledge Gaps. *Biosystems Engineering* **114**, 358-371.

Raun WR, Johnson GV, Lees HL, Sembiring H, Phillips SB, Solie JB, Stone ML, Whitney RW (1998). Microvariability in soil test, plant nutrient and yield parameters in Bermudagras. *Soil Science Society of America Journal* **62(3)**, 683–690.

Shibusawa S, Imade Anom SW, Sato S, Sasao A, Hirako S (2001). Soil mapping using the real-time soil spectrophotometer. In ‘Third European Conference on Precision Agriculture’. pp. 497–508. (Montpellier, France: ECPA).

Zhao Y, Billings SA, Wei H, Sarrigiannis P (2012) Tracking time-varying causality and directionality of information flow using an error reduction ratio test with applications to electroencephalography data. *Physical Review E* **86**, 051919.

Figure 1. Illustrated image of the tractor mounted on-line visible and near infrared spectroscopy (vis-NIRS) sensor (Mouazen 2006).

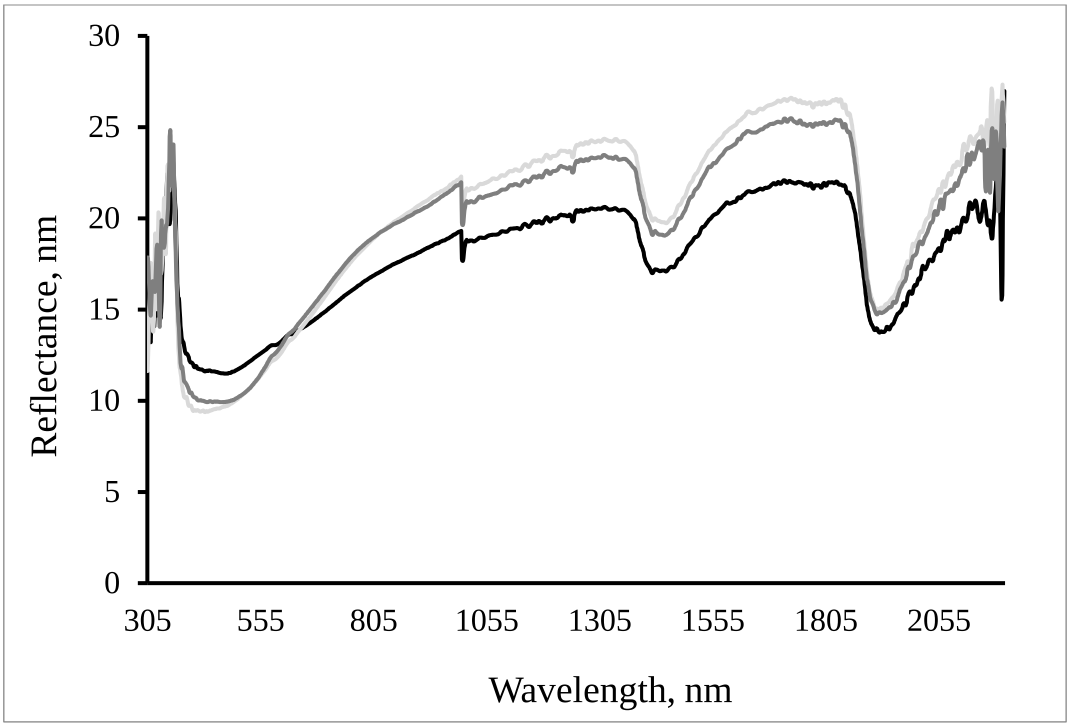
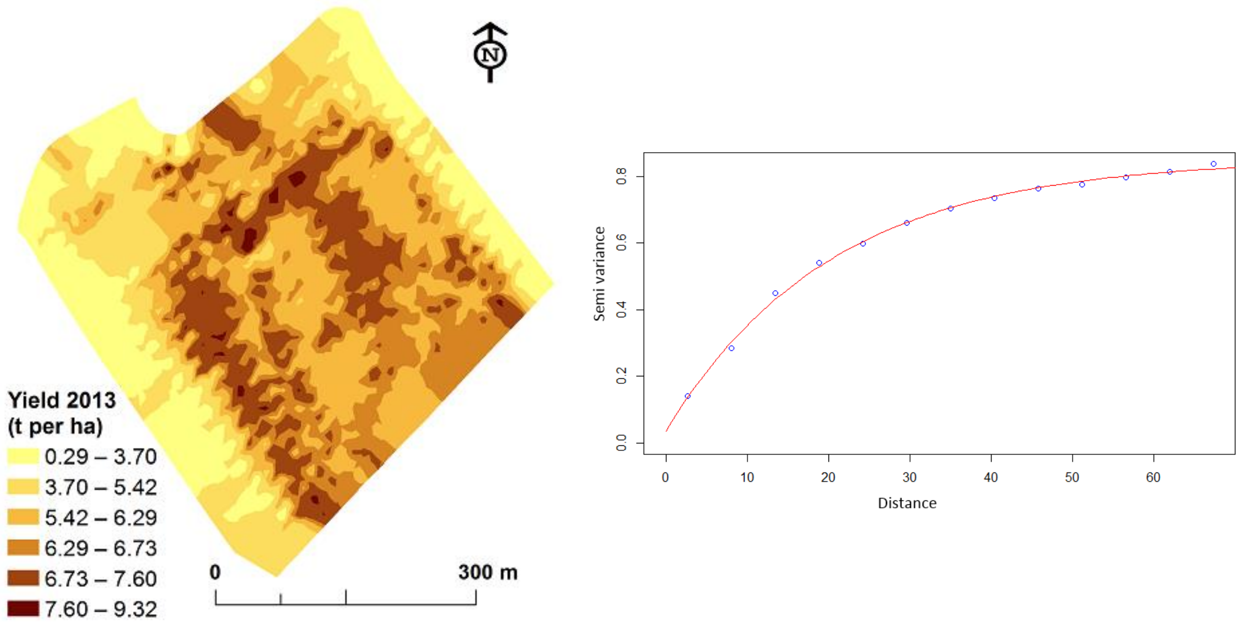


Figure 2. Examples of the raw on-line soil visible and near infrared (vis-NIR) spectra, collected with the on-line sensor. Showing slight deviations in relative absorbance, across the wavelengths, which is dependent on the soil properties.



(a) (b)

Figure 3. Interpolated yield map (a) and exponential semi-variogram of 0.036, 0.817 and 20.358, representing, nugget, sill and range, respectively (b) based on the 2013 harvest of wheat grain in tons per hectare. Lighter areas representing lower yield.

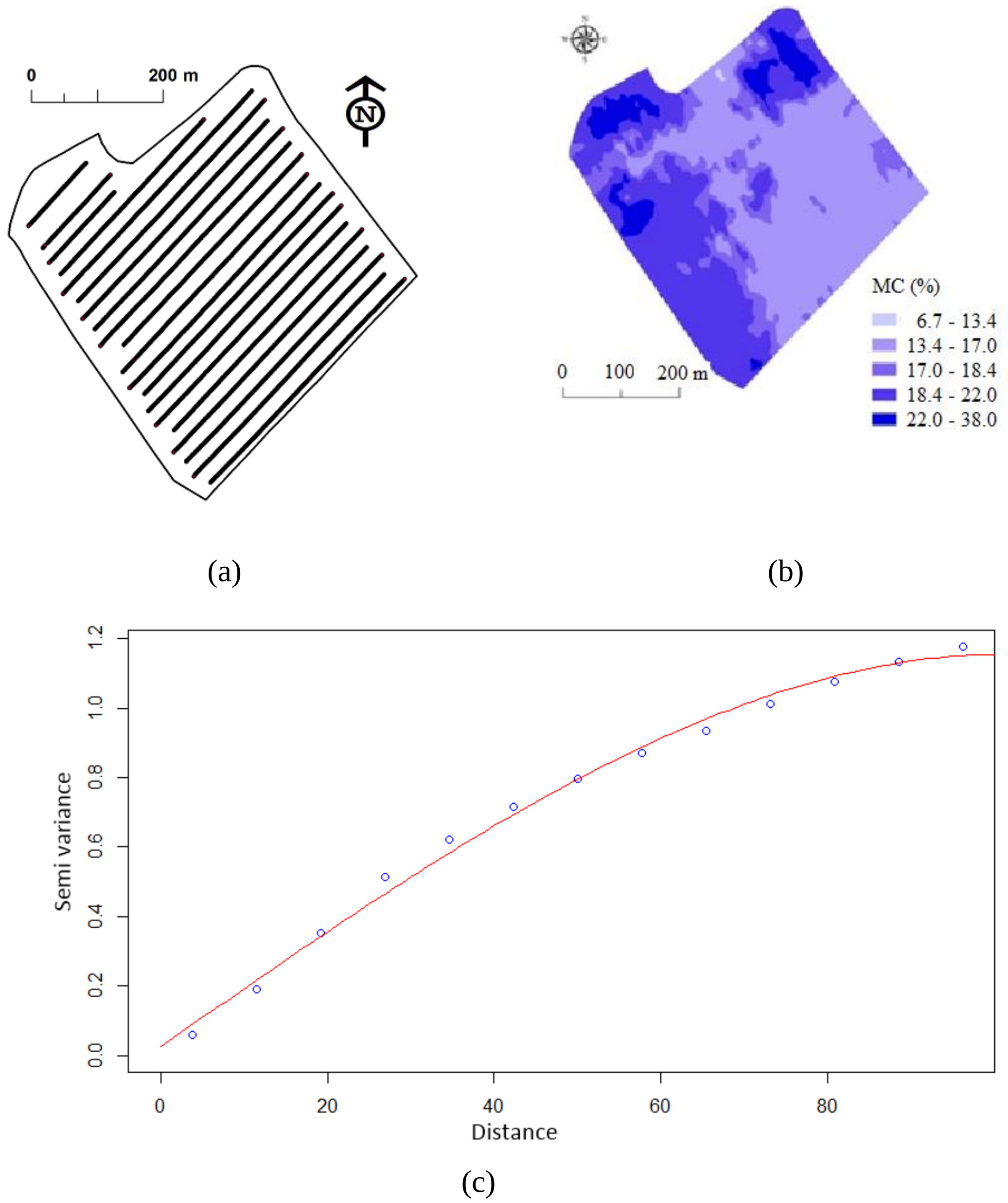


Figure 4. Measured transects (a), map of the soil moisture content (MC) measured with the on-line visible and near infrared spectroscopy (vis-NIRS) sensor after crop harvest in August, 2012 (b), and the spherical semi-variogram used for krigging of MC map with nugget, sill and range values of 0.036, 0.817 and 20.358, respectively.

Table 1: Pearson correlation (*r*) between on-line measured soil properties in 2012 and wheat yield harvested in 2013.

	Ca	CEC	MC	Mg	OC	P	pH	TN	Yield
Ca	1.000								
CEC	**0.733**	1.000							
MC	**0.519**	**0.748**	1.000						
Mg	**0.628**	**0.586**	0.476	1.000					
OC	**0.650**	0.441	0.436	0.176	1.000				
P	0.163	0.216	0.019	0.042	0.027	1.000			
pH	**0.747**	**0.590**	0.492	0.348	0.432	-0.013	1.000		
TN	**0.596**	0.411	0.269	0.167	**0.543**	**0.556**	0.307	1.000	
Yield	-0.321	-0.349	-0.209	-0.320	-0.199	-0.000	-0.152	-0.057	1.000

OC is organic carbon in %; P is extractable phosphorous in mg/l; MC is moisture content in %; TN is total nitrogen in %, CEC is cation exchange capacity in meq/100g; Ca is calcium in mg/l; Mg is magnesium in mg/l; and pH the log measurement of acidity.

Table 2: The correspondence between inputs variables in Volterra Non-linear Regressive with eXogenous inputs (VNRX) model and soil properties

Input	Property	Input	Property	Input	Property	Input	Property
u_1	Ca	u_2	CEC	u_3	MC	u_4	Mg
u_5	OC	u_6	P	u_7	pH	u_8	TN

OC is organic carbon in %; P is extractable phosphorous in mg/l; MC is moisture content in %; TN is total nitrogen in %, CEC is cation exchange capacity in meq/100g; Ca is calcium in mg/l; Mg is magneium in mg/l; and pH the log measurement of acidity.

Table 3: The first ten terms with corresponding error reduction ratio contribution (ERRC) values and coefficients based on the shortest distance approximation (SDA) re-sampling technique with a three m radius

Rank	Term	ERRC	Coefficient θ_i
1	CEC	15.68%	-0.0948
2	Mg	3.57%	-0.4840
3	Ca*CEC	1.13%	-0.0025
4	MC*Mg	0.72%	-0.0558
5	OC	0.78%	-0.2056
6	Mg*P	0.34%	-0.9615
7	Mg*TN	0.78%	5.0750
8	pH*pH	0.39%	-0.0670
9	constant	0.82%	0.1917
10	TN*TN	0.37%	-8.5096

OC is organic carbon in %; P is extractable phosphorous in mg/l; MC is moisture content in %; TN is total nitrogen in %, CEC is cation exchange capacity in meq/100g; Ca is calcium in mg/l; Mg is magnesium in mg/l; and pH the log measurement of acidity.

Table 4: Error reduction ratio contribution (ERRC) contribution of each soil property (input) to the crop yield (system output) with corresponding significance threshold based on the shortest distance approximation (SDA) re-sampling technique with a three m radius

Rank	Input	ERRC (%)	Significance threshold (%)	Significant
1	CEC	14.60	0.60	Yes
2	Mg	4.69	0.52	Yes
3	TN	1.00	0.50	Yes
4	Ca	0.98	0.43	Yes
5	OC	0.68	0.49	Yes
6	MC	0.62	0.47	Yes
7	pH	0.34	0.46	No
8	P	0.30	0.56	No
	Total	23.21	4.03	

OC is organic carbon in %; P is extractable phosphorous in mg/l; MC is moisture content in %; TN is total nitrogen in %, CEC is cation exchange capacity in meq/100g; Ca is calcium in mg/l; Mg is magnesium in mg/l; and pH the log measurement of acidity.

Table 5: Contribution of the top three significant soil properties in terms of the sum of error reduction ratio (SERR) on the crop yield, based on shortest distance approximation (SDA) and (CAA) sampling techniques calculated for different radius values.

Re-sampling technique	Re-sampling radius	Sampled number	Top three inputs		Total Contribution (SERR)
			Inputs	Contribution	
SDA	3	1377	CEC	14.60%	20.29%
			Mg	4.69%	
			TN	1.00%	
CAA	3	1377	CEC	16.54%	22.97%
			Mg	4.00%	
			TN	2.43%	
SDA	5	3605	CEC	9.20%	13.61%
			Mg	2.45%	
			TN	1.96%	
CAA	5	3605	CEC	12.90%	15.87%
			Mg	3.02%	
			Ca	2.65%	

TN is total nitrogen in %, CEC is cation exchange capacity in meq/100g; Mg is magnesium in mg/l.