European Journal of Soil Science, Volume 70, January 2019, pp. 140-150.
DOI:10.1111/ejss.12567

Rapid detection of alkanes and polycyclic aromatic hydrocarbons in oil-contaminated soil with visible near-infrared spectroscopy

R.K. DOUGLAS[a], S. NAWAR[b], M.C. ALAMAR[a], F. COULON[a], A.M. MOUAZEN[a,b]

[a]*School of Water, Energy and Environment, Cranfield University, Cranfield, MK43 0AL, UK, and* [b]*Department of Environment, Ghent University, Coupure 653, 9000 Gent, Belgium*

Running title: *Detection of ALK and PAH using soil spectroscopy*

Correspondence A.M. Mouazen, F. Coulon. E-mail: abdul.mouazen@ugent.be; f.coulon@cranfield.ac.uk

Published by Wiley. This is the Author Accepted Manuscript issued with:
Creative Commons Attribution Non-Commercial License (CC:BY:NC 4.0).
The final published version (version of record) is available online at DOI:10.1111/ejss.12567
Please refer to any applicable publisher terms of use.

Summary

Recent developments and applications of rapid measurement tools (RMT) such as visible near-infrared (vis–NR) spectroscopy confirmed that these technologies can provide 'fit for purpose' and cost effective data for risk assessment and management of oil-contaminated sites. While vis–NIR spectroscopy has been used more frequently to predict total petroleum hydrocarbon (TPH), it has had limited use for polycyclic aromatic hydrocarbons? (PAHs) and there have been none for alkanes. In the present study, the potential of vis–NIR spectroscopy (350–2500 nm) to measure PAHs and alkanes in 85 fresh (wet, unprocessed) oil contaminated soil samples collected from three sites in the Niger Delta, Nigeria was evaluated. The vis–NIR signal and laboratory measured alkanes and PAHs with sequential ultrasonic solvent extraction followed by gas chromatography-mass spectrometry (GC-MS) were then used to develop calibration models using partial least squares regression (PLSR) and random forest (RF) modelling tools. Prior to model development, the pre-processed spectra were divided into calibration (75%) and prediction (25%) sets. Results showed that the prediction performance of RF calibration models for both alkanes (a coefficient of determination [R^2] of 0.58, a root mean square error of prediction (RMSEP) of 53.95 mg kg^{-1} and a residual prediction deviation (RPD) of 1.59) and PAHs (R^2 = 0.71, RMSEP = 0.99 mg kg^{-1} and RPD = 1.99) outperformed PLSR (R^2 = 0.36, RMSEP = 66.66 mg kg^{-1} and RPD = 1.29, and R^2 = 0.56, RMSEP = 1.21 mg kg^{-1} and RPD = 1.55, respectively). The RF modelling approach accounted for nonlinearity of the soil spectral responses and therefore resulted in considerably greater prediction accuracy than the linear PLSR. Adoption of vis–NIR spectroscopy coupled with RF is recommended for rapid and cost effective assessment of PAHs and alkanes in

contaminated soil.

Keywords. Petroleum hydrocarbons, vis–NIR spectroscopy, chemometrics, site investigation.

Highlights

• We evaluated the potential of vis–NIR to estimate alkanes and PAHs in oil-contaminated soil.

• The prediction performance of RF models was better than PLSR models for both alkanes and PAHs.

• The spectral response to alkanes and PAHs in soil considerable non-linearity.

• Results suggest that RF-vis–NIR is a promising tool for rapid *in situ* assessment of soil alkanes and PAHs.

Introduction

Petroleum hydrocarbon (PHC) contamination in soil is an important worldwide issue and has attracted serious concerns about the risks to human health and ecosystem health. The major sources of PHC pollution in the soil environment are oil spills from production, storage and distribution of petroleum products. The PHCs encompass hundreds of various aromatic and aliphatic compounds as well as traces of heterocyclic compounds (nitrogen,

hydrogen, sulphur), which are well-known environmental contaminants (Coulon *et al*., 2010). The determination of PHCs in contaminated environmental matrices is a challenge to standardize because of the requirements of different jurisdictions. However, most modern risk assessment methodologies for contaminated sites dictate a risk-based approach and, hence, determination and quantification of particular species of contaminants and fractions are required (Ferguson, 1999). With millions of contaminated locations globally (Horta *et al*., 2015), there is a need for efficient, cost effective, portable and rapid tools for measurement and real-time analysis of PHCs in soil.

Over the last two decades, laboratory techniques have been developed for analysing soil contamination in the laboratory, which are time consuming and expensive (Okparanma & Mouazen, 2013; Chakraborty *et al*., 2015). Furthermore, laboratory techniques require prior sample analysis, extraction and sometimes required cleaning the contaminants non-volatile compounds from the GC injection ports and columns (Forrester *et al*., 2013). Among the laboratory techniques, gas chromatography with flame ionization detector (GC-FID) and gas chromatography-mass spectrometry (GC-MS) are the most common choices for the determination of PHCs in soil using extraction solvents such as dichloromethane or hexane, which pose some human health and environmental risk hazard (Okparanma & Mouazen, 2012). To analyse petroleum-contaminated soils rapidly, optical sensors are recommended (Okparanma *et al*., 2014a). Several studies have assessed the potential of optical techniques for the rapid estimation of PHC concentration in soil (e.g. Bray *et al.*, 2009; Okparanma & Mouazen, 2012, 2013; Okparanma *et al*., 2014a, 2014b; Wartini *et al*., 2017). For example, Bray *et al*. (2009) used the logistic regression method to predict total polycyclic aromatic hydrocarbon

(PAH) and benzo[a]pyrene with visible and near infrared (vis–NIR) spectroscopy, and achieved a good accuracy (90%). Okparanma & Mouazen (2012) used a vis–NIR sensor (350-2500 nm), coupled with partial least squares regression (PLSR) to quantify PAH in diesel-contaminated soil. Their results were good with a root mean square error of prediction (RMSEP) of 0.20 mg kg^{-1}, ratio of prediction deviation (RPD) of 2.75 and a coefficient of determination (R^2) of 0.89. In a later study, Okparanma & Mouazen (2013) assessed the applicability of the same vis–NIR sensor to predict phenanthrene in soil based on PLSR and reported R^2 values of 0.75 and 0.83, RPD values of 2.0 and 2.32, and RMSEP values of 0.21 and 0.25 mg kg^{-1} for validation and calibration, respectively. In another study with the same vis–NIR sensor coupled with PLSR, Okparanma *et al*. (2014a) recorded promising results for the measurement of PAH in the Niger delta, Nigeria with R^2, RPD, and RMSE values of 0.77 and 0.89, 1.86, and 3.12, and 1.16 and 1.95 mg kg^{-1} for validation and calibration, respectively. In monitoring hydrocarbon contamination, Okparanma *et al*. (2014b) investigated the applicability of vis–NIR coupled with PLSR for mapping PAH and the total toxicity equivalent concentration (TTEC) of PAH mixtures at different petroleum-discharge sites in the Niger Delta. They found that there were no significant (P >0.05) discrepancies between the soil maps PAH and TTECs obtained from vis –NIR-based prediction data. Wartini *et al.* (2017) predicted total recoverable hydrocarbon (TRH) concentration in 72 field contaminated soils with a coefficient of determination of cross-validation (R^2cv) of 0.75. More studies on the use of vis–NIR spectroscopy for the prediction of TPH and PAH in soil are described in a recent review by Douglas *et al*. (2017); however, the review included no studies on the predictions of alkanes. Although the prediction of TPH and PAH based on spectroscopy

and multivariate techniques has increased markedly, to the best of our knowledge there is nothing in the literature on the prediction of aliphatic fractions and alkanes (nC_{10-35}) in oil-contaminated sites.

The aim of this study was to report on the performance of a vis–NIR spectrometer for the detection of alkanes and PAH in oil-contaminated soil in the Niger Delta, Nigeria. The prediction performance of the linear PLSR model was compared with a nonlinear random forest (RF) model, to determine the best accuracy that can be achieved with this portable technology.

Materials and methods

The study area and sampling

The study area is in the Niger Delta, Nigeria (Figure 1). The studied fields are located in Bayelsa and Rivers State (Ikarama 6.4519^{o} and 6.4527^{o}E, 5.1538^{o} and 5.1542^{o}N; Kalabar: 6.4502^{o} and 6.4511^{o}E, 5.1369 and 5.1357^{o}N; Joinkrama: 6.1213 and 6.1224^{o}E, 4.9213^{o} and 4.9314^{o} N). It is characterized by a tropical rain forest climate with two seasons: the rainy season lasts for about seven months between April and October with an overriding dry period in August (known as the August break); the dry season lasts for about four months, between November and March. The temperature varies between 25 and 35 $^{\circ}$C in August. The regional geology of the Niger Delta is relatively simple; it consists of the Benin, Agbada (the kitchen of kerogen) and Akata formations, overlain by various types of Quaternary deposits (Wright *et al*., 1985). Soils of the area studied were classified according to the United State Department of Agriculture (USDA) (Soil Survey Staff, 2010) soil taxonomy into two orders, i.e. Inceptisols and Entisols, which include

four subgroups of *Typic Dystrudepts, Aeric Endoaquepts, Typic Udipsammerts and Typic Psammaqnents* (Udoh *et al*., 2013). Soil texture fractions were determined by the international pipette method (Piper, 1950); the results indicated different soil textures for the three sites. According to the USDA textural classification system (Soil Survey Staff, 1999), textures were clay and silty clay loam at the Ikarama site, silt loam at the Kalabar site, and clay loam and sandy clay loam for the Joinkrama site. A total of eighty five (n = 85) petroleum-contaminated soil samples were collected from the top 0–15 cm of three sites (Ikarama 31 samples, Kalabar 21 samples and Joinkrama 33 samples) with a shovel. Each sample taken from a sampling point was homogenized on-site with a hand trowel. We adopted a direct sampling approach to cover as much of the visible hot-spots in the contaminated sites. Soil samples were kept in air-tight centrifuge tubes and stored in cool boxes with ice blocks to avoid hydrocarbon volatilization and to preserve field-moist status until shipment to Cranfield University for further analysis.

Vis–NIR spectra acquisition and pre-processing

Prior to vis–NIR scanning of the soil in the laboratory, each soil sample was further homogenized using a spatula. To obtain optimal diffuse reflection, and therefore, a good signal-to-noise ratio, all plant and pebble particles were removed manually, and the surface was smoothed gently with a spatula before scanning. An ASD LabSpec2500® Vis–NIR spectrophotometer (Analytical Spectral Devices, Inc., Boulder, Colorado USA) with a spectral range of 350 to 2500 nm was used for spectral data acquisition in the laboratory. The equipment was started and allowed to warm up for at least 30 minutes and calibrated with a 100% Teflon white reference before soil spectral measurement, which was repeated at 30-minute intervals. The white reference measurement aimed to

avoid, and possibly remove dark current and effects from variation in ambient temperature and humidity as reported by Chakraborty *et al*. (2010). From each of the eighty five homogenized soil samples, three subsamples were packed into plastic Petri dishes for the acquisition of vis–NIR spectra. The vis–NIR spectra of all the samples were recorded with the spectrometer by placing the sample in direct contact with a high intensity probe with a built-in quartz-halogen bulb of 3000 $^{\circ}$K, enclosing a 20° angle with a detection fibre. The three replicates of each sample were scanned at three different places on the sample, and an average spectrum was obtained. The raw average spectra were subjected to pre-processing including successively, noise reduction, maximum normalization, first derivative and smoothing using the prospector-R package (Stevens and Ramirez, 2014). First, the spectral range outside 400–2345 nm was excised to remove the noise at both edges. Then, a moving average with five successive wavelengths was used to reduce noise. Maximum normalization followed, which is typically used to place all data at approximately the same scale. Spectra were then subjected first to Gap-segment derivative (gapDer) algorithms (Norris, 2001) with a second-order polynomial approximation. Finally, Savitzky–Golay smoothing was carried out to remove noise from the spectra and to decrease the detrimental effect on the signal-to-noise ratio that conventional finite-difference derivatives would have.

Hydrocarbon analysis

Hydrocarbon extraction was performed as described by Risdon *et al*. (2008) with some modifications. Total extractable and recoverable petroleum hydrocarbons (TERPH), aliphatic and aromatic fractions were identified and quantified using a GC-MS (Agilent 5973N, Santa Clara, California) system operated at 70 eV in positive ion mode. The GC

was fitted with a Restek RTX–5MS capillary column (30-m long, 0.25-mm internal diameter and 0.25-µm coating). 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 minute^{-1}, then to 310 °C at 6 ºC minute^{-1} and held at this temperature for 15 minutes. The mass spectrometer was operated using the full scan mode (range m z^{-1} 50-500) for quantitative analysis of target alkanes and PAHs. For each compound, quantification was performed by integrating the peak at specific m z^{-1}. External multilevel calibrations were carried out for both oil fractions, with values ranging from 0.5 to 2500 µg ml^{-1} for alkanes? and from 1 to 5 µg ml^{-1} PAHs. Internal standards for the alkanes were nonadecane-d_{40}, triacontane-d_{62} and napthalene d_8, phenanthracene-d_{10}, chrysene-d_{12} and perylene d_{12} (Sigma Aldrich, Gillingham, UK). For quality control, a 500 µg ml^{-1} standard diesel and mineral oil were analysed every 20 samples. In addition, duplicate soil control and reference material**s** were systematically used. The soil control was treated following the same procedure for samples without adding soil samples. The reference material was an uncontaminated soil of known characteristics, and was spiked with a standard diesel and mineral oil at a concentration equivalent to 16 000 mg kg^{-1}.

Model development

The pre-processed spectra and the laboratory (GC-MS) measured chemical data for alkanes and PAHs were used to develop calibration models. In this study we compared the prediction performance of two modelling techniques, namely PLSR and RF regression. Before the analyses, outliers were detected by box plot (Figure 2) and removed (5 and 2 samples of PAHs and alkanes were removed, respectively), after which the dataset was divided into calibration and prediction sets (58 and 23 for PAHs, and 65

and 18 for alkanes, respectively) using the Kennard–Stone algorithm (Kennard & Stone, 1969). The resulting calibration models developed with the calibration set were validated using the prediction samples.

It is well known that PLSR is the most commonly used multivariate regression technique available in standard statistical and chemometrics software. It is a bilinear modelling method where information in the original *x* data is projected on to a small number of underlying ('latent') variables called PLSR components. The *y* data are actively used in estimating the latent variables to ensure that the first components are those that are most relevant for predicting the *y* variables. Interpretation of the relation between the *x* data and *y* data is then simplified because this relation is concentrated on the smallest possible number of components (latent variables). More detailed information about the PLSR can be found in (Martens and Naes, 1989). In this study, alkanes and PAHs represented *y*, whereas the dependent variables (wavelengths) represented *x* and were used as regression generators for the independent variables (Mouazen *et al*., 2006). In this study, we used PLSR analysis with leave-one-out cross validation (LOOCV) to develop calibration models for alkanes and PAHs with the pls-R package (R Core Team, 2013), in order to annul the possible effect of under-fitting or over-fitting data (Efron & Tibshirani, 1993). The maximum number of components used in the PLSR was five for alkanes and six for PAHs.

Ensemble learning like RF well known as a method for classification and regression, which generates many classifiers and aggregates their results (Breiman, 2001). Tree diversity guarantees stability of the RF model, which is achieved in two ways: (i) a random subset of predictor variables is chosen to 'grow' each tree and (ii) 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 generalization error (Peters *et al*., 2007). The same datasets used in PLSR were used for RF, and all wavelengths were 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, two and three, 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).

Model evaluation

The accuracy of the calibration models for PLSR and RF analyses were evaluated with: (i) the coefficient of determination of prediction R^2, (ii) RMSEP and (iii) RPD, which is a ratio of standard deviation (SD) to RMSEP. In this study, we adopted the proposed classification system for RPD values of Viscarra Rossel *et al*. (2006), which divides the 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).

Results and discussion

Laboratory wet analysis

The distribution and concentrations of the aliphatic fractions and individual PAH across the three sites are summarized in Table 1. The three study sites followed the same trend: nC_{10}–nC_{12} had the smallest values at all the sites, whereas nC_{16}–nC_{21} dominated at all sites. The distribution of hydrocarbons confirms? that the hydrocarbon source at the three sites is weathered (degraded) (Brassington *et al*., 2010). More particularly, the concentration of aliphatic compounds at Site 1 (767.0 mg kg^{-1}) was 1.5 times greater than at Site 2 (498.1 mg kg^{-1}) and 1.1 times greater than Site 3 (671.2 mg kg^{-1}) (Table 1). Conversely, the concentration of aromatic compounds at Site 3 (321.8 mg kg^{-1}) was 97.23 times greater than at Site 2 (3.31 mg kg^{-1}) and 39.98 times greater than Site 1 (8.05 mg kg^{-1}) (Table 1).

Among the three sites studied, Joinkrama and Kalabar were the most and least contaminated sites with aliphatic hydrocarbons, respectively. The only exception was that the maximum concentration of the nC_{10}–nC_{12} in Kalabar was larger than at its counterpart in Ikarama. The concentrations of 3- and 4-ring PAHs ranged from 0.002 to 0.782 mg kg^{-1}, 0.003 to 0.514 mg kg^{-1} and 0.004 to 309.325 mg kg^{-1} at Sites 1, 2 and 3, respectively. The concentration of 5- to 6- ring PAHs ranged from 0.001 to 2.246 mg kg^{-1}, 0.000 to 0.016 mg kg^{-1} and 0.004 to 2.527 mg kg^{-1} at Sites 1, 2 and 3, respectively. The relatively large concentration of Benz[a]anthracene (309.3 mg kg^{-1}) at Site 3 cannot be explained because its degradation has not been documented elsewhere. Overall, Site 3 appeared to be the most contaminated compared to Sites 1 or 2.

A statistical summary of the concentrations of alkanes and PAHs determined by GC-MS, and used for the development of both PLSR and RF models is provided in Table 2 and Figure 2. The concentrations of alkanes varied between small to medium amounts with mean and maximum values of 151.6 and 551.2 mg kg^{-1}, respectively. There were only two samples with values above 512 mg kg^{-1}; both were outliers (Figure 2 a). The concentrations of PAHs ranged from 0.52 to 312.28 mg kg^{-1}, with a mean value of 9.11 mg kg^{-1}. Five outliers were detected (Figure 2 b) and were removed before modelling (Figure. 2 c).

Analysis of regression coefficients

Figure 3 depicts the regression coefficients plotted against wavelength; the coefficients resulted from the cross-validated PLSR analysis for alkanes and PAHs. Plots of the regression coefficients illustrate important wavelengths or bands that associate with properties or compounds to be predicted, in this case alkanes and PAHs. Figure 3(a) shows two absorption bands in the alkanes plot around 1716 and 2306 nm. The absorption band around 1716 nm in the first overtone region is characteristic of TPH. The absorption feature around 2306 nm is attributed to the long-chain C–H+C–C stretch combinations, which is related to $–CH_2$ aliphatic groups. This accords with the range reported by Wartini *et al*. (2017) for petroleum-contaminated soil (2300-2340 nm). For PAHs, two distinct spectral absorption peaks can be identified around 1688 and 1736 nm in the first overtone region of the NIR spectral range (Figure 3b). The absorption around 1688 nm is attributed to C–H stretching modes of ArCH associated with PAHs, whereas the absorption around 1736 nm is attributed to C–H stretching modes of terminal CH_3 and saturated CH_2 chemical group characteristic of TPH. The absorption bands around

1400 and 1900 nm in Figure. 3(a, b) are attributed to O–H stretching bands in the second and first overtone regions, respectively. The TPH absorption bands identified in the regression coefficients plots accord with the results reported elsewhere (Wartini *et al*., 2017; Okparanma *et al*., 2014a), whereas the PAHs absorption bands are similar to those of Okparanma *et al*. (2014a) and Workman and Weyer (2008). The absorption bands around 1394, 1873 and 1881 nm identified in this study compare well with the results of Stenberg *et al*. (2010) and Whalley and Stafford (1992), and they are associated with O–H stretching modes of water in the second (1394 nm) and first overtone (1873 and 1881 nm) regions, respectively. However, the largest absorption bands were those associated with water at the first and second absorption overtones of O–H, whereas those associated with PAHs were significantly smaller.

Figure 4 shows an average raw vis–NIR spectrum of oil-contaminated soils spectra collected from three sits, where smaller absorption features associated with hydrocarbons (1712, 1758, 2207 and 2302 nm) and larger features associated with water (1415 and 1914 nm) were identified. Interestingly, these wavelengths agree with spectral features of hydrocarbons and water observed in plots of the regression coefficients (Figure. 3). The wavelengths 1712 nm and 1758 nm are close to those reported by Okparanma and Mouazen (2013) (1712 and 1759 nm), Okparanma *et al*. (2014a) (1712 and 1752 nm) and Douglas *et al*. (2018) (1712 and 1758 nm) for hydrocarbon contaminated soils. The wavelength of 2207 nm is near to those reported by Chakraborty *et al*. (2015) and Forrester *et al*. (2013) (2220 nm), whereas the wavelength of 2302 nm is close to 2298, reported for hydrocarbon contamination in soils by Mullins *et al.* (1992). The wavelength of 2302 nm could also be attributed to soil minerals (Viscarra Rossel *et al*., 2006).

Prediction performance for alkanes and PAHs

The results of prediction performance of PLSR and RF for alkanes and PAHs are listed in Table 3 and shown in Figures. 5 and 6. The results show clearly that RF outperformed PLSR in both the cross-validation and prediction. For alkanes, RF cross-validation results (R^2 = 0.85, RMSEP = 55.71 mg kg^{-1} and RPD = 2.58) are typically better than the prediction results (R^2 = 0.58, RMSEP = 53.59 mg kg^{-1}, and RPD = 1.59). It is clear that PLSR performed poorly and resulted in R^2 of 0.49, RMSEP of 101.7 mg kg^{-1} and RPD of 1.41 in cross-validation, and of 0.36, 66.66 mg kg^{-1} and 1.29, respectively, in prediction (Table 3). With the RPD classification system of Viscarra Rossel *et al*. (2006) to evaluate prediction performance of the models, suggested that the predictions for alkanes based on an RF were between fair to excellent (RPD = 1.59–2.58), whereas the prediction performance of PLSR models was classified as poor to fair (RPD = 1.29–1.41). There is no other study yet on the use of vis–NIR spectroscopy to predict alkanes in soil, therefore, we could make no comparison of our results with independent literature. However, the prediction performance here suggests that there more research is needed to improve the model outputs, and to understand why the prediction was not in the good to excellent categories. One reason might be the limited number of samples used in the current research (85 samples for calibration and validation). Kuang and Mouazen (2013) showed that the prediction accuracy for soil total nitrogen and total carbon could be improved with the increase in number of samples that added (spiked) into a general calibration set.

Figure 6 shows scatter plots of measured against predicted PAHs values in cross-validation and prediction. Again, RF produced better results than PLSR. Unlike alkanes,

the prediction performance of PAHs with RF indicated excellent performance in cross-validation (R^2 = 0.89, RMSEP = 1.02 mg kg^{-1} and RPD = 2.99), and good performance in prediction (R^2 = 0.71, RMSEP = 0.99 mg kg^{-1} and RPD = 1.99). Results also showed that the PLSR model performed better for PAHs than for alkanes, with good performance in cross-validation (R^2 = 0.76, RMSEP = 0.81 mg kg^{-1} and RPD = 2.07) and fair performance in prediction (R^2 = 0.56, RMSEP =1.21 mg kg^{-1}, and RPD = 1.55) (Table 3 and Figure. 6). The better performance of RF compared to PLSR can be attributed to the fact that the RF modelling technique typically yields better results when the relation between reflectance and concentration is a nonlinear (typical in soils) (Nawar *et al*., 2016; Douglas *et al*., 2018), whereas the PLSR model fits only linear relations (Nawar *et al*., 2016). Results obtained with PLSR are not as good as those already reported in the literature. Okparanma *et al*. (2014a) reported an RPD range of 1.86-3.12 using soil samples from the Niger delta, whereas Okparanma and Mouazen (2013) reported a range of 1.67 - 3.20. An RPD value of 2.75 was reported by Okparanma and Mouazen (2012). The fair to good performance observed in this study with PLSR might also relate to the small number of samples used in the present study, compared to those reported elsewhere.

Conclusions

We have shown the potential application of visible and near infrared (vis-NIR) spectroscopy and chemometrics for the prediction of alkanes and polycyclic aromatic hydrocarbons (PAHs) in oil-contaminated soil samples collected from three contaminated sites in the Niger Delta, Nigeria. Our results revealed that prediction performance depended on the modelling techniques used, and that RF outperformed PLSR for the

prediction of both properties in both cross-validation and prediction. The RF models' prediction performance of alkanes and PAHs was classified as fair and good, respectively, whereas PLSR models' performance was poor for alkanes (and only fair for PAHs. The better performance of RF was its ability to deal with non-linearity in the dataset used in this study. Nevertheless, the small number of soil samples in this study might have affected the model performance at both the calibration and prediction stages. This was particularly so for RF at the prediction stage, wheeras the model provided much better results in cross-validation than in prediction. In contrast, the PLSR model performance slightly only deteriorated between cross-validation and prediction. Further work is being undertaken to improve the prediction accuracy of vis–NIR spectroscopy coupled with the RF nonlinear modelling approach by using the existing Nigerian contaminated soil spectral library and spiking technique.

Acknowledgements

The authors gratefully acknowledge the Petroleum Technology Development Fund (PTDF) of Nigeria (PTDF/OSS/PHD/DRK/711/14) and the Flemish Scientific Research (FWO) funded SiTeMan Odysseus I Project (Nr. G0F9216N).

The underlying data can be accessed at https://doi.org/10.17862/cranfield.rd.5794842.v1

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.

Brassington, K.J., Pollard, S.T.J. & Coulon, F. 2010. Weathered hydrocarbon wastes: a risk assessment primer. In: *Handbook of Hydrocarbon and Lipid Microbiology* (eds K.N. Timis, T. McGenity, J.R. Van Der Meer & V. De Lorenzo), pp. 2488-2499. Springer, Berlin.

Bray, J.G.P., Viscarra Rossel, R. & McBratney, A.B. 2009. Diagnostic screening of urban soil contaminants using diffuse reflectance spectroscopy. *Soil Research*, **47**, 433–442.

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

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.

Chakraborty, S., Weindorf, D.C., Li, B., Aldabaa, A.A.A., Gosh, R.K., Paul, S. & Ali, M.N. 2015. Development of a hybrid proximal sensing method for rapid identification of petroleum contaminated soils. *Science of the Total Environment*, **514**, 399–408.

Coulon, F., Whelan, M.J., Paton, G.I., Semple, K.T., Villa, R. & Pollard, S.J.T. 2010. Multimedia fate of petroleum hydrocarbons in the soil: oil matrix of constructed biopiles. *Chemosphere,* **81**, 1454–62.

Douglas, R. K., Nawar, S., Alamar, M.C., Coulon, F. & Mouazen, A.M. 2017. Almost 25 years of chromatographic and spectroscopic analytical method development for petroleum hydrocarbons analysis in soil and sediment: State-of-the-art, progress and trends. *Critical Reviews in Environmental Science and Technology*, **47(16)**, 1497–1527.

Douglas, R. K., Nawar, S., Alamar, M. C., Coulon, F. & Mouazen, A. M. 2018. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. *Science of the Total Environment*, **616-617**, 147–155.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7 (1)**: 1–26.

Efron, B., Tibshirani, R.J., 1993. An Introduction into the Bootsrap. Chapman & Hall 29 West 35th Street New York, NY 10001–2299.

Ferguson, C.C. 1999. Assessing risks from contaminated sites:Policy and practice in 16 European Countries. *Land Contamination & Reclamation*, **7**, 33–54.

Forrester, S.T., Janik, L.J., McLaughlin, M.J., Soriano-Disla, J.M., Stewart, R. & Dearman, B. 2013. Total Petroleum Hydrocarbon Concentration Prediction in Soils Using Diffuse Reflectance Infrared Spectroscopy. *Soil Science Society of America Journal*, **77(2)**, 450–460.

Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T.F.A., McBratney, A.B., Pallasser, R. & Pozza, L. 2015. Potential of integrated field spectroscopy and spatial

analysis for enhanced assessment of soil contamination: A prospective review. *Geoderma,* **241-242**, 180–209.

Kennard, R.W. & Stone, L.A. 1969. Computer aided design of experiments. Technometrics, **11**, 137–148.

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.

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: 14/01/2018).

Martens, H. & Naes, T. 1989. *Multivariate Calibration*, second ed. John Wiley and Sons, Chichester, UK.

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

Mullins, O.C., S. Mitra-Kirtley, and Y. Zhu. 1992. The electronic absorption edge of petroleum. *Applied. Spectroscopy,* **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, 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.

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

Okparanma, R.N. & Mouazen, A.M. 2012. Risk-based characterisation of hydrocarbon contamination in soils with Visible and near-infrared diffuse reflectance spectroscopy., In: Soil and Water Engineering. International Conference of Agricultural Engineering-CIGR-AgEng 2012: *Agriculture and Engineering for a Healthier Life,* Valencia, Spain, 8-12 July 2012. pp. C–0657.

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. 2014a. Analysis of petroleum-contaminated soils by diffuse reflectance spectroscopy and sequential ultrasonic solvent extraction-gas chromatography. *Environmental Pollution*, **184**, 298–305.

Okparanma, R.N., Coulon, F., Mayr, T. & Mouazen, A.M. 2014b. Mapping polycyclic aromatic hydrocarbon and total toxicity equivalent soil concentrations by visible and near-infrared spectroscopy. *Environmental Pollution*, **192**, 162–170.

Peters, J., Debaets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., DeBecker, P. & Huybrechts, W. 2007. Random Forest as a tool for ecological distribution modelling. *Ecological Modelling,* **207**, 304–318.

Piper, C. S. 1950. *Soil and plant analysis*. Interscience. Publ. Inc. New York.

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. *Analytical Chemistry*, **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.

Soil Survey Staff. 2010. Keys to Soil Taxonomy. Washington, D.C.: USDA – NRCS.

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

Stevens, A. & Ramirez Lopez, L. An introduction to the prospectr package; 2014, 1–22. At: https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf. Accessed: 22/04/2016).

Udoh, B.T., Esu, I.E., Ibia, T.O., Onweremadu, E.U. & Unyienyin, S.E. 2013. Agricultural Potential of the Beach Ridge Soils of the Niger Delta, Nigeria. *Malaysian Journal of Soil Science*, **17**, 17–37, (At:

http://www.msss.com.my/mjss/Full Text/Vol17/Udoh.pdf. Accessed: 14/1/2018)

Viscarra Rossel, R.A., McGlynn, R.N. & McBratney, A.B. 2006. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. *Geoderma*, **137**, 70–82.

Wartini, Ng., Brendan, P.M. & Budiman, M. 2017. Rapid assessment of petroleum-contaminated soils with infrared spectroscopy. *Geoderma,* **289**, 150–160.

Whalley, W.R. & Stafford, J.V. 1992. Real-time sensing of soil water content from mobile machinery: Options for sensor design. *Computers and Electronics in Agriculture*, **7**, 269–358.

Workman, Jr., J. & Weyer, L. 2008. *Practical Guide to Interpretive Near-infrared Spectroscopy*. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA.

Wright, J.B., Hasting, D.A., Jones, W.B. & Williams, H.K. 1985. *Geology and Mineral Resources of West Africa*, PP. 154–164. Springer, Dordrecht, Netherlands.

Figure captions

Figure 1 Soil sampling locations for the three contaminated sites in the Niger Delta (Bayelsa and Rivers State), Nigeria (After: Douglas *et al*., 2018).

Figure 2 Histograms and box-plots of concentrations for (a) alkanes with outliers, (b) polycyclic aromatic hydrocarbons (PAHs) with outliers and (c) PAHs without outliers, of the eighty five soil samples from the Niger Delta, Nigeria.

Figure 3 Plots of regression coefficient from partial least squares regression (PLSR) analysis for (a) alkanes and (b) polycyclic aromatic hydrocarbons (PAHs), based on visible and near infrared (vis–NIR) spectra of oil-contaminated soil samples from three sites in the Niger Delta, Nigeria. Wavelengths highlighted on the plot are the potential features of PAHs and alkanes.

Figure 4 Average raw visible and near infrared (vis–NIR) spectrum of oil-contaminated soil spectra from three crude oil spill sites in the Niger Delta region of Nigeria. Wavelengths of 1712, 1758, 2207 and 2302 nm are associated with hydrocarbons, whereas 1415 and 1914 nm are absorption features of water in the second and first overtones, respectively.

Figure 5 Values of alkanes measured with gas chromatography mass-spectrometry (GC-MS) plotted against predicted concentrations from visible and near infrared (vis–NIR) spectroscopy based on partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and random forest (RF) in (c) cross-validation and (d) prediction for samples from the Niger Delta, Nigeria.

Figure 6 Scatter plots of the measured polycyclic aromatic hydrocarbons (PAHs) using

gas chromatography mass-spectrometry (GC-MS) *versus* visible and near infrared (vis-NIR) spectroscopy predicted concentrations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and (B) random forest (RF) method in (c) cross-validation and (d) prediction for samples from the Niger Delta, Nigeria.

Table 1 Statistical summary of the concentrations of alkanes and polycyclic aromatic hydrocarbons (PAHs) for the three contaminated sites from the Niger Delta, Nigeria.

compound	LOQ/mg kg^{-1}	Ikarama				Kalabar				Joinkrama			
		N	Med.	Min.	Max.	*N*	Med.	Min.	Max.	*N*	Med.	Min.	Max.
			mg kg^{-1}				mg kg^{-1}				mg kg^{-1}		
nC10-nC12 Ali	0.02	31	6.59	1.52	31.46	21	11.34	2.65	35.52	33	12.45	0.59	73.77
nC12-nC16 Ali	0.02	31	21.42	4.70	83.19	21	18.42	6.89	52.84	33	27.59	1.92	154.14
nC16-nC21 Ali	0.02	31	106.4	26.26	371.53	21	105.49	32.76	241.39	33	83.23	5.38	314.32
nC21-nC35 Ali	0.02	31	80.52	15.07	280.78	21	89.73	20.10	168.38	33	39.00	3.65	128.97
∑Alkanes			214.97	47.55	766.96		224.98	62.4	498.13		162.27	11.54	671.20
Acenaphtylene	0.02	31	0.375	0.054	0.691	21	0.321	0.083	0.514	33	0.132	0.045	0.319
Fluorene	0.02	31	0.025	0.011	0.122	21	0.019	0.005	0.041	33	0.037	0.004	0.085
Anthracene	0.02	31	0.111	0.034	0.397	21	0.111	0.023	0.330	33	0.286	0.088	0.982
Phenantrene	0.02	31	0.124	0.038	1.121	21	0.100	0.021	0.364	33	0.104	0.013	0.859
Pyrene	0.02	31	0.059	0.014	0.545	21	0.093	0.030	0.262	33	0.120	0.019	1.070
Benzo[a]pyrene	0.02	31	0.049	0.005	0.948	21	0.068	0.016	0.495	33	0.445	0.024	1.940
Benzo[b] Fluoranthrene	0.02	31	0.099	0.006	0.957	21	0.062	0.016	0.420	33	0.460	0.037	2.527
Benzo[k]-fluoranthrene	0.02	31	0.028	0.006	2.246	21	0.030	0.004	0.516	33	0.695	0.004	2.150
Benz[a]anthracene	0.02	31	0.027	0.002	0.782	21	0.031	0.003	0.170	33	0.052	0.005	309.325
Dibenzo[a,h] anthracene	0.02	31	0.011	0.002	0.073	21	0.014	0.001	0.067	33	0.406	0.009	0.765
Benzo[g,h,i] perylene	0.02	31	0.007	0.001	0.076	21	0.010	0.000	0.066	33	0.323	0.008	0.805
Indeno [1,2,3-c,d] anthracene.	0.02	31	0.017	0.002	0.094	21	0.021	0.004	0.065	33	0.340	0.015	0.996
∑PAHs			0.932	0.175	8.052		0.88	0.206	3.310		3.399	0.271	321.823
TREPH			215.90	47.73	775.01		225.86	62.61	501.44		165.67	11.81	993.02

Med., median; Min., minimum; Max., maximum; LOQ, limit of quantification, defined as the lowest concentration at which an analyte can be reliably detected (Mitra, 2003); TREPH (∑Alkanes + ∑PAHs), total recoverable petroleum hydrocarbons; N, number of samples; Ali, aliphatic; Ikarama and Kalabar, soil sampling sites 1 and 2, respectively (Yenagoa Local Government Area Bayelsa State); Joinkrama, soil sampling site 3 (Ahoada East Local Government Area Rivers State).

Table 2 Statistical summary of concentrations of alkanes and polycyclic aromatic hydrocarbons (PAHs) for the soil samples measured with gas chromatography-mass spectrometry (GC-MS). Soil samples were collected from three petroleum-contaminated sites in the Niger Delta, Nigeria.

	N	Min.	Mean	Median	1st Qu.	3rd Qu.	Max.	St. dev
Alkanes/mg kg^{-1}	85	9.9	187.24	151.75	84.55	259.25	551.22	133.13
PAHs/ mg kg^{-1}	85	0.520	9.11	1.39	0.89	4.00	312.28	40.20

1st Qu., first quartile; 3rd Qu., third quartile; St. dev, standard deviation.

Table 3 Prediction performance of partial least squares regression (PLSR) and random forest (RF) models for alkanes and polycyclic aromatic hydrocarbons (PAHs) in oil-contaminated soils from three sites (Ikarama, Kalabar, and Joinkrama) in the Niger Delta, Nigeria, developed using visible and near-infrared (vis–NIR) spectroscopy.

			PLSR				**RF**			
Compound	Model	*N*	R^2	RMSEP /mg kg^{-1}	RPD	LV	R^2	RMSEP /mg kg^{-1}	RPD	*ntrees*
Alkanes	Calibration	65	0.49	101.71	1.41	6	0.85	55.71	2.58	500
	Prediction	18	0.36	66.66	1.29	4	0.58	53.95	1.59	200
PAHs	Calibration	58	0.76	0.81	2.07	6	0.89	1.02	2.99	500
	Prediction	23	0.56	1.21	1.55	4	0.71	0.99	1.99	200

R^2, coefficient of determination; RMSEP, is root mean square error of prediction; RPD, residual prediction deviation; LV, latent variable.

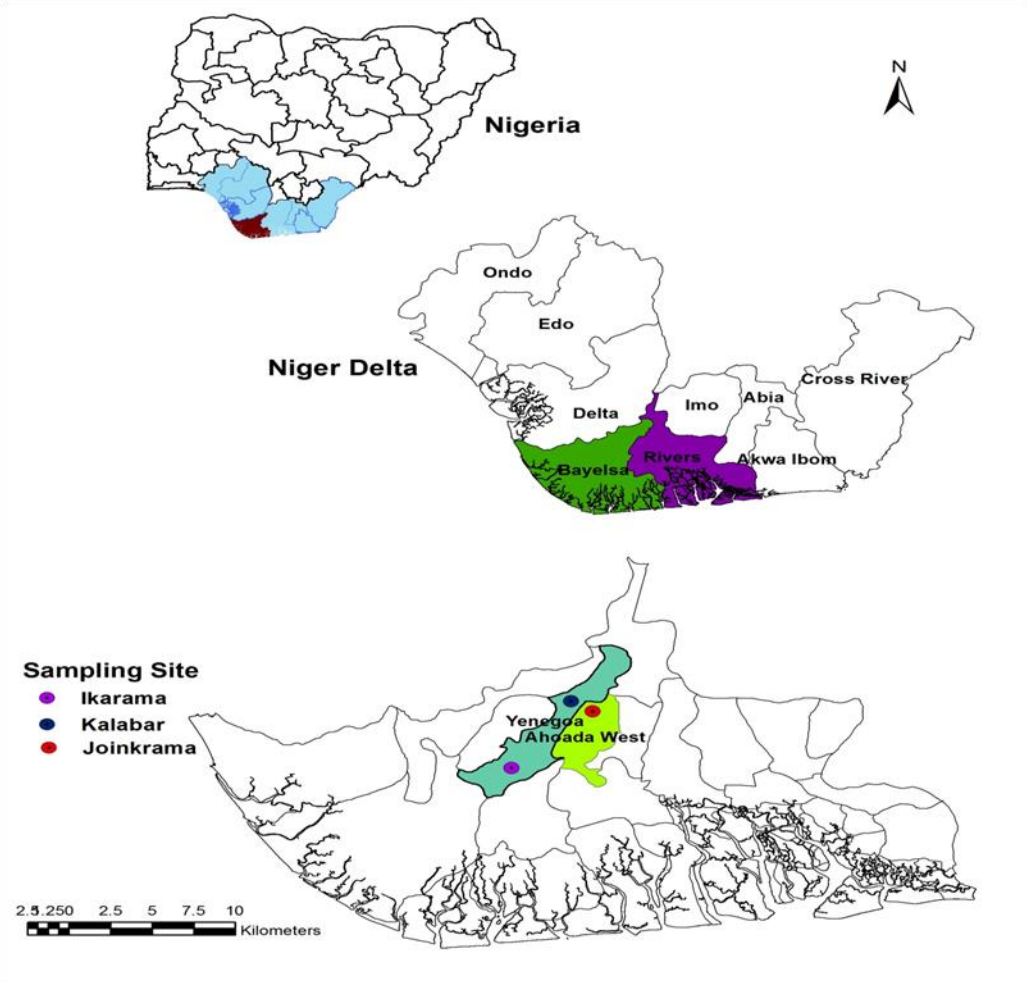

Figure 1 Soil sampling locations for the three contaminated sites in the Niger Delta (Bayelsa and Rivers State), Nigeria (After: Douglas et al., 2018).

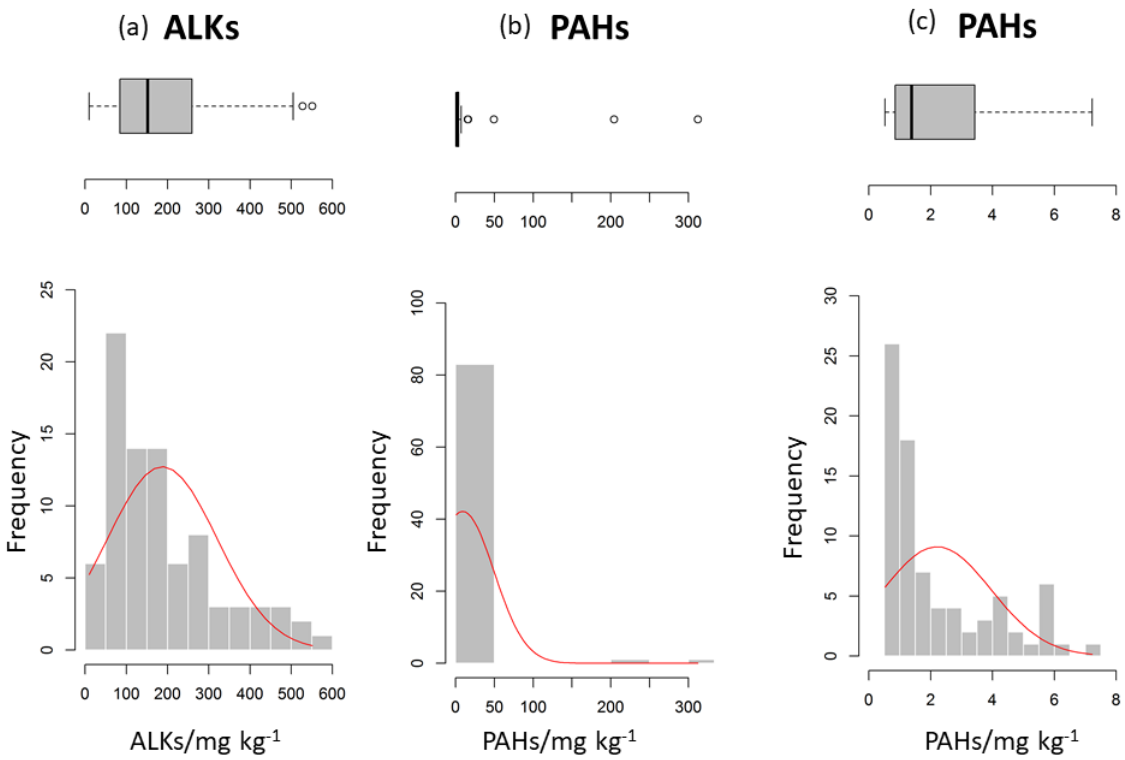

Figure 2 Histograms and box-plots of concentrations for (a) alkanes with outliers, (b) polycyclic aromatic hydrocarbons (PAHs) with outliers, and (c) PAHs without outliers, of the eighty five soil samples from the Niger Delta, Nigeria.

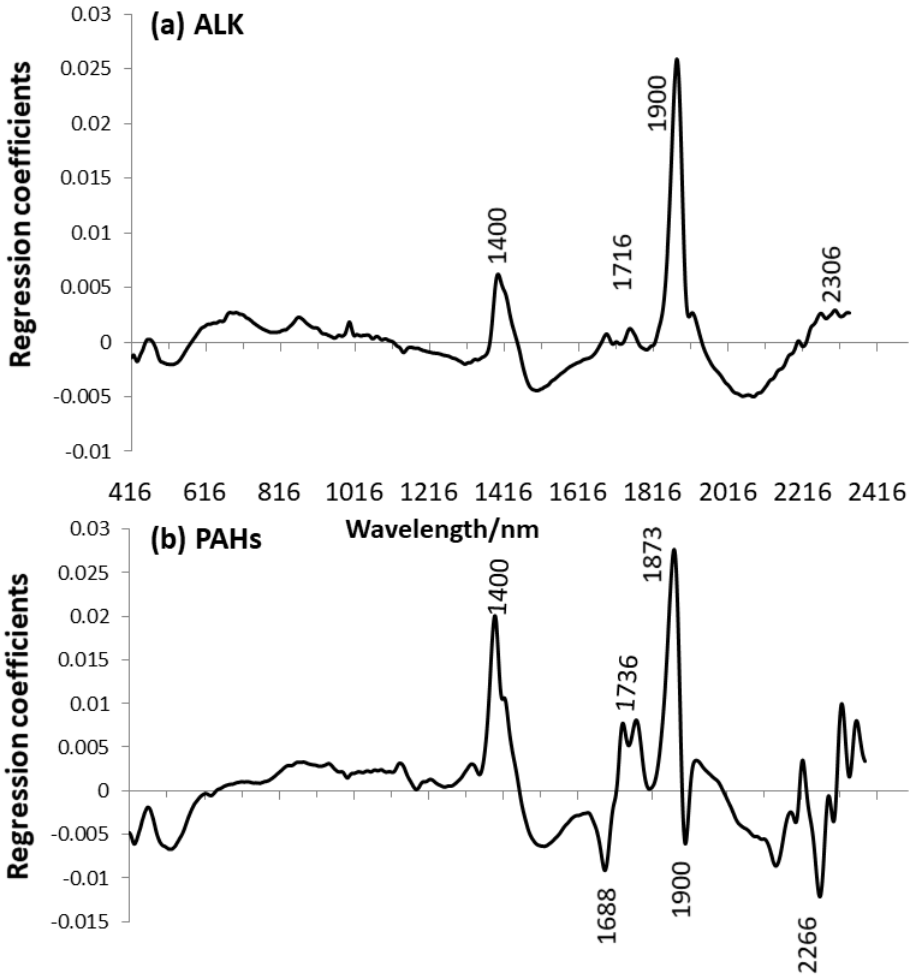

Figure 3 Plots of regression coefficient from partial least squares regression (PLSR) analysis for (a) alkanes and (b) polycyclic aromatic hydrocarbons (PAHs), based on visible and near infrared (vis–NIR) spectra of oil-contaminated soil samples from three sites in the Niger Delta, Nigeria. Wavelengths highlighted on the plot are the potential features of PAHs and alkanes.

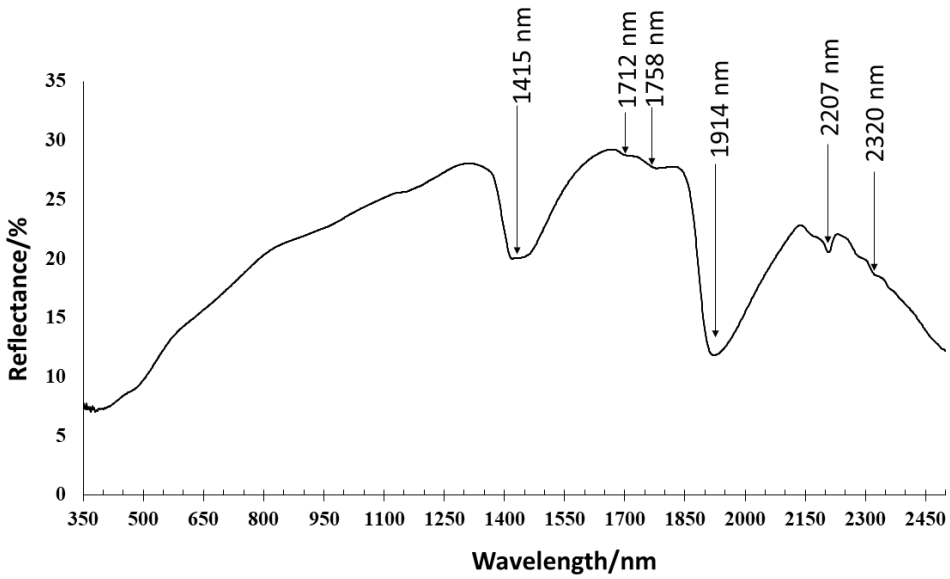

Figure 4 Average raw visible and near infrared (vis–NIR) spectrum of oil-contaminated soil spectra from three crude oil spill sites in the Niger Delta region of Nigeria. Wavelengths of 1712, 1758, 2207, and 2302 nm are associated with hydrocarbons, while 1415 and 1914 nm are absorption features of water in the second and first overtones, respectively.

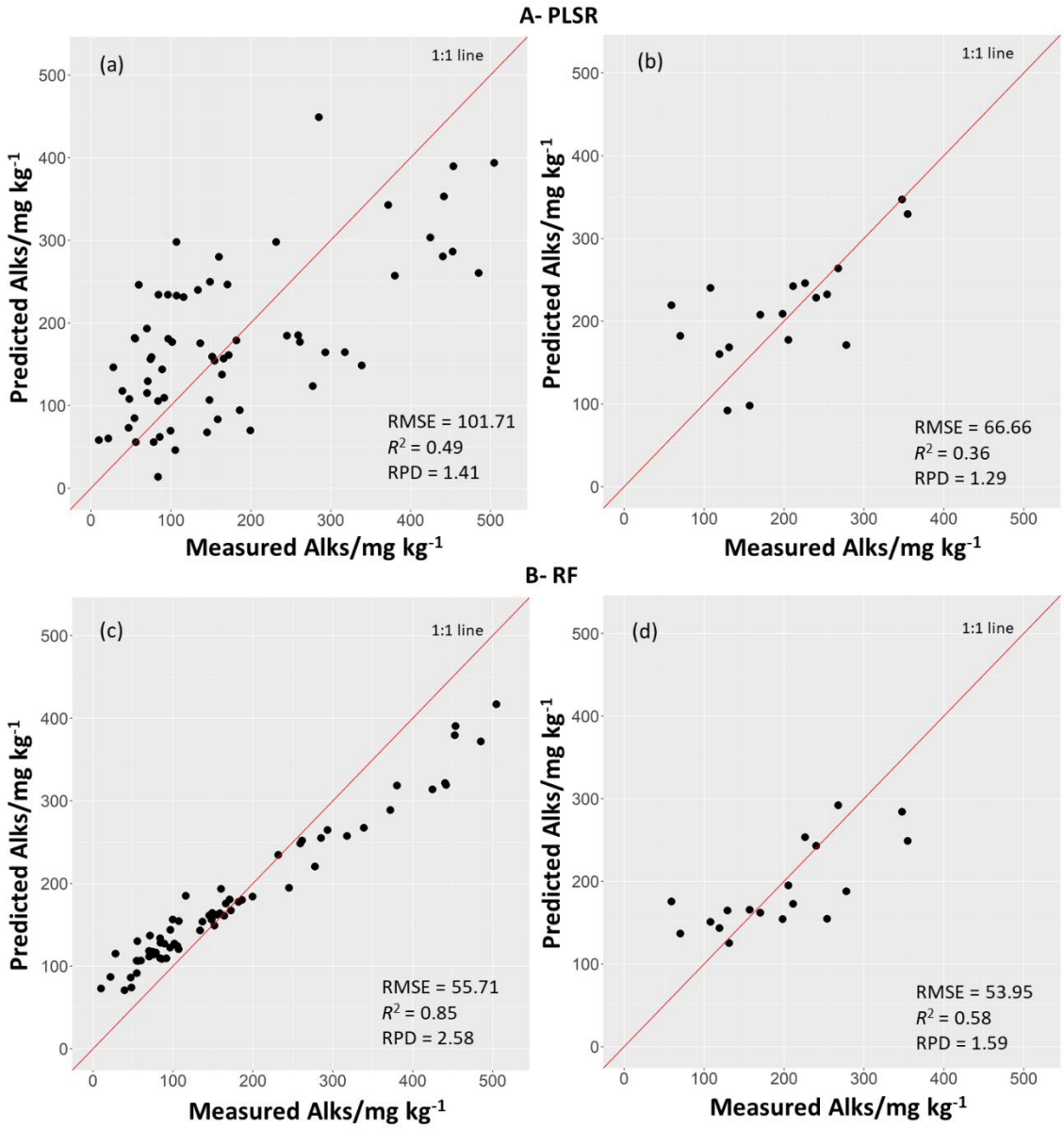

Figure 5 Values of alkanes measured with gas chromatography mass-spectrometry (GC-MS) plotted against predicted concentrations from visible and near infrared (vis–NIR) spectroscopy based on partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and random forest (RF) in (c) cross-validation and (d) prediction for samples from the Niger Delta, Nigeria.

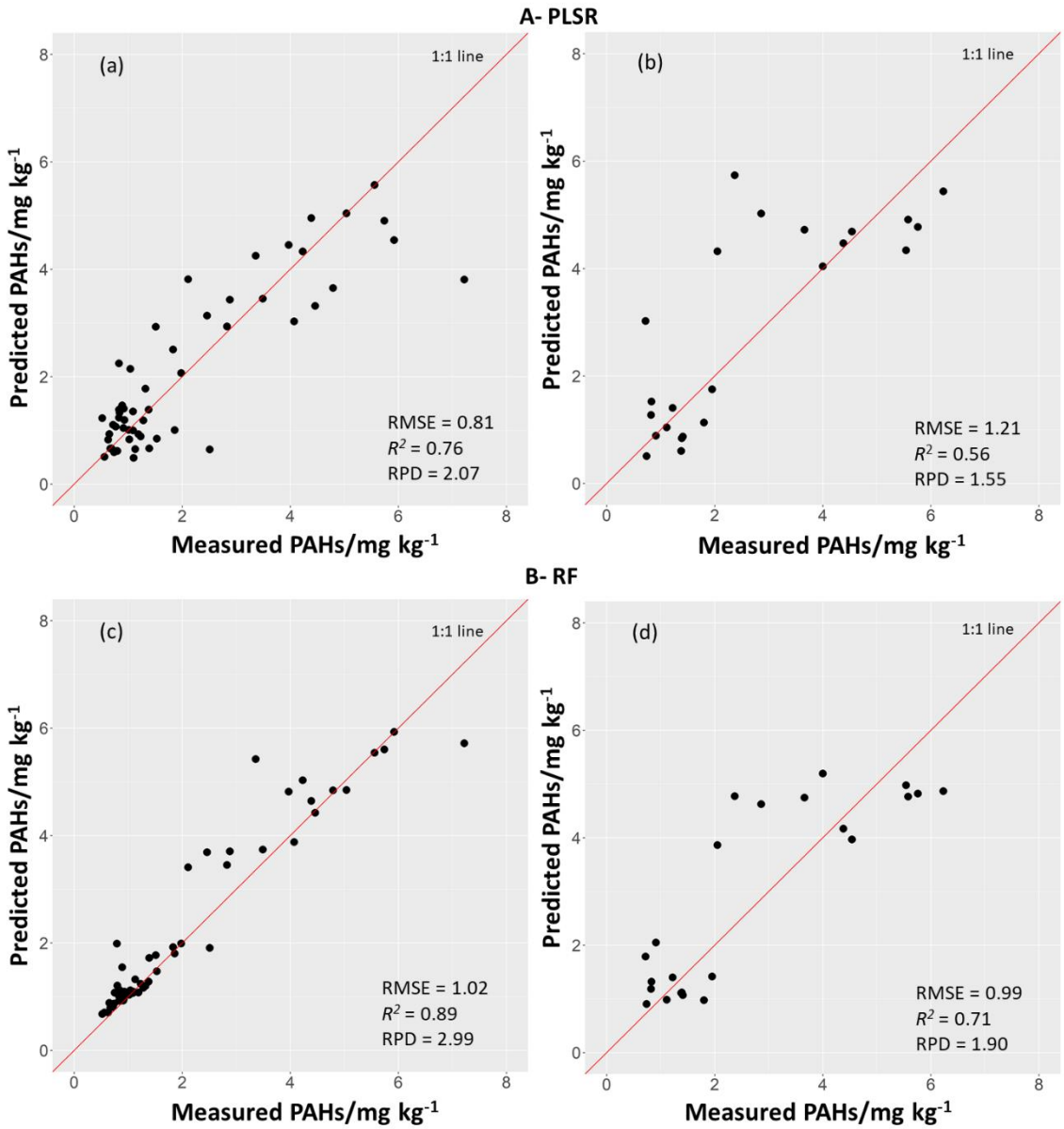

Figure 6 Scatter plots of the measured polycyclic aromatic hydrocarbons (PAHs) using gas chromatography mass-spectrometry (GC-MS) *versus* visible and near infrared (vis-NIR) spectroscopy predicted concentrations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and (B) random forest (RF) method in (c) cross-validation and (d) prediction for samples from the Niger Delta, Nigeria.