Science of the Total Environment, Volume 616-617, March 2017, pp147-155
DOI: 10.1016/j.scitotenv.2017.10.323

Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques

Douglas, R. K.[a*], Nawar, S.[a], Alamar, M. C[a]., Mouazen, A.M.[a,b], Coulon, F.[a]

[a]School of Water, Energy and Environment, Cranfield University, Cranfield, MK43 0AL, UK

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

E-mail of corresponding author: f.coulon@cranfield.ac.uk; Abdul.Mouazen@UGent.be;

Abstract: Visible and near infrared spectrometry (vis-NIRS) coupled with data mining techniques can offer fast and cost-effective quantitative measurement of total petroleum hydrocarbons (TPH) in contaminated soils. Literature showed however significant differences in the performance on the vis-NIRS between linear and non-linear calibration methods. This study compared the performance of linear partial least squares regression (PLSR) with a nonlinear random forest (RF) regression for the calibration of vis-NIRS when analysing TPH in soils. 88 soil samples (3 uncontaminated and 85 contaminated) collected from three sites located in the Niger Delta were scanned using an analytical spectral device (ASD) spectrophotometer (350-2500 nm) in diffuse reflectance mode. Sequential ultrasonic solvent extraction-gas chromatography (SUSE-GC) was used as reference quantification method for TPH which equal to the sum of aliphatic and aromatic fractions ranging between C_{10} and C_{35}. Prior to model development, spectra were subjected to pre-processing including noise cut, maximum normalization, first derivative and smoothing. Then 65 samples were selected as calibration set and the remaining 20 samples as validation set. Both vis-NIR spectrometry and gas chromatography

profiles of the 85 soil samples were subjected to RF and PLSR with leave-one-out cross-validation (LOOCV) for the calibration models. Results showed that RF calibration model with a coefficient of determination (R^2) of 0.85, a root means square error of prediction (RMSEP) 68.43 mg kg^{-1}, and a residual prediction deviation (RPD) of 2.61 outperformed PLSR (R^2 = 0.63, RMSEP = 107.54 mg kg^{-1} and RDP =2.55) in cross-validation. These results indicate that RF modelling approach is accounting for the nonlinearity of the soil spectral responses hence, providing significantly higher prediction accuracy compared to the linear PLSR. It is recommended to adopt the vis-NIRS coupled with RF modelling approach as a portable and cost effective method for the rapid quantification of TPH in soils.

Key words: Total petroleum hydrocarbons; vis-NIR spectroscopy; chemometric methods, Partial least squares regression, Random Forest regression

1. Introduction

Petroleum hydrocarbons contamination in soil is a worldwide significant environmental issue which has raised serious concerns for the environment and human health (Brevik and Burgess, 2013). Petroleum hydrocarbons encompass a mixture of short and long-chain hydrocarbon compounds. However the difference between the term petroleum hydrocarbons (PHC) as such and the term total petroleum hydrocarbons (TPH) should be noted. PHC typically refer to the hydrogen and carbon containing compounds that originate from crude oil, while TPH refer to the measurable amount of petroleum-based hydrocarbons in an environmental matrix and thus to the actual results obtained by sampling and chemical analysis (Coulon and Wu, 2017). TPH is thus a method-defined term. Among a range of techniques, gas chromatography is preferred for the measurement of hydrocarbon contamination in environmental samples, since it allows to detect a broad range of hydrocarbons and can provide both sensitivity and selectivity depending on the detector and hyphenated configuration used (Brassington et al., 2010; Drozdova et al., 2013). However, GC-based techniques can be time consuming and expensive and do not allowed rapid and broad scale analysis of petroleum contamination on-site (Okparanma and Mouazen, 2013; Okparanma et al., 2014).

Among potential rapid measurement techniques that can be carried out on-site, reflectance spectroscopy, including the visible and near infrared (vis-NIRS) and mid infrared ranges, is one of the most promising techniques for detecting and quantifying TPH (Okparanma and Mouazen, 2013). Reflectance spectroscopy measures the diffuse reflected electromagnetic energy from samples (i.e. soil or sediment) subjected to a light source; by modelling the sample spectral data against samples with known chemical composition and concentration levels, calibration models for quantifying key attributes can be established. However, to date very limited studies on the use of reflectance spectroscopy for the analyses of TPH in soil can be found in the literature.

There are also several factors affecting the measurement accuracy of reflectance spectroscopy, including among others the quality of the laboratory reference data and spectra, and adopted pre-processing and modelling techniques (Viscarra and Behrens, 2010; Nawar et al., 2016). Partial least-squares regression (PLSR) is the most common multivariate analysis method, as it is capable to model several response variables simultaneously while effectively addressing strongly collinear and noisy predictor variables (Wold, 2001). It is important to mention that PLSR is a linear approach that may not perform well when solving nonlinear behaviour, e.g., like those of soil. Random Forest (RF) is typically known as a hierarchical nonparametric method that estimates complex nonlinear relationships among independent and dependent variables. RF method was reported to be outperformed by PLSR, adaptive regression splines (MARS), artificial neural network (ANN) and support vector machine (SVM) for the analysis of soil organic carbon, clay content and pH (Viscarra and Brehen, 2010; Breiman, 2001) whereas Knox et al., (2015) reported that RF outperformed PLSR for the analysis of soil total carbon (TC) with residual prediction deviation (RPD) of 2.7 and 2.6 for RF and PLSR, respectively. For TPH analysis using vis-NIRS, a recent study by Chakraborty et al. (2015) showed PLSR outperformed both penalised spline regression (PSR) and RF modelling approaches; the authors reported residual prediction deviation (RPD) of 1.64, 1.86, and 1.96 for RF, PSR, and PLSR, respectively. This single study comparing the performance of RF with PLSR for the analysis of TPH may not confirm this trend to be correct, as previous work reported RF to outperform PLSR for modelling of other soil properties (Knox et al., 2015). Therefore, it is essential to evaluate the capability of the RF as a nonlinear modelling approach for modelling TPH content in the soil and to confirm whether or not TPH can be predicted with RF with higher accuracy than with PLSR. To the best of our knowledge, there is to date no study where RF modelling has been applied to estimate TPH in soils based on vis-NIR spectroscopy with a limited soil data set. Thus, the aim of this study is to

compare the performance of PLSR linear modelling technique with RF nonlinear technique to predict TPH in oil-contaminated soils from Niger Delta, Southern Nigeria, using vis-NIR spectroscopy.

2. Materials and methods

2.1 Study area and sample collection

The study area located in Bayelsa and Rivers State, Niger Delta, Southern Nigeria has a tropical rain forest climate characterised by two seasons: the rainy season lasts for about 7 months between April and October with an overriding dry period in August (known as August break); and the dry season lasts for about 5 months, between November and March. The temperature varies between 25°C and 35°C. The regional geology of the Niger Delta is relatively simple, consisting of Benin, Agbada (the kitchen of kerogen) and Akata Formations, overlain by various types of Quaternary deposits (Kogbe, 1989; Wright et al., 1985). A total of 85 representative spot sample points were collected randomly from three oil contaminated sites (Ikarama: 31 samples; Kalabar: 21 samples; and Joinkrama: 33 samples) in August 2015. The soil samples (approx. 5 kg) were collected in the top 15-cm soil layer using a shovel. In addition, three uncontaminated samples were collected (2 samples from Joinkrama, 1 sample from Kalabar) for control purpose. Fig. 1 shows the sampling location map. Soil samples were kept in air-tight centrifuge tubes and stored at 4 °C using ice block to avoid hydrocarbon volatilisation and preserve field-moist status until shipment to Cranfield University. The samples were then stored at -20^{o}C prior to GC-MS analysis.

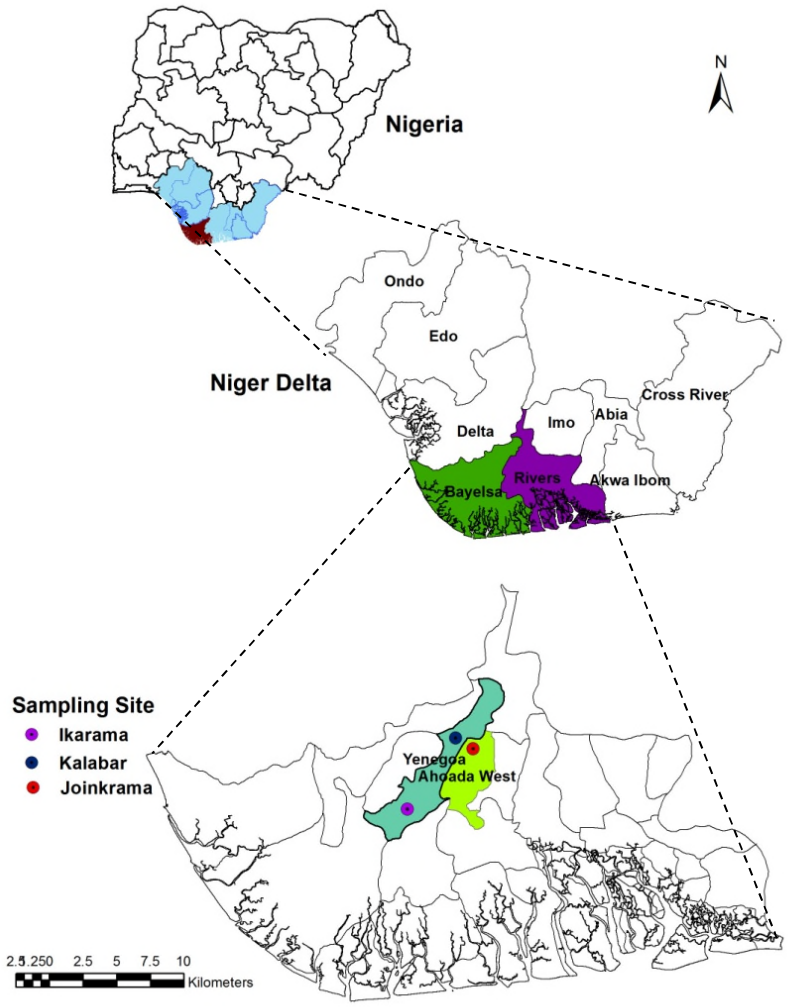

Figure 1 Soil sampling locations for the three sites.

2.2 Soil physiochemical properties

Prior to soil physiochemical properties analysis, soil samples were grouped based on the variation of the soil texture using the "Feel Method" (Thien, 1979). Then two representative samples were selected from each texture class with a total of 10 samples per site. Therefore soil physicochemical properties were determined on 30 soil samples selected to represent soil spatial variation in the study. This approach was used due to limited of amount of soil that could be transported back to the UK for analysis. Soil pH was measured following the Standard Operating Procedure (SOP) of the British

Standard BS ISO 10390:2005; the total organic carbon (TOC) was determined using a Vario III Elemental Analyser using SOP based on British Standard BS 7755 Section 3.8: 1995 and the particle size was determined using SOP based on British Standard BS 7755 Section 5.4:1995.

2.3 Soil scanning and spectral analysis

The diffuse reflectance spectra of the soil samples were measured using an ASD LabSpec2500® Vis–NIR spectrophotometer which covers a spectral range of 350–2500 nm (Analytical Spectral Devices, Inc., USA). With a spectral interval resolution varying of 3 nm at a wavelength of 700 nm and of 6 nm between 1400-1200 nm, the ASD LabSpec2500® spectrometer recorded a total 2151 spectral bands. The spectral measurements were made in the dark in order to both, control the illumination conditions and reduce the effects of stray light. The high-intensity probe has a built-in light source made of a quartz-halogen bulb of 2727 °C. The light source and detection fibres are assembled in the high-intensity probe enclosing a 35° angle. Before use, and after every 30 minutes, the instrument was calibrated by white-referencing with a white Spectralon disc of *ca.* 100% reflectance. Three subsamples (field- moist) from each soil sample were packed into plastic Petri dishes (1 cm height, 5.6 cm diameter) for vis-NIR DRS spectra measurement. To obtain optimal diffuse reflection, and hence, a good signal-to-noise ratio, all plant and pebble particles were removed and surface was smoothened gently with a spatula for scanning (Mouazen et al., 2005). Spectral measurements of all samples were recorded by placing the sample in direct contact with the high intensity probe. For each sample, 10 successive spectra measurements were acquired and further averaged in one representative spectrum of a soil sample. To avoid biased predictions due to noise, only 416-2384 nm spectral range was used to develop the calibration models. The raw average spectra were subjected to pre-processing including successively, noise cut, maximum normalization, first derivative and smoothing with R software (R Core Team, 2013). Maximum normalisation was then implemented to align all spectra to the same

scale or to obtain even distribution of the variances and average values. 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 (Okparanma and Mouazen, 2013). These routines were aimed at keeping useful chemical and physical information (Naes et al., 2002). The same pre-processed data was used for both PLSR and RF analyses.

2.4 Gas chromatography and hydrocarbons quantification

The petroleum hydrocarbons extraction method and GC-MS analysis used in this study followed the procedure described by Risdon et al. (2008) with some modifications. Briefly, 5 g of soil sample was mixed with 20 ml of dichloromethane (DCM): hexane (Hex) solution (1:1, v/v), shaken for 16 h at 150 oscillations per min over 16 h, and finally sonicated for 30 min at 20°C. After centrifugation, extracts were cleaned on Florisil® columns by elution with hexane. Deuterated alkanes and PAHs internal standards were added to extracts at appropriate concentrations. The final extract was diluted (1:10) for GC-MS analysis. Deuterated alkanes ($C10^{d22}$, $C19^{d40}$ and $C30^{d62}$) and PAH (naphthalene ^{d8}, anthracene ^{d10}, chrysene ^{d12} and perylene ^{d12}) internal standards were added to extracts at 0.5 µg ml^{-1} and 0.4 µg ml^{-1}, respectively. Aliphatic hydrocarbons and PAHs were identified and quantified using an Agilent 5973N GC-MS operated at 70 eV in positive ion mode. The column used was a Zebron fused silica capillary column (30 x 0.25 mm internal diameter, Phenomenex) coated with 5MS (0.25 µm film thickness). Splitless injection with a sample volume of 1 µL was applied. The oven temperature was increased from 60 °C to 220 °C at 20 °C min^{-1} then to 310 °C at 6 °C min^{-1} and held at this temperature for 15 min. The mass spectrometer was operated using the full scan mode (range *m/z* 50-500) for quantitative analysis of target alkanes and PAHs. For each compound, quantification was performed by integrating the peak at specific *m/z* using auto-integration method with Mass Selective Detector (MSD)

ChemStation software. External multilevel calibrations were carried out for both alkanes and PAH quantification ranging from 0.5 to 2500 µg ml^{-1} and from 1 to 5 µg ml^{-1}, respectively. For quality control, a 500 µg ml^{-1} diesel standard solution (ASTM C_{12}-C_{60} quantitative, Supelco) and mineral oil mixture Type A and B (Supelco) were analysed every 20 samples. The variation of the reproducibility of extraction and quantification of soil samples were determined by successive injections (n=7) of the same sample and estimated to ±8%. In addition, duplicate reagent control and reference material were systematically used. The reagent control was treated following the same procedure as the samples without adding soil sample. The reference material was an uncontaminated soil of known characteristics, and was spiked with a diesel and mineral oil standard at a concentration equivalent to 16,000 mg kg^{-1}. Relative standard deviation (RSD) values for all the soils was <10%. The limit of quantification (LOQ) of 0.02 mg kg^{-1} customarily used for PAH in Nigerian laboratories was adopted for this study because samples were collected from Nigeria. The LOQ was defined as the lowest concentration, at which an analyte can be reliably detected (Mitra, 2003). As such, any value below 0.02 mg/kg was considered unreliable and ignored from the computation. Finally, the TPH data was obtained by the sum of the aliphatic fractions and the PAH for each sample analysed.

2.5 Development of calibration models

A two dimensional data matrix was developed by combining the pre-processed spectra (predictor) of 85 soil samples and the TPH reference values (dependent variable) where the resolved spectral bands (wavelengths) were defined as X_i (the predictor variables), and TPH concentrations as Y_i (the response variables). The dataset was divided into 75% for calibration (65 samples) and 25% for prediction (independent validation) (20 samples). The selection was done by means of the Kennard-Stone algorithm which allows to select samples with a uniform distribution over the predictor space (Kennard and Stone, 1969). It is a stepwise procedure by maximizing the Euclidean distance based on the

important number of principal components to the objects already chosen. The analyses was performed using 'prospectr' packages in R (Stevens and Lopez, 2013).

2.5.1. Partial least squares regression (PLSR)

PLSR is a widely multivariate analysis method often used in chemometrics. This method is introduced in (Wold, 2001; Gelad and Kowalski, 1986). The algorithm uses a linear multivariate model to relate two data matrices – the predictor variables, X, and the response variables, Y. Information in the original X data is projected onto a small number of underlying orthogonal ("latent") variables called latent variables. In this study, the reflectance values for all 2151 spectral wavelengths comprise the set of X_i variables and the TPH reference values is the Y_i variables. PLSR with full cross-validation was used to relate the variation in a single-component variable (e.g. TPH) to the variation in a multi-component variable (e.g. wavelength) by means of using package 'pls' available in R software (R Core Team, 2013). The optimal number of latent variables (factors) for future predictions was determined on the basis of the number of factors with the smallest RMSEP. To develop the calibration model, 75% of the samples were used while the remaining 25% were used for prediction.

2.5.2. Random forest regression

Random forest (RF) is an ensemble learning method for classification and regression, which generates many classifiers and aggregates their results (Breiman, 2001).Tree diversity guarantees RF model stability, which 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 (Efron, 1979). Instead of testing the performance of all p variables, a modified algorithm is used for splitting at each node. The size of the subset of variables used to grow each tree (*mtry*) has to be selected by the user. Each tree grows until it reaches a predefined minimum

number of nodes (*nodesize*). The default *mtry* value is the square root of the total number of variables (Abdel-Rahman et al., 2014). Therefore, *ntrees* needs to be set sufficiently high. Consequently, RFs do not over fit when more trees are added, but produce a limited generalisation error (Peters et al., 2007). The same datasets used in PLS (75% calibration, 25% validation) were utilised for RF and all wavelengths have been included in the RF 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 500, 2, and 2, 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).

2.6 Evaluation of model performance

The performance of TPH prediction models were assessed using: (i) the coefficient of determination in prediction R^2, (ii) root mean square error of prediction (RMSEP), (iii) residual prediction deviation (RPD) which is a ratio of standard deviation (SD) to RMSEP, and (iv) the ratio of the performance to interquartile distance (RPIQ) which is expressed as the difference between the third and first per root mean square error (RMSE) (Bellon-Maurel and McBratney, 2011). In this study, we adopted (Viscarra et al., 2006) model classification criterion RPD < 1.0 indicates very poor model predictions, 1.0 ≤ RPD < 1.4 indicates poor, 1.4 ≤ RPD < 1.8 indicates fair, 1.8 ≤ RPD < 2.0 indicates good, 2.0 ≤ RPD < 2.5 indicates very good, and excellent if RPD > 2.5. In general, a good model prediction would have high values of R^2 and RPD, and small value of RMSEP.

3. Results and discussion

3.1. Soil chemical analyses

A summary of the soil samples physicochemical properties and TPH concentration determined by GC-MS is provided in Table 1 and Figure 2.

Table 1 Soil properties and TPH concentrations of the soil samples collected

	No	Min.	Mean	Median	1st Qu.	3rd Qu.	Max.	SD
TOC (%)	30	1.11	4.55	3.85	1.79	5.71	12.69	3.30
pH	30	5.20	6.25	5.95	5.73	6.73	8.20	0.83
Sand (%)	30	0.83	25	25	14	33	57	15
Silt (%)	30	19	45	49	34	57	71	14
Clay (%)	30	13	30	30	19	34	60	12
TPH (mg $kg^{-1)}$	85	16.07	252.59	213.69	120.66	339.27	666.33	165.51

TPH (mg kg^{-1}) = Total petroleum hydrocarbons; 1st Qu. = first quartile; 3rd Qu. = third quartile; SD = standard deviation.

The total organic carbon (TOC) content varies between low to medium with the mean and maximum values of 1.1% and 12.7%, respectively. The TOC content is larger than 2.0% for 70% of samples. Clay content ranged between 13% and 60%, with a mean value of 30%. Silt content is high with minimum and maximum values of 19% and 71%, and samples with silt content >40% comprised 66% of all soil samples. Soil texture varies between sandy clay loam to clay loam according to the United States soil texture classification (Soil Survey Staff, 1999).

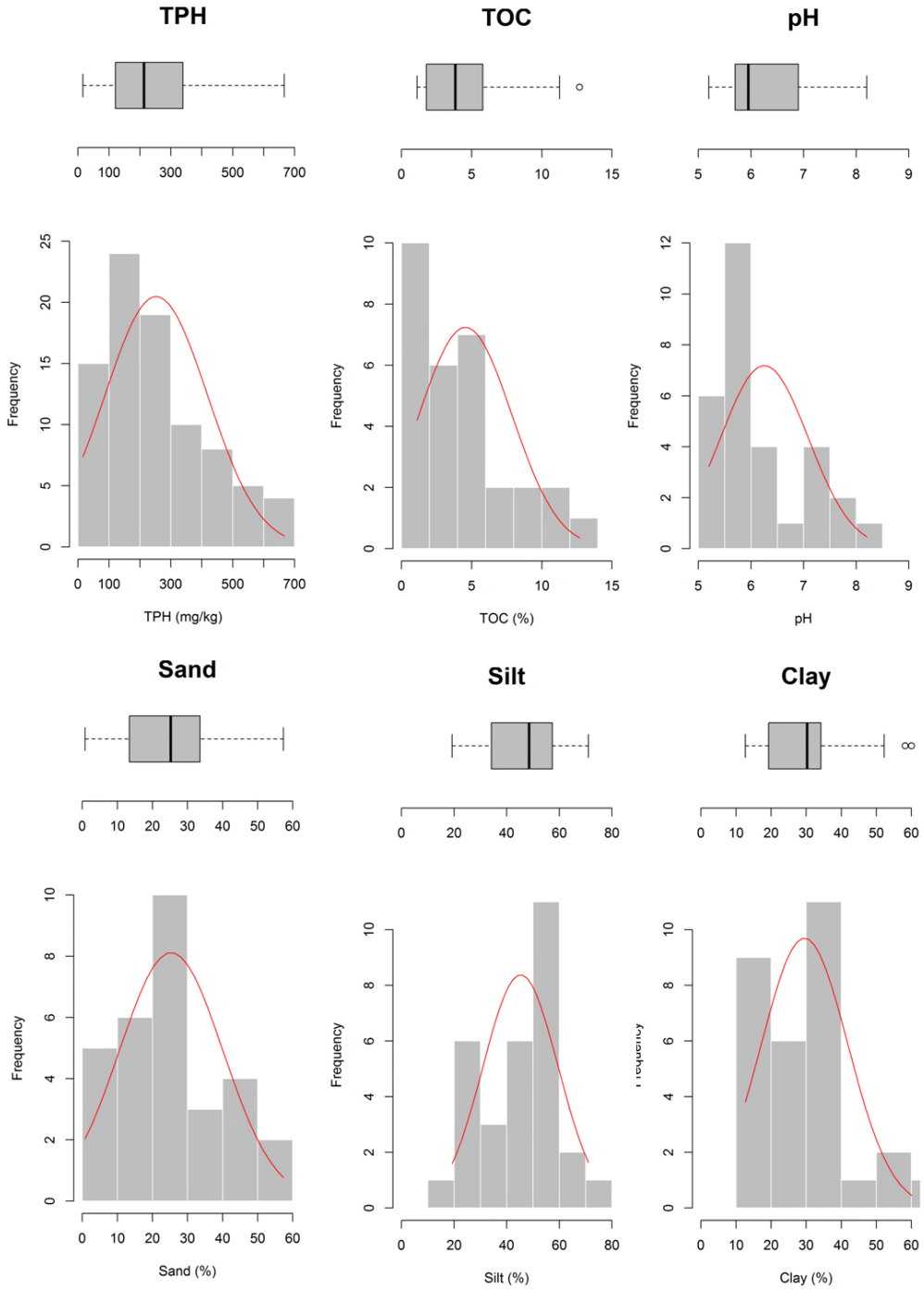

Figure 2. Histograms, box-plots with outliers of total petroleum hydrocarbon (TPH) of 85 soil samples, and total organic carbon (TOC), pH, sand, silt and clay content of selected soil samples (30).

Substantial variability was observed for soil pH ranging between 5.2 and 8.2. The TPH values ranged between 16 and 666 mg kg^{-1} with mean and standard deviations of 253 mg kg^{-1} and 166 mg kg^{-1},

respectively. No significant relationship was identified between TOC, pH, sand, silt, clay, and TPH content (randomization test *p*-values ranged between 0.38 to 0.9 and 0.11 at 0.05 or 0.01 significant level, respectively) (Figure 3).

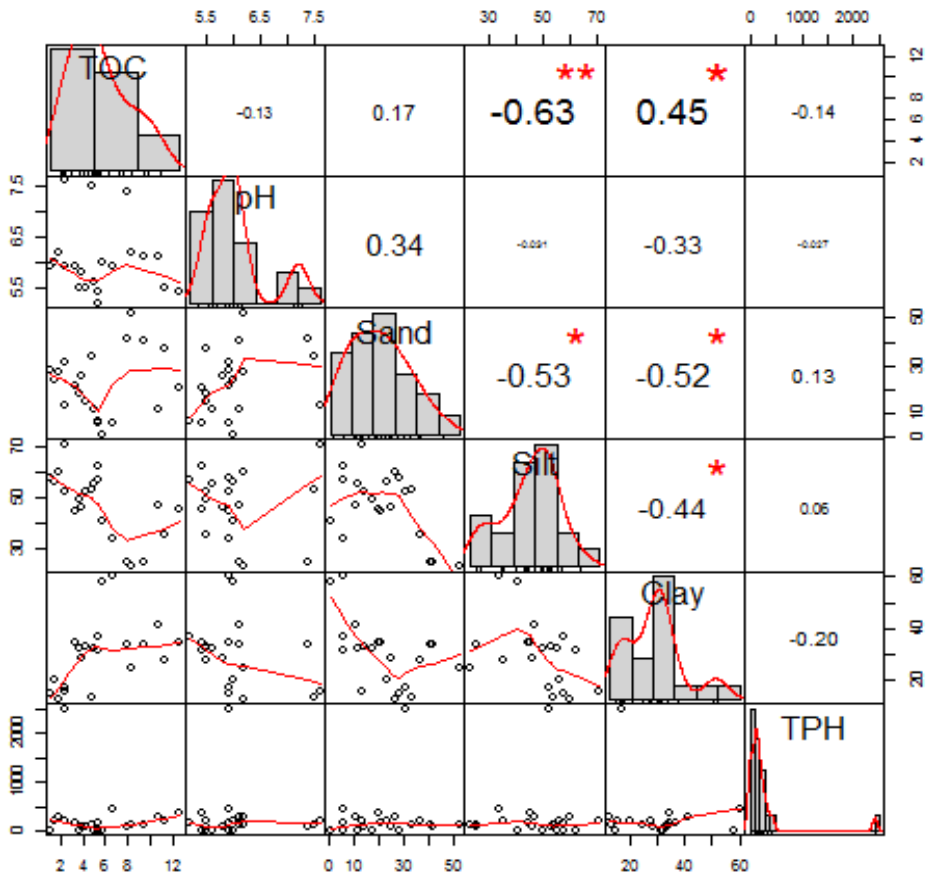

Figure 3. Scatterplot matrix for possible pairs of soil variables (lower diagonal), histograms with kernel density overlays for each the target variable (middle) and absolute value of the correlations at significance level of 0.05 (*) and 0.01(**) between the defined pairs of variables (upper diagonal). Soil variables are total petroleum hydrocarbon (TPH), total organic carbon (TOC), pH, sand, silt and clay content of the selected soil samples (30).

Table 2 shows the average concentrations of the hydrocarbon fractions and the TPH concentration in 85 soil samples. The alkanes and PAH distribution is medium/heavy-end skewed and unimodal with a higher proportion of nC_{16}-C_{21} hydrocarbons suggesting a mid-range distillate heavy oil product type. The average concentrations for the nC_{16}-C_{21} alkanes ranged between 5.4 and 372 mg kg^{-1} and the nC_{16}-

C_{21} PAHs between 0.1 and 2.0 mg kg^{-1} (Table 2). Site 1 had higher average TPH concentration, followed by site 3 and 2. The LOQ for every PAH is shown in Table 3. The lowest and highest LOQ in site 1 were 0.02 and 0.47 mg kg^{-1} for fluorene and acenaphtylene, respectively. In site 2, the lowest LOQ was 0.02 mg kg^{-1} and for Indeno[1,2,3-c,d]anthracene, whereas the highest was 0.26 mg kg^{-1} for Benzo[k]fluoranthrene. While the lowest LOQ was 0.04 mg kg^{-1} for fluorene, the highest was 1 mg/kg and for indeno[1,2,3-c,d]anthracene.

3.2. Spectral analysis of the oil-contaminated

Figure 4 shows the average raw spectra and average continuum removed spectra for uncontaminated samples (n=3) and contaminated samples (n=85). Figure 4 shows average raw reflectance spectra and continuum removed reflectance spectra for uncontaminated and contaminated soil samples, respectively. The average raw spectra and the average of continuum removed spectra for the 85 soil samples showed that oil contaminated soil samples with high TPH content (≥ 654 mg kg^{-1}) and uncontaminated soil sample with TPH below 0.04 mg kg^{-1} (as a control). Overall, the spectrum response (reflectance) pattern is similar for both contaminated and uncontaminated (control) samples, although the contaminated reflects relatively less light (energy). A similar phenomena was reported by Chakraborty et al. (2015) which was related to the higher absorbance of contaminated soils, particularly in the NIR range (700-2500 nm). This finding is in agreement with previous studies (Okparanma and Mouazen, 2013; Chakraborty et al., 2015; Hoerig et al., 2001).There are two distinct absorption peaks at 1415 nm and 1914 nm which are attributed to water absorption overtones, and a third adsorption peak at 2200 nm which is attributed to metal–hydroxyl stretching (Clark et al., 1990). Minima spectral absorption of oil-contaminated soil samples are observed around 1712 and 1758 nm in the first overtone region and around 2207 nm (stretch + bend) in the NIR range (Figure 3). Absorptions around 1712 and 1758 nm are attributed to C-H stretching modes of terminal CH_3 and saturated CH_2 groups

linked to TPH (Workman and Weyer, 2008; Forrester et al., 2010). Similar significant wavebands around 1712 and 1752 nm that were associated to vibrational C-H stretching modes of terminal CH_3 and saturated CH_2 functional chemical groups linked to TPH were reported elsewhere (Okparanma and Mouazen, 2013). The absorption band at 2207 nm can be attributed to either amides (C=O) absorption, or to crude oil spectral signatures (stretch + bend) and therefore linked to hydrocarbons (Mullins et al., 1992). However, these features are practically absent in the uncontaminated reflectance spectra (Fig. 4) which was also confirmed (Chakraborty et al., 2015). Therefore, the absorption bands of hydrocarbons around 1712 and 1758 nm and 2207 nm band can be used to discriminate uncontaminated from contaminated samples (Fig. 4).

Table 2 Hydrocarbon fractions concentration (mg kg^{-1}) and statistics across the three sites (n= 85).

Hydrocarbon fractions (mg/kg)		**Site 1**				**Site 2**				**Site 3**			
		N	Median	Minimum	Maximum	N	Median	Minimum	Maximum	N	Median	Minimum	Maximum
Aliphatic	nC10-nC12	31	6.6	1.6	31	21	11	2.7	36	33	12	0.6	74
	nC12-nC16	31	21	4.7	83	21	18	6.9	53	33	28	2.0	154
	nC16-nC21	31	106	26	372	21	105	33	241	33	83	5.4	314
	nC21-nC35	31	81	15	281	21	90	20	168	33	39	3.7	129
Aromatic	nC12-nC16	31	0.4	0.05	0.7	21	0.1	0.1	0.1	33	0.1	0.1	0.3
	nC16-nC21	31	0.3	0.1	2.1	21	0.3	0.1	1.0	33	0.6	0.2	1.8
	nC21-nC35	31	0.4	0.1	4.7	21	0.3	0.1	1.6	33	3.4	0.3	310
TPH	SUM	31	220	49	666	21	227	65.87	485	33	188	16	619

N=number of samples

Table 3 List of limit of quantification for every study PAH in the three sites.

PAH compounds	Number of rings	Site 1	Site 2	Site 3	LOQ used by laboratories in Nigeria
		LOQ (mg/kg)[a]	LOQ (mg/kg)[a]	LOQ (mg/kg)[a]	LOQ (mg/kg)[b]
Acenaphtylene	3	0.47	0.08	0.15	0.02
Fluorene	3	0.02	0.03	0.04	0.02
Anthracene	3	0.11	0.17	0.63	0.02
Phenantrene	3	0.14	0.08	0.08	0.02
Pyrene	4	0.11	0.06	0.24	0.02
Benz[a]anthracene	4	0.06	0.07	0.12	0.02
Benzo[a]pyrene	5	0.12	0.21	0.78	0.02
Benzo[b]fluoranthrene	5	0.30	0.17	0.54	0.02
Benzo[k]fluoranthrene	5	0.36	0.26	0.77	0.02
Dibenzo[a,h]anthracene	6	0.06	0.03	0.61	0.02
Benzo[g,h,i]perylene	6	0.07	0.03	0.81	0.02
Indeno[1,2,3-c,d]anthracene	6	0.05	0.02	1.00	0.02

LOQ (mg/kg)[a] and LOQ (mg/kg)[b] represents limit of quantification obtained for PAH from this current study and limit of quantification customarily used for PAH in Nigerian laboratories, respectively.

The absorption band at 2207 nm can be attributed to either amides (C=O) absorption, or to crude oil spectral signatures (stretch + bend) and therefore linked to hydrocarbons (Mullins et al., 1992). However, these features are practically absent in the uncontaminated reflectance spectra (Fig. 4) which was also confirmed (Chakraborty et al., 2015). Therefore, the absorption bands of hydrocarbons around 1712 and 1758 nm and 2207 nm band can be used to discriminate uncontaminated from contaminated samples (Fig. 4).

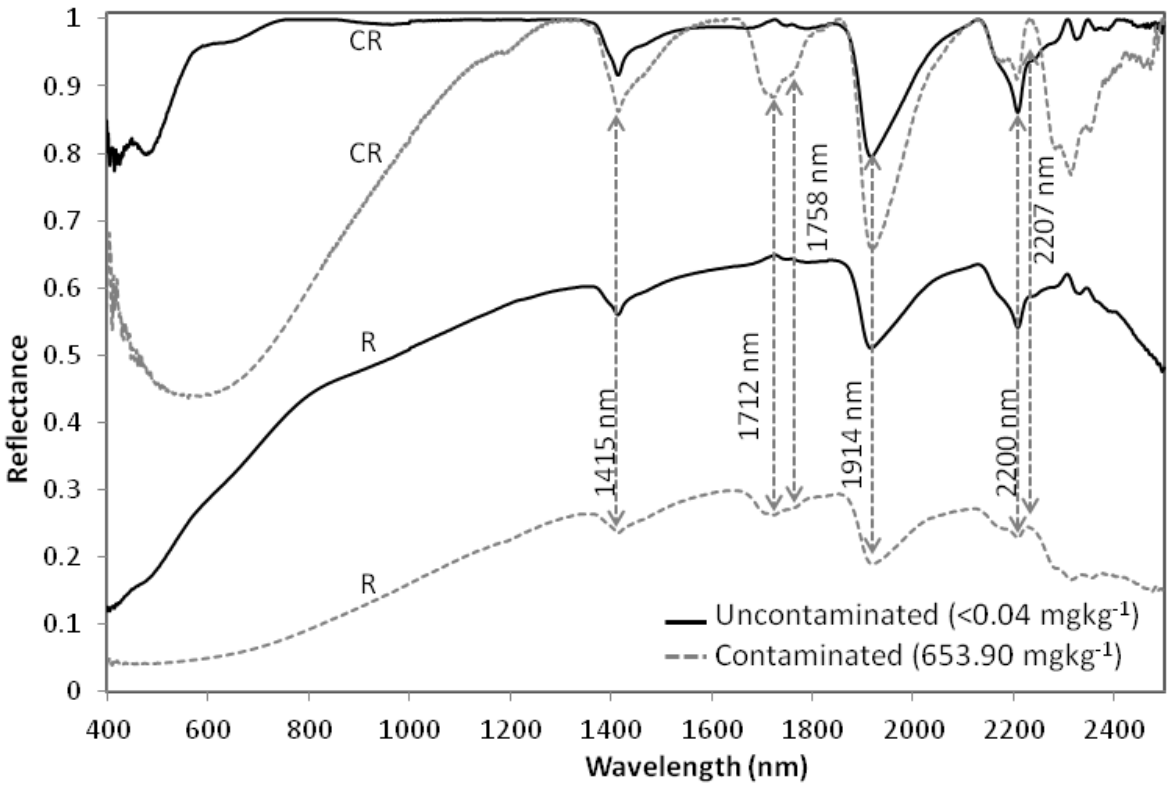

Figure 4. Average of raw (R) and continuum removed (CR) spectra of contaminated (85 samples) vs. uncontaminated soil samples (3 control samples).

The loadings (regression coefficients against wavelengths) based on the first two components (Comp1 and Comp2) resulted from the cross-validated PLSR analysis for TPH are shown in Figure 5. Notably, the numbers and intensities of significant wavelengths have changed, compared to the raw and continuum removed spectra shown in Fig. 3. Significant wavebands from around 1650 to 1850 and from 2250 to 2350 nm can be observed, which can be associated with the 1725 nm (two-stretch) and 2298 nm (stretch + bend) crude oil spectral signatures reported by Mullins et al. (1992). The 1758 nm wavelength is associated with TPH absorption in the first overtone, which is in line with observation of Workman and Weyer (2008) and Osborne et al. (2007) who indicated a significant wavelength for TPH absorption at 1752 nm. Moreover, typical spectral signatures at 1415 nm and 1914 nm were clearly

observed which are associated with the second and first overtones of water absorption bands around 1450 and 1940 nm reported elsewhere (Mouazen et al., 2007).

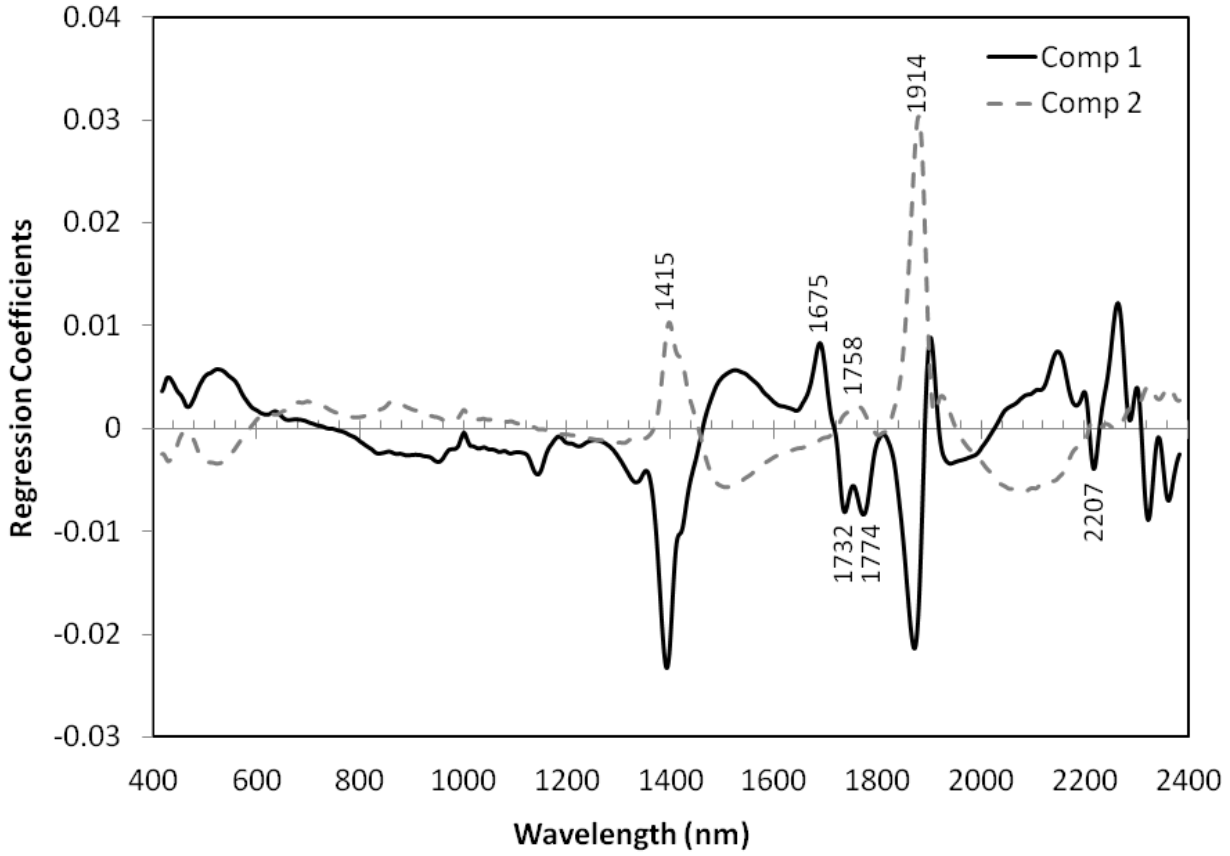

Figure 5. Regression coefficients based on the first and second components (Comp1 and Comp2) versus wavelengths resulted from cross-validated partial least squares regression (PLSR) analysis for total petroleum hydrocarbon (TPH) using visible and near infrared spectroscopy (vis-NIRS) for oil-contaminated soils from Niger Delta, Nigeria. Wavelengths highlighted on the plot are known features of TPH.

3.1 Model performance for estimating TPH from vis-NIR spectra

Table 4 and Figure 6 summarise the cross-validation and prediction models of TPH based on PLSR and RF analyses. In cross-validation, the RF model outperformed the PLSR and resulted in R^2 of 0.85, RMSE of 68.43 mg kg^{-1}, RPD of 2.61 and RPIQ = 3.96.

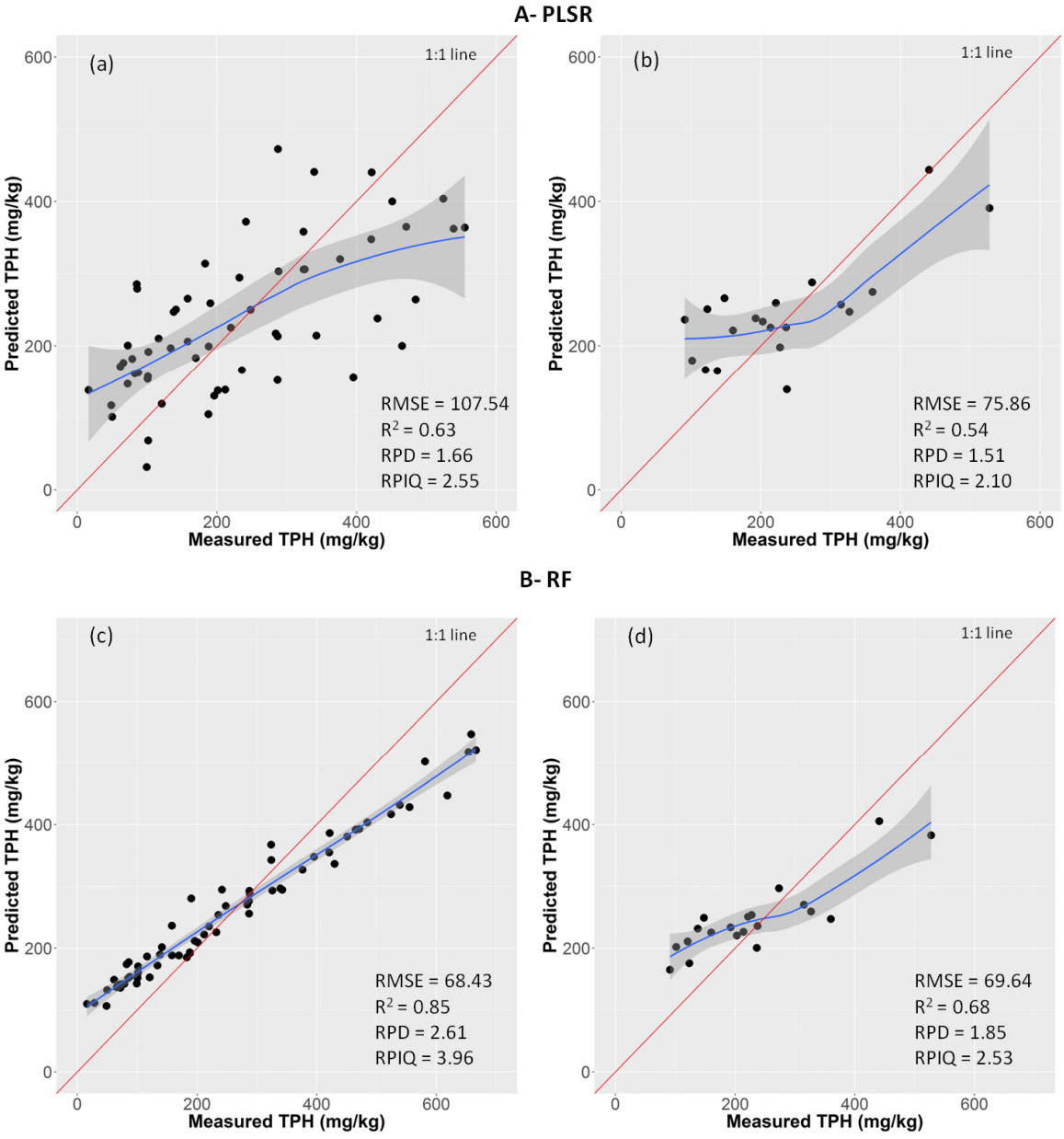

Figure 6. Scatter plots of laboratory measured total petroleum hydrocarbon (TPH) (mg kg^{-1}) by SUSE-GC versus predicted TPH with partial least squares (PLSR) in (a) calibration and (b) prediction models, and random forest (c) in calibration and (d) prediction model. These models were developed using soil samples from three oil-contaminated sites in Niger Delta, Nigeria.

Table 4 Summary results of partial least squares regression (PLSR) and random forest (RF) models in calibration (cross-validation) and prediction for TPH prediction in oil-contaminated soil samples collected from three petroleum-contaminated sites in Niger Delta, Nigeria.

		PLSR					RF				
	N	R^2	RMSEP (mg kg^{-1})	RPD	RPIQ	LV	R^2	RMSEP (mg kg^{-1})	RPD	RPIQ	ntrees
Calibration	65	0.63	107.54	1.66	2.55	8	0.85	68.43	2.61	3.96	500
Prediction	20	0.54	75.86	1.51	2.10	4	0.68	69.64	1.85	2.53	200

N= number of samples, R^2 = coefficient of determination, RMSEP = root mean square error of prediction, RPD = residual prediction deviation, LV = number of latent variables, 'ntrees' = number of trees, and RPIQ = ratio of performance to interquartile range.

The performance of PLSR is the lowest with R^2 of 0.63, RMSE of 107.54 mg kg^{-1}, RPD of 1.66 and RPIQ of 2.55. A similar trend to that of the cross-validation can be observed for the prediction set with both RF (R^2 = 0.68, RMSE = 69.64 mgkg^{-1}, RPD = 1.85 and RPIQ = 2.53) and PLSR (R^2 = 0.54, RMSE = 78.86 mg kg^{-1}, RPD = 1.51 and RPIQ = 2.10).

Our results for RF prediction are better than those reported by Chakraborty et al. (2015) using 108 contaminated soil samples (West Texas, USA) subjected to RF analysis alone (R^2 = 0.53, RMSE = 95.6 mgkg^{-1}, RPD = 1.48 and RPIQ = 1.91) and RF combined with penalized spline regression (PSR) (R^2 = 0.78, RMSE= 0.53 mgkg^{-1}, RPD = 2.19 and RPIQ = 0.75). Similarly, our RF results are better than those reported by Hoerig et al. (2001). For PLSR, Chakraborty et al. (2010 and 2015) reported slightly higher RPD values of 1.69 and 1.7, respectively, for in field-moist soils using PLSR. This difference with our results can be attributed to the combination of spectral treatment that represents an important phase in multivariate calibration and enhances the model performance (Nawar et al., 2016; Buddenbaum and Steffens, 2012; Mouazen et al., 2010). Moreover, Stenberg et al. (2010) and Wang et al. (2010) reported that the model performance depends to a large extent on the variability encountered in the dataset, including soil types, which was the case in our study (16 - 666 mg kg^{-1}), while this was not the case in the two studies conducted by Chakraborty et al. (2010 and 2015) where the original

TPH values were widely and non-normally distributed (44 to 48 mg kg^{-1} and 1.22-3.74×10^9 mg kg^{-1}, respectively). Also, the high variation of TOC (1.1-12.7%) in our study may increase the performance for estimating the TPH (Table 1). It is worth to note that the lower prediction performance observed in this study for PLSR compared to RF might be attributed to the non-linear behaviour of the spectral response of the data set. This feature was not accounted for by the linear PLSR model (Nawar and Mouazen, 2017). In contrast, the RF was capable to handle well the nonlinearity of the dataset of this study. According to RPD classification suggested by Viscarra Rossel et al. (2006), good predictions for TPH are obtained using RF (RPD = 1.85), whereas only fair prediction performance is obtained with PLSR (RPD = 1.51). These results are consistent with our study and previous studies (Okparanma et al. 2014).

4. Conclusions

In this study, we compared the performance of random forest (RF) and partial least squares regression (PLSR) modelling methods to predict total petroleum hydrocarbon (TPH) in fresh soil samples collected from three oil-contaminated sites in Niger Delta, Nigeria. Much better prediction results were achieved by RF with coefficient of determination (R^2), root mean square error of prediction (RMSEP) and ratio of prediction deviation (RPD) of 0.68 and 69.64 mg kg^{-1}, and 1.85, respectively, compared to PLSR with 0.54 and 75.86 mg kg^{-1}, and 1.51 values, respectively. The R^2, RPD, and RMSEP values obtained herein by RF models confirm its suitability as 'a good model prediction' for the estimation of soil properties. The better performance of RF may be attributed to the fact that RF had the advantage of handling the different sources of non-linearity that apparently exist in the studied dataset. There is a strong indication that vis-NIR spectroscopy signal acquisition followed by RF algorithm can be trusted for real application in hydrocarbon analysis in petroleum-contaminated sites where limited data are available. However we recommend that future work should compare other non-linear calibration

methods including artificial neural network, support vector machine, and PSR, among others, to select the best algorithm for the prediction of soil petroleum hydrocarbons.

Acknowledgements: The authors gratefully acknowledge the Petroleum Technology Development Fund (PTDF) of Nigeria (PTDF/OSS/PHD/DRK/711/14).

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.

Bellon-Maurel, V., McBratney, A., 2011. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in solids-Cretical review and research perspectives. Soil Biol. Biochem. 43, 1398-1410.

Brassington, K.J., Pollard, S.T.J., Coulon, F., 2010. Weathered hydrocarbon wastes: a risk assessment primer," in Handbook of hydrocarbon and Lipid Microbioloy In: Timmis, K.N., McGenity, T., Van Der Meer, J.R., De Lorenzo, V. (Eds.), Handbook of Hydrocarbon and Lipid Microbiology. Springer Berlin, 2488–2499.

Breiman, L., 2001. Random Forests. Mach. Learn 45, 5-32.

Brevik, E. C., Burgess, L.C., 2013. Soils and Human Health. Taylor Francis Press, Boca Raton, FL (Eds).

Buddenbaum, H., Steffens, M., 2012. The effects of spectral pretreatments on chemometric analyses of soil profiles using laboratory imaging spectroscopy. Appl. Environ. Soil Sci. 1–12.

Chakraborty, S., Weindorf, D. C., Morgan, C. L. S., Ge, Y., Galbraith, J. M., Li, B., Kahlon, C. S., 2010. Rapid identification of oil-contaminated soils using visible near-infrared diffuse reflectance spectroscopy. Journal of Environmental Quality. 39, 1378–1387.

117 Chakraborty, S., Weindorf, D.C., Li, B., Aldabaa, A.A.A., Gosh, R.K., Paul, S., Ali, M.N., 2015.
118 Development of a hybrid proximal sensing method for rapid identification of petroleum
119 contaminated soils. Science of the Total Environment. 514, 399-408.
120 Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G., Vergo, N., 1990. High spectral resolution
121 reflectance spectroscopy of minerals. J. Geophys. Res. 95, 12653-12680.
122 Coulon, F., Wu, G., 2017. Determination of petroleum hydrocarbon compounds from soils and
123 sediments using ultrasonic extraction In: Hydrocarbon and Lipid Microbiology Protocols McGenity
124 T.J et al. (eds.) Springer-Verlag Berlin Heidelberg, 31- 46.
125 Drozdova, S., Ritter, W., Lendl, B., Rosenberg, E., 2013. Challenges in the determination of petroleum
126 hydrocarbons in water by gas chromatography (hydrocarbon index). Fuels. 113, 527-536.
127 Efron, B., 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7 (1): 1-26.
128 Forrester, S., Janik, L., McLaughlin, M., 2010. An infrared spectroscopic test for total petroleum
129 hydrocarbon (TPH) contamination in soils, Proceedings of the 19th World Congress of Soil Science,
130 Soil Solutions for a Changing World, Brisbane, Australia, August 1–6, 13–16.
131 Geladi, P., Kowalski, B.P., 1986. Partial least-squares regression: A tutorial. Analytica Chimica Acta
132 185(1), 1-17.
133 Hoerig, B., Kuehn, F., Oschuetz, F., Lehmann, F., 2001. HyMap hyperspectral remote sensing to
134 detect hydrocarbons. Int. J. Remote Sens. 8, 1413–1422
135 Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137-
136 148.
137 Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.I., Myers, D.B., Harris, W.G., 2015.
138 Modeling soil carbon fractions with visible near-infrared (VNIR)and mid-infrared (MIR)
139 spectroscopy. Geoderma. 239-240, 229-239.

Kogbe, C.A., 1989. The Cretaceous and Paleogene sediments of Southern Nigeria. In: C.A. Kogbe (Ed.), Geology of Nigeria, Elizabethan Press, Lagos. 311-334.

Liaw, A., Wiener, M., 2015. Breiman and Cutler's Random Forests for Classification and Regression. R package version n 4.6-12 available on https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

Mitra,S., 2003. Sample Preparation Techniques in Analytical Chemistry''. Wiley and Sons, Inc., Publication, Hoboken, NJ, USA.

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 Tillage Res. 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 Till. Res. 93 (1), 13–27.

Mullins, O.C., Mitra-Kirtley, S., Zhu, Y., 1992. The electronic absorption edge of petroleum. Appl. Spectrosc. 46, 1405–1411.

Naes, T., Isaksson, T., Fearn, T., Davies, T., 2002. A user friendly guide to multivariate calibration and classification. NIR Publications. Chichester, UK.

Nawar, S., Buddenbaum, J., Hill, J. K., 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.

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

Okparanma, R. N., Mouazen, A. M., 2013. Combined effects of oil concentration, clay and moisture contents on diffuse reflectance spectra of diesel-contaminated soils'', Water, Air and Soil Pollut. 224 (5), 1539-1556.

Okparanma, R.N., Coulon, F., Mouazen, A.M., 2014. Analysis of petroleum-contaminated soils by diffuse reflectance spectroscopy and sequential ultra sonic solvent extraction-gas chromatography. Environmental Pollut. 184, 298-305.

Osborne, B.G., Fearn, T., Hindle, P.H., 2007. Practical NIR Spectroscopy with Applications in Food and Beverage Analysis, second ed. Longman Group UK Limited, England.

Peters, J., DeBaets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., DeBecker, P., Huybrechts, W., 2007. Random Forests as a tool for ecohydrological distribution modelling. Ecological Modelling. 207, 304 – 318.

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/).

Risdon, G.C., Pollard, S.J.T., Brassington, K.J., McEwan, J.N., Paton, G.I., Semple, K.T., Coulon, F., 2008. Development of an analytical procedure for weathered hydrocarbon contaminated soils within a UK risk-based framework. Anal. Chem. 80, 7090–7096.

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.

186 Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and Near Infrared
187 Spectroscopy in Soil Science. Adv. Agron. 107, 163-215.
188 Stevens, A., Ramirez Lopez, L., 2013. An introduction to the prospectr package (At: https://cran.r-
189 project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf (Accessed :22 April 2016).
190 Thien, S.J., 1979. A flow diagram for teaching texture by feel analysis. Journal of Agronomic
191 Education. 8:54-55.
192 Viscarra Rossel, R.A., Walvoort, D.JJ., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible,
193 near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous
194 assessment of various soil properties. Geoderma. 131, 59-75.
195 Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse
196 reflectance spectra. Geoderma. 158, 46–54.
197 Wang, J., He, T., Lv, C., Chen, Y., Jian, W., 2010. Mapping soil organic matter based on land
198 degradation spectral response units using Hyperion images. Int. J. Appl. Earth Obs. Geoinf. 12,
199 S171–S180.
200 Wold, S., 2010. Personal memories of the early PLS development. Chemometrics and Intelligent
201 Laboratory Systems 58, 83–84.
202 Workman, Jr., J., Weyer, L., 2008. Practical Guide to Interpretive Near-infrared Spectroscopy. CRC
203 Press, Taylor and Francis Group, Boca Raton, FL, USA.
204 Wright, J.B., Hasting, D.A., Jones, W.B., Williams, H.K., 1985. Geology and Mineral Resources of
205 West Africa, Allen and Unwin Limited, UK, 107.