

Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques

Said Nawar^{a,*}, Abdul M. Mouazen^{a,b}

^aCranfield Soil and AgriFood Institute, School of water, Energy and Environment, Cranfield University, Cranfield, MK43 0AL, UK

^bDepartment of Soil Management, Ghent University, Coupure 653, 9000 Gent, Belgium

*correspondence; E-Mail: s.m.nawar@cranfield.ac.uk;
Tel.: +44 (0) 1234 750111 (ext 8334); Fax +44 (0) 1234 752 971.

Abstract

The development of accurate visible and near infrared (vis-NIR) spectroscopy calibration models for selected soil properties based on mobile measurements is essential for site specific soil management at fine sampling scale. The objective of the present study was to compare the mobile and laboratory prediction performance of vis-NIR spectroscopy for total nitrogen (TN), total carbon (TC) and soil moisture content (MC) of field soil samples based on single field (SFD), two-field dataset (TFD), UK national dataset (UND) and European continental dataset (ECD) calibration models developed with linear and nonlinear data mining techniques including spiking. Fresh soil samples collected from fields in the UK, Czech Republic, Germany, Denmark and the Netherlands were scanned with a fibre-type vis-NIR spectrophotometer (tec5 Technology for Spectroscopy, Germany), with a spectral range of 305-2200 nm. After dividing spectra into calibration (75%) and validation (25%) sets, spectra in the calibration set were subjected to three multivariate calibration models, including the partial least squares regression (PLSR), multivariate adaptive regression splines (MARS) and support vector machines (SVM), with leave-one-out cross-validation to establish calibration models of TN, TC and MC. Results showed that the best model performance in cross-validation was obtained with MARS methods

for the majority of dataset scales used, whereas the lowest model performance was obtained with the SFD. The effect of spiking was significant and the best model performance in general term was obtained when local samples collected from two target fields in the UK were spiked with the ECD, with coefficients of determination (R^2) values of 0.96, 0.98 and 0.93, root mean square error (RMSE) of 0.01, 0.1 and 1.75, and ratio of performance to interquartile distance (RPIQ) of 7.46, 6.57 and 3.98, for TC, TN and MC, respectively. Therefore, these results suggest that ECD vis-NIR MARS calibration models can be successfully used to predict TN, TC and MC under both laboratory and mobile measurement conditions.

Keywords: Diffuse reflectance spectroscopy, Spectral library, Soil properties, Data mining, Spiking.

Introduction

Visible and near-infrared (vis-NIR) diffuse reflectance spectroscopy has attracted increasing interest among soil scientists in recent times, and has been proposed as a possible method of soil analysis. It provides higher soil sampling density for mapping purposes compared with conventional laboratory analysis (Shepherd and Walsh, 2002; Wetterlind et al., 2010). This technique also allows for in field (in situ) non mobile (Viscarra Rossel and Chen, 2011; Brodský et al., 2013) and mobile measurement with high soil-sampling resolution (Maleki et al., 2008; Kuang and Mouazen, 2013). Literature show that Vis-NIR spectroscopy has been used successfully for modelling and mapping of soil properties, under both mobile and non-mobile measurement conditions, (i.e., Shibusawa et al., 2001; Mouazen et al., 2005; Kuang and Mouazen, 2013; Kuang et al., 2015). However, compared to laboratory spectral measurements that is done under controlled conditions, mobile spectroscopy analyses are affected by

environmental factors such as ambient light, soil moisture content, soil structure, temperature, contamination by stones and excessive residues (Mouazen et al., 2007; Stenberg et al., 2010). One way to overcome these negative influences is by adopting effective spectral data preprocessing and advanced data mining techniques when developing calibration models for mobile spectroscopy.

Precision agriculture aims at optimizing management of within field variability for sustainable increase in land productivity (Bongiovanni and Lowenberg-Deboer, 2004). Variable-rate fertilizer application, which requires reliable soil information at a high spatial resolution, is required to achieve this goal (Wetterlind et al., 2010; Mouazen and Kuang, 2016). Successful implementation of the mobile vis-NIR spectroscopy was reported for sensor based and map-based variable rate phosphorus fertilization (Maleki et al., 2008; Mouazen and Kuang, 2016) and variable rate N-fertilizer application (Halcro et al., 2013). However, accurate recommendation maps for variable rate applications will depend on accurate prediction of soil properties with vis-NIR spectroscopy, which can be achieved by minimizing error in reference and spectra measurements, minimizing influences of ambient conditions and finally by adopting effective spectra pre-processing and advanced data mining techniques.

In order to gain the full advantage of the use of vis-NIR, different techniques have been applied to the development of calibration models, such as spiking of samples from target site or dataset into existing spectral library (Brown, 2007; Sankey et al., 2008; Wetterlind et al., 2010; Kuang and Mouazen, 2011, 2013; Guerrero et al., 2014). The geographical scale of soil samples collected was reported to have influence on model performance (Sudduth and Hummel 1996). This will be especially evident when predicting variations on a small scale (Brown, 2007). Combining global and local samples by adding a few local ones to a more general soil library

(spiking) and recalibrating was proposed by Brown (2007) as another way to increase the accuracy of soil organic carbon (SOC) prediction, as opposed to local-sample calibrations. Sankey et al. (2008) also reported improved prediction results for clay content, SOC and inorganic carbon, using the same global calibration set spiked with local samples from three highly variable landscape study sites in Montana, US, compared with global or local calibrations alone. Kuang and Mouazen (2013) used spiking technique with different dataset ratios for mobile vis-NIR modelling at European field scale. They achieved good results using partial least square regression (PLSR) models for soil moisture content (MC), SOC, and total nitrogen (TN) with residual prediction deviation (RPD) (calculated as standard deviation of measured soil properties divided by root mean square error of prediction) of 2.76 to 3.96, 1.88 to 2.38, and 1.96 to 2.52, respectively. Limited works have been reported on the combined effect of spiking and samples scale on model predictive performance in soil analysis (Sankey et al., 2008; Guerrero et al., 2010; Wetterlind and Stenberg, 2010; Guerrero et al., 2014), particularly for mobile collected vis-NIR spectral data, where no reports could be found in the literature.

As a linear multivariate analysis, PLSR is the most commonly used technique for soil spectral analysis (Conforti et al., 2013, 2015). However, the accuracy of linear-regression techniques in spectroscopic analysis tends to decrease due to the non-linear nature of the relationship between spectral data and the dependent variable (Araújo et al., 2014). Data-mining techniques, such as artificial neural network (ANN) (Mouazen et al., 2010; Kuang et al., 2015), multivariate adaptive regression splines (MARS) (Bilgili et al., 2010; Nawar et al., 2015) and support-vector machines (SVM) (Morellos et al., 2016; Nawar et al. 2016), were reported to improve the accuracy of the calibration models. As a nonlinear method based on the machine learning theory, SVM was proposed by Vapnik (1998) to be capable of modelling linear and nonlinear relationships and

solving calibration problems with high performance (Suykens and Vandewalle, 1999). SVM has been used successfully for modelling soil properties based on reflectance spectroscopy (e.g. Viscarra Rossel and Behrens, 2010; Vohland et al., 2011; Peng et al., 2014; Nawar et al., 2016), and has gained extensive application in soil spectroscopy, because of its advantages and high performance (Viscarra Rossel and Behrens, 2010; Vohland et al., 2011). As nonparametric method MARS estimates complex nonlinear relationships among independent and dependent variables (Friedman, 1991), and has been effectively applied in different fields (Luoto and Hjort, 2005; Bilgili et al., 2010; Felicísimo et al., 2012; Samui, 2012) and generally exhibits higher performance results for modelling soil properties (e.g. Shepherd and Walsh, 2002; Bilgili et al., 2010; Nawar et al., 2016). However, these linear and non-linear modelling methods were not compared in the literature for soil analyses at different geographical scales including spiking. This is particularly true for modelling of mobile collected vis-NIR soil spectra.

The aim of this paper was to compare the predictive performance of the vis-NIR spectroscopy of TN, TC and MC for field samples based on single field, two-field, national and continental scales, using PLSR and two multivariate data-mining techniques, namely, SVM and MARS. The calibrations were made using laboratory and mobile collected soil spectra for predicting within-field variation in named soil properties.

2. Materials and methods

2.1. Experimental sites

The experimental work was carried out in two fields in Yorkshire, UK. Hagg field is located at Cawood, north Yorkshire with longitudes of -1.172° and -1.166° W, and latitudes of 53.936° and 53.941° N with total area of about 21 ha (Fig. 1). This field is characterised by fosters cooler

summer conditions with regular rainfall (average annual 600 mm), and the mean air temperatures range from 1°C to 6°C in winter and from 8°C to 18°C in summer, allowing more consistent crop growth. The field is cultivated with vegetables crops (leeks, cabbage, carrots and onions). The soil type is a slightly acidic sandy loam with impeded drainage to the south and eastern margins of the field. The soil organic matter (SOM) ranges between 0.7 and 1.95% at 0–30 cm depth. Hessleskew field is located in Sancton, north Yorkshire between longitudes -0.590° and -0.586° W, and latitudes 53.844° and 53.844° N and with total area of about 12 ha (Fig. 1). Elevations in the area barely reach 212 m above sea level. The annual rainfall and air temperature are in the same range of the Hagg field. The field is characterised by freely draining with texture varies between clay loam to clay. The SOM ranges between 0.9 and 2.1%, and it is cultivated with cereal crops in rotation.

2.1. Mobile soil measurement and collection of soil samples

The mobile measurement system designed and developed by Mouazen (2006) was used to measure both fields. It consists of a subsoiler, which penetrates the soil to the required depth, making a trench, whose bottom is smoothened by the downwards forces acting on the subsoiler (Mouazen et al., 2005). The subsoiler was retrofitted with the optical unit and attached to a frame. This was mounted onto the three point linkage of the tractor (Mouazen et al., 2005). 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. The sampling interval of the instrument was 1 nm. A differential global positioning system (DGPS) (EZ-Guide 250, Trimble, USA) was used to record the position of mobile measured spectra with sub-meter accuracy. Each scan consisted of three spectra that were averaged in one spectrum representing about 1.2 m travel distance. Outlier spectra were visually

detected and removed from further analyses. These outlier spectra represented 2.5 and 4% out of 12766 and 14276 soil spectra collected from Hessleskew and Hagg fields, respectively. During the measurement at each line, 3 or 4 soil samples were collected from the bottom of the trench and the sampling positions were carefully recorded with a DGPS. The locations of soil samples were selected to cover the soil variation within both fields. A total of 122 and 149 soil samples were collected during the mobile measurement from Hessleskew and Hagg fields, respectively (Fig. 1). Measurements were carried out in 2015 and 2016 for the former and latter fields, respectively, pulling the sensor at 12 m gap between adjacent transects (Fig. 1).

(Fig. 1)

2.2. Laboratory chemical and optical measurements

Each sample was divided into two parts; one part was dried for 24 h at 105 °C and the other part was left fresh (wet). The dried soil sample was analysed for total carbon (TC) according to the British Standard BS 7755 Section 3.8:1995 using combustion method, which is identical to ISO 10694:1995. Total nitrogen was determined by the Dumas method, where the soil samples are heated to 900°C in the presence of oxygen gas as described by British Standard BS EN 13654-2:2001. Soil MC was determined by oven drying of the soil samples at 105 °C for 24 h. The precision of standard laboratory analyses for TN, TC, and MC indicated that uncertainties values (mean values and standard deviations) were $0.35 \pm 0.025\%$, $3.8 \pm 0.26\%$, and $20 \pm 1.22\%$, respectively.

The fresh part of soil sample was placed in a glass container and mixed well. Three Petri dishes of 2 cm in diameter and 2 cm deep were used for three replicate measurements. Each soil sample was then placed into these Petri dishes and pressed gently before levelling with a spatula to

ensure a smooth surface; and therefore maximum light reflection and a large signal-to-noise ratio (Mouazen et al., 2005). Soil samples were scanned by the same spectrometer used in the mobile measurements. A total of ten scans were collected from each replicate, and these were averaged into one spectrum for each sample.

2.3. Spectra pretreatment

The same pretreatment of soil spectral data was carried out for all soil properties investigated using R packages (prospectr; <https://cran.r-project.org/web/packages/prospectr>). First, noise was removed at both edges of each spectrum and the spectra were cut to 370–1979 nm. Then, the number of wavelengths was reduced by averaging five successive wavelengths. Maximum normalization was followed, which is typically used to get all data to approximately the same scale, or to get a more even distribution of the variances and the average values. Spectra were then subjected to Savitzky–Golay first derivation (Martens and Naes, 1989) with a second-order polynomial approximation. Finally, the Savitzky-Golay smoothing was carried out to remove noise from spectra and to decrease the detrimental effect on the signal-to-noise ratio that conventional finite-difference derivatives would have. Spectra modification due to the different pretreatments followed in this study is shown in Fig. 2.

(Fig. 2.)

2.4. Spectral dataset with different scale

The following four different data sets were considered:

1- Single-field dataset (SFD; n=122 for Hessleskew and 149 for Hagg), where samples from one field either Hessleskew or Hagg fields were used.

2- Two-field dataset (TFD; n=271), where samples from both fields were merged into one

dataset;

3- UK national dataset (UND; n=89), where samples collected from four fields in the UK were used; and

4- European continental dataset (ECD; n=529), where samples from sixteen fields collected from five European countries were used. These included samples from Germany (two fields), Denmark (five fields), the Netherlands (one field), Czech Republic (four fields) and the UK (four fields), where 151, 147, 43, 99, and 89 samples were collected, respectively (Kuang and Mouazen, 2011, 2013).

Spiking was used to introduce the local variability of the two experimental fields into the existing data sets (e.g., UND and ECD). A total of 85 and 110 samples were spiked from the Hessleskew and Hagg fields, respectively, into the UND and ECD. After spiking, the total number of samples for UND and ECD used to develop calibration models were 174 and 614, and 199 and 639, for Hessleskew and Hagg fields, respectively. A principal component analysis (PCA), was performed on 724 samples of the ECD including the two experimental fields of the current studies to explore similarity between different groups of soil samples.

2.5. Development of calibration models

The four datasets discussed above were subjected to PLS, MARS and SVM analyses with the leave-one-out cross validation using R software (R Core Team, 2013), which resulted in four groups of models for each modelling technique for each soil property, e.g., for SFD, TFD, UND and ECD. Before running the analysis the entire dataset of each target field (Hesseleskew or Hagg) was divided into 75% for calibration, and 25% for prediction. This was done for both the laboratory and mobile collected soil spectra. The 75% soil samples were also spiked into the

207 UND and ECD. The same prediction samples (25%) were used for both laboratory and mobile
208 validation of obtained results of the four models for each soil property.

209 PLSR is a popular multivariate regression method that has a good capacity for estimating
210 attributes resulting from the spectral characteristics of the soil (Vasques et al., 2008; Song et al.,
211 2013; Yu et al., 2016). To determine the optimal number of latent factors, leave-one-out cross-
212 validation (LOOCV) was used (Efron and Tibshirani, 1993) to prevent over- or under-fitting the
213 data, which may produce models with poor performance. The root mean squared error of cross
214 validation (RMSECV) of the predictions and the coefficient of determination (R^2) were
215 ascertained to identify the optimal cross-validated calibration model. Generally, a model with the
216 highest cross-validated R^2 value and lowest RMSECV value was selected.

217 SVM is a kernel-based learning method originated from statistical learning theory (Vapnik
218 1995). Kernel-based learning methods use an implicit mapping of the input data into a high
219 dimensional feature space defined by a kernel function (Karatzoglou and Feinerer, 2010). The ε -
220 SVM-Regression (ε -SVMR) uses training data to obtain a model represented as a so-called ε -
221 insensitive loss function, which maps independent data with maximum ε deviation from
222 dependent training data. Error within the predetermined distance ε from the true value is ignored,
223 error greater than ε is penalised (Vohland et al. 2011). In consequence, the model reduces the
224 complexity of the training data to a significant subset of so-called support vectors. In the current
225 study, the SVM models were developed using libsvm algorithm with ε -SVM and radial basis
226 function (RBF) kernel using e1071 package in R (Meyer et al., 2015). The optimal parameters of
227 SVM (cost, epsilon, $\text{gam}(\gamma)$ and $\text{sig2}(\sigma^2)$) were adopted to regulate the models. For each
228 combination of cost, $\text{gam}(\gamma)$ and $\text{sig2}(\sigma^2)$ parameters, the root mean square error of cross-
229 validation (RMSECV) was calculated, and the optimum parameters were selected to produce

smaller RMSECV. The leave-one-out cross validation (LOOCV) was used for the validation process.

MARS is data mining technique developed by Friedman (1991). MARS was applied as a regression method in several disciplines (Shepherd and Walsh, 2002; Bilgili et al., 2010; Nawar et al., 2015) and was generally reported to show better performance than others traditional statistical methods for soil analysis based on vis-NIR spectra (Nawar et al., 2016). The MARS analysis uses basis functions to model the predictor and response variables (Hastie et al., 2009). To construct the basis functions, MARS splits the data into sub-regions (splines) with different interval ending knots where the regression coefficients change and fits the data in each sub-region by using a set of adaptive piecewise linear regressions. These basis functions are subsequently used as new predictor variables for modelling purposes. Each basis function may contain nonlinear and variable interaction factors of the second and third order as well as linear combinations. To measure lack of fit, avoid over-fitting, and improve prediction, the redundant basis functions are removed one at a time using a backward stepwise procedure. Additional basis functions in the model provide greater flexibility but also add complexity. However, MARS attempts to minimise model complexity by selecting the optimal model based on the lowest generalised cross-validation value.

In order to evaluate the model performance for the prediction of named soil properties three statistical parameters were used: R^2 , root mean square error (RMSE) and the ratio of the performance to interquartile distance (RPIQ) (Bellon-Maurel et al., 2010) defined as fellow:

$$RPIQ = \frac{IQ}{RMSE}$$

where IQ is the difference between the third and first quartiles ($IQ = Q3 - Q1$).

Arbitrary groups have been used for simplification of interpretation: (1) excellent models ($RPIQ > 2.5$), (2) very good models ($2.5 > RPIQ > 2.0$), (3) good model ($2.0 > RPIQ > 1.7$), fair ($RPIQ > 1.7 > RPIQ > 1.4$), and very poor model ($RPIQ < 1.4$). This classification was adopted in this study to compare between different models in cross-validation and in laboratory and mobile prediction.

3. Results

3.1 Laboratory measured soil properties and spectral datasets

The results of the descriptive statistical analyses for soil parameters are shown in Table 1. For Hessleskew field, the TN is low, with a mean value of 0.25%, whereas the mean values of TC and MC ranges from 2.121% to 18.9%, respectively. In the Hagg field, the TN is also low, with a mean value of 19.21%. TC is smaller than that of Hessleskew field, ranging between 1.34% and 3.18%, whereas MC was higher than that in Hessleskew, ranging between 11.53% and 24.64% (Table 1). The small range of laboratory analysis values of the studied soil properties indicates these fields are not the ideal case study fields. This small soil variability within both fields may be reduced the prediction capability of the established calibration models (Kuang and Mouazen, 2011).

(Table 1)

The PC similarity maps of principal component (PC) 1 and PC2 shows clear separation between different groups of samples collected from the five European countries. A clear overlap of the two study field spectra with those of the UND collected can be observed (Fig. 3), explain the effect of geographical origin on soil samples on vis-NIR spectroscopy analysis for soil properties.

(Fig. 3)

3.2. Performance of calibration models in cross-validation

The calibration methods considered in this study provide different prediction accuracies of TN, TC and MC. Tables 2 and 3 summarise the cross-validation results for calibration models developed with different datasets. Figures (4 and 5) illustrate the performance of prediction models in terms of RPIQ for Hessleskew and Hagg field, respectively. Cross-validation results indicate that MARS has over-performed SVM and PLSR for modelling three studied soil properties in cross-validation, followed, successively by SVM and PLSR, which is the least performing method. Among MARS models, the best results for TN is obtained with ECD with $R^2 = 0.96$, RMSECV = 0.01%, and RPIQ = 7.46 for Hagg field (Table 3), and $R^2 = 0.96$, RMSECV = 0.01%, and RPIQ = 7.41 for Hessleskew field (Table 2). The best results for TC are achieved using a MARS model developed with the ECD in both Hagg field ($R^2 = 0.98$, RMSECV = 0.10% and RPIQ = 6.57) and Hesselskew field ($R^2 = 0.98$, RMSECV = 0.06% and RPIQ = 4.87). Similar model performance to those of TN and TC is also observed for MC, with the best results in cross-validation were calculated for the MARS models developed with the TFD (not with ECD as for TN and TC) for the Hagg field ($R^2 = 0.93$, RMSE = 0.53% and RPIQ = 5.15) and Hessleskew field ($R^2 = 0.92$, RMSECV = 0.53% and RPIQ = 5.11).

3.3. Performance of calibration models for laboratory prediction

The developed calibration models were validated using laboratory scanned spectra of the prediction sets. Tables 2 and 3 summarise the accuracy of the laboratory prediction for the studied soil properties at the different calibration scales. The cross-validation results show considerably better results than those obtained for laboratory predictions. Among the studied modelling techniques, MARS models generally provide the best results for the studied soil

properties. For TN, the best performance is obtained with MARS using the ECD dataset with $R^2 = 0.87$, RMSE = 0.03%, and RPIQ = 5.21 for Hessleskew field (Table 2; Fig. 4), and $R^2 = 0.77$, RMSE = 0.03%, and RPIQ = 3.57 for Hagg field (Table 3; Fig. 5). PLSR-SFD based modelling has resulted in the least appreciable results for TN in Hessleskew field ($R^2 = 0.50$, RMSE = 0.01%, and RPIQ = 1.21), followed by SVM with SFD ($R^2 = 0.51$, RMSE = 0.01%, and RPIQ = 1.41, as shown in Table 2). However, SVM showed better performances for predicting TN for TFD, UND and ECD, compared to PLSR using the corresponding datasets. For Hagg field, the worst results are obtained with PLSR-SFD based modelling ($R^2 = 0.54$, RMSE = 0.02%, and RPIQ = 2.36, Table 3). Again SVM over-performed PLSR models, but under-performed MARS models (Tables 2 and 3).

The best results for TC laboratory prediction are achieved using the MARS model based on the ECD ($R^2 = 0.88$, RMSE = 0.19% and RPIQ = 5.94) for Hessleskew field, as well as for Hagg field ($R^2 = 0.85$, RMSE = 0.46% and RPIQ = 4.34). The lowest prediction performance is obtained with PLSR models particularly for the TFD in Hagg field and Hessleskew field with R^2 , RMSE, and RPIQ of 0.52 and 0.57, 0.24% and 0.23, and 1.95 and 1.98, respectively (Tables 2 and 3). SVM models show the same trend as for the cross-validation where it outperform PLSR for all datasets in both fields, and the best results are obtained with TFD in Hagg field ($R^2 = 0.89$, RMSE = 0.14% and RPIQ = 3.26) and with ECD in Hessleskew field ($R^2 = 0.79$, RMSE = 0.25% and RPIQ = 3.19).

Results for MC show a similar trend to those of TN and TC, as the best performing prediction is obtained with the MARS models based on the TFD (not with the ECD, as for TC and TN) in both fields. The worst performing models are for PLSR with TFD in Hessleskew field ($R^2 = 0.48$, RMSE = 1.38% and RPIQ = 1.69) and ECD in Hagg field ($R^2 = 0.57$, RMSE = 1.72% and

RPIQ = 2.26). SVM generates acceptable results, with the best model accuracy obtained with ECD in Hagg field ($R^2 = 0.77$, RMSE = 1.72% and RPIQ = 3.0), and the lowest accuracy obtained with SFD in Hessleskew field ($R^2 = 0.53$, RMSE = 1.32% and RPIQ = 0.90) (Table 2). Similar trend to that of the cross-validation can be concluded for the laboratory prediction. For example, MARS has over-performed SVM and PLSR for the three studied soil properties followed, respectively by SVM and PLSR, which is the least performing method. The only exception is for MC, where depending on the dataset, SVM over-performed PLSR in two cases only out of four. Here, it is not necessary the case that the ECD models outperformed corresponding models developed with the other three datasets.

3.4. Performance of calibration models for mobile prediction

The mobile collected spectra were used to predict soil TN, TC and MC using the calibration models developed in advance, as explained above. The laboratory reference measurement values were compared with the mobile predicted concentration values at the same positions. Tables 2 and 3 summarise the accuracy of the mobile measurement for the studied soil properties at the different-scale datasets.

For TN, the best results are obtained with MARS, followed by SVM and PLSR, respectively, which is similar for the trends in cross-validation and laboratory prediction. The highest MARS model performance is resulted with the ECD dataset with $R^2 = 0.79$, RMSE = 0.02%, and RPIQ = 3.26 for Hessleskew field (Table 2; Fig. 4), and $R^2 = 0.73$, RMSE = 0.03%, and RPIQ = 3.16 for Hagg field (Table 3; Fig. 5). The least acceptable results are produced by SVM using SFD of Hessleskew field ($R^2 = 0.48$, RMSE = 0.02%, and RPIQ = 1.17), followed by PLSR-SFD for the Hagg field ($R^2 = 0.49$, RMSE = 0.03%, and RPIQ = 1.96, Table 3). However, SVM show a good

performance for predicting TN with TFD and less evident with UND, for both fields, outperforming PLSR for both datasets.

Like for TN, the best results for TC are achieved using the MARS-ECD model for Hagg field ($R^2 = 0.81$, RMSE = 0.29% and RPIQ = 4.33) and Hessleskew field ($R^2 = 0.81$, RMSE = 0.31% and RPIQ = 3.57). The mobile prediction of PLSR models are of lower degree of accuracy particularly for SFD of Hagg field with R^2 , RMSE, and RPIQ of 0.50, 0.23%, and 1.97, respectively, which is comparable to the results of SVM-SFD (Table 3). The best results for SVM are obtained with ECD, which is a similar trend to that in the cross-validation modelling. Generally MARS models provided considerably better performance in mobile prediction than both SVM and PLSR, which performed almost equally. The one exception is when the UND is used, with which both PLSR and SVM performed better than TC.

The mobile prediction of MC follows the same trend of the cross-validation, as the best results are obtained with MARS for seven out of eight cases in both fields (Tables 2 and 3). The best mobile prediction performance is resulted from the MARS-ECD modelling in Hessleskew field ($R^2 = 0.73$, RMSE = 1.91% and RPIQD = 1.94) and MARS-UND in Hagg field ($R^2 = 0.72$, RMSE = 1.91% and RPIQ = 2.58). The results of PLSR-TFD model of Hessleskew field are the least accurate over all models ($R^2 = 0.48$, RMSE = 1.36% and RPIQ = 1.58). Depending on the data set used, SVM generated slightly better results than PLSR with RPIQ values ranging between 1.76 and 2.19 for Hagg field (Fig. 5), and 1.32 and 2.66 for the Hessleskew field (Fig. 4).

In the majority of models, MARS has over-performed SVM and PLSR for the three studied soil properties in mobile prediction, followed, respectively by SVM and PLSR, which is the least

performing method. The only exception is for MC, where depending on the dataset, SVM over-performed PLSR in three cases out of four.

(Table 2)

(Table 3)

3.5. Analysis of uncertainty versus RMSE

To investigate the accuracy and quality of results obtained with laboratory prediction, we compared the RMSE with uncertainty calculated for standard laboratory chemical analyses (Figs. 6 and 7). The majority of models result in smaller RMSE than the uncertainty ($0.35 \pm 0.025\%$, $3.8 \pm 0.26\%$, and $20 \pm 1.22\%$ for TN, TC, and MC, respectively). This is indeed the case in for Hessleskew field, particularly for TN and TC with MARS and SVM (Fig. 6), whereas RMSE becomes larger than the uncertainty level for MC when the three modelling techniques are applied for the UND and ECD datasets. Similarly in Hagg field, RMSE is larger than the uncertainty level for not only MC but also TN and TC, for both the UND and ECD datasets (Fig. 7).

(Fig. 4)

(Fig. 5)

4. Discussion

4.1. Comparison performance of modelling techniques

This study compared the performance of the MARS, SVM and PLSR models for the prediction of TC, TN and MC based on different datasets. The variations in RPIQ and RMSE values

obtained from models in cross-validation, laboratory and mobile prediction are shown in Tables 2 and 3. MARS models have over-performed SVM and PLSR for modelling the three studied soil properties in cross-validation, laboratory and mobile prediction. Examining the values of RPIQ in cross-validation, laboratory and online prediction, respectively, of MARS models allows confirming that the predictive performance of the MARS model in this study was high. This result is in line with results reported by Nawar et al. (2016) for non-mobile measurement, who found that MARS performed SVM and PLSR. Based on the adapted RPIQ groups, models in laboratory prediction show excellent accuracy for TN and TC, but only good for MC, whereas the accuracy of mobile prediction are excellent, very good and good for TN, TC and MC, respectively. Laboratory and mobile predictions of TN and TC are better than those obtained by other researchers for PLSR models (e.g. Kuang and Mouazen, 2011, 2013) ($R^2 = 0.75\text{--}0.93$ and $0.74\text{--}0.89$; $RMSE = 0.03\text{--}0.06\%$ and $0.7\text{--}1.1\%$, for TN, and TC, respectively), while for MC the results of Kuang and Mouazen (2013) were better ($RMSE = 0.32\text{--}0.74\%$ and $0.36\text{--}0.84\%$ for laboratory and online predictions, respectively. This is because the MARS modelling technique typically yields better results when a nonlinear relationship between reflectance and concentration exists, whereas the PLSR model fits only linear relationships (Brown et al., 2006; Mouazen et al., 2010; Viscarra Rossel and Behrens, 2010; Nawar et al., 2016).

Generally, after MARS SVM was the best performing method in cross-validation for the three studied soil properties. This was not always the case for laboratory and mobile prediction, where depending on the dataset PLSR overperformed SVM and even MARS (Tables 2 and 3). These results are supported by the findings of Viscarra Rossel and Behrens (2010), who showed that SVM provided more robust predictions of soil organic carbon than PLSR models. Similarly, and Nawar et al. (2016) also showed that the non-linear nature of SVM led to a better results than

PLSR. Morellos et al. (2016) also found SVM to overperform two linear methods, namely, PLSR and principal component regression for modelling OC, TN and MC. This is because SVM is a non-linear and flexible method, capable to model complex, non-linear and linear relationships between variables (Viscarra Rossel and Behrens 2010). However, in the current work we prove this to be the case for mobile prediction too.

Although the results of PLSR models showed the weakest performance among the three models in cross-validation, SVM and PLSR performed equally in the laboratory and mobile prediction depending on the data set. This was especially the case with the Hagg field. At the Hessleskew field, the variation in concentrations of studied soil properties was slightly higher (Table 1) than that in the Hagg field. This resulted in slightly higher not only RPIQ and R^2 values but also RMSE values, which is in line with findings of Kuang and Mouazen (2011).

The small prediction datasets used in the current work (37 and 39 for Hessleskew and Hagg fields, respectively), might have affected the prediction performance in the laboratory and mobile predictions, as a small data set size was reported to lead to various counterintuitive and unfamiliar side effects, which can significantly impact the validation results and lead to a very poor performance (Klement et al., 2008).

Model performance for TN and TC in cross-validation, laboratory and mobile prediction behaved differently from models for MC. It is interesting and rather odd to observe that TC and TN models to overperform MC model, which is not a typical finding in vis-NIR spectroscopy for soil analyses, as MC used to over-perform all other soil properties models (Kuang and Mouazen, 2013; Mouazen et al., 2010; Morellos et al., 2016). No other reason could be found for the odd MC prediction behaviours except that the current field were of stony soils that might contributed

to the error of MC laboratory measurement with the drying method.

Based on Figs. 6 and 7, we observed that relatively small values of RMSE for TN and TC, having large range of concentrations, indicate unbiased predictions. In addition, the majority of RMSE values lie within the uncertainties levels with an average deviation $< 5\%$ for TN and TC and 6.35% for MC. These deviations are much smaller for MARS compared to PLSR, indicating good prediction performance of MARS models for both the laboratory and mobile predictions.

(Fig. 6)

(Fig. 7)

4.2. Influence of geographical scale on Models' performance

The dataset (scale) has shown to have a considerable influence on the performance of calibration and prediction of TN, TC, and MC. Spiking local samples in the UND and ECD libraries almost always improved the model performance, particularly in cross-validation, compared with those obtained using the SFD and TFD (Tables 2 and 3), which is in agreement with the results presented by Brown (2007) and Sankey et al. (2008), although the current work is based on a smaller scale datasets of a smaller variability. The improvement was mainly expressed as an increase in R^2 and RPIQ values, but a decrease in RMSE (Tables 2 and 3). This finding is in agreement with Brown (2007) findings, who reported a great reduction in RMSE for predictions of SOC in upland soil samples from a catchment in Uganda, by adding local samples to a global library. Furthermore, it is also in line with the findings of Kuang and Mouazen (2013) that spiking local soil samples into ECD datasets proved to be an efficient way to improve the prediction accuracy of target field samples.

Examining the RPIQ values obtained with different dataset suggest that spiking of laboratory scanned spectra into UND and ECD is a successful strategy to obtain accurate mobile predictions of soil properties. Figs. 4 and 5 and Tables (2 and 3) illustrate how spiking coupled with MARS have led to the highest mobile prediction performance compared to PLSR and SVM, particularly at ECD scale, for which RPIQ values were of good to very good values according to the adopted RPIQ groups. Compared to published results of spiking of a ECD with target field samples obtained with PLSR analyses (Kuang and Mouazen, 2013) for TN (RPD = 1.96–2.52), OC (RPD = 1.88–2.38), and MC (RPD = 2.76–3.96), comparable results with slight improvements were obtained in the current work for mobile MARS-ECD prediction, only for TN and TC. Compared with artificial neural network (ANN)-ECD (nonlinear) model performance for the mobile prediction of SOC reported by Kuang et al. (2015), slight improved results can be observed with MARS-ECD obtained in the current work (RPIQ for TC = 3.57 – 4.33), as RPD values reported by Kuang et al. (2015) were improved from 1.93 for PLSR analysis to 2.28, for ANN. Furthermore, the current results outperform those obtained by other researchers (Pietrzykowski and Chodak, 2014) for TN based on non-mobile measurement and PLSR analysis only (RPIQ = 2.34), and combined PLSR with genetic algorithm (RPIQ = 4.12; Ludwig et al., 2017). This is also true for TC based on PLSR (RPIQ = 31.3; Sarkhot et al., 2011) and SVM and PLSR (RPIQ = 2.03 and 2.59, respectively; Wijewardane et al., 2016). This is at least a step forward the current work have shown that there will be always an opportunity to improve the mobile prediction performance by adopting new data mining techniques (e.g., MARS in the current work). Considering the narrow variation range of soil properties in the two measured fields (Table 1), the concept of spiking a general dataset with samples from measured target field seems to be a successful calibration procedure for mobile vis–NIR measurement of soil TN, TC

and MC. The alternative concept to spiking would be to collect a sufficient number of soil samples from a target site (e.g. SFD or TFD) and produce calibration models for mobile prediction of soil properties as proposed by Mouazen et al. (2007). In this concept, one should be lucky to encounter sufficient spatial variability in the target site, to enable establishing calibration models (Guerrero et al., 2010; Wetterlind et al., 2010). This concept is definitely more expensive than the spiking, as a much larger number of soil samples should be collected from a site (> 5 sample per ha, or > 100 samples per field), compared to spiking, where depending of the spatial variability only few samples (e.g., 1–2 sample per ha) would be sufficient to account for the specific soil variability of the target site. Kuang and Mouazen (2012) suggested that among different sample number datasets studied (e.g., 25, 50, 75 and 100 samples), the number of soil samples should be chosen according to the accuracy required, although 50 soil samples was considered appropriate to establish calibration models of TN, SOC and MC. This is correct assuming that considerable spatial variability exists in the dataset to enable correlation between soil spectra and target soil properties to be established, which is not the case for all fields where for some fields variability could be very small. Therefore, spiked ECD calibration models with target field specific samples are successful prediction tools for mobile vis–NIR spectroscopy sensing of the studied soil properties.

Similar to the mobile prediction, spiking of the ECD in particular seems to provide the best prediction performance for laboratory scanned (non-mobile) soil spectra. Spiking local (target site) soil samples into global or ECD models for non-mobile calibrations proved to be an efficient approach to improve the prediction accuracy of a target field for some soil constituents (Shepherd and Walsh, 2002; Janik et al., 2007; Waiser et al., 2007; Minasny et al., 2009; Viscarra Rossel et al., 2009; Wetterlind and Stenberg, 2010; Guerrero et al., 2010; Kuang and

496 Mouazen, 2013).

497 In line with findings of Kuang and Mouazen (2011), it is believed that a fundamental factor that
498 leads to improved model performance (in both mobile and laboratory predications) is the wide
499 range of variability of TN, TC, MC which can be secured with the ECD. Stenberg et al. (2010)
500 reported that the model performance depends on to a large extent on variability encountered in
501 the dataset, including soil types, which is the case in the ECD of the current work. This may be
502 attributed to the fact that different soil types can have considerably different physical and
503 chemical characteristics influencing soil spectra, which is attributed to differences in the parent
504 materials. Thus, with large soil heterogeneity, regression can be more successful and this may
505 influence the model accuracy. The high performance for estimating the soil properties in the
506 current research obtained with MARS may be attributed to the wide range of variation of the
507 large-scale dataset (ECD), shown in Table 1.

508 In the present study, spiked UND and ECD models did not result in substantially lower RMSE
509 values compared with the results obtained with when using local calibrations alone (e.g., SFD
510 and TFD). This corresponds to the small and variable differences for TN, TC and MC between
511 the calibrations with the spiked UND and ECD libraries, and the local-only samples, which is in
512 agreement with observations found by Brown (2007). Moreover, these results were consistent
513 with the observations for TN, OC, and MC made by Kuang and Mouazen (2013), who obtained
514 substantially better predictions for some sites using the ECD library spiked with local samples,
515 compared with using only local samples. These authors confirmed that although higher R^2 and
516 RPD values can be obtained with spiked ECD calibration models, higher RMSE values should
517 be expected.

518

519 **5. Conclusions**

520 In this study, the partial least squares regression (PLSR), support vector machine (SVM) and
521 multivariate adaptive regression splines (MARS) methods were compared for the laboratory
522 (non-mobile) and mobile prediction of soil total nitrogen (TN), total carbon (TC) and moisture
523 content (MC) in two fields (e.g., Hessleskew and Hagg) using four different datasets of different
524 geographical scales. Generally, the accuracy of the PLSR, SVM and MARS models varied in
525 accordance with variations in the calibration scales. Results showed the majority of the non-
526 linear calibration methods (particularly MARS) to outperform the linear PLSR in cross-
527 validation modelling. However, PLSR provided acceptable accuracy for the prediction of the
528 studied soil properties. The most important finding was that in general results of cross-validation
529 to improve with increasing the scale of the data set from one field dataset (SFD), to two fields
530 (TFD), and finally spiking with large data set collected from some fields in the UK (UND), and
531 from several fields in Europe (ECD). However, the validation of model performance in
532 prediction does not necessarily follow the same trend. Higher prediction results were observed
533 for the individual and two-field data set as compared to the large EU data set. Another interesting
534 finding was that the two-field data set resulted in skewness of predicted values towards the field
535 with high concentration of a given soil property.

536 The ECD calibrations seem to be the best option for mobile predictions of the studied soil
537 parameters as compared to field or field scale. The ECD calibrations based on 529 samples
538 outperformed UND (89 samples), TFD (234 samples) and the SFD (122 and 149 for Hessleskew
539 and Hagg, respectively) models for TN, TC and MC at both fields. The potential for good
540 calibration was obtained particularly for TC and TN. There was a tendency for better predictions

when spiking the ECD dataset compared to spiking the UND dataset, particularly for TN and TC.

The predictive models for estimating TN, TC and MC may become more accurate through the selection of an optimal data set to spike with UND and ECD libraries. Future research should focus on the potential for integration of data-mining techniques with the optimal spiked libraries for improving the prediction accuracy at different scales. We believe that these estimation models should be subjected to further examination and optimisation prior to their broad application in soil TN, TC and MC modelling and mapping.

Acknowledgements

Authors acknowledge the financial support received through Tru-Nject project (Nr. 36428-267209), which was jointly sponsored by Innovate UK and Biotechnology and Biological Sciences Research Council (BBSRC).

References

- Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *Eur. J. Soil Sci.* 65, 718–729.
- Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC - Trends Anal. Chem.* 29, 1073–1081.
- Bilgili, A.V., van Es, H.M., Akbas, F., Durak, A., Hively, W.D., 2010. Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. *J. Arid Environ.* 74, 229–238.
- Bongiovanni, R., Lowenberg-Deboer, J., 2004. Precision Agriculture and Sustainability. *Precis. Agric.* 5(4), 359–387.
- British Standard BS 7755 Section 3.8, 1995. Soil quality. Chemical methods. Determination of organic and total carbon after dry combustion (elementary analysis). Equivalent to ISO 10694:1995. London, UK: The British Standards Institution.
- British Standard BS EN 13654-2, 2001. Soil improvers and growing media. Determination of

571 nitrogen (Dumas method). Equivalent to ISO 5725:1994. London, UK The British
572 Standards Institution.

573 Brodský, L., Vašát, R., Klement, A., Zádorová, T., Jakšík, O., 2013. Uncertainty propagation in
574 VNIR reflectance spectroscopy soil organic carbon mapping. *Geoderma* 199, 54–63.

575 Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil characterization and
576 landscape modeling in a 2nd-order Uganda watershed. *Geoderma* 140, 444–453.

577 Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global
578 soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* 132, 273–290.

579 Conforti, M., Buttafuoco, G., Leone, A.P., Aucelli, P.P.C., Robustelli, G., Scarciglia, F., 2013.
580 Studying the relationship between water-induced soil erosion and soil organic matter using
581 Vis–NIR spectroscopy and geomorphological analysis: A case study in southern Italy.
582 *Catena* 110, 44–58.

583 Conforti, M., Castrignanò, A., Robustelli, G., Scarciglia, F., Stelluti, M., Buttafuoco, G., 2015.
584 Laboratory-based Vis–NIR spectroscopy and partial least square regression with spatially
585 correlated errors for predicting spatial variation of soil organic matter content. *Catena* 124,
586 60–67.

587 Efron, B., Tibshiran, R., 1993. An introduction to the bootstrap. Chapman & Hall, Inc., New York.

588 Felicísimo, Á.M., Cuartero, A., Remondo, J., Quirós, E., 2012. Mapping landslide susceptibility
589 with logistic regression, multiple adaptive regression splines, classification and regression
590 trees, and maximum entropy methods: a comparative study. *Landslides* 10, 175–189.

591 Friedman, J.H., 1991. Multivariate Adaptive Regression Splines 19, 1–67.

592 Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R.A., Maestre, F.T., Mouazen, A.M.,
593 Zornoza, R., Ruiz-Sinoga, J.D., Kuang, B., 2014. Assessment of soil organic carbon at local
594 scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking
595 subset. *Eur. J. Soil Sci.* 65, 248–263.

596 Guerrero, C., Zornoza, R., Gómez, I., Mataix-Beneyto, J., 2010. Spiking of NIR regional models
597 using samples from target sites: effect of model size on prediction accuracy. *Geoderma* 158,
598 66–77.

599 Halcro, G., Corstanje, R., Mouazen, A.M., 2013. Site-specific land management of cereal crops
600 based on management zone delineation by proximal soil sensing. In J. Stafford (Ed.),
601 Precision Agriculture 2013, Proceedings of the 10th European Conference on Precision
602 Agriculture (pp. 475–481). Wageningen: Wageningen Academic Publishers.

603 Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning: Data
604 Mining, Inference, and Prediction. Springer-Verlag, New York, pp. 744 pp.

605 Janik, L.J., Skjemstad, J.O., Shepherd, K.D., Spouncer, L.R., 2007. The prediction of soil carbon
606 fractions using mid-infrared-partial least square analysis. *Aust. J. Soil Res.* 45, 73–81.

607 Karatzoglou, A., Feinerer, I., 2010. Kernel-based machine learning for fast text mining in R.
608 *Comput. Stat. Data Anal.* 54, 290–297.

609 Klement, S., Madany Mamlouk, A., Martinetz, T., 2008. Reliability of Cross-Validation for
610 SVMs in High-Dimensional, Low Sample Size Scenarios, in: Artificial Neural Networks -

611 ICANN 2008. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 41–50.

612 Kuang, B., Mouazen, A.M., 2011. Calibration of visible and near infrared spectroscopy for soil
613 analysis at the field scale on three European farms. *Eur. J. Soil Sci.* 62, 629–636.

614 Kuang, B., Mouazen, A.M., 2012. Influence of the number of samples on prediction error of
615 visible and near infrared spectroscopy of selected soil properties at the farm scale. *Eur. J.*
616 *Soil Sci.*, 63(3): 421–429.

617 Kuang, B., Mouazen, A.M., 2013. Effect of spiking strategy and ratio on calibration of on-line
618 visible and near infrared soil sensor for measurement in European farms. *Soil Tillage Res.*
619 128, 125–136.

620 Kuang, B., Tekin, Y., Mouazen, A.M., 2015. Comparison between artificial neural network and
621 partial least squares for on-line visible and near infrared spectroscopy measurement of soil
622 organic carbon, pH and clay content. *Soil Tillage Res.* 146, 243–252.

623 Ludwig, B., Vormstein, S., Niebuhr, J., Heinze, S., Marschner, B., Vohland, M., 2017.
624 Estimation accuracies of near infrared spectroscopy for general soil properties and enzyme
625 activities for two forest sites along three transects. *Geoderma* 288, 37–46.

626 Luoto, M., Hjort, J., 2005. Evaluation of current statistical approaches for predictive
627 geomorphological mapping. *Geomorphology* 67, 299–315.

628 Martens, H., Naes, T., 1989. Multivariate calibration. John Wiley & Sons, Chichester, UK.

629 Maleki, M.R., Mouazen, A.M., De Ketelaere, B., Ramon, H., De Baerdemaeker, J., 2008. On-
630 the-go variable-rate phosphorus fertilisation based on a visible and near-infrared soil sensor.
631 *Biosyst. Eng.* 99, 35–46.

632 Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., Lin, C.-C.,
633 2015. Package ‘e1071’: Misc Functions of the Department of Statistics, Probability Theory
634 Group (Formerly: E1071), TU Wien. Available online:
635 <https://mran.revolutionanalytics.com/web/packages/e1071/e1071.pdf> (accessed
636 27.10.2015).

637 Minasny, B., McBratney, A.B., Pichon, L., Sun, W., Short, M.G., 2009. Evaluating near infrared
638 spectroscopy for field prediction of soil properties. *Aust. J. Soil Res.* 47, 664–673.

639 Morellos, A., Pantazi, X-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G,
640 Wiebensohn, J., Bill, R., Mouazen, A.M., 2016. Machine learning based prediction of soil
641 total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy.
642 *Biosyst. Eng.* doi:10.1016/j.biosystemseng.2016.04.018.

643 Mouazen, A. M., 2006. Soil sensing device. International publication, Published under the Patent
644 Cooperation Treaty (PCT). World Intellectual Property Organization, International Bureau,
645 International Publication Number; W02006/015463; PCT/ BE 2005/000129;
646 G01N21/00G01N21/00

647 Mouazen, A.M., De Baerdemaeker, J., Ramon, H., 2005. Towards development of on-line soil
648 moisture content sensor using a fibre-type NIR spectrophotometer. *Soil Tillage Res.* 80,
649 171–183.

650 Mouazen, A.M., Kuang, B., 2016. On-line visible and near infrared spectroscopy for in-field

phosphorous management. *Soil Tillage Res.* 155, 471–477.

Mouazen, A.M., Kuang, B., De Baerdemaeker, J., Ramon, H., 2010. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma* 158, 23–31.

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.

Nawar, S., Buddenbaum, H., Hill, J., 2015. Estimation of soil salinity using three quantitative methods based on visible and near-infrared reflectance spectroscopy: a case study from Egypt. *Arab. J. Geosci.* 8, 5127–5140.

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., Mouazen, A.M., 2016. Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. *Soil Tillage Res.* 155, 510–522.

Peng, X., Shi, T., Song, A., Chen, Y., Gao, W., 2014. Estimating Soil Organic Carbon Using Vis/NIR Spectroscopy with SVMR and SPA Methods. *Remote Sens.* 6, 2699–2717.

Pietrzykowski, M., Chodak, M., 2014. Near infrared spectroscopy — A tool for chemical properties and organic matter assessment of afforested mine soils. *Ecol. Eng.* 62, 115–122.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

Samui, P., 2012. Multivariate Adaptive Regression Spline (Mars) for Prediction of Elastic Modulus of Jointed Rock Mass. *Geotech. Geol. Eng.* 31, 249–253.

Sankey, J.B., Brown, D.J., Bernard, M.L., Lawrence, R.L., 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. *Geoderma* 148, 149–158.

Sarkhot, D. V., Grunwald, S., Ge, Y., Morgan, C.L.S., 2011. Comparison and detection of total and available soil carbon fractions using visible/near infrared diffuse reflectance spectroscopy. *Geoderma* 164, 22–32.

Shepherd, K.D., Walsh, M.G., 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Sci. Soc. Am. J.* 66, 988–998.

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

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* 107, 163–215.

Song, Y., Ji, J., Mao, C., Ayoko, G.A., Frost, R.L., Yang, Z., Yuan, X., 2013. The use of reflectance visible-NIR spectroscopy to predict seasonal change of trace metals in suspended solids of Changjiang River. *Catena* 109, 217–224.

Sudduth, K.A., Hummel, J.W. 1996. Geographic operating range evaluation of a NIR soil sensor. *Transactions of the ASAE*, 39, 1599–1604.

690 Suykens, J., Vandewalle, J., 1999. Least squares support vector machine classifiers. *Neural.*
691 *Process. Lett.* 293–300.

692 Vapnik, V.N., 1995. The nature of statistical learning theory. *Information Science and Statistics.*
693 Springer, New York.

694 Vapnik, V.N., 1998. *Statistical Learning Theory.* John Wiley and Sons, Inc., New York.

695 Vasques, G.M., Grunwald, S., Sickman, J.O., 2008. Comparison of multivariate methods for
696 inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma* 146, 14–
697 25.

698 Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse
699 reflectance spectra. *Geoderma* 158, 46–54.

700 Viscarra Rossel, R.A., Cattle, S.R., Ortega, A., Fouad, Y., 2009. In situ measurements of soil
701 colour, mineral composition and clay content by vis-NIR spectroscopy. *Geoderma* 150,
702 253–266.

703 Viscarra Rossel, R.A., Chen, C., 2011. Digitally mapping the information content of visible–near
704 infrared spectra of surficial Australian soils. *Remote Sens. Environ.* 115, 1443–1455.

705 Vohland, M., Besold, J., Hill, J., Fründ, H.-C., 2011. Comparing different multivariate
706 calibration methods for the determination of soil organic carbon pools with visible to near
707 infrared spectroscopy. *Geoderma* 166, 198–205.

708 Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T., 2007. In Situ Characterization of
709 Soil Clay Content with Visible Near-Infrared Diffuse Reflectance Spectroscopy. *Soil Sci.*
710 *Soc. Am. J.* 71, 389.

711 Wetterlind, J., Stenberg, B., 2010. Near-infrared spectroscopy for within-field soil
712 characterization: Small local calibrations compared with national libraries spiked with local
713 samples. *Eur. J. Soil Sci.* 61, 823–843.

714 Wetterlind, J., Stenberg, B., Söderström, M., 2010. Increased sample point density in farm soil
715 mapping by local calibration of visible and near infrared prediction models. *Geoderma* 156,
716 152–160.

717 Wijewardane, N.K., Ge, Y., Morgan, C.L.S., 2016. Moisture insensitive prediction of soil
718 properties from VNIR reflectance spectra based on external parameter orthogonalization.
719 *Geoderma* 267, 92–101.

720 Yu, X., Liu, Q., Wang, Y., Liu, X., Liu, X., 2016. Evaluation of MLSR and PLSR for estimating
721 soil element contents using visible/near-infrared spectroscopy in apple orchards on the
722 Jiaodong peninsula. *Catena* 137, 340–349.



DSPACE

<https://dspace.org/>

Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques

Nawar, Said; Mouazen, Abdul Mounem

2016-12-22

Elsevier

<http://dx.doi.org/10.1016/j.catena.2016.12.014>

Nawar S, Mouazen AM, Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques, CATENA, Volume 151, April 2017, Pages 118–129