CRANFIELD UNIVERSITY

REWARD KOKAH DOUGLAS, B. Eng. (Hons), MSc.

IMPLEMENTATION OF SPECTROSCOPY AS A RAPID MEASUREMENT TOOL (RMT) TO INFORM RISK ASSESSMENT AT PETROLEUM CONTAMINATED SITES IN THE NIGER DELTA, NIGERIA.

SCHOOL OF WATER, ENERGY AND ENVIRONMENT

PHD THESIS
Academic Year: 2017 - 2018

Supervisor: Dr M. Carmen Alamar, Prof Frederic Coulon,
Prof Abdul M. Mouazen
April 2018

CRANFIELD UNIVERSITY

SCHOOL OF WATER, ENERGY AND ENVIRONMENT

PHD THESIS

Academic Year 2017 - 2018

REWARD KOKAH DOUGLAS, B. Eng. (Hons), MSc.

IMPLEMENTATION OF SPECTROSCOPY AS A RAPID MEASUREMENT TOOL (RMT) TO INFORM RISK ASSESSMENT AT PETROLEUM CONTAMINATED SITES IN THE NIGER DELTA, NIGERIA.

Supervisor: Supervisor: Dr M. Carmen Alamar, Prof Frederic Coulon, Prof Abdul M. Mouazen

April 2018

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

ABSTRACT

The recent developments and applications of rapid measurement tools (RMT) such as visible near-infrared (vis-NR) spectroscopy can provide ‘fit for purpose’ and cost effective data for informing risk assessment and managing oil-contaminated sites. Infrared spectroscopy discriminates between chemical compounds by detecting the specific vibrational frequencies of molecular bonds, producing a unique infrared ‘spectral signal’ thereby enhancing its identification and quantification applying chemometrics. The aim of the research was therefore to assess the potential of vis-NIR and mid-infrared (MIR) diffuse reflectance spectroscopy (DRS) techniques as RMT to inform risk decision support for remediation of petroleum contaminated sites. The objectives of the study were to: critically review chromatographic and spectroscopic methods for petroleum hydrocarbon analysis in soils; evaluate vis-NIR sensitivity to detect hydrocarbon concentration differences throughout weathering; predict TPH, PAH and alkanes concentrations in contaminated soils using vis-NIR and MIR DRS coupled with regression techniques. The study further evaluated which spectroscopy technique (vis-NIR or MIR); and which modelling method (RF or PLSR) performs best. In this study, a series of 13 soil mesocosms was set up where each soil sample collected was spiked with 10 ml of Alaskan crude oil and allowed to equilibrate at room temperature for 48 h before analysis. The mesocosms were incubated for two years at room temperature in the dark. Soils scanning and gas chromatography coupled to mass spectrometry (GC-MS) analysis were carried out at T0, 4, 12, 16, 20 and 24 months. Prior to scanning, soil samples were air-dried at room temperature (21^{o}C) to reduce the effect of moisture. The soil scanning was done simultaneously using an AgroSpec spectrometer with a spectral range of 305 to 2200 nm (tec5 Technology for Spectroscopy, Germany) and Analytical Spectral Devices LabSpec2500 spectrometer (ASD Inc, USA) with a spectral range of 305 to 2500 nm to assess and compare the sensitivity and response of the two instruments to weathering and hydrocarbon composition change overtime against GC-MS data. Partial least squares (PLS) and random forest (RF) regression models showed that ASD LabSpec2500 performed better than tec5 which may be attributed to the shorter wavelength spectra range of the tec5 spectrometer and therefore not detecting all significant hydrocarbon signals (e.g., 2207, 2220, 2240 and 2460 nm). The sensitivity of the two spectral devices to weathering and hydrocarbon composition change was, however, comparable; and the predicted concentrations were also comparable to the hydrocarbons concentrations determined by GC-MS. The results (coefficient of determination, R^2=0.9; ratio of prediction deviation, RPD=3.79 and root mean square error of prediction, RMSEP=108.56 mg/kg) demonstrate that visible-near infrared diffuse reflectance spectroscopy (vis-NIR DRS) is a proven tool for rapid site investigation and monitoring without the need of collecting soil samples and lengthy hydrocarbon extraction for further analysis..To this end, 85 soil samples collected from three crude oil spill sites in the Niger Delta, Nigeria. Prior to spectral measurement, soil physiochemical properties such as pH, total organic carbon and textural analysis were carried out. Soil samples were scanned (field-moist) and assessed using ASD LabSpec2500 (wavelength 350-2500 nm) and MIR data was acquired with Agilent 4300 handheld Fourier transform infrared (FTIR) spectrometer (Agilent Technologies, Santa Clara, CA, United States) with a spectral range of 4000-650 cm^{-1}. Specifically, detailed analysis of the hydrocarbon content including total petroleum hydrocarbons (TPH), aliphatic and aromatic hydrocarbon fractions were determined and quantified by GC-MS, vis-NIR and MIR DRS. MIR over-performed vis-NIR with RF modelling method performing better than PLSR in predicting TPH, PAH and alkanes. However, PLSR-vis-NIR produced slightly better results than PLSR-MIR in predicting TPH and alkanes.

Overall, vis-NIR (wavelength 350-2500 nm) laboratory-scale study yields better TPH prediction than the field-scale study. The minimised moisture content may have improved the results, as laboratory-scale samples were air-dried. Based on the results, MIR spectroscopy coupled with RF is recommended for the analysis of hydrocarbon contaminated soil. Finally, spectroscopy approach was proposed as RMT for contaminated soil investigation and risk prioritisation.

Keywords: Niger Delta, soil contamination, spectroscopy, chemometrics, site investigation, risk-decision

ACKNOWLEDGEMENTS

I would want to express my sincere appreciation to my supervisors: Dr M. Carmen Alarmar, Prof. Frederic Coulon, and Prof. Abdul M Mouazen for their close guidance; and wealth of experience in spectroscopy, environmental chemistry and land contamination that led to the successful completion of the PhD research. The Chair, Dr. Giuseppina Di Lorenzo and the Reviewer, Dr. Toby Waine of the thesis committee are acknowledged. I would also want to thank the Petroleum Technology Development Fund (PTDF) of Nigeria who sponsored this research programme through the Foreign Scholarship Scheme.

I am also grateful to the GC-MS technical team, particularly Dawn Fowler and the technical support team of the Environmental Analytical Facility for their guidance during the practical session in the laboratories. My thanks are due Dr Said Nawar for his guidance and wealth of experience in modelling that enrich the thesis. My colleague, Sabrina Cipullo is acknowledged. I also appreciate Mr. Keith Hurley and Mrs. Alison Waters within the IT Department of Cranfield University for their support with Excel and data analysis and thesis formatting irrespective of their tightened schedules. K & A GEOSOLUTION NIG. LTD staff are appreciated for their support during my field trips in Nigeria.

I specially thank Chief Bestman A. Festus, Dr. Raphael T Samuel, Dr. Reuben Okparanma and Dr. Kenneth S Okiongbo for their advice and encouragement throughout this programme. My friends, Engr. Perez Araka, Engr. Ayebatin Fou, Ichebadu V Orlu, Alfred W Opukumo, Ayibawari Egai, Alaye T Theophilus, Mrs Felicia Adike amongst others are greatly appreciated. Barr. Jude Amiditor Rex-Ogbuku, you are highly acknowledged.

My heartfelt appreciation goes to my wife, Mrs Eseimokumo Douglas and children for their unflinching love, support and understanding throughout the PhD. All my family members, I appreciate you all for your support and words of encouragement.

I am deeply grateful to my Christian families: Deeper Life Bible Church (Buchinghamshire region, England) and Christ Church Aguobiri (Bayelsa Nigeria).

LORD, I give you all the glory!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ANN	Artificial neural network
ASTM	American Society of Testing and Materials
BRT	Boosted regression tree
BTEX	Benzene, toluene, ethylbenzene, xylene
CWA	Chemical warfare agent
DCM	Dichloromethane
DRS	Diffuse reflectance spectroscopy
ED-XRF	Energy dispersive X-ray fluorescence
EPA	Environment protection agency
EU	European Union
FGC	Field gas chromatography
FID	Flame ionisation detector
GC	Gas chromatography
GC × GC	Two dimensional gas chromatography
GDP	Gross domestic product
H	High
IRS	Infrared spectroscopy
L	Low
LC	Liquid chromatography
LMW	Low molecular weight
LOOCV	Leave-one-out cross-validation
LOD	Limit of detection
M	Medium
MARS	Multivariate adaptive regression spline
MC	Moisture content
MLR	Multiple linear regression
MPLC	Medium pressure liquid chromatography
MIR	Mid-infrared
MIRS	Mid-infrared spectroscopy
MMW	Medium molecular weight
MS	Mass spectrometry
MTBE	Methyl tertiary butyl ether
OCLC	Open-column liquid chromatography
OLR	Ordinal logistic regression

OPEC	Organisation of Petroleum Exporting Countries
PLSR	Partial least squares regression
PAH	Polycyclic aromatic hydrocarbon
PGC	Portable gas chromatography
PHC	Petroleum hydrocarbon
PHLC	High performance liquid chromatography
PSR	Penalised spline regression
QTOF	Quadrupole time-of-flight
R^2	Coefficient of determination
RENA	Remediation by Natural Attenuation
RFR	Random forest regression
RI	Retention index
RMSEP	Root mean square error of prediction
RPD	Residual prediction deviation
RS	Raman spectroscopy
SimDis	Simulated distillation
SFC	Supercritical fluid chromatography
SOM	Soil organic carbon
SUSE-GC	Sequential ultrasonic solvent extraction gas chromatography
SVM	Support vector machine
TAME	Tert-amyl methyl ether
TC	Total carbon
TGC	Transportable gas chromatography
TPH	Total petroleum hydrocarbon
TREPH	Total recoverable petroleum hydrocarbon
TIC	Toxic industrial chemicals
TN	Total nitrogen
TOF	Total time-of-flight
TTEC	Total toxicity equivalent concentration
TLC	Thin layer chromatography
UNEP	Union Nations Environment Programme
USDA	United State Department of Agriculture
Vis-NIR	Visible and near infrared spectroscopy
VOC	Volatile organic compound
WD-XRF	Wavelength dispersive X-ray fluorescence
XRFS	X-ray fluorescence spectroscopy

CHAPTER 1 : Overview of the project

1.1 Introduction

Crude oil exploration in Nigeria started in the Niger Delta in 1956 at Oloibiri, Bayelsa State, by Shell D’Arcy; and commercial production started two years later (Kadafa, 2012; Amu, 1997). Bayelsa State is located within the southern end of Nigeria and is within the Niger Delta Basin. It lies between longitude 006^{o}05′ and 006^{o}04′ East of the prime meridian and latitudes 04^{o}23.3′ and 04^{o}38.2′ North of the equator within the coastal area of the Niger Delta. Nigeria’s economy is greatly reliant on earnings from the oil and gas sector, which accounts for *ca*. 35 percent of Gross Domestic Product (GDP); while petroleum exports revenue represents over 90 percent of the total exports revenue (OPEC, 2015; Akpabio and Akpan, 2010). These records show that the oil and gas industry in the Niger Delta region has contributed enormously to the growth and development of Nigeria. However, since the beginning of the establishment of oil and gas industry in the region, several oil spill incidents have been reported; and, to date, it has been estimated that 13 million tons of hydrocarbons have been spilled in the region due to pipeline fatigue, well blowout (a case in which control of well is lost during drilling operations), pipeline vandalism, and sabotage (Ambitunni et al., 2014; Nwilo and Badejo, 2006). Similarly, Ite et al. (2013) reported that the number of contaminated sites in the Niger Delta region is in excess of 2000. Furthermore, the United Nation Environment Programme (UNEP) reported in 2011 that in Ogoniland alone (a small part of the Niger Delta), over 69 sites were heavily contaminated with crude oil (concentration > 139,000 mg/kg affecting soil, air and water quality criteria and posing a serious human health threat; Figure 1-1). However, the Nigeria legislation dealing with soil contamination from oil operations handled by the Environmental Guidelines and Standards for Petroleum Industries in Nigeria (EGASPIN, 1992) set out minimum requirements in terms of hydrocarbon contaminations in soil (target value=50 mg/kg and intervention value=5000 mg/kg).

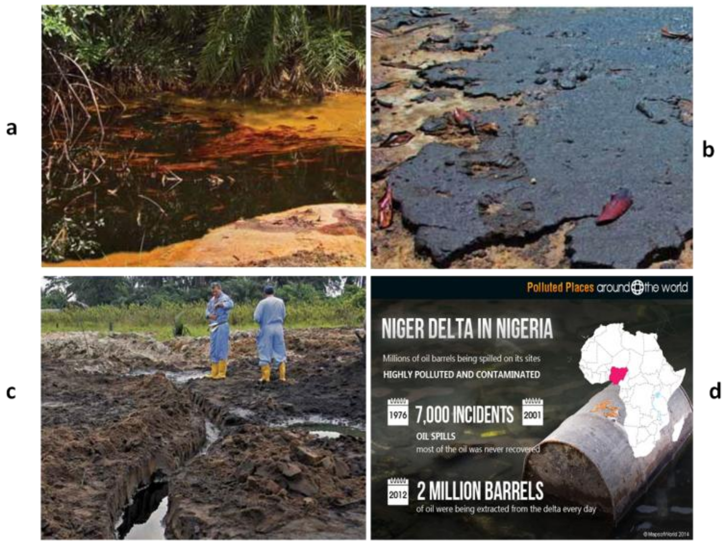

a
b
c
d
Polluted Places around the world
NIGER DELTA IN NIGERIA
Millions of oil barrels being spilled on its sites
HIGHLY POLLUTED AND CONTAMINATED
1976
7,000 INCIDENTS
2001
OIL SPILLS
most of the oil was never recovered
2012
2 MILLION BARRELS
of oil were being extracted from the delta every day
©MapsofWorld 2014

Figure 1-1**: Oil spill incidents in the Niger Delta region of Nigeria:(a) oil spill site within Ogoniland; (b) soil caked into a crust of dried crude oil; (c) trench made from remediation by enhanced natural attenuation (RENA) site to a nearby watercourse; and (d) oil spill incidents between 1976 and 2001, and the number of barrels extracted per day from the Niger Delta, Nigeria. Source: a-c (UNEP, 2011); d: MapsoftWorld, 2014.**

Although the full cost of oil-contaminated sites clean-up in Nigeria is not yet known, UNEP (2010) estimated at least $1bn for the first ten years of the thirty year clean-up programme for oil-contaminated sites in Ogoniland (1, 000 km^2) in the Niger Delta (112, 110 km^2). While urgent attention to clean-up these sites is needed, Nigeria lacks the necessary funds, like most nations, to address all contaminated sites accordingly (Sam et al., 2016). In order to tackle this problem, decision-makers must prioritise their clean-up activities to maximise the benefit derived from limited funds. To facilitate this, there is a need for rapid measurement tools (RMT) that can be easily used by staff, even with little training, to help with the identification and risk prioritisation of contaminated sites and inform swiftly risk decision making. To date, only Okparanma et al. (2014a) have used visible and near-infrared (vis-NIR) diffuse reflectance spectroscopy (DRS) as a rapid and cost-effective technique for offset analysis of petroleum-contaminated soils collected from oil spill sites in the Niger Delta region of Nigeria. The authors only predicted PAH in genuine oil-contaminated soils. Although the authors suggested that the methodology may be useful for rapid assessment of the spatial variability of petroleum hydrocarbons in petroleum release soils to inform risk assessment and remediation, their study did not involve alkanes prediction and could not cover the large number of petroleum hydrocarbon contaminated land sites in the region. Thus, there is need for alkanes prediction because they are toxic hence be considered for risk assessment in fresh spill sites. Similarly, mid-infrared (MIR) DRS has also been previously demonstrated as a rapid and cost-effective technique for the analysis of oil-contaminated soil (Wartini et al., 2017; Webster et al., 2016; Horta et al., 2015). However, this technique has not been applied to analyse contaminated soils from the Niger Delta, Nigeria to date. Because of the advantages of vis-NIR and MIR techniques such as portability, ease to use, and rapid assessment of hydrocarbon contaminations in soil, over the slow and expensive analytical chemistry methods [e.g., gas chromatography-mass spectrometry (GC-MS)], DRS spectroscopy is a candidate rapid measurement tool (RMT) for contaminated sites with hydrocarbons.

Infrared spectroscopy is a non-destructive technique that has been and could be further used for the analysis of petroleum-contaminated sites. This involves a first step where reflectance spectra from the soil sample are acquired; the second step consists of modelling this spectral data against same samples with reference hydrocarbon (e.g., total petroleum hydrocarbon [TPH], polycyclic aromatic hydrocarbon [PAH]) concentrations (Wartini et al., 2017; Chakraborty et al., 2015; Okparanma et al., 2014a, 2014b; Okparanma and Mouazen, 2013, 2012; Schewartz et al., 2012; Chakraborty et al., 2010; Forrester et al., 2010; Bray et al., 2009; Malley et al. 1999). Similarly, the potential applications of MIR for the detection of total petroleum hydrocarbon in soils have been reported (Wartini et al., 2017; Horta et al., 2015; Forrester et al., 2013, 2010). Currently, only a United States Patent (Janik et al., 2015) reported on the successful use of MIR-PLS for the determination of PAH concentration in soil. Thus, more work on the use of MIR for PAH prediction in soil is needed to support the findings of Janik et al. (2015).

However, there are some aspects that have not been addressed by the above studies: (a) the evaluation of vis-NIR spectroscopy sensitivity to changes in TPH due to weathering; (b) the prediction of alkanes in petroleum-contaminated soils by vis-NIR; and (c) the detection of alkanes in petroleum-contaminated soils by MIR spectroscopy. These gaps of research necessitate the current project aim and objectives detailed below. This chapter presents a schematic on how the formulated objectives guided to achieve the research aim (Figure 1-2).

1.2 Research aim and objectives

1.2.1 Aim of study

The aim of this research is to assess the potential of the vis-NIR and MIR diffuse reflectance spectroscopy (DRS) techniques as rapid measurement tool (RMT) to inform decision support for remediation of petroleum contaminated sites.

1.2.2 Objectives of study

In order to achieve the aim of the study the following objectives were formulated:

1. To critically review chromatographic and spectroscopic methods for petroleum hydrocarbon analysis in soils.

2. To evaluate vis-NIR sensitivity to detect hydrocarbon concentration differences throughout weathering.

3. To predict TPH, PAH and alkanes concentrations in contaminated soils using vis-NIR DRS coupled with regression techniques.

4. To predict TPH, PAH and alkanes concentrations in contaminated soils using MIR DRS coupled with regression techniques.

Within the study objectives the following sub-objectives were evaluated: evaluate which spectroscopy technique (vis-NIR or MIR); and which modelling method (RF or PLSR) performs best.

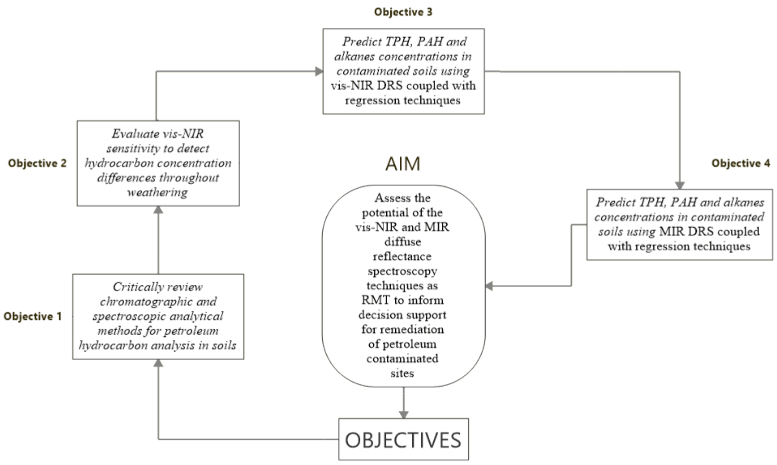

Objective 3
Predict TPH, PAH and alkanes concentrations in contaminated soils using vis-NIR DRS coupled with regression techniques
Objective 2
Evaluate vis-NIR sensitivity to detect hydrocarbon concentration differences throughout weathering
AIM
Objective 4
Assess the potential of the vis-NIR and MIR diffuse reflectance spectroscopy techniques as RMT to inform decision support for remediation of petroleum contaminated sites
Predict TPH, PAH and alkanes concentrations in contaminated soils using MIR DRS coupled with regression techniques
Objective 1
Critically review chromatographic and spectroscopic analytical methods for petroleum hydrocarbon analysis in soils
OBJECTIVES

Figure 1-2: Schematic showing how the objectives were used to achieve the aim. Where TPH: total petroleum hydrocarbons; PAH: polycyclic aromatic hydrocarbon, MIR: mid-infrared spectroscopy, vis-NIR: visible near-infrared spectroscopy, DRS: diffuse reflectance spectroscopy.

1.3 Rationale and significance of the study

The challenges of oil spillage, such as soil, surface and groundwater contamination (Davies and Abolude, 2016; Dyck et al., 2013; Hua et al., 2012; UNEP, 2011), and its negative impacts on human beings and other species. These impacts cannot be adequately addressed without the application of rapid, portable, and cost-effective techniques; as traditional laboratory methods can be slow, labour-intensive, expensive, use toxic extraction solvents, and require both skilled operators and prior soil preparation. This research hopes to position reflectance spectroscopy as a RMT for cost-effective *in situ* technique to facilitate *on-site* risk assessment pertaining soil PHC contamination.

1.4 Thesis structure

The thesis is organised into six chapters, demonstrating how vis-NIR spectroscopy was used for field- and laboratory-scale analysis of oil-contaminated soils (Figure 1-3).

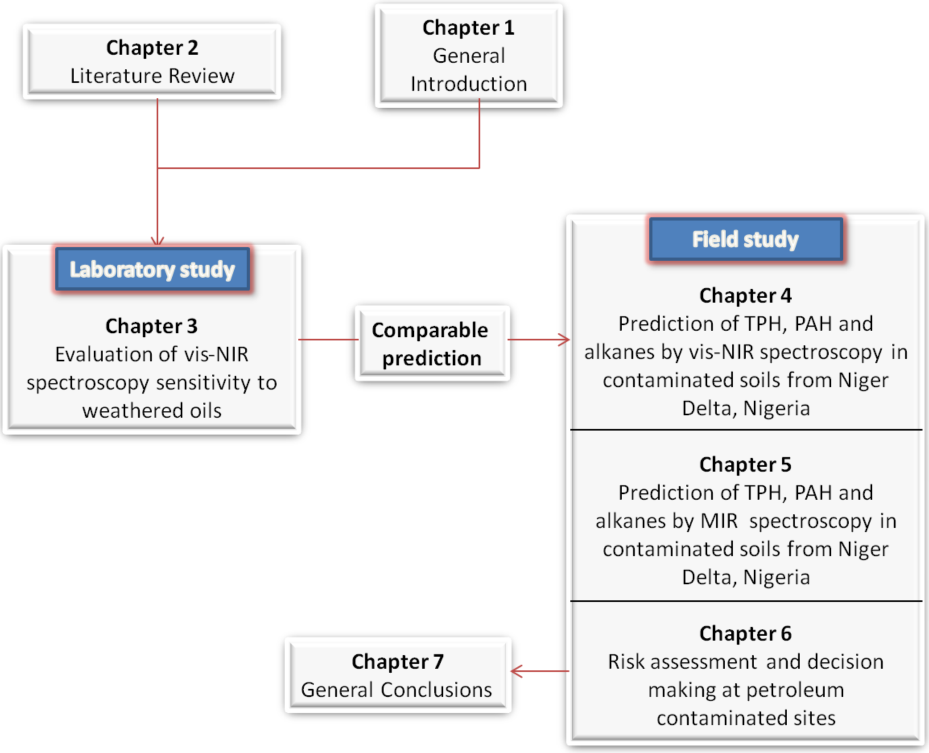

Chapter 2
Literature Review
Chapter 1
General Introduction
Laboratory study
Chapter 3
Evaluation of vis-NIR spectroscopy sensitivity to weathered oils
Comparable prediction
Field study
Chapter 4
Prediction of TPH, PAH and alkanes by vis-NIR spectroscopy in contaminated soils from Niger Delta, Nigeria
Chapter 5
Prediction of TPH, PAH and alkanes by MIR spectroscopy in contaminated soils from Niger Delta, Nigeria
Chapter 6
Risk assessment and decision making at petroleum contaminated sites
Chapter 7
General Conclusions

Figure 1-3: Representation of the thesis structure. Where TPH: total petroleum hydrocarbons; PAH: polycyclic aromatic hydrocarbon, MIR: mid-infrared spectroscopy, vis-NIR: visible near-infrared spectroscopy.

The contents of each chapter are summarised as follows:

Chapter 1: This chapter presents a general introduction, which also contains the research gaps, aim and objectives, and brief description of the thesis structure.

Chapter 2: This chapter provides a review of chromatographic and spectroscopic techniques for the measurement of petroleum hydrocarbons in soil and sediment samples. A comprehensive analytical framework based on spectroscopic techniques integration and data fusion for the rapid measurement of PHC is presented. This chapter has been published in Critical Reviews in Environmental Science and Technology (Douglas et al., 2017).

Chapter 3: In this chapter, the potential of vis-NIR to quantify and discriminate different weathering groups of soils was assessed at lab-scale. Principal component analysis (PCA) on soil spectra followed by partial least squares regression (PLSR) and random forest (RF) analyses using the spectra and gas chromatography profiles of the samples with leave-one-out cross-validation (LOOCV) were carried out. Results show that vis-NIR reflectance spectroscopy is sensitive to changes in hydrocarbon due to weathering; however, the sensitivity decreases over time. Chapter 3 has been published in Science of the Total Environment (Douglas et al., 2018b).

Chapter 4: In this chapter, PLSR and RF prediction performance was compared when developing vis-NIR calibration models for the estimation of TPH, PAH and alkanes in fresh (wet non-processed) soils. Results showed that RF modelling technique outperformed PLSR with excellent and good prediction accuracies, respectively. It was concluded that vis-NIR spectroscopy coupled with RF modelling approach can be a promising method for rapid, cost-effective and *in situ* quantification of TPH, PAH and alkanes in soils. Chapter 4 has been published in Science of the Total Environment (Douglas et al., 2018a). In Chapter 4, the potential of vis-NIR spectroscopy for the prediction of alkanes and PAH concentrations in contaminated soils has also been investigated. Again, RF outperformed PLSR for the prediction of both soil properties. The result of this work was published in European Journal of Soil Sciences (Douglas et al., 2018c).

Chapter 5: This chapter implements the use of MIR spectroscopy for the prediction of TPH, PAH and alkanes in fresh (wet non-processed) soil samples, with the aim to compare it with vis-NIR spectroscopy results presented in Chapter 4. Furthermore, PLSR and RF modelling methods were compared. The results showed a much better prediction performance of RF-MIR models when predicting TPH, PAH and alkanes than RF-vis-NIR models. Thus, the use of MIR spectroscopy coupled with RF is recommended as the best optical method for the measurement of PHCs in soils. This chapter was written as a research paper, which has been submitted to Journal of Environmental Management.

Chapter 6: This chapter presents the general conclusions drawn from all chapters, and highlights the research implications. This chapter also demonstrates how reflectance spectroscopy can be implemented for rapid and cost-effective investigation and risk prioritization of oil-contaminated sites. Furthermore, this chapter concludes with providing recommendations for future work.

1.5 Publications

In the course of writing the thesis, four articles were published in peer-review journals, and one was submitted.

1. Almost 25 years of chromatographic and spectroscopic analytical method development for petroleum hydrocarbons analysis in soil and sediment: start-of-the-art, progress and trends. *Crit. Rev Environ Sci Technol.*, 47(16), 1497–1527.
2. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. *Sci. Total Environ*. 616 (2018) 147-155.
3. Evaluation of vis-NIR reflectance spectroscopy sensitivity to weathering for enhanced assessment of oil contaminated soils. *Sci. Total Environ*. 626, 1108-1120.
4. Rapid detection of alkanes and polycyclic aromatic hydrocarbon in oil-contaminated soils using visible near-infrared spectroscopy. *European. J. Soi. Sci*: doi:10.1111/ejss.12567.
5. Rapid prediction of total petroleum hydrocarbon, polycyclic aromatic hydrocarbon and alkanes contamination in soils by a handheld mid-infrared spectroscopy. *J. Envron. Manage.* (Submitted).

1.6 References

Akpabio, E.M., Akpan, N.S., 2010. Governance and Oil Politics in Nigeria’s Niger Delta: The Question of Distributive Equity. *J. Hum. Ecol*. 30, 111–121.

Ambituuni, A., Amezaga, J., Emeseh, E., 2014. Analysis of safety and environmental regulations for downstream petroleum industry operations in Nigeria: Problems and prospects. Environ. Dev. 9, 43–60. doi:http://dx.doi.org/10.1016/j.envdev.2013.12.002

Amu, L.A.O., 1997. A Review of Nigeria’s Petroleum Industry. NNPC, Lagos.

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

Chakraborty, S., Weindorf, D.C., Li, B., Aldabaa, A.A.A., Ghosh, 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. *J. Environ. Qual.*, 39, 1378–1387.

Davies, O., and Abolude, D., 2016. Polycyclic aromatic hydrocarbons (pahs) of surface water from Oburun Lake, Niger Delta, Nigeria. *Appl. Sci. Res*. 13, 20-24.

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., Cipullo, S., Alamar, M.C., Coulon, F., Mouazen, A.M., 2018b. Evaluation of vis-NIR reflectance spectroscopy sensitivity to weathering for enhanced assessment of oil contaminated soils. *Sci. Total Environ*., 626, 1108-1120.

Dyck, R., Islam, M.S., Zargar, A., Mohapatra, A., Sadiq, R., 2013. ‘‘Application of data fusion in human health risk assessment for hydrocarbon mixtures on contaminated sites’’. *Toxicol.*, 313(2-3), 160-173.

Environmental Guidelines and Standards for the Petroleum Industries in Nigeria (EGASPIN) (1992) issued by the Department of Petroleum Resources, Nigeria (Revised edition, 2002).

Forrester, S., Janik, L., McLaughlin, M., and Gilkes, R.J., 2010. An infrared spectroscopic test for total petroleum hydrocarbon (TPH) contamination in soils., in: Proceedings of the 19th World Congress of Soil Science: Soil Solutions for a Changing World, Brisbane, Australia, 1-6 August 2010. Working Group 1.5 Soil Sense: *Rapid Soil Measurements*. pp. 13–16.

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 Sci. Soc. Am. J., 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.

Hua, Y., Luo, Z., Cheng, S., Xiang, R., 2012. ‘‘Health risks of organic contaminated soil in an out-of-service oil refinery site’’. *J. Earth Sci. China*, 23(1), 3-4.

Ite, A.E., Ibok, U.J., Ite, M.U., Petters, S.W., 2013. Petroleum Exploration and Production: Past and Present Environmental Issues in the Nigeria’s Niger Delta. *Am. J. Environ. Prot*. 1, 78–90.

Janik, L.J., Loibner, A,P., Kattner, J., Edelmann, E., 2015. US 9134227 B2: Method for determining polycyclic aromatic hydrocarbon contaminant concentration. Available at:http://patents.google.com/patent/US9134227 (Accessed:10 July 2018).

Kadafa, A.Y., 2012. Environmental Impacts of Oil Exploration and Exploitation in the Niger Delta of Nigeria. *Global Journal of Sci. Frontier Res. Envt. and Earth Sci*. 12 (3).

Malley, D.F., Hunter, K.N., Webster, G.R.B., Malley, D.F., Hunter, K.N., Webster, G.R.B., Barrie, G.R., 1999. Analysis of Diesel Fuel Contamination in Soils by Near-Infrared Reflectance Spectrometry and Solid Phase Microextraction-Gas Chromatography. *Soil Sediment Contam.*, 8, 481–489.

Nwilo, P.C., Badejo, O.T., 2006. Impacts and management of oil spill pollution along the Nigerian coastal areas. Adm. Mar. Spaces Int. Issues 119.

Okparanma, R.N., Coulon, F., and Mouazen, A.M., 2014a. Analysis of petroleum-contaminated soils by diffuse reflectance spectroscopy and sequential ultrasonic solvent extraction-gas chromatography. *Env. Pollut.*, 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. *Environ. Pollut.*, 192, 162–170.

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, Soil Pollut*., 224, 1539.

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.

OPEC, 2015. Nigeria: Facts and Figures [WWW Document]. URL http://www.opec.org/opec_web/en/about_us/167.htm

Sam, K., Coulon, F., Prpich, G., 2016. Working towards an integrated land contamination framework for Nigeria. *Sci. Total Environ.* doi:10.1016/j.scitotenv.2016.07.075.

Schwartz, G., Ben-Dor, E., Eshel, G., 2012. Quantitative analysis of total petroleum hydrocarbons in soils: comparison between reflectance spectroscopy and solvent extraction by 3 certified laboratories. *Appl. Environ. Soil Sci.*, 2012, 1-11.

UNEP, 2011. Environmental Assessment of Ogoniland. UNEP, Switzerland.

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

Webster, G.T., Soriona-Disla, J.M., Kirk, J., Janik, L.J., Forester, S.T., McLaughlin, M.J., Stewart, R.J., 2016. Rapid prediction of total petroleum hydrocarbons in soil using a handheld mid-infrared instrument. *Talanta* 160, 410-416.

CHAPTER 2 : 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

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

[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

Abstract: This review provides a critical insight into the selection of chromatographic and spectroscopic techniques for semi-quantitative and quantitative detection of petroleum hydrocarbons in soil and sediment matrices. Advantages and limitations of both field screening and laboratory-based techniques are discussed and recent advances in chemometrics to extract maximum information from a sample by using the optimal pre-processing and data mining techniques are presented. An integrated analytical framework based on spectroscopic techniques integration and data fusion for the rapid measurement and detection of on-site petroleum hydrocarbons is proposed. Furthermore, factors influencing petroleum hydrocarbons analysis in contaminated samples are discussed and recommendations on how to reduce their influence provided.

Key words: Analytical techniques, multi-sensor and data fusion, contaminated soil, petroleum hydrocarbons

2.1 Introduction

Land contamination from either poor historical industrial practices or incidents is a widespread and well-recognised environmental issue. In the EU alone, *ca.* 342,000 sites are affected by industrial activity leading to soil contamination (Van Liedekerke et al., 2014). Petroleum hydrocarbons (PHC) are common contaminants found in the environment. PHC encompass hundreds of various aromatic and aliphatic compounds as well as traces of heterocyclic compounds (containing sulphur, nitrogen, oxygen), which are well-known environmental contaminants (Cozzolino, 2015; Coulon et al., 2010). When the focus is about PHC, the difference between the terms PHC and total petroleum hydrocarbons (TPH) should be noted. PHC typically refer to the hydrogen and carbon containing compounds that originate from crude oil, whereas TPH refer to the measurable amount of petroleum-based hydrocarbons in an environmental matrix and, therefore, to the actual results obtained by sampling and chemical analysis (Coulon and Wu, 2017). Thus, TPH is a method-defined term and therefore the estimates of TPH concentrations will vary depending on the analytical method used to measure it. Historically this has been a significant source of inconsistency, as laboratories have different interpretations of the term TPH.

Over the last two decades, numerous field and laboratory techniques have been developed for the identification and quantification of TPH and polycyclic aromatic hydrocarbons (PAHs), as well as for the fractionation and quantification of aliphatic and aromatic hydrocarbons (Coulon and Wu, 2017; Li et al., 2015; Forester et al., 2013; Schwartz et al., 2012; Brassington et al., 2010). Field-based spectroscopic techniques offer rapid, non-destructive and cost-effective means of defining levels and distribution of PHC on-site before undertaking more costly and lengthy laboratory-based chemical analysis. In addition, they can provide real-time monitoring data and, therefore, be useful for initial site assessment and inform future sampling campaign for detailed risk assessment of the contaminated sites. However, one drawback of these field-based techniques is that they often fail to determine and quantify the entire range of PHC in soil or sediment. Therefore, choosing which technique to use is an important process to enable effective site investigation (Gałuszka et al., 2015); equally important it is to understand the type and quality of data generated (i.e., qualitative, semi-quantitative and quantitative). Moreover, their interpretation needs to be carefully evaluated before conclusions on a best technique to adopt can be drawn.

In contrast, laboratory techniques provide accurate analytical measurement and determination of hydrocarbons. They are, however, comparatively more expensive and require extra time for sample extraction and analysis (Forrester et al., 2013). Laboratory-based techniques include gas chromatography with flame ionization detector (GC-FID), GC coupled with mass spectrometry (GC-MS) or two-dimensional gas chromatography with FID (GC×GC-FID), GC×GC coupled with time-of-flight mass spectrometry (GC×GC/TOFMS), GC interfaced with quadrupole time-of-flight (GC-QTOF) tandem mass spectrometry. Raman spectroscopy (RS), infrared spectroscopy (IRS) and high performance liquid chromatography (HPLC) coupled with either fluorescence or ultraviolet visible detection. Among these techniques, GC-FID and GC-MS are the most common choices for PHC fingerprinting analysis of environmental matrices. The particular advantage of GC-FID is that the quantitative response of the FID is approximately the same for equal weights of any hydrocarbon, so that in a first approximation, relative peak areas can be used directly for the determination of weight percentage values (Malley et al. 1999). Similarly, GC-MS is used for more comprehensive analysis due to its ability to resolve and specify a broad range of hydrocarbon compounds, including hydrocarbon biomarkers (Coulon and Wu, 2017; Brassington et al., 2010; Barnes, 2009; Wang and Fingas, 1995). While GC-MS and GC-FID are mature techniques with excellent performance, there are still fascinating new developments such as GC×GC, Time-of-flight mass spectrometry (TOF-MS), and GC-QTOF amongst others.

Given the relative difficulty (and expense) of the GC techniques described above, there has recently been considerable efforts in finding satisfactory rapid measurement techniques to be used in the field. Optical methods such as visible and near-infrared spectroscopy (vis-NIRS), mid-infrared (MIR) and X-ray fluorescence (XRF) spectroscopy have been identified as suitable techniques for implementation in the laboratory and/or the field (Okparanma and Mouazen, 2012; Chakraborty et al., 2010). Portable vis-NIR spectrophotometers have been one of the most popular instruments used for on-site determination of a wide range of analytes since the 1990s (McMahon, 2007). They offer quick, cost-effective measurement and do not require sample preparation (He et al., 2007; Viscarra Rossel et al., 2006; McCarty et al., 2002). Likewise, MIR spectroscopy has been used for the detection of PHC (Horta et al., 2015). Although this technology is field-deployable, soil type and moisture can affect the measurement accuracy. Field portable X-ray fluorescence (PXRF) spectrometers also offer many advantages over traditional techniques including speed, portability, wide dynamic range of elemental quantification, little/no need for sample preparation and simplicity (Weindorf et al., 2014).

Research into multi-sensor and data fusion for the determination of soil properties has made significant advances (O'Rourke et al., 2016; Wang et al., 2015) during the last decade, yet it is still young for PHCs in environmental samples. Research needs to embrace combination of techniques for PHCs that are portable, rapid, and requires no consumables, making it attractive and economic. A multi-sensor and data fusion approach is the next step that may open new windows for new applications, where the performance of the current spectroscopic methods can be maximised (Mouazen et al., 2016). To the best of our knowledge, there is no study yet reviewing and/or demonstrating the potential of field-portable multi-sensor and data fusion for the analysis of PHC in contaminated soil and sediment.

This chapter provides (i) a critical review of the main laboratory and field chromatographic and spectroscopic techniques used in the determination of PHCs and fractions; (ii) insights into the advantages and limitations of both techniques; and (iii) discussion on the potential of optical, field-portable integrated framework of XRF+MIR, vis-NIR+MIR or XFR+vis-NIRS+MIR for timely, cost-effective and more accurate analysis of PHCs in soil and sediment.

2.2 Overview of analytical techniques for petroleum hydrocarbons detection

PHCs are separated into saturated and aromatic fractions; both fractions consist of highly complex mixture of hydrocarbons. The saturated fraction is composed of n-alkanes, branched alkanes and cycloalkanes and may also contain unsaturated hydrocarbons (alkenes). The aromatic fraction contains mainly compounds with two or more fused aromatic rings with or without a degree of alkylation. It may also contain polar non-hydrocarbons such as thiophenes, dibenzothiophenes and the oxygen-analogous aromatic heterocycles due to similar physico-chemical properties and therefore they are difficult to separate from the aromatic hydrocarbons. A number of analytical techniques have been developed for the quantification of PHCs in soil samples. This review, however, focuses only on chromatographic and spectroscopic techniques (Table 2-1).

Table 2-1: Most common chromatographic and spectroscopic techniques for determining petroleum hydrocarbon contaminants (PHCs) in soil and sediment samples.

Technique	Targeted analytes	Sample matrix	Sample* preparation	Measurement scale	Limit of detection	Advantages	Limitations	Reference

GC-MS	TPHs, PAHs	Sediment and soil	Either air or chemically dried samples
			SPME extraction using hexane and/or DCM, or acetone followed by Hex: DCM (1:1); use silica gel, florisil or alumina to clean up extract.
				Laboratory	1.0 mg/kg for individual PAH
					1.5 mg/kg Benzo(a)anthracene in sediment
					TPH = 50 mg/kg in soil
						Relatively sensitive and specific to quantify PAHs,
						Assess sediment quality for total PAHs,
						Detect signatures of priority PAHs in sediments.
							Untimely, high-labour sampling demanded,
							Uneconomic in assessing large-scale contamination,
							Use toxic solvent for extraction purposes (e.g., Soxhlet),
							Suitable for thermally stable analytes,
							Costly and time consuming analysis
								Wang and Fingas, 1995;
								Brassington et al.,2010;
								Poster et al., 2006;
								Okparanma and Mouazen, 2013;
								Chimezie et al. 2005
								Risdon et al., 2004

Technique	Targeted analytes	Sample matrix	Sample* preparation	Measurement scale	Limit of detection	Advantages	Limitations	Reference

GC-FID	TPHs, PAHs	Soil	Dry sample either in an oven at 105^{o}C or chemically using anhydrous Na_2SO_4, extract sample using hexane and DCM, use silica gel or alumina to clean up extract.	Laboratory	TPH = 10 mg/kg in soil
					PAH = 330 µg/kg in soil, TPH =2.30 mg/kg in soil/sediment matrix
						Simple,
						Detect wide Measure array of hydrocarbon compounds,
						Sensitive and selective,
						Applied both qualitatively and quantitatively.
							Costly and time consuming analysis,
							Instrument calibration difficulties,
							Effect of sample matrix, Suitable for thermally stable analytes
								Brassington et al., 2010;
								TPHCWG, 1998;
								Vallejo et al., 2001; Cortes et al., 2012

Vis-NIR spectroscopy	TPHs, PAHs	Soil and sediment	Air dry sample, crush and sieve to remove stones and plant residues. Field level: no sample preparation.	Laboratory and field	NA	Rapid, simple, inexpensive, Expedited site investigation, No prior site investigation Portable	Relatively fair accuracy, Affected by moisture content, Does not measure TPH directly hence has no LOD.	Deeks et al., 2014 Okparanma et al., 2014b
MIRS	TPH	Soil	Air dry sample and sieve	Laboratory and field	NA	Excellent detector for hydrocarbon	Affected by moisture content,	Horta et al., 2015; Sorak et

Technique	Targeted analytes	Sample matrix	Sample* preparation	Measurement scale	Limit of detection	Advantages	Limitations	Reference
						levels. Portable	Does not measure TPH directly hence has no LOD.	al., 2012
Portable Field gas chromatography	Volatile and semi-volatile hydrocarbons including TPHs PAHs	Soil, soil-gas	-	Field	[a]1-10 mg/kg in soil	Portable, lightweight, compact, durable, highest quality amongst other analytical techniques	Expensive due to the 'fit for purpose' gas carrier	Harris, 2003
Immunoassay	PAHs,	Soil	-	Field	TPH = 10-50 mg/kg in soil	Portable, quick, sensitive, economic, It complements chromatography procedures	Less affinity for hydrocarbons with rising soil clay content, Soil matrix effects	TPHCWG, 1998; Weisman, 1998

* There is no single, generic protocol for the analysis of hydrocarbons by GC. The methods vary considerably depending on the nature of the sample and the goals of the analysis. Readers are referred to the references provided for additional information on the extraction method. MIRS= mid-infrared spectroscopy, PXRF= portable X-ray fluorescence, TPH =total petroleum hydrocarbon, PAH=polycyclic aromatic hydrocarbon, NA=not available, GC-MS= Gas chromatography mass spectrometry, GC-FID= Gas chromatography coupled to flame ionization detection, [a]Source: United States Environmental Protection Agency (EPA 510-B-97-001) (Expedited Site Assessment Tools for Underground Storage Tank Sites. A Guide for Regulators.

2.3 Chromatographic techniques

2.3.1 Gas chromatography

Most environmentally important hydrocarbons are relatively volatile and thermally stable. Therefore, gas chromatographic techniques requiring the target compounds to be vaporised without destruction, have been established as the most important method for hydrocarbon separation. Gas chromatography (GC) is perhaps the most robust analytical instrument used for determining the structural composition and quantification of volatile mixtures such as TPH in environmental samples. The ability to couple highly sensitive detectors such as the flame ionization detector (FID) and mass spectrometry (MS) makes it a choice for highly sensitive petroleum analysis.

The principles are common to all chromatographic separation methods: the analytes of interest carried along by a mobile phase interact with a stationary phase and separate through these interactions. The separated analytes are detected as they elute. In GC, the mobile phase is helium the carrier gas. The stationary phase is typically a thin film chemically bonded to a narrow-bore capillary column. Most common coating used in hydrocarbon analysis are nonpolar stationary phases such as polydimethylsiloxanes or slightly more polar polysiloxanes in which a certain proportion (e.g., 5%) the methyl groups is substituted by phenyl groups. Important physical parameters influencing the separation characteristics of the analytical columns include the column length, its inner diameter and the film thickness of the stationary phase.

Flame ionization detection (FID) is the most used in gas chromatography than any other method for signal detection. This is because the burning of carbon compounds produces ions that will be detected by the FID. The success of FID resides mainly in its very low noise level, linear response over a very wide concentration range, and its sensitivity and its response varies very little with factors such as detector temperature and carrier gas flow rate (Weisman, 1998). However the FID response depends on the number of ions produced by a compound. Since this varies considerably between hydrocarbon classes, FID response factors vary accordingly (Karasek and Clement, 2003). The ability of a chromatographic method to successfully separate, identify and quantify species is determined by many factors as critically reviewed by Hibbert (2012). For example the observed GC retention times mainly depend on the temperature, flow rate and column length settings and, therefore, they are not ideal parameters for identification purposes. Instead, retention index (RI) also known as Kovats retention index is used to convert retention times into system-independent constants (Marriot et al., 2012; Song et al., 2002). Temperature oven optimisation may also be required to resolve specific target compounds such as diastereomers with very similar physical properties. While a comprehensive listing of all factors and solutions for optimising GC is beyond the scope of this review, there are several references on the gas chromatographic theory and principles, instrumentations and applications available -see Dettmer-Wilde and Engewald (2014).

Gas chromatography mass spectrometry (GC-MS) is a hyphenated analytical technique commonly used for environmental analysis due its specific and distinct monitoring capacity, especially when applied in the selective ion mode (Yang et al., 2015; Brassington et al., 2010; Wang and Fingas, 1995). The identification and characterisation of petroleum compounds by GC-MS is achieved by comparing retention time and a query mass spectrum with reference mass spectra in a library via spectrum matching. Versions of the NIST library, currently containing over 276,000 reference spectra, and search algorithms are available from all major MS manufacturers (Yang et al., 2015). Such method has been used to assess the PAHs in tar-contaminated soils (Lorenzi et al., 2010) and monitor bioremediation of PAH-contaminated soil via in-vessel composting using fresh organic waste (Zhang et al., 2011).

The MS analyzer can serve as both a selective and universal detector in the analysis of hydrocarbons. Electron impact at 70 eV is the most common mode, whereby an electron is stripped from the parent molecule (M) generating an M+ ion, which may undergo further fragmentation. Most hydrocarbons will only have one charge, so the mass is equivalent to the m/z ratio. Other methods, such as chemical, supersonic, and field ionization, are amenable to interfacing with gas chromatography and are soft ionization techniques that preferentially yield parent ions with limited fragmentation (Giri et al., 2017).

Recent studies also demonstrated that the performance compound identification depends on multiple factors including the mass spectrum library, spectral similarity measure and weight factors. They further showed that the compound identification based on mass spectra only has limited accuracy and the high accuracy compound identification can be achieved by incorporating compound separation information into mass spectrum matching. Since retention time in GC depends on experiment condition dependent, combination of retention index with mass spectrum is becoming more widely used (Marriot et al., 2012).

In comprehensive two-dimensional gas chromatography (GC × GC), the entire sample is subjected to two distinct analytical separations resulting in an enhanced separating capacity most useful for the characterisation of complex mixtures of organic compounds (Li et al., 2015). Additionally, it has been reported that combining two-dimensional gas chromatography (GC×GC) and TOF-MS can facilitates the identification of compounds by providing adequate spectrum acquisition speed, producing robust structural information without mass spectral skewing across the chromatographic peak (Li et al., 2015; Tran et al. 2010). This system has a high resolution of many co-elution substances including tricyclic and pentacyclic terpanes (Avila et al., 2010; Tran et al. 2010). Li et al. (2015) also used GC×GC-TOF-MS with a reverse-phase column system (one-dimensional polar column coupled with two-dimensional nonpolar column) in addition to the normal-phase system (one-dimensional nonpolar column coupled with two-dimensional polar column) to separate and identify components of crude oils. While the normal phase system is useful for separating hydrocarbons, especially high molecular weight compounds between C_{25}-C_{35} (Tissot and Welte, 1984), the reverse-phase system allows a greater separation for medium-low molecular weight cycloalkanes, which are normally very difficult to separate from aromatic hydrocarbons in normal phase system (Li et al., 2015). It also allows the identification of suitable biomarkers including steranes and terpanes (Li et al., 2015).

Both high-temperature and comprehensive two-dimensional GC provide relatively recent methodological advances for PHC analysis offering greater resolution and characterisation of complex mixtures of hydrocarbons. Specifically, high-temperature GC is a key technique in extending the molecular application range of gas chromatography.

Simulated distillation (SimDis) GC utilises fused silica column that considerably lowers the elution temperature of the analytes, which results in a decrease in the final oven temperature while ensuring a complete separation of the mixture (Boczkaj et al., 2011). This removes the chance of breakdown of less thermally stable mixture components and bleeding of the stationary phase thereby improving the detector signal. SimDis GC method permits the characterisation of the effective carbon number distribution of the constituent classes of soil extracts by a non-polar GC as a surrogate distillation column, where fractions are distilled using linear temperature profile (Pollard et al., 2004). SimDis GC of various fractions of soil extracts were achieved (Pollard et al., 2004) with a modified American Society for Testing and Materials (ASTM) method D2887-89 (ASTM, 1992). The authors performed SimDis GC using a Hewlett-Packard 5890 GC equipped with an on-column temperature controlled injector, an aluminium clad and Quadrex column coated with phenyl silicone. An oven temperature programmed at 55-420^{o}C was employed at a linear rate of 10^{o}C/min.

2.3.2 Liquid chromatography

Hydrocarbon analysis can be performed by various liquid chromatography techniques, such as thin layer chromatography (TLC), open-column liquid chromatography (OCLC), medium pressure liquid chromatography (MPLC) and high-performance liquid chromatography (HPLC) (Pan et al., 2013; Barman et al., 2000). Due to the nonpolar nature of the PHC normal-phase LC is commonly used (Chibwe et al., 2017). Accordingly, solvents or solvent systems used for isocratic or gradient elution are typically nonpolar in normal phase separations. A broad variety of detectors can be used for PHC, including spectroscopic (UV-Vis, fluorescence, infrared), bulk property (refractive index, evaporative light-scattering, dielectric constant, flame-ionisation), mass spectrometric and element specific detectors. UV-Vis detection provides excellent sensitivity for aromatic but is not applicable to saturated hydrocarbons. The utility of common atmospheric pressure ionisation interfaces used for on-line coupling of LC and mass spectrometry in the analysis of volatile and/or nonpolar compounds is rather limited. TLC coupled to flame-ionisation detection (FID) is an important compound group screening method in hydrocarbon analysis (Cavanagh et al., 1995). TLC-FID is a promising method for analysing oil fractions including aromatics. It has been used to separate solvent-extractable petroleum organics on silica-coated quartz rods into paraffins, aromatics and polar constituents (Dunn et al., 2000). Napolitano et al., (1998) also used TLC-FID as a quick way of measuring PHCs in soils.

HPLC separation is limited to aromatics but has a high sensitivity (Pan et al 2013). Greater interference due to co-elution is therefore more likely to occur for HPLC separation compared to GC separation. This will be especially marked in a heavily hydrocarbon contaminated environmental sample where there will be a large number of different PAHs (Coulon et al., 2012). HPLC techniques are applied much less, for oil-fingerprinting analysis in comparison to GC methods (Yang et al., 2015). The major disadvantage of HPLC applications for quantification is the lack of universal detector, which yields same response for all class of chemical constituents (Sarowha et al., 1997).

2.3.3 Portable/Field Gas Chromatography

The quest to cut down the expensive delays associated to laboratory-based GCs triggered portable designs of GCs. For instance, field gas chromatographs (FGCs), portable GC FROG 4000, and HAPSITE chemical identification system, among others are available. The FGCs measure constituent-definite analysis of soil-gas, soil, and water samples for volatile and semi-volatile hydrocarbons. FGCs are the only field measuring techniques for methyl tertiary butyl ether (MTBE) and they are of two types, namely, (person)-portable gas chromatographs (PGCs) and transportable gas chromatographs (TGCs) (EPA, 1997). PGCs are portable analytical devices used for hydrocarbons analysis. The PGCs possesses in-house batteries and carrier gas provider thus making the equipment portable. However, there is limited power supply due to the features (EPA, 1997). Portable gas chromatographs such as Portable FROG 4000 and Portable-GC-TMS weighs 2.2 kg and 4.5 kg, respectively (Koshy and Sudhakar, 2013). However, “fit for purpose” lightweight cylinders to supply the carrier gas have been recommended; thus, it tends to attract high cost (Deeks et al., 2014). Portable GC FROG 4000 has been applied onsite for analysing volatile organic compounds (VOCs) in soil, air and water in ppm and sub ppm in less than 5 min for benzene, toluene, ethylbenzene and xylene (BTEX) (California Geotechnical Services, 2016). With benchtop quality analysis, GC FROG 4000 satisfy the needs of various applications including site characterisation and assessment, soil characterisation, groundwater monitoring, Brownfield remediation, Superfund clean-up and leak detection (California Geotechnical Services, 2016).

HAPSITE is the only field-portable GC-MS for on-site detection, recognition and quantification of VOCs, toxic industrial chemicals (TICs) and chemical warfare agents (CWAs) [low molecular weight synthetic compounds that act very fast and are deadly at low concentration levels] (www.inficon.com). HAPSITE has the ability to detect and identify VOCs in parts per million to parts per trillion range. The results (obtained in minutes) from HAPSITE may be useful for the investigation of problems triggered by a very low concentration of contaminants that are essential for critical decision-making affecting human life, health and safety. Operators of HAPSITE require minimal training (www.inficon.com).

In conclusion, PGCs are field deployable and have less analysis run time. However, the method is not sensitive to many aliphatic compounds. HAPSITE measures very low contaminant’s concentration and is timely. HAPSITE results are useful for health and safety decision making. In addition, with European Standard ENISO 22155:2016, it is possible to measure volatile aromatic and halogenated hydrocarbons and selected aliphatic ethers in soil. ENISO 22155:2016 requires static headspace method for quantitative gas chromatographic measurements, and it is useful for all soil types. The limit of detection (LOD) depends upon the detection system used and the quality of the solvent (methanol grade) used for the extraction. In this method, the following LOD applies (expressed based on dry matter): typical LOD using GC-FID for volatile aromatic hydrocarbons is 0.2 mg/kg, aliphatic ethers such as methyl tert-butyl ether (MTBE) and tert-amyl methyl ether (TAME) is 0.5 mg/kg. Using GC-electron capture detection, the typical LOD for volatile halogenated hydrocarbons is 0.01 to 0.2 mg/kg (ENISO 22155:2016). However, there remain many obstacles to overcome so that a greater community of users can adequately and economically deploy this type of instrumentation. This instrumentation is still bulky (vacuum system, gas canister etc.), power hungry, and somewhat fragile.

TGCs are not person-portable (but transportable – heavy weight) and they separate well the constituents due to the presence of long capillary columns. TGCs can generate results comparable to laboratory quality (Koshy and Sudhakar, 2013). They accurately identify and quantify the constituents in samples.

Truly most common applications of field GC are the measurement of VOCs in air, such as BTEX, and chemical-warfare agents. Due to the huge demand for rapid, on-site analysis of environmental contaminants, there is a need for technological advancement in developing the already existing fast scanning techniques including GC×GC/TOFMS, and GC-QTOF tandem mass spectrometry to achieve analysis of contaminants in the field. This would help real-time decisions and cost-effective solutions to the challenges encountered during site investigation. The use and demand for field GC-MS will continue to grow as these instruments are miniaturised and performance remains at lab-quality. As more of these newer instruments enter the market, the costs will invariably drop to refuel the instrument development cycle.

2.4 Spectroscopic techniques

A number of spectroscopic techniques exist for the analysis of environmental contaminants (e.g., TPHs and PAHs). However, this current study focuses on the applications of XRFS, IR, vis-NIRS and MIRS for the analyses of TPH, PAH or both.

2.4.1 X-ray fluorescence spectroscopy (XRFS)

XRFS is a well-known laboratory technique (Hou et al., 2004). XRF functions on the principle that electrons embedded in the inner energy shell of an atom cleave from their shell upon excitation by X-rays. Electrons from the elevated, external, energy shells due to the discharge of excess energy in the form of an X-ray photon, occupy nearly instantly the voids in the inner shells created by electrons being cleaved following the X-ray excitation (Weindorf et al., 2014). The associated wavelength of the XRF thus depends on the energy level of the electrons in the interior shells. Moreover, the fluorescence emission is dependent upon the atom’s principal inner shell electrons taking part in excitation (Hou et al., 2004). Consequently, XRF detectors can measure the X-ray spectrum of any element, though they cannot efficiently measure elements with atomic numbers less than twelve (Horta et al., 2015). Hou et al. (2004) have previously reported the application of XRF for the analyses of PHCs in soil, water and liquid samples. With recent technological improvements, portable XRF spectrometers have now become available; they have been used to rapidly measure soil contaminants (with minimal sample preparation required) and offer a number of strengths relative to traditional laboratory-based methods (Horta et al., 2015). Both, wavelength dispersive x-ray fluorescence (WD-XRF) or energy dispersive x-ray fluorescence (ED-XRF) are commonly used as portable XRF instruments. The former is of higher resolution with fewer spectral overlaps and lower background intensities, but it is more expensive and prone to error than the latter. The ED-XRF analyser is designed to detect a group of elements all at once. One of the most advantages of XRF as a portable handheld device "gun-shaped meter" is that it can be taken to the field for analysis of soils *in situ*. The time of scanning is short, typically ranging between 60 to 90s. In addition, portable XRF instruments are operated by rechargeable Li-ion batteries that enable 6–12 h field measurements; thus, requiring no conventional electrical power supply on site. Indeed, XRF has been reported to be an accurate, non-destructive, and cost-effective method (Ulmanu et al., 2011); its use in environmental surveying has also been described (Hou et al., 2004).

Aside from its ability to mainly quantify and screen soil nutrients, the XRF technique has been used in combination with vis-NIR diffuse reflectance spectroscopy (DRS) to produce an optimised model for the swift measurement of soil HCs in Texas (Chakraborty et al., 2015). The authors concluded that the synergistic use of vis-NIR and XRF technique is viable for a quick and cost-effective quantification of petroleum contamination in soil.

2.4.2 Infrared spectroscopy

The electromagnetic spectrum of IRS consists of three regions i.e., near infrared (14000 – 4000 cm^{-1} or 750-2500 nm), mid-infrared (4000-400 cm^{-1}or 2500-25000 nm) and far infrared (400-10 cm^{-1} or 25000-1000000 nm). In this section the application of IRS (lab-based), vis-NIR and MIR techniques for the analysis of PHCs will be discussed separately.

IRS is generally applied for the measurement of organic compounds in soil, though some inorganic compounds may equally produce infrared signals (Horta et al., 2015). IR uses the stretching and bending modes of vibrations linked with molecules when they absorb energy in the infrared region of the electromagnetic spectrum for property clarification (Weisman, 1998). In this method, spectra of hydrocarbon compounds are generated from the carbon-hydrogen (e.g., C-H) linkages of saturated CH_2 and terminal CH_3 functional chemical groups, which are observed within the MIR spectral range of 3000-2900 cm^{-1} or a particular waveband of 2930 cm^{-1} (Weisman, 1998). To start with, samples are extracted using an eluting solvent with no C-H bonds. Prior to IR analysis, the eluate is passed through silica gel to eliminate biogenic polar compounds. Subsequently, the absorbance of the eluate is measured at the particular waveband, and further compared with a calibration curve made using petroleum hydrocarbon standards at known concentrations (Weisman, 1998).

IRS techniques were often employed for the detection of TPH in soils before the development of GC-based techniques (Current and Tilotta, 1997) due to its official acceptance by EPA (EPA method 418.1) (EPA, 1978) and International Organization for Standardisation (ISO) (e.g. ISO/TR 11046) (Becker et al., 2002). Currently, the use of IRS-based systems is scanty due to the ban of Freon (1,1,2-trichlorotrifluoethane, CFE) for solvent extraction (Forrester et al., 2010; Becker et al., 2002; Weisman, 1998). Furthermore, the ISO for France has replaced ISO/TR 11046 with ISO/DIS 16703, which suggests using GC-FID detector instead of IRS technique to follow extraction using non-halogenated solvent (ENISO16703:2011). In addition, the technique has been reported to be insensitive to unsaturated fractions of weathered hydrocarbons, showing no measurable adsorption bands at screening wavelength (Whittaker et al., 1995; Fan et al., 1994). IRS methods face with problems of interference (positive and negative); however, multivariate calibration annuls it. Sample porosity also affects IRS signal intensity (Forrester et al., 2010). Nevertheless, IRS methods provide quantitative responses, by employing calibrated standards with the analyser being positioned at the desired wavelength. Via a programmed calibration, concentration in parts per million (ppm) of the whole hydrocarbon can be determined (Deeks et al. 2014). IR-based techniques are simple, fast, and cost-effective with LOD of ~10mg/kg in soil, though they are not portable for field measurement (Weisman, 1998).

2.4.3 Visible and near-infrared (Vis-NIR) spectroscopy

The principle of near infrared (NIR) spectroscopy is based on the absorption of energy (generated by a light source) by substances, which result from fundamental vibrations of molecules that take place in the MIR range. Fundamental vibrations are of different modes but not limited to the stretching and bending of bonds that entails C-H, O-H, N-H and S-H chemical bonds (Osborne et al., 1993). However, in the NIR range (780 – 2500 nm) overtones and combinations of fundamental vibration are generated (Kuang et al., 2012). Infrared spectroscopy discriminates between chemical compounds by detecting the specific vibrational frequencies of molecular bonds, producing a unique infrared ‘spectral signal’ thereby enhancing its identification and quantification applying chemometrics.. In the visible (vis) range (400–780 nm), absorption bands related to soil colour are due to electron excitations, which assist the measurement of soil organic matter content and moisture content (Kuang et al., 2012; Viscarra Rossel et al., 2009).

In the late 1980s, the spectral characteristics of hydrocarbons were first documented (Cloutis, 1989). The spectra of hydrocarbons emanated primarily from either a combination or overtones of fundamental vibrations in the MIR region e.g., C-H stretching modes of aliphatic CH_2 and terminal CH_3 or aromatic functional groups (Aske et al., 2001). A comparison of average spectra between petroleum contaminated soils and non-contaminated soils is shown in Figure 2-1 (Chakraborty et al., 2015). These are similar in terms of optical intensity only in the visible range (Clark et al., 1990). However, in the NIR range the reflectance decreases with increasing contamination leading to increased absorbance and thus less reflectance than the non-contaminated samples (Hoerig et al., 2001).

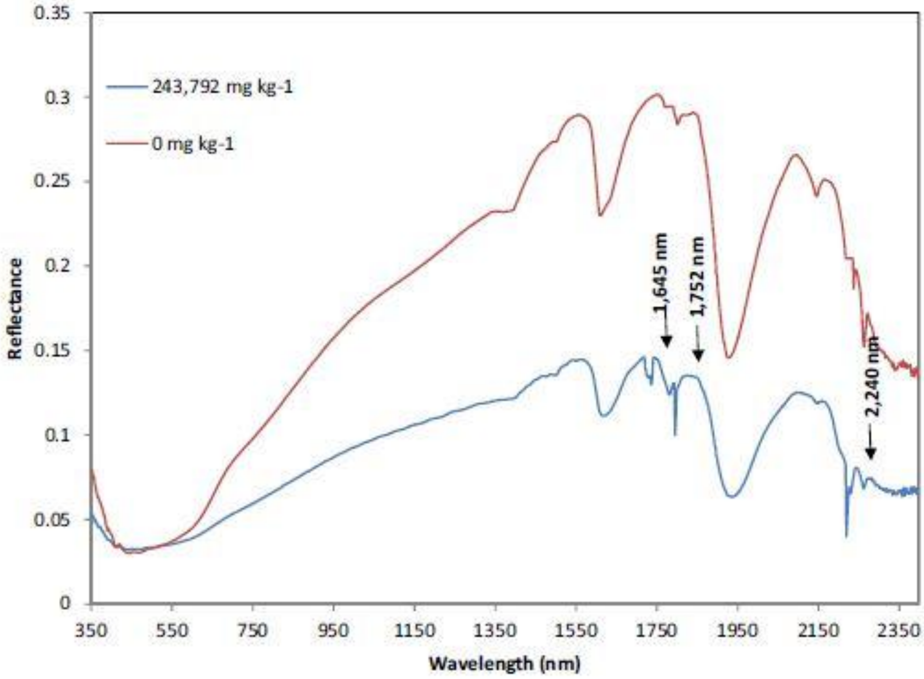


Figure 2-1: Visible and near infrared (vis-NIR) average reflectance for hydrocarbon contaminated (blue spectrum) and non-contaminated (red spectrum) soils (Chakraborty et al., 2015)

A considerable amount of literature has been published on the application of vis-NIR spectroscopy for the rapid estimation of soil PHCs (Okparanma et al., 2014a; Okparanma and Mouazen, 2013; Chakraborty et al., 2010; Bray et al., 2009; Malley et al. 1999). For example, Okparanma et al. (2014a) assessed the ability of vis-NIR diffuse reflectance spectroscopy (vis-NIR DRS) (350-2500 nm) for the measurement of petroleum hydrocarbon contamination in soils. The authors used sequential ultrasonic solvent extraction-gas chromatography (SUSE-GC) to measure PAH in soil samples. Both, the SUSE-GC measured data and the vis-NIR soil spectral data were pulled into one data matrix, and further subjected to a partial least square regression analysis. Prediction models with R^2 values ranging between 0.77 and 0.89, residual prediction deviation (RPD) values ranging between 1.86 and 3.12, and root mean square error ranging between 1.16 and 1.95 mg/kg were obtained. Though the PAH concentrations were low, vis-NIR reflectance response was also provided. Okparanma et al. (2014a) recommended that the method may be promising for quick evaluation of the spatial variability of PAHs in petroleum-contaminated soils and could assist site risk assessment.

The opportunity of employing vis-NIR DRS for the mapping of PAHs and the total toxicity equivalent concentration (TTEC) of PAH mixtures in different petroleum-discharge sites in the Niger Delta, Nigeria, was investigated by Okparanma et al. (2014b). The t-test results showed no significant (p>0.05) discrepancies between the GC-MS measured and vis-NIRS predicted PAH and TTECs maps (kappa coefficients = 0.19-0.56). The authors concluded that vis-NIR technique had good potential for monitoring hydrocarbon contamination in petroleum-discharged area. Okparanma and Mouazen (2013) assessed the applicability of vis-NIR DRS (350-2500 nm) to evaluate phenanthrene in 150 diesel-seeded soils. They used PLSR with cross-validation and obtained RPD values of 2.0 and 2.32, root mean square error of prediction (RMSEP) values of 0.21 and 0.25 mg/kg and R^2 values of 0.75 and 0.83 for validation and calibration, respectively. Other studies have investigated the capability of vis-NIR spectroscopy to assess PAHs in artificially contaminated soils (Okparanma and Mouazen, 2012; Bray et al., 2009; Malley et al., 1999). Using PLSR, Okparanma and Mouazen (2012) achieved an RMSEP of 0.2010 mg/kg, RPD of 2.75 and an R^2 of 0.89 for the calibration model. They suggested the potential of the technique to quantitatively characterise PAH in diesel-contaminated soils. With an ordinal logistic regression method, Bray et al. (2009) predicted total PAHs and benzo[a]pyrene using the vis-NIR technique. Their results showed good accuracy (90%) and a moderate to high false-positive rate at the low and high total PAH threshold, respectively.

NIR reflectance spectroscopy (1100-2498 nm) in combination with a step-by-step multiple linear regression were employed to predict the concentration of TPH in field diesel-contaminated soils (Malley et al., 1999), reporting a low accuracy and high prediction error. The low performance was attributed to (but not limited to) the small number of samples used and the inconsistency in the reference laboratory results. Chakraborty et al. (2010) evaluated the performance of vis-NIR DRS (350-2500 nm) to quantify PHCs contamination in soils. To achieve their research objective, 46 contaminated and control samples were collected from Louisiana, USA, after which the soil was scanned with a vis-NIR DRS as either ‘field-moist intact’ or ‘air-dry’ samples. Using both PLS regression and boosted regression tree (BRT) calibration models, the authors obtained a R^2 of 0.64 and a RPD of 1.70 as the best result for the prediction of TPH content from the field-moist scans, since the air-dried scans yielded 0.57 and 1.25

for R^2 and RPD, respectively. Authors concluded that there is the possibility of using vis-NIR DRS as a proximal soil-sensing tool for PHCs. However, this is true only for the analysis with moist soil samples, where the prediction performance was acceptable.

The applicability of vis-NIR for the analysis of TPH content in control soil samples seeded (spiked) with diesel and crude oil, and control soil mixed with oil was examined by Forrester et al. (2010), who reported a RMSE range of 4500-8000 mg/kg out of the TPH range of 0-100 000 mg/kg. Although the authors made no conclusions on their result, we concluded that since the RMSE upper limit is 8% of the upper TPH range, the result achieved is of small error. Thus, NIRS is a suitable screening tool for TPH measurement in soil. To examine the detection ability of vis-NIR for TPH in soils, Schwartz et al. (2012) utilised contaminated soils with a definite concentration of petroleum. Hydrocarbon analysis was carried out in three different certified laboratories; hence, the exact procedure was kept confidential. However, all the laboratories used the general methodology for the adjusted EPA 418.1 method. The measured results from these different certified laboratories in Israel were compared, and authors observed discrepancies between them; Laboratory A: 4575, 5288, 4932; Laboratory B: 6179, 6292, 6236; and Laboratory C: 3730, 4480, 4111 (represents minimum, maximum, and average concentrations of TPH (ppm), respectively. However, a satisfactory correlation from the plot of reflectance spectroscopy (4617 ppm) and the laboratories TPH (4500 ppm) versus projected TPH (5674 ppm) results was established. Consequently, they inferred that the accuracy of the vis-NIR spectroscopy technique was as promising as the commercial laboratories, and therefore it could be a feasible on-line sensing tool.

In a recent study, Chakraborty et al. (2015) combined XRF technique with vis-NIR diffuse reflectance spectroscopy (DRS) to produce an optimised model to predict PHC in soils from Texas, USA. Using a combined penalised spline regression (PSR) and random forest regression (RFR) modelling approach, authors obtained a R^2 of 0.78 and RPD of 2.19 and concluded that the this synthesised modelling methodology produced a better result compared to individual model based analysis, which resulted in RPD of 1.64, 1.86, and 1.96 for RFR, PSR and PLSR analyses, respectively.

Despite the potential advantage of this technique for measuring soil properties and detecting PHC in soils, only few studies have been carried out on contaminated sediment and soil samples. Therefore, further research is needed to boost the application and opportunities for spectroscopy in the future. Especially, vis-NIR DRS holds promising potential for rapid and cost-effective measurement of PHCs in soils, to inform risk assessment and decision support for remediation of agricultural lands. It is also important to mention that this technology offers portable systems that can be taken to the field to enable *in situ* measurement of PHCs, which is a fundamental requirement for accurate site-specific land reclamation, based on high sampling resolution data (Okparanma et al., 2014b). However, it should be noted that, vis-NIRS results can be affected by soil factors such as moisture content, soil types, ambient lights, etc.). Therefore, accounting for these external factors affecting the prediction performance is a key step for successful implementation of this sensing technology as a portable tool for field screening of PHCs in soils. Furthermore, it is important to note that accuracy reported so far by different research groups indicates that these detection methods are at a semi-quantitative stage, where more works to improve performance is needed.

2.4.4 Mid-infrared (MIR) spectroscopy

The principle of mid-infrared DRS is that molecules possess definite frequencies, and they vibrate in accordance with different energy levels (Horta et al., 2015). The fundamental vibrations of molecules when subjected to energy (e.g., light source) take place in the MIR range, which lead to absorption of light, to various degrees, with a specific energy quantum corresponding to the difference between two energy levels. As the energy quantum is directly related to frequency, the resulting absorption spectrum produces a characteristic shape that can be used for analytical purposes (Stenberg et al., 2010). Spectroscopy in the MIR range (2500-25000 nm) can rapidly capture soil information (Horta et al., 2015) important for soil contaminants assessment.

MIR spectroscopy (MIRS) is one of the rapid and cost-effective techniques developed for soil analyses (Bellon-Maurel and McBrtney, 2011). MIRS has been demonstrated to be a better measurement tool for soil total carbon, organic carbon, and inorganic carbon than vis-NIR spectroscopy (McCarty and Reeves, 2006; McCarty et al., 2002). MIR spectroscopy yields more informative spectra and peaks compared to NIR, which is characterised by broad bands of overtones and combinations (Soriano-Disla et al., 2014; Reeves, 2010). However, the superior performance of MIRS to vis-NIR is yet to be established in all soil science research (Vohland et al., 2014), although some literature indicated MIR spectroscopy to overcome the vis-NIR spectroscopy.

The potential application of MIR for the detection of PHCs in soils has been reported to be an excellent tool for hydrocarbon concentration in soils (Wartini et al., 2017; Webster et al., 2016; Horta et al., 2015). However, MIR accuracy and reproducibility are influenced by sample inhomogeneity and thus requires extra sample preparation (Horta et al., 2015). The applicability of MIR for the quantification of TPH in a control sample spiked with diesel and crude oil, and a control soil mixed with oil was examined by Forrester et al. (2010); they reported a smaller RMSE range of 2000-4000 mg/kg compared to NIRS (2000-8000 mg/kg), out of the TPH range of 0-100 000 mg/kg. Furthermore, Forrester et al. (2013) used real contaminated soil samples (205) to demonstrate the ability of MIR spectroscopy to detect TPH in soils, reporting RMSE<1000 mg/kg for the 0-15000 mg/kg of TPH content range, and recommended that this accuracy might be satisfactory in terms of screening. This study also presents an overview of analytical techniques, analyte, multivariate analyses and accuracy of different methods available for the analysis of soil contaminants (Table 2-2).

Table 2-2: Analytical techniques, multivariate analysis and machine learning results for the measurement of petroleum hydrocarbon contamination in soils and sediments.

Technique	Targeted analytes	Number of samples	Spectral range (nm)	Modelling technique	Statistical parameters	Sample origin	References
Vis-NIR DRS	TPH	46	350-2500	PLSR	R^2 = 0.79, RPD = 1.64, RMSEP = 0.353 mg/kg	USA	Chakraborty et al., 2010
				BRT	R^2 = 0.38, RPD = 1.38, RMSEP = 0.42 mg/kg		
Vis-NIR DRS	PAH	150	350-2500	PLS	R^2 = 0.89, RPD = 2.75, RMSEP = 0.2010 mg/kg	UK	Okparanma and Mouazen, 2012
Vis-NIR DRS	PAH	150	350-2500	PLSR	R^2 = 0.75-0.83, RPD = 2.0-2.32, RMSEP = 0.21-0.25 mg/kg	UK	Okparanma and Mouazen, 2013
Vis-NIR	PAH	137	350-2500		R^2 = 0.77-0.89, RPD = 1.86-3.12,	Nigeria	Okparanma

Technique	Targeted analytes	Number of samples	Spectral range (nm)	Modelling technique	Statistical parameters	Sample origin	References
DRS					RMSEP = 1.16-1.95 mg/kg		et al., 2014a
Vis-NIR DRS	TPH	108	350-2500	RFR PSR PLSR	RPD=1.64 RPD=1.86 RPD=1.96	USA	Chakraborty et al., 2015
PXRF+Vis-NIR DRS	TPH	108	350-2500	PSR	R^2 = 0.78, RPD = 2.19	USA	Chakraborty et al., 2015
MIR	TPH	205	2170-3330	PLSCV	RMSE<1000mg/kg for 0-15000mg/kg	Not stated	Forrester et al., 2013
MIR	TPH	67		PLSR	R^2 =0.99, RMSE <200 mg/kg	Australia	Webster et al., 2016
GC-MS	PAH	150	Not applicable	PLSR	R^2_P =0.89, RPD:1.52-2.79, RMSE=0.201mg/kg		Osborne et al., 1993
GC-FID	TPH	26	Not applicable	Stepwise MLR	R^2_P: 0.68-0.72, RPD: 0.84-1.00		Malley et al., 1999

Technique	Targeted analytes	Number of samples	Spectral range (nm)	Modelling technique	Statistical parameters	Sample origin	References
FTIR	TPH	172	400-2500	PLSR	R^2cv=0.81 RMSECV=4,500-8000 mg/kg	Not stated	Forester et al., 2010
FTIR	PAH	65	350-2500	OLR	Accuracy (65.90.25%), FPR (0.57-0.91) FNR (0.03-0.13)	Wales	Bray et al., 2010

Vis-NIR DRS = visible and near-infrared diffuse reflectance spectroscopy, PXRF = portable x-ray fluorescence, MIR = mid-infrared, GC-MS= gas chromatography-mass spectrometry, GC-FID = gas chromatography-flame ionization detector, PLSR= partial least squares regression, PLSCV = partial least squares cross-validation, ANN = artificial neural network, PSR= penalized spline regression, RFR = random forest regression, BRT= boosted regression tree, MLR = multiple linear regression, R = coefficient of determination, RPD = residual prediction deviation, RMSEP = root mean square error of prediction, TPH = total petroleum hydrocarbon, PAH = polycyclic aromatic hydrocarbon, GEMAS = geochemical mapping of agricultural soils and grazing land of Europe, FTIR = Fourier transform infrared spectroscopy, OLR = ordinal logistic regression, FPR=false-positive rate, FNR=false-negative rate.

Like in the vis-NIR spectroscopy case, advances in MIR spectroscopy have made portable systems available for *in situ* measurement of different soil properties, including PHCs. However, it should be noted here that, although sharp and clear signatures of organic pollutants can be obtained with MIR spectroscopy, which is encouraging for accurate measurements. MIR spectroscopy is susceptible to soil moisture content (MC), limiting field applications. In comparison with NIR spectroscopy, the effect of water on spectral response is more severe with MIR spectroscopy. This necessitates advanced data mining techniques to remove the influence of MC. Recent studies on the use of the vis-NIR spectroscopy for the measurement of other soil properties proved that MC effect can be removed by adopting direct standardisation of external parameter orthogonalisation techniques (Ji et al., 2015). These techniques are yet to be tested in the MIR spectroscopy, particularly for PHC contamination detection. Furthermore, other approaches that can be adopted to remove the water effect in soil samples is to classify spectra into different soil water classes, for each specific calibration models of soil consistent is developed (Mouazen et al., 2006). From the brief review on MIR spectroscopy, we can conclude that this technique is commonly used for the measurement of various soil PHCs contaminants. However, MIR accuracy and reproducibility are affected by sample heterogeneity, thus requiring extra sample processing. The technique is field-deployable (Sorak et al., 2012) though MC effect is a limiting factor in the field. Nevertheless, advanced data mining approaches can remove the influence on MC.

Another IR method for analysis of soil hydrocarbon is attenuated total reflectance (ATR) spectroscopy. Quantitative analysis of TPH in oil-contaminated soils by ATR-IR spectroscopy was documented (Guryanova et al., 2016). The authors demonstrated the feasibility of oil hydrocarbon contamination analysis of soil ATR-IR spectroscopy with an infrared fibre-based immersion probe without any sample pre-treatment. Multivariate modelling of ATR-IR spectroscopic data of samples was implemented using PLS regression method. The authors reported calibration results in terms of root mean square error, RMSE =1.51 mg/kg and coefficient of determination, R =0.824. They concluded that the proposed methodology can be utilised to develop a portable real-time in-situ field analyser presenting a viable alternative to laboratory analysis of collected soil samples. However, the authors further recommend modification of the probe to sample interface so as to enlarge the contact surface to enhance oil hydrocarbon determination; and increase the number and variety of soils in the calibration to enhance analysis accuracy.

2.5 Multi-sensor and data fusion approach

A multi-sensor and data fusion approach has recently been introduced in digital agriculture, as a tool to improve soil and crop management (Grunwald et al., 2015; Kuang et al., 2012). Also, recent reports confirmed that this approach was extended to the environmental sector e.g., to measure and manage PHC in soils (e.g., Chakraborty et al., 2015; Horta et al., 2015). However, it is worth to stress that multi-sensor and data fusion approach is more common for field measurement scenarios, which allow overcoming the major shortcoming of these technologies regarding accuracy. In this sense, it can be hypothesised that by the integration of more than one field sensor and advanced data fusion modelling, improvement in calibration accuracy is expected compared to that provided by individual sensing technology. Although field measurement methods have been used independently for environmental analysis, they are yet to be integrated into single use (data fusion) for swift and better environmental analysis. While multi-sensor is the use of more than one sensor (hardware) when collecting multi-data layer from one sample or spot, data fusion is the integration and modelling of the multi-data layer from different sources to produce more accurate (reliable) quantitative assessment of a property, which could not be attained from a single source (Horta et al., 2015).

Data fusion, as a methodology for environmental analysis, is new and has so far attracted little attention in the literature. Fused XRF data and vis-NIR spectra was used to produce an optimised model for swift and more accurate measurement of soil PHC in Texas (Chakraborty et al., 2015). Using spectral libraries and field validation, Horta et al. (2015) reported that the synergistic use of vis-NIR and XRF spectrometry data is possible for better soil contaminant analysis, nevertheless, they recommended also the need to develop unique calibration methods. However, portable sensing technologies are not restricted to vis-NIRS, but MIRS (Sorak et al., 2012), micro spectroscopy, XRFS, and others that may well be integrated and their multi-layer data analysed. Currently, there is no study yet integrating XRFS, MIRS, and vis-NIRS optical sensors for the evaluation of soil PHCs. Thus, a field-portable integrated framework of XRF+MIR, vis-NIR+MIR or XFR+vis-NIRS+MIR to analyse PHCs in soils and sediments has never been proposed. The synergistic use of these combinations, albeit complexity and increased capital cost, is portability, requirement for little or no consumable, and minimum or no samples preparation. With these advantages, the higher capital cost would be recovered within a short period of time, as cost of analysing TPH per sample can be high. Either of the hybrids would benefit environment regulators and remediation experts. The workflow for the "newly integrated approach of multi-sensor and data fusion'' based on chemometrics, or machine learning is illustrated in Figure 2-2.

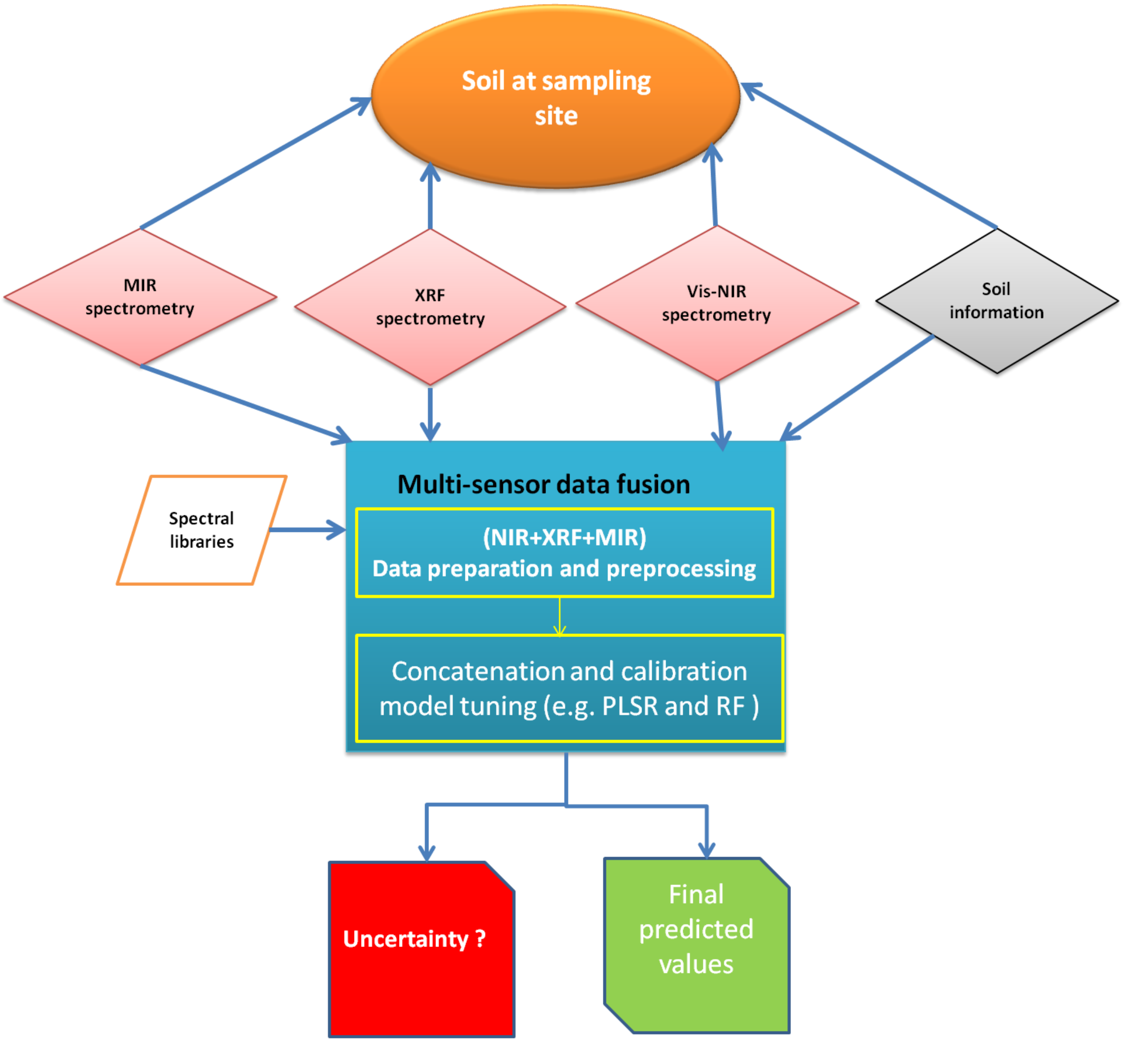


Figure 2-2: The integrated concept of multi-sensor and data fusion for the measurement of petroleum hydrocarbons (PHC) in soil and sediments.

In this approach the three spectrometers are transferred to the field (*in situ* measurement), or soil samples are brought to the laboratory (laboratory-based analysis). The multi-data layers obtained from the three sensing technologies are pooled together in one matrix, subjected to data pre-processing, before multivariate statistics (e.g., PLSR) and machine learning (e.g. artificial neural network ANN), support vector machine (SVM), and random forest (RF) modelling techniques are used to establish calibration models to predict PHCs in soils. From the few successes made in earlier studies (Chakraborty et al., 2015; Wang et al., 2015) (Table 2-3) with data fusion technology, it is expected that the multi-sensor and data fusion outlined in the present paper would be effective and feasible for analysing soil PHCs contaminants. This has to be validated with experimental work in the future. However, data fusion has limitations since each technique has different requirements on sample preparation but spectral measurements are acquired at the same time. These are currently key challenges against the implementation of the approach. However, it is hopeful that this will be a reality subject to methodological and technological advancement.

Table 2-3: Comparison of data fusion approach and performance for targeted analytes in soil.

Technique	**Targeted analytes**	**Multivariate technique**	**Sample matrix**	**Statistical parameters**	**Reference**
Vis-NIRS	TPH	PSR	**Soil**	R^2=0.70, RMSE=0.75 mg/kg, RPD=1.86	Chakraborty et al., 2015
Vis-NIRS+ XRF	TPH	PSR	**Soil**	R^2=0.73, RMSE=0.59 mg/kg, RPD=1.96	
Vis-NIRS+ XRF	TPH	PSR +RFR	Soil	R^2=0.78, RMSE=0.53 mg/kg, RPD=2.19	
XRF	TN	RFR	Soil	R^2=0.9, RMSEP=0.02 mg/kg, RPD=3.20	Wang et al., 2015
XRF	TC	RFR	Soil	R^2=0.77, RMSE=0.336 mg/kg, RPD=2.11	

Technique	**Targeted analytes**	**Multivariate technique**	**Sample matrix**	**Statistical parameters**	**Reference**
Vis-NIRS	TPH	PSR	**Soil**	R^2=0.70, RMSE=0.75 mg/kg, RPD=1.86	Chakraborty et al., 2015
Vis-NIRS+ XRF	TPH	PSR	**Soil**	R^2=0.73, RMSE=0.59 mg/kg, RPD=1.96	
Vis-NIRS	TN	RFR	Soil	R^2=0.90, RMSE=0.019 mg/kg, RPD=3.23	
Vis-NIRS	TC	RFR	Soil	R^2=0.81, RMSE=0.331 mg/kg, RPD=2.33	
Vis_NIRS+PXRF	TN	RFR	Soil	R^2=0.91, RMSE=0.019 mg/kg, RPD=3.39	
Vis-NIRS+PXRF	TC	RFR	Soil	R^2=0.83, RMSE=0.319 mg/kg, RPD=2.42	

R = coefficient of determination, RPD = residual prediction deviation, RMSE = root mean square error, RMSEP = root mean square error of prediction, PSR = penalised spline regression, RFR = random forest regression, Vis-NIRS=visible and near infrared spectroscopy, XRF = X-ray fluorescence, TN = total nitrogen, TC = total carbon, TPH = total petroleum hydrocarbon.

2.6 Decision making in selecting a detection techniques: advantages and limitations

The first step towards decision making on the best technique to measure a source of contamination is driven by time, cost, and the final application of results. For example, in cases where time is not a crucial factor and accuracy is more appealing, laboratory measurement techniques are the most appropriate option, as accuracy is higher than field techniques. However, with recent advances in sensing technologies and IT infrastructure, field equipment become available, which may soon become real competitors of current tradition laboratory analytical techniques (e.g., the gas chromatography); particularly, if the current challenges, e.g. accuracy, can be overcome or at least minimised with advanced modelling techniques. One example of potential solution is the multi-sensor and data fusion approach. A wide range of field measuring techniques is available for quick measurement of PHCs in soil (Okparanma and Mouazen, 2013), although no single technique measures the whole range of PHCs. Thus, the detection of these contaminants depends on the samples and the analytical technique employed (Deeks et al., 2014). Therefore, choice of technique is important in conducting effective measurement of PHCs in soil and sediment. Field measuring techniques should be cost-effective, timesaving, portable, and provide sufficient accuracy in detecting and monitoring PHCs contamination levels in soil and sediment, rather than expensively analysing samples later in the laboratory (Barnes, 2009). These advantages allow field techniques to enable collecting a high number of samples per field area in a relatively short period, which is a crucial requirement for precision land reclamation (Okparanma et al., 2014b). This is because by enabling high sampling resolution to be collected, better spatial sample coverage and thus a better understanding and characterisation of the contaminated area can be obtained. The accuracy and limits of detection in field measuring methods are advancing and some may be applied to detect low concentration or even targeted chemicals (Deeks et al., 2014). However, the analytical quality of these techniques may be less accurate, and at a semi-quantitative range, compared to laboratory analysis for the measurement of PHCs in soils and sediments.

Table 2-4 shows the factors influencing the decision making process in selecting analytical techniques. Factors such as analysis run time, analysis cost per sample, operational skills and limitations were considered for decision-making. Among the techniques, there are currently no analysis cost per sample for vis-NIR, PXRF and portable GC-MS methods. Thus, there is an active research need in this area to enhance the decision-making process for analytical methods in environmental analysis. There is also need for research into the analysis run time of vis-NIR spectroscopy. In the area of operational skills, field GC, portable GC-MS, vis-NIR spectroscopy and PXRF require medium to high skill. While headspace FIDs and headspace PIDs require low to medium skill, Immunoassay test kits requires medium skill. To select the best analytical technique for environmental analysis, the highlighted research needs have to be addressed.

Table 2-4: Factors influencing decision in selecting analytical techniques.

Technique	Analysis run time (min)	Analysis cost per sample	Expertise needed	Limitations
Headspace analysis: PIDs	[a]1-30	[a]£0.69 - 6.89	L-M	Less sensitive to detect aromatic hydrocarbons, High amount of organic content can affect the measurements
Headspace analysis: FIDs	[a]1-30	[a]£0.69-£6.89	L-M	Less sensitive to aliphatic hydrocarbons, High organic content can affect the measurements
Field Gas Chromatographs	[a]10-60	[a]£13.78 - £48.22	M-H	A skilled operator is needed
Portable GC/MS:	[b]10	na	M-H	A skilled operator is needed, Requires prior sample extraction, on-site carrier gas, Insensitivity issues, particularly microchip GCs
Vis-NIR Spectrophotometer	na	na	M-H	Comparable accuracy for heavy metals and hydrocarbon
Portable x-ray fluorescence	30 s - 2 min	na	M-H	A skilled operator is needed
Immunoassay test kit	[a]30-45	[a]£13.78 - £41.34	M	Cross-reactivity may impact interpretation of result

Key: L = low, M = medium, H = high, na = not available, [a] Source: United States Environmental Protection Agency (EPA 510-B-97-001) (Expedited Site Assessment Tools for Underground Storage Tank Sites. A Guide for Regulators, [b] Source: Harris (2003).

2.7 Conclusion

A plethora of chromatographic and spectroscopy techniques and extraction methods for the analysis of petroleum hydrocarbons (PHCs) in soil and sediments are available in the literature. This literature review has discussed both laboratory and field techniques, and showed that no method is problem-free, but there are issues of different magnitudes. For example, it has been documented that both near infrared spectroscopy (NIRS) and mid infrared spectroscopy (MIRS) are affected by moisture content (MC), which has to be accounted for in field measurement protocols by adopting appropriate modelling techniques.

The high selective and sensitive of gas chromatographic lab-based techniques makes them the preferred choice for the identification and quantification of hydrocarbon contamination in environmental samples. However, they can be time-consuming and required a high level of expertise. In contrast, field portable GC techniques offer direct on-site analysis of samples for quick detection and measurement.

Recent advances made with field spectroscopy methods (e.g., X-ray fluorescence [XRF], mid-infrared [MIR] and visible and near infrared [vis-NIR]) suggest that the development of field techniques towards practical applicability still have to follow; and the literature provides rather proof-of-concepts-studies so far. However, these field-portable methods and the implementation of a multi-sensor and data fusion approach improve PHCs prediction accuracy over individual sensing technologies. We believe that there are huge research opportunity for improved field measurements of contaminants in soil and sediment if data fusion from different optical sensors could be integrated. The best spectroscopy combination candidates from environmental prospective, and which have not been investigated yet, include XRF+MIR, vis-NIR+MIR or XFR+vis-NIRS+MIR. There is the chance that this synergy—rather than a single technique—could produce more reliable and accurate information for the mapping of contaminants in petroleum release sites. MIR and vis-NIR spectroscopy are candidate techniques for analysing PHCs, while XRF is widely known for the analysis of heavy metals and inorganic compounds. However, the fusion of XRF elemental data and vis-NIR spectra has shown to improve the quantification accuracy of soil TPH.

When optimal sensor combination, data mining and modelling technique is established, and when a successful technique to remove the negative influences on moisture content is implemented, high sampling resolution per unit field area can be collected *in situ*; this will assist in contaminated site remediation, contaminated land management, and risk assessment of petroleum hydrocarbon on human and welfare health.

This approach would be useful in the future. However, in order to test the feasibility and potential application of the combination of Vis-NIR, MIR and XRF spectrometry spectral data for rapid and cost-effective analysis of soil PHCs contaminants, there is need for technological advancement in the proposed synergistic method together with special calibration approaches; and a pilot study needs to be conducted. In addition, multivariate modelling needs to be carried out with the conjoint data, using nonlinear analytical methods including artificial neural network and support vector machine, instead of commonly applied linear methods like PLS regression. Further work is also imperative in the area of analysis run time for generic field measurement methods like Vis-NIRS. This would facilitate selection of techniques for petroleum hydrocarbons detection in soil.

2.8 References

American Society for Testing and Materials., 1992. Standard testing methods for Boiling Range Distribution of Petroleum Fraction by Gas Chromatography D2887-89, ASTM Annual Book of Standards, Vol. 05.02, ASTM, Philadelphia, 483.

Aske, N., Kallevik, H., Sjöblom, J., 2001. Determination of saturate, aromatic, resin, and asphaltenic (SARA) components in crude oils by means of infrared and near-infrared spectroscopy. *Energy & Fuels* 15, 1304–1312.

Avila, B.M.F., Aguirar, A., Gomes, A.O., Azevedo, D.A., 2010. Characterization of Extra Heavy Gas Oil Biomarkers using Comprehensive Two-dimensional Gas Chromatography Coupled to Time-of-Flight Mass Spectrometry, *Org.Geochem.*, 41, 863.

Barman, B.N., Cebolla, V.L., Membrado, L., 2000. Chromatographic techniques for petroleum and related products. *Critical Reviews in Analytical Chemistry*, 30:75-120.

Barnes, B., 2009. Framework for the use of rapid measurement techniques (RMT) in the risk management of land contamination. Environment Agency,Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol BS32 4UD, UK, pp. 1-90.

Becker, R., Koch, M., Wachholz, S., Win, T., 2002. Quantification of total petrol hydrocarbons (TPH) in soil by IR-spectrometry and gas chromatography – conclusions from three proficiency testing rounds. *Accreditation and Quality Assurance*, 7, 286–289.

Boczkaj, G., Przyjazny, A., Kaminski, M., 2011. A new procedure for the determination of distillation temperature distribution of high-boiling petroleum products and fractions. *Anal. Bioanal Chem*, 399, 3253-3260.

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*, pp. 2488–2499.

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

Cavanagh, J.E., Juhasz, A.L., Nichols, P.D., Franzmann, P.D., McMeekin, T.A., 1995.

Analysis of microbial hydrocarbon degradation using TLC-FID. *Journal of Microbiological Methods* 22, 119-130.

California Geotechnical Services., 2016. Portable GC Frog 4000 for Environmental Remediation, Air Quality Testing, and More. (Available at: http://www.geotechnical.net/portable-gc-frog-4000.shtml, accessed 10.04.2016).

Chakraborty, S., Weindorf, D.C., Li, B., Aldabaa, A.A.A., Ghosh, 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. *J. Environ. Qual.,* 39, 1378–1387.

Chibwe, L., Davie-Martin, C.L., Aitken, M.D., Hoh E., Massey Simonich, S.L., 2017. Identification of polar transformation products and high molecular weight polycyclic aromtaic hydrocarbons (PAHs) in contaminated soil floolowing remediation. Sci. *Total Environ*., 599-600, 1099-1107.

Chimezie, A., Anthony, O., Pete, P., Herbert, C., Ukpo, G., Ogah, C., 2005. GC/MS analysis of polynuclear aromatic hydrocarbons in sediment samples from the Niger Delta region.*Chemosphere* 60, 990-997.

Clark,R.N., King, T.V.V., Klejwa, M., Swayze, G.A., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals.*J. Geophys. Res*., 95 (B8), 12653-12680.

Cortes, J.E., Suspes, A., Roa, S., Gonzalex, C., Castro, H.E., 2012. Total Petroleum Hydrocarbons by Gas Chromatography inColombian Waters and Soils. *American Journal of Environmental Science*, 8(4), 396-402.

Cloutis, E.A., 1989. Spectral reflectance properties of hydrocarbons: remote-sensing implications. *Science,* 245, 165–168.

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.

Coulon F., Brassington K.J., Bazin R., Linnet P.E., Thomas K.A., Mitchell T.R., Lethbridge G., Smith J.W.N., Pollard S.J.T., 2012. Effect of fertiliser formulation and bioaugmentation on biodegradation and leaching of crude oils and refined products in soils. *Environmental Technology*, 33, 1879-1893.

Coulon F., Wu G., 2017. Determination of petroleum hydrocarbon compounds from soils and sediments using ultrasonic extraction. In: *Hydrocarbon and Lipid Microbiology Protocols* McGenity T.J et al. (eds.) Springer-Verlag Berlin Heidelberg, pp 31-46.

Cozzolino, D., 2015. Near infrared spectroscopy as a tool to monitor contaminants in soil, sediments and water – state of the art, advantages and pitfalls. *Trends Environ. Anal. Chem. 9, 1-7.*

Current, R. W., Tilotta, D. C., 1997. Determination of total petroleum hydrocarbons in soil by on-line supercritical fluid extraction-infrared spectroscopy using a fibre-optic transmission cell and a simple filter spectrometer. *Journal of Chromatography A*, 785, 269–277.

Deeks, L., Coulon, F., Tibeth, M., Kirk, G., Mouazen, A.M., Tothill, L., and Walton, C., 2014. Practical guidance document for field screening technologies of hydrocarbons and associated metals in soil and water. *Energy Institute Technology Publication*, ISBN 9780852937044. Available at:http://publishing.energyinst.org/publication/ei-technical-publications/environment/refinery-emissions/practical-guidance-on-technologies-for-field-screening-hydrocarbons-and-associated-metals-in-soil-and-water.

Dettmer-Wilde, K., Engewald W. (ed)., 2014 Practical gas chromatography – A comprehensive reference, Springer Heidelberg, ISBN 978-3-642-54639-6, 893.

Dunn, K., Chilingarian, G.V., Lian, H., Wang, Y.Y., Yen, T.F., 2000. Chapter 11: Analysis of Asphalt and its components by Thin-Layer Chromatography. *Development in Petroleum Science*, 40, pp. 305-317.European Standard ENISO 22155 (2016). Soil quality-gas chromatographic determination of volatile aromatic and halogenated hydrocarbons and selected ethers-static headspace method.

EPA., 1997. Chapter VI Field methods for the analysis of petroleum hydrocarbons.In:

EPA Expedited Site Assessment Tools for Underground Storage Tank Sites: A 87 Final draft guide for regulators. United States Environmental Protection Agency Office of Underground Storage Tanks, OSWER, Washington DC, pp. VI-1 to VI-52.

EPA Method 418.1., 1978. Total Recoverable Petroleum Hydrocarbons by IR. Government Printing Office, Washington, DC, USA.

Fan, C. Y., Krishnamurthy, S., Chen, C. T., 1994. A critical review of analytical approaches for petroleum contaminated soil. In: T. A. O'Shay and K. B. Hoddinott (Editors). Analysis of soil contaminated with petroleum constituents. American Society for Testing and Materials ASTM STP 1221, Philadelphia, PA. pp. 61–74.

Forrester, S., Janik, L., McLaughlin, M., Gilkes, R.J., 2010. An infrared spectroscopic test for total petroleum hydrocarbon (TPH) contamination in soils., in: Proceedings of the 19th World Congress of Soil Science: Soil Solutions for a Changing World, Brisbane, Australia, 1-6 August 2010. Working Group 1.5 Soil Sense: *Rapid Soil Measurements*. pp. 13–16.

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 Sci. Soc. Am. J., 77(2)*, 450-460.

Gałuszka, A., Migaszewski, Z.M., Namieśnik, J., 2015. Moving your laboratories to the field – Advantages and limitations of the use of field portable instruments in environmental sample analysis. *Environ. Res.*, 140, 593–603.

Grob, R.L., Garry, E.F., 2004. Modern practice of gas chromatography 4th edition.Wiley and Sons, NewYork.

Grunwald, S., Vasques, G.M., Rivero, R.G., 2015. Fusion of Soil and Remote Sensing Data to Model Soil Properties. *Adva., Agro.*, 131, 1-191.

Giri, A., Coutriade, M., Racaud, A., Okuda, K., Dane, J., Cody, R.B., Focant, J-F., 2017. Molecular characterisation of volatiles and petrochemical base oils by photo-ionization GC×GC-TOF-MS. *Anal. Chem.*, 89, 5395-540.

Guryanova, A., Ermakov, V., Galyanin, V., Artyushenko, V., Sakharova, T., Usenov, I., Bykov, D., Bogomolov, A., 2016. Quantitative analysis of total hydrocarbons and water in oil-contaminated soils with attenuated total reflectance infrared spectroscopy. *J. Chemometrics*, 31, 1-10.

Harris, C.M., 2003. Today's chemist at work. American Chemical Society, 33-38.

He, Y., Tang, L., Wu, X., Hou, X., Lee, Y, I., 2007. Spectroscopy: the best way toward green analytical chemistry? *Appl. Spectrosc. Res*., 42, 119-138.

Hibbert, D.B., 2012. Experimental design in chromatography: A tutorial review. *J. Chromatogr*. B, 919, 2-13.

Hoerig, B., Kuehn, F., Oschuetz, F., Lehmann, F., 2001. HyMap hyperspectral remote sensing to detect hydrocarbon. *Int. J. Remote Sens.*, 8, 1413-1422.

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.

Hou, X., He, Y., Jones, B.T., 2004. Recent Advances in Portable X-Ray Fluorescence Spectrometry. *Appl. Spectrosc. Rev*., 39, 1–25.

Ji, W., Li, S., Chen, S., Shi, Z., Viscarra Rossel, R.A., Mouazen, A.M., 2015. Prediction of soil attributes using the Chinese soil spectral library and standard spectra recorded at field conditions. *Soil Tillage Res*., 155, 492-500.

Koshy, V.J., Sudhakar, P., 2013. Gas Chromatographs for Environmental Field Analysis. Available in www.envirotech-online.com. Accessed 18 April 2017.

Kuang, B., Mahmood, H.S., Quraishi, Z., Hoogmoed, W.B., Mouazen, A.M., van Henten, E.J., 2012. Sensing soil properties in the laboratory, in situ, and on-line: a review. In Donald Sparks, edittors: *Advances in Agronomy*, 114, AGRON, UK: Academic Press, 155-224.

Li, S., Cao, J., Hu, S., 2015. Analyzing hydrocarbon fractions in crude oils by two-dimensional gas chromatography/time-of-flight mass spectrometry under reversed-phase column system. *Fuel,* 158, 191–199.

Lorenzi, D., Cave, M., Dean, J.R., 2010. An investigation into the occurrence and distribution of polycyclic aromatic hydrocarbons in two siil size fractions at a former industrial site in NE England, UK using in situ PFE-GC-MS. *Environmental Geochemistry and Health*, 32,553-565.

Malley, D.F., Hunter, K.N., Webster, G.R.B., Malley, D.F., Hunter, K.N., Webster, G.R.B., Barrie, G.R., 1999. Analysis of Diesel Fuel Contamination in Soils by Near-Infrared Reflectance Spectrometry and Solid Phase Microextraction-Gas Chromatography. *Soil Sediment Contam.*, 8, 481–489.

Marriot, P.J., Chin, S,T., Maikhunthod, B., Schmarr, H.G., Bieri, S., 2012. Multidimensional gas chromatography. *Trends in Analytical Chemistry*, 34, 1-21.

McCarty, G.W., Reeves, J.B., 2006. Comparison of near infrared and mid infrared diffuse reflectance spectroscopy for field-scale measurement of soil fertility parameters. *Soil Sci.*, 171, 94–102.

McCarty, G.W., Reeves, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Sci. Soc. Am. J.*, 66, 640–646.

McMahon, G., 2007. Portable instruments in the laboratory, Analytical Instrumentation: A Guide to Laboratory, Portable and Miniaturized Instruments. John Wiley & Sons, Ltd., Chichester, UK.

Mouazen, A.M., Steffens, M., Borisover, M., 2016. Reflectance and fluorescence spectroscopy in soil science-Current and future research and developments. *Soil Tillage Res.*, 155, 448-449.

Mouazen, A.M., De Baerdemaeker, J., Ramon, H., 2006. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil Tillage Res*., 80, 171–183.

Napolitano, G.E., Richmong, J.E., Stewart, A.J., 1998. Characterisation of petroleum-contaminated soils by Thin-Layer Chromatography with Flame Ionisation Detection. *J of Soil Contamination*, 7:6. 709-724.

O'Rourke, S.M., Minasny, B., Holden, N.M., McBratney, A.B., 2016. Synergistic Use of Vis-NIR, MIR, and XRF Spectroscopy for the Determination of Soil Geochemistry. *Soil Sci. Soc. Am. J.*, 80:888-899.

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. *Env. Pollut.,* 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. *Environ. Pollut.*, 192, 162–170.

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, Soil Pollut*., 224, 1539.

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.

Osborne, B.G., Fearn, T., Hindle, P.H., 1993. Practical NIR spectroscopy with applications in food and beverage analysis. *Longman scientific and technical*. Addison-Wesley Longman Ltd: Harlow UK.

Pan, D., Wang, J., Chen C., Huang, C., Cai, Q., Yao, S., 2013. Ultrasonic assisted extraction combined with titatnium plate based solid phase extraction for the anlysis of PAHs in soil samples by HPLC-FLD. Talenta, 108:117-122.

Pasquini, C., 2003. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *J. Braz. Chem. Soc*., 14, 198–219.

Pollard, S.J.T., Hrudey, S.E., Rawluk, M., Fuhr, B.J., 2004. Characterisation of weathered hydrocarbon wastes at contaminated sites by GC-simulated distillation and nitrous oxide chemical ionisation GC/MS, with implications for bioremediation. *Journal of Environmental Monitoring,* 6, 713-718.

Poster, D.L., Schantz, M.M., Sander, L.C., Wise, S.A., 2006. Analysis of polycyclic aromatic hydrocarbons (PAHs) in environmental samples: a critical review of gas chromatographic (GC) methods. *Anal. Bioanal. Chem*. 386, 859–881.

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

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.

Sarowha, S.L.S., Sharma, B.K., Sharma, C.D., Bhagat, S.D., 1997. Characterisation of petroleum Heavy Distillates using HPLC and Spectroscopic Methods, *Energy Fuel*., 11, 566.

Schwartz, G., Ben-Dor, E., Eshel, G., 2012. Quantitative analysis of total petroleum hydrocarbons in soils: comparison between reflectance spectroscopy and solvent extraction by 3 certified laboratories. *Appl. Environ. Soil Sci.*, 2012, 1-11.

Song Y.F., Jing X., Fleischmann S., Wilke B-M., 2002. Comparative study of extraction methods for the determination of PAHs from contaminated soils and sediments. *Chemosphere.* 48, 993-1001.

Sorak, D., Herberholz, L., Iwascek, S., Altinpinar, S., Pfeifer, F., Siesler, H.W., 2012. New development s and applications of handheld Raman, mid-infrared, and near-infrared spectrometers. *Appl. Spectrosc. Rev*., 47 (2), 83-115.

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

Tissot, B.P., and Welte, D.H., 1984. Petroleum formation and occurrence . Berlin Heidelberg: Springer-Verla.

Tran, T.C., Logan, G.A., Grosjean, E., Ryan, D., Marriott, P.J., 2010. Use of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry for the characterisation of biodegradation and unresolved complex mixtures in petroleum. *Geochim Cosmochim Acta*, 74:6468-84.

Ulmanu, M., Anger, I., Gament, E., Mihalache, M., Plopeanu, G., Ilie, L., others., 2011. Rapid determination of some heavy metals in soil using an X-ray fluorescence portable instrument. *Res. J. Agric. Sci*., 43, 235–241.

Vallejo, B., Izquierdo, A., Blasco, R., del Campo, P.P., de Castro, M.D.L., 2001. Bioremediation of an area contaminated by a fuel spill. *J. Environ. Monit*. 3, 274–280.

Van Liedekerke, M., Prokop, G., Rbl-Berger, S., Kibblewhite, M., Lowagie, G., 2014. Progress in the management of contaminated sites in Europe, Reference Report by the Joint Research Centre of the Eurpean Commission 72.

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.

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

Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B., 2014. Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. *Geoderma*, 223-225, 88–96.

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

Wang, Z., 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., Soriona-Disla, J.M., Kirk, J., Janik, L.J., Forester, S.T., McLaughlin, M.J., Stewart, R.J., 2016. Rapid prediction of total petroleum hydrocarbons in soil using a handheld mid-infrared instrument. *Talanta* 160, 410-416.

Weindorf, D.C., Bakr, N., Zhu, Y., 2014. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications. *Adv. Agron.*, 128, 1–45.

Weisman, W., 1998. Analysis of petroleum hydrocarbons in environmental media. In: W. Weisman (Editor). Total Petroleum Hydrocarbon Criteria Working Group (TPHCWG) Series. *Amherst Sci. Publ. Amhers*t, MA 1–98.

Whittaker, M., Pollard, S.J.T., Fallick, T.E., 1995. Characterisation of refractory wastes at heavy oil-contaminated sites: A review of conventional and novel analytical methods. *Environ. Technol*., 16, 1009–1033.

Yang, C., Wang, Z.D., Hollebone, B., Brown, C.E., Yang, Z.Y., Landriault, M., 2015. Chapter 5: Chromatographic fingerprinting analysis of crude oil and petroleum products. In: Fingas, M. (Ed.), Handbook of Oil Spill Science and Technology. John Wiley & Sons, Inc, Hoboken, NJ, pp. 95–163

Zhang, Y., Zhu, Y.G., Houot, S., Qiao, M., Nunan, N., Garnier, P., 2011. Remediation of polycyclic aromatic hydrocarbon (PAH) contaminated soil through composting with fresh organic wastes. *Environmental Science and Pollution Research*, DOI 10. 1007/s11356-011-0521-5.

CHAPTER 3 : Evaluation of vis-NIR reflectance spectroscopy sensitivity to weathering for enhanced assessment of oil contaminated soils

Douglas, R. K.[a*], Nawar, S.[a], S, Cipullo., 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

Abstract: This study investigated the sensitivity of visible near-infrared spectroscopy (vis-NIR) to discriminate between fresh and weathered oil contaminated soils. The performance of random forest (RF) and partial least squares regression (PLSR) for the estimation of total petroleum hydrocarbon (TPH) throughout the time was also explored. Soil samples (n=13) with 5 different textures of sandy loam, sandy clay loam, clay loam, sandy clay and clay were collected from 10 different locations across the Cranfield University's Research Farm (UK). A series of soil mesocosms was then set up where each soil sample was spiked with 10 ml of Alaskan crude oil (equivalent to 8450 mg/kg), allowed to equilibrate for 48 h (T2d) and further kept at room temperature (21^{o}C). Soils scanning was carried out before spiking (control TC) and then after 2 days (T2d) and months 4 (T4m), 8 (T8m), 12 (T12m), 16 (T16m), 20 (T20m), 24 (T24m), whereas gas chromatography mass spectroscopy (GC-MS) analysis was performed on T2d, T4m, T12m, T16m, T20m, and T24m. Soil scanning was done simultaneously using an AgroSpec spectrometer (305 to 2200 nm) (tec5 Technology for Spectroscopy, Germany) and Analytical Spectral Device (ASD) spectrometer (350 to 2500 nm) (ASDI, USA) to assess and compare their sensitivity and response against GC-MS data. Principle component analysis (PCA) showed that ASD performed better than tec5 for discriminating weathered versus fresh oil contaminated soil samples. The prediction results proved that RF models outperformed PLSR and resulted in coefficient of determination (R^2) of 0.92, ratio of prediction deviation (RPD) of 3.79, and root mean square error of prediction (RMSEP) of 108.56 mg/kg. Overall, the results demonstrate that vis-NIR is a promising tool for rapid site investigation of weathered oil contamination in soils and for TPH monitoring without the need of collecting soil samples and lengthy hydrocarbon extraction for further quantification analysis.

Keywords: visible near-infrared diffuse reflectance spectroscopy; weathering; hydrocarbon; land management; chemometrics.

3.1 Introduction

Globally petroleum hydrocarbons are used widely but their uses have caused contamination of soil, water and air mainly during oil production activities, storage and distribution of petroleum products and spillage accidents (ATSDR, 1999). Petroleum hydrocarbons are a complex mixture of aliphatic and aromatic hydrocarbon compounds, among which certain compounds can pose a significant risk to human health and or the environment (Cipullo et al., 2018; Wartini et al., 2017). While there have been a great deal of studies that have been carried out on developing and validating analytical framework for characterising and quantifying petroleum hydrocarbons in soil matrices, they often require soil sampling and then rely on lengthy extraction procedure that needs to be carried out in the laboratory (Douglas et al., 2017; Paiga et al., 2012). There is a need for rapid measurement of petroleum hydrocarbons in soil to allow better and swifter site characterisation and increased confidence in prioritising remediation actions. Most importantly, the concept of taking 'the lab to the field' for measuring hydrocarbon contamination in soil without compromising data quality and information needs to be demonstrated (Douglas et al., 2017; Horta et al., 2015). To this end, field-based techniques offer rapid, non-destructive and cost-effective means of defining levels and distribution of petroleum hydrocarbons on-site. They also provide real-time monitoring data useful for initial site assessment and inform future sampling campaign for detailed risk assessment of the contaminated sites. However, one drawback of field-based techniques is that they often fail to determine and quantify the entire range of petroleum hydrocarbons, the aliphatic and aromatic hydrocarbon fractions, in soil (Douglas et al., 2017).

Once petroleum hydrocarbon are discharged to the environment, they undergo physical, chemical and biological processes that further alter their composition, toxicity, availability, and distribution in the environment. Such weathering (degradation) processes include adsorption, volatilization, dissolution, biotransformation, photolysis, oxidation, and hydrolysis (Jiang et al., 2016; Brassington et al., 2007). These processes shift the chemical composition of the hydrocarbons towards recalcitrant, asphaltenic products of increased hydrophobicity (Coulon et al., 2010). Weathered hydrocarbons are highly complex mixture and are known soil contaminants, which in the face of 40 years of petroleum research, are still not adequately understood or appropriately characterise for informing contaminated land risk categorisation (Coulon et al., 2010). Recently, research has been intensified in developing robust analytical technique for the identification of weathered hydrocarbons, which are the main sources of the organic carcinogens or suspected carcinogens that drive quantitative risk assessment (e.g., Benz[a]anthracene, benzo[a]pyrene, chrysene) at oil-contaminated sites (Environment Agency, 2005). Analytical methods including gas chromatography mass spectroscopy (GC-MS), gas chromatography coupled with flame ionization detector (GC-FID), gravimetric analysis, and infrared spectroscopy are available for analysing weathered hydrocarbons; however, the choice of technique may be influenced by the risk assessment being used during the remediation of contaminated land (API, 2001).

Table 3-1: Previous results of visible near-infrared (vis-NIR) technology performance for the analysis of petroleum-contaminated soils at field-scale

Targeted analyte	N	Spectral range (nm)	Modelling method	Statistical parameters	References
TPH	85	350-2500	RF	R^2=0.68, RMSEP=69.64 mg/kg, RPD=1.85	Douglas et al., 2018a
			PLSR	R^2=0.54, RMSEP=75.86 mg/kg, RPD=1.51	
PAH	85	350-2500	RF	R^2=0.71, RMSEP=0.99 mg/kg, RPD=1.99	Douglas et al., 2018b
			PLSR	R^2=0.56, RMSEP=1.12 mg/kg, RPD=1.55	
TPH	108	350-2500	PSR	R^2=0.70, RMSEP=0.75 mg/kg, RPD=1.86	Chakraborty et al., 2015
			RF	R^2=0.61, RMSEP=0.70 mg/kg, RPD=1.64	
			PLSR	R^2=0.73, RMSEP=0.59 mg/kg, RPD=1.96	
TPH	164	350-2500	FD (PSR)	R^2=0.87, RMSEP=0.528 mg/kg, RPD=2.78	Chakraborty et al., 2014
			SNV-DT (PSR)	R^2=0.80, RMSEP=0.66 mg/kg, RPD=2.21	
			FD (RF)	R^2=0.58, RMSEP=0.95 mg/kg, RPD=1.56	
			SNV-DT (RF)	R^2=0.58, RMSEP=0.94 mg/kg, RPD=1.57	
PAH	137	350-2500	PLSR	R^2=0.89, RMSEP=1.16 mg/kg, RPD=3.12	Okparanma et al., 2014
PAH	150	350-2500	PLSR	R^2=0.89, RMSEP=0.20 mg/kg, RPD=2.75	Okparanma et al., 2013b
TPH	205	2000-2500	PLSR	R^2=0.63, RMSEP=5224 mg/kg, RPD=1.5	Forrester et al., 2013
TPH	45	1560-1800	PLSR	R^2=0.94, RMSECV=1590 mg/kg, Bias=0.003	Hauser et al., 2013

Targeted analyte	N	Spectral range (nm)	Modelling method	Statistical parameters	References
TPH	46	350-2500	PLSR	R^2=0.64, RMSEP=0.34 mg/kg, RPD=1.70	Chakraborty et al., 2010
TPH	26	1100-2498	SMLR	R^2=0.71, SEP=770 mg/kg, RPD=1.80	Malley et al., 1999

N=number of samples, TPH=total petroleum hydrocarbon, PAH=polycyclic aromatic hydrocarbon, R^2 = coefficient of determination, RMSEP = root mean square error of prediction, SEP= standard error of prediction, RPD = residual prediction deviation, RF=random forest, SMLR, = stepwise multiple linear regression, PLSR=partial least square regression, PSR=penalized spline regression, FD = first derivative pre-processing, SNV-DT= standard normal variate pre-processing followed by detrending.

Infrared spectroscopy, including visible and near-infrared (vis-NIR) or mid-infrared (MIR) spectroscopy, has been shown to be a suitable rapid acquisition method for the measurement of hydrocarbon concentration in soil without the need of any sample preparation (Douglas et al., 2018a; Horta et al., 2015; Okparanma and Mouazen, 2013a; Chakraborty et al., 2010). Infrared spectroscopy for PHCs detection in soils is a proven technology (Okparanma et al., 2014a; Forester et al., 2013; Okparanma and Mouazen, 2013; Charaborty et al., 2010; Bray et a., 2009; Malley et al., 1999). More details on previous works on the use of vis-NIR spectroscopy for quantifying hydrocarbons in soils can be found in Table 3-1. However, to the best of our knowledge, the application of vis-NIR-based techniques to differentiate between freshly contaminated *versus* weathered crude oil contaminated soils has not been investigated. Furthermore, no attempts to implement the vis-NIR spectroscopy to quantify the total petroleum hydrocarbon (TPH) in soil, across different stages of weathering can be found in the literature.

The objectives of this study were (i) to investigate the sensitivity of two portable vis-NIR spectrophotometers (ASD and tec5) for the discrimination between weathered and fresh oil spill in soils using principal component analysis (PCA), and (ii) to quantify TPH in these soils during weathering, using partial least squares regression (PLSR) and random forest (RF) modeling methods.

3.2 Materials and methods

3.2.1 Study area and soil sampling

A total of thirteen (n=13) surface soil samples (0-15 cm) with approximately 5 kg per sample were collected using a shovel from 10 sites located in Bedfordshire, namely, Avenue, Downings, Orchard, Mound, Wood, Copse, Ivy ground, Near Warden, Showground, and Sandpit; all from the Cranfield University's Research Farm, Bedfordshire, UK (Figure 3-1). Samples were taken with Ziploc bags to the laboratory and stored in the freezer at 4 ^{o}C prior to utilisation. Two and three samples were collected for Avenue and Ivy ground fields, respectively, while one sample was collected from each of the remaining five fields. The collected soil samples were subjected to soil physical and chemical analyses. The soil moisture content (MC) was measured by oven-drying soil samples at 105 ± 5^{o}C for 24 h. 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.

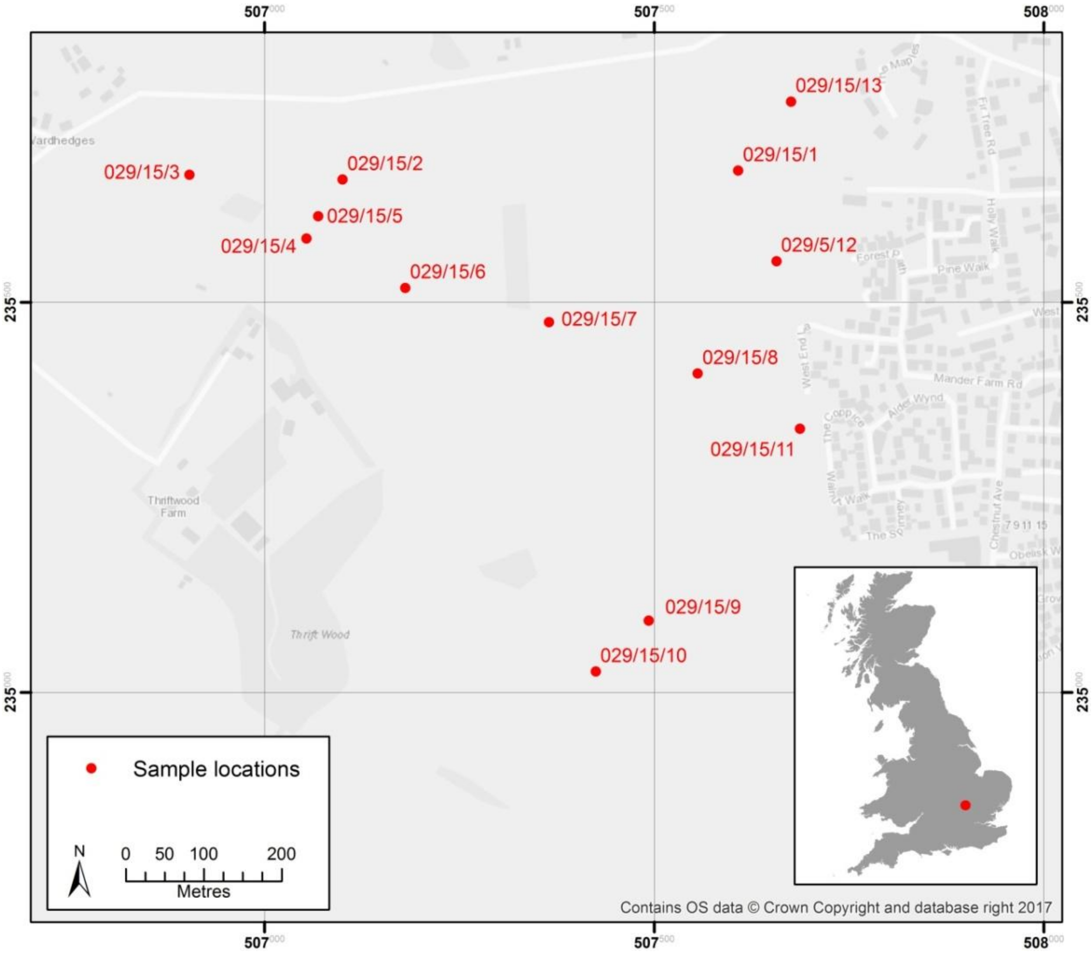


Figure 3-1: Location of the study area and sampling points collected from 10 sites in Bedfordshire, UK.

3.2.2 Mesocosms setup

Using 1 kg soil, 13 soil mesocosms (representing all the 13 samples) were set up. Each soil sample was spiked with 10 ml of Alaskan crude oil (equivalent to 1845 mg/kg) and allowed to equilibrate at room temperature (21 ^{o}C) for 48 h. Alaskan crude oil was the crude oil available at Cranfield and relatively close to Nigerian oil composition – so for practicality the Alaskan crude oil was used. Also it is difficult to obtain crude oil samples from Nigeria due to security issues. Vis-NIR scanning was performed on pristine soil (control (TC) - pristine samples dried at room temperature to reduce moisture effect) and then after 2 days (T2d) and months 4 (T4m), 8 (T8m), 12 (T12m), 16 (T16m), 20 (T20m), 24 (T24m); whereas gas chromatography mass spectroscopy (GC-MS) analysis was performed on T2d, T4m, T12m, T16m, T20m, and T24m. Therefore, data of T8m was excluded from the quantitative analysis of TPH. The experiment set up with UK samples was to mimic environmental conditions as hydrocarbons are weathered/aged in most of the spill sites in the Niger Delta, Nigeria where soil samples were collected. The idea was to take portable versions to site to acquire on-site data in the future.

3.2.3 Optical measurement and spectra pre-processing

Soil spectral measurements were done in the laboratory using two vis-NIR spectrophotometers, namely, an AgroSpec vis-NIR spectrometer with a spectral range of 305-2200 nm (tec5 Technology for Spectroscopy, Germany) and an ASD LabSpec2500® (Analytical Spectral Devices, Inc., USA), which covers a spectral range of 350–2500 nm. Both spectrometers are portable, but use different detectors; ASD uses monochromatic detector while tec5 is equipped with a diode array detector.

Spectral measurement by ASD LabSpec2500® spectrometer in this study followed the protocols described by Douglas et al. (2018a). Before scanning, samples were air-dried in order to eliminate the effect of moisture content on soil spectral analysis (Mouazen et al., 2006). After removal of all plants and pebble materials, three subsamples were prepared from each soil sample; these were placed into 3 different Petri dishes (1 cm height x 5.6 cm in diameter), and the surface was smoothened gently with a spatula before scanning (Mouazen et al., 2005). This was done to achieve optimal diffuse reflection and, thus, a good signal-to-noise ratio. A high-intensity probe was used for scanning of soil samples, which has a built-in light source made of a quartz-halogen bulb of 2727 °K. The light source and detection fibres are assembled in the high-intensity probe enclosing a 35° angle. The device was calibrated using almost 100 % white Spectralon disc before use, and after every 30 min. The spectral measurements were made in the dark in order to both, control the illumination conditions and reduce the effects of stray light. The three replicates of each sample were scanned at three different spots, and an average spectrum was obtained for further analysis. A total of 10 scans were acquired from each replicate, and the average spectrum of the three replicates was considered as the sample spectrum.

Prior to multivariate analysis, three standardised spectral pre-treating approaches (including maximum normalization, first derivative, and smoothing) were carried out using R software (R Core Team, 2013). Maximum normalisation divides each row (spectrum) by its maximum absolute value to achieve an even distribution of the variances; the first derivative removes the baseline shift to improve the accuracy of quantification (Okparanma et al. 2014; Demetriades-Shah et al., 1990); and smoothing reduces the impact of noise (Okparanma and Mouazen, 2013b). These routines were aimed at keeping all useful chemical and physical information in the spectra for analysis.

3.2.4 Gas chromatography and peak integration

Chemical analysis for TPH concentration was carried out using sequential ultrasonic solvent extraction-gas chromatography (SUSE-GC) as 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) and 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 polycyclic aromatic hydrocarbons (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 spectrometry 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 analyzed 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. Relative standard deviation (RSD) values for all the soils was <10%. From the results obtained for alkanes and PAHs, TPH was obtained for each sample, and further used for modelling purposes.

3.3 Multivariate analyses

3.3.1 Principal component analysis (PCA)

PCA was used for qualitative vis-NIR discrimination of soil samples based on the spectral properties of the different contaminated weathering groups. PCA is a multivariate technique that reduces the dimensionality of large multivariate datasets. PCA helps to transform the wavelengths (independent variables) into principle components (PCs). Plotting the PCs enables one to examine interrelationships among different variables, and detect and interpret sample patterns, groupings, similarities, or differences (Mouazen et al., 2006; Martens and Naes, 1989). The pre-processed spectra have been used in the PCA; the results showed a similarity map of principal PCs, as well as the loadings that can be used to investigate the significant wavebands for hydrocarbons. The PCA was performed using FactorMine R-package (R Core Team, 2013).

3.3.2 Quantitative assessment of TPH using PLSR and RF methods

The pre-processed vis-NIR soil spectra for both ASD and tec5 spectrophotometers coupled with the reference laboratory TPH measured by SUSE-GC were used to develop calibration models for quantifying TPH through 2 years weathering period. The total number of samples used for both PLSR and RF modelling were 78, obtained from 13 soil samples scanned at six occasions through 24 months. Sixty (n=60) samples were selected for calibration while eighteen (n=18) for prediction (validation). The same calibration and validation datasets used in PLSR were utilized for RF analysis. The selection of the samples in the calibration and prediction set was done based on the Kennard-Stone algorithm (Kennard and Stone, 1969). Kennard-Stone was used because it has been widely used in chemometrics and soil spectroscopy and showed good performance in separating samples into calibration (cross-validation) and independent validation (prediction) data sets (Viscarra Rossel and Webster, 2012; de Groot et al., 1999). Two groups of calibration models for TPH were developed, one for tec5 and the second one for ASD spectral data. The intension was to evaluate the effect of the spectral range of the prediction accuracy of TPH in the soil during 2 years weathering period.

PLSR is a commonly used multivariate regression technique available in standard statistical and chemometrics software. It is a combination of both the independent variables (TPH values) and the dependent variables (wavelengths), which are used as regression generators for the independent variables. In this study, we use PLSR with leave-one-out cross validation (LOOCV) to develop TPH prediction model, using pls package (R Core Team, 2013). It is documented that LOOCV annul the possible effect of model under- or over-fittings (Efron and Tibshirani, 1993).

Random forest is a nonparametric and nonlinear classification and regression algorithm using assembly learning strategy that integrates hundreds of individual trees (Breiman, 2001). A bootstrap sample is first drawn from the training dataset to build each tree. At each node split, the candidate set of the regressor is a random subset of all the regressors. The final prediction of a new observation is the average of the predicted values from all the trees in the forest. The tuning parameters of RF have been defined based on function implemented in the R software package and were set to 500, 2, and 2 for the number of trees (*ntree),* the number of predictor variables used to split the nodes at each partitioning (*mtry*), and the minimum size of the leaf (*nodesize*), respectively. Models were developed with R program using the software package randomForest Version 4.6-12 (Liaw and Wiener, 2015), based on Breiman and Cutler's Fortran code (Breiman, 2001).

$$\mathrm{RMSEP} = \frac{\sqrt{\sum_{i=1}^{Np}(\hat{y} \ - y_i)^2}}{\mathrm{Np}}$$

Equation 3-1: Root mean square error of prediction

$$\mathrm{RPD} = \frac{\mathrm{SD}}{\mathrm{RMSE}}$$

Equation 3-2: Residual prediction deviation

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

Equation 3-3: Ratio of performance to interquartile range

Where $\hat{y} = predicted\ values, y_i = measured\ values,$ SD=standard deviation of the measured reference values, N=number of samples in the set, RPIQ=ratio of performance to interquartile range, IQ=difference between the third and the first quartiles (IQ=Q_3-Q_1)

3.4 Evaluation of model performance

The performance of TPH prediction models was assessed by means of three parameters: (i) the coefficient of determination in prediction R^2, (ii) root mean square error of prediction (RMSEP), and (iii) residual prediction deviation (RPD), which is a ratio of standard deviation (SD) to RMSEP. In this study, we adopted the model classification criterion of Viscarra Rossel et al. (2006): 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 best model performance would have the highest values of R^2 and RPD, and smallest value of RMSEP.

3.5 Results and discussion

3.5.1 Soil physiochemical properties

Soil physio-chemical properties (*viz*. partial size distribution, TOC, and MC) of the different soil samples are presented in Table 2. Clay content ranged between 14% and 57%, silt between 15% and 27%, and sand between 16% and 63%. However, examining the soil texture type according to the United State Department of Agriculture (USDA) classification system, indicates the majority of soils in the study fields are on the heavy side of the texture triangle. TOC was high with minimum and maximum of 1.62 and 4.48%, respectively. Results indicated a high variation in soil texture and TOC among the soil samples. Tamburini et al (2017) demonstrated that the general effects of physical soil characteristics do not generate dramatic interferences with spectral signals. Despite a slight worsening of the prediction capacity, the possibility to gather all samples and build a unique calibration model has permitted to encompass the two principal sources of spectral offsets and shifts in their calibration model, increasing its robustness and reliability with unknown samples. Future improvements of this application could permit performing NIR analysis of soils directly in field by potentially using a probe connected to the NIR instrument (Tamburini et al., 2017).

Studies have reported on the effect of soil factors potentially soil moisture content on vis-NIR. For example, moisture content affects vis-NIR measurement (Malley et al., 1999). A study by Horta et al. (2015) concluded that effect of moisture content on vis-NIR happens to cause more attenuation than soil structure. However, since soil samples were scanned after air drying, the effect of MC was excluded from spectral analysis. It has been reported that small particle size (high clay content) can result in a better model performance (Fontán et al., 2010) of soil organic carbon, whereas prediction was reported to be less accurate in coarse soil textures (Stenberg, 2010). Since the majority of soil textures of the samples analysed in this work were on the heavy side of the texture triangle, the similarity in texture is assumed to have minor effect on prediction accuracy of TPH.

Table 3-2: *S*oil physio-chemical properties of 13 surface soil samples (0-15 cm) collected from ten different locations across the Cranfield University's Research Farm, Bedfordshire, UK.

Location name	Sample No.	Clay %	Silt %	Sand %	TOC %	Texture
Avenue	1	17	20	63	2.02	Sandy loam
	2	30	19	51	1.67	
Downings	3	28	19	53	2.3	Sandy clay loam
Orchard	4	33	26	41	2.32	Clay loam
Mound	5	16	21	63	1.96	Sandy loam
Wood	6	42	25	33	2.28	Clay
Copse	7	38	26	36	2.7	Clay loam
	8	57	27	16	4.48	Clay
Ivy ground	9	57	27	16	4.48	Clay
	11	57	27	16	4.48	Clay
Near warden	10	57	25	18	3.1	Clay
Showground	12	24	17	59	1.87	Sandy clay loam
Sand pit	13	14	15	71	1.62	Sandy loam

TOC=total organic carbon.

3.5.2 Spectral data analysis

Illustrative raw air dry soil spectra and pre-processed soil spectra changes overtime are presented in Figure 3-2 (note that only T2d, T12m and T20m are shown for clarity). In both Figure 3-2a and Figure 3-2c, the control soil (TC) reflects higher than the contaminated soils or, in other words, absorb less light energy due to the lighter colour of samples without oil added.

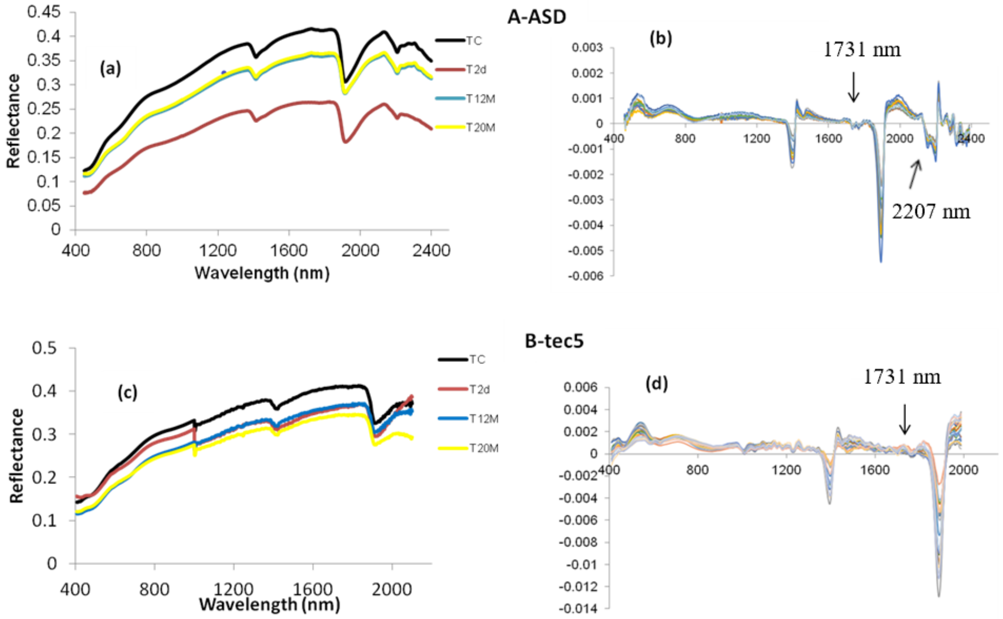


Figure 3-2: Illustrative example of visible and near infrared (vis-NIR) soil spectra overtime: Control pristine soil (TC), and contaminated soil after 48 hours (T2d), 12 months (T12m) and 20 months (T20m); Panels a & b showed raw spectra and pre-processed spectra obtained with ASD spectrometer; Panels c & d showed raw spectra and pre-processed spectra obtained with the tec5 spectrophotometer.

It is clearly demonstrated that reflectance decreased or absorption increased when adding crude oil, due to the darker color. Among the contaminated soils, the spectral reflectance increased (i.e., less absorbance) as weathering of hydrocarbons in soils progresses. Thus, T2d samples had the highest absorbance, and this decreased with weathering time. In terms of equipment performance, a better discrimination between groups' average spectra was achieved with the ASD spectrometer compared to tec5 spectrometer (Figure 3-2).

The behaviour of control and contaminated spectra observed herein is in line with the conclusions drawn by Hoerig et al. (2001). Both ASD and tec5 spectrophotometers captured hydrocarbon features around 1731 nm in the first overtone region (Figure 3-2b and Figure 3-2d), which is linked with TPH. Our result is not far from those identified by other scientists e.g., 1732, 1758 nm (Douglas et al. 2018a), 1752 nm (Chakraborty et al., 2015), 1712, 1752 nm (Okparanma et al., 2014a). An absorption band of hydrocarbons around 2207 nm in the combination region (Figure 3-2) was also observed in the ASD spectra, a wavelength that is close to those reported by other researchers e.g., 2240 nm by Chakraborty et al. (2015), and 2460 nm by Forrester et al. (2013). The other absorption bands are associated with other soil properties, e.g., water, clay mineralogy, and organic carbon. More details about the hydrocarbon signatures in soils are presented (section 3.2, Figure 3-4).

3.5.3 Qualitative discrimination of weathering groups by PCA

In order to examine the variability between spectra of the contaminated soils overtime, spectra were subjected to PCA, with the aim to extract distinctive spectral features that can assemble similar weathered contaminated soils together in distinguished groups. If this can be achieved, we can claim that the vis-NIR spectrometers used in this study can differentiate weathered *versus* fresh oil spill in soils. A scatter diagram of component score for the first and second principal components (PC-1, PC-2) is shown in Figure 3-3a for the ASD spectrometer and Figure 3-3b for the tec5 spectrometer. With the ASD spectrometer, PC1 accounted for 94.50% while PC 2 accounted for 5.10% of variance, with a total of 99.6%. However, a slightly less variance was accounted for by the PCA performed on the tec5 spectra (Figure 3-3b), with PC1 accounting for 93.30% and PC2 accounting for 5.12%, which sums up to 98.42% of the total variance. It is noteworthy that the separation patterns of the various weathering group soils achieved with the two portable vis-NIR instruments are different; with ASD (Figure 3-3a) providing the best visual separation in the principal component space. The separation was particularly clear between the non-contaminated (TC) and freshly contaminated samples at T2d, obtained with the ASD spectrometer. Different weathering groups were formed along the PC1 of the ASD-PCA plot, showing different degree of overlap between soil groups of different weathering time, where overlap becomes more evident after month 12 and up to month 24 in Figure 3-3a. Soil samples at T2d and T4m are better separated from the remaining weathering groups (Figure 3-3a). Few samples from T4m overlapped with those of T2d, whereas one T4m and few T8m samples were in the neighbourhood of the T12m and T24m samples. In the case of the T2d and T4m samples, there is less compositional resemblance reflected on different spectral signature, whereas more compositional resemblance exists within the T12m to T24m samples, resulting in smaller spectral differences of the same sample throughout weathering time, and hence the increase of sample overlap. The tec5-PCA plot shows less clear separation between different weathering groups (Figure 3-3b) compared to the ASD-PCA plots. Separation here occurs along the diagonal access between PC1 and PC2 (Figure 3-3b). It is obvious that TC samples are clearly separated from the other groups, and that more clear overlap exists between the remaining groups compared to the ASD-PCA plots. For example, it is odd to observe that T4m samples are closer to TC samples, in comparison with T2d samples, which were further away from TC samples. Furthermore, samples of T24m and T20m are closer to TC samples than the remaining groups with smaller weathering time (e.g., T4m, T8m, T12m and T16m).

Overall, we can conclude that, the ASD spectrometer provided logical and clearer separation of the different weathering groups and that instrument's sensitivity to weathering reduces overtime due to the reduction of the TPH concentration (see discussion below). On the other hand, the clear separation observed between the contaminated and TC samples indicate that the two groups are compositionally dissimilar. This is in agreement with the results reported by Chakraborty et al. (2010), who assessed the ability of vis-NIR spectroscopy to distinguish contaminated and non-contaminated soils qualitatively using PCA.

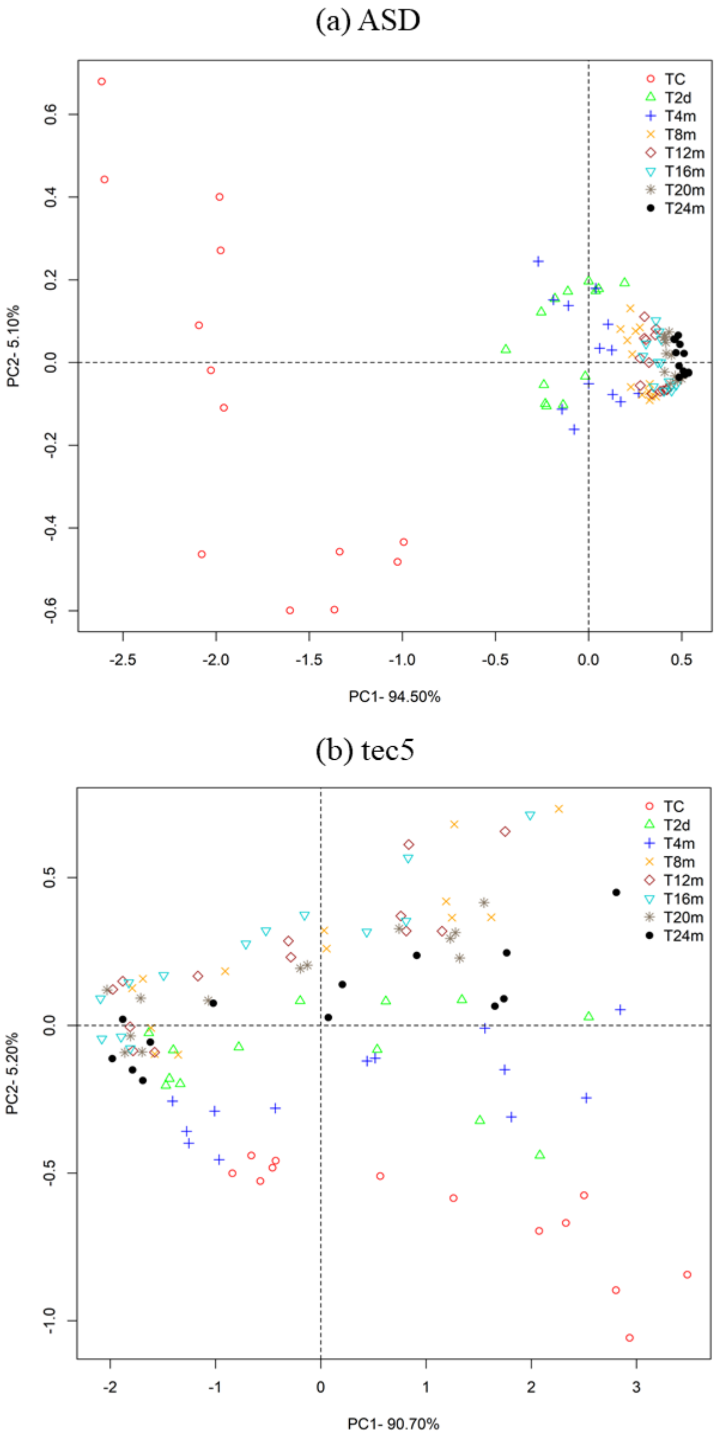


Figure 3-3: **Principal component analysis of the soil scanning profile overtime obtained using** (a) **ASD and** (b) **tec5 spectrophotometers (TC: control samples (pristine); and contaminated soil samples after 48 h (T2d), and months 4 (T4m), 8 (T8m), 12 (T12m), 16 (T16m), 20 (T20m) and 24 (T24m).**

Furthermore, PCA loadings were produced to investigate potential wavelengths associated with diesel originated hydrocarbon contamination (Figure 3-4). In the PCA loadings, an absorption minimum was observed at 1730 nm in both ASD and tec5 spectrometers, which is attributed to C-H stretching modes of terminal CH_3 and saturated CH_2 groups linked to TPH in the first overtone region. This result is in line with observations from others researchers (Okparanma et al., 2014; Workman and Weyer, 2008). Furthermore, typical spectral signatures around 1452 nm and 1950 nm were clearly observed in both ASD and tec5 spectrometers. These are associated with the second and first overtones of water absorption around 1450 nm and 1950 nm, previously reported (Mouazen et al., 2005; Mouazen et al., 2006). Absorption features around 2279 and 2340 nm were also observed in ASD spectrometer alone. These are associated with metal-OH bend and O-H stretch combination and characteristic of clay minerals. The results obtained here are similar to those at 2200 and 2300 nm, reported in the literature (Viscarra Rossel et al., 2006b; Clark et al., 1990). The absorption band at 2207 nm can be attributed to either amides (C=O) absorption (Viscarra Rossel and Behrens, 2010) or crude oil spectral signatures (stretch+bend) (Mullins et al., 1992). Furthermore, this band can be linked to the hydrocarbon concentration that can be effective to discriminate between weathering groups (Figure 3-4a). Therefore, the ASD showed a high capability to discriminate between the weathering group, and this is because its full vis-NIR range spectrum including all the effective waveband associated with hydrocarbons.

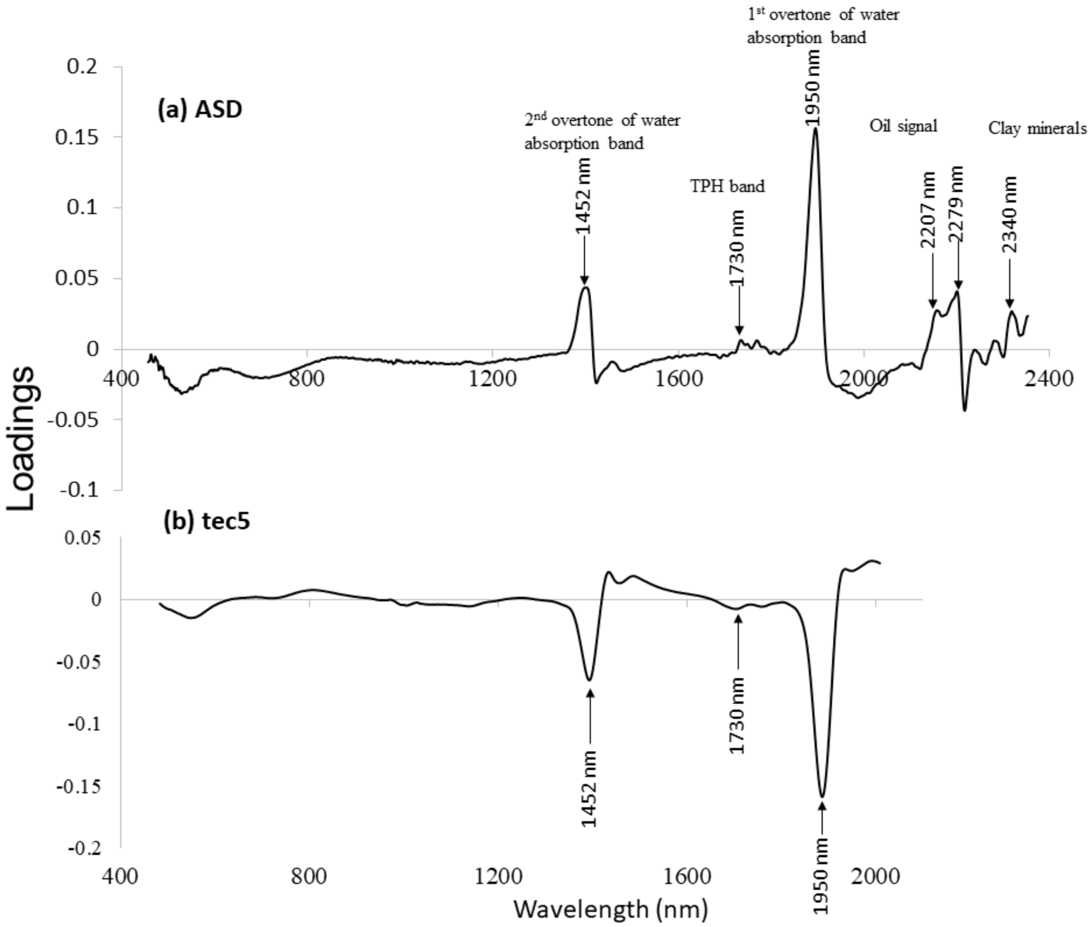


Figure 3-4: Principal Component analysis loadings of the spectral patterns showing the wavelengths associated with hydrocarbon fractions, water and mineralogy.

3.5.4 Soil TPH analysis

The petroleum hydrocarbon profiles and change overtime are illustrated in Figure 3-5. Chromatogram showed a well-developed series of n-alkanes distribution with carbon band range C_{10} – C_{36}, but with about 85 % of the mixture existing within the range C_{12} - C_{28} (Figure 3-5; T2d). The distribution confirms that the hydrocarbon source is weathered (degraded) over time. After month 16 and 24, the most prominent residual hydrocarbon fractions were the aliphatic fractions C_{16}-C_{35} and C_{35}-C_{40}, and the aromatic fractions C_{12}-C_{16} and C_{16}-C_{21}, respectively.

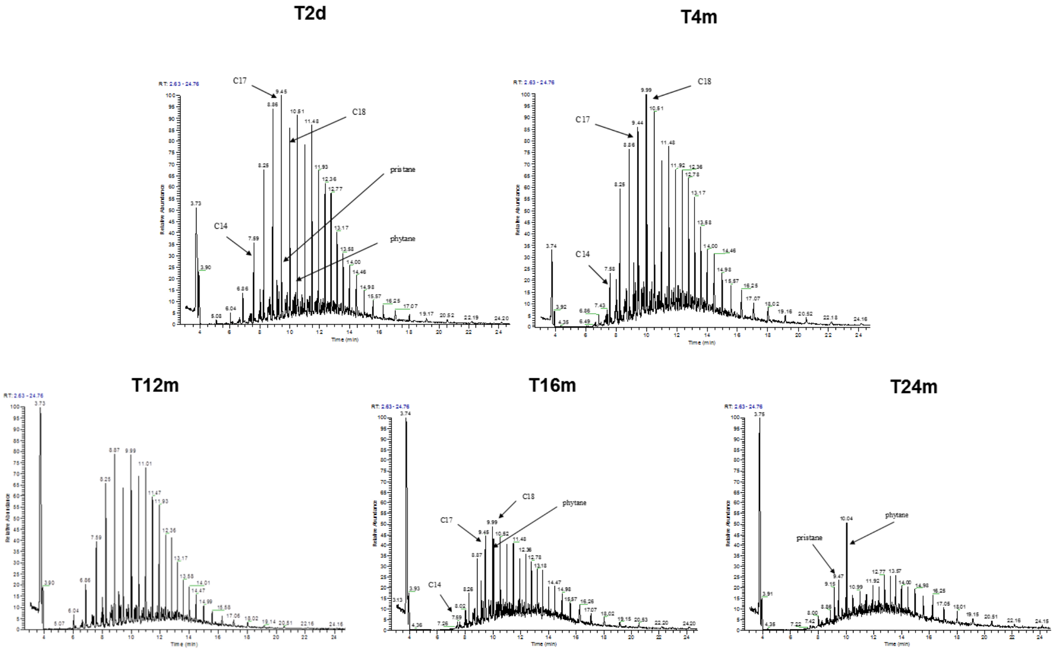


Figure 3-5: Illustrative gas chromatography-mass spectrometry (GC-MS) chromatogram showing petroleum hydrocarbons fingerprint change overtime. Results are shown for contaminated soil samples after 48 h (T2d), after months 4 (T4m), 12 (T12m), 16 (T16m), and 24 (T24m).

Summary statistics of the aliphatic and aromatic fractions as well as the TPH concentrations which equal to sum of aliphatic and aromatic fractions are provided in Table 3-3. These TPH values were used for the vis-NIR spectra modelling. Samples were divided into calibration and prediction sets. In the calibration set, the minimum and maximum concentrations of TPH were 187.5 and 1761.5 mg/kg, respectively. The minimum and maximum concentrations of TPH in the prediction set were 186.7 and

1362.4 mg/kg, respectively (Table 3-4). The largest reduction in both the aliphatic and aromatic fractions were obtained after month 16 where 50% and 38% of the aliphatic and aromatic fractions, respectively, were degraded. Further to this, TPH reduction reached 72% by month 20 and 85% by month 24.

Table 3-3: Descriptive statistics of aliphatic and aromatic fraction concentrations (mg/kg) in 13 soil samples overtime (n = 78). Results are shown for diesel contaminated soil samples after 48 h (T2d), and months 4 (T4m), 12 (T12m), 16 (T16m), 20 (T20m) and 24 (T24m).

Hydrocarbon fractions		T2d			T4m			T12m		
		Med	Min	Max	Med	Min	Max	Med	Min	Max
Aliphatic	nC10-nC12	48.24	1.55	121.02	40.90	0.87	119.20	32.51	0.44	82.67
	nC12-nC16	63.20	32.97	99.04	50.54	4.19	105.84	37.56	4.54	81.10
	nC16-nC35	34.09	0.14	161.69	18.64	0.15	107.60	25.95	0.21	75.43
	nC35-nC40	0.56	0.02	4.18	0.78	0.04	4.68	0.73	0.01	13.57
	Total	1259	1113	1642	887.11	813.70	1214.75	880	721	1055
Aromatic	nC10-nC12	nd	nd	Nd	nd	nd	nd	nd	nd	nd
	nC12-nC16	3.32	3.25	3.96	3.33	3.26	4.29	3.36	3.10	3.53
	nC16-nC21	3.97	3.26	30.11	4.06	3.21	12.20	3.64	3.10	14.03
	nC21-nC35	6.49	3.50	15.24	5.72	3.36	16.51	3.92	3.10	16.63

	Total	82.68	71.59	134.37	81.71	64.77	100.47	56.09	43.10	94.93
	TPH	1343.28	1190.78	1716.49	963.83	884.15	1315.2	959.9	802.45	1101.3

		T16m			T20m			T24m		
		Med	Min	Max	Med	Min	Max	Med	Min	Max
Aliphatic	nC10-nC12	21.65	0.64	56.75	6.38	0.14	23.50	2.43	1.01	6.99
	nC12-nC16	29.05	13.82	50.42	13.83	1.80	31.91	7.48	1.05	18.81
	nC16-nC35	19.96	0.16	75.91	10.80	0.03	44.85	3.28	0.01	20.88
	nC35-nC40	0.25	0.01	2.19	0.24	0.01	2.35	0.02	0.01	0.76
	Total	678	628	774	326	233	421	162	133	185
Aromatic	nC10-nC12	nd	nd	Nd	nd	nd	nd	nd	nd	nd
	nC12-nC16	3.24	2.92	3.55	3.30	2.22	3.42	2.37	2.29	3.41
	nC16-nC21	4.29	2.05	7.64	3.77	3.31	6.29	3.32	3.10	4.40
	nC21-nC35	4.47	2.94	7.46	3.54	3.10	5.27	3.34	3.10	6.90

Total	59.12	53.07	62.83	47.35	43.41	62.50	46.20	43.38	50.76
TPH	733.87	687.62	833.98	380.53	279.68	465.40	207.8	178.47	232.64

nd= not detected, med=median, min=minimum, max=maximum.

Table 3-4: Statistical summary of total petroleum hydrocarbons (TPH) concentrations of the collected soil samples measured with gas chromatography-mass spectrometry (GC-MS) for the different weathering stages in cross-validation and independent validation.

	N	Minimum	Mean	Median	1st Qu.	3rd Qu.	Maximum	St. dev
TPH (mg/kg)								
Cross-validation	60	187.5	773.70	789.20	383.60	990.10	1761.50	133.13
Independent validation	18	186.7	800.40	838.20	372.50	1121.4	1362.40	40.20

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

3.5.5 Models performance for estimating TPH

Table 3-5 and Figure 3-6, and Figure 3-7 summaries the cross-validation and prediction results of TPH based on PLSR and RF analyses obtained with both the ASD and tec5 spectrophotometers. Generally, the RF models outperformed the PLSR in cross-validation and prediction for both ASD and tec5 measurements. The results of prediction based on ASD spectra indicated that RF model resulted in R^2 of 0.92, RMSEP of 108.56 mg/kg, RPD of 3.79, and RPIQ of 6.90, which outperformed PLSR model (R^2 = 0.83, RMSEP = 164.87 mg/kg, RPD = 2.49, RPIQ = 4.54). This was also the case for tec5 spectra as the RF model (R^2 = 0.22, RMSEP = 352.71 mg/kg, RPD = 1.16, and RPIQ = 2.13) outperformed PLSR (R^2 = 0.11, RMSEP = 422.50 mg/kg, RPD = 0.97, and RPIQ = 1.77). The current results for both PLSR and RF prediction are better than those reported by Douglas et al. (2018a, 2018b) using 85 genuine contaminated soil samples collected from the Niger Delta region of Nigeria. Furthermore, our results for RF prediction are better than those reported by Chakraborty et al. (2015) using 108 contaminated soil samples (West Texas, USA) with i) RF modelling method only (R^2 = 0.61, RMSE = 0.70 mg/kg, RPD = 1.64 and RPIQ = 0.57), and ii) RF combined with penalised spline regression (PSR) RF+PSR (R^2 = 0.78, RMSE = 0.53 mg/kg, RPD = 2.19 and RPIQ = 0.75). Also, the PLSR prediction in the current study are better than the results reported by Chakraborty et al. (2015 and 2010), who achieved RPD values of 1.96 and 1.7, respectively, for field-moist soils (Table 3-1). A possible reason for the observed difference in the present study may be attributed to the combination of spectral pre-processing (maximum normalization, 1st derivative and smoothing) that represents a vital step in multivariate calibration and improves the model performance (Nawar et al. 2016; Buddenbaum and Steffens, 2012; Mouazen et al., 2010). According to Viscarra et al. (2006) model classification for RPD, excellent and very good predictions for TPH were achieved with RF-ASD (RPD = 3.79) and PLSR-ASD (2.49), respectively, whereas using tec5, poor and very poor results were obtained with RF-tec5 (RPD = 1.16) and PLSR-tec5 (RPD = 0.97), respectively.

Table 3-5: Summary results of partial least squares regression (PLSR) and random forest (RF) models in calibration (cross-validation) and prediction (independent validation) for total petroleum hydrocarbons (TPH) prediction in oil-contaminated soil samples using ASD and tec5 spectrophotometers.

		PLSR					RF				
Instrument		R^2	RMSEP (mg/kg)	RPD	RPIQ	LV	R^2	RMSEP (mg/kg)	RPD	RPIQ	*ntrees*
ASD	Calibration (n=60)	0.92	113.42	3.60	5.34	6	0.98	44.07	9.28	13.76	500
	Prediction (n=18)	0.83	164.87	2.49	4.54	4	0.92	108.56	3.79	6.90	200
tec5	Calibration (n=60)	0.83	164.26	2.47	3.70	8	0.92	111.65	3.63	5.45	500
	Prediction (n=18)	0.11	422.50	0.97	1.77	8	0.22	352.71	1.16	2.13	200

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 scatter plots of GC-MS measured *versus* ASD and tec5 predicted TPH concentrations (based on PLSR and RF models) are shown in Figure 3-6 and Figure 3-7, respectively. Both the ASD and tec5 instruments quantitatively discriminated the soils at their various stages of weathering; however, a better discrimination was achieved with the ASD instrument. The results herein support the qualitative separation of the various soil groups by PC score plots illustrated in Figure 3-3.

The TPH wavelength regression coefficients plots shown in Figure 3-8 illustrate important wavebands around 1452, 1730, and 1950 nm for both ASD and tec5 spectrometers. The 1730 nm wavelength is attributed TPH absorption in the first overtone, which is close to the previous findings (Douglas et al., 2018a; Okparanma et al., 2014; Workman and Weyer, 2008; Osborne et al., 2007). The significant spectral signals around 1452 and 1950 nm are associated with water absorption bands in the second and first overtones, respectively, which accord findings reported in previous studies (Douglas et al., 2018a; Mouazen et al., 2007). In the ASD spectra, the spectral signature at 2207 nm may be due to the effect of hydrocarbon in the combination region around 2220 nm (Chakraborty et al., 2015 Forrester et al., 2013). Interestingly, the absorption feature around 2279 nm and 2340 nm is the same with the one observed in the PCA loadings (Figure 3-4a). This is characteristic of clay minerals around 2300 nm (Clark et al., 1990). The low performance of tec5 in separating the different weathering groups (Figure 3-4b). The quantitative assessment of TPH may be attributed to the high latent variables (LV), compared to that of ASD (Table 3-5).

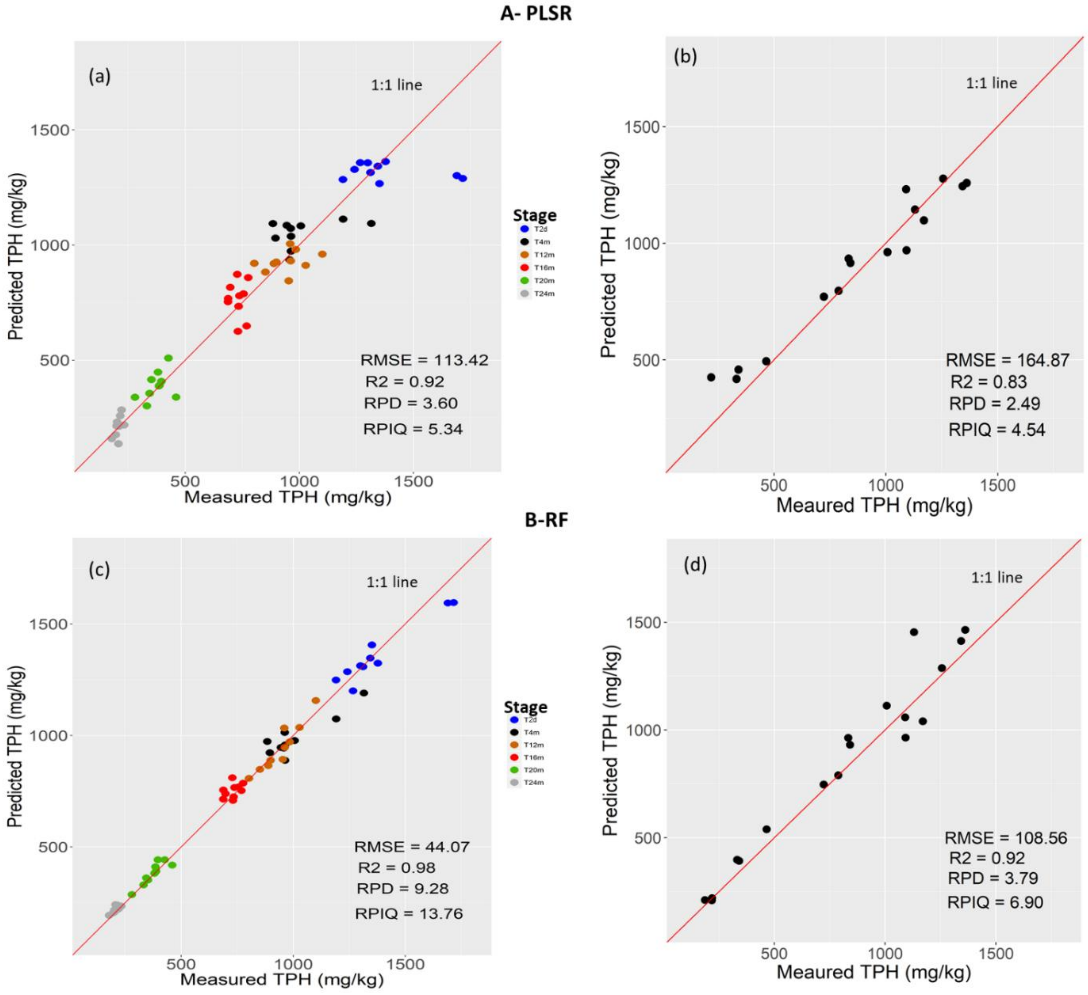


Figure 3-6: Scatter plots of measured total petroleum hydrocarbons (TPH) using gas chromatography-mass-spectrometry (GC-MS) *versus* visible and near infrared (vis-NIR) ASD spectrometer predicted concentrations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and (B) random forest (RF) in (c) cross-validation and (d) prediction. Results show clear separation of diesel contaminated groups of different weathering stages of 48 h (T2d), and months 4 (T4m), 12 (T12m), 16 (T16m), 20 (T20m) and 24 (T24m).

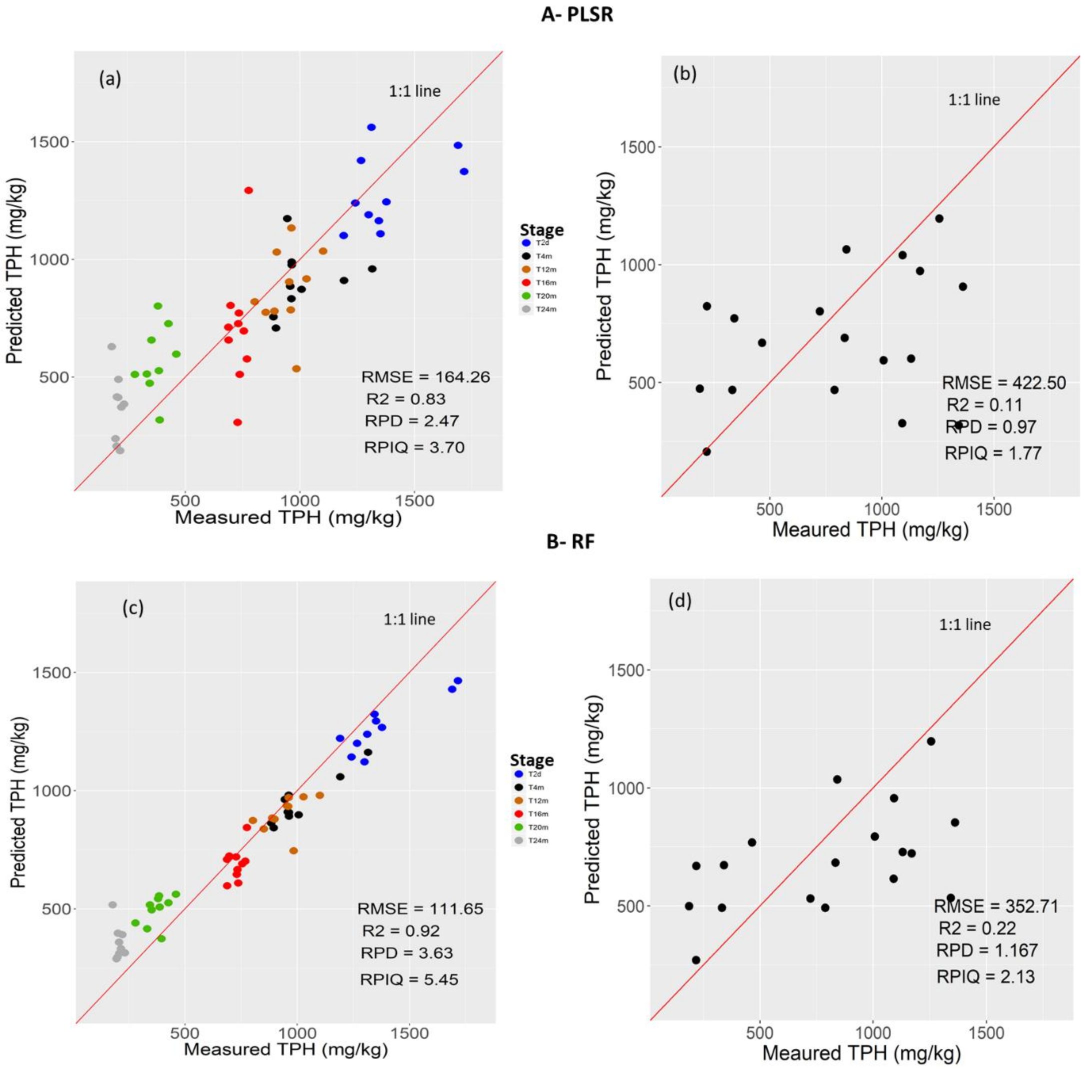


Figure 3-7: Scatter plots of measured total petroleum hydrocarbons (TPH) using gas chromatography-mass-spectrometry (GC-MS) *versus* visible and near infrared (vis-NIR) tec5 spectrometer predicted concentrations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and (B) random forest (RF) in (c) cross-validation and (d) prediction. Results show clear separation of diesel contaminated groups of different weathering stages of 48 h (T2d), and months 4 (T4m), 12 (T12m), 16 (T16m), 20 (T20m) and 24 (T24m).

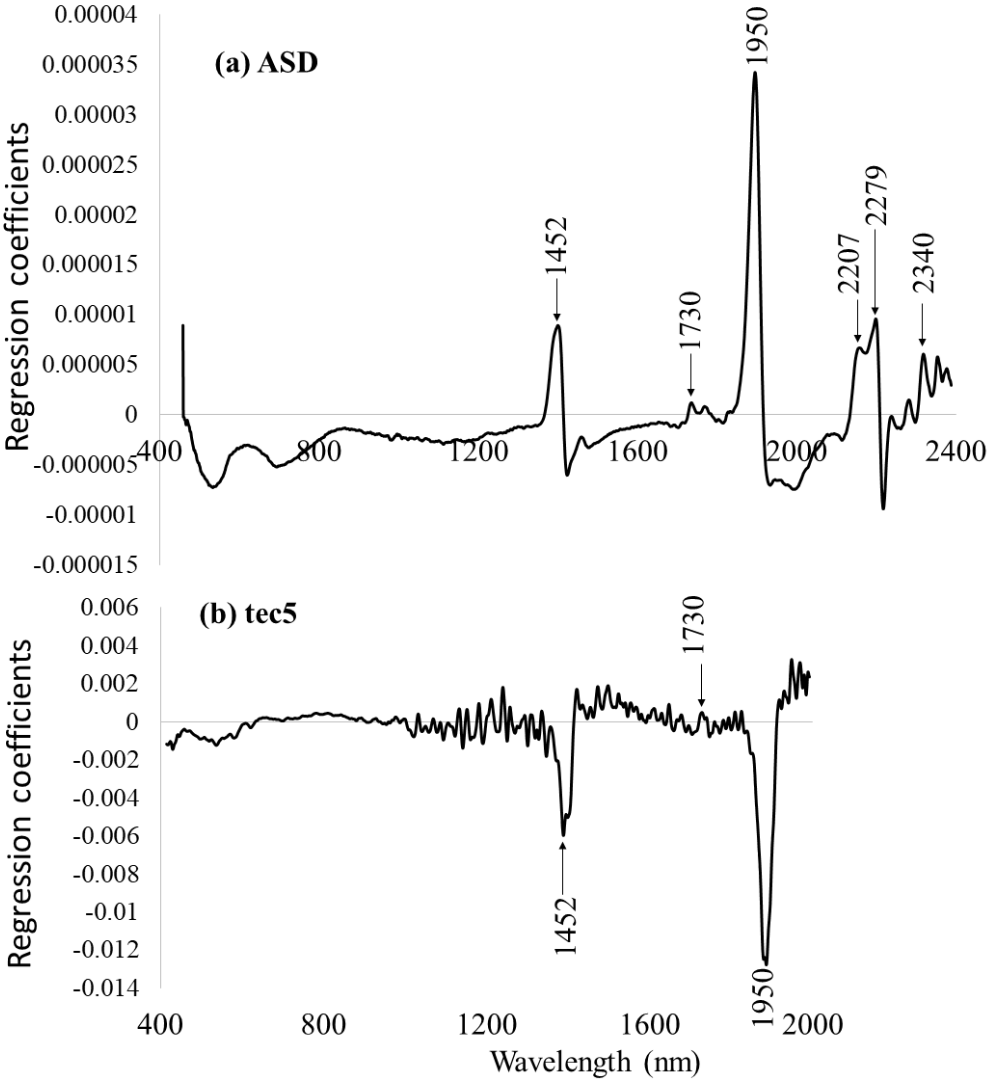


Figure 3-8: Regression coefficients plots resulted from partial least squares regression (PLSR) analysis for total petroleum hydrocarbons (TPH) based on visible and near infrared (vis-NIR) spectra of oil-contaminated soil samples using (a) ASD and (b) tec5 spectrophotometers. Wavelengths highlighted on the plot are the potential features for TPH.

3.6 Conclusion

This pilot study evaluated visible and near infrared (vis-NIR) diffuse reflectance spectroscopy sensitivity to hydrocarbon concentration differences attributed to weathering for enhanced assessment of crude oil contamination in soils. It compared the performance between a full vis-NIR range of 350-2500 nm spectrometer (e.g., ASD) with a short range of 305-2200 nm spectrometer (e.g., tec5), using two calibration methods of random forest (RF) and partial least squares regression (PLSR). From the results reported the following conclusions can be drawn:

- Principal component analysis (PCA) showed reasonable separation between the different weathered soil groups over time. This was true for the ASD spectrometer only, which was attributed to the large wavelength range of 350-2500 nm, compared to the short wavelength range (305-2200 nm) of the tec5 spectrometer. However, since total petroleum hydrocarbon (TPH) content is soil samples decreases with time due to weathering, the sensitivity of the ASD spectrometer for detecting changes due to weathering in soils decreases, particularly after 8 months of contamination.
- Both RF and PLSR analyses supported the PCA results for the ASD spectrometer in separation between different weathering groups, which was again much better that the separation obtained with the tec5 spectrometer. However, the RF model provided clearer separation than PLSR.
- Both RF and PLSR demonstrated that TPH can be estimated throughout time up to two years weathering. However, better estimation of TPH was obtained with RF-ASD model (R^2 = 0.92, RPD = 3.79, RMSE = 108.56 mg/kg), compared to PLSR-ASD model (R^2 = 0.83, RPD = 2.49, RMSE = 164.87 mg/kg).

Overall, the results demonstrated the potential of vis-NIR spectroscopy with a spectral range of 350-2500 nm for the successful estimation and discrimination of different weathering groups in oil-impacted soils. It is a rapid measurement tool for quick on-site investigation and monitoring through weathering (up to 2 years), without the need for collecting soil samples and lengthy hydrocarbon extraction associated to traditional laboratory analysis.

3.7 References

API., 2001. Risk-based Methodologies for Evaluating Petroleum Hydrocarbon Impacts at Oil and Natural Gas E&P Sites, API Publication 4709, API Publishing Services, Washington DC. Available at http://api-ep.api.org/industry/index.cfm?bitmask=002007001005009000.

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.

Brassington, K.J., Hough, R.L., Paton, G.I., Semple, K.T., Risdon, G.C., Crossley, J., Hay, I., Askari, K., Pollard, S.J.T, 2007. Weathered hydrocarbon wastes: a risk assessment management primer. *Crit. Rev Environ Sci Technol.,* 37,199–232.

British Standard BS 7755 Section 5.4., 1998. Determination of particle size distribution in mineral soil material-Method by sieving and sedimentation which is identical to ISO 11277:1998.

British Standard BS 7755 Section 3.8., 1995. Determination of organic and total organic after dry combustion (elementary analysis) which is identical to ISO 10694:1995.

British Standard BS ISO 10390, 2005. Determination of pH.

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

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., Li, B., Ali, M.N., Majumdar, K., Ray, D.P., 2014. Analysis of petroleum contaminated soils by spectral modeling and pure response profile recovery of n-hexane. *Environ. Pollut*. 190, 10–18.

Chakraborty, S., Weindorf, D. C., Zhu, Y., Li, B., Morgan, C. L. S., Ge, Y., Galbraith, J. M., 2012. Assessing spatial variability of soil petroleum contamination using visible near-infrared diffuse reflectance spectroscopy. *J. Environ. Pollut* 14, 2886–2892.

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. *J. Environ. Qual*. 39, 1378–1387.

Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. *J. Geophys. Res*. 95, 12653–12680.

Cipullo S., Prpich G., Campo P., Coulon F., 2018. Assessing bioavailability of complex mixtures in contaminated soils: progress made and research needs. *Sci. Total Environ*. 615, 708–723.

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. *Chem*., 81, 1454–62.

Demetriades-Shah, T.H., Steven, M.D., Clark, J.A., 1990. High Resolution Derivative Spectra in Remote Sensing. *Remote Sens. Environ*. 33:55–64.

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., Coulon, F., Mouazen, A.M., 2018b. Rapid detection of alkanes and polycyclic aromatic hydrocarbons in oil-contaminated soils using visible near-infrared spectroscopy. *Eur. J. Soil Sci*. doi:10.1111/ejss.12567.

Drozdova, S., Ritter, W., Lendl, B., Rosenberg, E., 2013. Challenges in the determination of petroleum hydrocarbons in water by gas chromatography (hydrocarbon index). *Fuels*. 113, 527–536.

Fontán, J. M., Calvache, S., López-Bellido, R. J. & López-Bellido, L., 2010. Soil carbon measurement in clods and sieved samples in a Mediterranean Vertisol by Visible and Near-Infrared Reflectance Spectroscopy. *Geoderma*, 156, 93–98.

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 Sci. Soc. Am. J.* 77, 450–460.

Environment Agency., 2005. The UK approach for evaluating human health risks from petroleum hydrocarbons in soils, Science Report P5-080/TR3, Environment Agency, Almondsbury, Bristol.

Geladi, P., Kowalski, B.P., 1986. Partial least-squares regression: A tutorial. Analytica Chimica Acta 185(1), 1–17.

Hauser, A., Ali, F., Al-Dosari, B., Al-Sammar, H., 2013. Solvent-free determination of TPH in soil by near-infrared reflectance spectroscopy. Int. J. Sustain. Dev. Plan. 8, 413–421.

Hoerig, B., Kuehn, F., Oschuetz, F., Lehmann, F., 2001. HyMap hyperspectral remote sensing to detect hydrocarbons. Int. *J. Remote Sens.* 8, 1413–1422.

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.

Jiang Y., Brassington K.J., Prpich G., Paton G.I., Semple K.T., Pollard S.J.T., Coulon F., 2016. Insights into the biodegradation of weathered hydrocarbons in contaminated soils by bioaugmentation and nutrient stimulation. Chemosphere. 161: 300–307.

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

Malley, D.F., Hunter, K.N., Webster, G.R.B., 1999. Analysis of Diesel Fuel Contamination in Soils by Near-Infrared Reflectance Spectrometry and Solid Phase Microextraction-Gas Chromatography. *J. Soil Contam*. 8, 481–489.

Martens, H., T. Naes., 1989. Multivariate calibration. 2nd ed. John Wiley & Sons, Chichester, UK.

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.

Mouazen, A.M., Karoui, R., De Baerdemaeker, J., Ramon, H., 2006. Characterization of soil water content using measured visible and near infrared spectra. *Soil Science Society of America Journal*, 70, 1295–1302.

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.

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

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. *Environ. Pollut*. 184, 298–305.

Okparanma, R. N., Mouazen, A. M., 2013a. 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., Mouazen, A. M., 2013b. 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.

Paíga, P., Mendes, L., Albergaria, J.T., Delerue-Matos, C.M., 2012. Determination of total petroleum hydrocarbons in soil from different locations using infrared spectrophotometry and gas chromatography. *Chem*. 66, 711–721.

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.

Schwartz, G., Ben-Dor, E., and Eshel, G., 2012. Quantitative analysis of total petroleum hydrocarbons in soils: comparison between reflectance spectroscopy and solvent extraction by 3 certified laboratories. *Appl. Environ. Soil Sci*., 2012, 1–11.

Stenberg, B., 2010. Effects of soil sample pre-treatments and standardised rewetting as interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon. *Geoderma*, 158(1-2), 15–22.

Tamburini, E., Vincenzi, F., Costa, S., Mantovi, P., Pedrini, P., Castaldelli., 2017. Effects of moisture and particle size on quantitative determination of total organic carbon (TOC) in soil using near-infrared spectroscopy. Sensors, 17, 2366, 1-15.

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

Viscarra Rossel, R.A., Walvoort, D.JJ., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*. 131, 59–75.

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

Wold, S., 2010. Personal memories of the early PLS development. Chemometrics and Intelligent Laboratory Systems 58, 83–84.

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

CHAPTER 4 : Rapid prediction of total petroleum hydrocarbons, polycyclic aromatic hydrocarbons and alkane 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

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), polycyclic aromatic hydrocarbon (PAH) and alkanes 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, PAH and alkane in soils. Eighty eight soil samples (3 uncontaminated and 85 contaminated) collected from three sites located in the Niger Delta, Nigeria. All plant and pebble particles were removed and surface was smoothened gently with a spatula for scanning and 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, PAH and alkanes. Prior to model development, spectra were subjected to pre-processing including noise cut, maximum normalization, first derivative and smoothing. Then for TPH (65, 20), PAH (58, 23) and alkanes (65, 18) samples were selected as calibration and validation set, respectively. Both vis-NIR spectrometry and gas chromatography profiles of the respective 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; R^2 = 0.89, RMSEP = 1.02 mg/kg and RDP =2.99; R^2 = 0.85, RMSEP = 55.71

mg/kg and RDP =2.58 outperformed PLSR (R^2 = 0.63, RMSEP = 107.54 mg/kg and RDP =2.55), (R^2 = 0.76, RMSEP = 0.81 mg/kg and RDP =2.07) and (R^2 = 0.49, RMSEP = 101.71 mg/kg and RDP =1.41) for TPH, PAH and alkanes, respectively in cross-validation (calibration). 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, PAH and alkanes in soils.

Key words: Total petroleum hydrocarbon; polycyclic aromatic hydrocarbon; alkanes; vis-NIR spectroscopy; chemometric methods, Partial least squares regression, Random Forest regression.

4.1 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 (PHC) encompass hundreds of various aromatic (PAH) and aliphatic (alkanes) compounds as well as traces of heterocyclic compounds (nitrogen, hydrogen, sulphur), which are well-known environmental contaminants (Cozzolino, 2015; Coulon et al., 2010). However the difference between the term 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). The determination of PHCs in contaminated environmental matrices is a challenge to standardise due to 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 and fractions are required (Ferguson, 1999). With millions of contaminated locations globally (CRCCARE, 2015; Horta et al., 2015), there is a need for efficient, cost-effective, portable and rapid measurement tools for 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 (Chakraborty et al., 2015; Okparanma and Mouazen, 2013). Also, laboratory techniques require prior sample analysis, extraction and sometimes clean-up steps (Forrester et al., 2013). Among the laboratory techniques, gas chromatography with flame ionisation 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 and Mouazen, 2012). To rapidly analyse petroleum-contaminated soils, optical sensors are recommended (Viscarra Rossel and Behren, 2010).

A considerable number of studies have assessed the potential of optical techniques for the rapid estimation of PHC concentration in soils (e.g. Wartini et al., 2017; Okparanma et al., 2014a, 2014b; Okparanma and Mouazen, 2013, 2012; Bray et al., 2009). Details of these studies can be found in Chapter 2 of the thesis. These works have focused on developing models to predict TPH and PAH concentrations. Of these literatures reviewed, none generated models for the prediction of the alkane fraction ($n\mathrm{C}_{10}$-$_{35}$) in oil-contaminated soils, which is a crucial research need fulfilled in this study.

There are 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 (Nawar et al., 2016; Viscarra et al., 2010). 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, multivariate 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, PAH and alkanes contents in soil and to confirm whether or not PHCs can be predicted with RF with higher accuracy than with PLSR. To the best of our knowledge: (1) 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 dataset, (2) no published works on PAH and alkanes in soils using RF models. Thus, the aim of this study is to compare the performance of PLSR linear modelling technique with RF nonlinear technique to predict TPH, PAH and alkanes in oil-contaminated soils from Niger Delta, Southern Nigeria using vis-NIR spectroscopy.

4.2 Materials and methods

4.2.1 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). 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 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. Figure 4-1 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 in a freezer at -20^{o}C prior to GC-MS analysis.

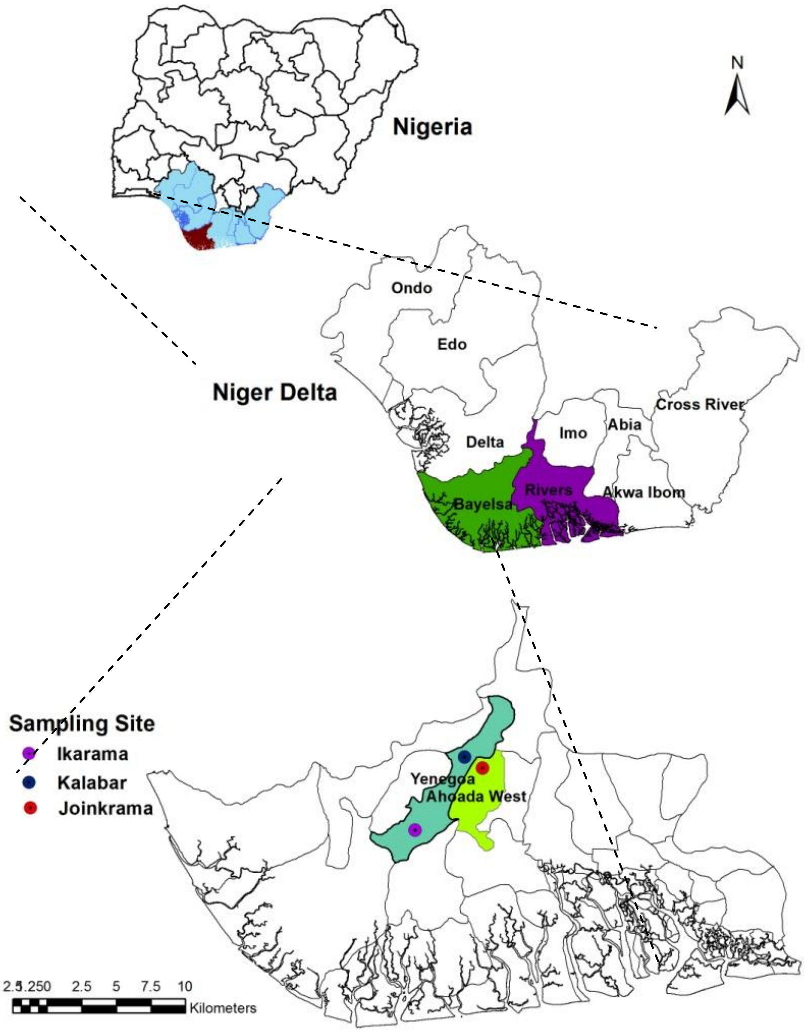


Figure 4-1: Soil sampling locations for the three sites in the Niger Delta, Nigeria.

4.2.2 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.

4.2.3 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 almost 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, samples were properly mixed with spatula, 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.

4.2.4 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. Relative standard deviation (RSD) values for all the soils was <10%. The limit of quantification (LOQ) of 0.02 mg/kg 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.

4.3 Development of calibration models

A two dimensional data matrix was developed by combining the pre-processed spectra (predictor) of the soil samples and the TPH, PAH and alkanes reference values (dependent variables) where the resolved spectral bands (wavelengths) were defined as X_i (the predictor variables), and TPH, PAH and alkanes concentrations as Y_i (the response variables). For TPH, the dataset was divided into 75% (65 samples) for cross-validation (calibration) and 25% (20 samples) for independent validation (prediction). For PAH and alkanes, outliers were detected and removed (4 and 2 samples of PAH and alkanes, respectively), after which the dataset was divided into calibration and prediction sets (58 and 23 for PAH, and 65 and 18 for alkanes), respectively. 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 maximising 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).

4.3.1 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, PAH and alkanes reference values the Y_i variables. PLSR with full cross-validation was used to relate the variation in a single-component variable (e.g. TPH, PAH and alkanes) 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 for modelling soil TPH. While modelling soil PAH (58 and 23) and alkanes (65 and 18) samples were used for calibration and prediction dataset, respectively.

4.3.2 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 PLSR 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 for TPH modelling while 500, 2 and 3 for PAH and alkanes. 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).

4.4 Evaluation of model performance

The performance of TPH, PAH and alkanes prediction models were assessed using: (i) the coefficient of determination in prediction R^2, (ii) root mean square error of prediction (RMSEP), and (iii) residual prediction deviation (RPD) which is a ratio of standard deviation (SD) to RMSEP. 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.

4.5 Results and discussion

4.5.1 Soil chemical analyses

A summary of the soil samples physicochemical properties (TOC, pH, Sand, Silt, and Clay) and TPH, PAH and alkanes concentration determined by GC-MS is provided in Table 4-1. The total organic carbon (TOC) content varies between low to medium with the mean and maximum values of 1.1% and 12.69%, 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). Histograms and box-plots of soil properties and TPH are showed in Figure 4-2**.**

Table 4-1: Soil properties, TPH, PAH and alkane concentrations of the soil samples collected from oil spill sites in the Niger Delta, Nigeria.

Soil properties and PHCs	No	Min	Mean	Median	1st Qu.	3rd Qu.	Max	SD
TOC (%)	30	1.11	4.55	3.85	1.79	5.71	12.69	3.3
pH	30	5.2	6.25	5.95	5.73	6.73	8.2	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)	85	16.07	252.6	213.69	120.66	339.27	666.33	165.51
PAH (mg/kg)	85	0.52	9.11	1.39	0.89	4	312.28	40.2
Alkanes (mg/kg)	85	9.9	187.2	151.75	84.55	259.25	551.22	133.13

TPH (mg/kg) = Total petroleum hydrocarbons; PAH (mg/kg) = polycyclic aromatic hydrocarbon, Alkanes (mg/kg), 1st Qu. = first quartile; 3rd Qu. = third quartile; SD = standard deviation, PHC=petroleum hydrocarbon.

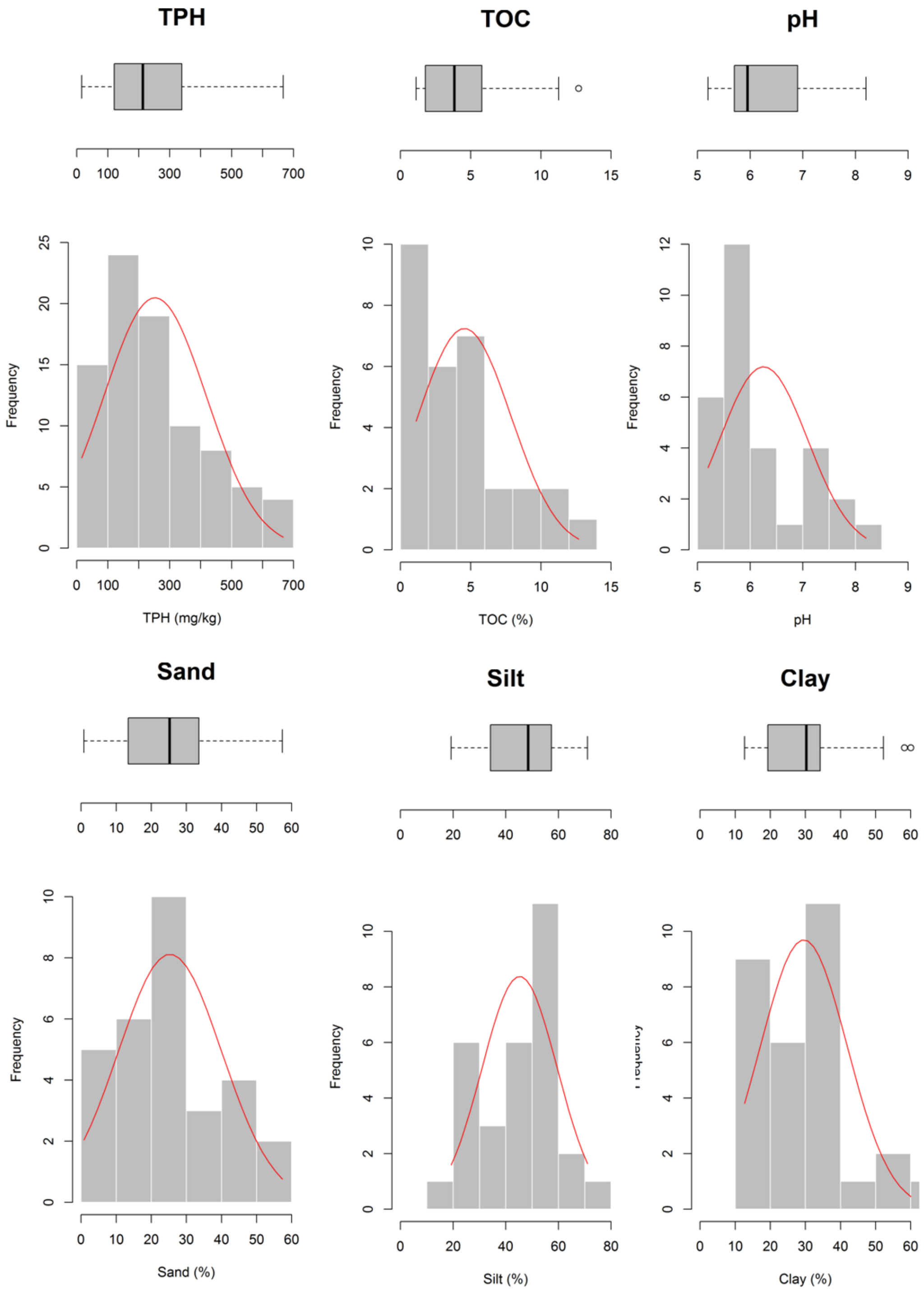


Figure 4-2: Histograms, box-plots with outliers of total petroleum hydrocarbon (TPH) of 85 soil samples, and total organic carbon (TOC), pH, sand, silt and clay content of selected soil samples (30).

Substantial variability was observed for soil pH ranging between 5.2 and 8.2. The TPH values ranged between 16 and 666 mg/kg with mean and standard deviations of 253 mg/kg and 166 mg/kg, respectively. No significant relationship was identified between TOC, pH, sand, silt, clay, and TPH content (randomisation test *p*-values ranged between 0.38 to 0.9 and 0.11 at 0.05 or 0.01 significant level, respectively) (Figure 4-3).

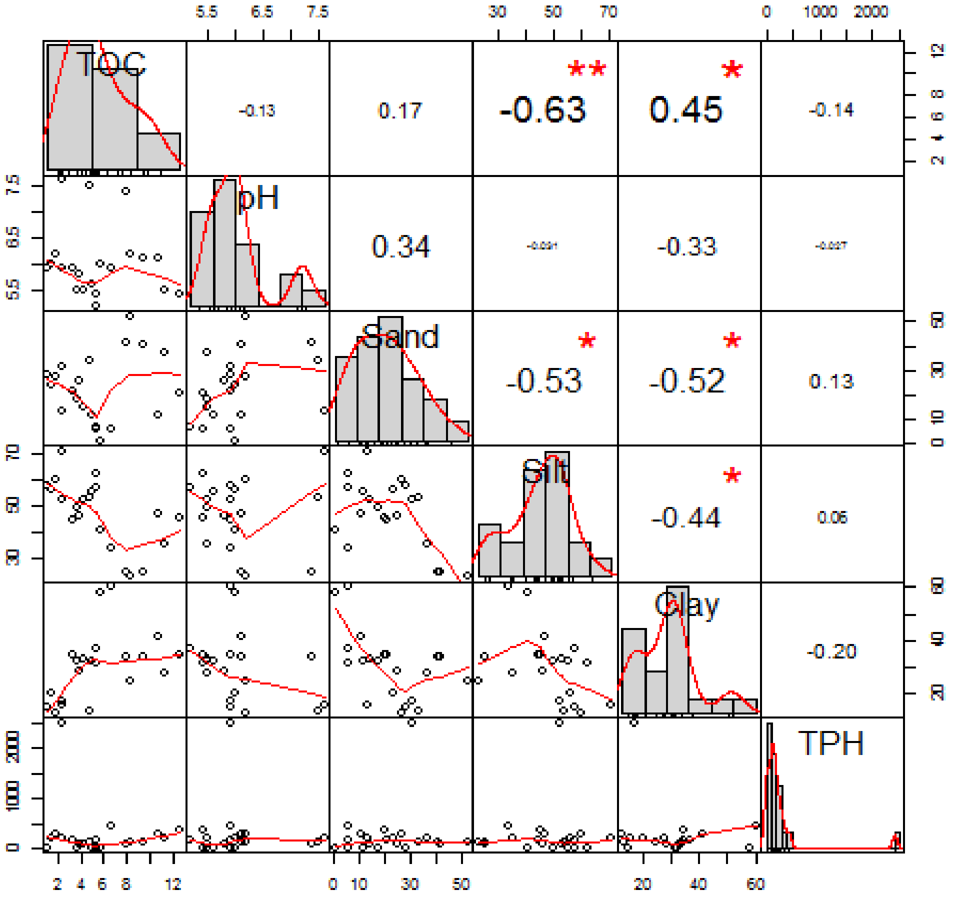


Figure 4-3: Scatterplot matrix for possible pairs of soil variables (lower diagonal), histograms with kernel density overlays for each the target variable (middle) and absolute value of the correlations at significance level of 0.05 (*) and 0.01(**) between the defined pairs of variables (upper diagonal). Soil variables are total petroleum hydrocarbon (TPH), total organic carbon (TOC), pH, sand, silt and clay content of the selected soil samples (30).

The concentrations of alkanes varied between small to medium amounts with mean and maximum values of 151.6 and 551.2 mg/kg, respectively. There were only two samples with values above 512 mg/kg; both were outliers (Figure 4-4a). The concentrations of PAHs ranged from 0.52 to 312.28 mg/kg, with a mean value of 9.11 mg/kg. Four outliers were detected (Figure 4-4b) and were removed before modelling (Figure 4-4c).

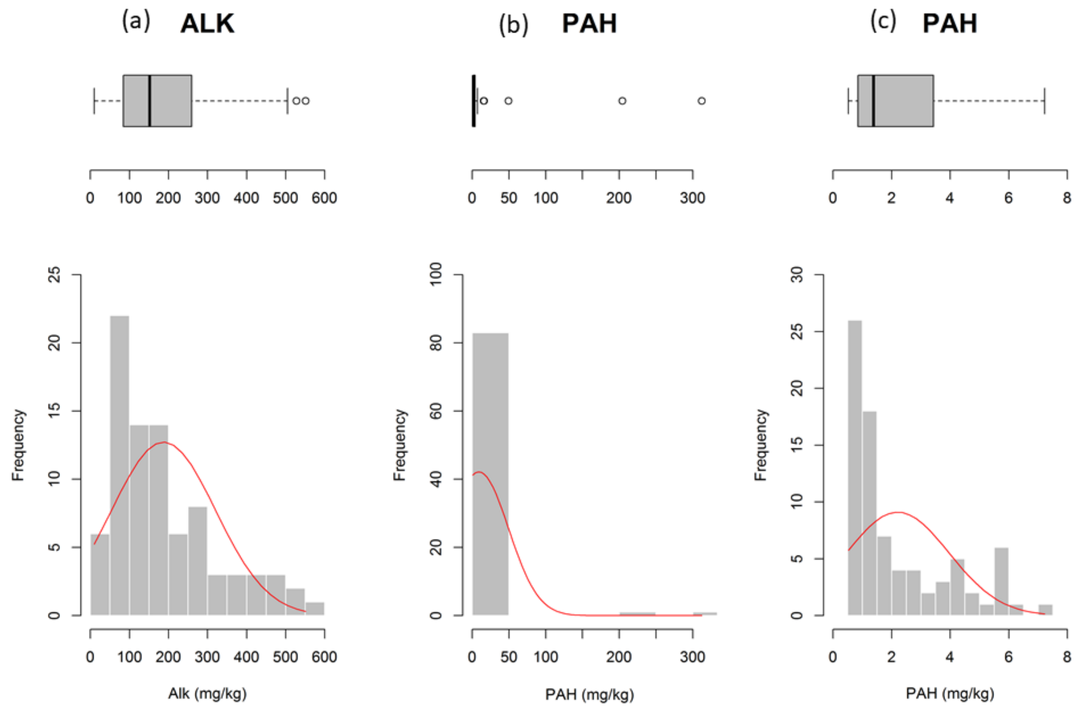


Figure 4-4: Histograms and box-plots of concentrations for (a) alkanes with outliers, (b) polycyclic aromatic hydrocarbons (PAH) with outliers, and (c) PAH without outlier, of the collected eighty five soil samples collected from genuine oil-contaminated sites in the Niger Delta, Nigeria.

Table 4-2 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 and the nC_{16}-C_{21} PAHs between 0.1 and 2.0 mg/kg (Table 4-2). Site 1 had higher average TPH concentration, followed by site 3 and 2. The LOQ for every PAH is shown in Table 4-3. The lowest and highest LOQ in site 1 were 0.02 and 0.47 mg/kg for fluorene and acenaphtylene, respectively. In site 2, the lowest LOQ was 0.02 mg/kg and for Indeno[1,2,3-c,d]anthracene, whereas the highest was 0.26 mg/kg for Benzo[k]fluoranthrene. While the lowest LOQ was 0.04 mg/kg for fluorene, the highest was 1 mg/kg and for indeno[1,2,3-c,d]anthracene.

The distribution and concentrations of the aliphatic fractions and individual PAH across the three sites are summarised in Table 4-4. 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) was 1.5 times greater than at Site 2 (498.1 mg/kg) and 1.1 times greater than Site 3 (671.2 mg/kg) (Table 4-4). Conversely, the concentration of aromatic compounds at Site 3 (321.8 mg/kg) was 97.23 times greater than at Site 2 (3.31 mg/kg) and 39.98 times greater than Site 1 (8.05 mg/kg) (Table 4-4).

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, 0.003 to 0.514 mg/kg and 0.004 to 309.325 mg/kg at Sites 1, 2 and 3, respectively. The concentration of 5- to 6- ring PAHs ranged from 0.001 to 2.246 mg/kg, 0.000 to 0.016 mg/kg and 0.004 to 2.527 mg/kg at Sites 1, 2 and 3, respectively. The relatively large concentration of Benz[a]anthracene (309.3 mg/kg) 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.

Table 4-2: Hydrocarbon fractions concentration (mg/kg) and statistics across the three sites (n= 85).

Hydrocarbon fractions (mg/kg)		Site 1				Site 2				Site 3			
		N	Med	Min	Max	N	Med	Min	Max	N	Med	Min	Max
	nC10-nC12	31	6.6	1.6	31	21	11	2.70	36	33	12	0.6	74
Aliphatic	nC12-nC16	31	21	4.7	83	21	18	6.90	53	33	28	2	154
	nC16-nC21	31	106	26	372	21	105	33	241	33	83	5.4	314
	nC21-nC35	31	81	15	281	21	290	20	168	33	39	3.7	129
∑Alkanes			214.6	47.3	767		424	62.6	498		162	11.7	671
	nC12-nC16	31	0.4	0.1	0.7	21	0.1	0.1	0.1	33	0.1	0.1	0.3
Aromatic	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	0.3	0.1	4.7	21	0.3	0.1	1.6	33	3.4	0.3	310
∑PAH			1.0	0.2	7.4		0.7	0.3	2.7		4.1	0.6	312.1
∑TPH			220	49	666		227	65.87	485		188	16	619

N = number of samples, Med = median, Min = minimum, Max = maximum, TPH = total petroleum hydrocarbon, PAH = polycyclic aromatic hydrocarbon.

Table 4-3: List of limit of quantification for every study polycyclic aromatic hydrocarbon (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.

Table 4-4: Statistical summary of the concentrations of alkanes (mg/kg) and polycyclic aromatic hydrocarbons (PAHs (mg/kg)) for the three contaminated sites from the Niger Delta, Nigeria.

Compound	LOQ(mg/kg)	Ikarama				Kalabar				Joinkrama			
		N	Med.	Min.	Max.	N	Med.	Min.	Max.	N	Med.	Min.	Max.
			mg/kg				mg/kg				mg/kg		
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]-	0.02	31	0.028	0.006	2.246	21	0.030	0.004	0.516	33	0.695	0.004	2.150

Compound	LOQ(mg/kg)	Ikarama				Kalabar				Joinkrama			
		N	Med.	Min.	Max.	N	Med.	Min.	Max.	N	Med.	Min.	Max.
fluoranthrene													
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

N=number of samples, Med=median, Min=minimum, Max=maximum, Ali=aliphatic, PAH=polycyclic aromatic hydrocarbon, TREPH=total recoverable petroleum hydrocarbon.

4.5.2 Spectral analysis of the oil-contaminated

Figure 4-5 shows average raw reflectance spectra and continuum removed reflectance spectra for uncontaminated (n=3) and contaminated (n=85) 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) and uncontaminated soil sample with TPH below 0.04 mg/kg (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 (Chakraborty et al., 2015; Okparanma and Mouazen, 2013; 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 4-5). 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 (Forrester et al., 2010; Workman and Weyer, 2008). 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 (Figure 4-5) 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 (Figure 4-5).

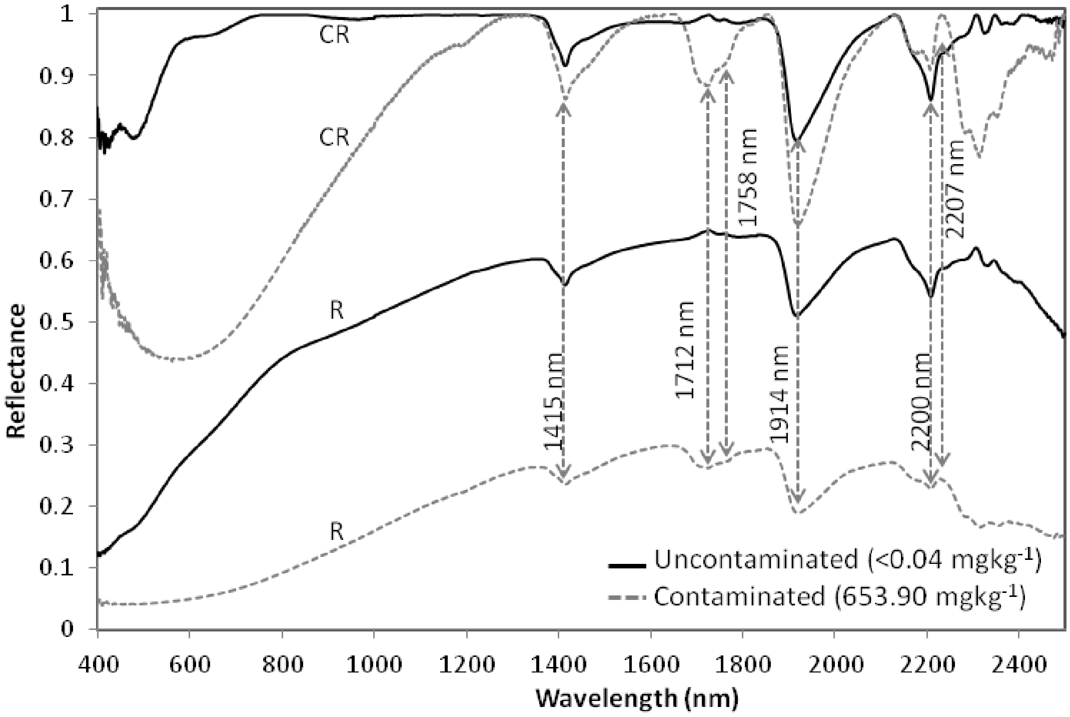


Figure 4-5: Average of raw (R) and continuum removed (CR) spectra of contaminated (85 samples) version uncontaminated soil samples (3 control samples). (1712, 1758) and 2207 nm are known features total petroleum hydrocarbon (TPH) and alkanes, respectively.

The loadings (regression coefficients against wavelengths) based on the first two components (Comp1 and Comp 2) resulted from the cross-validated PLSR analysis for TPH are shown in Figure 4-6. Notably, the numbers and intensities of significant wavelengths have changed, compared to the raw and continuum removed spectra shown in Figure 4-5. 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).

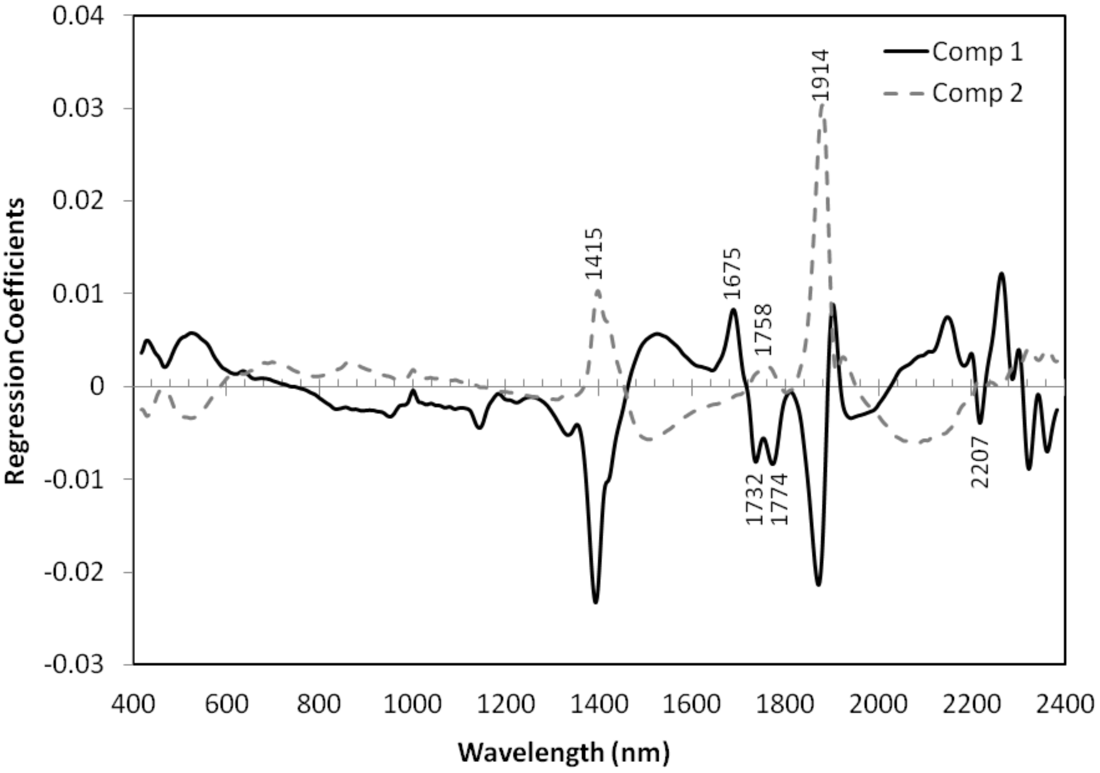


Figure 4-6: **Regression coefficients based on the first and second components (Comp1 and Comp2) versus wavelengths resulted from cross-validated partial least squares regression (PLSR) analysis for total petroleum hydrocarbon (TPH) using visible and near infrared spectroscopy (vis-NIRS) for oil-contaminated soils from Niger Delta, Nigeria. Wavelengths (1732, 1758, 1774), 1675 and 2207 nm are associated with TPH, polycyclic aromatic hydrocarbon (PAH) and alkanes, respectively.**

The regression coefficients versus wavelength for alkanes and PAHs are shown in Figure 4-7. The coefficients resulted from the cross-validated PLSR analysis. 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 4-7(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 agrees 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 4-7b). 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 4-7(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.

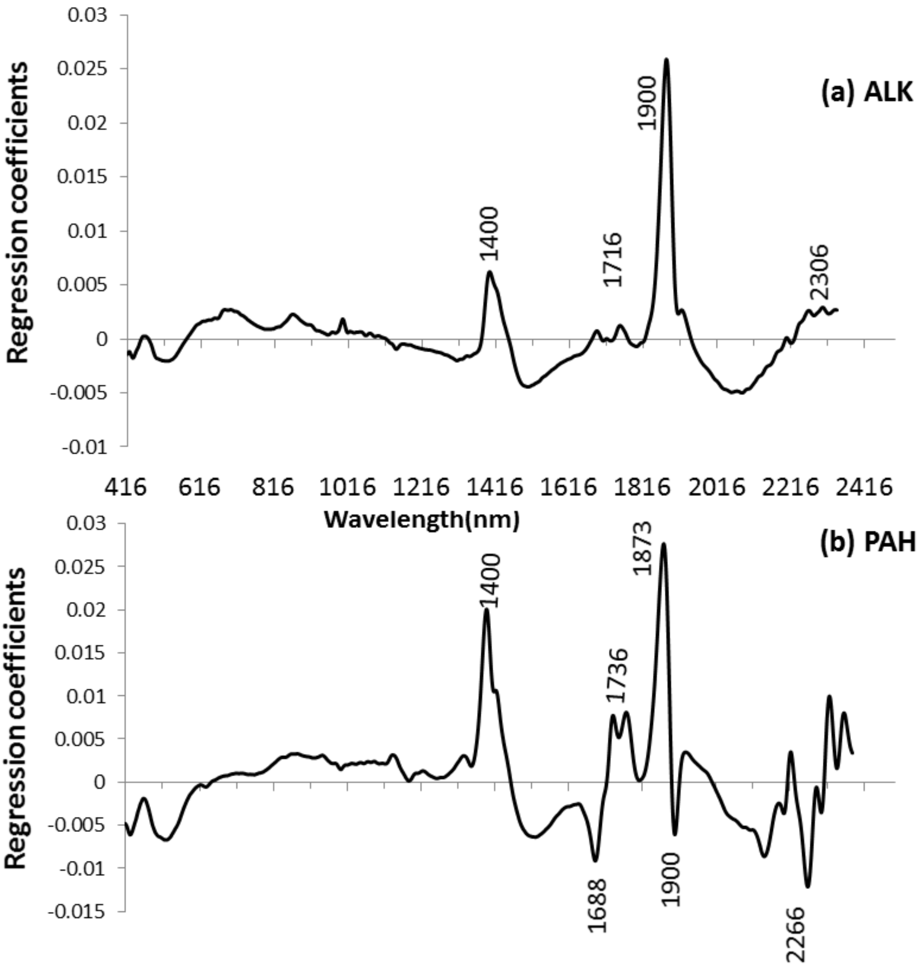


Figure 4-7: **Regression coefficient plots from partial least squares regression (PLSR) analysis for (a) alkanes and (b) polycyclic aromatic hydrocarbon (PAH), based on visible and near infrared (vis-NIR) spectra of oil-contaminated soil samples collected from three sites in the Niger Delta, Nigeria. Wavelengths: 2306 and 1688 nm are associated with alkanes and PAH, respectively.**

4.5.3 Model performance for estimating TPH, PAH and alkanes from vis-NIR spectra

Table 4-5, Figure 4-8, Figure 4-9, and Figure 4-10 summarise the calibration and prediction models of TPH, PAH and alkanes based on PLSR and RF analyses. Figure 4-8, Figure 4-9 and Figure 4-10 represent the scatter plots of predicted TPH, PAH and alkanes versus laboratory measured TPH, PAH and alkanes, respectively. For TPH analysis, RF model (R^2 of 0.85, RMSE of 68.43 mg/kg, RPD of 2.61 and RPIQ = 3.96) outperformed the PLSR model (R^2 of 0.63, RMSE of 107.54 mg/kg, RPD of 1.66 and

RPIQ of 2.55) in calibration. A similar trend to that of the calibration was observed for the prediction set with both RF (R^2 = 0.68, RMSE = 69.64 mg/kg, RPD = 1.85 and RPIQ = 2.53) and PLSR (R^2 = 0.54, RMSE = 78.86 mg/kg, 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 mg/kg, RPD = 1.48 and RPIQ = 1.91) and RF combined with penalized spline regression (PSR) (R^2 = 0.78, RMSE= 0.53 mg/kg, RPD = 2.19 and RPIQ = 0.75). For PLSR, Chakraborty et al. (2015 and 2010) reported slightly higher RPD values of 1.96 and 1.70, 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 (TPH value:16 - 666 mg/kg), while this was not the case in the two studies conducted by Chakraborty et al. (2015 and 2010) where the original TPH values were widely and non-normally distributed (1.22 to 3.74×10^9 mg/kg and 44 to 48 mg/kg, respectively). Also, the high variation of TOC (1.1-12.7%) in our study may increase the performance for estimating the TPH (Table 4-1). 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 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).

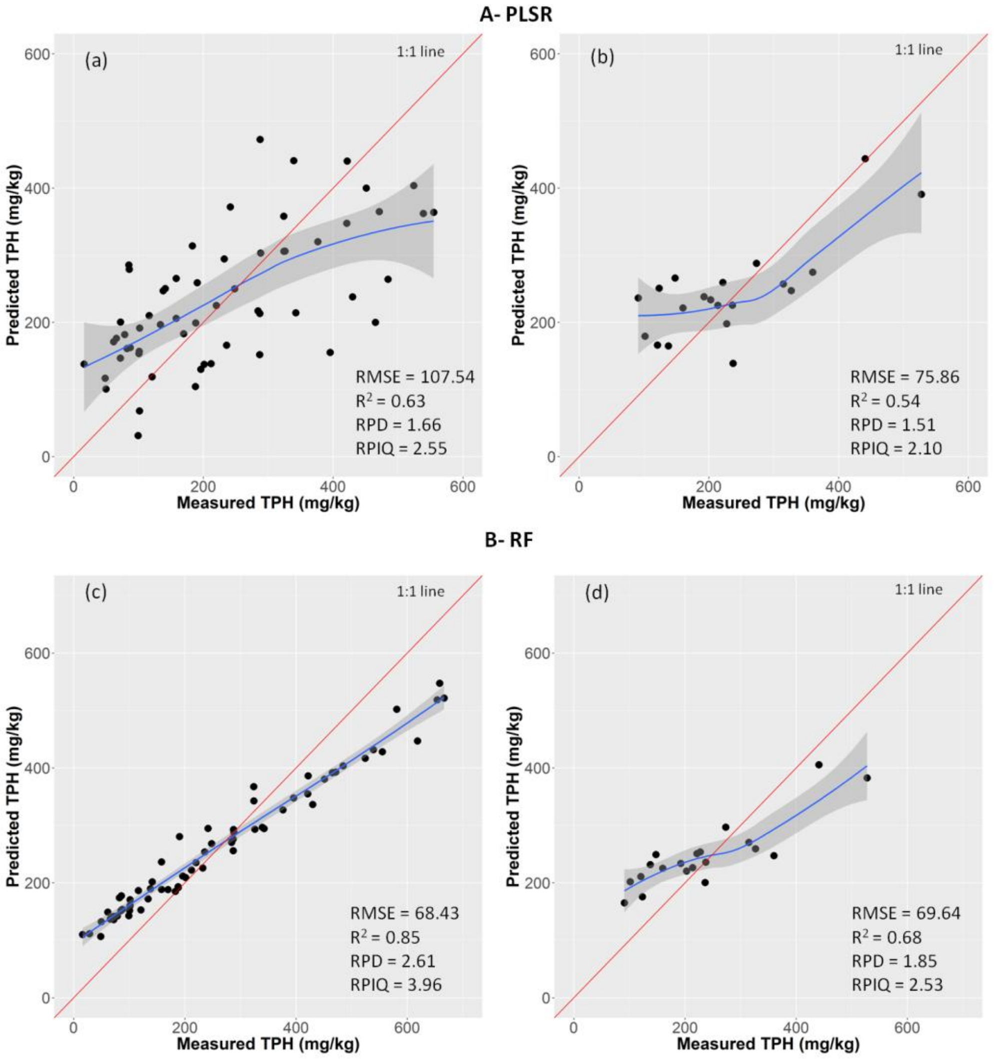


Figure 4-8: **Scatter plots of laboratory measured total petroleum hydrocarbon (TPH) (mg/kg) by SUSE-GC *versus* predicted TPH with partial least squares (PLSR) in (a) calibration and (b) prediction models, and random forest (c) in calibration and (d) prediction model. These models were developed using soil samples from three oil-contaminated sites in Niger Delta, Nigeria.**

Table 4-5: Summary results of partial least squares regression (PLSR) and random forest (RF) models in calibration (cross-validation) and prediction for total petroleum hydrocarbon (TPH), polycyclic aromatic hydrocarbon (PAH) and alkanes prediction in oil-contaminated soil samples collected from three petroleum-contaminated sites in Niger Delta, Nigeria, developed using visible and near-infrared (vis-NIR) spectroscopy.

			PLSR				RF			
				RMSEP				RMSEP		
Compound	Model	N	R^2	(mg/kg)	RPD	LV	R^2	(mg/kg)	RPD	ntrees
TPH	Calibration	65	0.63	107.54	1.66	8	0.85	68.43	2.61	500
(mg/kg)	Prediction	20	0.54	75.86	1.51	4	0.68	69.64	1.85	200
PAH	Calibration	58	0.76	0.81	2.07	6	0.89	1.02	2.99	500
(mg/kg)	Prediction	23	0.56	1.21	1.55	4	0.71	0.99	1.90	200
Alkanes	Calibration	65	0.49	101.71	1.41	6	0.85	55.71	2.58	500
(mg/kg)	Prediction	18	0.36	66.66	1.29	4	0.58	53.95	1.59	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, and ‘ntrees’ = number of trees.

The model performance of PLSR and RF for polycyclic aromatic hydrocarbon (PAH) and alkanes are summarised in Table 4-5 showing that RF outperformed PLSR in both the calibration and prediction. The prediction performance of PAH with RF indicated excellent performance in calibration (R^2 = 0.89, RMSEP = 1.02 mg/kg and RPD = 2.99), and good performance in prediction (R^2 = 0.71, RMSEP = 0.99 mg/kg and RPD = 1.90). Results also showed that the PLSR model performed better for PAHs than for alkanes, with good performance in calibration (R^2 = 0.76, RMSEP = 0.81 mg/kg and RPD = 2.07) and fair performance in prediction (R^2 = 0.56, RMSEP =1.21 mg/kg, and RPD = 1.55) (Table 4-5). 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) (Douglas et al., 2018a; Nawar et al., 2016), 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 be related to the small number of samples used in the present study, compared to those reported elsewhere.

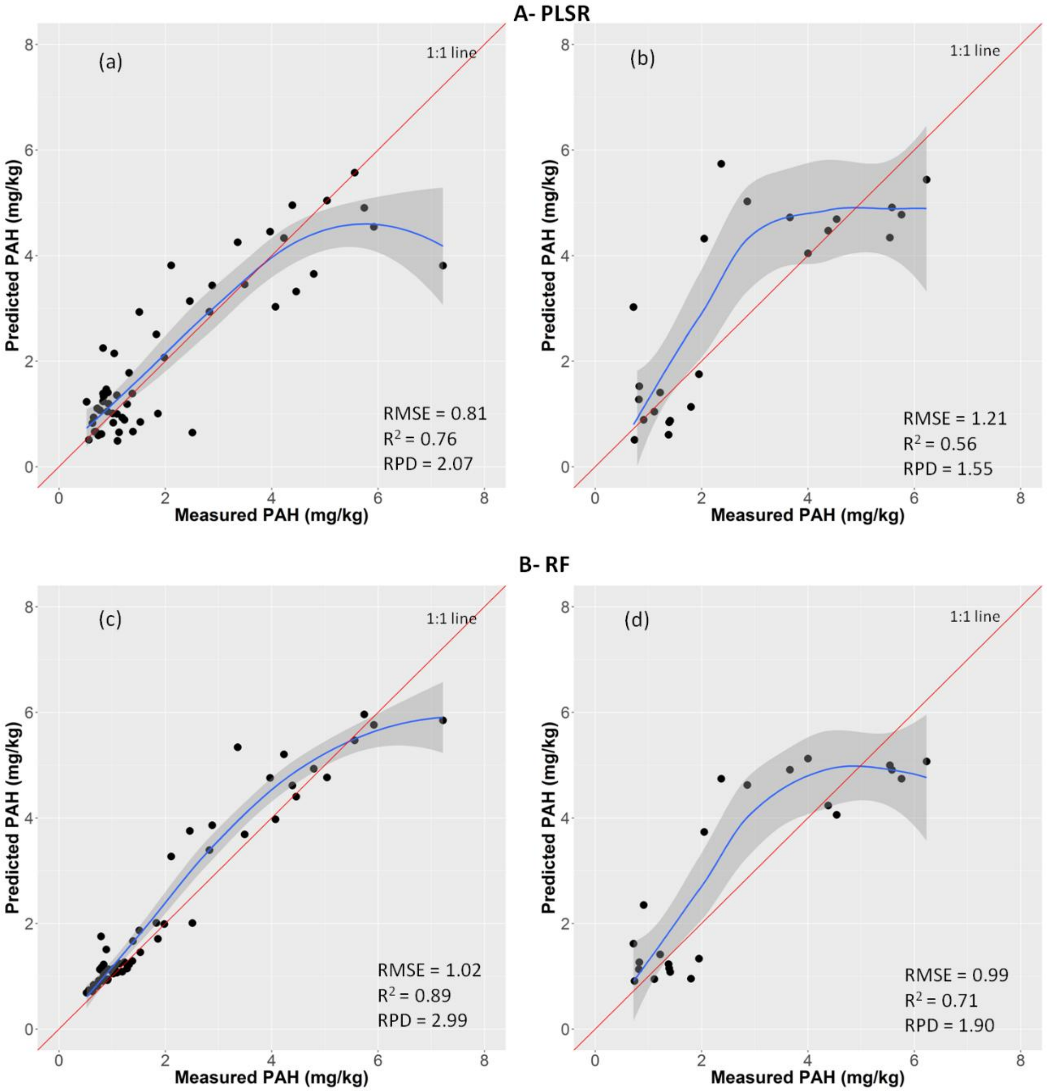


Figure 4-9: **Scatter plots of the measured polycyclic aromatic hydrocarbon (PAH) using gas chromatography mass-spectrometry (GC-MS) versus visible and near infrared (vis-NIR) spectroscopy predicted concentrations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and (B) random forest (RF) method in (c) cross-validation and (d) prediction for samples from the Niger Delta, Nigeria. The blue lines and the grey areas represent the regression line and 95% confidence interval, respectively.**

For alkanes, RF calibration results (R^2 = 0.85, RMSEP = 55.71 mg/kg and RPD = 2.58) are typically better than the prediction results (R^2 = 0.58, RMSEP = 53.59 mg/kg, 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 and RPD of 1.41 in calibration, and of 0.36, 66.66 mg/kg and 1.29, respectively, in prediction (Table 4-5). With the RPD classification system of Viscarra Rossel et al. (2006) to evaluate prediction performance of the models, 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 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. Also, very low hydrocarbons concentrations and mixing of soils from three different sites in the same calibrations might have influenced the model prediction accuracy.

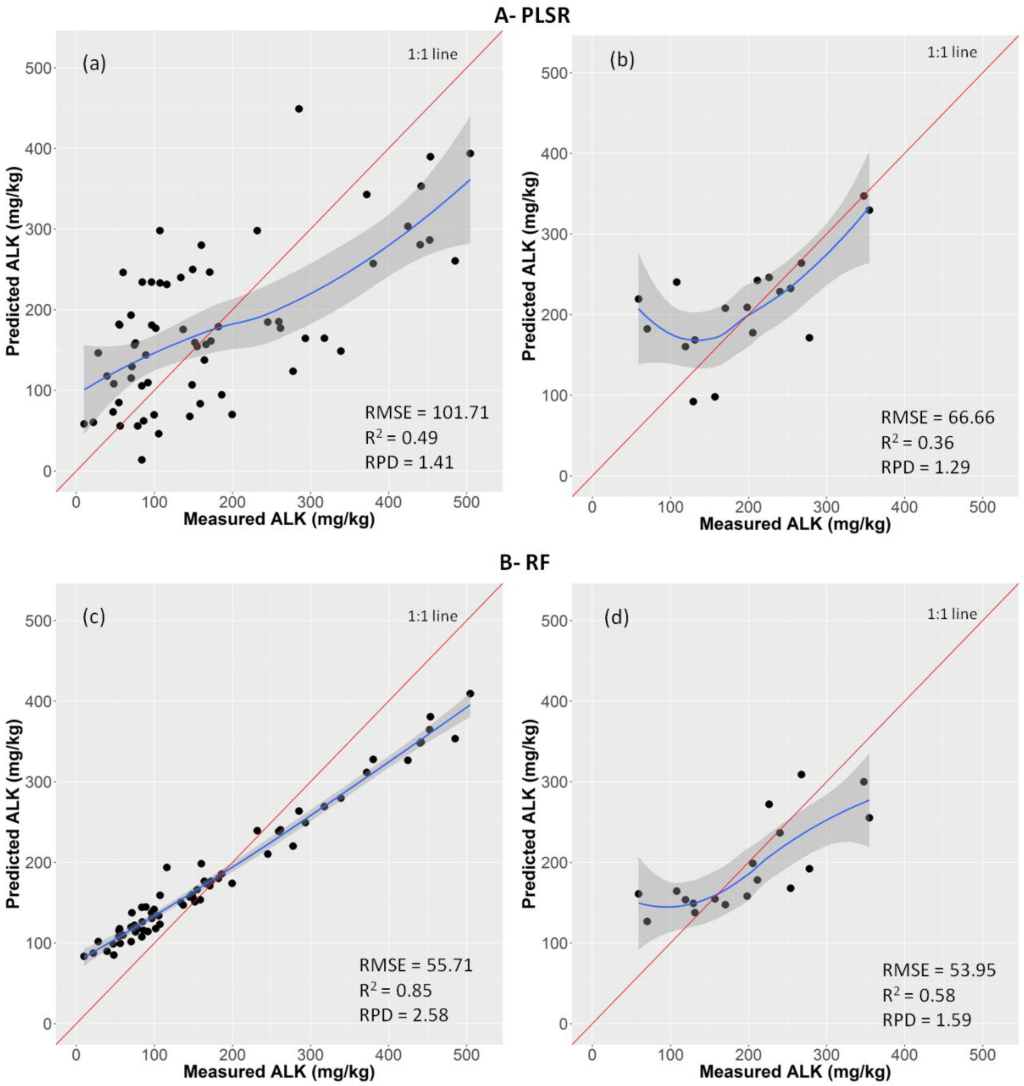


Figure 4-10: **Scatter plots of measured alkanes using gas chromatography mass-spectrometry (GC-MS)** ***versus*** **visible and near infrared (vis-NIR) spectroscopy predicted concentrations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and (B) random forest (RF) in (c) cross-validation and (d) prediction for samples from the Niger Delta, Nigeria. The blue lines and the grey areas represent the regression line and 95% confidence interval, respectively.**

4.6 Conclusion

This study compared the performance of nonlinear random forest (RF) and linear partial least squares regression (PLSR) modelling methods coupled with visible and near infrared (vis-NIR) spectroscopy spectral signals to predict total petroleum hydrocarbon (TPH), polycyclic aromatic hydrocarbons (PAH) and alkanes in genuine oil-contaminated soil samples collected from three petroleum release sites in the 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, and 1.85, respectively, compared to PLSR with 0.54 and 75.86 mg/kg, and 1.51 values, respectively for modelling TPH. The R^2, RPD, and RMSEP values obtained herein by RF models confirm its suitability as ‘a good model predictor’ for the estimation of soil TPH.

RF models for chemometric analysis of PAH and alkanes also outperformed PLSR with R^2 of 0.71, RMSEP of 0.99 mg/kg, RPD of 1.90 and R^2 of 0.58, RMSEP of 53.95 mg/kg, RPD of 1.59 in prediction, respectively. The RF models’ prediction performance of PAHs and alkanes was classified as good and fair, respectively, whereas PLSR models’ performance was fair for PAH and poor for alkanes. 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, whereas the model provided much better results in calibration than in prediction. In contrast, the PLSR model performance slightly only deteriorated between calibration and prediction. Overall, 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. 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.

4.7 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.

CRCCARE (Cooperative Research Centre for Contamination Assessment and Remediation of the Environment)., 2015. Annual Report 2014–2015. Cooperative Research Centre for Contamination Assessment and Remediation of the Environment, University of Newcastle, Callaghan NSW 2308, Australia (At:http://www.crccare.com/files/dmfile/CRCCARE_AnnualReport_2014-15.pdf. Accessed: 14/01/2018).

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., Ali, N., Majumdar, K., Ray, D.P., 2014. Analysis of petroleum contaminated soils by spectral modeling and pure response profile recovery of n-hexane. Environ. Pollut. 190, 10–18. Chromatography. *Environ. Pollut.* 184, 298–305.

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.

Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. *J. Geophys. Res*. 95, 12653-12680.

Coulon, F., Wu, G., 2017. Determination of petroleum hydrocarbon compounds from soils and sediments using ultrasonic extraction In: Hydrocarbon and Lipid Microbiology Protocols McGenity T.J et al. (eds.) Springer-Verlag Berlin Heidelberg, 31- 46.

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.

Cozzolino, D., 2015. Near infrared spectroscopy as a tool to monitor contaminants in soil, sediments and water – state of the art, advantages and pitfalls. *Trends in Environmental Analytical Chemistr,* 9, 1–7.

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., 2018a. 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.

Drozdova, S., Ritter, W., Lendl, B., Rosenberg, E., 2013. Challenges in the determination of petroleum hydrocarbons in water by gas chromatography (hydrocarbon index). *Fuels*. 113, 527-536.

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

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

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.

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.

Geladi, P., Kowalski, B.P., 1986. Partial least-squares regression: A tutorial. *Analytica Chimica Acta* 185(1), 1-17.

Hoerig, B., Kuehn, F., Oschuetz, F., Lehmann, F., 2001. HyMap hyperspectral remote sensing to detect hydrocarbons. Int. *J. Remote Sens*. 8, 1413–1422.

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.

Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.I., Myers, D.B., Harris, W.G., 2015. Modeling soil carbon fractions with visible near-infrared (VNIR)and mid-infrared (MIR) 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.

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 available on https://cran.r-project.org/web/packages/randomForest/ randomForest.pdf.

Maleki, M.R., Mouazen, A.M., Ramon, H., De Baerdemaeker, J., 2007. Optimisation of soil VIS-NIR sensor-based variable rate application system of soil phosphorus. *Soil and Tillage Research*, 94, 239–250.

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.

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., 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.

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.

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

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

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., 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.

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.

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/).

Saari, E., erämäki, P., Jalonen, J., 2008. Measurement uncertainty in the determination of total petroleum hydrocarbons in soil by GC-FID 92, 3–12.

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

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

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and Near Infrared Spectroscopy in Soil Science. *Adv. Agron*. 107, 163-215.

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

Thien, S.J., 1979. A flow diagram for teaching texture by feel analysis. *Journal of Agronomic Education*. 8:54-55.

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., Walvoort, D.JJ., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*. 131, 59-75.

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

Wang, J., He, T., Lv, C., Chen, Y., Jian, W., 2010. Mapping soil organic matter based on land degradation spectral response units using Hyperion images. Int. J. Appl. Earth Obs. Geoinf. 12, S171–S180.

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.

Wold, S., 2010. Personal memories of the early PLS development. Chemometrics and Intelligent Laboratory Systems 58, 83–84.

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, Allen and Unwin Limited, UK, 107.

CHAPTER 5 : Rapid prediction of total petroleum hydrocarbon, polycyclic aromatic hydrocarbon and alkanes contamination in soils by a handheld mid-infrared spectroscopy

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

[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

Abstract: Rapid analysis of petroleum-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 rapid, non-destructive and inexpensive prediction of total petroleum hydrocarbons (TPH), polycyclic aromatic hydrocarbons (PAH) and alkanes in petroleum-contaminated soils. Partial least squares (PLS) regression and random forest (RF) modelling techniques were compared for the prediction of TPH, PAH and alkanes in eighty five (n=85) genuine petroleum-contaminated soil samples collected from 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, 2.72 and 1.96, and root mean square error of prediction (RMSEP) values of 63.80, 0.83 and 65.88 mg/kg for TPH, PAH, and alkanes, 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: Petroleum-hydrocarbon contamination; soil pollution; reflectance spectroscopy, mid-infrared; random forest regression, partial least squares regression

5.1 Introduction

Soil pollution with petroleum hydrocarbons (PHCs) severely impacts the environment and wellbeing of the people, and reduce the agronomic potential of agricultural, grasslands and forest lands. PHCs cover a wide range of aliphatic and aromatic compounds with different concentrations. These include benzene, toluene, ethylbenzene, and xylenes (BTEX), polycyclic aromatic hydrocarbons (PAHs), and oxygen-, nitrogen- and sulphur-containing compounds (Ritchie et al., 2001), which can contaminate the soil through oil spills or leaks as a result of exploration, production, storage and transportation of petroleum products. PHCs are not only widespread contaminants in the soil and water domains but also toxic for human health and other species (Li et al., 2014; Sammarco et al., 2013). Thus, BTEX can further enter the human body via ingestion of contaminated food or intake of contaminated water, vapour inhalation from the impacted soil, and dermal contact (Yang et al., 2017). Similarly, PAH fractions contain toxic constituents that can be adsorbed and further accumulated in the soil, and which may then leak to groundwater 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 techniques 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 (Brassington et al., 2010; Wang and Fingas, 1995). 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 (Chakraborty et al., 2015; Horta et al., 2015; Viscarra Rossel et al., 2011). 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 (Douglas et al., 2018a, 2018b; Okparanma and Mouazen, 2013).

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., Douglas et al., 2018a, 2018b; Chakraborty et al., 2015; Okparanma et al., 2014; Chakraborty et al., 2010), only few studies have used the 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) and MIR (root mean square error of prediction of cross validation [RMSEcv] = 2000-4000 mg/kg) laboratory-based spectroscopy for 0-100 000 mg/kg TPH range; whereas Forrester et al. (2013) predicted TPH concentration in 205 naturally contaminated soils by laboratory-based MIR (RMSE = 601 mg/kg and ratio of prediction deviation [RPD] = 3.4) and NIR (RMSE = 564 mg/kg and RPD = 3.7) methods. Webster et al. (2016) utilised a handheld MIR field instrument for the prediction of TPH in three sites (n = 194), reporting RMSEP of 1225, 1293, and 1091 mg/kg and RPD values of 13, 8, and 10 for Sites 1 (laboratory constructed soils), 2 (field contaminated soils), and 3 (laboratory constructed soils), respectively, where all samples were air-dried before scanning. Wartini et al. (2017) used portable MIR and vis-NIR spectroscopy 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, 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). Hitherto, there is no study yet on the application of MIR spectroscopy for the prediction of alkanes and PAHs in soils indicating an important research gap which is addressed in this study. From the scanty literature on the use of MIR for PHCs analysis in soils, no study was reported on the performance of the MIR methodology using field-moist soils. Therefore, the focus of the current study was to assess the potential of a handheld MIR instrument for the prediction of TPH, PAH and alkanes in eighty five genuine petroleum-contaminated soil samples (field-moist) collected from oil spill sites in the Niger Delta region of Nigeria. PLSR and RF regression models were developed to estimate soil TPH, PAH and alkanes. These modelling methods were evaluated to determine which method could perform best based on independent validation (prediction) datasets.

5.2 Materials and methods

5.2.1 Study area and soil sampling

The area of study is 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), in the Niger Delta province of Nigeria. The soil sampling location map is shown (Chapter 4, Figure 4-10). The geology as a part of the Niger Delta is simple and consisting of Benin, Agbada and Akata formations, overlain by various types of Quaternary deposits (Kogbe, 1989; Wright et al., 1985). More details of the study area including geology, sampling method, number of samples collected, sampling depth, mass of samples, and sample preservation can be found in Chapter 4, section 4.2.1.

5.2.2 Hydrocarbon analysis

Chemical analysis for hydrocarbon concentrations in soil was carried out using sequential ultrasonic solvent extraction gas chromatography (SUSE-GC) as described in 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) and shaken for 16 h at 150 oscillations per min over 16 h; and finally sonicated for 30 min at 20°C. Identification and quantification of aliphatic hydrocarbons and PAHs were carried out by an Agilent 5973N GC-MS operated at 70 eV in positive ion mode as described in Chapter 3 section 3.4 of the thesis. The validation methodology was set against a robust and validated GC-MS method previously reported (Risdon et al., 2008).

5.2.3 MIR spectra collection and pre-processing

MIR diffuse reflectance spectra of field-moist soil samples were collected 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. A total of 32 scans were acquired per sample and these were later averaged to produce a reflectance spectrum for each individual sample using Microlab software V5.0 supplied with the spectrometer (Agilent Technologies, Santa Clara, CA, United States). This instrument was calibrated with the standard background, a silver reference cap provided by the manufacturer. A total of eighty five (n=85) field-moist oil-contaminated soil samples were placed in a 5-cm diameter plastic Petri dishes without compression and levelled using a stainless-steel blade.

All collected spectra were converted from reflectance (R) to absorbance by log (1/R), smoothed using the Savitzky-Golay (SG) algorithm with a window size of 11 and polynomial of order 2, and normalised using maximum normalization transformations. SG algorithm was used to remove instrument noise within the spectra by smoothing the data using the polynomial regression, while normalisation of the spectra was implemented to align all spectra to the same scale and to obtain even distribution of the variances and average values (Rinnan et al., 2009). Baseline corrections were finally implemented using ‘modpolyfit’ methods in chemometric R- package (R Core Team, 2013), before modelling.

5.3 Modelling

The data matrix including the processed MIR spectra and the SUSE-GC TPH, PAH and alkanes reference values were 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). Kennard-Stone algorithm was used to select the calibration and validation sets based on its capability to select samples with a uniform distribution over the predictor space (Kennard and Stone, 1969). After outlier removal and division of samples into calibration and validation sets, the former set was subjected to both PLSR and RF analyses to establish calibration models for TPH, PAH and alkanes.

5.3.1 Partial least squares regression (PLSR)

PLSR is a widely used multivariate analysis method in spectroscopy, which was introduced previously by Wold et al. (2001). The algorithm uses a linear multivariate model to relate two data matrices – the predictor variables, X (MIR spectra in this case), and the response variables, Y (TPH, PAH or alkanes). Information in the original X data is projected onto a small number of underlying orthogonal ("latent") variables called latent variables. PLSR with full cross-validation was used to relate the variation in a single-component variable (TPH, PAH or alkanes) to the variation in a multi-component variable (e.g., wavelength), using package 'pls' available in R software (R Core Team, 2013). The optimal number of latent variables (factors) for future predictions was determined as the number of factors that resulted in the smallest RMSEP.

5.3.2 Random forest regression

Random forest (RF) developed by Breiman (2001), is an ensemble learning method commonly used for classification and regression analyses. The algorithm works by growing an ensemble of regression trees based on binary recursive partitioning, where the algorithm first begins with a number of bootstrap samples (*ntree*) from the predictor space (original data) (Breiman, 2001). Each bootstrap sample will then grow a regression tree with a modifying operation, in which subsequently a number of the predictors (*mtry*) are randomly sampled, and the algorithm chooses the best split from among those sampled variables rather than considering all variables. In this work, an *ntree* of 500 and 200, and an *mtry* of 2 were used to develop the TPH, PAH and alkanes models. The same datasets with the same spectra pre-processing used in PLSR analysis (77% calibration, 23% validation) were utilised for RF. The RF models were performed using the R software package, named Random Forest Version 4.6-12 (Liaw and Wiener, 2015).

5.4 Evaluation of model performance

The performance of models for the prediction of TPH, PAH and alkanes was assessed using: (i) the coefficient of determination in prediction R^2, (ii) RMSEP, and (iii) RPD, which is a ratio of standard deviation (SD) to RMSEP. We adopted the Chang et al. (2001) RPD classification criterion, where: RPD < 1.4 indicates no predictive ability, 1.4 <RPD < 2.0 indicates limited predictive ability, and RPD > 2.0 indicates accurate predictive ability. In general, a good model prediction would have high values of R^2 and RPD, and small RMSEP values.

5.5 Results and discussion

5.5.1 Laboratory analysis of TPH, PAH and alkanes

Table 5-1 displays the summary statistics of TPH, PAH and alkanes concentrations acquired using SUSE-GC from the three study sites (Ikarama, S1; Kalabar, S2; and Joinkrama, S3). Among the sites, S3 happened to be 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).

Table 5-1: Statistical summary of total petroleum hydrocarbons (TPH), alkanes and polycyclic aromatic hydrocarbons (PAH) concentrations of the collected soil samples measured with sequential ultrasonic solvent extraction gas chromatography (SUSE-GC).

	N	Min.	Mean	Median	1st Qu.	3rd Qu.	Max.	St. dev.
TPH (mg/kg)	85	16.07	252.59	213.69	120.66	339.27	666.33	165.51
Alkanes (mg/kg)	85	9.9	187.24	151.75	84.55	259.25	551.22	133.13
PAH (mg/kg)	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.

5.5.2 Spectra of soils

MIR absorption spectra of oil contaminated soils from the three sites are compared with an uncontaminated soil spectrum in Figure 5-1A and B, for raw and maximum normalised spectra, respectively. 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 Figure 5-1A and B were similar, and differences can be attributed to soil physico-chemical properties and level of oil contamination. Absorbance peaks (Figure 5-1A) 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 (TREPH) 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}). 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 a 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 Figure 5-1A and B, which is a clear characteristic to differentiate the contaminated samples from the uncontaminated sample.

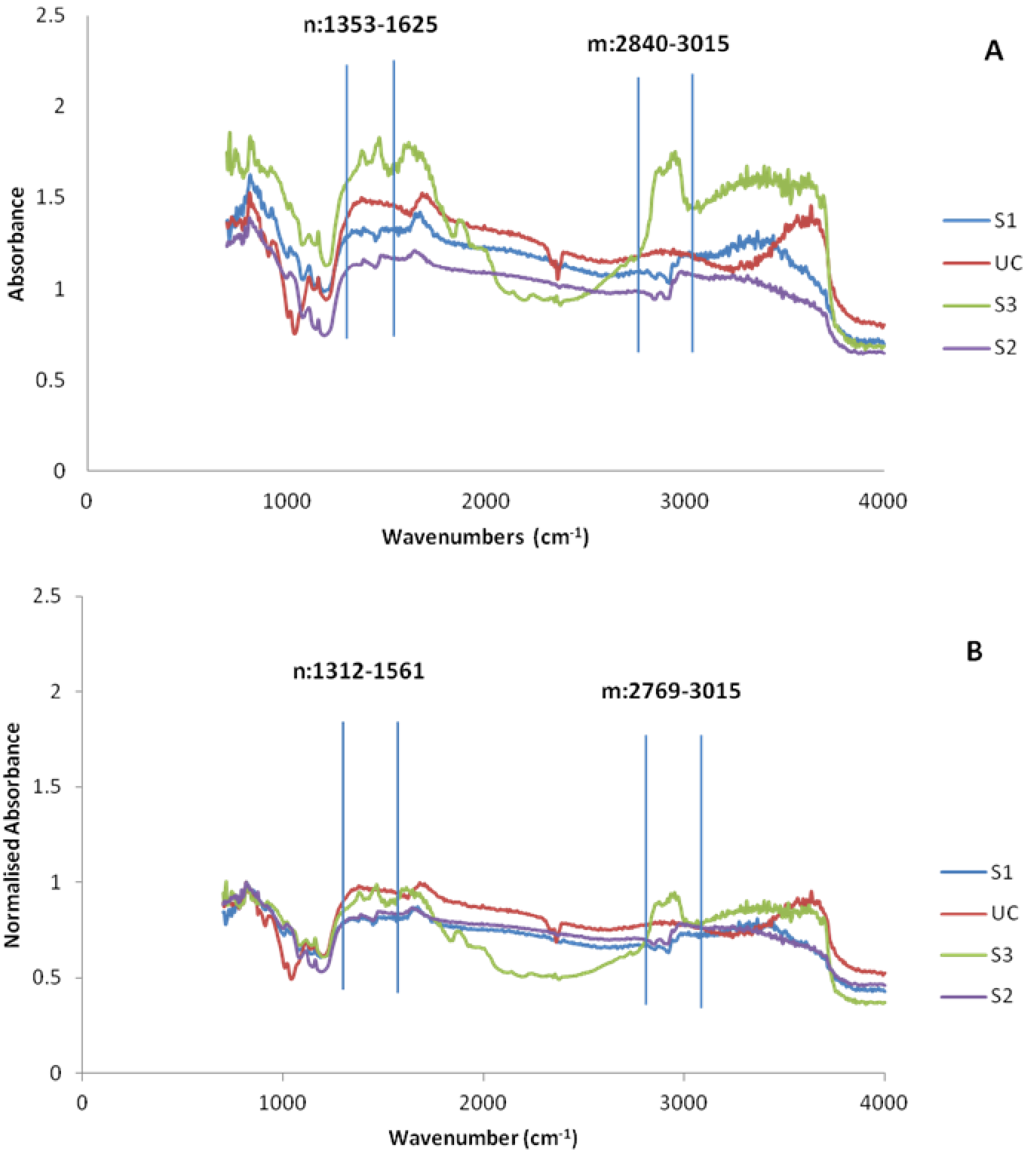


Figure 5-1: (A) Raw mid infrared (MIR) absorbance spectra and (B) maximum normalized MIR absorbance spectra of site 1 (S1) oil-contaminated soil, site 2 (S2) oil-contaminated soil, site 3 (S3) oil-contaminated soil, and uncontaminated (UC) soil spectrum. All samples were collected from the Niger Delta, 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.

5.5.3 Models performance for predicting TPH, PAH and alkanes

Table 5-2 shows the modelling results in calibration and prediction of TPH, PAH and alkanes 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, $R^2 = 0.79$, RPD = 2.27, RMSEP = 0.83 mg/kg and $R^2 = 0.72$, RPD = 1.96, RMSEP = 68.88 mg/kg for TPH, PAH and alkanes, 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, which can be attributed to the non-linear behaviour of the MIR spectral response of the data set that is not accounted for by the linear PLSR model, whereas RF is capable to handle this nonlinearity (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 affects the prediction accuracy of PHCs by reducing the intensity of peaks related to these contaminants (Hazel et al 1997); and the non-linearity effect becomes much stronger with high moisture contents (Webester 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 non-processed) soil samples with high soil moisture content.

Our results for RF prediction are better than those reported by Wartini et al. (2017) for cross-validation prediction of TRH in laboratory spiked soil samples using a field portable MIR coupled with PLSR (RMSE and R^2 of 1592 mg/kg and 0.89, respectively). 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 5-2).

The superior performance of MIR over that of vis-NIR may be attributed to fundamental vibrations of molecules that take place in the MIR spectral region, which generates more intense peaks (Soriano-Disla et al., 2014; Reeves, 2010). These findings, therefore, support the use of a portable MIR instrument to predict TPH, PAH and alkanes in fresh oil contaminated soil samples.

Table 5-2: Summary results of partial least squares regression (PLSR) and random forest (RF) models in calibration (cross-validation) and prediction for total petroleum hydrocarbon (TPH), polycyclic aromatic hydrocarbon (PAH) and alkanes (ALK) prediction in oil-contaminated soil samples collected from three petroleum-contaminated 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)	RPD	RPIQ	LV	R^2	RMSEP (mg/kg)	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	1.9	200	
	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	200	
	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	1.8	2.09	8	0.79	0.83	2.27	3.83	200	

Instrument		**PLSR**					**RF**					Property
	Present study	R^2	RMSEP (mg/kg)	RPD	RPIQ	LV	R^2	RMSEP (mg/kg)	RPD	RPIQ	'ntrees'	
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	2.1	4	0.68	69.64	1.85	2.53	200	
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		4	0.58	53.95	1.59		200	
	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		4	0.71	0.99	1.99		200	

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, TPH =total petroleum hydrocarbon, PAH = polycyclic aromatic hydrocarbon, ALK = alkanes.

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 soils, field contaminated and laboratory constructed soils. The poor result in the current study may be attributed to the small dataset used. In another study, Wartini et al. (2017) reported R^2 and RMSEcv of 0.89 and 1592 mg/kg, 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.

Figure 5-2, Figure 5-3, and Figure 5-4 show the scatter plots of measured *versus* predicted TPH, PAH and alkanes for both the cross-validation and prediction sets obtained with both the RF and PLSR models. 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 (Figure 5-2) compared to more scattered points around the 1:1 lines for PAH (Figure 5-3) and alkanes (Figure 5-4). According to RPD classification suggested by Chang et al. (2001), accurate predictions for TPH and PAH were obtained using RF (RPD = 2.35 and 2.27, respectively), whereas a limited prediction for the alkanes with RPD of 1.96 was observed. These results are consistent with previous studies for estimating TPH based on MIR (Wartini et al., 2017; Webster et al., 2016). 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 as mentioned in (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 that suggests 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 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 non-processed) field sample conditions.

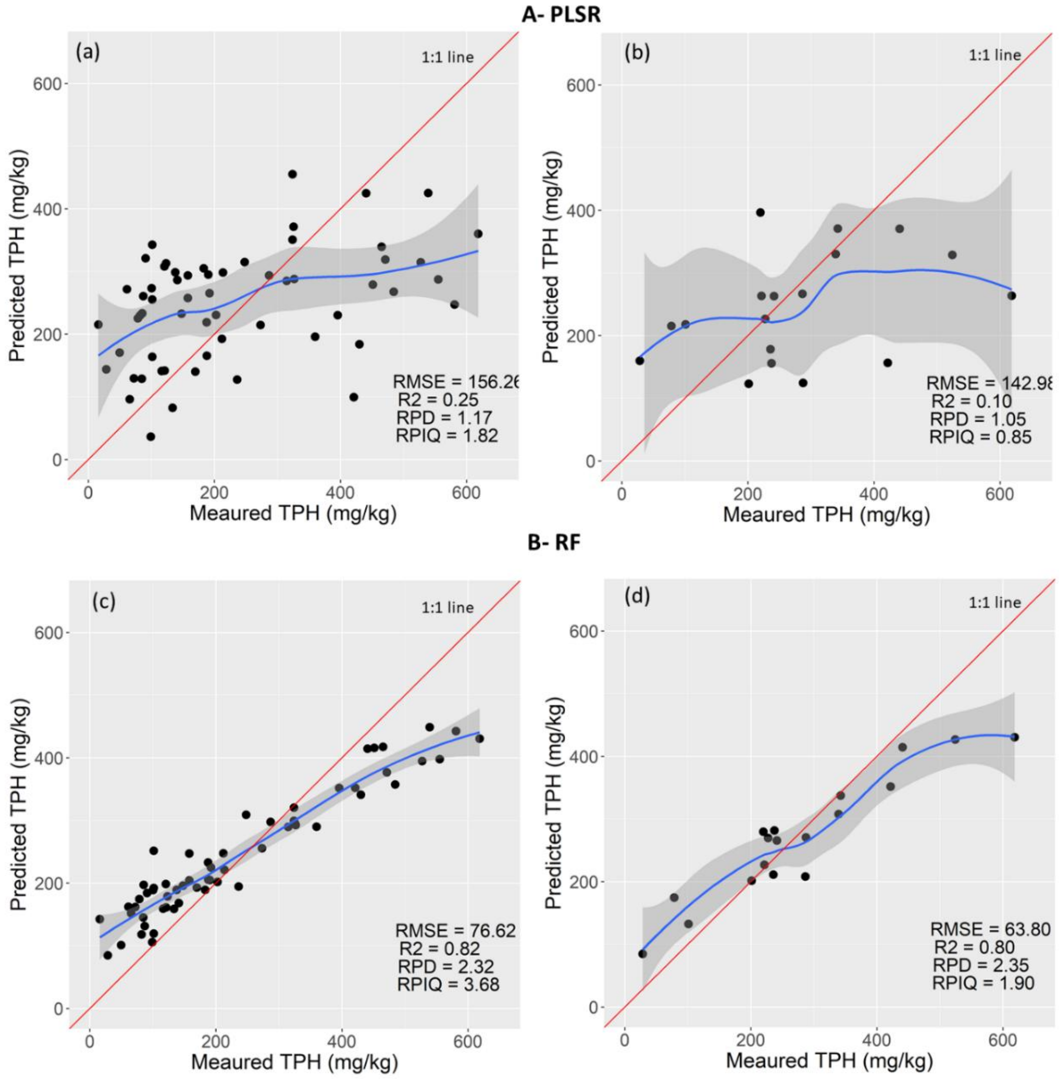


Figure 5-2: Scatter plots of sequential ultrasonic solvent extraction gas chromatography (SUSE-GC) measured total petroleum hydrocarbon (TPH) *versus* mid-infrared (MIR) spectroscopy predicted concentrations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction (b), and (B) random forest (RF) in (c) cross-validation and (d) prediction. The blue lines and grey areas represent the regression line and 95% confidence interval, respectively.

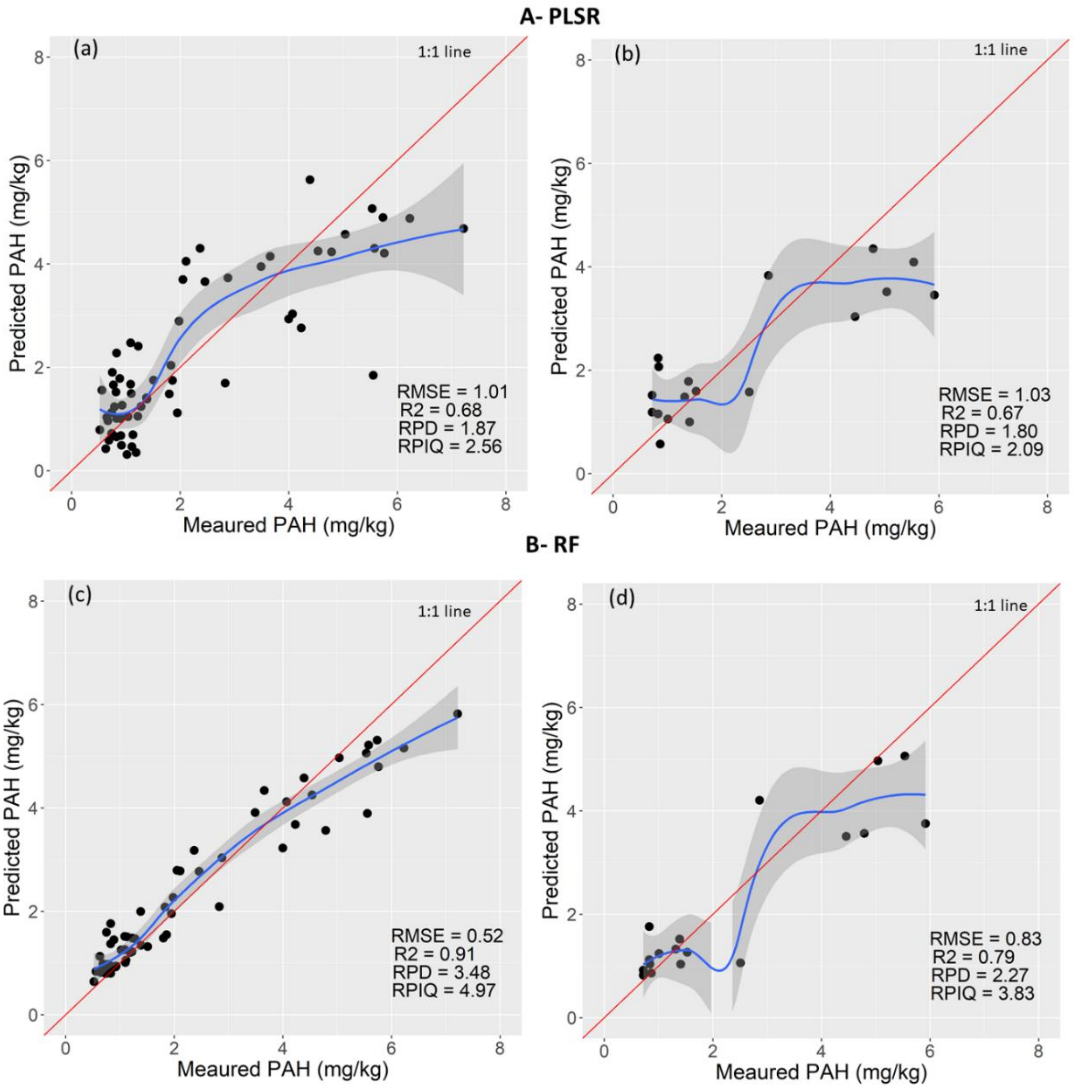


Figure 5-3: Scatter plots of sequential ultrasonic solvent extraction gas chromatography (SUSE-GC) of measured polycyclic aromatic hydrocarbon (PAH) *versus* mid-infrared (MIR) spectroscopy predicted concentations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and (B) random forest (RF) method in (c) cross-validation and (d) prediction. The blue lines and grey areas represent the regression line and 95% confidence interval, respectively.

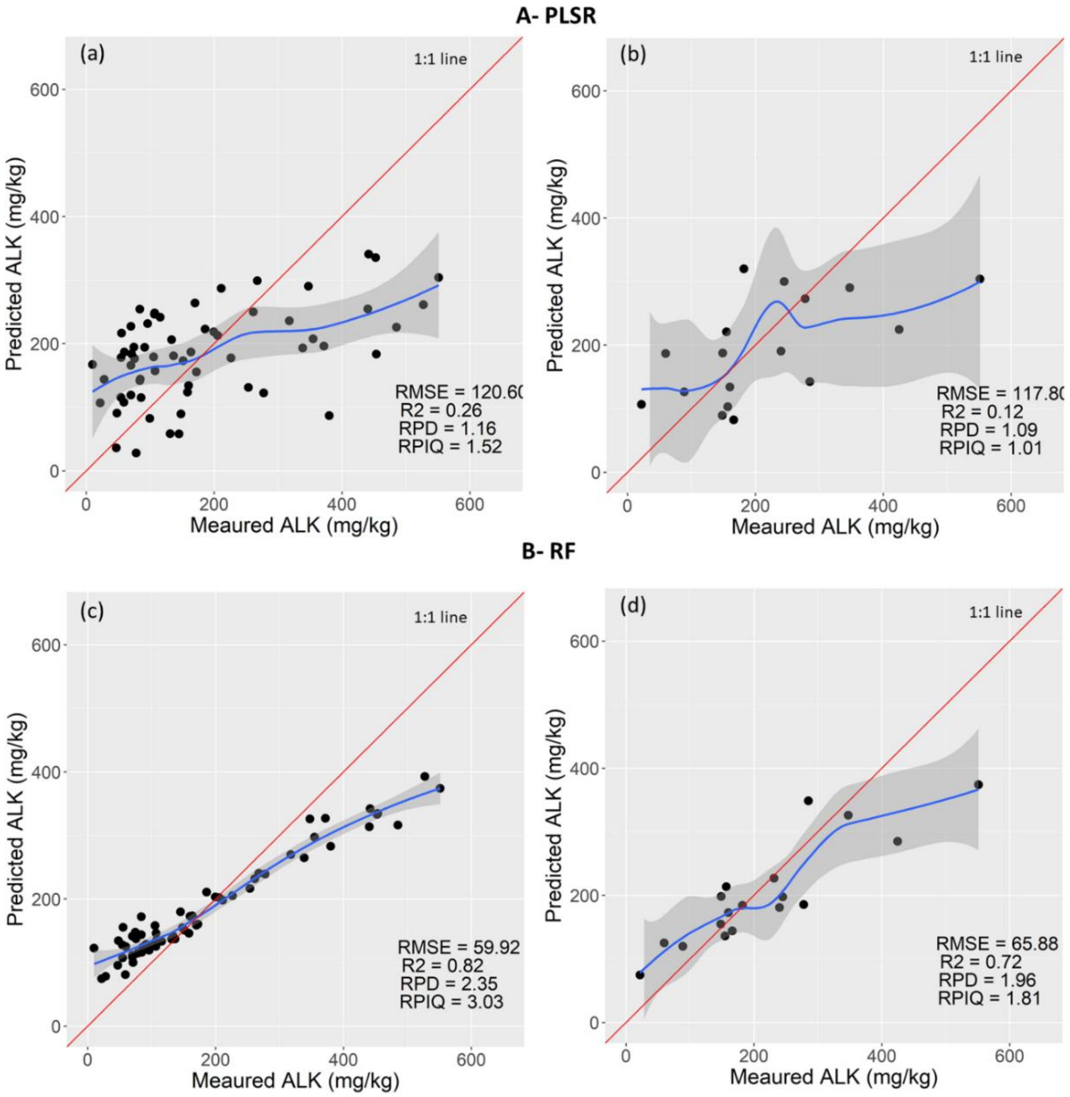


Figure 5-4: **Scatter plots of sequential ultrasonic solvent extraction gas chromatography (SUSE-GC) measured alkanes *versus* mid-infrared (MIR) spectroscopy predicted concentations based on (A) partial least squares regression (PLSR) in (a) cross-validation and (b) prediction, and (B) random forest (RF) method in (c) cross-validation and (d) prediction. The blue lines and grey areas represent the regression line and 95% confidence interval, respectively.**

5.6 Conclusion

This study investigated the potential of a handheld mid infrared (MIR) spectrophotometer for the measurement of total petroleum hydrocarbon (TPH), polycyclic aromatic hydrocarbon (PAH) and alkanes in soils. MIR soil spectra were acquired from eighty five (n=85) petroleum-contaminated soil samples 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 for TPH, PAH and alkanes was compared using an independent prediction dataset. 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); PAH (R^2 = 0.79, RPD = 2.27, RMSEP = 0.83 mg/kg) and alkanes (R^2 = 0.72, RPD = 1.96, RMSEP = 65.88 mg/kg). Results also showed that MIR spectroscopy performs better than visible and near infrared spectroscopy- based on previously published work on the same samples. Results obtained herein suggest RF-MIR spectroscopy as a good approach for the analysis of TPH, PAH and alkanes in soils based on a limited number of 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. Further work on the development of Nigerian soil-spectral library is needed to support the assessment of soil PHC contamination variability in the numerous contaminated sites in the Niger Delta, Nigeria.

5.7 References

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.

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*. In press.

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 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.

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

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.

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.

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.

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 http://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., 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., 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., 2010. Personal memories of the early PLS development. Chemometrics and Intelligent Laboratory Systems 58, 83–84.

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*, D4015003-1-10.

CHAPTER 6 : General conclusions, research implications and recommendations for future work

6.1 Introduction

The overarching aim of this PhD research was to implement spectroscopy as rapid measurement tool (RMT) to support risk assessment and/or remediation in petroleum contaminated sites in the Niger Delta, Nigeria. To achieve this aim, laboratory-and field-based studies were designed to assess the detection and sensitivity of spectroscopy to both aliphatic and aromatic hydrocarbon fractions in different soils contaminated with crude oil products. The ability of two vis-NR spectroscopy (wavelength 350-2500, 305-2200 nm) to differentiate fresh from weathered oil contaminated soils was investigated (Chapter 3). Translation to real-time field application was assessed using vis-NIR (wavelength 350-2500 nm) and MIR (wavenumber 4000-650 cm^{-1}) portable spectrometers on genuine contaminated soils with crude oil in the Niger Delta, Nigeria (Chapter 4 and 5, respectively). The original concept of the field-scale study was to acquire diffuse reflectance spectra *on-site* but due to logistics of transporting the equipment to Nigeria, soil samples were collected and shipped to Cranfield University for spectral measurement.

6.2 Overview of the key findings and contribution to knowledge

Recognising the UNEP report of 2011 and other studies (Sam et al., 2017; Ambitunni et al., 2014; Ite et al., 2013; Kadafa, 2012) pertaining soil total petroleum hydrocarbon (TPH) contamination in the Niger Delta region of Nigeria, and the twin problems of funding and trained personnel in handling contaminated sites; this research underline the need for the establishment of RMT to rapidly diagnose contaminated sites and assist *on-site* site characterisation to inform contaminated land risk categorisation in the numerous petroleum contaminated sites in the region. Consequently, a comprehensive review of chromatographic and spectroscopic analytical techniques for rapid determination of hydrocarbons in soil and sediment matrices. The review critically discussed both laboratory and field measuring methods with their associated issues and pointed out solutions highlighting explicitly the pros and cons and research needs. For instance, the critical review shows that both the vis-NIR and MIR spectroscopy are influenced by soil moisture, however this bias can be minimised using spectroscopic pre-processing methods like first derivative which is independent of soil moisture content (Wu et al., 2009). Although both MIR and vis-NIR have been widely used for analysing petroleum hydrocarbons in soil and even been supplemented with X-ray fluorescence (XRF) there is to date no systematic studies that have looked into technology integration (combination of two or more sensors) and data fusion method to improve hydrocarbon prediction. This chapter proposed the first integrated analytical framework based on spectroscopic techniques integration and data fusion coupled with multivariate modelling to improve prediction accuracy and rapid measurement of petroleum hydrocarbons in soil and sediment. This chapter provides further recommendation on undertaking real-time field measurement using vis-NIR spectroscopy to increase the selection of techniques for hydrocarbon detection in soil.

Next, vis-NIR spectroscopy was utilised to successfully investigate and distinguish fresh from weathered oil contamination in soils and to quantify TPH concentrations. This was the first study of its kind. The sensitivity of two portable vis-NIR spectrophotometers, namely, ASD and tec5 with wavelength ranges of 350-2500 and 305-2200 nm, respectively to crude oil contamination in soil was assessed (Objective 2). The key novelty is that this study demonstrates vis-NIR spectroscopy (ASD with wavelength 350-2500 nm) as effective and sensitive tool to hydrocarbon concentration differences due to weathering as no study has previously investigated. This was achieved through the application of nonlinear random forest (RF) regression technique, enabling to account for nonlinearity of the soil spectral responses. The finding has created new avenue for scientists to explore the potential of vis-NIR spectroscopy for rapid soil hydrocarbon contamination assessment to support environmental and human health risks assessment *on-site* without involving soil sampling, tedious and time-consuming traditional laboratory analysis.

Principal component analysis (PCA) showed reasonable discrimination between the different soil groups with ASD spectral data only (Figure 3-3). This supports the qualitative separation achieved with PCA based separation of control, crude oil contaminated and heavy crude oil contaminated soil samples which is good agreement with Chakraborty et al. (2015). However, the instrument's sensitivity decreases over time as TPH levels in soil decreases due to weathering of hydrocarbons [TPH = 1761.5 mg kg^{-1} and 186.7 mg/kg for start (after 48 h) and end of experiment (after 24 month), respectively]. From spectral perspective, reflectance increased as weathering of hydrocarbons in soil continuous, while the control sample (no TPH) had the highest reflectance (Figure 3-2). The result supports the conclusion that soil absorbance increases with increasing oil concentrations (Okparanma and Mouazen, 2013) and conversely the average reflectance decreased in the contaminated soil in comparison to uncontaminated soil (Chakraborty et al., 2015). Furthermore, partial least squares regression (PLSR) and RF modelling techniques provide a quantitative separation of the oil-contaminated soil samples based on weathering pattern. This was the first study of its kind to qualitatively discriminate and quantitatively estimate hydrocarbon concentration differences in soil due to weathering by vis-NIR spectroscopy. Interestingly, the separation achieved with PLSR and RF was in agreement with the PCA; however, RF separates better than PLSR (Figure 3-6, Figure 3-7). This is because RF is a non-linear regression technique that handles the nonlinearity of the spectral response. Better quantitative estimation of TPH was obtained using RF-ASD calibration model than PLSR-ASD model. Therefore for effective soil TPH modelling, RF regression technique is recommended.

The performance of PLSR and RF modelling techniques to predict TPH, PAH and alkanes in genuine contaminated soils with crude oil using vis-NIR spectroscopy (Objective 3). Soil samples were collected from oil spill sites in the Niger Delta, Nigeria. Results showed again that RF modelling technique outperformed PLSR for all hydrocarbon groups including aliphatic and aromatic hydrocarbon compounds. The most striking outputs of the study are i) the good to fair prediction of TPH despite the limited dataset, and ii) the fact that no case study on hydrocarbon prediction has previously reported the superior performance of RF modelling method over PLSR.

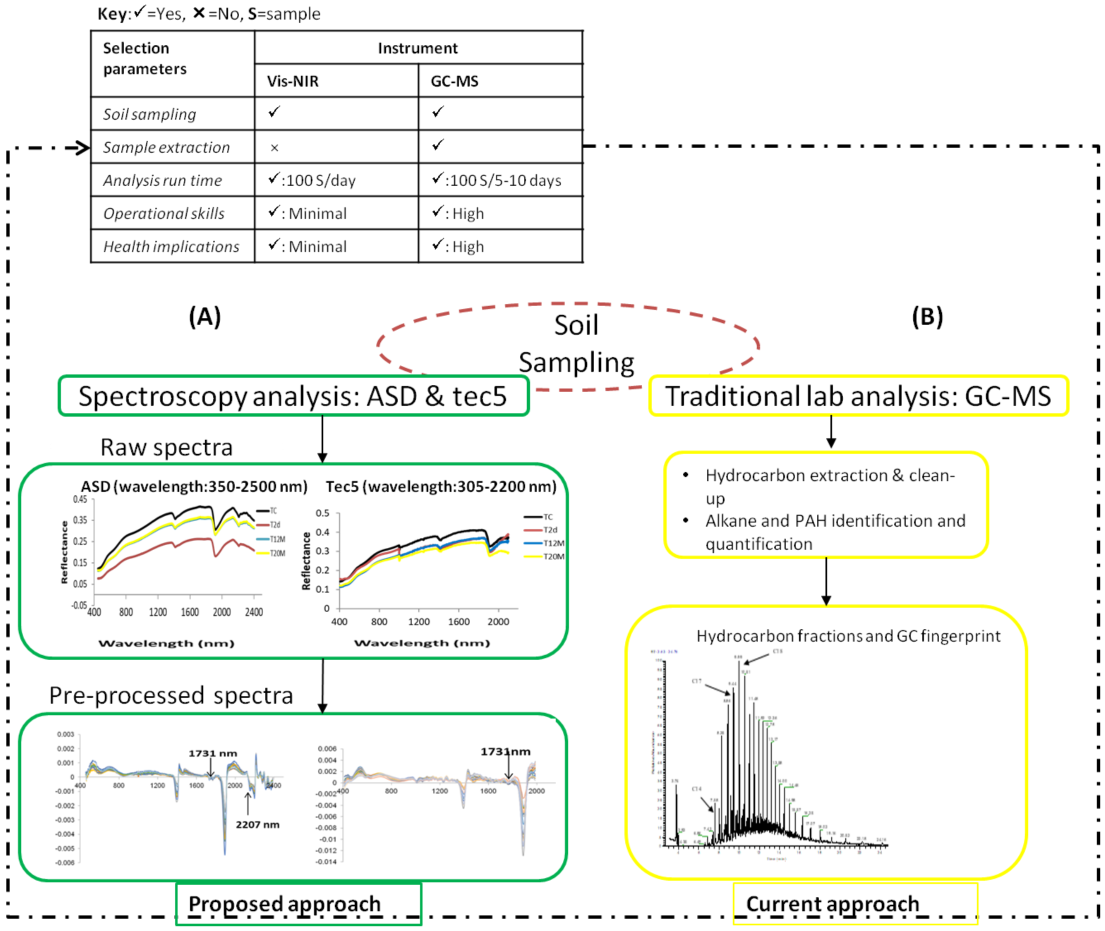
Good (RPD=1.99) and fair (1.55) predictions for PAH was achieved with RF and PLSR models, respectively while respectively, fair (RPD=1.59) and poor (RPD=1.29) alkanes prediction results were obtained using RF and PLSR models. This PhD research is the first to report on the assessment of aliphatic fractions (alkanes) using spectroscopy, however the model predictions were only fair to poor and this is likely due to the low concentration of alkanes. The alkanes calibration models developed in this study has open up the opportunity of exploiting diffuse reflectance spectroscopy to successfully predict the full range of hydrocarbons (TPH, PAH and alkanes) in soil, thus this is a significant advancement in the application of spectroscopy. Based on this study, vis-NIR and non-linear regression RF technique is strongly recommended for quantitative determination of soil hydrocarbons concentrations.

A portable MIR spectroscopy (wavenumber 4000-650 cm^{-1}) was evaluated for predicting the concentrations of TPH, PAH and alkanes in genuine contaminated soils with crude oil using non-linear RF and linear PLSR methods (Objective 4). The aim was to compare the predictive capability of the MIR and vis-NIR sensing techniques for predicting soil TPH, PAH and alkanes. Results showed that the MIR over-performed vis-NIR with RF modelling method performing better than PLSR for predicting TPH, PAH and alkanes. However, PLSR-vis-NIR produced slightly better results than PLSR-MIR in predicting TPH and alkanes. This may be attributed to the low concentration of alkanes used in the development of MIR models. Other reason could be the sensitivity of MIR to moisture content, which is larger than the sensitivity of vis-NIR spectroscopy. So far, MIR technique has not been used for the prediction of aliphatic (alkanes) fractions in contaminated soil. Thus, this study has filled the research gap identified in chapter 1. This chapter demonstrated that MIR spectral technique is superior to vis-NIR technique for detecting and quantifying rapidly oil contamination in soil and therefore being identified as a promising tool for field screening.

6.3 Spectroscopy approach for soil identification and prioritisation

The prediction results of the aliphatic and aromatic hydrocarbons in air-dried and field-moist soil samples in this study confirm that spectroscopy approach alone can be used for hydrocarbon contamination assessment in soil. The motivation of this approach is to provide rapid and cost-effective identification of hydrocarbon contaminated soil/site and support risk assessment and/or remediation plans. The proposed approach is presented in Figure 6-1, panel A. This approach would overcome much the cost and time-consuming efforts associated with conventional laboratory analysis of hydrocarbon contamination in soil. The guiding attributes in selecting spectroscopy as RMT over the commonly used analytical wet chemistry methods was based on: i) ease of use, ii) sample preparation, and iii) analysis run time. These attributes amongst others are presented in Figure 6-1. The use of vis-NIR and MIR spectroscopy for soil scanning require just little operational skills, whereas analytical techniques (e.g., laboratory-based GC) need expert operators. Similarly, little or no sample preparation is required when using either vis-NIR or MIR for soil contamination assessment while laboratory-based GC technique involves lengthy sample preparation protocols. Because GC-based techniques involve sample preparation and use extraction solvents (Risdon et al., 2008), there are concerns of occupational health and safety. Analysis run time refers to the total cycle time [time expended by the analytical system to go from one analysis to the other (Harris, 2003)] and the time used in preparing the sample for analysis. High analysis run time is associated with analytical techniques involving lengthy prior sample preparation protocols, potentially laboratory-based GC techniques. For example, the conventional GC analysis time (excluding sample preparation) is 30 minutes or even more (Barman et al., 2000). It is pertinent to also mention that vis-NIR or MIR spectroscopy analysis time is between 30-40 seconds. Thus, the shorter analysis run time of vis-NIR or MIR would position spectroscopy potentially better technique for cost-effective soil analysis, rapid investigation and decision-making at contaminated sites. Portability of the instrument is another factor for the selection of RMT techniques for field applications, as the selection of the best technique to measure the source of contamination is driven by time (Douglas et al., 2017). Portable versions of spectrophotometers are being previously developed and used globally for various applications including soil studies (Pimstein et al., 2011). Portability would enhance quick identification and risk mapping of hot spots of hydrocarbons for in-field applications.

Despite the above advantages of the vis-NIR over the laboratory-based GC techniques, vis-NIR results can be influenced by soil moisture content, soil types, ambient lights, etc.). Therefore, accounting for these external factors influencing the prediction performance is key for successful implementation of this sensing technology as a portable tool for field screening of petroleum hydrocarbons in soils. Also, it is pertinent to note that accuracy reported so far by different research groups show that spectroscopy detection methods are at a semi-quantitative stage, where more work is needed to improve the performance.



Key:✓=Yes, ✕=No, S=sample

Selection parameters	Instrument	
	Vis-NIR	GC-MS
Soil sampling	✓	✓
Sample extraction	×	✓
Analysis run time	✓:100 S/day	✓:100 S/5-10 days
Operational skills	✓: Minimal	✓: High
Health implications	✓: Minimal	✓: High

(A)
Soil Sampling
(B)
Spectroscopy analysis: ASD & tec5
Traditional lab analysis: GC-MS
Raw spectra
ASD (wavelength:350-2500 nm)
Tec5 (wavelength:305-2200 nm)
Reflectance
Wavelength (nm)
TC
T2d
T12M
T20M
Pre-processed spectra
1731 nm
2207 nm
1731nm
Hydrocarbon extraction & clean-up
Alkane and PAH identification and quantification
Hydrocarbon fractions and GC fingerprint
C17
C18
C14
Proposed approach
Current approach

Figure 6-1: Hydrocarbon contamination assessment flowchart: proposed approach (panel A) and current approach (panel B).

Previous studies have reported on the effect of soil factors potentially soil moisture content on vis-NIR and MIR results. For example, moisture content affects vis-NIR measurement (Malley et al., 1999). Similarly, MIR has been affected to a greater degree by soil moisture (Soriano-Disla et al., 2014; Hazel et al., 1997). A study by Horta et al. (2015) concluded that effect of moisture content on vis-NIR happens to cause more attenuation than soil structure. However, spectra pre-processing method (first derivative) has been reported to be independent of soil moisture content (Wu et al., 2009). Because spectral pre-processing takes few minutes only and even can be automated to be done in few seconds, the proposed approach could be most resource-effective; and would provide rapid real-time soil contamination assessment to support site-specific risk prioritisation at oil spill sites. In conclusion, there is room to optimise the proposed approach for enhanced investigation of soil contamination by hydrocarbons rather than the onerous analytical GC-based techniques.

6.4 Implications of the research

The Niger Delta region of Nigeria has faced serious petroleum hydrocarbon originated land contamination challenges due to the activities of oil and gas industries in the region for over five decades, which has posed significant threat to ecological systems and human health. Thus, this study evaluated RMT for identification of contaminated sites to facilitate site prioritisation and to inform risk assessment and remediation of contaminated sites in the region. This study will potentially impact on:

- Risk assessment and remediation experts in risk decision making and management of oil-contaminated sites in the Niger Delta, Nigeria, and can be used elsewhere in the World. This is because the implementation of RMT would enhance rapid diagnosis of oil spill sites (both fresh and weathered) to support risk classification procedures that by-pass time-consuming and costly wet chemistry analytical laboratory protocols. Since money and trained personnel are among the critical challenges facing the Nigerian Government to handle contaminated sites, there is no doubt that both the regulators (the Government) and the regulated (the oil operators) would benefit as the use of these techniques are cost-effective and did not require high operational skills.
- Diffuse reflectance spectroscopy can be used as a viable technique for the quantification of TPH, PAH and alkanes in soil science research, both for laboratory and field applications; thus it could replace time-consuming and expensive traditional wet chemistry methods for analysing petroleum hydrocarbons in soil.

6.5 Limitations of the research

During this PhD research, challenges encountered with funding and access to petroleum contaminated sites in the region limited the number of soil samples that could be collected and analysed. These factors may have affected the robustness of the models in predicting TPH, as a small dataset (85 samples) was used for developing the calibration models. In addition, mixing of soils from three different sites in the same calibrations might have affected the model performance. However, the results achieved were good in most of the case at the exception of the alkanes to draw reasonable conclusions.

6.6 Recommendations for future work

Considering the achievements and limitations of this PhD research, the following recommendations are made for future work:

- More work is needed in this area of research to cover large variability of TPH concentrations, soil types and properties in the Niger Delta, Nigeria.
- Diffuse reflectance spectral data should be acquired *on-site* with handheld equipment to identify contaminated sites and facilitate risk assessment and remediation for petroleum contaminated land sites in the Niger Delta region of Nigeria and elsewhere faced with similar challenges.
- Based on the better performance of nonlinear RF modelling method over the linear PLSR approach in this study, RF modelling is recommended for the prediction of soil TPH, PAH and alkanes instead of the commonly used PLSR modelling method. Among the nonlinear modelling techniques, future study should compare RF with other non-linear modelling techniques including support vector machine (SVM), artificial neural network (ANN), and penalised spline regression (PSR) to select the best technique for the prediction of TPH, PAH and alkanes in contaminated soil.
- It is suggested to test the potential of multi-sensor and data fusion for more accurate predictions of soils contaminated with hydrocarbons. Sensing methods should include different combinations of XRF, vis-NIR and MIR spectroscopy.
- Diffuse reflectance spectroscopy approach (Figure 6-1, panel A) should be implemented in practice for cost-effective, rapid identification and risk assessment of petroleum contaminated sites in the Niger Delta region of Nigeria.

6.7 References

Ambituuni, A., Amezaga, J., Emeseh, E., 2014. Analysis of safety and environmental regulations for downstream petroleum industry operations in Nigeria: Problems and prospects. Environ. Dev. 9, 43–60. doi:http://dx.doi.org/10.1016/j.envdev.2013.12.002

Barman, B.N., Cebolla, V.L., Membrado, L., 2000. Chromatographic techniques for petroleum and related products. *Critical Reviews in Analytical Chemistry*, 30:75-120

Chakraborty, S., Weindorf, D.C., Li, B., Aldabaa, A.A.A., Ghosh, 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.

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.

Harris, C.M., 2003. Today's chemist at work. American Chemical Society, 33-38.

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.

Ite, A.E., Ibok, U.J., Ite, M.U., Petters, S.W., 2013. Petroleum Exploration and Production: Past and Present Environmental Issues in the Nigeria's Niger Delta. Am. J. Environ. Prot. 1, 78–90.

Kadafa, A.Y., 2012. Environmental Impacts of Oil Exploration and Exploitation in teh Niger Delta of Nigeria. Global Journal of Sci. Frontier Res. Envt. and Earth Sci. 12 (3).

Malley, D.F., Hunter, K.N., Webster, G.R.B., Malley, D.F., Hunter, K.N., Webster, G.R.B., Barrie, G.R., 1999. Analysis of Diesel Fuel Contamination in Soils by Near-Infrared Reflectance Spectrometry and Solid Phase Microextraction-Gas Chromatography. *Soil Sediment Contam.*, 8, 481–489.

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, Soil Pollut*., 224, 1539.

Pimstein, A., Notesco, G., Ben-Dor, E., 2011. Performance of Three Identical Spectrometers in Retrieving Soil Reflectance under Laboratory Conditions. *Soil Sci. Soc. Am. J.* 75:746-759.

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.

Sam, K., Coulon, F., Prpich, G., 2017. Management of petroleum hydrocarbon contaminated sites in Nigeria: Current challenges and future direction. Land Use Policy 64 (2017) 133-144.

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.

Wu, C.Y., Jacobson, A.R., Laba, M., Baveye, P.C., 2009. Alleviating moisture content effects on the visible near-infrared diffuse-reflectance sensing of soils. Soil Science, 174, 456-465.