Science of the Total Environment, Volume 665, 15 May 2019, Pages 253-261
DOI:10.1016/j.scitotenv.2019.02.065

The application of a handheld mid-infrared spectrometry for rapid measurement of oil contamination in agricultural sites

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

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

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

Corresponding authors:
E-mail: abdul.mouazen@ugent.be; Tel + 32 9 264 6037
Email: f.coulon@cranfield.ac.uk; Tel: +44 (0) 1234 754981

Abstract

Rapid analysis of oil-contaminated soils is important to facilitate risk assessment and remediation decision-making process. This study reports on the potential of a handheld mid-infrared (MIR) spectrometer for the prediction of total petroleum hydrocarbons (TPH), including aliphatic (alkanes) and polycyclic aromatic hydrocarbons (PAH) in limited number of fresh soil samples. Partial least squares regression (PLSR) and random forest (RF) modelling techniques were compared for the prediction of alkanes, PAH, and TPH concentrations in soil samples (n = 85) collected from three contaminated sites located in the Niger Delta, Southern Nigeria. Results revealed that prediction of RF models outperformed the PLSR with coefficient of determination (R^2) values of 0.80, 0.79 and 0.72, residual prediction deviation (RPD) values of 2.35, 1.96, and 2.72, and root mean square error of prediction (RMSEP) values of 63.80, 83.0 and 65.88 mg kg^{-1} for TPH, alkanes, and PAH, respectively. Considering the limited dataset used in the independent validation (18 samples), accurate predictions were achieved with RF for PAH and TPH, while the prediction for alkanes was less accurate. Therefore, results suggest that RF

calibration models can be used successfully to predict TPH and PAH using handheld MIR spectrophotometer under field measurement conditions.

Keywords

Mid-infrared reflectance spectroscopy; petroleum-hydrocarbon contamination; random forest; partial least squares regression.

1. Introduction

Soils contaminated with petroleum hydrocarbons (PHCs) severely impact the environment and wellbeing of the people, and reduce the agronomic potential of agricultural, grasslands and forest lands. PHCs are common environmental contaminants found in soils, waters and sediments (Sammarco et al., 2013; Li et al., 2014). PHCs consist of a complex mixture of aliphatic and aromatic compounds with different concentrations that are considered toxic for human and livestock (Ritchie et al., 2001; Yang et al., 2017). Especially, the polycyclic aromatic hydrocarbon (PAH) fraction contains toxic compounds that can be either adsorbed and further accumulated in the soil, or leach to groundwater and, subsequently, causing significant food chain contamination (Chen et al., 2013). Therefore, environmental pollution as a result of oil spill requires immediate attention and actions to reduce contamination levels and reclaim contaminated lands (Pinedo et al., 2013). The first step towards achieving this urgent goal is by rapid detection methods of PHCs in soils that offer *in situ* measurement with high sampling density to allow spatial and temporal assessment.

Chromatographic techniques, particularly gas chromatography-mass spectrometry (GC-MS), have been a common choice for the measurement of PHCs in environmental samples due to their relative selectivity and sensitivity (Wang and Fingas, 1995; Brassington et al., 2010). However,

GC-MS methods for soil hydrocarbon analysis depend on the use of toxic extraction solvents such as hexane, acetone, dichloromethane (Douglas et al., 2018a, 2018b; Okparanma and Mouazen, 2013). Overall, traditional techniques for the measurement of soil contaminants in the laboratory are slow, expensive and require specific expertise (Viscarra Rossel et al., 2011; Chakraborty et al., 2015; Horta et al., 2015). Thus, there is need for rapid, accurate, and cost-effective measurement tools for PHC concentrations in soils for in-field applications, where there is no need for the use of toxic extraction solvents. The most obvious candidates that offer all the advantages over traditional analytical methods of PHCs are the optical methods (Okparanma and Mouazen, 2013; Douglas et al., 2018a, 2018b).

There are a number of studies that have successfully used optical sensors for the analysis of petroleum-contaminated soils. In analysing soils, these sensors use electromagnetic energy, especially those in the visible and near-infrared (vis-NIR) and mid-infrared (MIR) regions. Both, vis-NIR and MIR spectroscopy, have been used for the analysis of oil-contaminated soils. While the majority of studies were reported on the use on the vis-NIR spectroscopy (e.g., Chakraborty et al., 2010; Okparanma et al., 2014, Chakraborty et al., 2015; Douglas et al., 2017), only few studies have used MIR spectroscopy. For example, Forrester et al. (2010) have successfully determined total petroleum hydrocarbon (TPH) in spiked minerals with both vis-NIR (root mean square error of prediction [RMSEP)] = 4500-8000 mg kg^{-1}) and MIR (root mean square error of prediction of cross validation [RMSEcv] = 2000-4000 mg kg^{-1}) laboratory-based spectroscopy for 0-100 000 mg kg^{-1} TPH range; whereas Forrester et al. (2013) predicted TPH concentration in 205 naturally contaminated soils by laboratory-based MIR methods (RMSE = 601 mg kg^{-1} and ratio of prediction deviation [RPD] = 3.4) and NIR (RMSE = 564 mg kg^{-1} and RPD = 3.7). More recently, Webster et al. (2016) used a handheld MIR instrument in reflectance mode to predict

TPH of three different soil types including a carbonate dominated clay, a kaolinite dominated clay and a loam from Padova Italy, north Western Australia and southern Nigeria, respectively. All samples (Nr = 194) were air-dried before scanning. Successful partial least squares regression (PLSR) predictions, with coefficient of determination (R^2) of 0.99 and RMSE < 200 mg kg^{-1}, were obtained for TPH concentrations ranging between 0 and 3,000 mg kg^{-1}. These predictions were carried out using a set of independent samples for each soil type. Prediction models were also tested for the full concentration range (0 - 60,000 mg kg^{-1}) for each soil type model, obtaining R^2 and RMSE values of 0.99 and < 1,255 mg kg^{-1}, respectively. Portable MIR and vis-NIR spectroscopy were used by Wartini et al. (2017) for rapid prediction of total recoverable hydrocarbon (TRH) in air-dried contaminated soils (n=126), resulting in RMSE of calibration (RMSEcal) values of 1592 and 1881 mg kg^{-1}, respectively. More details on available studies can be found in a recent review of chromatography and spectroscopy for PHCs analysis published by Douglas et al. (2017). To the best of our knowledge, there is no study yet on the application of MIR spectroscopy for the prediction of alkanes and PAHs using limited number of fresh (unprocessed) soil samples. This is essential requirement, for the implementation of field spectroscopy, to explore whether MIR spectroscopy being sensitive to moisture content is capable to predict alkanes and PAHs in fresh (field-moist) soil samples. Therefore, the current study aims at assessing the potential of a handheld MIR instrument for the prediction of alkanes, PAH and TPH in fresh and genuinely contaminated samples (n=85), collected from three agricultural sites in the Niger delta, Nigeria. The prediction performance of the nonlinear machine learning random forest (RF) and the linear partial least squares regression (PLSR) methods was compared.

2. Materials and methods

2.1. Study area and soil sampling

The soil samples were collected from three selected contaminated sites located in the Niger Delta province of Nigeria. For more details on the studied area and samples used (e.g., sampling location, method for sample collection, sampling depth, mass of samples, and sample preservation) readers are referred to Douglas et al. (2018a, b).

2.2. Hydrocarbon analysis

Chemical analysis for hydrocarbon concentrations in soil was carried out at Cranfield University, UK. The sequential ultrasonic solvent extraction gas chromatography (SUSE-GC) (Agilent 5973N GC-MS) was operated at 70 eV in positive ion mode for the analysis as described in Risdon et al. (2008) with some modifications. Briefly, a mixture of 20 mL of hexane (Hex): dichloromethane (DCM) solution (1:1, v/v) was added to 5 g soil sample and was shaken for 16 h at 150 oscillations per min, and finally sonicated for 30 min at 20°C. The validation methodology was set against a robust and validated GC-MS method previously reported (Risdon et al., 2008).

2.3. Quality assurance/quality control (QA/QC)

The present study used the quality assurance (QA)/quality control (QC) protocol prescribed by Risdon et al. (2008). The Method is also mCERTS standards (UK Environment Agency validation). The sample extracts were cleaned on Florisil® columns by elution with hexane. The recovery from the extraction method obtained by spiking dried samples with 1 mL of a surrogate solution containing o-terphenyl (oTP), squalane (Sq), heptamethylnonane (HMN) and 2-fluororbiphenyl (2-Fb) at a concentration of 200 μg mL^{-1} each in acetone, was > 98%. To extract the appropriate concentrations, Deuterated alkanes and PAHs internal standards were added; then

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 introduced to extracts at 0.5 µg mL^{-1} and 0.4 µg mL^{-1}, respectively. Aliphatic hydrocarbons (alkanes) and aromatic hydrocarbons (PAHs) were identified and quantified by GC-MS (Agilent 5973N), operated at 70 eV in positive ion mode. The mass spectrometer was operated under the full scan mode (range *m/z* 50-500) for quantitative analyses of alkanes and PAHs. For each compound, quantification was carried out by integrating the peak at specific *m/z* using auto-integration method by Mass Selective Detector (MSD) ChemStation software. External multilevel calibrations were performed for both oil fractions, and quantification ranged from 0.5 to 2500 µg mL^{-1} and from 1 to 5 µg mL^{-1}, respectively. For QC purpose, a 500 µg mL^{-1} diesel standard and mineral oil were analysed after 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%. The limit of quantification (LOQ) of 0.02 mg kg^{-1} customarily used for PAH in Nigerian laboratories was adopted for the present study, since samples were collected from crude oil spill sites in the Niger Delta, Nigeria.

2.4. MIR spectra collection and pre-processing

The field-moist soil samples were scanned using an Agilent 4300 handheld Fourier transfer infrared (FTIR) spectrometer (Agilent Technologies, Santa Clara, CA, United States), with spectral wavenumber range of 4000 cm^{-1} to 650 cm^{-1} at 8 cm^{-1} resolution and ~2 cm^{-1} sampling interval. The instrument was equipped with a deuterated triglycine sulfate (DTGS) thermal detector, and a zinc selenide (ZnSe) beam splitter which has high mid-infrared throughput and a wide spectral range (Eid et., 2018). This detector relies on only the amount of heat energy

delivered, and its response is independent of wavelength and provides slow and linear response over a very wide range of FT-IR throughput (Theocharous and Birch, 2006). Before any measurement, the spectrometer was warmed up at least for 30 min, and then the three performance tests (signal-to-noise, Stability, and laser frequency calibration) on the instrument using the validated software were performed (Agilent Technologies, 2015). Before spectral measurement, all plant debris were removed from the field-moist oil-contaminated soil samples (n=85), thoroughly mixed and placed in a 5-cm diameter plastic Petri dishes. Samples were levelled using a stainless-steel blade. The sample preparation was carried out to enhance the accuracy and reproducibility of the instrument as MIR is affected by sample heterogeneity. For each spectrum, the number of scans (co-added scans) was 32 while the resolution was set to 8 cm^{-1}. Prior to each measurement, a single beam spectrum (background spectrum) was taken with a silver-plated reference cap provided by the manufacturer. The background scan provides a baseline profile of the system conditions with no sample loaded on the instrument, and helps to avoid the negative effects of changes in the environment (e.g. changes in local atmospheric composition) and potential instrument drift (Agilent Technologies, 2015). To eliminate these effects, the measured spectrum was divided internally by the collected background (Hutengs et al., 2018). The spectral data were collected using the Microlab software V5.0 supplied with the spectrometer. The collected raw spectra in reflectance (R) format were firstly converted into absorbance by calculating log (1/R). Then three successive pre-treatment steps were carried out. Smoothing using the Savitzky-Golay (SG) algorithm with polynomial of 2 and windows size of 11 was adopted to remove noise. Smoothing was followed by maximum normalization (Rinnan et al., 2009). Finally, the baseline corrections were implemented using 'modpolyfit' method in chemometrics R- package (R Core Team, 2013), before modelling.

2.5. Modelling

The data matrix including the processed MIR spectra and the SUSE-GC TPH, PAH and alkanes reference values was used to develop PLSR and RF prediction models. Five samples out of the eighty five samples were detected as outliers by principal component analysis (PCA) and removed before the modelling. The remaining 80 samples were divided into two sets: 77% of them for calibration (62 samples) and the remaining 23% for prediction (18 samples) using Kennard-Stone algorithm (Kennard and Stone, 1969). After outliers removal and division of samples into calibration and validation sets, the former set was subjected to both PLSR (Wold, 1982) and RF (Breiman, 2001) analyses to establish calibration models for TPH, alkanes and PAH.

The PLSR analysis was performed using the pls-R package (R Core Team, 2013). A two-dimension matrix composites of full MIR spectra (800 independent variables, X), coupled with the references measured data (3 dependent variables, Y) was subjected to leave-one-out cross-validation (LOOCV). The number of optimal LVs (8) was defined by plotting the resulted RMSEcv against the used LVs of the models, and where the drop-in error value was not significant any more (Wold et al., 2001).

RF was carried out by generating some bootstrap samples or resamples with replacement (*ntree*) from the calibration data. Then, each resample is grown to a regression tree with a modifying process. In this process, numbers from the predictors (*mtry*) tend to be arbitrarily sampled, and the algorithm chooses the best split through these sampled variables rather of all of the variables. The final models were grown to 500 trees (*ntree* = 500), and the minimum number of splitting variables (*mtry*) was set to 2. The final prediction is then calculated as the mean values of the individual predictions of each decision tree. The final TPH, alkanes and PAH models were tested

rigorously using 77% of the data with jack-knife cross-validation, and 23% of the data for the independent validation. In cross-validation, a sample was omitted for testing, and the remaining samples were used for prediction one at a time until all samples in the dataset (77%) were tested. The RF models were performed using the random Forest-R package (Liaw and Wiener, 2015).

2.6. Evaluation of model performance

The prediction performance of TPH, alkanes and PAH models were assessed using three parameters, namely, RMSEP, the ratio of standard deviation (SD) to RMSEP (RPD), and the coefficient of determination in prediction (R^2). Based on the Chang et al. (2001) RPD classification criterion, the performance was classified into four classes: RPD < 1.4 indicates no predictive ability, 1.4 < RPD < 1.8 indicates limited predictive ability, 1.8 < RPD < 2.0 indicates good predictive ability, and RPD > 2.0 indicates accurate predictive ability.

3. Results and discussion

3.1. Laboratory analysis of TPH, alkanes and PAH

Table 1 displays the summary statistics of TPH, alkanes and PAH concentrations acquired using SUSE-GC from the three study sites (Ikarama, S1; Kalabar, S2; and Joinkrama, S3). Among the sites, S3 happened to be the most contaminated. More details of the hydrocarbon concentrations including limit of quantification of the every studied PAH across the sites can be found in Douglas et al. (2018a, and b). However, the hydrocarbon concentration ranges of the samples (Nr = 80) used in this study were 16.07-618.54 mg kg^{-1} for TPH, 9.9-551.22 mg kg^{-1} for alkanes, and 0.52-7.22 mg kg^{-1} for PAH (Table 1).

(Table 1)

3.2. Spectra of soils

MIR absorption spectra of oil contaminated soils from the three sites are compared with an uncontaminated soil spectrum in Fig. 1A and B, for raw and maximum normalised spectra, respectively. This control sample (TPH = < 0.04 mg kg^{-1}) is one of the three uncontaminated samples were collected from the three sites. The comparison shows clear differences between contaminated and non-contaminated spectra, as well as among contaminated spectra themselves. However, the overall shape of the MIR spectra in all the samples presented in Fig. 1A and B were similar, and differences can be attributed to soil physico-chemical properties and level of oil contamination. Absorbance peaks (Fig. 1A) between 1353-1625 cm^{-1} were identified to be associated with aromatic functional groups, while peaks between 2840-3015 cm^{-1} are linked to total recoverable petroleum hydrocarbon (TRH) concentration (aliphatic-CH_2, -CH_3). Absorption peaks around 1353-1625 cm^{-1} observed in the current study are close to those reported by Wartini et al. (2017), which were attributed to aromatic C, C=C conjugated with C=O (1580-1630 cm^{-1}). Also, the 1353-1625 cm^{-1} absorbance peaks are attributable to the vibrations of C-H bending in CH_3, CH_3 out of plane bending, and CH_2 wagging and twisting (Daimay et al., 1991). Similarly, the significant absorption peaks around 2840-3015 cm^{-1} are not far from 2990-2810 cm^{-1} reported by Wartini et al. (2017). Significant absorbance range of 3000-2800 cm^{-1} obtained by a PCA was reported by Webster et al. (2016) to be associated with TPH concentrations. Forrester et al. (2013) identified the wavenumber of 2730 cm^{-1} to be potentially specific to TPH absorption in soils, whereas the same research group (Forrester et al., 2010) found the spectral range of 2700-3000 cm^{-1} to be characteristic features of alkyl-CH_3 stretching vibrations. The aforementioned absorbance signals of hydrocarbons are practically absent in the uncontaminated absorbance curve (UC) in Fig. 1A and B, which is a clear characteristic to differentiate the contaminated samples from the uncontaminated sample.

(Fig. 1.)

3.3. Models performance for predicting TPH, alkanes and PAH

Table 2 shows the modelling results in cross-validation and prediction of TPH, alkanes and PAH using both PLSR and RF prediction methods. Results indicate that RF-MIR models outperformed PLSR in prediction (using prediction set) for the three hydrocarbon components with R^2 = 0.8, RPD = 2.35, RMSEP = 63.80 mg kg^{-1}, R^2 = 0.72, RPD = 1.96, RMSEP = 68.88 mg kg^{-1}, and R^2 = 0.79, RPD = 2.27, RMSEP = 0.83 mg kg^{-1} for TPH, alkanes, and PAH, respectively. The highest prediction accuracy is obtained for TPH, for which RPD values obtained with the RF models were 1.19 and 1.04 times better than alkanes and PAH models, respectively. Lower prediction performance was observed for PLSR compared to RF. The reason behind this is that RF can model the linear and nonlinear response of the MIR spectral data, whereas PLSR is capable to handle only the linear response (Nawar and Mouazen, 2017). This in line with Douglas et al. (2018a) findings for the prediction TPH based on vis-NIR spectroscopy. It has been previously reported that MIR spectra are sensitive to moisture content, which reduces the intensity of the PHCs related peaks leading to low estimation accuracy (Hazel et al., 1997); and the non-linearity effect becomes much stronger with high moisture contents (Webster et al., 2016). Having said that, it can be claimed that results presented in the current work are of strong prediction capability, although the analysis were based on fresh (wet) soil samples with high soil moisture content.

The results achieved in the current study based on RF prediction are better than those reported by Wartini et al. (2017) for cross-validation of TRH in laboratory spiked soil samples using a field portable MIR coupled with PLSR (RMSE and R^2 of 1592 mg kg^{-1} and 0.89, respectively). Also,

the models accuracy in our study is better than those of Webster et al. (2016), who reported RMSE = 1225 mg kg^{-1} for TPH prediction using a handheld MIR. Compared to the prediction results of a portable vis-NIR spectrophotometer for TPH (Douglas et al., 2018a), and alkanes and PAH (Douglas et al., 2018b), where the same samples were studied, results obtained herein with the MIR are more accurate (Table 2). The superior performance of MIR over that of vis-NIR may be attributed to the fact that fundamental molecular vibration occurs in the MIR spectral region, which generates more intense peaks (Reeves, 2010; Soriano-Disla et al., 2014). These findings, therefore, support the use of a portable MIR instrument to predict TPH, alkanes and PAH in fresh oil contaminated soil samples.

(Table 2)

The performance of the PLSR models in the current research is considered poor, compared with previous works by Webster et al. (2016), who reported site specific TPH prediction models with RPD values of 8-13 for three groups of diesel contaminated air-dried and ground soils, field contaminated and laboratory constructed soils. The poor result in the current study may be attributed to very low hydrocarbons concentrations and mixing of soils from three different sites in the same calibrations might have influenced the model prediction accuracy. In another study, Wartini et al. (2017) reported R^2 and RMSEcv of 0.89 and 1592 mg kg^{-1}, respectively, for TRH in processed (air-dried) soils; however, no independent predictions were provided to be able to compare them with results from the present study. It can be challenging to put the results of PAH and alkanes into context with the other studies, since there are no RF-MIR prediction models yet reported in the open literature.

Figures 2, 3, and 4 show the results of the cross-validation and prediction of TPH, alkanes, and PAH using RF and PLSR models. As mentioned earlier the concentration range (for the 80

samples used in the analysis) of TPH, alkanes, and PAH were 16.07-618.54 mg kg^{-1}, 9.9-551.22 mg kg^{-1}, and 0.52-7.22 mg kg^{-1}, respectively. Visually, these scatter plots demonstrate a relatively compact data cloud in all the three plots; indicating a better fit. Among the three studied hydrocarbon components, the RF prediction of TPH was more accurate, as the measured *versus* predicted points are close to the 1:1 line (Fig. 2) compared to more scattered points around the 1:1 lines for alkanes (Fig. 3) and PAH (Fig. 4). The predictions for TPH and PAH with RF models can be classified as accurate (RPD = 2.35 and 2.27, respectively), whereas a limited prediction for the alkanes with RPD of 1.96 was observed (Chang et al., 2001). These results are in line with those reported by other research groups for estimating TPH based on MIR (Webster et al., 2016; Wartini et al., 2017). The limited prediction of alkanes in this study with both RF and PLSR might be attributed to the small range of the concentration, as well as the limited number of samples in the prediction set (18). The dataset size (e.g., sample number) has shown also to have a considerable influence on the prediction performance of TPH (Douglas et al., 2018a) and organic carbon (Nawar and Mouazen, 2017).

It was reported that a small dataset size leads to a negative effect, that is difficult to measure, and may result in very poor performance (Klement et al., 2008). However, the prediction performance here with RF was much better than that obtained with PLSR. Therefore, the current work confirms previous findings and provides additional evidence suggesting that advanced data mining methods (e.g., RF in the current work) have the capability to improve MIR spectroscopy prediction performance for PHCs estimation. Moreover, the use of a handheld MIR spectrometer coupled with RF method has been proved to be a promising tool for field investigation and estimation of the TPH, PAH and alkanes with limited number of soil samples scanned in fresh (wet unprocessed) field sample conditions.

(Fig. 2)

(Fig. 3)

(Fig. 4)

4. Conclusions

This study investigated the potential of a handheld mid infrared (MIR) spectrophotometer for the measurement of total petroleum hydrocarbon (TPH), alkanes, and polycyclic aromatic hydrocarbon (PAH) in fresh (unprocessed) soil samples of relatively small number (n=85), collected from three oil spill sites in the Niger Delta region of Nigeria. Random forest (RF) and partial least squares regression (PLSR) prediction models were developed and the prediction performance was compared. The prediction results showed that RF models outperformed PLSR for the estimation of TPH (coefficient of determination [R^2] = 0.80, ratio of prediction deviation [RPD] = 2.35, and root mean square error of prediction [RMSEP] = 63.80 mg kg^{-1}); alkanes (R^2 = 0.72, RPD = 1.96, RMSEP = 65.88 mg kg^{-1}) and PAH (R^2 = 0.79, RPD = 2.27, RMSEP = 0.83 mg kg^{-1}). Results also showed that MIR spectroscopy performs better than visible and near infrared spectroscopy-based on previously published work using the same samples. This study has demonstrated that the MIR when coupled with RF non-linear calibration method provided accurate prediction of soil TPH, alkanes, and PAH using limited but fresh (unprocessed) soil samples. It is, therefore, concluded that handheld MIR spectrometer coupled with RF modelling can be very useful in quantifying soil hydrocarbon and would provide a rapid and cost-effective means of contaminated site investigation to enhance on-site risk prioritisation; and to support timely pollutant management decision-making and remediation with a potential future field application. Future work will focus on improving the prediction accuracy of the MIR by implementing spiking of the current limited samples into an existing Nigerian contaminated soil

spectral library, followed by modelling using machine learning (e.g., RF) techniques. Further research into developing models for the prediction of hydrocarbons from MIR and vis-NIR signals is necessary so as to select the best performing tool for quantitative analysis of hydrocarbon in soils.

Acknowledgements

We acknowledge the Petroleum Technology Development Fund (PTDF) of Nigeria (PTDF/OSS/PHD/DRK/711/14) for the financial support provided for the PhD project of the first author. We also acknowledge the Flemish Scientific Research (FWO) funded SiTeMan Odysseus I Project (Nr. G0F9216N) for supporting Dr. Nawar stipend.

References

Agilent Technologies, 2015. *Agielnt MicroLab Software Operation Manual*. Fifth edition. Agilent Technologies, Hewlett-Packard-Strasse 8 76337 Waldbronn, Germany. 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.

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. Sci. Total Environ. 514, 399–408.

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.

Chang, C-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R. 2001. Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. Soil Sci. Soc. Am. J. 65:480–490.

Chen, M., Huang, P., Chen, Li., 2013. Polycyclic aromatic hydrocarbons in soils from Urumqi, China: Distribution, source contribution, and potential health risks. Environ. Monit. Assess. 189(7), 5639–5651.

Daimay L-V, Norman, B. C, William, G. F, Jeanette G. G., 1991. In book: The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecules, 9-28.

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. Crit. Rev Environ Sci Technol. 47(16), 1497–1527.

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

Douglas, R.K., Nawar, S., Alamar, M.C., Mouazen, A.M., Coulon, F., 2018b. Rapid detection of alkanes and polycyclic aromatic hydrocarbons in oil-contaminated soils using visible and near-infrared spectroscopy. European Journal of Soil Sci. doi:10.1111/ejss.12567.

Eid, S.M., Abd El-Rahman, M.K., Elghobashy, M.R., and Kelani, K.M., 2018. Attenuated Total Reflectance Fourier Transformation Infrared spectroscopy fingerprinted online monitoring of

the kinetics of circulating Butyrylcholinesterase enzyme during metabolism of bambuterol. Anal. Chim. Acta 1005: 70–80.

Forrester, S.T., Janik, L.J., McLaughlin, M.J., Soriano-Disla, J.M., Stewart, R., and Dearman, B. (2013). Total Petroleum Hydrocarbon Concentration Prediction in Soils Using Diffuse Reflectance Infrared Spectroscopy. Soil Sci. Soc. Am. J., 77(2), 450–460.

Forrester, S., Janik, L., McLaughlin, M., 2010. An infrared spectroscopic test for total petroleum hydrocarbon (TPH) contamination in soils, Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia, August 1–6, 13–16.

Hazel, G., Buchholtz, F., Aggarwal, I.D., Nau, G., Ewing, K.J., 1997. ‘‘Multivariate analysis of mid-IR FT-IR spectra of hydrocarbon-contaminated wet soils’’, Appl. Spectrosc. 51 (7), 984–989.

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

Hutengs, C., Ludwig, B., Jung, A., Eisele, A., and Vohland, M., 2018. Comparison of portable and bench-top spectrometers for mid-infrared diffuse reflectance measurements of soils. Sensors 18, 1–17.

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

Klement, S., Madany Mamlouk, A., Martinetz, T., 2008. Reliability of cross-validation for SVMs in high-dimensional, low sample size scenarios. Artificial Neural Networks-ICANN 2008. Springer Berlin Heidelberg, Berlin, Heldelberg, 41–50.

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 (Accessed :28 April 2016)

Li, J., Lu, H., Fan, X., 2014. Stochastic goal programming based groundwater remediation management under human-health-risk uncertainty. J. Hazard. Mater., 279, 257–267. 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.

Theocharous, E., and Birch, J. R., 2006. Detectors for Mid- and Far-infrared Spectrometry: Selection and Use. In : Handbook of Vibrational Spectroscopy, Chalmers, J.M., Ed., John Wiley & Sons, Ltd: Chichester, UK, p 392.

Okparanma, R. N., Mouazen, A. M., 2013. Determination of Total Petroleum Hydrocarbon (TPH) and Polycyclic Aromatic Hydrocarbon (PAH) in soils. A Review, Appl. Spectrosco. Rev, 46 (6), 458–486.

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.

Pinedo, J., Ibanez, R., Lijzen, J.P.A., Irabien, A., 2013. Assessment of soil pollution based on total petroleum hydrocarbons and individual oil substances. J. Environ. Manage. 130, 72–79.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (URL https://www.R-project.org/).

Reeves, J.B., 2010. Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? Geoderma 158, 3–14

Rinnan, A., Van Den Berg, F., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal. Chem. 28, 1201–1222.

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.

Ritchie, G.D., Still, K.R., Alexander, A.F., Nordholm, C.L., Wilson, J., Rossi III, D., Mattie, R., 2001. A review of the neurotoxicity risk of selected hydrocarbon fuels. J. Toxico. Environ. Health Part B: Crit. Rev. 4, 223–312.

Sammarco, P.W., Kolian, S.R., Wraby, R.A.F., Bouldin, J.L., Sabra, W.A., Porter, S.A., 2013. Distribution and concentrations of petroleum hydrocarbons associated with BP/Deepwater Horigin Oil Spil, Gulf of Mexico. Mar. Pollut. Bull. 73(1), 129–143.

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.

Soriano-Disla, J.M., Janik, L.J., Allen, D.J., McLaughlin, M.J., 2107. Evaluation of the performance of portable visible-infrared instruments for the prediction of soil properties. Biosyst. Eng. 161, 24–36.

Soriano-Disla, J.M., Janik, L.J., Rossel, R.A.V., Macdonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical and biological properties. Appl. Spectrosco. Rev. 49 (2), 139–186.

Viscarra Rossel, R.A., Chappell, A., de Caritat, P., McKenzie, N.J., 2011. On the soil information content of visible-near infrared reflectance spectra. Europ. J. Soil Sci. 62, 442–453.

Wang, Z., and Fingas, M. (1995). Differentiation of the source of spilled oil and monitoring of the oil weathering process using gas chromatography-mass spectrometry. J. Chromatogr A, 712 (2), 321–343.

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

Webster, G.T., Soriano-Disla, J.M., Kirk, J., Janik, L.J., Forrester, S.T., McLaughlin, M.J., Stewart, R.J., 2016. Rapid prediction of total petroleum hydrocarbons in soil using a hand-held mid-infrared field instrument. Talanta 160, 410–416.

Wold, S. Sjostrom, M. Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometer. Intell. Lab. Syst. 58, 109–130.

Wold, H., 1982. ‘‘Soft modeling: the basic design and some extensions’’, in Systems under Indirect Observation, Part 2, Joreskog, K.G and Wold, H (Eds), North Holland, Amsterdam, 1-54.

Wright, J.B., Hasting, D.A., Jones, W.B., Williams, H.K., 1985. Geology and Mineral Resources of West Africa, Allen and Unwin Limited, UK, 107.

Yang, Z.H., Lien, P.J., Huang, W.S., Surampalli, R.Y., Kao, C.M., 2017. Development of the Risk Assessment and Management Strategies for TPH-Contaminated Sites using TPH fraction Methods. J. Hazard, Toxi and Radioa Waste, 21(1), 1–10.

Figure captions:

Fig. 1. (A) Raw mid infrared (MIR) absorbance spectra and (B) maximum normalized MIR absorbance spectra of site 1 (S1) oil-contaminated soil, site 2 (S2) oil-contaminated soil, site 3 (S3) oil-contaminated soil, and uncontaminated (UC) soil spectrum. All samples were collected from the Niger Delta, Nigeria. Absorbance peaks between 1353-1625 cm^{-1} were identified to be associated with aromatic functional groups, while peaks between 2840-3015 cm^{-1} are linked to total recoverable hydrocarbon (TRH) concentrations. These features were not observed in the UC soil spectrum.

Fig. 2 Scatter plots of the measured total petroleum hydrocarbon (TPH) *versus* mid-infrared (MIR) spectroscopy predicted concentrations in cross-validation (a, and c), and in prediction (b, and d) based on (A) partial least squares regression (PLSR), and (B) random forest (RF) modelling methods. The grey areas and the blue lines represent the 95% confidence interval and regression line, respectively.

Fig. 3 Scatter plots of the measured alkanes *versus* mid-infrared (MIR) spectroscopy predicted concentations in cross-validation (a, and c) and in prediction (b, and d) based on (A) partial least squares regression (PLSR), and (B) random forest (RF) modelling methods. The grey areas and the blue lines represent the 95% confidence interval and the regression line, respectively.

Fig. 4 Scatter plots of the measured polycyclic aromatic hydrocarbon (PAH) *versus* mid-infrared (MIR) spectroscopy predicted concentations in cross-validation (a, and c) and in prediction (b, and d) based on (A) partial least squares regression (PLSR), and (B) random forest (RF)

modelling methods. The grey areas and the blue lines represent the 95% confidence interval and the regression line, respectively.

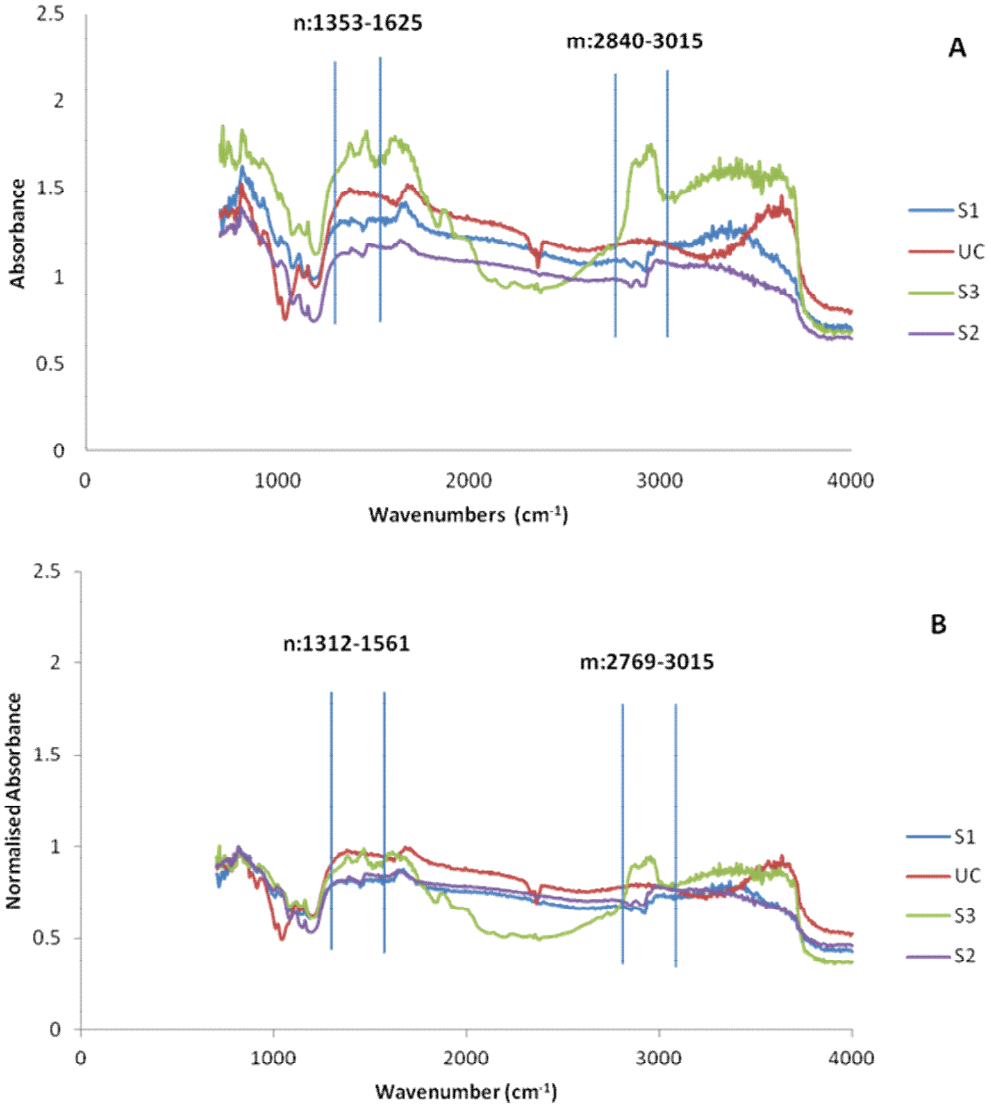

Fig. 1. (A) Raw mid infrared (MIR) absorbance spectra and (B) maximum normalized MIR absorbance spectra of site 1 (S1) oil-contaminated soil, site 2 (S2) oil-contaminated soil, site 3 (S3) oil-contaminated soil, and uncontaminated (UC) soil spectrum. All samples were collected from the Niger Delta, Nigeria. Absorbance peaks between 1353-1625 cm^{-1} were identified to be associated with aromatic functional groups, while peaks between 2840-3015 cm^{-1} are linked to total recoverable hydrocarbon (TRH) concentrations. These features were not observed in the UC soil spectrum.

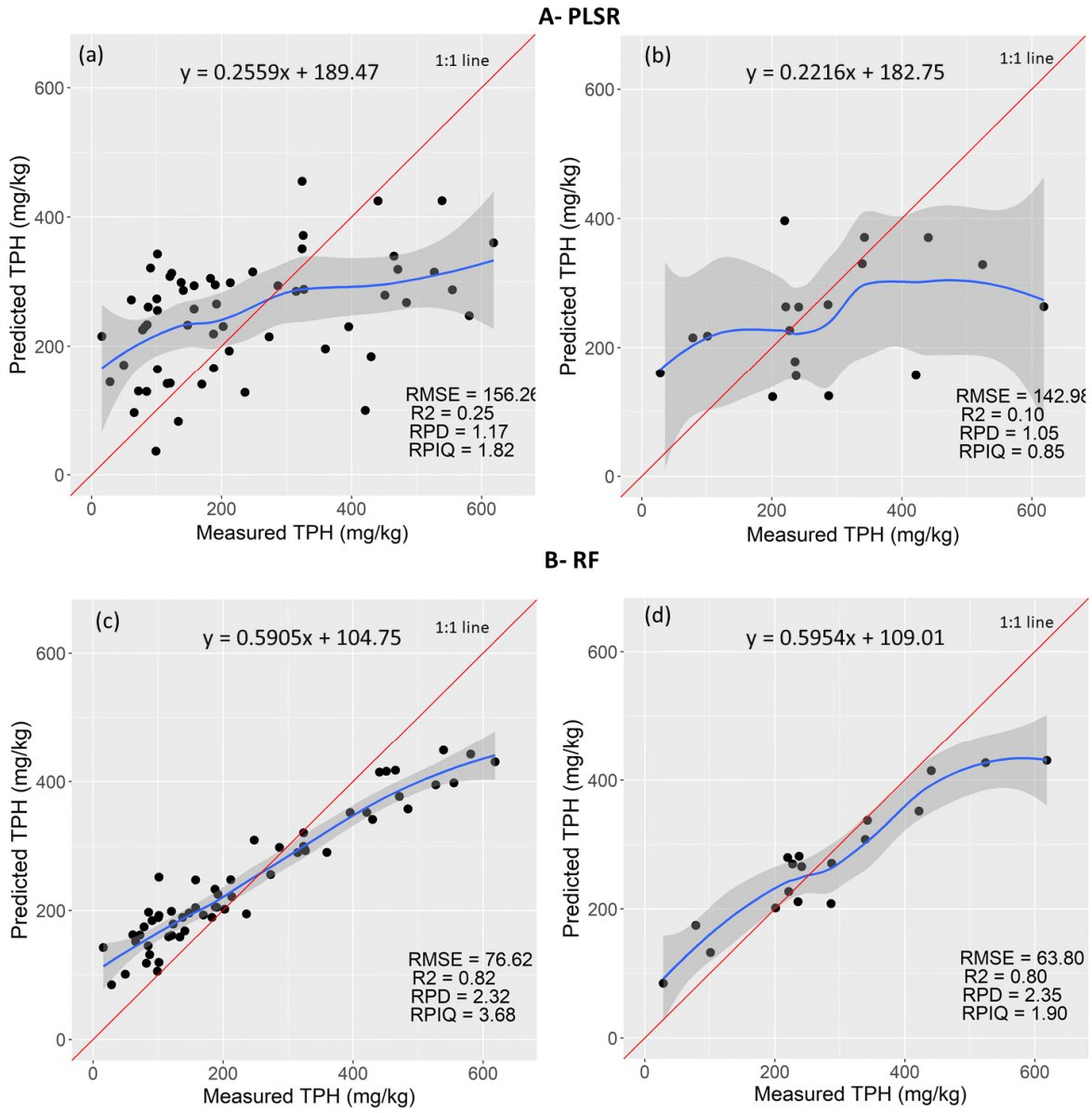

Fig. 2 Scatter plots of the measured total petroleum hydrocarbon (TPH) *versus* mid-infrared (MIR) spectroscopy predicted concentrations in cross-validation (a, and c), and in prediction (b, and d) based on (A) partial least squares regression (PLSR), and (B) random forest (RF) modelling methods. The grey areas and the blue lines represent the 95% confidence interval and regression line, respectively. The plot was based on without outliers of the TPH data.

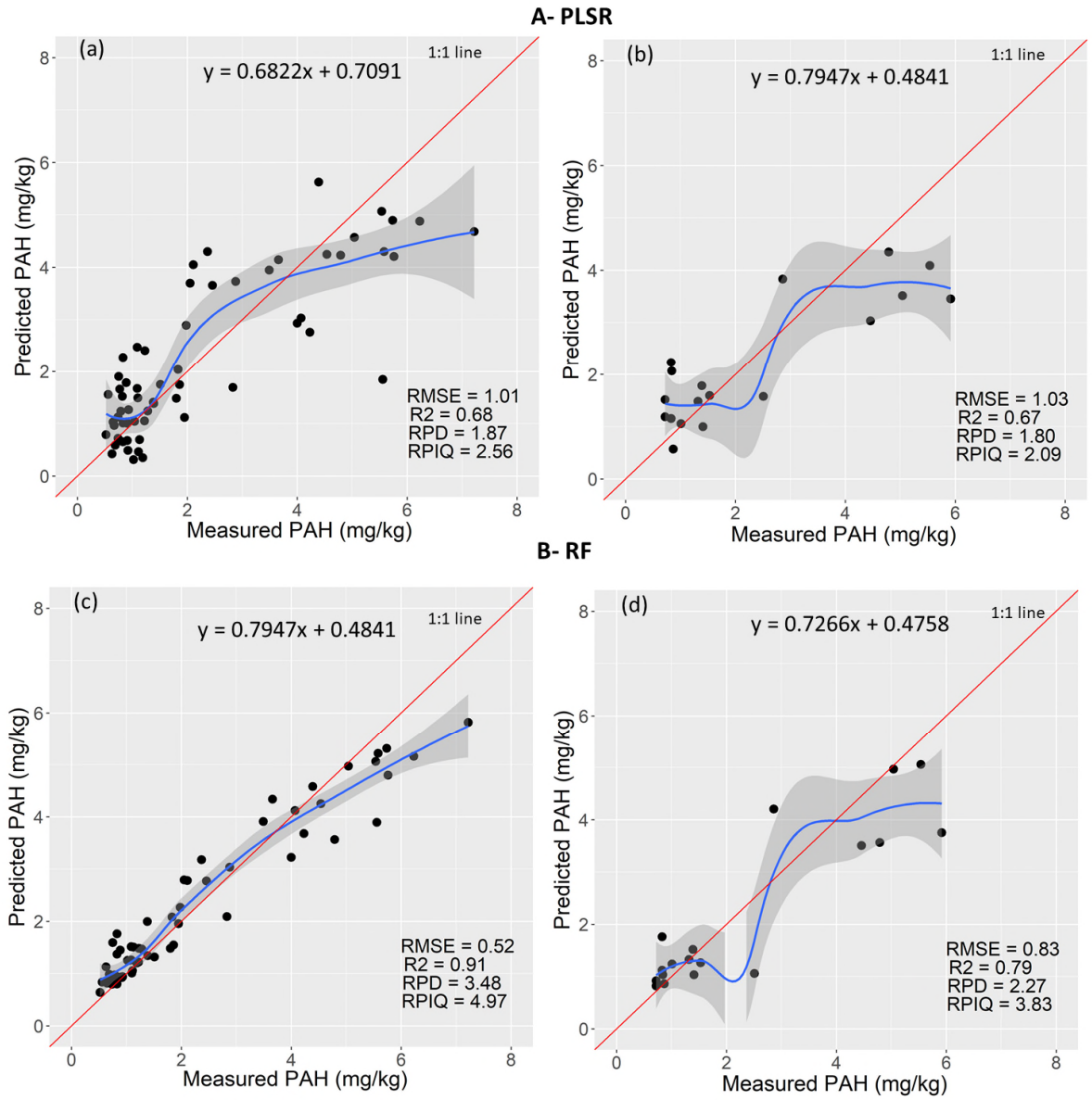

Fig. 3 Scatter plots of the measured alkanes *versus* mid-infrared (MIR) spectroscopy predicted concentations in cross-validation (a, and c) and in prediction (b, and d) based on (A) partial least squares regression (PLSR), and (B) random forest (RF) modelling methods. The grey areas and the blue lines represent the 95% confidence interval and the regression line, respectively. This plot was based on without outliers of PAH data.

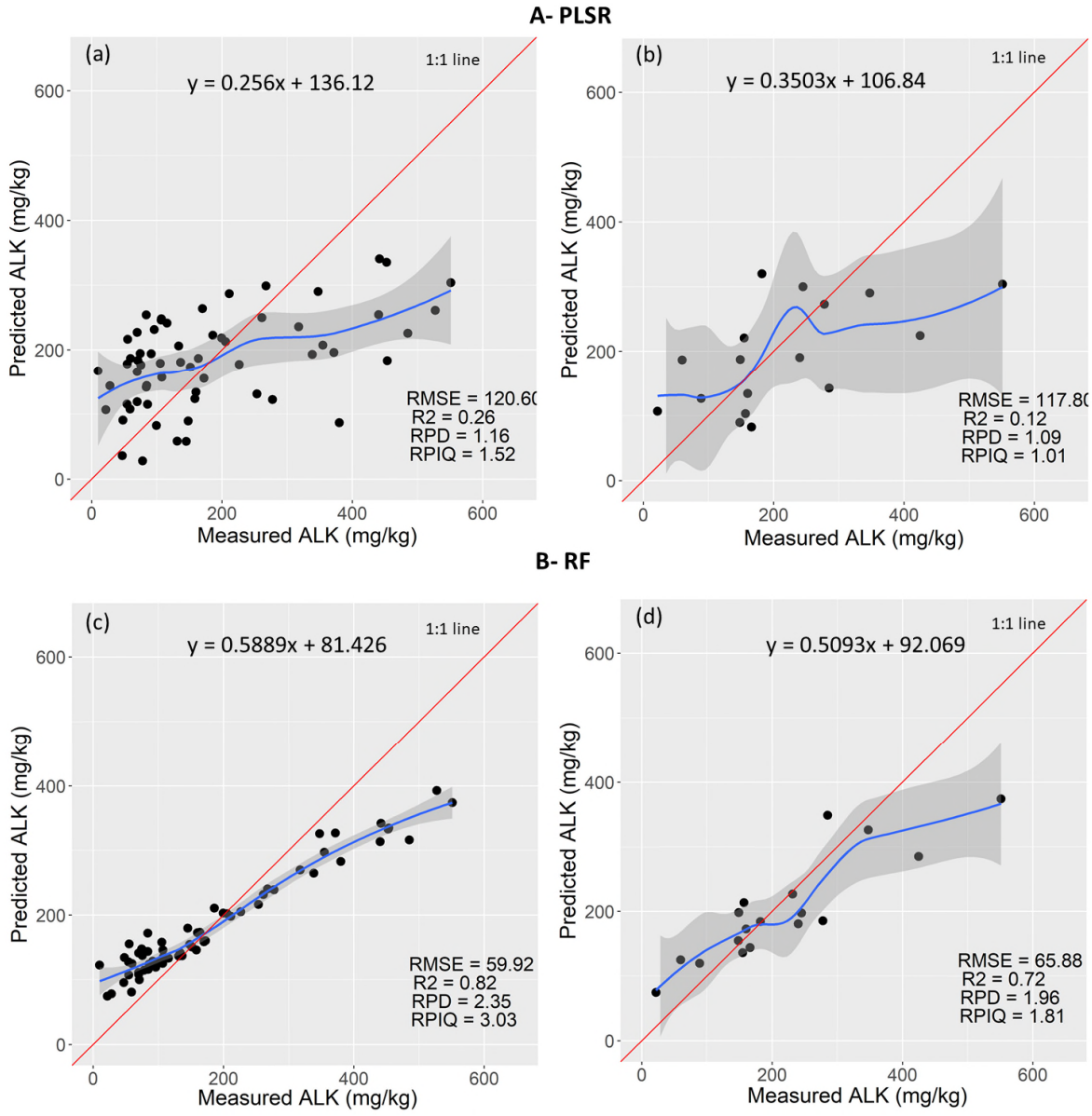

Fig. 4 Scatter plots of the measured polycyclic aromatic hydrocarbon (PAH) *versus* mid-infrared (MIR) spectroscopy predicted concentations in cross-validation (a, and c) and in prediction (b, and d) based on (A) partial least squares regression (PLSR), and (B) random forest (RF) modelling methods. The grey areas and the blue lines represent the 95% confidence interval and the regression line, respectively. This plot was based on without outliers of the alkanes data.

Table 1
The descriptive analysis of the tested soil samples for the three measured contaminate compounds total petroleum hydrocarbons (TPH), alkanes and polycyclic aromatic hydrocarbons (PAH) measured using sequential ultrasonic solvent extraction gas chromatography (SUSE-GC).

This hydrocarbon data coupled with vis-NIR spectral signal were previously used by Douglas et al. (2018a, b) for the prediction of TPH, alkanes, and PAH.

	N	Min.	Mean	Median	1st Qu.	3rd Qu.	Max.	St. dev.
TPH (mg $kg^{-1)}$	85	16.07	252.59	213.69	120.66	339.27	666.33	165.51
Alkanes (mg kg^{-1})	85	9.90	187.24	151.75	84.55	259.25	551.22	133.13
PAH (mg kg^{-1})	85	0.52	9.11	1.39	0.89	4.00	312.28	40.20

N= number of samples; Min. = Minimum; 1st Qu. = first quartile; 3rd Qu. = third quartile; St. dev. = standard deviation.

Table 2

The results of calibration (cross-validation) and prediction for total petroleum hydrocarbon (TPH), alkanes (ALK), and polycyclic aromatic hydrocarbon (PAH) models based on of random forest (RF) and partial least square regression (PLSR) methods in naturally oil-contaminated soil samples collected from three sites in the Niger Delta, Nigeria. Results compare the RF and PLSR mid infrared (MIR) prediction performance of the present study with those obtained from visible near infrared (vis-NIR) spectroscopy analyses reported previously by Douglas et al. (2018a) and Douglas et al. (2018b).

Instrument		**PLSR**					**RF**					Property
	Present study	R^2	RMSEP (mg kg^{-1})	RPD	RPIQ	LV	R^2	RMSEP (mg kg^{-1})	RPD	RPIQ	ntrees	
MIR	Calibration (n=62)	0.25	156.26	1.17	1.82	8	0.82	76.62	2.32	3.68	500	TPH
	Prediction (n=18)	0.10	142.98	1.05	0.85	8	0.80	63.8	2.35	1.90	500	
	Calibration (n=62)	0.26	120.6	1.16	1.52	8	0.82	59.92	2.35	3.03	500	ALK
	Prediction (n=18)	0.12	117.8	1.09	1.01	8	0.72	65.88	1.96	1.81	500	
	Calibration (n=62)	0.68	1.01	1.87	2.56	8	0.91	0.52	3.48	4.97	500	PAH
	Prediction (n=18)	0.67	1.03	1.80	2.09	8	0.79	0.83	2.27	3.83	500	
Vis-NIR	***Previous study*** [a]											
	Calibration (n=65)	0.63	107.54	1.66	2.55	8	0.85	68.43	2.61	3.96	500	TPH
	Prediction (n=20)	0.54	75.86	1.51	2.10	8	0.68	69.64	1.85	2.53	500	
Vis-NIR	***Previous study*** [b]											
	Calibration (n=65)	0.49	101.71	1.41		6	0.85	55.71	2.58		500	ALK
	Prediction (n=18)	0.36	66.66	1.29		6	0.58	53.95	1.59		500	
	Calibration (n=58)	0.76	0.81	2.07		6	0.89	1.02	2.99		500	PAH
	Prediction (n=23)	0.56	1.21	1.55		6	0.71	0.99	1.99		500	

Previous study[a] =Douglas et al., 2018a; Previous study[b] =Douglas et al., 2018b; R^2 = coefficient of determination; RMSEP = root mean square error of prediction; RPD = residual prediction deviation; LV = latent variables; ntrees = number of trees; and RPIQ = ratio of performance to interquartile range.