Comparison between random forest and partial least squares regression of on-line vis-NIR spectroscopy measurements of soil total nitrogen

S. Nawar[1, 2], A.M. Mouazen[2]

[1] Cranfield Soil and AgriFood Institute, School of water, Energy and Environment, Cranfield University, Cranfield, MK43 0AL, UK

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

Abstract.

Accurate and detailed spatial soil information about within field variability is essential for variable-rate applications of farm resources. Soil total nitrogen (TN) is important fertility parameter that can be measured with visible and near infrared (vis-NIR) spectroscopy, whose the calibration method may considerably affect the measurement accuracy. This study aims at comparing the performance of local farm scale calibrations with those based on spiking of local samples into an European continental dataset (ECD) for TN estimation using two modelling techniques, namely, random forest (RF) and partial least squares regression (PLSR). An on-line sensor platform equipped with a mobile, fiber type, vis-NIR spectrophotometer (AgroSpec from tec5 Technology for Spectroscopy, Germany), with a measurement range of 305–2200 nm was used to acquire soil spectra in diffuse reflectance mode from two fields in the UK. After dividing spectra into calibration (75%) and validation (25%) sets, spectra in the calibration set were subjected to RF and PLSR with leave-one-out cross-validation to establish calibration models of TN. Results showed that RF outperformed PLSR models for both datasets used, whereas the lowest model performance was obtained with the local dataset. The effect of spiking of local samples into the ECD was significant, and resulted in high coefficients of determination (R^2) values of 0.97, low root mean square error (RMSE) of 0.01, and high residual prediction deviations (RPD) of 5.78. The on-line predicted maps showed considerable spatial similarity with measured TN. Therefore, these results suggest that spiked ECD based vis-NIR RF calibration models can be successfully used to predict TN under on-line measurement conditions.

Keywords. Vis-NIR spectroscopy, spiking, random forest, partial least squares regression, soil mapping.

Introduction

Estimation of nitrogen status in the soil is crucial from both agriculture and environment points of view. Soil total nitrogen (TN) has for long been identified as a factor that is important to soil fertility and crop production (Kucharik et al., 2001; Muñoz & Kravchenko, 2011). It is well known that mineral nitrogen is the macronutrient that often limits the growth of plants and significantly affects yield (Michopoulos et al., 2008). But, mineral nitrogen is either provided to the soil by chemical or organic fertilizer, or became available from the mineralization of TN. Therefore, spatial predictions of soil TN contents is needed for a wide range of agricultural and environmental applications (Muñoz & Kravchenko, 2011; Wang et al., 2013).

Traditional laboratory analysis methods for TN are laborious, time consuming, costly and destructive (Wang et al., 2015). Therefore, there is a growing demand for a quick, cost-effective, nondestructive and sufficiently accurate approach for predicting TN in situ (Wang et al., 2015).

Proximal soil sensing (PSS) techniques have the potential to eliminate the aforementioned constraints (Viscarra Rossel et al., 2011; Kuang et al., 2012). As common PSS technique, 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 offers the possibility of collecting high spatial resolution data, compared with conventional laboratory analyses (Shepherd & Walsh, 2002; Mouazen et al., 2007; Wetterlind et al., 2010). This spectroscopy technique allows for both in field (in situ) non-mobile (Viscarra Rossel & Chen, 2011; Brodský et al., 2013) and mobile (on-line) measurements (Maleki et al., 2008; Kuang & Mouazen, 2013). However, compared to laboratory analysis that is done under controlled conditions, field spectroscopy analysis are affected by ambient and experimental conditions that need to overcome for accurate prediction to be achieved (Mouazen et al., 2007; Stenberg et al., 2010). One way to reduce these negative influences is by adopting advanced modelling techniques, particularly those approaches that account for nonlinearity in the soil.

As a linear multivariate analysis, partial least squares regression (PLSR) is the most commonly used technique for soil spectral analysis (Viscarra Rossel et al., 2006; Vohland et al., 2011; Conforti et al., 2015). However, the accuracy of PLSR as a linear multivariate regression technique tends to decrease due to the non-linear nature of the relationship between spectral data and the dependent variable (Araújo et al., 2014). Non-linear regression introduced in the literature as better alternatives to PLSR for spectroscopic analyses of soils (Morellos et al., 2016; Nawar et al., 2016). Among those models artificial neural networks (ANN) (Mouazen et al., 2010), support vector regression (SVR) (Vohland et al., 2011), boosted regression trees (Brown et al., 2007), multivariate adaptive regression splines (MARS) (Nawar et al., 2016) and random forests (RF) (Viscarra Rossel & Behrens, 2010) were proven to provide improved prediction performances. Recently, RF has received growing attention in vis-NIR spectral analyses in different domains. It is an ensemble learning technique introduced by (Breiman, 2001) as a combination of tree predictors. It has many advantages such as resistance to noise, ability to be used even when the predictor variables are higher than observations, suffering of small overfitting, and provides an assessment of variable importance (Díaz-Uriarte & Alvarez de Andrés, 2006; Prasad et al., 2006; Ishwaran, 2007). Accordingly, RF can handle nonlinear and hierarchical behaviors when introducing variability to the general spectral library for predicting local samples. Although few studies on the use of RF for the analyses of soil properties under laboratory non-mobile measurement conditions have been reported, to our best knowledge no study of RF modelling based on on-line collected vis-NIR spectra can be found in the literature.

This study aims at comparing the performance of local farm scale calibrations with those based on spiking of local samples into an European continental dataset (ECD) for TN estimation using two modelling techniques, namely, random forest (RF) and partial least squares regression (PLSR) analyses.

2. Materials and Methods

2.1. Experimental sites

Two experimental fields were used in this study, namely, Hessleskew (longitudes -0.590° and -0.586° W, and latitudes 53.844° and 53.844° N) with total area of about 12 ha, and Hagg (longitudes of -1.172° and -1.166° W, and latitudes of 53.936° and 53.941° N) with total area of about 21 ha, both located in Yorkshire, the UK (Fig. 1). Hessleskew field is cultivated with cereal crops in rotation, where Hagg field is cultivated with vegetables crops (leeks, cabbage, carrots and onions). The soil texture for the Hessleskew and Hagg fields is clay and sandy loam, respectively, according to United States Department of Agriculture (USDA) textural soil classification system (Soil Survey Staff, 1999).

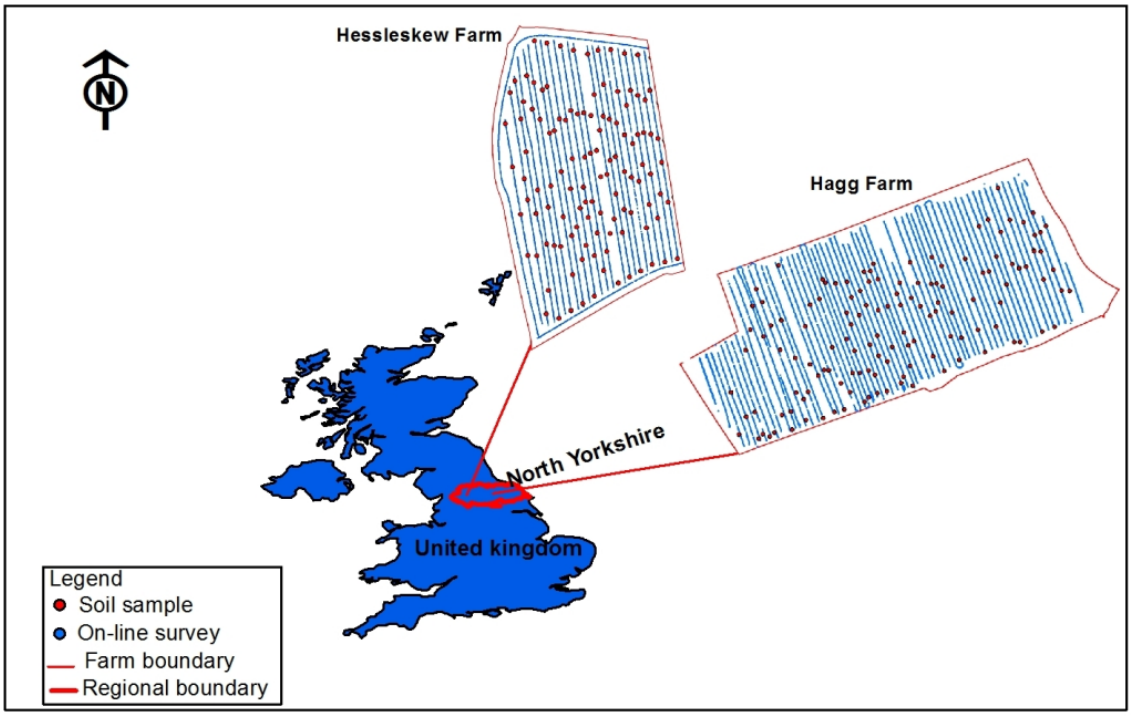

Fig. 1. Location of the Hessleskew and Hagg fields in Yorkshire, UK, showing sampling points collected along on-line survey transects.

2.1. On-line soil measurement and collection of soil samples

Both fields were scanned using the on-line system designed and developed by Mouazen (2006). It consists of a subsoiler, which penetrates the soil to any depth between 5 and 50 cm, 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 (Fig. 2). This was mounted onto the three point linkage of a 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 the on-line measured spectra with sub-meter accuracy. The on-line 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, setting the subsoiler tip at 15 cm average depth. A total of 122 and 149 soil samples were collected during the on-line measurement from Hessleskew and Hagg fields, respectively and used for the calibration and validation modelling described below.

2.2. Laboratory chemical and optical measurements

Fresh soil samples were used in the laboratory spectral and chemical analyses. Each soil sample was placed in a glass container and mixed well then divided into two parts. The first part was used to fill three Petri dishes of 2 cm in diameter, and 2 cm deep, representing three replicated measurements. Each soil in the Petri dishes was 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 on-line measurements. A total of ten scans were collected from each replicate, and these were averaged into one spectrum. The second part of each sample was air dried before it was analyzed for TN using the Dumas method (British Standard BS EN 13654-2:2001).

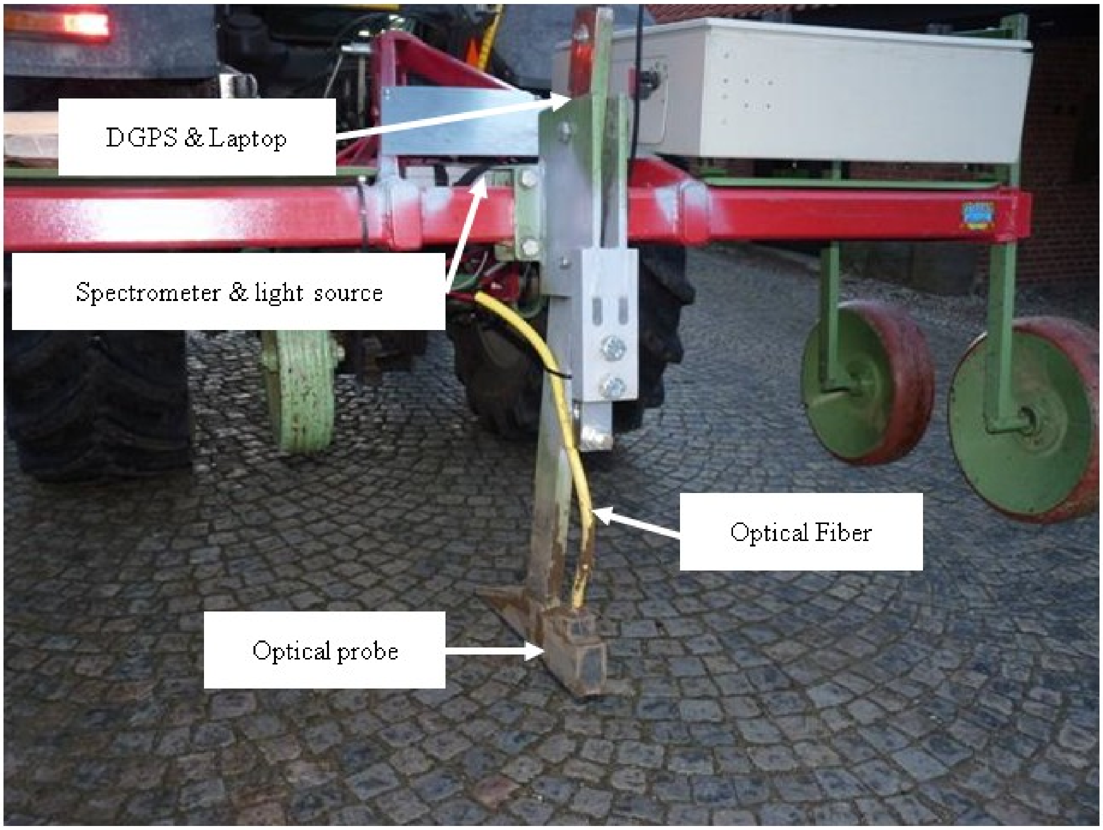

Fig. 2. The on-line visible and near infrared (vis-NIR) spectroscopy sensor developed by Mouazen (2006)

2.3. Spectra pretreatment

The same pretreatment of soil spectra of on-line and laboratory measurements was carried out using R packages (Stevens & Ramirez Lopez, 2013). First, the spectral range outside 370–1979 nm was cut to remove the noise at both edges. Then, moving average with five successive wavelengths was used to reduce noise. Maximum normalization was followed, which is typically used to get all data to approximately the same scale. Spectra were then subjected to first derivation using Gap–segment derivative (gapDer) algorithms (Norris, 2001) 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.

2.4. Dataset set selection and modelling

The following two different data sets (Nawar & Mouazen, 2017) were considered in this study:

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

2- European continental dataset (ECD; n=528), 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 150, 147, 43, 99, and 89 samples were collected, respectively (Kuang & Mouazen, 2011, 2013).

Kennard-Stone calibration sample algorithm (Kennard & Stone, 1969) was used to select the calibration set (75%), and the remaining samples (25%) were assigned to the prediction set, which was used for assessing the model's prediction performance. Based on a spectral distance measure, the Kennard-Stone algorithm selects a set of samples having a uniform distribution over the spectral space. Here, we defined the Mahalanobis distance in the normalised score space of the principal components explaining more than 95% of the spectral variation.

Spiking was used to introduce the local variability of the two experimental fields (e.g., SFD) into the existing data sets (e.g., ECD). A total of 85 and 110 samples that have been selected from the Hessleskew and Hagg fields, respectively, using Kennard-Stone were spiked into the ECD.

2.4.1. Partial least squares regression

PLSR is a widely multivariate analysis method often used in chemometrics. This method is introduced first by Geladi and Kowalski (1986). PLSR is very often used to conduct quantitative spectral analyses. The algorithm uses a linear multivariate model to relate two data matrices – the predictor variables, X, and the response variables, Y. Successive orthogonal (latent) factors are selected, thereby maximizing X and Y covariance – between (X, spectra and Y, the measured value). As opposed to multiple linear regression, PLSR is appropriate for managing data with severe co-linearity in the independent variables, particularly in cases where the sample size is small. The number of latent factors is defined by using leave-one-out cross-validation (LOOCV) (Efron and Tibshirani, 1994). This helped to prevent over- or under-fitting of the data, which may produce models with poor performance. PLSR with leave-one-out cross-validation was used to establish models to predict TN using package 'pls' available in R software (R Core Team, 2013).

2.4.2. Random Forest Regression

Random Forest (RF) is an ensemble learning method generally used for data classification and regression algorithm developed by Breiman (2001). The algorithm worked by growing an ensemble of regression trees based on binary recursive partitioning, where the predictor space at each tree node was partitioned based on binary splits on a subset of randomly selected predictors. RF does not need any data pretreatment which is one of its main advantages. RF runs very fast, and that is a very important factor in the field of on-line and in situ measurements of soil properties. Tree diversity guarantees RF model stability that is achieved by two means: (1) a random subset of predictor variables is chosen to grow each tree and (2) each tree is based on a different random data subset, created by bootstrapping, i.e., sampling with replacement. The data portion used as training subset is known as the “inbag” data, whereas the rest is called the “out-of-bag” data. The latter are not used to build the tree, but provide estimates of generalization errors (Breiman, 2001). The mean square error calculated from prediction with the test dataset averaged over all trees is called the out-of-bag (oob) error. As forest size increases, this generalization error always converges (Breiman, 2001). For each tree in the bootstrapped set, a modified algorithm is used for splitting at each node instead of testing the performance of all p variables. The size of the subset of variables used to grow each tree (mtry) has to be selected by the user. The default mtry value is the square root of the total number of variables (Abdel Rahman et al., 2014). Therefore, ntree needs to be set sufficiently high. Consequently, RFs do not overfit when more trees are added, but produce a limited generalisation error (Prasad et al., 2006; Peters et al., 2007). The same datasets used in PLSR (e.g., 75% calibration, 25% validation) were utilised for analysis. The optimal number of trees to be grown (*ntree),* number of predictor variables used to split the nodes at each partitioning (*mtry*), and the minimum size of the leaf (*nodesize*) were set to 100, 2, and 3, respectively. These parameters were determined by the tune RF function implemented in the R software package, named Random Forest Version 4.6-12 (Liaw and Wiener, 2015), based on Breiman and Cutler's Fortran code (Breiman, 2001).

2.5. Evaluation of model accuracy

Model performance for the prediction of named soil properties were evaluated by means of the coefficient of determination (R^2), root mean square error of prediction (RMSEP) and ratio of residual prediction deviation (RPD). Viscarra Rossel et al. (2006) classified the RPD values referring to accuracy of modelling into six classes: excellent (RPD > 2.5), very good (RPD = 2.5–2.0), good (RPD = 2.0–1.8), fair (RPD = 1.8–1.4), poor (RPD = 1.4–1.0), and very poor model (RPD < 1.0). This classification was adopted in this study to compare between different RF models in cross-validation and in laboratory and on-line prediction.

2.6. Soil TN mapping

Two types of maps were developed, namely, comparison and full-data points maps. The comparison maps were developed to compare on-line predicted with laboratory reference measurement of TN based on randomly collected soil samples in the field. This comparison also included maps of predicted values based on laboratory scanned spectra of the same randomly selected samples. To developed the full-point maps, all on-line predicted data points (1200–1500 point per ha) were used. The comparison maps were developed based on the inverse distance weighing (IDW) interpolation method in ArcGIS 10.4.1 (ESRI, USA) software. To produce the latter maps, the geoR package in R program was used for the estimation of the parameters of the model variogram, so as to typify the spatial variation by a fitting variogram model. After a satisfactory variogram was obtained, it was applied in ArcGIS using Geostatistical analyst toolbox to predict the variable qualities at un-sampled sites to make a prediction map and produce the full-data point maps of on-line measured TN.

3. Results and discussion

3.1 Laboratory measured soil properties and spectra

The descriptive statistics for soil TN is shown in Table 1. For Hessleskew field, TN concentration is low, with minim and maximum of 0.19 and 0.34%, respectively. The mean and median is equal (0.25%), wheraes the 1st quartile and 3rd quartile are 0.23 and 0.26%, respectively. TN in the Hagg field is similar to Hessleskew field with minimum, mean, and maximum values of 0.13%, 0.21, and 0.35%, respectively (Table 1). The small range of variability in TN indicates these fields are not the ideal case study fields, as the smaller the variability is the less successful results to be expected for the prediction capability of the calibration models established (Kuang & Mouazen, 2011).

Table 1. Descriptive statistics for soil total nitrogen (TN) in % for Hessleskew, Hagg fields, and European continental dataset (ECD).

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	St.dev
Hessleskew	**(n=122)**						
	0.19	0.23	0.25	0.25	0.26	0.34	0.02
Hagg	**(n=149)**						
	0.13	0.19	0.21	0.21	0.24	0.35	0.04
ECD	**(n=528)**						
	0.03	0.11	0.14	0.15	0.17	0.30	0.04

3.2. Performance of calibration models in cross-validation

Table 2 summarises the cross-validation results for TN calibration models developed with SFD and ECD datasets, whereas Fig. 3 shows the scatterplots for the measured versus predicted values. Examining Tables 2 and Fig. 3 one can conclude that the best results in cross-validation are achieved with the ECD using the RF calibration (R^2 = 0.97, RMSEcv = 0.01%, and RPD = 5.72 for Hagg field, and R^2 = 0.96, RMSEcv = 0.01%, and RPD = 4.83 for Hessleskew field). PLSR calibration models developed with the SFD has resulted in the lowest results with R^2, RMSE, and RPD of 0.56 and 0.62, 0.01 and 0.02%, and 1.48 and 1.62, for Hessleskew and Hagg field, respectively (Table 2 and Fig. 4).

Adopting Viscarra Rossel et al. (2006) classification system of RPD for the evaluation of model prediction performance, reveals that results achieved in this study is excellent for RF modelling in cross-validation (RPD = 4.83–5.72). These results are in line with those reported by Nawar & Mouazen (2017) based on multivariate adaptive regressing splines (MARS) method (R^2 = 0.96, RMSEcv = 0.01% and the ratio of performance to the interquartile range (RPIQ) = 6.57). However, RF model performance based on the ECD dataset obtained in this study is better than that obtained by Kuang and Mouazen (2013) using PLSR (RMSEcv and RPD values of 0.023% and 2.61). This is because the RF modelling technique typically yields better results when a nonlinear relationship between reflectance and concentration exists (typical in soils), whereas the PLSR model fits only linear relationships (Viscarra Rossel & Behrens, 2010; Nawar et al., 2016).

Table 2. Hessleskew and Hagg fields results for soil total nitrogen (TN) in cross-validation, laboratory and on-line predictions based on partial least square regression (PLSR) and random forest (RF) calibration models developed with single field dataset (SFD), and spiked European continental dataset (ECD).

		SFD			ECD		
		RMSE (%)	R^2	RPD	RMSE (%)	R^2	RPD
Hessleskew Field							
PLSR	n.LV						
Cross- valid.	6	0.01	0.56	1.48	0.02	0.73	1.91
Lab Valid.	6	0.01	0.55	1.51	0.03	0.73	1.96
On-line Valid.	6	0.02	0.53	1.48	0.03	0.68	1.79
RF	n.tree						
Cross- valid.	100	0.01	0.84	2.49	0.01	0.96	4.83
Lab Valid.	100	0.01	0.84	2.53	0.02	0.82	2.38
On-line Valid.	100	0.01	0.73	1.95	0.03	0.62	1.63
Hagg Field							
PLSR	n.LV						
Cross- valid.	6	0.02	0.62	1.62	0.02	0.75	2.02
Lab Valid.	6	0.02	0.61	1.63	0.03	0.55	1.5
On-line Valid.	6	0.02	0.62	1.64	0.03	0.65	1.71
RF	n.tree						
Cross- valid.	100	0.01	0.84	2.49	0.01	0.97	5.72
Lab Valid.	100	0.02	0.79	2.19	0.02	0.84	2.54
On-line Valid.	100	0.02	0.76	2.07	0.02	0.83	2.37

nLV=number of latent variables, n.tree = number of trees , RPD = residual prediction deviation, RMSE = root mean square error.

3.3. Performance of calibration models for laboratory prediction

The RF and PLSR calibration models were validated under laboratory conditions using laboratory scanned spectra of the prediction sets. Similar trend to that of the cross-validation can be concluded for the laboratory prediction for TN, where the best predictions are achieved with RF-ECD based model (R^2 = 0.84, RMSE = 0.02%, and RPD = 2.54) in Hagg field, whereas slightly less performance was achieved in Hessleskew field (R^2 = 0.82, RMSE = 0.02%, and RPD = 2.38) (Table 2 and Fig. 3). The poorest results are produced by PLSR-SFD based models (R^2 = 0.55 and 0.61, RMSE = 0.010 and 0.02%, and RPD = 1.51 and 1.63) in Hessleskew and Hagg fields, respectively. These results are supported by the findings of Nawar & Mouazen (2017), who showed that MARS provided robust predictions of TN with ECD (R^2, RMSE, and

RPIQ of 0.87, 0.03%, and 5.21, respectively). However, the RF ECD prediction results in this research are comparable to those reported by Kuang & Mouazen (2013) (RPD of 2.52, respectively). Kuang & Mouazen (2011) reported higher prediction accuracy for non-mobile measurements coupled with PLSR (R^2 = 0.93, RMSEP = 0.06% and RPD = 3.88) than that achieved in this study, probably due to the wide concentration range of TN in the prediction set (0.06 to 1.22%) compared to (0.13 to 0.35%) in the current work.

3.4. Performance of calibration models for on-line prediction

The on-line collected spectra were used to predict soil TN using the calibration models developed in advance, as explained above. Tables 2 summarises the accuracy of the on-line measurement for TN based on SFD and ECD datasets. As for cross-validation and laboratory prediction again RF method when combined with spiking of individual field samples into the ECD has resulted in the best on-line prediction (R^2 = 0.83, RMSE = 0.02%, and RPD = 2.37) in Hagg field, and R^2 = 0.62, RMSE = 0.03%, and RPD = 1.63 in Hessleskew field. Again the lowest results are produced by PLSR-SFD based models with R^2 = 0.53, RMSE = 0.02%, and RPD = 1.48 in Hessleskew field (Fig. 3). Generally ECD models provide considerably better performance in on-line prediction than SFD, which perform almost equally in both fields. Generally, RF outperformed PLS models for on-line prediction for TN, which in line with those results reported by Kuang & Mouazen (2013), who reported lower prediction accuracy using PLSR with RMSE and RPD values of 0.08% and 2.22. Similar accuracies (R^2 = 0.79, RMSE = 0.02%, and RPIQ = 3.26) were reported by Nawar & Mouazen (2017) for almost equal prediction set size (n=39). Our results were better in terms or RMSE from the previous researchers, which reflect the accuracy of the RF model produced in the current study for on-line prediction of TN.

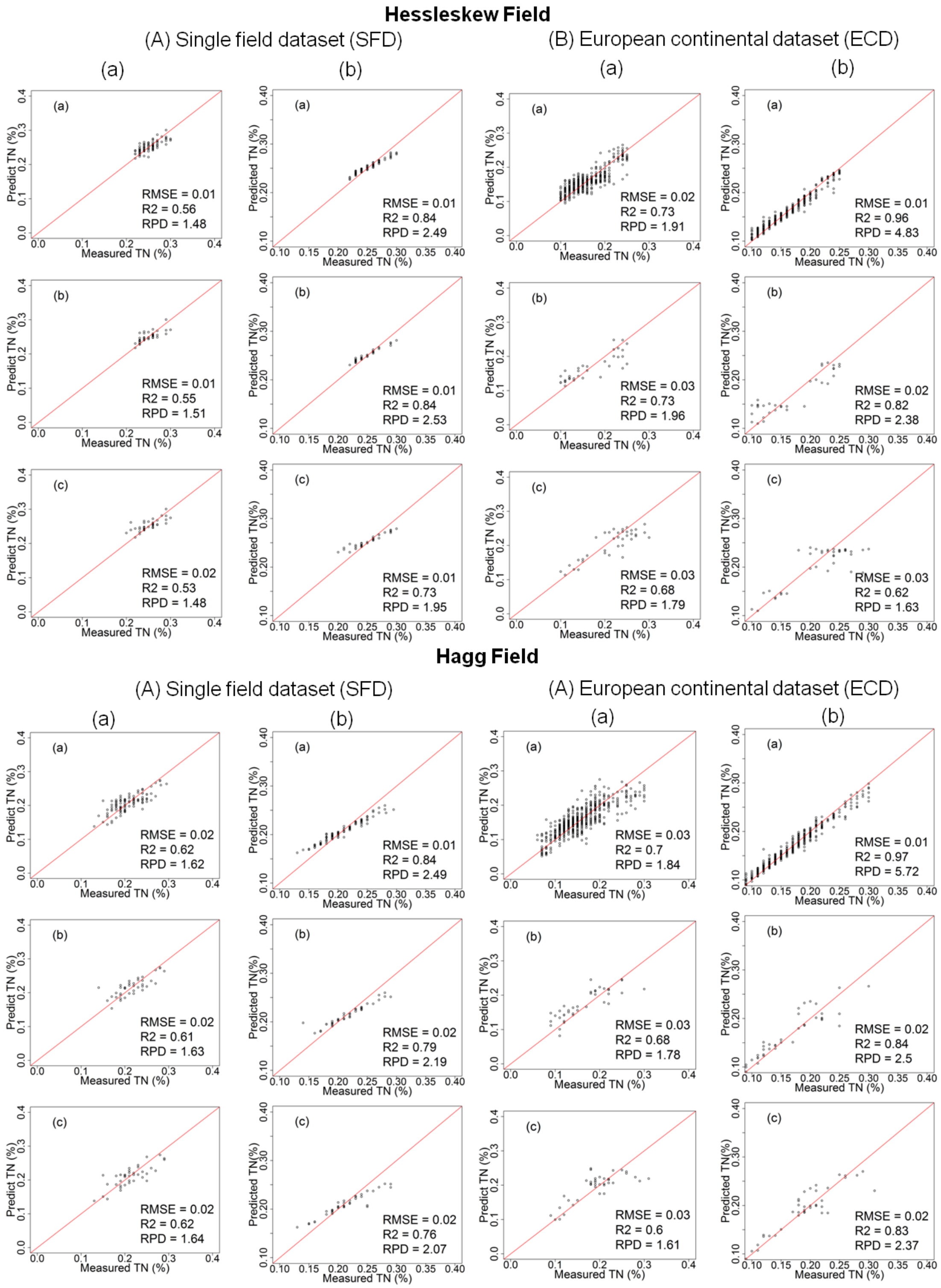

Fig 3. Scatter plots of visible and near infrared (vis-NIR) spectroscopy-predicted versus laboratory-analysed total nitrogen (TN) in Hessleskew and Hagg fields in cross validation (a), lab prediction (b) and on-line prediction (c), using (A) single field dataset (SFD) and (B) spiked European continental dataset (ECD). Results compares between (a) partial least squares regression (PLSR) and (b) random forests (RF) models.

3.5. Influence of dataset on models' prediction performance

The dataset (scale) has shown to have a considerable influence on the performance of calibration and prediction of TN (Nawar & Mouazen, 2017), using other linear and nonlinear modelling tools. The same observation is applied in the current work for RF modelling. Spiking local samples in the ECD library almost always improved the model performance, particularly in cross-validation, compared with those obtained using the SFD (Table 2), which is in agreement with the results presented by Kuang & Mouazen (2013), who concluded that with the increase in the sample set scale (e.g., ECD) and concentration range, one should expect not only increases in RPD and R^2 but in RMSE too. Examining the RPD values obtained with both datasets suggest that spiking of laboratory scanned spectra into ECD is a successful strategy to obtain accurate on-line predictions of soil TN. Figure 4 and Table 2 illustrate how spiking coupled with RF have led to the highest on-line prediction performance compared to SFD for both TN models. Moreover, the majority of on-line prediction models are classified as very good to good prediction performance according to Viscarra Rossel et al. (2006) classification scale of RPD. Compared to MARS-ECD model performance for on-line prediction reported by Nawar & Mouazen (2017) (RPD =2.35), and artificial neural network (ANN)-ECD model performance reported by Kuang et al. (2015) (RPD = 2.28), only slight improved results are achieved in this study with RF-ECD (RPD = 2.37). The current work confirms previous findings and provides additional evidence that the nonlinear modelling methods of soil spectroscopy is a better solution than the linear PLSR techniques, and that there will be always an opportunity to improve the on-line prediction performance by adopting new data mining techniques (e.g., RF in the current work). Moreover, the concept of spiking a general dataset (ECD) with samples from measured target field (SFD) in particular with narrow variation range seems to be a successful calibration procedure for on-line vis–NIR measurement of soil TN.

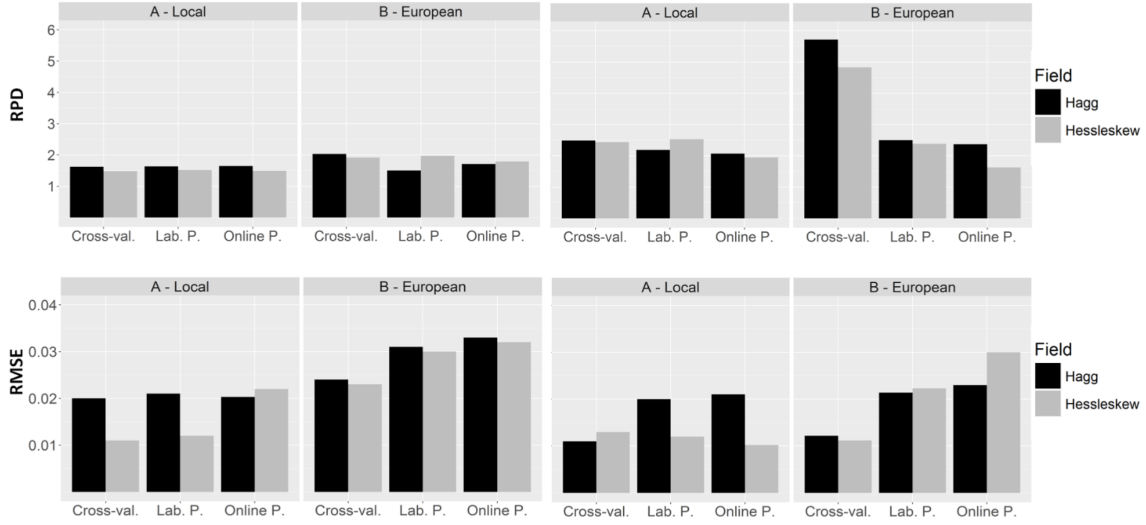

Fig. 4. Comparison between (a) partial least squares regression (PLSR) and (b) random forests (RF) analyses model performances in Hessleskew and Hagg fields for total nitrogen (TN) prediction in cross-validation (cross-val.) and laboratory prediction (Lab. P.) and online prediction (Online P.). Results are shown for residual prediction deviation (RPD) and root mean square error (RMSE).

A possible explanation for improved models performance (in both on-line and laboratory predictions) may be the wide range of variability TN, which can be secured with e.g., the ECD. This is supported by the fact that a vis-NIR model performance depends to a large extent on variability encountered in the dataset including soil types (Stenberg et al., 2010; Wang et al., 2010), which is the case in the ECD of the current work. Thus, with large soil heterogeneity, regression can be more successful, and this may influence the model accuracy (Kuang & Mouazen, 2011). The high performance for estimating TN in the current work obtained with RF method may be attributed to the wide range of variation of the large-scale dataset (ECD). However, the higher R^2 and RPD values of the spiked ECD model, the values of RMSE did not result in substantially lower compared with the results obtained with the local calibrations alone (e.g., SFD and TFD) (Fig. 4). This is maybe attributed to the small and variable differences for TN between the calibrations with the spiked ECD library, and the local samples, which is in agreement with observations found by Kuang & Mouazen (2013).

3.5 Soil TN maps

Figures 5 and 6 show the TN maps of laboratory reference values, laboratory predicted, and on-line predicted values for Hessleskew and Hagg fields, respectively. In order to allow objective comparisons between reference versus laboratory and on-line maps, exactly the same number of classes (6 classes) was considered along with similar range in the three maps. The comparison exhibits large spatial similarity, along with high and low areas match almost perfectly. No distinguished spatial differences between the on-line and laboratory predicted maps can be observed. This particular demonstrates the high quality of the on-line measured spectra, that displays the sensor stability as well as robustness.

3.6. Full points' TN maps

Based on the spherical semivariagram parameters shown in Table 3, full-point maps based on PLSR and RF predictions of TN using all on-line measured spectra are shown in Figures. 7 and 8. These maps showed high spatial variability for both TN in both fields, which encouraged the need for the on-line soil sensor for the characterisation of within field spatial variability of soil properties, as zones with different levels of concentrations should be managed differently in precision agriculture. However, from visual examination, it is rather difficult to conclude on whether or not the RF maps were better than those of the PLSR. But a rather more detailed characterisation of spatial distribution of TN could be observed with RF, particularly for Hagg field (Fig. 8). No clear difference between RF and PLSR maps could be observed in Hessleskew Field, which may be attributed to the small spatial variability of soil TN.

Discussing these maps of TN with the English farmers revealed that the variation in TN for Helsesskew field is attributed to the fact that this field has received large amount of manure (Fig. 7). The variation in Hagg field is attributed to the different type of crops grown within the field in a single cropping season, each receiving different amount of chemical fertilizers (Fig. 8).

The high measurement accuracy of RF demonstrates the ability of RF to approximate the non-linear behavior of soil TN, and recommend RF for more accurate quantitative estimation of TN than PLSR. However, further research is needed to evaluate which of the two groups of full-point maps should be used to develop VR Nitrogen application maps.

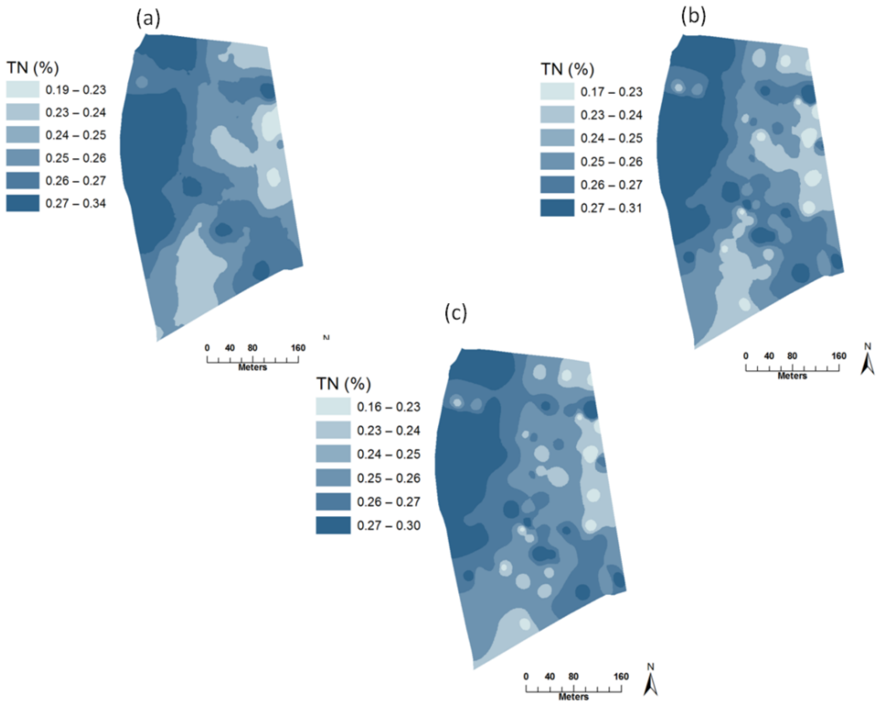

Fig 5. Total nitrogen (TN) maps of Hessleskew field based on laboratory measured (a), laboratory predicted (b), and on-line predicted values (c).

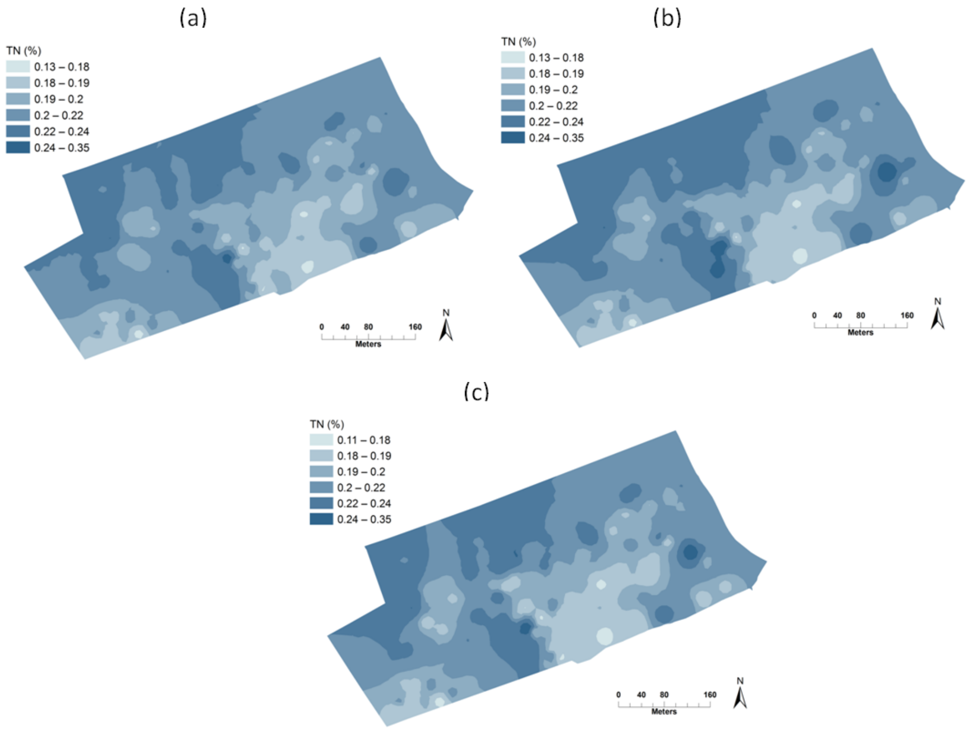

Fig 6. Total nitrogen (TN) maps of Hagg field based on laboratory measured (a), laboratory predicted (b), and on-line predicted values (c).

Table 3 Semivariagram parameters used for the development of partial least squares regression (PLSR) and random forest (RF) predicted full-point maps of total nitrogen (TN) in Hessleskew and Hagg fields

Model	Field	Model fit	Nugget (C0)	Sill (C0 + C1)	Range	Proportion (C1/C0 + C1)
PLSR	Hessleskew	spherical	0.018	0.07	44.23	0.76
RF	Hessleskew	spherical	0.022	0.08	33.85	0.77
PLSR	Hessleskew	spherical	0.010	0.03	42.84	0.75
RF	Hessleskew	spherical	0.005	0.01	62.25	0.66
PLSR	Hagg	spherical	0.004	0.01	60.96	0.77
RF	Hagg	spherical	0.004	0.01	44.41	0.66
PLSR	Hagg	spherical	0.003	0.01	55.26	0.76
RF	Hagg	spherical	0.003	0.01	51.87	0.77

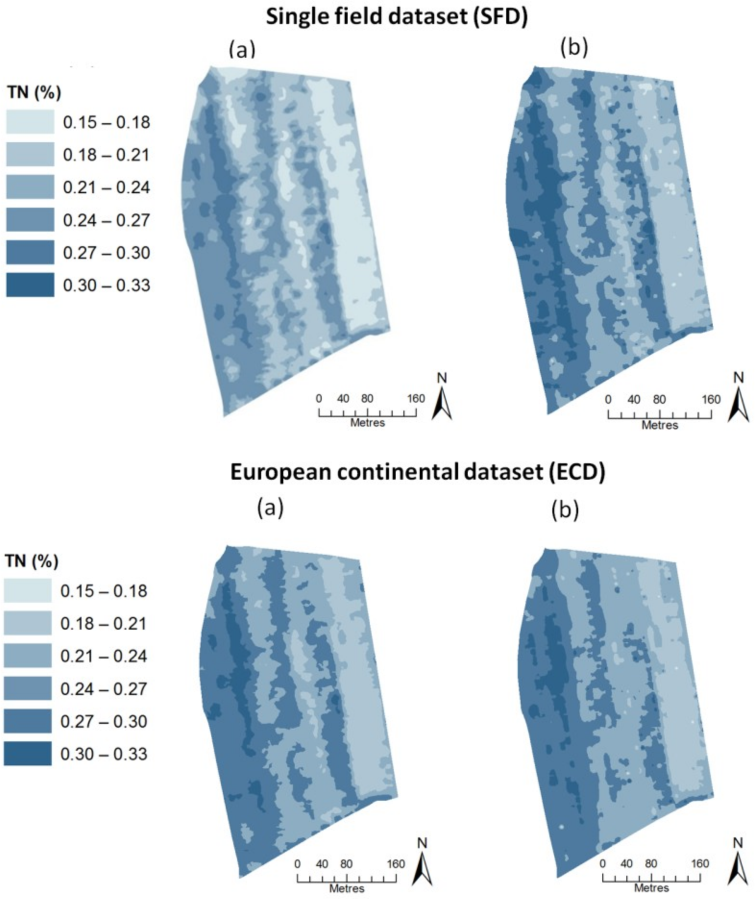

Fig 7. Full-point maps of on-line predicted total nitrogen (TN) using local dataset and spiked European dataset of Hessleskew field based on (a) partial least squares regression (PLSR) and (b) random forests (RF) modelling techniques.

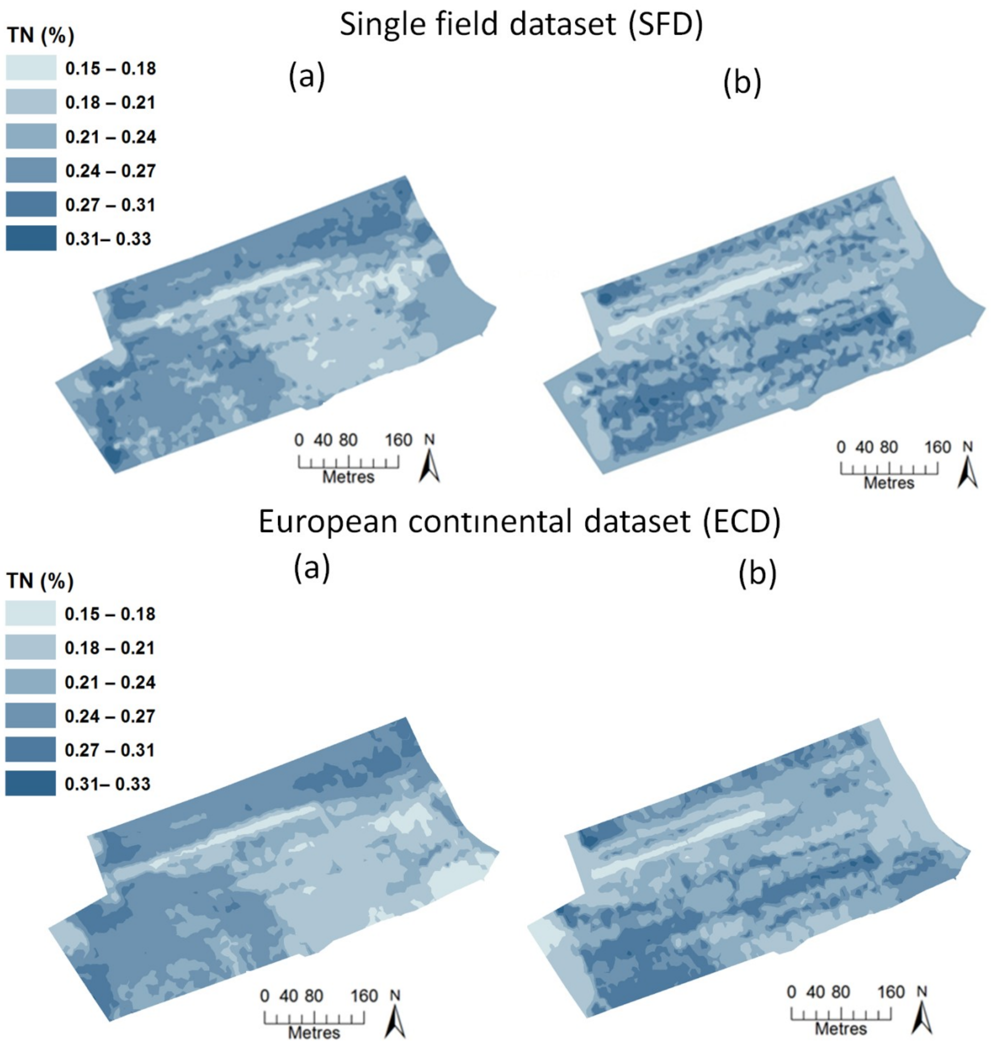

Fig 8. Full-point maps of Hagg field based on (a) partial least squares regression (PLSR) and (b) random forests (RF) predictions of total nitrogen (TN) using local dataset and European dataset.

4. Conclusions

In this study, the performance of random forest (RF) modelling technique was compared with that of partial least squares regression (PLSR) for modelling of the visible and near infrared (vis-NIR) spectral data for laboratory (non-mobile) and on-line (mobile) prediction of soil total nitrogen (TN) in two fields (e.g., Hesselskew and Hagg) using two different datasets of different geographical scales. Generally, the performance of both RF and PLSR models varied in accordance with variations in the calibration scales (dataset). The most important finding was that spiking of selected samples collected from a target field into a spectral library with large variability (e.g., an European calibration dataset (ECD), has resulted in the best laboratory and on-line predictions of TN. The on-line prediction accuracies of TN are classified as good to very good models and were found to be better than those achieved so far with other researches using similar datasets but with different nonlinear modelling approaches. This may well confirm the superiority of the RF technique compared to PLSR for on-line predictions of TN in handling the nonlinear relationship that exists in soils between the response variable and predictor variables. The on-line predicted maps showed considerable spatial similarity with measured TN.

Further work is being undertaken to optimize the selection of the calibration dataset for spiking of the target field samples into existing spectral library. The concept of spiking of general calibration models needs to be tested for other soil properties with other dataset than those reported in this study.

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). The authors also acknowledge the FWO funded Odysseus Project (Nr. G0F9216N).

References

Abdel Rahman, A.M., Pawling, J., Ryczko, M., Caudy, A.A., & Dennis, J.W. 2014. Targeted metabolomics in cultured cells and tissues by mass spectrometry: Method development and validation. Analytica Chimica Acta 845, 53–61.

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. European Journal of Soil Science 65, 718–729.

Breiman, L. 2001. Random forests. Machine Learning 45, 5–32.

British Standard BS EN 13654-2, 2001. Soil improvers and growing media. Determination of nitrogen (Dumas method). Equivalent to ISO 5725:1994. London, UK The British Standards Institution.

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

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

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

Díaz-Uriarte, R., & Alvarez de Andrés, S. 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics 7, 1–13.

Efron, B, Tibshiran, R., 1994. An introduction to the bottstrap. Chapman &Hall, Inc., New York.

Geladi P, Kowlaski B (1986) Partial least square regression: A tutorial. Analytica Chimica Acta 35, 1–17.

Ishwaran, H. 2007. Variable importance in binary regression trees and forests. Electronic Journal of Statistics 1, 519–537.

Kennard, R.W., & Stone, L.A. 1969. Computer Aided Design of Experiments. Technometrics 11, 137–148.

Kuang, B., Mahmood, H.S., Quraishi, M.Z., Hoogmoed, W.B., Mouazen, A.M., & van Henten, E.J. 2012. Sensing soil properties in the laboratory, in situ, and on-line. A review. Advances in Agronomy 114, 155–223.

Kuang, B., & Mouazen, A.M. 2011. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. European Journal of Soil Science 62, 629–636.

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

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

Kucharik, C.J., Brye, K.R., Norman, J.M., Foley, J.A., Gower, S.T., & Bundy, L.G. 2001. Measurements and Modeling of Carbon and Nitrogen Cycling in Agroecosystems of Southern Wisconsin: Potential for SOC Sequestration during the Next 50 Years. Ecosystems 4, 237–258.

Liaw, A. & Wiener, M., 2015. Breiman and Cutler's Random Forests for Classification and Regression. R package version n 4.6-12, (At:"https://cran.r-project.org/web/packages/randomForest/randomForest.pdf (Accessed :28 April 2016).

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

Michopoulos, P., Baloutsos, G., & Economou, A. 2008. Nitrogen cycling in a mature mountainous beech forest. Silva Fennica 42, 5–17.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., & Mouazen, A.M. 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosystems Engineering 152, 1–13.

Mouazen, A. M., 2006. Soil sensing device. International publication, Published under the Patent Cooperation Treaty (PCT). World Intellectual Property Organization, International Bureau, International Publication Number; W02006/015463; PCT/ BE 2005/000129; G01N21/00GO1N21/00.

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

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 and Tillage Research 93, 13–27.

Muñoz, J.D., & Kravchenko, A. 2011. Soil carbon mapping using on-the-go near infrared spectroscopy, topography and aerial photographs. Geoderma 166, 102–110.

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 and Tillage Research 155, 510–522.

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

Norris, K. 2001. Applying Norris Derivatives. Understanding and correcting the factors which affect diffuse transmittance spectra. NIR news 12, 6–9.

Peters, J., Baets, B. De, Verhoest, N.E.C., Samson, R., Degroeve, S., Becker, P. De, & Huybrechts, W. 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling 207, 304–318.

Prasad, A.M., Iverson, L.R., & Liaw, A. 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 9, 181–199.

R Development Core Team (2013) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Shepherd, K.D., & Walsh, M.G. 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. Soil Science Society of America Journal 66, 988–998.

Soil Survey Staff, 1999. Soil Taxonomy - A basic system of soil classification for making and interpreting soil surveys; second edition. Agricultural Handbook 436; Natural Resources Conservation Service, USDA. Washington DC, USA.

Stenberg, B., Rossel, R. a V, Mouazen, a M., & Wetterlind, J. 2010. Visible and Near Infrared Spectroscopy in Soil Science. Advances in Agronomy, Vol 107 107, 163–215.

Stevens, A., & Ramirez Lopez, L. 2013. An introduction to the prospectr package. (At: https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf (Accessed :22 April 2016).

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J., & Lobsey, C. 2011. Proximal soil sensing: An effective approach for soil measurements in space and time. Advances in Agronomy 113, 237–282.

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

Viscarra Rossel, R.A., & Chen, C. 2011. Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils. Remote Sensing of Environment 115, 1443–1455.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., & Skjemstad, J.O. 2006.

Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75.

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

Wang, D., Chakraborty, S., Weindorf, D.C., Li, B., Sharma, A., Paul, S., & Ali, M.N. 2015. Synthesized use of VisNIR DRS and PXRF for soil characterization: Total carbon and total nitrogen. Geoderma 243–244, 157–167.

Wang, J., He, T., Lv, C., Chen, Y., & Jian, W. 2010. Mapping soil organic matter based on land degradation spectral response units using Hyperion images. International Journal of Applied Earth Observation and Geoinformation 12, S171–S180.

Wang, K., Zhang, C., & Li, W. 2013. Predictive mapping of soil total nitrogen at a regional scale: A comparison between geographically weighted regression and cokriging. Applied Geography 42, 73–85.

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