University of Reading

Soil Characterization using Visible Near Infrared Diffuse Reflectance Spectroscopy (VNIR DRS)

A thesis submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy (PhD) in Environmental Science

School of Archeology, Geography and Environmental Science

Kofoworola Amudat Olatunde

October 2017

Declaration

I confirm that this is my own work and the use of all materials from other sources has been properly and fully acknowledged.

Signed: Kofoworola Amudat Olatunde

Date:

Abstract

Soil analysis for agriculture, pollution assessment/remediation or resource exploration requires rapid procedures that are reliable, fast and cheap. This study examined the potential of visible near infrared diffuse reflectance spectroscopy (VNIR DRS) for soil analysis with emphasis on soil organic carbon (SOC) and extractible total petroleum hydrocarbon (ETPH). Preliminary laboratory studies were conducted to determine if visible near infrared diffuse reflectance spectral data contain adequate information for characterising SOC and ETPH and to identify sensitive wavelength regions best suited for quantitative modeling. It was concluded that VNIR DR spectra contained adequate information to quantify both SOC and ETPH in soils. Modelling with the whole spectrum was also found to give better predictions compared to modelling with portions of the spectrum. A 76/24 dataset split pattern was identified as an optimal split, ensuring that significant proportions of datasets are used for both model calibration and testing. Soil is inherently heterogeneous varying in space. The prediction performances of VNIR DRS models were observed to reduce with an increase in the geographical size of soil collection sites indicating that the best calibration models will likely be generated from spectral data derived from local soils with similar geology. Performance of VNIR DRS for characterizing ETPH was affected by SOC content. Higher model performances were observed at low organic carbon content, though all models developed for the SOC range studied (0.94 – 26.5%) had good prediction qualities (RPD > 2). Similarly, model performances were also affected by the type of petroleum hydrocarbon product that had been used to contaminate soils. This study provides evidence that VNIR DRS can be an important analytical approach to soil analysis particularly when and where costs and time are limiting conditions.

Acknowledgements

The origin of my being comes from Allah, who has been my helper and my source of strength and to whom I am immensely grateful.

I would like to thank my supervisors Prof. Chris Collins and Dr Kevin White for their support, guidance, constructive criticisms and advice over the years.

I am also grateful to the management of the tertiary education trust fund (TETFUND) of the Federal Republic of Nigeria for providing the funds for this study.

Thank you to all the staff of the University of Reading who have assisted me through this project: Karen Gutteridge, Anne Dudley, Dr Chris Speed, Dr Geoff Warren, Elizabeth Wyeth and a special thank you to Alice Ughi for her endurance and determination at those times in the laboratory when instruments test our patience.

My profound gratitude also goes to my parents and siblings for their concern and support to my continuing success.

Finally, I will like to thank my darling husband, Mr I.B. Olatunde; children Royhan and Habeeb for their support, encouragement and understanding. I would never have been able to pull this through without you.

Contents

List of Figures

List of Tables

Chapter 1

1.1. Introduction

The soil is a precious natural resource which supports agriculture and underpins a healthy population. No two soils are the same and variations occur in time (as biological and chemical processes break down or combine compounds over time) and in space (both vertically and horizontally). Overall, the soil is a mixture of minerals, organic matter, gases, liquids and living organisms performing four important functions. It serves as a sink for carbon sequestration and as such helps to mitigate the effect of global warming. It is a habitat for hundreds of organisms, thereby functioning as a gene bank. It is a regulator of water movement in the landscape and also serves as an environmental filter of nutrients and pollutants that may leach into the environment (Kettler, 2016).

An accurate knowledge of soil components, including macronutrients (nitrogen, phosphorus, and potassium), is needed for efficient agricultural production, including site-specific crop management (SSCM), where soil amendment rates are adjusted spatially based on local requirements (Kim et al., 2006). Of particular importance is the organic carbon component of the soil, which improves the physical properties of soil by increasing both the cation exchange capacity (CEC) and water-holding capacity (Leeper and Uren, 1993). It also contributes to the structural stability of soils by helping to bind particles into aggregates and has been widely accepted as a major indicator of soil health. Apart from the contribution of soil organic carbon (SOC) to the physical quality of the soil, its presence also supplies plants with nutrients such as N, P, K, Ca, Mg and Na required for growth. In addition, the structure of micro and macro organisms within the soil is largely influenced by the quantity of SOC (Thiele-Brunh et al.,

2012) which acts as an energy stock for their survival, and the microbial community in turn, and is essential for carbon cycling.

Unfortunately, soils have been subjected to severe degradation and/or pollution due to anthropogenic activities. One particularly prevalent form of pollution of soils derives from petroleum hydrocarbon production, transport and usage, which has attendant adverse environmental and human health consequences. Petroleum hydrocarbon contamination of soils is particularly worrisome in developing countries such as Nigeria, Gabon and Mauritania, with inadequate preventive and remediative environmental policies, leading to large scale pollution that is often allowed to remain in the soils for extended periods before a clean-up or remediation procedure is carried out.

When petroleum hydrocarbon contamination in soil is left unattended, it clogs soil pores, thereby suffocating soil microbial community and consequently destroying soil quality by releasing toxic components into it. According to Tang et al. (2011), a total petroleum hydrocarbon content of 0.5% within the soil will inhibit the activity of luminescent bacteria while a 1.5% content is considered to be a critical toxic level for plants and earthworms. Given that petroleum hydrocarbons are potential soil contaminants and neurotoxins for humans and animals (Schwartz *et al*., 2009), there is need for sustainable remediation techniques to restore the soil to a healthy productive condition. This cannot, however, be achieved without an adequate characterization of the type and extent of hydrocarbon contamination.

Thus, analysis of soils for organic carbon content and extent of petroleum hydrocarbons in cases of contamination is of importance both for agricultural and remediation applications respectively. Most often, economic decisions on whether or not to plant, apply fertilizers and/or deploy

specific remediation procedures are usually made based on the results of soil analysis. Conventional soil analytical procedures rely on laboratory testing of dried and ground soils. Multiple soil cores are collected in the field and transported to the laboratory where they are dried. After drying, the samples are passed through a sieve to achieve the required consistency and finally tested for physical and/or chemical properties These methods are often laborious, costly, time consuming and inadequate when high spatial and temporal resolutions are required (Chakraborty *et al*., 2012) taking days for soil preparation, analysis and result reporting, with the possibility of inaccurate results due to equipment constraints and/or mishandling of samples. These analytical challenges have prompted an increase in demand for alternative effective methods of soil characterisation.

One of the emerging alternatives is the visible near - infrared diffuse reflectance spectroscopy (VNIR-DRS). Over the past two decades, research on the use of spectroscopy in soil science to quantify soil properties such as soil mineralogy, organic matter, plant nutrients and soil texture has increased (Brown *et al*., 2006; Wetterlind *et al*., 2008b) due to the fact that it permits rapid and cost-effective quantification as compared to traditional chemical analyses (Schwartz *et al*., 2009). By utilizing the electromagnetic spectrum, radiation reflected from soils can be regressed against measurements of soil components determined by conventional soil analysis. The constructed statistical models, if stable and strong, can be successfully utilized to quantify the total petroleum hydrocarbon content and other parameters, such as organic carbon and total soil carbon, from unknown soil samples.

Soil spectra generated in visible near - infrared region usually have complex absorption patterns due to the overlapping absorption of soil constituents and scattering effects caused by soil structure and other soil properties, such as moisture and colour (Stenberg, 2010). Specific data in

the spectra therefore require mathematical extraction before they can be correlated with soil properties. The difficulty in interpreting Vis-NIR spectra has, however, been overcome to a large extent by the use of advanced chemometrics and data-processing techniques (Pasquini, 2003).

Univariate descriptive tools such as mean, median, standard deviation, and Normal distribution can be useful, but they limit the interpretation and understanding of a process by only looking at one variable at a time, potentially leading to incorrect conclusions (CAMO, 2011). Multivariate analysis, however, makes it possible to find patterns and relationships between several variables simultaneously and, in addition, make it possible to predict the effect a change in one variable will have on other variables. Multivariate analysis also generally solves the problem of interference from compounds closely related to the target compound, thereby eliminating the need for selectivity (Naes *et al*., 2002). Hence, the efficient analysis of soil spectra requires the use of multivariate calibrations (Martens and Naes, 1989).

Common calibration methods for soil applications are based on linear regressions. They include stepwise multiple linear regression (SMLR) (Viscarra Rosell *et al*., 2006), principal component regression (PCR), and partial least squares regression (PLSR). SMLR is usually used due to the inadequacy of more conventional regression techniques such as multiple linear regression (MLR) to accommodate spectral data with large numbers of collinear variables and a lack of awareness among soil scientists of the existence of full spectrum data compression techniques such as PCR and PLSR. Both of these techniques (PCR and PLSR) are related and, in most situations, make similar predictions. They can also cope with data containing large numbers of predictor variables that are highly co-linear. However, PLSR gives better prediction results when there are fewer components than PCR (Naes *et al*., 2002).

1.2. Soil and Soil Analysis

The characterization of soil conditions and its spatial and temporal changes as a result of natural and anthropogenic activities are vital for managing of soil health for agricultural purposes. Soil analysis, an investigation into the condition of soil, be it physical observation or extensive laboratory procedures, have been used for many years to assess and manage soil fertility.

Soil nutrient assessment is an important aspect of precision agriculture. This ensures that current soil composition with respect to variable properties (e.g., pH, organic matter, and soil nutrient levels), field (e.g., slope and elevation) and crop parameters (e.g., yield and biomass) are managed in such a way as to optimize inputs such as fertilizers and herbicides (Whelan and Taylor, 2013). Due to the effects of farm inputs, such as fertilizers, on soil and ground water quality, soil analysis have been widely used to determine where insufficient or excess nutrient levels occur (Hengl et al., 2017; Ilori et al., 2014; Zeng et al., 2012).

Declining soil health accentuated by intensive farming, erosion and/or resource exploration causes environmental and economic problems. Assessment of physical and soil chemical attributes can therefore serve as tracers to monitor such activities. However, soil characterization, especially for large scale monitoring, is usually limited due to the costly and time consuming intensive manual sampling and preparation associated with conventional soil analysis (Kim et al., 2006). In more recent times, reflectance spectroscopy has been proposed as a time and cost effective tool for mapping chemical and physical attributes of the soil. These soil attributes can be determined based on diagnostic absorption characteristics inherent in a soils reflectance spectrum (Chabrillat et al., 2013)

1.2.1. Total Organic Carbon

Soil carbon is a generic name given to all forms of carbon in the soil. It is a major pool of the carbon cycle that is continuously being fed by biomass and non-biomass sources such as quartz and feldspars and clayey materials. Carbon that is fixed by plants is transferred to the soil through dead plant and animal materials. This dead organic matter creates a substrate which decomposes adding to the carbon content of the soil.

Soils contain carbon in both organic and inorganic forms. However, in most soils, the majority of carbon is held as soil organic carbon (SOC). This differs from soil organic matter (SOM) which is the totality of biomass such as tissues from dead plants and animals, materials less than 2 mm in size, and soil organisms present in the soil at various stages of decomposition. The term soil organic carbon is used to refer to carbon stored in organic constituents in the soil such as tissues from dead plants and animals, products from tissue decomposition and soil microbial biomass.

The organic matter content of the soil, of which carbon constitute a major part, serves as a chelate to a great number of nutrient cations and trace elements that are of importance to plant growth. It buffers soil from strong changes in pH, prevents nutrient leaching and is essential to organic acids that make minerals available to plants (Leu, 2007).

A number of methods have been used to determine total organic carbon in soils. The wet oxidation methods, such as the Walkley and Black method and the Mebius method are the most accessible, but have the disadvantage of being laborious, involving the use of toxic reagents (Pimentel et al., 2006). Auto analyzers allow for rapid and reliable determinations by dry combustion (Sato et al., 2014), however, the cost of each determination is higher than the cost of wet combustion methods (Segnini et al., 2008). Also, the gravimetric dry combustion method is

rarely used in routine soil organic carbon analysis, due to low sample throughput and difficulty in automation (Miyazawa *et al*., 2000). There is no consensus on the ideal or standard method for soil carbon analysis. Recovery efficiencies using these different methods depend on the type of substrate that comprises the organic matter (Conyers *et al*., 2011), hence making comparison of experimental results difficult. Each method also relates differently with soil texture and mineralogy (Zinn *et al.,* 2007).

1.2.2. Petroleum Hydrocarbon Contamination of Soils

Petroleum or crude oil is an oily liquid which occurs naturally and consists primarily of hydrocarbons, the remaining fraction consists of oxygen, nitrogen or sulphur and trace amount of organometallic compounds (NRC, 1985). It is found occasionally in springs or pools but usually obtained from beneath the earth's surface through the use of drilling wells. The physical and chemical compositions of petroleum vary markedly depending on the source. The colour of crude oil can be black, red, amber or brown.

Crude oil is a raw material for almost every industry, either as fuel or a component in the manufacture of medicinal drugs, plastics, detergents, synthetic rubbers as well as many solvents (Speight, 1990). In the United Kingdom, an average of 1651.25 thousand barrels day^{-1} of crude oil was consumed between 2008 and 2011. In the energy sector, fuel oils from petroleum account for a large proportion of raw materials, with an average of 60,909 thousand tonnes of crude oil day^{-1}, accounting for 44 % of total fuels used in the generation of energy in 2011 (DECC, 2013). Due to this large dependence and consequent exploration of petroleum, release of contaminants into the environment during petroleum extraction, refinement, and transportation, is a common

problem. Petroleum may also get into the environment through natural sources such as earthquakes and geological seepage from the ocean floor. Other avoidable causes of petroleum spills in the environment include vandalism/sabotage, theft, illegal dumping and equipment failure. It is estimated that more than 9 million gallons of crude oil have spilled in the United States since 2010 (Harrington, 2016). Similarly, the Niger Delta region of Nigeria has a devastating history of oil spills, with petroleum hydrocarbons in soil at depths of at least 5 m (UNEP, 2011). Sadly, there is a lack of systematic scientific information about how much spillage has occurred over time. A 2014 report on petroleum contamination within the Niger Delta region of Nigeria by the Shell Petroleum Development Company of Nigeria Limited records as much as $1000m^3$ of crude oil spilled in one month, with sabotage being the main cause of spills (figure 1.1).

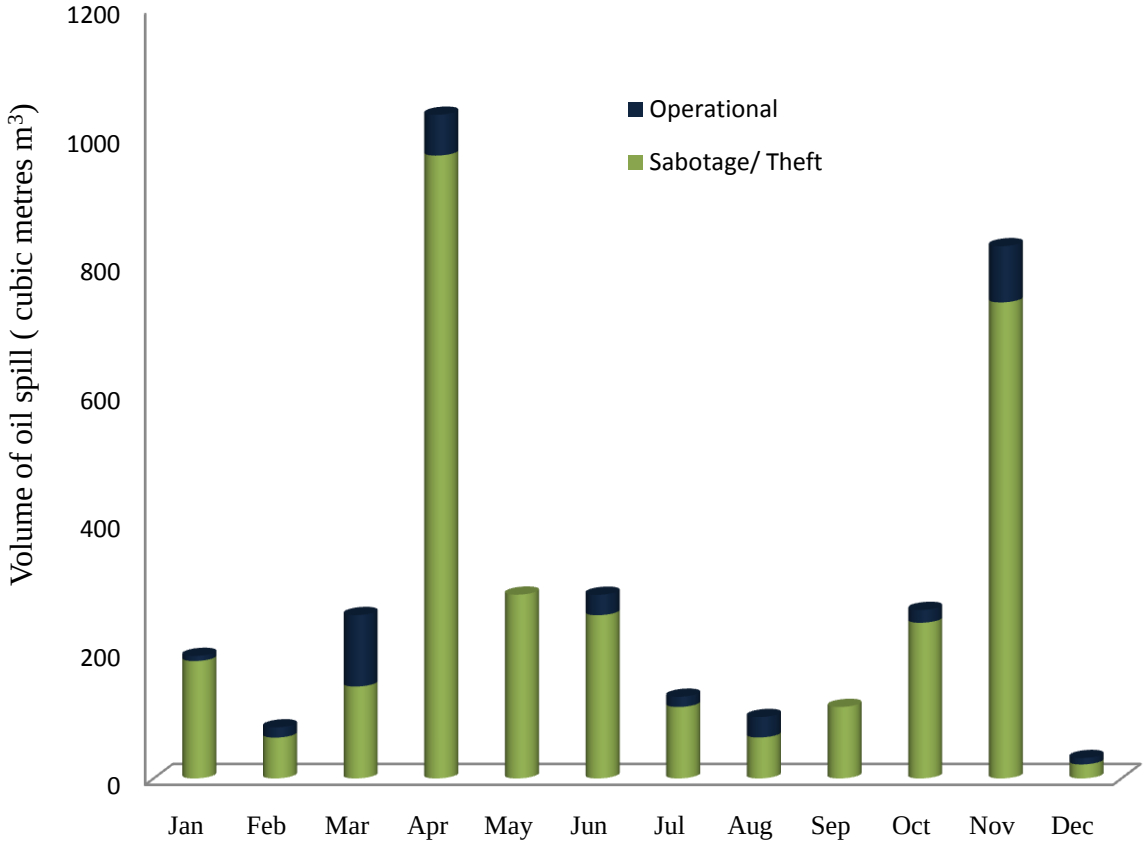


Figure 1.1: Volume of crude oil spilled within the Niger Delta region, Nigeria in 2014 (SPDC, 2014)

The use of petroleum for such a broad spectrum of activities has led to the release of petroleum products into the environment through accidental spills, long-term leakages, operational failures, broken oil well service lines, leaking storage tanks or crumbling infrastructure, and underground gasoline storage tanks at local fuel stations. In some cases, where oil production is occurring concurrently with crop production, agricultural soils are affected (Ite *et al*., 2013), but in other instances, the contamination may take place in wildlife refuges or national parks (Chakraborty *et al*., 2012). Consequently, many soil and water environments are contaminated with petroleum hydrocarbons.

The toxicity of petroleum hydrocarbon compounds in the environment have been extensively studied (Kharaka and Dorsey, 2005; Ugochukwu and Ertel, 2008). Soil petroleum contamination endangers local and regional ecological systems, food chains, and even creates the risk of explosion in urban areas (Fine et al., 1997). Release of petroleum hydrocarbons during exploration and transportation of petroleum and its products cause damage to terrestrial and aquatic environments. Aliphatic hydrocarbons, when released into the soil, cause changes in soil microbial population and physical properties (Kadafa, 2012). Petroleum hydrocarbon contamination of soils also causes displacement of soil oxygen, leading to asphyxiation of soil microorganisms, a limiting factor to soil fertility and hence crop productivity (Onwurah *et al*., 2007).

The persistence of some petroleum hydrocarbons in the environment is a matter of significant public, scientific and regulatory concern because of their potential toxicity, carcinogenicity and ability to bioconcentrate up the trophic ladder (Marmiroli et al., 2006). Continued presence of petroleum hydrocarbons may cause phytotoxicity leading to crop failure. If this does not occur, and the crops are able to acclimatize, bioaccumulation of the pollutants to dangerous levels may

occur such that the consuming public is at risk. The continued presence of these recalcitrant pollutants in the soil could also result into groundwater contamination.

The effect of petroleum hydrocarbon contamination of soils on the economic wellbeing of a community cannot be overemphasized. Some past spills in the Niger Delta region of Nigeria have necessitated the complete relocation of some communities, loss of ancestral homes, pollution of fresh water, loss of forest and agricultural land, destruction of fishing grounds and reduction of fish population, destroying the major source of income for the people (Ukoli, 2005). Understanding the type and extent of pollution is a key aspect of pollution management. Techniques that can help delineate polluted sites include carefully collected samples of soil and rapid site characterization techniques. Rapid site characterization techniques allow for rapid definition of the physical and chemical characteristics of the contaminated site (WDNR, 2003). On-site, rapid techniques for analyzing total petroleum hydrocarbon is especially important where a response action is needed immediately after an accident to minimize major environmental damage, and also for urban development where delays to construction can be costly (Forrester *et al*., 2012).This helps to reduce costs associated with their management, and the likely impacts of future pollution incidents (Okparanma *et al*., 2014). For effective management of these compounds in the environment, knowledge of their concentration and distribution is vital.

1.2.2.1. Characterization of Petroleum Contaminated Soils

Total petroleum hydrocarbons (TPH) is a gross parameter commonly used to quantify environmental contamination originating from petroleum products such as fuels, oils, lubricants, waxes, and others (ESD, 1993). The conventional wet chemistry methods for quantifying TPH

in soil samples are based on extracting the contaminant from a known quantity of soil (Schwartz *et al*., 2012). The TPH level in the extracted solution is then determined by gravimetry, IR spectroscopy or GC measurement. Total petroleum hydrocarbon (TPH), a mixture of different hydrocarbons, is generally used as an indicator of petroleum contaminated soils.

Apart from being expensive, the wet chemistry methods may also give incorrect TPH readings due to the fact that extraction yields can be strongly matrix dependent, soil moisture content dependent, and the extraction method development and optimization can be quite complicated (Schwartz *et al*., 2012). Lastly, there is an added analytical procedure that involves silica gel clean-up to remove five to six-ring alkylated aromatics in the extraction mixture (Xie *et al*., 1999). This increases the time required for analysis. Measurement of petroleum hydrocarbons in contaminated soils using conventional methods also requires rigorous field sampling, making wide-scale quantitative assessment difficult (Dent and Young, 1981). Gas-chromatography based laboratory methods for total petroleum hydrocarbon (TPH) quantification lack field-portability (Forrester et al., 2010). In addition, a lack of standardized methods has resulted in high variability in TPH results across commercial laboratories (Malley *et al*., 1999). Hence, there is a strong need for an innovative, fast, environmentally friendly, and cheaper sensing technology to identify petroleum contaminated areas for remediation and to monitor restoration on an ongoing basis (Prince, 1993).

1.3. Spectroscopy and Soil Characterization

Spectroscopy, the interaction of electromagnetic radiation with matter is useful in many ways to determine both the identity of compounds and their concentration in mixtures. Environmental analysts have used visible and ultraviolet spectroscopic methods for years, in which common

colorimetric tests for properties of water, such as acidity, have been reduced to simple kit forms, using visual colour matching or hand-held portable colorimeters (Kebbekus and Mitra, 1998). Infrared spectroscopy is also being used in compound characterization and the development of long range IR sensors, while X-ray fluorescence is being used to determine the atomic composition of solid materials, having the advantage of operating on solids without prior dissolution (Kebbekus and Mitra, 1998).

The spectral reflectance of a soil is a cumulative property derived from the inherent spectral behaviour of the various heterogeneous components (minerals, organic matter, and water molecules) present in the soil. Soil parameters are estimated through either direct or indirect relationships of soil reflectance with the chemical, physical, and biological characteristics of the soil matrix (Chabrillat et al., 2013).

Across the electromagnetic spectrum, three regions have proved sensitive to soil properties. They include

1. The visible region (400 to 780nm). This region provides information on soil colour, iron content and composition, soil water, and organic matter.
2. The near infrared (NIR) (780 to 2500 nm). This region provides information on phyllosilicates, most sorosilicates, hydroxides, some sulphates, amphiboles, carbonates, soil water, and organic matter.
3. The mid-infrared (mid-IR) (2500 to 25000 nm). This spectral region provides information on quartz, feldspars, silicate minerals, mafic, clay, carbonate mineral group, and organic compounds (Chabrillat *et al*., 2013).

1.3.1. Visible Near Infrared Diffuse Reflectance Spectroscopy (VNIR DRS)

Reflectance spectroscopy is commonly applied for quantitative determination in many disciplines. When radiation is incident upon a sample, it may be reflected, absorbed or transmitted, and the relative contribution of each phenomenon depends on the chemical composition and physical characteristics of the sample (Nicola et al., 2007). Reflection is due to three different phenomena. Specular reflection occurs on glossed surfaces, external diffuse reflection is induced by rough surfaces, whereas internal reflection occurs when infrared radiation enters an attenuated total reflectance crystal made of a highly refractive infrared transmitting material.

In diffuse reflectance spectroscopy, incident radiation is focused onto a rough or irregular surface material, such as a powder, fibre or granules. The diffusely scattered light is then collected and measured using a spectrometer that is optimized to increase collection of the diffuse reflectance component and decrease the specular component. Figure 1.2 shows a pictorial representation of both specular and diffuse reflectances from a surface.

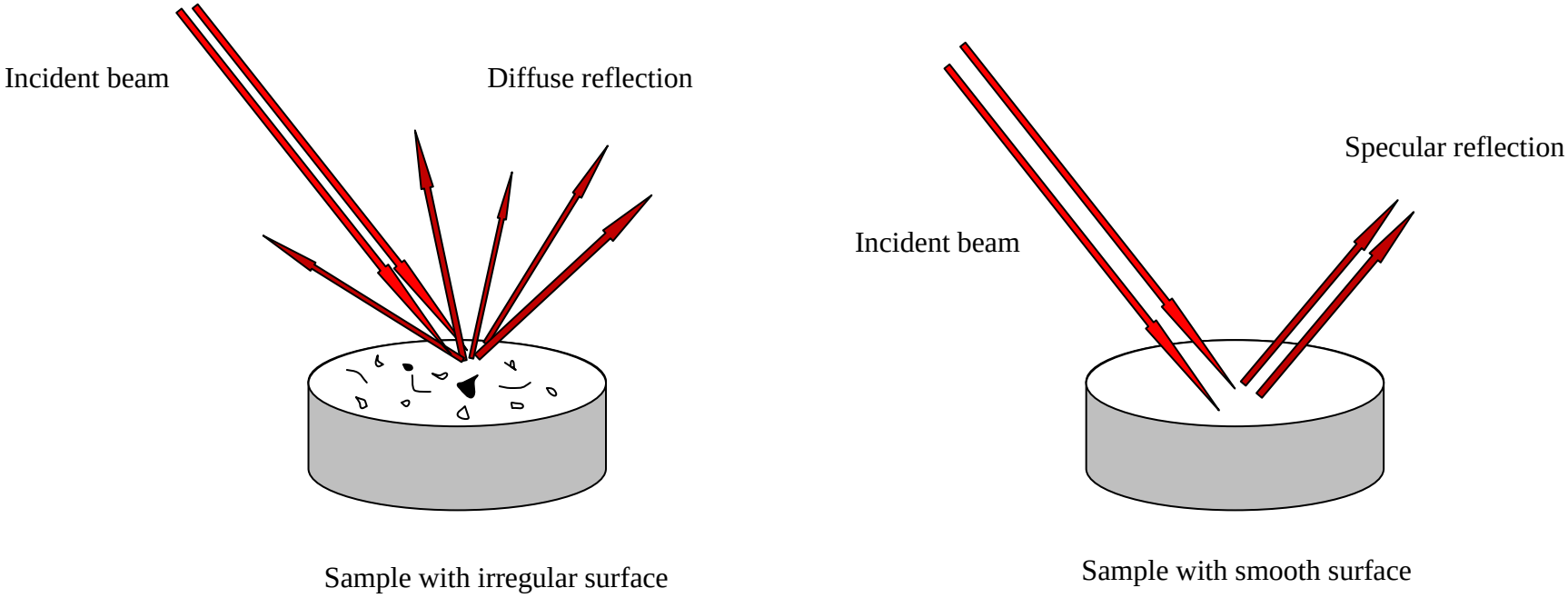


Figure 1.2: Pictorial representation of diffuse and specular reflectances

To generate a soil spectrum, radiation containing all relevant frequencies is directed onto a soil sample. Depending on the constituents present in the soil, the radiation will cause individual molecular bonds to vibrate, either by bending or stretching, and they will absorb a portion of light with a specific energy corresponding to the difference between two energy levels. The remaining radiation, reflected back into space and read by a spectrophotometer produces a spectrum with a characteristic shape that can be used for analytical purposes (Miller, 2001).

Visible near infrared diffuse reflectance spectroscopy (VNIR DRS) involves measuring the reflected electromagnetic energy from the soil samples in the VIS-NIR region (350 -2500 nm), and modelling the spectral data against samples with known concentration levels. Information about the soil components hidden within the spectral data is extracted using multivariate statistical techniques, also called chemometrics (Schwartz *et al*., 2012). Measurements within the visible near infrared region is usually characterised by a small number of broad, overlapping and weak absorption features that are difficult to interpret. However, this region contains useful information on organic and inorganic materials in the soil (Stenberg et al., 2010). Measurements within the VNIR are also particularly advantageous because the NIR radiation can typically penetrate deeper into a sample than mid infrared radiation (Nagy and Konya, 2009).

According to Reeves *et al*. (2002) the predictive accuracy of VNIR DRS in soil analytical studies reported in a number of studies suggests that it might replace standard laboratory methods for some applications (soil organic C, inorganic C, dithionate–citrate extractable Fe, cation exchange capacity, total N, 1.5 MPa water, basal respiration rate, sand, silt, and Mehlich III extractable Ca) (Chang et al., 2001; Brown et al., 2005, 2006; Stenberg et al., 2010).

1.3.1.1. Soil Organic Carbon

Soil organic carbon is the most frequently estimated soil property obtained from visible near infrared diffuse reflectance calibrations (Stenberg, 2010). Due to the complex nature of soil organic matter, absorption features associated with organic carbon occur at a wide range of wavelengths with fundamental absorptions occurring in the mid infrared and their overtones and combinations occurring in the visible near infrared. Bands around 1100, 1600, 1800, 2000, 2200 and 2400nm have been identified as being important for both soil organic carbon and total soil nitrogen calibrations (Stenberg, 2010; Martin et al., 2002).

It has been suggested that the visible near infrared interacts better with soil organic carbon than the infrared region alone and, therefore, gives a better prediction when used in calibrating models (Viscarra Rossel et al., 2006). This can be due to the fact that absorptions attributed to organic carbon in the visible region are clearer and easily identified when compared to weak and hidden absorptions in the infrared region. In addition, the organic carbon content of soil influences its colour which is an important factor for prediction in the visible region (Udelhoven et al., 2003). Including the visible region (350-780nm) in the calibrations for soil organic carbon have resulted into better results for Australian soils (Islam et al., 2003), Norwegian soils (Fystro, 2002), US land resource areas (Chang et al., 2001) and South eastern Australian soil(Dunn et al., 2002). However, other soil properties, such as moisture, texture and mineralogy, may also influence the colour of soils, implying that colour may only be useful as a discriminator within a limited region of sampling.

1.3.1.2. Total Petroleum Hydrocarbons

The spectral properties of petroleum hydrocarbons were identified at the late 1980s, though it was argued that these properties are only visible at concentrations of 4% wt and above

(Cloutis, 1989). Schwartz *et al*. (2012) compared diffuse reflectance spectroscopy as a tool for TPH assessment with results from three commercial certified laboratories using traditional methods. The reflectance spectroscopic method was found to be as good as the commercial laboratories in terms of accuracy. In addition, large variations were found between the results of the three commercial laboratories, both internally and between laboratories, emphasizing the need for an alternative rapid, environmentally friendly and cost effective screening tool for soil analysis that is reproducible.

Over the last two decades, the main focus of various studies on visible near infrared diffuse reflectance spectroscopy of soils have been on basic soil composition, such as soil organic matter (SOM), texture, clay mineralogy, soil fertility, structure, microbial activity. Advantages of using visible near infrared diffuse reflectance spectroscopy for soil analysis include

1. There is no need for sample preparation, such as drying and crushing, and therefore the samples are not affected in any way as compared to the destructive sampling of conventional soil analysis
2. No hazardous chemicals are required
3. Measurements takes a few seconds and results are available for prompt decision making
4. Several soil properties can be estimated from a single scan
5. The approach can be used both in the laboratory and *in-situ* (Viscarra Rossel et al., 2006)

1.3.2. Spectral Pre-processing

Spectral pre-processing involves the use of certain techniques to remove any irrelevant information present within spectra which affects the extraction of information present within the spectra or cannot be handled properly by regression techniques. Issues associated with spectra requiring pre-processing include noise, variations in pathlength, poor spectral resolution of equipment and overlap of individual spectra for the constituents in the sample (Naes *et al*., 2002).

1.3.2.1. Averaging

Averaging spectra is usually carried out during the acquisition of the spectrum to reduce the thermal noise of the detector. Averaging wavelengths is also done to reduce the number of wavelengths or to smooth the spectrum. Modern spectrophotometers typically have an optical resolution finer than 10 nm. While this certainly increases statistical processing time it does not necessarily improve the information content of the spectra for studies measuring the relatively broad absorption bands of soil constituents (Nicolai *et al*., 2007).

1.3.2.2. Smoothing

Smoothing techniques, such as moving average filters and the Savitzky-Golay algorithm, are employed to remove random noise from spectra (Naes *et al*., 2002). There have been reservations as to the use of smoothing techniques even though it obviously improves the visual aspect of spectrum. Smoothing can remove fine resolution information at a stage where it is not clear yet whether this information is useful, and since the multivariate regression techniques typically used for calibrations already incorporate an explicit model for additive noise, the need for smoothing have been questioned (Nicolai *et al*., 2007).

1.3.2.3. Normalisation

Normalization is used to compensate for additive (baseline shift) and multiplicative (tilt) effects in the spectral data, which are induced by physical effects, such as non-uniform surface scattering, which is dependent on the wavelength of the radiation, the particle size and the refractive index. Among the most common Normalization methods offered by chemometric software, multiplicative scatter correction (MSC) is the most popularly used technique (Naes *et al*., 2002). The technique also attempts to remove the effects of scattering by linearising each spectrum to the average spectrum of the sample (Nicolai *et al*., 2007).

1.3.2.4. Derivatisation

Derivatives of reflectance spectra are usually calculated by difference to perform baseline corrections and remove weak signals. Second derivatives are mostly used as they improve spectral resolution and the peak positions are same as the original (Chakraborty *et al*., 2012). During processing, caution should be exercised to ensure that the parameters of the algorithm (interval width, polynomial order) are carefully selected to avoid amplification of spectral noise (Nicolai *et al*., 2007).

1.3.2.5. Log Transformation

Reflectance measurements are frequently converted to Log_{10} (1/R), which are then used in a similar way to optical density readings (Nicolai *et al*., 2007). This approach has been widely used to correct for the non-linearity within spectra when measuring transmittance or reflectance. One other transformation used to account for nonlinearity within spectra is the Kubelka–Munck transformation $(1–R)^2/ 2R$.

1.3.2.6. Standard Normal Variate (SNV)

The standard normal variate pre-processing tool removes scatter effects by centering and scaling each individual spectrum. It is sometimes used in combination with de-trending (DT) to reduce multicollinearity, baseline shift and curvature within spectroscopic data especially diffuse reflectance spectra. It also removes multiplicative interferences of scatter and particle size effects from spectral data (CAMO, 2011).

1.3.2.7. Detrending (DT)

Detrending (DT) is used *t*o remove nonlinear trends in spectroscopic data. I*t is often used in combination with the s*tandard normal variate (SNV) *to* reduce multicollinearity, baseline shift and curvature. The detrend *tool* calculates a baseline function as a least squares fit of a polynomial to the sample spectrum. As the polynomial order of the detrend increases, additional baseline effects are removed (CAMO, 2011).

1.3.3. Multivariate Calibrations

Multivariate calibration is an area of chemometrics that focuses on finding relationships between one set of measurements that are easy and cheap to acquire and another set of measurement that is either labour intensive or expensive to obtain. The aim is to find good relationships such that expensive measurements can be easily and rapidly predicted with high accuracy from the cheaper measurements. In routine use, multivariate predictions save time, money and eliminate error associated with the laboratory analyst.

Diffuse reflectance spectra of soils in the visible near infrared usually contain broad overlapping absorption features. These measurements are also affected by the problem of non-selectivity as no single wavelength provides sufficient information to describe the

analyte being measured (Naes et al., 2002). This characteristic lack of specificity is worsened by scattering effects caused by soil structure or certain constituents such as quartz (Sparks, 2010). All these effects result in complex absorption patterns that require specific statistical tools for data extraction. Over the last few decades, efforts have been made to extract relevant information from visible near infrared diffuse reflectance spectra of soils and predict soil properties. Most calibration methods used by researchers are based on linear regressions such as partial least squares (PLSR), stepwise multiple linear regression (SMLR) and principal component analysis (PCR). Data-mining techniques such as Multivariate Adaptive Regression Splines (MARS), neural networks (NN) and the boosted regression trees(BRT) have also been employed to identify and describe correlations between soil properties of interest (Vasques, 2008).

Chemometric and data-mining approaches are easy to use, requiring no extensive understanding of spectroscopy, and can be used to quickly produce predictive models given spectral datasets with associated laboratory data for calibration. Statistical modelling is, however, prone to over fitting, as with any high-dimensional statistical method. Cross-validation alone can often give over-optimistic estimates of model predictive accuracy (Dardenne *et al*., 2000). To overcome this, independent datasets are used to validate diffuse reflectance models using a randomly selected 25–33% sample of a given data set (Brown *et al*., 2006).

1.3.3.1. Multiple linear Regression

Multiple linear regression is an approach for modelling the relationship between a scalar dependent variable Y and more than one explanatory variables X_1 - X_n. If the regression modelling involves only one independent variable, it is referred to as a simple linear

regression. Linear regression models are often fitted using the least squares approach. Multiple linear regressions have the following properties:

- The number of X variables must be smaller than the number of samples
- In the case of collinearity among X variables, the b-coefficients are not reliable and the model is not stable
- It tends to over-fit when noisy data is used (Xin, 2009)

The ability of X variables to vary independently of each other is a crucial requirement for variables used as predictors in multiple linear regressions. When variables are correlated, as in spectral data, procedures are required in which hidden information in multiple X variables that best describes the variation in Y are used as a basis of the regression modelling. Therefore, multiple linear regressions may be of little use in chemometrics due to the collinear nature of spectral data.

1.3.3.2. Principal Component Analysis

Principal components analysis (PCA) is a data reduction technique usually applied to environmental data, where datasets may be large and difficult to interpret, and where complex inter-relationships between variables are difficult to identify and visualize. It is used to explain the variations and groupings within a data set using new variables (principal components) which have been calculated from linear combinations of the original variables. The first principal component, or factor, accounts for the greatest variability in the data, and there can be a number of new components each accounting for less data variability than the previous (Webster, 2001). Factor loadings are correlation coefficients between the original variables and the factors and are used to describe the processes that control data variability.

Factor scores indicate how strongly individual samples are associated with each of the factors, and thus can be used to investigate similarity between samples, where samples with a similar composition will have similar scores and may, therefore, have similar contaminant sources and/or behaviour (Reid and Spencer, 2009).

The use of PCA in the analysis of environmental data is varied and widespread. It has been applied to the study of heavy metal distribution in sediments (Spencer, 2002) and soil acidification in forest soils (Boruvka *et al.,* 2005). It has also been used to examine spatial variability of contaminants (Backe *et al*., 2004), discriminate between contaminant sources (Chakraborty *et al*., 2010; Kim *et al*., 2006; Mudge and Duce, 2005), discriminate between samples according to the variable soil types, organic carbon levels, oil grades, and oil concentrations (Chakraborty, 2012; Li *et al*., 2004; Christensen *et al*., 2004), and to identify key variables for environmental monitoring purposes (Carlon *et al*., 2001; Shin and Lam, 2001).

1.3.3.3. Partial Least Square Regression (PLSR)

Partial Least Square Regression is a chemometric technique that employs statistical rotations to overcome the problem of dimensionally correlated predictors. In principal component regressions, principal components are chosen so that they describe as much variation in the predictor variables as possible, irrespective of the strength of the relationship between the predictor and the response variable. However, in partial least square regressions, the order of the regressor channels (wavelengths in the case of spectral reflectance data) is ignored. This means that the same results will be obtained when the regressors are shuffled (Chakraborty. 2012). Like principal component regression (PCR), the partial least square (PLS) starts by generating linear combinations of the predictor variables. In PLS, variables that show a high

correlation with the response variables are given extra weight because they will be more effective in prediction. In this way, latent variables are derived that are highly correlated with the response variable and are able to explain the variation in the predictor variables (Miller and Miller, 2000). For soil characterization, Partial Least Squares (PLS) regression using the 1st derivatives of soil reflectance is commonly used to reduce high-dimensional spectral data obtained from near infrared and middle infrared detectors (Udelhoven *et al*., 2003). Studies by Hazel *et al*. (1997) demonstrated the adverse effect of moisture on relatively small sample sets and showed that it was possible to counter the negative effects of soil moisture by the use of PLS regression methods.

1.3.3.4. Support Vector Machine Regression (SVMR)

The support vector machine regression is a non parametric regression characterized by the usage of kernels, absence of local minima, and sparseness of the solution and capacity control. The performance of the SVM algorithm depends on a good setting of three meta-parameters: capacity parameter *C*, a free parameter ϵ that serves as a threshold, and the kernel parameters. This regression tool has an advantage of avoiding difficulties of using linear functions in the high dimensional feature space, as the optimization problem is transformed into dual convex quadratic programmes (Drucker et al., 1997)

1.4. Gaps in Research

Research on the use of the VNIRDRS for analysis of soil is on the increase. There is increasing evidence from these studies that the approach could be a potentially cheaper, faster, safer and efficient technique for soil analysis. However, there have been differences in the levels of accuracies reported by individual studies on analysis of both soil organic carbon and total petroleum hydrocarbons. There is need for additional research to advance this

approach to a mature analytical technique, hence an investigation into its potential for analyzing soil organic carbon and total petroleum hydrocarbons in soils.

There is relatively large body of work within the literature on the application of VNIRDRS for the analysis of soil organic carbon. These studies have been conducted using soils obtained within a defined area or region (i.e. soils expected to have similar reflectances). However, soil is inherently complex and soil composition varies in space. As such, spectral reflectance of soils collected from differing sites may vary despite having similar contents of organic carbon. It is, therefore, important to extend research on this technique to capture the complex variability of soils in an attempt to assess the spatial viability of this approach to soil analysis.

Spectral reflectances are accumulations of responses from the various soil components. They are also influenced by variations in the composition of these soil components, as some of them react similarly across the electromagnetic spectrum because they possess similar functional groups (as is with the case with petroleum hydrocarbons and organic carbon compounds). It should, therefore, be of interest as to whether diffuse reflectance models predict soil parameters based upon spectral absorption of such soil components or whether they are built indirectly upon correlations with other soil components that can also influence soil reflectance. Therefore, this study seeks to investigate how concentrations of soil organic carbon influence the predictive accuracies of models calibrated for total petroleum hydrocarbons. This study also seeks to further understand the effect petroleum hydrocarbon sources will have on the application of VNIR DRS to analyzing total petroleum hydrocarbons in soils.

1.5. Objectives of Study

The objectives of this study include

1. To investigate the potential of predicting total organic carbon (TOC) in soils using visible near infrared diffuse reflectance spectroscopy.
    a. Assess the spatial viability of VNIR DRS in analyzing TOC in soils
    b. Compare the predictive potentials of VNIR DRS to MIR in analyzing TOC in soils
2. To investigate the potential of predicting petroleum hydrocarbon contamination in soils using visible near infrared diffuse reflectance spectroscopy.
    a. Assess the effect of TOC concentrations on the performance of models developed for extractible total petroleum hydrocarbons (ETPH)
    b. Assess the effect of petroleum hydrocarbon contamination sources on the performance of models developed for ETPH
3. Study the performances of VNIR DRS regression models
    a. Investigate the effect of sample split patterns on the calibration and predictive accuracies of models
    b. Investigate the calibration and predictive accuracies of models generated using different regression tools
4. Propose the most appropriate approaches for the use of spectral analysis for SOC and TPH in soil samples.

References

ANDOR. (2013). *Diffuse reflectance spectroscopy and fluorescence spectroscopy as techniques for identification of lung cancers: application note.* Available at http://www.andor.com/learning-academy.

Backe, C., Cousins I.T. and Larsson P. (2004). PCB in soils and estimated soil–air exchange fluxes of selected PCA congeners in the south of Sweden. *Environmental Pollution* 128: 59-72.

Boru, G., Vantoai, T., Alves, J., Hua, D. and Knee, M. (2003). Response of soybean to oxygen deficiency and elevated root zone carbon dioxide concentration. *Annal of botany* 91:447-453

Boruvka, L., Vecek, O. and Jenlika, S. (2005). Principal component analysis as a tool to indicate the origin of potentially toxic elements in soil. *Geoderma* 128:289-300.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D. and Reinsch, T.G. (2006).Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* 132: 273-290.

CAMO 2011.*What is multivariate analysis: An introduction to the principles and common models used in multivariate data analysis*.

Carlon, C., Critto, A., Marcomini, A. and Nathanail, P. (2001). Risk based characterization of contaminated industrial site using mulivariate and geostatistical tools. *Environmental Pollution* 111: 417-427.

Chakraborty, S., Weindorf, D. C., Morgan, C. L. S., Ge, Y., Galbraith, J., Li, B. and Kahlon, C.S. (2010).Rapid identification of oil-contaminated soils using visible near-infrared diff use reflectance spectroscopy. *Journal of Environmental Quality* 39:1378-1387

Chakraborty, S., Weindorf, D. C., Zhu, Y., Li, B., Morgan, C., Ge, Y. and Galbraith, J. (2012). Spectral reflectance variability from soil physicochemical properties in oil contaminated soils. *Geoderma* 177–178: 80–89

Chang, C., Laird, D.A., Mausbach, M.J., and Hurburgh, C.R. (2001). Near-infrared reflectance spectroscopy: principal components regression analyses of soil properties. *Soil Science Society American Journal* 65:480-490

Christensen, J.H., Hansen, A.B., Tomasi, G., Mortensen, J. and Andersen, O. (2004). Integrated methodology for forensic oil spill identification. *Environmental Science and Technology* 38(10): 2912-2918.

Cloutis, E. A. (1989). Spectral reflectance properties of hydrocarbons: remote-sensing implications. *Science* 245(4914): 165–168

Conyers, M.K., Poile, G.J., Oaetes, A.A., Waters, D. and Chan, K.Y. (2011).Comparison of three carbon determination methods on naturally occurring substrates and the implication for the quantification of soil carbon. *Soil Research* 49: 27-33

Dardenne, P., Sinnaeve, G. and Baeten, V. (2000). Multivariate calibration and chemometrics for near infrared spectroscopy: which method? *Journal of Near Infrared Spectroscopy* 8 (4): 229–237.

Davis, A.M. and Grant, A. (1987). Review: Near infra-red analysis of food. *International Journal of Food Science and Technology* 22:191-207

DECC, 2013. Department of energy and climate change: energy consumption in the UK. Available at www.gov.uk/government/publications/energy-consumption-in-the-uk

Dent, A. and Young, A. (1981). *Soil survey and land evaluation*. George Allen & Unwin Publ., Boston, MA.

Drucker, Harris; Burges, Christopher J. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. (1997); "Support Vector Regression Machines", in *Advances in Neural Information Processing Systems 9, NIPS 1996*, 155–161, MIT Press.

Dunn, B.W., Beecher, H.G., Batten, G.D. and Ciavarella, S. (2002). The potential of near-infrared reflectance spectroscopy for soil analysis - a case study from the Riverine Plain of south-eastern Australia. *Australian Journal Experimental Agriculture* 42 (5): 607– 614.

Eilers, P.H.C., Marx, B.D., (1996). Flexible smoothing with B-spline and penalties (with comments and rejoinder). *Statistical Science* 11: 89–121.

Environmental Sciences Division. (1993). Use of Gross Parameters for Assessment of Hydrocarbon Contamination of Soils in Alberta, Oxford, UK.

Faber, N.K. (1999). Multivariate sensitivity for the interpretation of the effect of spectral pre-treatment methods on near-infrared calibration model predictions. *Annals of Chemistry* 71(3):557-65.

Forrester, S., Janik, L., McLaughlin, M. (2010). *An infrared spectroscopic test for total petroleum hydrocarbon (TPH) contamination in soils*. Proceedings of the 19th 452 world congress of soil science, soil solutions for a changing world, Brisbane, Australia, August 1–6, pp. 13–16.

Fine, P., Graber, E.R. and Yaron, B. (1997). Soil interactions with petroleum hydrocarbons: abiotic process. *Soil technology* (10):133-153

Fystro, G. (2002). The prediction of C and N content and their potential mineralisation in heterogeneous soil samples using Vis-NIR spectroscopy and comparative methods. *Plant Soil* 246:139–149

Harrington R. (2016). *Here's how much oil has spilled from US pipelines since 2010*. Business insider, UK

Hazel, G., Bucholtz, F., Aggarwal, I.D., Nau, G. and. Ewing, K.J. (1997). Multivariate analysis of mid-IR FT-IR spectra of hydrocarbon-contaminated wet soils. *Applied Spectroscopy* 51:984–989.

Hengl, T., Leenaars, J., Shepherd, K. (2017). Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems* 109: 77 - 102.

Hensel, P. (2002) Oil Spills in Mangroves, National Oceanic and Atmospheric Administration. NOAA Ocean Service, Office of Response and Restoration. Washington

Ilori, E.G., Okonjo, P.N., Ojeh, E., Isiekwu, M.U. and Aondona, O. (2014). Assessment of soil nutrient status of an oil palm plantation. *Agricultural Journal* 9(2):127 - 131

Islam, K., Singh, B. and McBratney, A.(2003). Simultaneous estimation of several soil properties by ultra-violet, visible, and nearinfrared reflectance spectroscopy. *Australian Journal of Soil science Research* 41 (6): 1101-1114.

Ite, A.E., Ibok, U.J., Ite, M.U. and Petters, S.W. (2013). Petroleum Exploration and Production: Past and Present Environmental Issues in the Nigeria's Niger Delta. *American Journal of Environmental Protection* 1(4): 78-90

Kadafa, A.A. (2013). Environmental impacts of oil exploration and exploitation in the Niger Delta of Nigeria. *Global Journal of Science Frontier Research Environment & Earth Sciences* 12 (3)

Kharaka, Y. K., and Dorsey, N. S. (2005). Environmental issues of petroleum exploration and production: Introduction. *Environmental Geosciences*1 2 (2):61-63

Kim, M., Kennicutt I., and Qian, Y. (2006). Molecular and stable carbon isotopic characterization of PAH contaminants at McMurdo Station, Antarctica. *Marine Pollution Bulletin* 52: 1585-1590.

Kebbekus, B. B. and Mitra, S. (1998).*Environmental Chemical Analysis*. Blackier Academic & Professional. Pp 54

Kettler T. (2016). *Soil Genesis and Development*, Lesson 6 - Global Soil Resources and Distribution. Plant and soil sciences e-library

Leu A, (2007). *Increasing soil carbon, crop productivity and farm profitability*. The natural farmer. The newspaper of the northeast organic farming association.

Leeper, G.W. and Uren, N.C. (1993). *Soil Science: an introduction*. (Fifth Edition). Melbourne University Press, Carlton.

Leyval, C and Binet, P. (1998). Effect of polyaromatic hydrocarbons in soil on arbuscular mycorrhizal plants. *Journal of environmental quality* 27:402-407

Li, J.F., Fuller, S., Cattle,J.,Way, C.P. and Hibbert, D.B. (2004). Matching fluorescence spectra of oil spills with spectra from suspect sources. *Analytica Chimica Acta* 514(1): 51-56.

Malley, D.F., Hunter, K.N. and Webster, G.R.B. (1999). Analysis of diesel fuel contamination in soils by near-infrared reflectance spectrometry and solid phase micro extraction-gas chromatography. *J. Soil Contam.* 8:481-489.

Marmiroli N., Marmiroli, M. and Maestri, E. (2006). *Phytoremediation and phytotechnologies: A review for the present and the future*. In: Twardowska I, Allen HE, Haggblom MH (ed). Soil and water pollution monitoring, protection and remediation. Springer, Netherland

Martens, H., and Naes, T. (1989). Multivariate Calibration John Wiley & Sons, Chichester, UK. 419 pp

Martin, P. D., Malley, D. F., Manning, G., and Fuller, L. (2002). Determination of soil organic carbon and nitrogen at the field level using near-infrared spectroscopy. *Can. J. Soil. Sci*. 82:413-422.

Miller, C.E. (2001). *Chemical principles of near-infrared technology*. In: Williams P, Norris K (eds) Near-infrared technology in the agricultural and food industries, 2nd edn. American Association of Cereal Chemists Inc, Minesota, pp 19–37

Miller, J. N. and Miller, J. C. (2000). *Statistic and chemometrics for analytical chemistry* (4th ed.). Pearson Education, Harlow.

Miyazawa, M., Pavan, M.A., Oliveira, E.L., Ionashiro, M. andSilva,A.K. (2000). Gravimetric determination of soil organic matter. *Brazilian Archives of Biology and Technology* 43:475-478.

Mudge, S.M. and Duce, C.E. (2005). Identifying the source, transport path and sinks of sewage derived organic matter. *Environmental Pollution* 136: 209:220.

Naes, T., Isaksson, T., Fearn, T. and Davies, T. (2002). *A User-friendly Guide to Multivariate Calibration and Classification*. NIR publications, Charlton, Chichester, UK.

Miller, C.E. (2001).*Chemical principles of near-infrared technology*. In: Williams P, Norris K (eds) Near-infrared technology in the agricultural and food industries, 2nd edn. American Association of Cerial Chemists Inc, Minesota, pp 19–37

Konya, J. and Nagy, N. M. (2009). Isotherm equation of sorption of electrolyte solutions on solids: How to do heterogeneous surface from homogeneous one. *Periodica Polytechnica: Chemical Engineering*53(2):55-60

Nicolai, B.M., Beullens, K., ElsBobelyn, E., Peirs, A., Saeys, W., Theron, K. and Lammertyn, J. (2007). *Non-destructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review*. *Postharvest Biology and Technology* 46(2): 99-118.

National Research Council. (1985). *Oil in the Sea: Inputs, Fates, and* Effects. Washington, DC: The National Academies Press

Okparanma R. N. and Mouazen, A. M. (2013).Determination of total petroleum hydrocarbon (TPH) and polycyclic aromatic hydrocarbon (PAH) in soils: a review of spectroscopic and non-spectroscopic techniques. *Applied Spectroscopy Reviews*48 (6): 458-486

Okparanma, R.N., Coulon, F. and Mouazen, A.M.(2014). Analysis of petroleum-contaminated soils by diffuse reflectance spectroscopy and sequential ultrasonic solvent extraction-gas chromatography. *Environmental Pollution*184: 298-305

Onwurah, I.N.E., Ogugua, V.N.,Onyike, N.B., Ochonogor, A.E. and Otitoju, A.F. (2007). Crude oil spills in the environment: effects and some innovative clean up technologies. *International journal of environmental research* 1: 30-320

Pasquini, C. (2003). Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian chemical Society* 14(2): 198-219

Pimentel, D. (2006). Soil erosion: a food and environmental threat. *Environment, Development and Sustainability* 8: 119-137.

Prince, R.C. (1993). Petroleum spill bioremediation in marine environments. *Critical Reviews in Microbiology*19:217–242

Reid, M.K. and Spencer K.L. (2009). Use of principal components analysis (PCA) on estuarine sediment datasets: The effect of data pre-treatment. *Environmental Pollution* 157: 2275-2281

Reimann, C. and Filzmoser, P. (2000). Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology* 39:1001–1014.

Reeves J.B., McCarty, G. and Mimmo, T. (2002). The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soils. *Environ. Pollut.* 116: S277-S284.

Samarasekara, V.N. (1994). The impact of agriculture and industry on a wetland ecosystem: The case of Koggala Lagoon, Sri Lanka. *Coastal Manage Trop Asia* 3: 15-19

Schwartz. G., Eshel, G., Ben-Haim, M. andBen-Dor, E. (2009).Reflectance spectroscopy as a rapid tool for qualitative mapping and classification of hydrocarbons soil contamination. Tel Aviv, Israel, Available at http://www.earsel6th.tau.ac.il/~earsel6/CD/PDF/earsel-ROCEEDINGS/3080%20Schwartz.pdf (verified 1 June 2010).

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. *Applied and Environmental Soil Science* 2012, Article ID 751956.

Shell Petroleum Development Company (SPDC) of Nigeria 2014 data on oil spill. Available at http://www.shell.com.ng/sustainability/environment/oil-spills.html

Shin, P.K.S. and Lam, W.K.C. (2001). Development of a marine sediment pollution index. *Environmental Pollution* 113: 281-291.

Sparks, D. L. (2010). *Advances in agronomy*. Academic press. 232pp

Spencer, K.L. (2002). Spatial variability of metals in the inter-tidal sediments of the Medway Estuary, Kent, UK.*Marine Pollution Bulletin* 44: 933-944

Speight, J.G. (1990). *Fuel Science and Technology Handbook*. Chap. 3, Marcel Dekker, New York, 80.

Stenberg B., ViscarraRossel, R. A., Mouazen, A. M. and Wetterlind, J. (2010).*Visible and Near Infrared Spectroscopy in Soil Science*. In Donald L. Sparks, editor: Advances in Agronomy, Vol. 107, Burlington: Academic Press, 2010, pp. 163-215. http://dx.doi.org/10.1016/S0065-2113(10)07005-7

Tang J., Wang M., Wang F., Sun Q. and Zhou Q. (2011). Eco-toxicity of petroleum hydrocarbon contaminated soil.*J Environ Sci (China)* 23(5):845-51.

Thiele-Brunh, S., Bloem, J., de Vries, F. T., Kalbitz, K. andWagg, C. (2012).Linking soil biodiversity and agricultural soil management.*Environmental Sustainability*, 4: 523- 528.

Udelhoven, T., Emmerling, C. and Jarmer, T. (2003). Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: a feasibility study. *Plant Soil* 251 (2): 319-329.

Ugochukwu, C. N. and Ertel, J. (2008). Negative impacts of oil exploration on biodiversity management in the Niger Delta area of Nigeria. *Impact assessment and project appraisal* 26 (2). 139-147

Ukoli, M.K. (2005). Environmental factors in the management of the oil and gas industry in Nigeria.www.cenbank.org

United States Energy Information Administration. (2012). Petroleum Consumption in OECD Countries.

Vasques, G.M., Grunwald, S. and Sickman, J.O. (2008). Comparison of multivariate methods for inferential modelling of soil carbon using visible/ near-infrared spectra. *Geoderma* (146): 14-25.

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

Wisconsin Department of Natural Resources. (2003). Understanding chlorinated hydrocarbon behaviour in groundwater: Investigation, assessment and limitations of monitored natural attenuation.

Webster, R. (2001). Statistics to support soil research and their presentation. *European Journal of Soil Science* 52: 331-340.

Wetterland, J., Stenberg, B. and Soderstrom, M. (2008). The use of near infrared (NIR) spectroscopy to improve soil mapping at the farm scale. *Precision Agric* 9:57–69.

Whelan, B. and Taylor,J. (2013). *Precision agriculture for grain production systems*. Csiro publishing.

Xie, G., Barcelona,M. J. and Fang, J. (1999). Quantification and interpretation of total petroleum hydrocarbons in sediment samples by a GC/ MS method and comparison with EPA 418.1 and a rapid field method. *Analytical Chemistry* 71(9): 1899–1904

Xin Yan. (2009). *Linear Regression Analysis*: *Theory and computing*. World Scientific Publishing

Zhang, C. (2006). Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland. *Environmental Pollution* 142: 501-511.

Zheng, Z., He, X. and Li, T. (2012). Status and evaluation of the soil nutrients in tea plantation. *Procedia Environmental Sciences* 12(A): 45-51

Zinn, Y.L., Lal, R., Bigham, J.M. and Resck, D.V.S. (2007). Edaphic controls on soil organic carbon retention in the Brazilian Cerrado: soil texture and mineralogy. *Soil Science Society of America Journal* 71: 1204-1214.

Chapter 2. Method Development

2.1. Introduction

Developing models for future prediction of soil parameters from visible near infrared diffuse reflectance spectra require data from a reference analysis, albeit limited in number due to the costs implications. The calibration statistics are, therefore, dependent on reference measurements and can never be better than the quality of the chemical reference analyses (Wetterlind et al., 2013). This means that errors present in the reference chemical analysis to which the spectra are correlated will be included in the calibration model, reducing the potential of obtaining good predictions. The analytical methods used within this research are illustrated in table 2.1a, while table 2.1b details he softwares used for data analysis. This chapter describes how the extraction/GC-FID analyses for petroleum hydrocarbons were optimised prior to modelling studies. It also details the method for obtaining visible near infrared diffuse reflectance spectra of diesel contaminated soils to assess if they contain adequate information for characterising petroleum hydrocarbon contamination in soils. Lastly, it explores the potential of different wavebands within the visible near infrared diffuse reflectance spectral data, in terms of their performance in estimating petroleum hydrocarbon contamination.

Table 2.1a: Analytical methods

Measurement	Methods	Instrumentation
Extractible total petroleum hydrocarbons (ETPH)	Extraction/ GC-FID analysis	Gas Chromatograph/Flame ionization detector (GC-FID)
Soil organic carbon	Milling/ flash combustion	Organic elemental analyser (Thermo Scientific Flash 2000)
Spectral	Scanning	GER3700 V-SWIR spectrophotometer

Table 2.1b: Data analytical softwares

Procedure	Software
Principal component analysis (PCA), Multivariate regressions	Unscrambler-X 10.3 analytical software developed by CAMO.
Continuum removal on spectra	DISPEC software developed by NASA Jet Propulsion Laboratory.

2.2. Optimising Extraction/Cleanup/Gas Chromatography Procedure

2.2.1. Introduction

Petroleum hydrocarbon contamination in soils is conventionally analyzed using an approach based on a solvent extraction and silica gel solid-phase clean-up/fractionation process. Gas chromatographic (GC) analysis using a flame ionization detector (FID) is then used to identify and quantify both target aliphatic and polynuclear aromatic hydrocarbon analytes in the extract (MADEPH, 2009).

The term 'extractible total petroleum hydrocarbon (ETPH)' is used to describe the quantity of petroleum hydrocarbons that can be successfully extracted from a contaminated environmental sample. This definition is specifically used in contrast to total petroleum hydrocarbons (TPH), which is the totality of hydrocarbon compounds that can be attributed to all crude oil fractions, some of which may not be extractable from the matrix. The objective of this study was to optimise the accuracy of the extraction/gas chromatographic procedure for analysis of extractible total petroleum hydrocarbon.

2.2.2. Methodology

The MADEPH 2009 procedure is a set of Quality Assurance and Quality Control protocols for the analysis of petroleum hydrocarbons recommended by the Massachusetts Department of Environmental Protection, State Of Connecticut. It quantifies extractable aliphatic hydrocarbons within two specific ranges: C_9 through C_{18}, and C_{19} through C_{36}, while extractable aromatic hydrocarbons are collectively quantified within the C_{11} through C_{22} range. The sum of the concentrations in these three ranges gives the total extractible petroleum hydrocarbon. A modified version of the MADEPH 2009 procedure was optimized as follows.

a. Establishment of retention time windows and optimisation of GC-FID system operation

Serial injections of standard solutions of aliphatic and aromatic petroleum hydrocarbons were made into the GC-FID. The retention time window for each hydrocarbon analyte was calculated thus:

$$Rt\ window = Mean \pm 3SD$$

Gas chromatography analytical conditions are as follows:

Column:	30m × 0.25mm × 0.5µm nominal	
Oven temp:	Initial temp 60^0C, hold 1 min ramp to 290^0C at 8^0C/min hold 6.75 min	
Total run:	36.5 min	
Gas flows:	Carrier:	Helium at 2 mL/min
	Oxidiser:	Air at 400 mL/min
	Fuel:	Hydrogen at 35 mL/min
	Makeup:	Nitrogen at 45 mL/min
Injector temp:	285^0C	
Detector Temp:	315^0C	

b. Optimisation of silica gel solid-phase clean/fractionation process

High purity grade Silica gel (70 – 230 mesh) and anhydrous sodium sulphate were activated in separate beakers by adding analytical hexane to each, ensuring that it is submerged and then sonicating in a water bath for 30 minutes. The hexane was allowed to evaporate and the reagents were then heated overnight at 130^0C. 6ml SPE Cartridges were packed first with 1g

of activated silica gel and then 0.5g anhydrous sodium sulphate and a polyethylene frit (20 μm porosity) was placed at the top.

200ppm standard solutions of petroleum hydrocarbon analytes (14 aliphatic and 17 aromatic compounds) were prepared from certified reference materials using analytical grade hexane as solvent. 0.5ml of standard solutions was cleaned and fractionated by using variable volumes (2.5, 3, and 3.5) ml of hexane to elute the aliphatic hydrocarbon fraction while the aromatic fraction was then eluted by washing the column down slowly with 9ml of 3% isopropyl alcohol in hexane. The fractions were concentrated to 1ml under a gentle blow of nitrogen and transferred to 1.5 ml GC vials for analysis. The recoveries from the extraction for ETPH analytes were calculated thus.

$$\%\ Recovery = \frac{Concentration\ in\ extract}{Concentration\ of\ spiking\ solution} \times 100$$

c. Assessment of the Efficiency of Extraction Procedure

10 ml of an hexane: acetone (1:1) mixture was added to 1g each of a sand blank(50 – 70 mesh white quartz CRM) and an agricultural soil (pH= 6.2) both previously spiked with 200ppm petroleum hydrocarbon standard containing 14 aliphatic and 17 aromatic compounds. The mixture was sonicated for 15 minutes and shaken at 60 rpm for 60 minutes. 4 ml of deionised water was added and the vials were left in the freezer. The entire top layer was pipetted and the extract evaporated to 1 ml under a gentle blow of nitrogen. Extracts were transferred to GC vials and analysed.

2.2.3. Results and Discussion

Table 2.2 (a, b) illustrates the retention times of ETPH marker compounds comprising of 14 aliphatic hydrocarbons (C_9-C_{36}) and 16 aromatic hydrocarbons (C_{11}-C_{22}).The retention time for each analyte was observed to be affected by the matrix. Chromatograms of extracts derived from the sand blank and standards solutions had ETPH analytes appearing at earlier retention times compared with extracts from agricultural soils. Similarly, the retention time windows also varied after every GC - FID equipment maintenance operation. Consequently, retention time windows were updated at every start of an experiment and when maintenance operations are carried out on the equipment.

Percent recoveries of aliphatic ETPH analytes increased with an increase in the volume of Hexane used for SPE fractionation (figure 2.1). The highest recoveries (65% - 89%) for aliphatic ETPH analytes were obtained when 3.5ml Hexane was used. However, a test of means (analysis of variance (ANOVA)) showed no significant difference ($p<0.05$) between the elutions of 3ml Hexane and 3.5ml Hexane for all aliphatic analytes considered. Therefore, 3.5ml Hexane was considered an adequate elution volume for fractionating aliphatic ETPH analytes, while still avoiding premature releases of the aromatic analytes through the column. Percent recoveries for aromatic analytes range between 50% - 81%. 9ml isopropyl alcohol was therefore considered adequate for elution of aromatic analytes. It is recommended thatpercent recoveries should fall within 40 and 140% for all ETPH ranges and target PAH analytes, except for n-nonane which must be between 30 and 140% (MADEPH, 2009).

Table 2.2a: Retention times (Rt) of aliphatic ETPH marker compounds

Compound	Matrix	
	Pure standards/ Soil blank	Agricultural soil
n-Nonane	5.72	5.91
n-Decane	7.65	7.82
n-Dodecane	11.22	11.39
n-Tetradecane	14.32	14.50
n-Hexadecane	17.04	17.20
n-Octadecane	19.43	19.71
n-Nonadecane	20.53	20.80
n-Eicosane	21.57	21.72
n-Docosane	23.53	23.62
n-Tetracosane	25.22	25.35
n-Hexacosane	26.80	26.91
n-Octacosane	28.25	28.34
n-Triacontane	29.58	29.67
n-Hexatriacontane	34.90	35.01

Table 2.2b: Retention times (Rt) of aromatic ETPH marker compounds

Compound	Matrix	
	Pure standards/ soil blank	Agricultural soil
Naphthalene	9.16	9.33
Acenaphthylene	13.29	13.48
Acenaphthene	13.77	13.96
Fluorene	15.24	15.43
Phenanthrene	17.73	17.96
Anthracene	17.86	18.06
Fluoranthene	20.95	21.15
Pyrene	21.48	21.68
benzo(a)anthracene	24.78	24.99
Chrysene	24.90	25.09
benzo(b)fluoranthene	27.52	27.77
benzo(k)fluoranthene	27.57	28.35
benzo(a)pyrene	28.25	29.66
ineno(1,2,3-cd)pyrene	30.89	31.03
dibenzo(a,h)anthracene	31.30	32.75
benzo(g,h,i)pyrelene	35.05	35.09

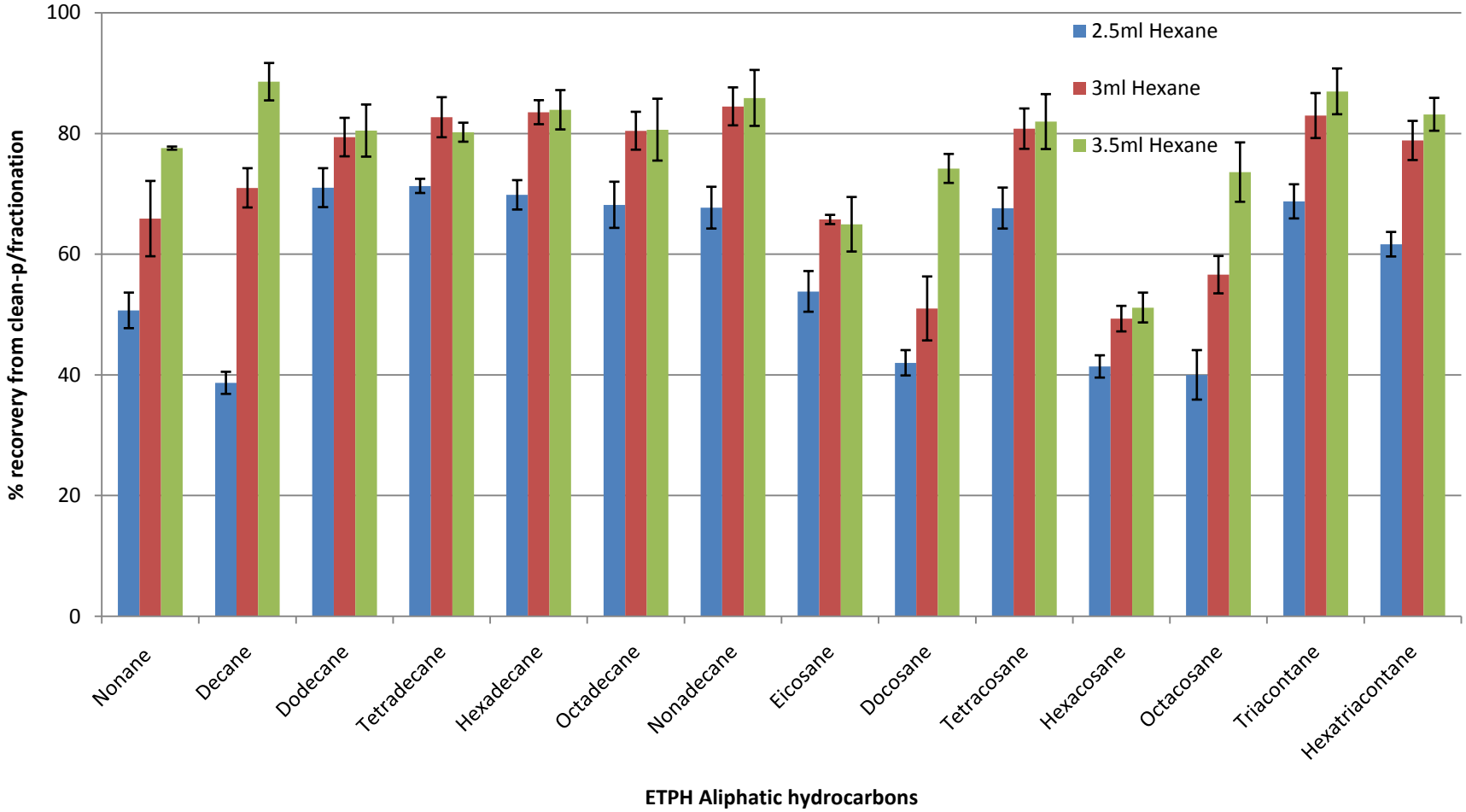


Figure 2.1a: Percent recoveries of ETPH aliphatic marker compounds from clean-up/fractionation process. Error bars indicate standard deviations

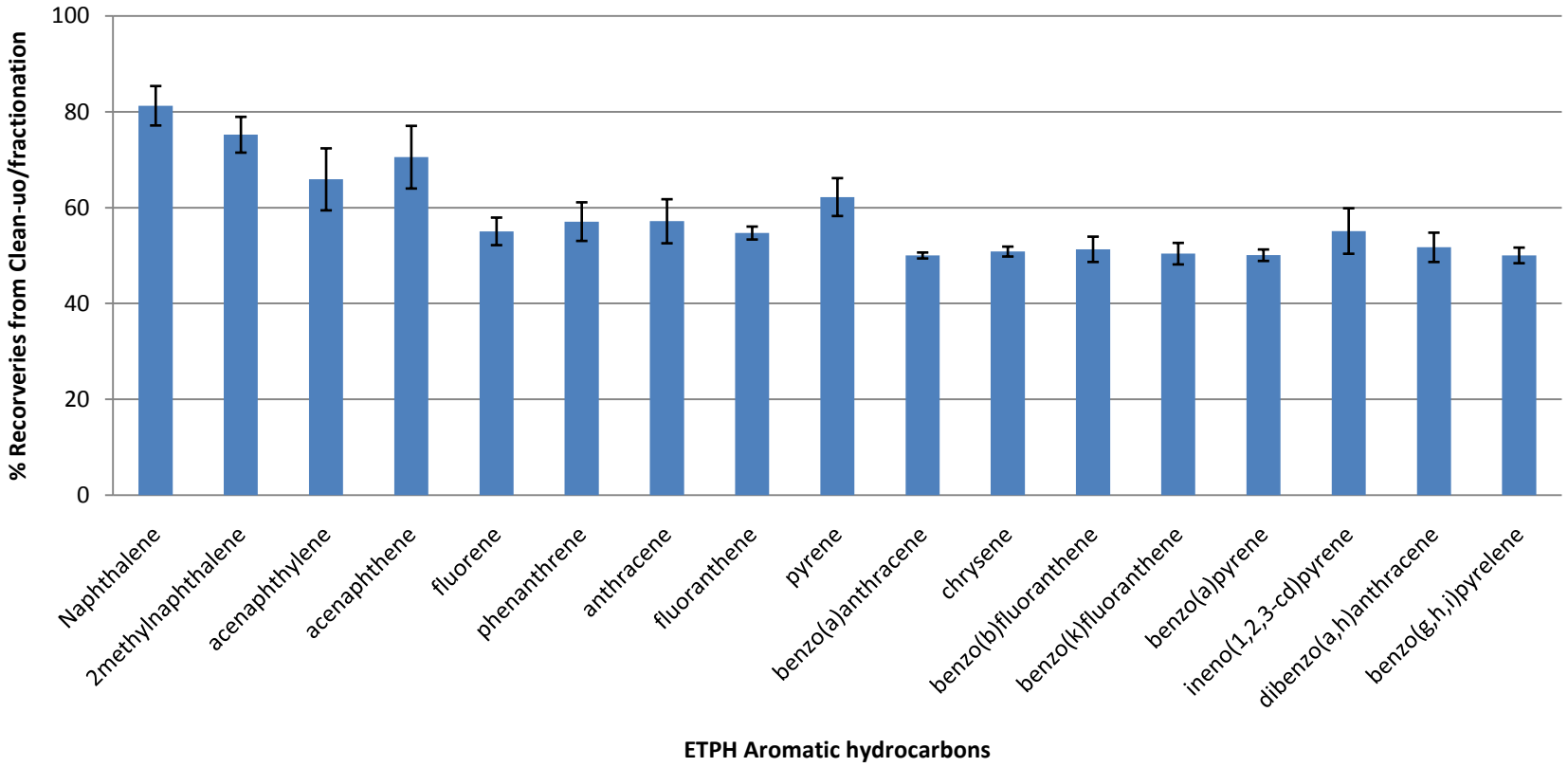


Figure 2.1b: Percent recoveries of ETPH aromatic marker compounds from clean-up/fractionation process. Error bars indicate standard deviations

The percent recoveries of all ETPH marker compounds from the acetone: hexane extraction process ranged between 68.6% – 94%with an average of 88.6% for aliphatics and 74 % for aromatics (Figure 2.2). The MADEPH 2009 protocol recommends percent recoveries between 40 and 140% for all ETPH marker compounds, except for n-nonane which must be between 30 and 140%; results outside this range are not acceptable. Recoveries from extraction/fractionation were monitored during subsequent experiments to ensure the required performance standards are met. The extraction/ silica gel clean-up procedure had an overall efficiency of 69.4%. To monitor overall efficiency for subsequent experiments, recommended ETPH surrogates (Chloro-octadecane and O-Terphenyl) were included in analyses.

Figure 2.3 illustrates a chromatogram of an extract from a diesel contaminated soil. Eleven (11) aliphatic ETPH analytes (C_9-C_{26}) were identified while all ETPH aromatics were below the detection limit (0.05ppm) of the Chromatograph.

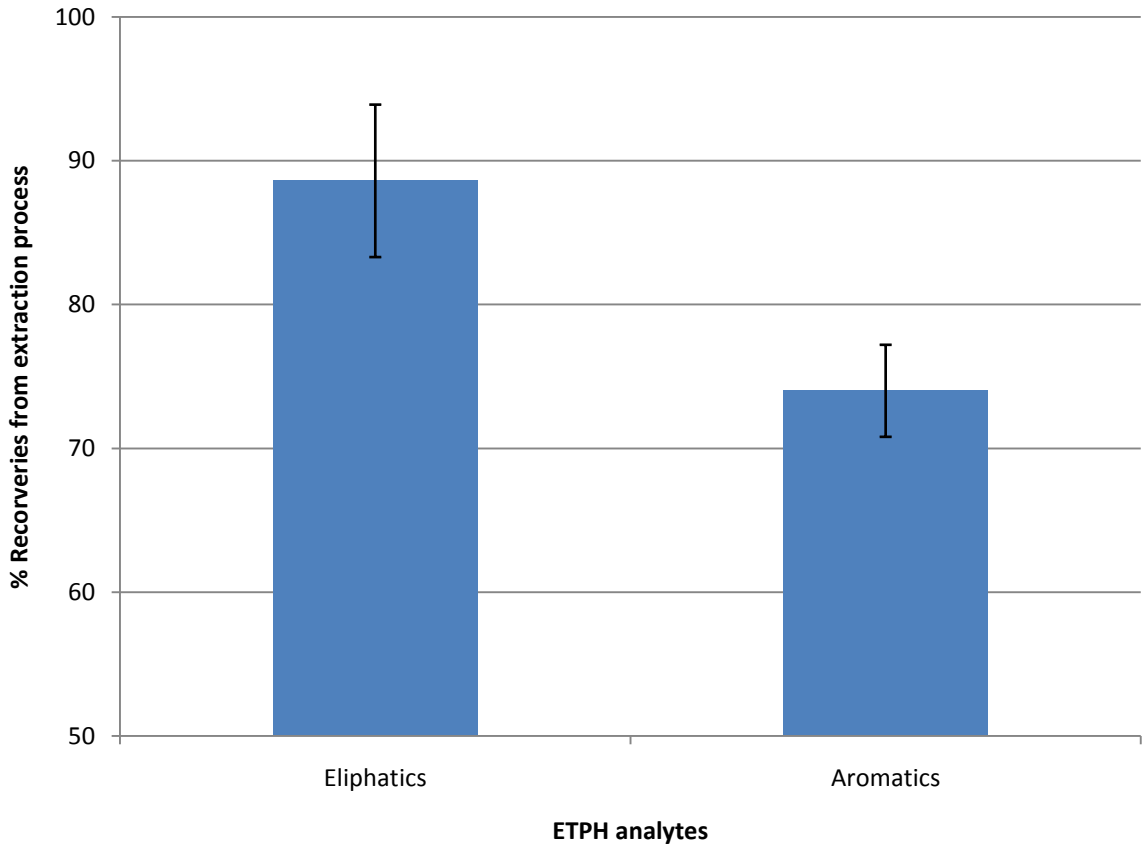
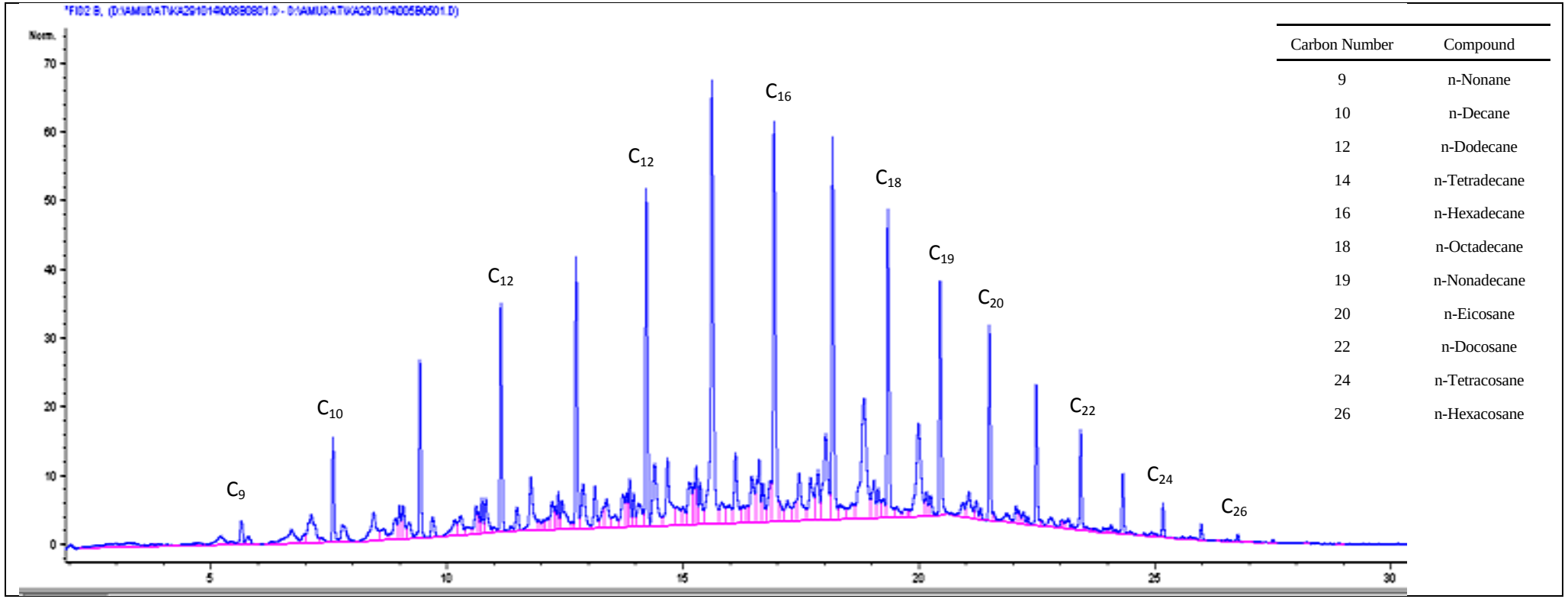


Figure 2.2: Average recoveries of ETPH analytes from extraction. Error bars indicate standard deviations



Carbon Number	Compound
9	n-Nonane
10	n-Decane
12	n-Dodecane
14	n-Tetradecane
16	n-Hexadecane
18	n-Octadecane
19	n-Nonadecane
20	n-Eicosane
22	n-Docosane
24	n-Tetracosane
26	n-Hexacosane

Figure 2.3: Aliphatic fraction of a Diesel sample showing peaks C_9 (Nonane) to C_{26} (Hexacosane)

2.3. Assessment of variability in VNIR DR Spectra of diesel contaminated soils

2.3.1. Introduction

Radiation when directed on soil causes functional groups within soil components to vibrate either by bending or stretching, leading to the absorption of a part of the radiation. The non-absorbed radiation is reflected back and can be measured as a reduced signal relative to the incident radiation. The quantity of reflected radiation and the wavelengths at which these occur depends on the chemical matrix. The resultant reflectance spectrum produces a characteristic shape that can be used for detection and possible quantification of a wide range of molecules (Stenberg et al., 2010).

Obtaining the VNIR DR spectra of petroleum hydrocarbon contaminated soils is a non destructive process and can be adapted for fieldwork such that measurements can be carried out *in-situ*. It also takes less time and requires less cost compared with conventional wet chemistry procedures for petroleum hydrocarbon analysis described above. This approach can, therefore, be a particularly useful approach for characterising petroleum contaminated soils, especially when finance and time constraints limit soil analysis. Diffuse reflectance within the visible near infrared is, however, characterised by broad absorptions, scattering effects and large numbers of multi-collinear variables and, as such, require statistical tools to extract information from them.

This section details the method used to obtain visible near infrared reflectance spectra of diesel contaminated soils to identify possible variations that can be attributed to increasing levels of diesel contamination. It also measures the strength of the relationship between these spectral variations and petroleum hydrocarbon contamination in soil.

2.3.2. Methodology

a. Soil Processing

Agricultural soil from the University of Reading farm in Sonning, United Kingdom (51°28'21"N; -0°54'13"E) was collected and oven dried at 20^0C for 7 days in a large soil oven. The soil was homogenised and then sieved to 2mm. 250g subsamples of soil were mixed thoroughly with varying quantities of diesel fuel to give concentrations of 112, 225, 450, 1000, 2000, 4000, 8000, 10000, 20000, 30000, 40000 and 50000 ppm (wt/wt). Uncontaminated soils were used as a control. Contaminants in soils were thoroughly mixed by agitation with a spatula for 1 minute. Samples were then allowed to equilibrate for 2 days prior to spectral and laboratory analysis.

b. Spectral Reflectance Measurement

Diffuse reflectance spectra of soil samples were obtained in a controlled dark laboratory with a GER3700 V-SWIR spectrophotometer (350 - 2500nm) coupled with a light source made of a quartz-halogen bulb. The spectrophotometer has one Si array (350 - 1050 nm) and two peltier-cooled InGaAs detectors (1050 - 1900 nm and 1900 - 2500 nm). Spectral sampling interval of the instrument was 3nm at (350 - 1050 nm), 7nm at (1050 - 1900 nm) and 9.5nm at (1900 - 2500 nm). The light source illuminated the sample at 45^0. The spectrophotometer measured the reflectance spectra at nadir, scanning an area with a diameter of about 60mm.

Sample spectral radiances were compared with those of a reference panel of Spectralon to determine spectral reflectance, expressed as a percentage. Scans were taken from the soil samples, tightly packed and levelled in borosilicate petri - dishes (100mm ×15mm). Replicate measurements were collected at four positions by rotating the petri dish at an angle of 90^0 to

account for any directional surface scattering effects. Each scan was an average of 16 internal scans. The 64 replicate scans were then averaged to produce a single scan per sample.

c. Processing of Spectra

Noisy portions in the spectrum (350 – 400nm, 1905 – 1928nm and 2450 – 2500nm) were removed because of low sensitivity at these wavelengths. This was followed by averaging adjacent 5 nm wavelengths to reduce the impact of noise. An index, the normalized difference index (NDI), a modification of the normalized vegetation index (NDVI) was developed to generate values that describes the observed variation within the spectra using two wavelengths points at which reflectances were observed to vary with increasing diesel contamination.

Green plans absorb radiation in the red portion of the visible region for photosynthesis. They however scatter radiation in the infrared (IR) region because the energy level in the IR region is not sufficient for photosynthesis and absorption will only lead to overheating of the plant causing tissue damage. His phenomenon is exploited by the NDVI which quantifies vegetation by measuring the difference in plant spectral reflectances between near-infrared (where vegetation strongly reflects) and red portion of the visible region (where plants strongly absorbs) (Esau et al., 2016).

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

The spectral absorptions of hydrocarbon based oils are apparent around 1647nm, 1712nm, and 1759nm in the first overtone of the near infrared band (Okparanma and Mouazen, 2013). The derived index (NDI_1) therefore utilized spectral reflectances at two points of the spectrum where soil reflectances were observed to vary strongly in relation to diesel contamination

The derived indices were plotted against diesel contamination to observe variations within the spectra. It should be noted that the derived indices do not give full information from the spectra. It is therefore used as a graphical indicator, providing an estimate of the gradient of the spectral reflectance curve between the two measurement points, for qualitative purposes rather than for quantitative purposes.

$$NDI_1 = \frac{(R_{(2129nm)} - R_{(1728nm)})}{(R_{(2129nm)} + R_{(1728nm)})}$$

$R_{(2129nm)}$ refers to the wavelength at which the highest reflectance was observed in each spectrum

$R_{(1728nm)}$ refers to the wavelength at which the lowest reflectance associated with diesel contamination (i.e. maximum absorption) was observed

The use of the raw reflectance is not recommended for the identification and measurement of individual absorption features since the area, depth and symmetry of a feature will be strongly influenced by its position on the background continuum (White et al., 2007). During spectral measurement, there may be a shift in absorption minimum due to a wavelength dependent scattering impacting a slope on the spectrum. However, removing the continuum slope corrects the feature minimum to that of the true feature centre (Clark and Roush, 1984).

The continuum removal technique involves fitting a maxima curve over a spectrum to form a continuum or Hull. A mathematical function is then fitted to the spectrum to represent absorptions due to processes other than those of interest. These effects can then be removed by dividing or subtracting the measured spectra from the continuum function (Clark and Roush, 1984).This allows for a comparison of individual absorption features on each spectrum from a common baseline (figure 2.4).

To quantitatively measure the strength of the relationship between variations observed within spectra and petroleum hydrocarbon contamination in soils, the continuum removal technique was applied to VNIR DR spectra of diesel contaminated soils using the DISPEC software developed by NASA Jet Propulsion Laboratory. The depth and areas of two absorption features at ~1640 – 1760nm and ~2240 – 2340nm, observed to have a positive relationship with increasing ETPH concentrations, were then calculated using the DISPEC software and regressed against petroleum hydrocarbon contamination.

2.3.3. Results and Discussion

Figure 2.4 illustrates the visible near infrared diffuse reflectance spectra of diesel contaminated soils. Overall soil reflectances were observed to decrease with increasing diesel contamination. Once the raw reflectances were continuum removed, all spectra can be viewed from the same baseline and specific absorptions associated with petroleum hydrocarbons were observed at wavelength regions ~1640nm – 1760nm, most likely caused by CH_2 and CH_3 stretches and bends in the first overtone region of the fundamental C-H stretch between 3300 -3400nm. Another absorption linked with petroleum hydrocarbons is also observed at ~2240nm – 2440nm which is probably a consequence of the C-H stretch of a polyaromatic hydrocarbon and a C-H stretch of terminal CH_3 and CH_2 saturated hydrocarbons (Okparanma and Mouazen, 2013).

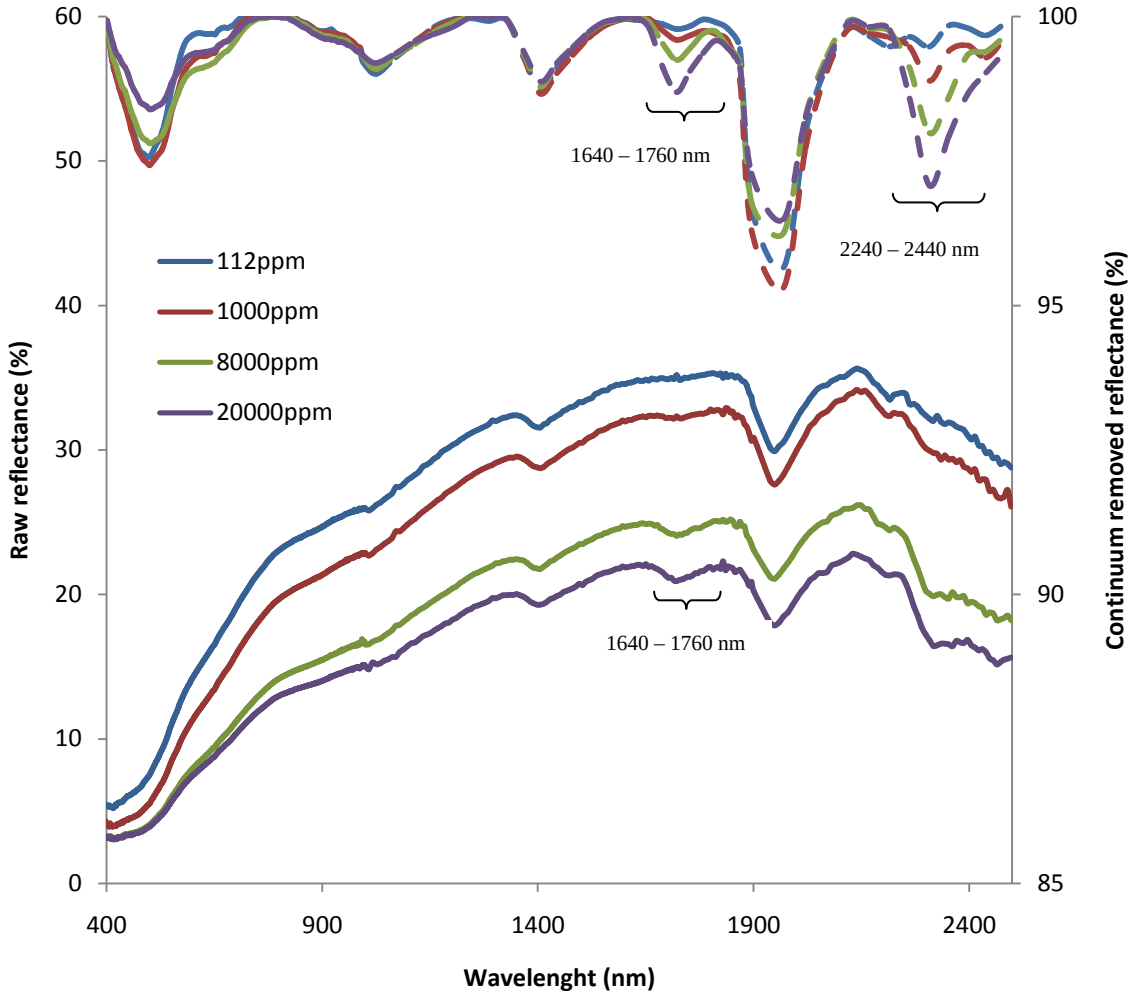


Figure 2.4: VISNIR spectral reflectance curves of diesel contaminated soils. Solid lines are spectral reflectance curves; dashed lines are hull differenced (continuum removed) spectra.

Figure 2.5 shows the indices generated from reflectance spectra of diesel contaminated soils as a scatter plot against the spiked diesel contamination. The data indicate increases in index with increasing soil contamination. However, there seem to be no increase in the index beyond 30000ppm diesel contamination; possibly indicating that this technique has reached saturation at these levels. This implies that using the visible near infrared diffuse reflectance spectra may not be appropriate for distinguishing between soils with diesel concentrations greater than 30000ppm; at these levels soils should be considered highly polluted. While it should be noted that this index is used qualitatively, it indicates that diffuse reflectance spectra of soils can be a useful tool in pollution studies for identifying petroleum hydrocarbon contamination in soils, as acceptable levels fall below 30000ppm. Acceptable clean-up levels of total petroleum hydrocarbons in soils vary across locations depending on existing legislation as well as land use patterns: 1000ppm in Massachusetts (MADEPH, 2014), 50 – 500ppm in Oklahoma (ODEQ, 2012), 59 – 2300ppm in Wyoming (WDEQ, 2000), 100ppm in Alabama and Arkankas, 100 – 2000ppm in Alaska, <100ppm in Arizona (NSCEP,1993). The most frequently used soil cleanup standard for total petroleum hydrocarbons in soils is 100ppm, although the standards and guidelines in different states range from background concentrations up to 10000ppm TPH in soil (Michelsen and Boyce, 1993).

The areas of two identified absorption features at ~1640 – 1760nm and ~2240 – 2340nm regressed against laboratory measured petroleum hydrocarbon values show strong positive relationships ($R^2 > 0.70$) with increasing ETPH concentrations (figure 2.6).This show that visible near infrared diffuse reflectance spectroscopy, coupled with continuum removal, can be used not only to identify petroleum hydrocarbon contamination in soils but also to quantify concentrations.

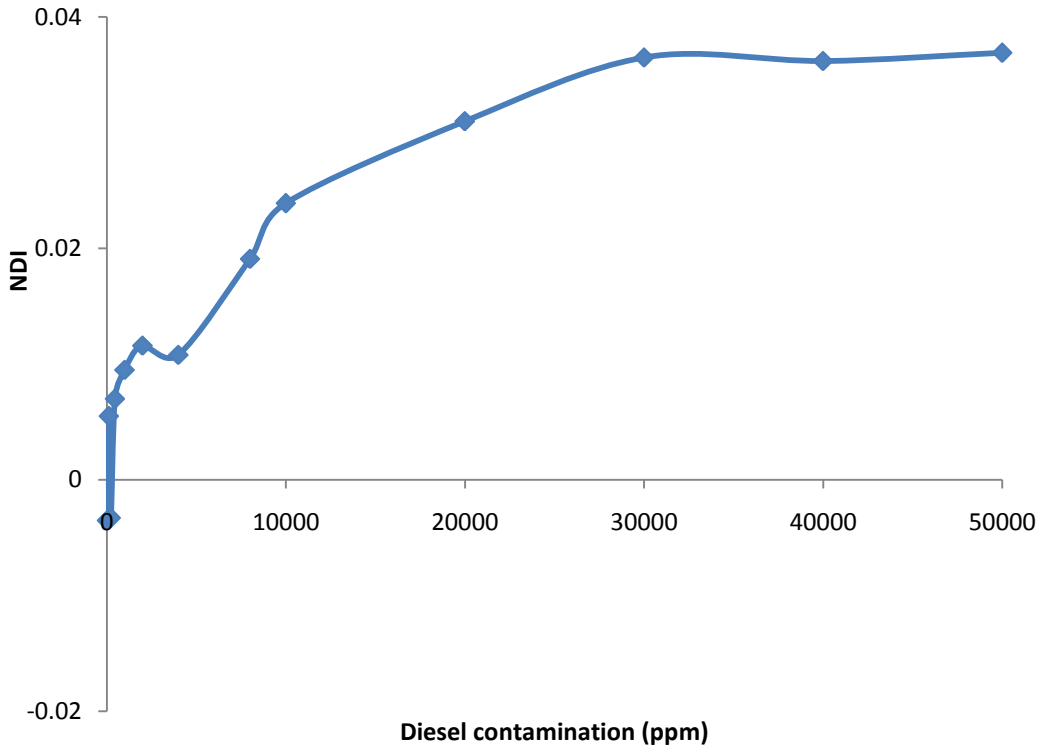


Figure 2.5: Soil reflectance index versus levels of diesel contamination

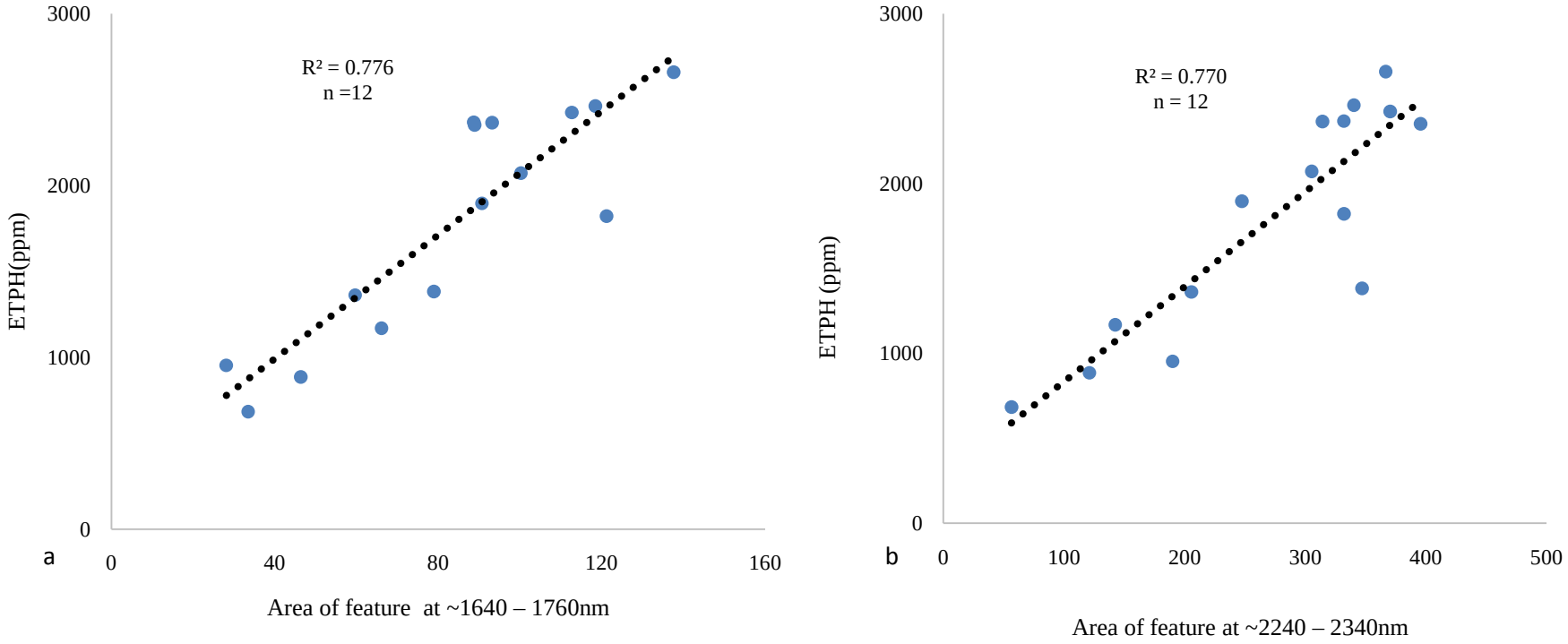


Figure 2.6: Relationship between ETPH and area of absorption features at (a) ~1640 – 1760nm and (b) ~2240 – 2440nm. Where n = number of samples

2.4. Modeling petroleum hydrocarbon content of soil using VNIR DR spectra

2.4.1. Introduction

Visible near infrared spectra are characterised by a large number of multi-collinear variables with no single wavelength providing sufficient information for the analyte being measured, thus requiring multivariate regression tools that are able to accommodate the collinear nature of the variables. Two broad absorption features at ~1640 – 1760nm and ~2240 – 2440nm, whose areas have strong positive correlations (R^2> 0.70) with laboratory measured ETPH values, have been identified from continuum removed reflectance spectra of diesel contaminated soils in the previous section of this chapter. This study, therefore, aims to develop and compare prediction models generated from portions of the visible near infrared spectra in a bid to identify sensitive wavelength regions best suited for quantitative modeling of petroleum hydrocarbon contamination in soils.

2.4.2. Methodology

Two groups of agricultural soils: Group Y consisting 43 topsoil samples from the University of Reading farm in Sonning, (51°28'21"N; -0°54'13"E) and Group Z consisting of 51 topsoil samples collected from the Royal Horticultural Society experimental plots at Wisley, Surrey,(51°19'24.34"N; 0°28'27.81"W)both in the United Kingdom, were air dried and then sieved to 2mm. They were then contaminated in the laboratory with different amounts of diesel fuel and the contaminants thoroughly mixed by agitation with a spatula for 1 minute. Samples were then allowed to equilibrate for 2 days prior to spectral and laboratory analysis. Contaminated soils tightly packed and levelled in borosilicate petri – dishes were scanned for diffuse reflectance spectra in a dark room with a GER3700 VNIR spectrophotometer. Scanned samples were then analysed for extractible total petroleum hydrocarbon (ETPH)

content using the optimised extraction/GC-FID procedure described in Section 2.1 of this chapter.

Partial least square regression (PLSR) models were developed to quantify ETPH in soils from both VNIR DR spectra and ETPH using different wavelength regions within the spectra. Prior to model calibration, soil spectra were processed by averaging 5 consecutive wavelengths to reduce variable size, and application of the standard normal variate/ detrending tool to reduce multi-collinearity, baseline shift and curvature. All spectral pre-treatment and multivariate calibration/validations were carried out using the Unscrambler-X 10.3 analytical software developed by CAMO.

The predictive accuracies of PLSR models were compared using the coefficient of determination (R^2), root mean square error (RMSE), relative percent difference (RPD) values and the percentage prediction error (% PE).

$$RMSEp = \sqrt{\frac{\sum(TPHpred - TPHmeas)^2}{n}} \quad -----(1)$$

$$\%\, PE = \frac{RMSEp}{Xh_v} \times 100 \quad ----------(2)$$

$$RPD = {}^{SD}/_{RMSEp} \quad ----------(3)$$

Where n is the number of validation samples, SD is the standard deviation of the predicted validation values and Xh_v is the largest measured data point within the validation set.

According to ViscarraRossel *et al*. (2006) and Chang *et al*. (2001), very poor models show RPD < 1.0; poor models: 1.0 ≤ RPD ≤ 1.4; fair models: 1.4 ≤ RPD ≤ 1.8; good models : 1.8 ≤ RPD ≤ 2.0 and very good models : 2.0 ≤ RPD ≤ 2.5; and excellent models have RPD > 2.5.

2.4.3. Results and Discussion

Model quality statistics are illustrated in table 2.3. Modeling ETPH using only the wavelength regions physically identified from the continuum removed spectra to have strong positive correlations with diesel contamination did not improve model quality. The best models were obtained when the whole spectrum was used for model calibrations (R^2 = 0.83; 0.95 and RPD = 2.19; 4.29 for soil groups Y and Z respectively). The lowest percent prediction errors were also obtained when the whole spectrum was used for model calibration (figure 2.7). Including both absorption features (1640 – 1760nm and 2240 – 2440nm) in the modeling gave better model statistics than when each was used individually for modeling. There was, however, no established pattern regarding which absorption feature was better at predicting ETPH.

Model calibrations using specific diagnostic spectral features are usually biased by our prior understanding of the origin of the features. In addition, by using only small parts of the spectrum, the full information content of the spectra are not utilised (Hurley et al., 2002). Though the large number of variables, and their multi-collinearity, is a common problem inherent to visible near infrared diffuse reflectance spectra of soils (Stenberg et al., 2010), the whole spectrum can be investigated with multivariate statistical tools, such as the partial least squares regression which is able to extract significant signals and to create reliable models (Haenlein and Kaplan, 2004).

Table 2.3: Performance of ETPH – VNIR partial least square regression models of diesel contaminated soils using different wavelength regions

Data set	Spectral band	Calibration		Validation		
		R^2_c	$RMSE_c$ Log_{10} ppm	R^2_{cv}	$RMSE_{cv}$ Log_{10} ppm	RPD
Y	1640 - 1760nm	0.80	0.30	0.79	0.30	1.85
	2240 – 2440nm	0.72	0.22	0.70	0.23	1.83
	1640 - 1760nm, 2240 – 2440nm	0.79	0.16	0.77	0.17	2.05
	400 – 2500nm	0.83	0.14	0.80	0.15	2.19
Z	1640 - 1760nm	0.89	0.17	0.89	0.17	1.86
	2240 – 2440nm	0.94	0.21	0.94	0.21	2.82
	1640 - 1760nm, 2240 – 2440nm	0.94	0.14	0.94	0.15	3.52
	400 – 2500nm	0.95	0.11	0.95	0.11	4.29

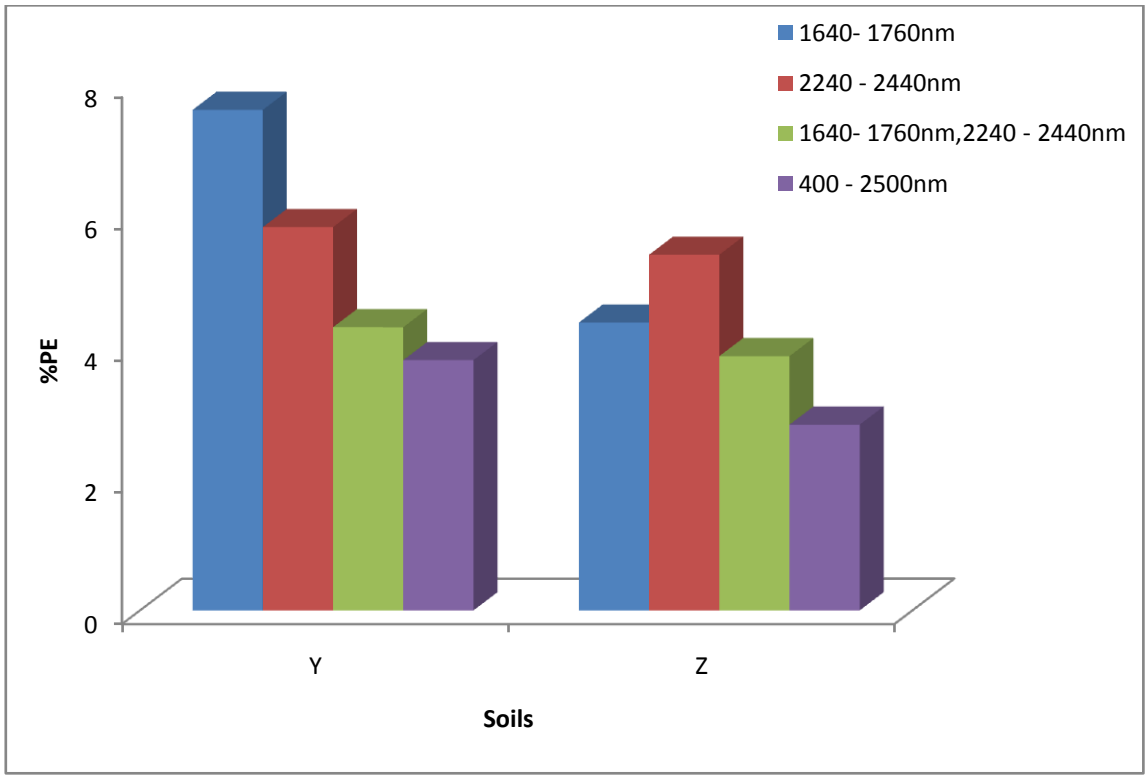


Figure 2.7: Percent prediction error statistic of ETPH-VNIR partial least square regression models

2.5. Conclusion

Optimizing the analytical procedures to ensure reproducibility is important as the robustness of future modeling will depend on the reliability of results obtained from both the reference and spectral measurements. The optimised extraction/silica gel clean-up procedure for ETPH analysis had percent recoveries from the clean-up/fractionation process ranging between 65% - 89% for aliphatics and 50% - 81 % for aromatics. The Acetone: Hexane extraction processes had recoveries ranging between 68.6% – 94%with an average of 88.6% for aliphatics and 74 % for aromatics while the overall extraction/silica gel/clean-up procedure had an efficiency of 69.4%.

A normalised index derived using reflectance values at two wavelength points on the spectra identified petroleum hydrocarbon contamination up to a limit of about 30000ppm. Applying continuum removal on soil spectra obtained from the diesel contaminated soils, two broad absorptions at ~1640 – 1760nm and ~2240 – 2340nm were identified,whose areas were positively correlated (R^2> 0.70) with the degree of diesel contamination.

Selecting the most suitable wavebands within a visible near infrared diffuse reflectance spectrum for estimating petroleum hydrocarbon contamination is usually cumbersome, as the analysis of a single feature is typically not sufficient to explore such rich information. Using data from two groups of soils, ETPH - PLSR models developed from whole spectra were found to have better prediction potentials compared to PLSR models derived from two absorption features ~1640 – 1760nm and ~2240 – 2340nm. This implies that modeling petroleum hydrocarbon contamination in soils is better done by incorporating all information provided by several spectral features, using regression tools that can accommodate multi-collinearity, to ensure that subtle information are not discarded and the correlations between absorptions at all wavelengths are taken into consideration.

References

Bullard, J.E. and White, K. (2002) Quantifying iron oxide coatings on dune sands using spectrometric measurements: an example from the Simpson-Strzelecki Desert, Australia. *Journal of Geophysical Research*, 107 (B6): 1-11

Clark, R.N. and Roush, R.N. (1984). Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research*, 89 (B7): 6329–6340

Esau, I., Miles, V., Davy, R., Miles, M. and Kurchatova, A. (2016). Trends in normalized difference vegetation index (NDVI) associated with urban development in northern West Siberia. *Atmos. Chem. Phys* (16): 9563–9577

Haenlein, M. and Kaplan, A.M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics* 3(4), 283–297

Hurley, P. D., Oliver, S., Farrah, D., Wang, L and Efstathiou, A. (2012) Principal component analysis and radiative transfer modelling of Spitzer Infrared Spectrograph spectra of ultraluminous infrared galaxies. *Monthly Notices of the Royal Astronomical Society*, 424 (3): 2069-2078.

MADEPH, 2009.Recommended reasonable confidence protocols quality assurance and quality control requirements extractable petroleum hydrocarbons. State of Connecticut Department of Environmental Protection.

National Service Centre for Environmental Publication (NSCEP). (1993). *State Summary of Soil and Groundwater Cleanup Standards for Hydrocarbons*. Available at https://nepis.epa.gov.

Oklahoma Department of Environmental Quality (ODEP). (2012). Risk – based levels for total petroleum hydrocarbons (TPH). Available at http://www.deq.state.ok.us/factsheet.

Okparanma R. N. and Mouazen, A. M. (2013). Determination of total petroleum hydrocarbon (TPH) and polycyclic aromatic hydrocarbon (PAH) in soils: a review of spectroscopic and non-spectroscopic techniques. *Applied Spectroscopy Reviews,* 48 (6) 458–486

Stenberg Bo, ViscarraRossel, R. A., Mouazen, A. M. and Wetterlind J. (2010).Visible and Near Infrared Spectroscopy in Soil Science. In Donald L. Sparks editon: *Advances in Agronomy,* Vol. 107, Burlington

White, K., Walden, J. and Gurney, S. D. (2007) Spectral properties, iron oxide content and provenance of Namib dune sands. *Geomorphology,* 86 (3-4). pp. 219-229.

Wyoming Department of Environmental Quality (WDEP). (2000). *Soil clean up levels.* Available at http://deq.wyoming.gov/media/attachments

Chapter 3. Performance of visible near infrared diffuse reflectance spectrometry for soil analysis: effect of data split patterns and regression techniques

3.1. Introduction

The use of visible near infrared diffuse reflectance (VNIR DR) spectra of soils, coupled with multivariate statistics, has been shown to be a potentially viable analytical procedure for the rapid determination of the chemical composition of soils. However, VNIR DR spectra are usually characterised by broad overlapping absorption bands, large multiple collinear variables, and scattering effects / pathlength variation. They therefore require pre-processing and multivariate calibrations to extract relevant information on soil composition.

Statistical spectral pre-processing involves the application of statistical tools to spectra in a bid to remove or minimise noise and interference, emphasise pertinent weak signals and correct for scattering effects, baseline shifts and variations in path-length. It is the basis for feature extraction and model development as it plays an important role in the modeling process and the accuracy of the analysis (Li et al., 2010). Several preprocessing techniques, such as normalization, multiplicative scattering correction (MSC), standard normal variate (SNV), detrending (DT), derivatization and data smoothing filters (e.g. Savitsky – Golay smoothing),have been shown to be effective in reducing the number of latent factors or principal components within models when compared to unprocessed spectra. The proper choice of preprocessing can be difficult to determine prior to model validation, and there may not be a single “best” spectral data pre-processing technique, the decision on which technique to use depending on the quality of the raw spectra (Shi *et al*., 2012). However, as a minimum requirement, the ideal pre-processing should maintain or reduce the effective model complexity.

While this important step is meant to improve the quality of datasets for consequent modeling, it could also reduce data quality if not properly used. For example, using the first derivative spectral transformation prior to modeling may not produce consistent results when working with datasets obtained from a varied geographical location (Brown et al., 2005). This is due to the fact that spectral signatures of soil organic matter are highly complex, varying by depth, vegetative cover and geographical location. The wavelengths at which these absorptions occur within the visible near infrared region are not entirely unique and may be masked by absorptions associated with iron oxides and some clay minerals (Clark, 1999). More importantly, the first derivative of an absorption feature will shift due to the influence of other broad absorptions, making it an unreliable predictor (Kokaly and Clark, 1999).Therefore, preprocessing of spectra, if needed, must be done with caution to ensure that only the right preprocessing tools are used in a bid to avoid an unintentional screen-out of absorption features that may be relevant to the soil component being modeled.

In principle, during model calibrations, data sets are divided into two groups, one for calibration or model building and the other for model validation. As the aim of any regression modeling is to make accurate predictions in the future, the validation of models estimates the uncertainty of such future predictions. If the uncertainty is reasonably low, the model can be considered valid. The use of a validation set obtained from the same process as the calibration set but independent of model calibration provides a real time assessment of the ability of the model to predict future values of the response variable. However, there is usually a lack of reasonably large data sets due to high costs of reference analysis and, in most situations, cross-validation results in lieu of test validation have been reported by most studies(Russell, 2003; Udelhoven et al., 2003) while only calibration results are reported by some studies(Reeves et al., 2002). Cross-validation does not

accurately measure the predictive ability of the models as it introduces pseudo – randomness in the process (Krstajic *et al*., 2014). Instead, cross-validation better measures the robustness of a particular model, as estimated by an internal sub-setting of the calibration dataset, rather than giving information as to the future use of the model (Esbensen and Geladi, 2010).

While there is no fixed value for the number of samples to be used in calibrations (Broad *et al.*, 2001), between 20 and 30 samples have been proposed for feasibility studies and initial calibrations (Williams, 2001). The size of an independent randomly selected validation dataset, relative to the calibration data set, becomes crucial when the dataset is limited. It therefore becomes important to identify the appropriate data split pattern that can produce a reasonable accurate calibration model and one that ensures that constructed models are adequately validated.

The choice of which regression tool to use during multivariate calibrations can also be a challenging task. Due to the heterogeneous nature of soils and the multi-collinear nature of VNIR DR spectra of soils, the appropriate model must be able to accommodate large data sets of such nature, as well as emphasise the absorption features relevant to the soil parameter being measured, while downplaying noise and interferences within the spectra. Several statistical methods have used been in combination with spectral data for estimating different soil compositions; the most frequently used being partial least square regression (PLSR).Other methods used include multiple linear regression (MLR) and principal component regression (PCR). However, it still remains unclear as to which statistical method is more suitable and can overcome the problems posed by the heterogeneous nature of soils and multi-collinear properties of spectra during multivariate calibrations.

The aim of this study is to examine the performance of visible near infrared diffuse reflectance spectroscopy as a potential soil analytical procedure under different dataset split patterns and regression techniques. Specific objectives include

1. Identify the reasonably appropriate dataset split pattern that produces the best calibration performance.
2. Study and compare the predictive accuracies of different regression techniques for analysis of soil carbon and extractible total petroleum hydrocarbons in soils.

3.2. Effect of dataset split patterns on VNIR DR model quality

3.2.1. Methodology

To assess the effect of data split patterns on model quality, and to identify the appropriate data split pattern that is capable of producing a reasonably accurate calibration model, four data split patterns are studied (Table 3.1). Visible near infrared diffuse reflectance spectra of diesel contaminated soils were pre-processed using the standard normal variate coupled with detrending technique (SNV-DT) and regressed against extractible total petroleum hydrocarbon (ETPH) concentrations.

Soils used for this study were carefully homogenized prior to contamination, at varying levels, with diesel fuel in the laboratory. This was done to ensure that variations within soil spectra to be modeled originate mostly from petroleum hydrocarbons within the soils. The calibration datasets for each of the data split patterns studied was carefully selected to cover the range of ETPH concentrations to be predicted. The spectra of soil samples were also carefully studied and it was

ensured that the calibration data set is representative of the variations observed within the population.

Randomly extracting the calibration set from the total population though simple and commonly used in statistical studies does not guarantee that representative samples are included in the calibration. It also does not make certain that the samples on the boundaries of the population are included in the calibration. Similarly, many sample selection methods such as linear algebra techniques and the Kennard–Stone (KS) algorithm focuses only on the spatial distribution of samples while the reference data set is not taken into consideration for multivariate calibrations (Zhang et al., 2014). The calibration set as selected within this study ensures that both the reference and spectral data sets are included in the sample selection process to improve the modeling efficiency.

All spectra pre-processing and model calibrations were carried out using the Unscrambler-X 10.3 analytical software developed by CAMO Software, Oslo, Norway.

Table 3.1: Data split patterns studied

Data split pattern	% of dataset used for calibration	% of dataset used for validation
Full cross-validation	100	100
Split 90/10	90	10
Split 76/24	76	24
Split 50/50	50	50

Four partial least square regression (PLSR) models were developed using these dataset split patterns. Qualities of calibrated PLSR models were assessed using the coefficient of determination R^2, relative percent difference (RPD) values, root means square error (RMSE) values and number of latent factors for the optimum model.

R^2: This is the square of Pearson's product moment correlation coefficient. It measures the relationship between the modelled and the measured data on a positive scale between 0 and 1, the higher the better explanation of the dependent variable by the independent variable. It also explains the proportion of variation within the predicted response variable that can be explained by the constructed model.

RMSE: This is a measure of the average deviation between modelled and measured data. When models generated from the same data are compared, the lower the RMSE, the better. However, comparing RMSE of models generated from different sets of data may be misleading, as the RMSE values are unique to each calibration/validation data set and most likely will reflect the range of values used in the modelling

$$RMSE = \sqrt{\frac{\sum(TPHp - TPHm)^2}{n}}$$

Where *TPHp* are the modelled values, *TPHm* are the laboratory measured values and *n* is the number of validation samples.

RPD: This is a measure of predictive accuracy.

$$RPD = {}^{SD}/_{RMSE}$$

where SD is the standard deviation of the predicted values. According to Viscarra Rossel et al. (2006) and Chang et al. (2001), RPD < 1.0 indicate very poor models; $1.0 \leq RPD \leq 1.4$ indicate poor models; $1.4 \leq RPD \leq 1.8$ indicate fair models which can be improved upon: ; $1.8 \leq RPD \leq 2.0$ good models suitable for quantitative predictions while very good or excellent models have RPD > 2.0.

In a bid to solve the problem of multi-collinearity associated with reflectance spectra, the partial least square regression transforms all variables into linear combinations of the variables, referred to as latent factors. Choosing the model with the optimum number of latent factors is essential. If too much latent factors are used in a model, the model equation becomes data dependent and will not give good predictions. Likewise, if too few latent factors are used, the model equation will not be large enough to capture variability within the calibration data and, as such, will not give good predictions (Figure 3.1). Therefore, the optimum number of latent factors will be that with the first local minimum in root mean square error (Esbensen et al., 2000).

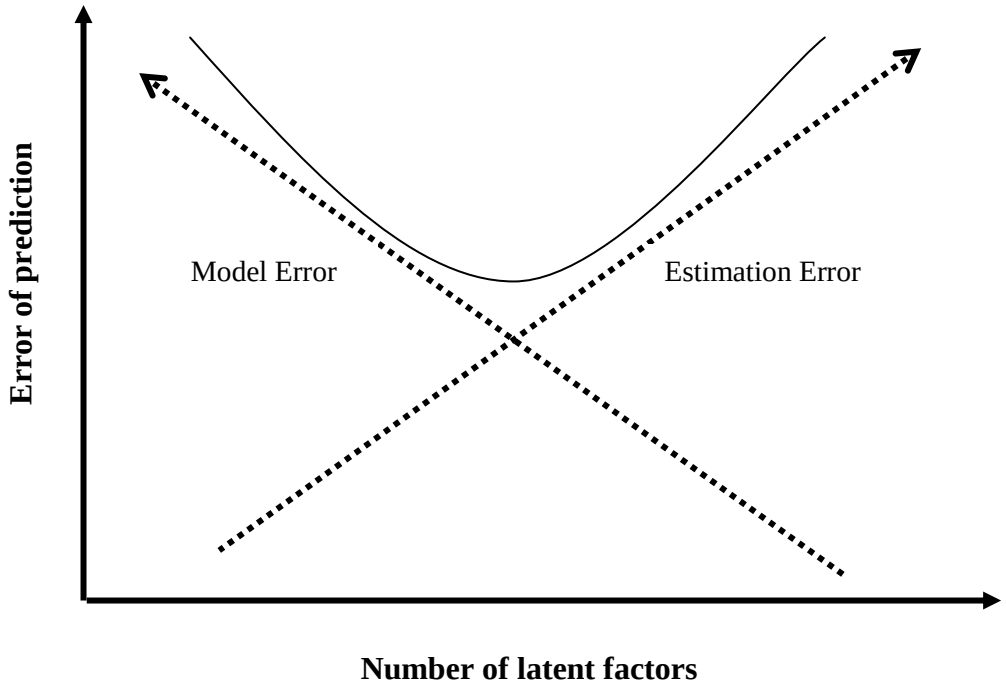


Figure 3.1: Relationship between model quality and number of latent factors (Esbensen et al., 2000)

3.2.2. Results and Discussion

Table 3.2 illustrates the quality statistics of PLSR models derived from visible near infrared diffuse reflectance spectra of diesel contaminated soils using different data split patterns. All models were able to describe a very large proportion (> 90%) of the variance observed in both the predictor (spectral data) and response (ETPH) variables. For each of the dataset split patterns, an increase in validation R^2 and RPD values was observed with an increase in latent factors up to point 6 after which it starts to slight decline. Therefore, for each of the dataset split patterns, the optimum model chosen is that with the least number of latent factors and with the first minimum in root mean square error of validation to avoid the problem of over-fitting.

The aim of multivariate calibrations is to generate models that can adequately estimate future values. Therefore, it is important that a validation is performed to increase confidence in the predictive ability of models. All four optimum models developed in this study had relative percent difference (RPD) values greater than 2. However, the predictive quality of PLSR models, as assessed using the RPD and RMSE values, were observed to decline after reaching a maximum at the 76/24 split pattern (figure 3.2). This was considered as the optimal data split pattern. This observation aligns with the conclusion of Mourad et al. (2005), who proposed a validation proportion size greater than 15% but not more than 40% and that using approximately 25% of the entire dataset for validation is statistically optimal. An important point to look out for when selecting both the calibration and validation sets is to ensure that both datasets are representative of the variations observed, both within the spectra as well as the soil property being considered, as each calibration is unique and applicable only for the purpose and range of distribution for which it was calibrated (Fearn, 2005; Stenberg *et al.,* 2012).

Table 3.2: Quality statistics of PLSR models developed using different dataset split patterns

Data split pattern	Latent factor	Calibration		Validation		
		R^2_c	$RMSE_c$ (Log_{10}ppm)	R^2_v	$RMSE_v$(Log_{10}ppm)	RPD
Full cross-validation	6	0.92	0.06	0.80	0.09	2.19
Split 90/10	4	0.86	0.08	0.85	0.06	2.30
Split 76/24	4	0.86	0.08	0.94	0.06	3.81
Split 50/50	6	0.90	0.06	0.92	0.07	3.45

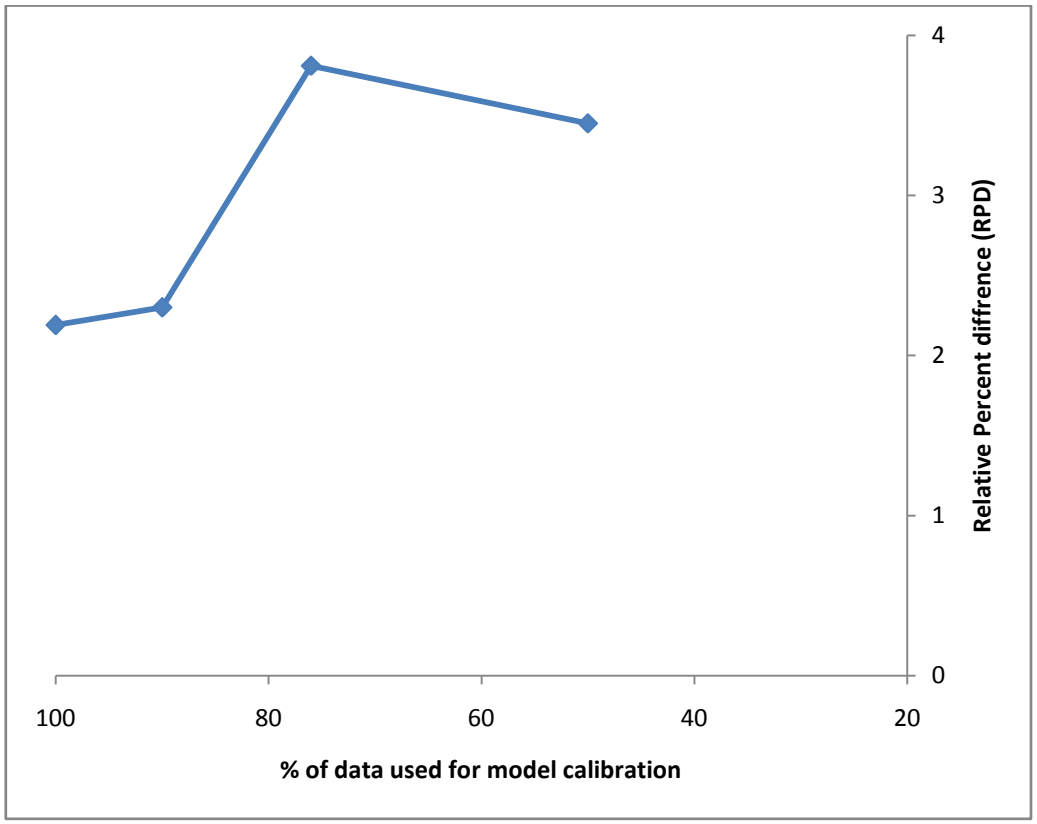


Figure 3.2: Relative percent difference (RPD) for ETPH (Log_{10}ppm) estimation using partial least square regression (PLSR) and different dataset split patterns. Predictive quality of PLSR model observed to decline after reaching a maximum at the 76/24 split pattern.

3.3. Performance of VNIR DR using different regression methods

3.3.1. Methodology

The performance of VNIR DR spectra were assessed and compared using four regression methods - multiple linear regression (MLR), principal component regression (PCR), partial least square regression (PLSR) and support vector machine regression (SVMR).Regression models were developed to predict extractible total petroleum hydrocarbons (ETPH) concentrations using VNIR DR spectra of diesel contaminated soils and data derived using the modified MADEPH analytical procedure for extractible petroleum hydrocarbon analysis. Descriptive statistics of ETPH concentrations in soil samples are presented in Table 3.3.

Visible near infrared spectra were first processed by removing wavelengths less than 400nm and greater than 2450nm, where high noise effects become significant, due to low signal/noise ratio. They were then smoothed by averaging five consecutive wavelength points to give a total of 200 wavelength variables. The raw spectra were then pre-processed using a combination of the standard normal variate transformation (SNV) and the detrending tool prior to multivariate calibrations. Models were calibrated with 76% of the dataset and validated with 24% of the dataset independent of calibration, in line with earlier identified data split pattern.

Prior to regression modeling, a principal component analysis was carried out on the data sets to identify possible outliers. Four samples with very high leverages and residuals are considered possible outliers and excluded from further modelling.

Due to the inability of the multiple linear regression (MLR) algorithm to handle datasets with variable sizes larger than the sample size (one of its setbacks), wavelengths that best describes the variations between soil spectra were identified from continuum removed soil spectra and

principal component analysis, and were used in calibrating the MLR model.Wavelength variables used for calibrating ETPH-MLR models include ~1150 – 1550nm, ~1650 – 1750nm, ~1800 – 2100nm and ~2200 – 2449nm. The SVMR model was calibrated using a capacity parameter (C) and the Gaussian radial basis function (RBF) of 100 and 0.01 respectively, searched for by the grid searching method. The best SVMR model used 28 support vectors and explained 93 % of the variation within the ETPH content of soils.

Model performance was assessed using the coefficient of determination R^2, relative percent difference (RPD) values, root means square error (RMSE) values and number of latent factors for the optimum model. All spectral pre-processing and model calibrations were carried out using the Unscrambler-X 10.3 analytical software developed by CAMO.

Table 3.3: Descriptive statistics of the extractible total petroleum hydrocarbon content (Log_{10} ppm) contents of soil samples.

	ETPH					
	N	Min	Max	Mean	Median	SD	Skewness
Whole data set	151	2.36	4.01	3.50	3.62	0.41	-0.50
Calibration dataset	112	2.36	4.01	3.49	3.60	0.42	-0.50
Validation data set	35	2.65	3.95	3.53	3.64	0.37	-0.44

Where N: number of samples; Min: minimum value; Max: maximum value; SD: Standard deviation

3.3.2. Results and Discussion

The best models calibrated and used for the prediction of ETPH values each for MLR, PCR, PLSR and SVMR are presented in figure 3.3(a-d). All four regression methods generated models with very good prediction statistics having high coefficient of determinations (R^2) between measured and predicted ETPH values (MLR = 0.90, PCR = 0.94, PLSR = 0.94, SVMR = 0.96; $p<0.05$) .Comparing the four models using the validation RMSE and RPD values, the SVMR model had the lowest RMSEv of 0.07 Log_{10}ppm and the highest RPD value of 4.67. Similar performances was recorded for PCR model (RMSEv = 0.10Log_{10}ppm; RPD = 3.91) and PLSR model (RMSEv = 0.08 Log_{10} %SOC; RPD = 4.14) while MLR preformed the least (RMSEv = 0.12 Log_{10}ppm; RPD = 2.95).

The MLR model had the worst performance, although it had very good model quality statistics. In general, the use of simple models such as the MLR is preferred in studies, and some non-linear problems within datasets have been adequately solved with linear regression methods instead of non-linear regressions (Thissen et al., 2004; Shao and He 2011). Some studies have also reported poor quality models calibrated using the MLR method especially when working with datasets having nonlinear relationships and multi-collinear variables as observed with visible near infrared diffuse reflectance spectra and ETPH contents of soils (Shi et al., 2012). While linear regression methods such as the PLSR and the PCR are able to overcome the problem of multicollinearity through variable compression and statistical rotations (Esbensen et al., 2000), the MLR can only solve this problem by using a lower number of variables. This may, however, result into a removal of wavelengths that may be significant to the prediction of the soil parameter under investigation.

Within this study, PLSR and the PCR models performed better than the MLR model despite all three being linear regression methods. Linear regression methods are based on the assumption that the relationships between variables are linear. However, such an assumption is not always valid, thus the linear approaches like the MLR may fail to adequately represent such relationships (Vohland et al. 2011). The PLSR and PCR regression methods are capable of modelling non linear relationships by considering models with more latent factors. This should, however, be done with caution as increasing the number of latent factors or principal components within a model increases model complexityand results into over-fitting, a situation in which the model equation is data dependent and not always able to give good predictions (Esbensen et al., 2000; Rennan et al., 2009). This can be visualised in figure 3.4, where PLSR model quality parameters were observed to improve up to a maximum of 6 latent factors but declined thereafter.

Several studies have reported that the SVMR, a nonlinear regression method, is more capable of modelling the nonlinear relationships between soil components and soil reflectances and as such should have better estimation accuracies when compared to linear regressions such as the MLR and the PLSR (Rossel and Behrens 2010; Stevens et al. 2010, Thissen et al., 2004). However, not much difference was observed between the predictive qualities of SVMR, PLSR and PCR within this study. Similar results have been reported by Shi et al. (2012) and Vohland et al. (2011).

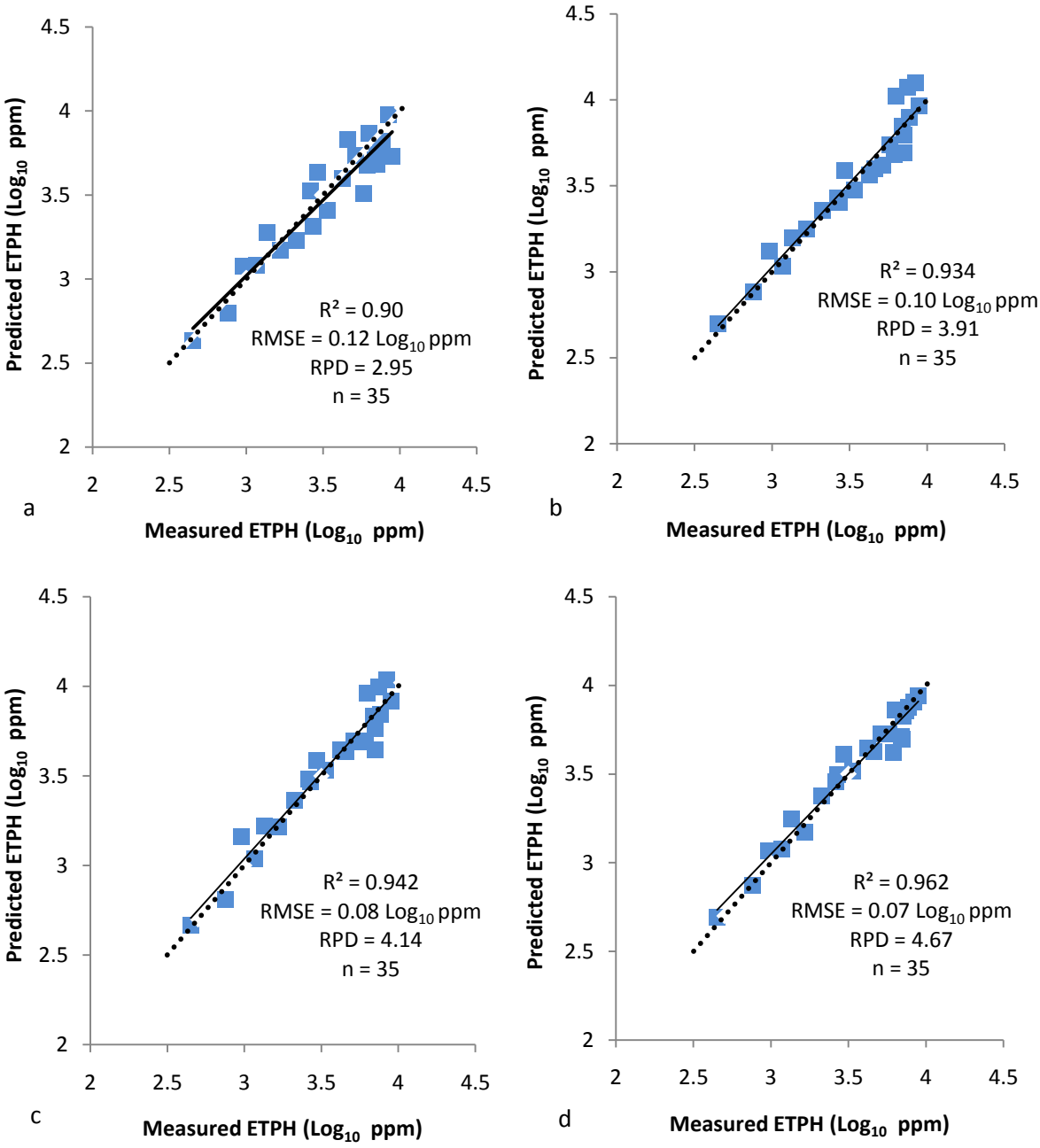


Figure 3.3: Scatter plots of the measured vs. predicted extractible total petroleum hydrocarbon (ETPH) contents (Log_{10} ppm) for the (a) stepwise multiple linear regression (b) principal component regression (c) partial least squares regression and (d) support vector machine regression methods. Solid lines represent the regression lines while dashed lines represent the 1:1 lines. n = number of validaion samples.

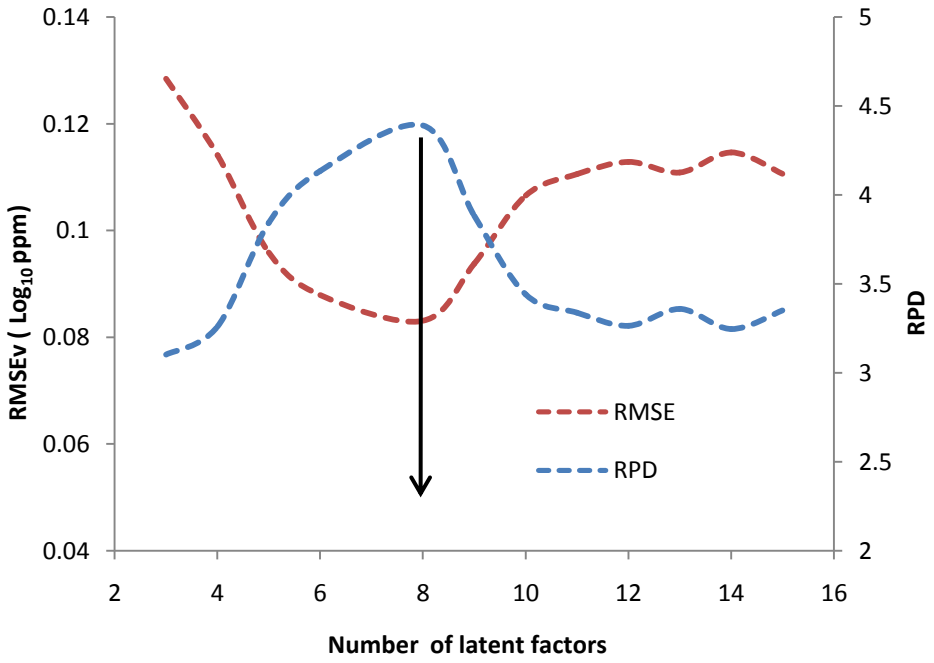


Figure 3.4: Validation root mean square errors (RMSEv) and relative percent difference (RPD) values for ETPH estimation using partial least squares regression (PLSR) with different latent factors. Drop down arrow indicates optimum PLSR model,

3.4. Conclusion

This study aimed to evaluate the performance of visible near infrared diffuse reflectance spectroscopy under different dataset split patterns for the estimation of extractible total petroleum hydrocarbons. Using partial least square regression indicated that using 76% of the dataset for model calibration produced the best model. This ensures that a considerable proportion of the dataset are used both to build and validate the models.

All the four regression methods studied gave good prediction statistics for the dataset used (R^2: 0.90 – 0.96; RPD: 2.95 – 4.67). Furthermore very little difference was observed in the performance of the support vector machine regression, principal component regression and the partial least square regression, indicating that linear regression models such as the partial least square regression and principal component regression, like the support vector machine regression are also capable of modeling nonlinear relationships

This study to access the performance of visible near infrared diffuse reflectance spectroscopy utilized agricultural soil collected from a single farm and homogenized before analysis. However, soils are inherently heterogeneous with composition varying in space and contributing to its overall spectral reflectance. Therefore, additional research is required, incorporating soils from wider geographical locations with differing land use types to enlarge the calibration/validation dataset.

References

Brown, D., Bricklemyer, R. and Miller, P. (2005). Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. *Geoderma*129: 251–267.

Clark, R.N.(1999). *Spectroscopy of rocks and minerals, and principles of spectroscopy*. In: Rencz, N. (Ed.), Remote sensing for the earth sciences: Manual of remote sensing. John Wiley and Sons, New York, pp. 3 - 52.

Esbensen, K., Guyot, D. and Westand, F. (2000). *Multivariate data analysis in practice*, 4th edition, CAMO, Norway.

Esbensen, K. And Geladi, P. (2010). Principles of proper validation: Use and abuse of re-sampling for validation. *Journal of Chemometrics* 24: 168–187

Kokaly, R.F., Clark, R.N., 1999. Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. *Remote Sens. Environ*. 67 (3): 267-287.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). *Cross-validation pitfalls when selecting and assessing regression and classification models*. *Journal of Chem. informatics*, *6*(1):10.

Li, R.H., Zhou, S.L., Song, J.B., Ye, F. and Zhu, Q. (2004). Quantitative determination of index areas and participating factors for farmland gradation (in Chinese). *Act Pedologica Sinica* 41:517–521.

Mourad, M., Bertrand-Krajewski, J. and Chebbo, G. (2005). Calibration and validation of multiple regression models for storm water quality prediction: data partitioning, effect of data sets size and characteristics. *Water Science and Technology* 52(5): 61-68.

Reeves, J.B., McCarty, G.W. and Mimmo, T. (2002). The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soils. *Environ. Pollut*. (116): 277–284.

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

Russel, C.A. (2003). Sample preparation and prediction of soil organic matter properties by near infrared reflectance spectroscopy. *Communications in soil science and plant analysis* 34(11):1557-1572.

Shao, Y.N. and He, Y. (2011). Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy. *Soil Res* 49:166–172.

Shi, T., Cui, L., Wang, J., Fei, T., Chen, Y. and Wu, Y. (2012). Comparison of multivariate methods for estimating soil total nitrogen with visible near-infrared spectroscopy. *Plant Soil* 366:363–375.

Stevens, A., Udelhoven, T., Denis, A. and Tychon, B. (2010). Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* 158:32–45.

Tissen, U., Pepers, M., Ustun, B. and Buydens, L. (2004). Comparing support vector machines to PLS for spectral regression application. *Chemometrics and intelligent laboratory systems* 73(2): 169-179.

Udelhoven, T., Emmerling, C. and Jarmer, T. (2003). Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: a feasibility study. *Plant Soil* 251:319–329.

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

Williams, P.C. (2001). *Implementation of near infrared technology*: in Near - infrared technology in the agricultural and food industries, ed. P. C. Williams and K. Norris, Chapter 8:145 - 170

.

Chapter 4. Rapid characterization of soil organic carbon quality using visible near infrared diffuse reflectance spectroscopy (VNIR DRS)

4.1. Introduction

Soil analysis for organic carbon content (SOC) is a key component of sustainable soil fertility management. Organic carbon in the soil enhances soil structure, water retention as well as serving as the driving force of all biological activities, providing energy and nutrients for soil organisms (Craswell and Lefroy, 2001). Most often, decisions whether or not to deploy specific soil management procedures are usually made based on the results of soil carbon analysis.

Conventional procedures for characterising soil organic carbon, such as chromate oxidation and combustion, are expensive and time consuming (McCarty *et al*., 2002; Watson *et al*., 2000; Nelson and Sommers, 1996). Loss-on-ignition, which is a cheaper and more rapid procedure, has also been shown not to be totally accurate due to the decomposition of certain mineral fractions such as kaolinite, and iron oxyhydroxides (i.e. ferrihydrite) at the high temperatures required by this method (Lal *et al*., 2001). In addition, these methods are unsuitable for large scale soil monitoring where a large number of samples are required for accurate assessments (McCarty and Reeves, 2006). Thus, these analytical challenges have prompted increase in demand for alternative cheaper, faster and effective methods of soil organic carbon characterisation (McBratney *et al.*, 2006; Shepherd and Walsh, 2002).

One of the emerging alternatives is the use of spectroscopy. Over the past two decades, there has been a rise in research on the use of spectroscopy, combined with chemometrics, to quantify a variety of soil properties such as soil mineralogy, organic matter, plant nutrients, soil texture and soil pollutants, such as petroleum hydrocarbons (Brown *et al*., 2006; Wetterlind *et al*., 2008b).

The approach permits rapid and cost-effective quantification, as compared to the traditional costly and time consuming chemical analyses (Schwartz *et al*., 2009). This approach has also been shown to yield comparable results to the traditional methods (Bellon-Maurel and McBratney, 2011). By utilizing the electromagnetic spectrum, radiation reflected from soils can be regressed against measurements of soil components such as organic carbon determined by conventional soil analysis. The constructed statistical models, if stable and strong, can then be successfully utilized to quantify the unknown soil parameter in new soil samples.

Several studies have reported good calibration models for analysis of soil organic carbon using VNIR DRS (Knox *et al*., 2015; McDowell et al., 2012; Ge et al., 2011; Sarkhot et al., 2011; Reeves, 2010; Vasques et al., 2009, 2010; Chang et al., 2001; Brown et al., 2006 and McCarty et al., 2002). Most of these studies have been done on samples having similar soil composition and spectral characteristics. However, soils are extremely variable and the relationship between spectra and soil attributes can be complex and variable in space. It is therefore not certain if the good predictions of organic carbon reported in past studies will also be achieved in studies done on soils obtained from a wider geological area, or if the VNIR DRS approach to analysis of carbon may only be suited for local spectroscopic models built from soils within a given geographical entity or soil type.

It has been suggested that mid infrared spectroscopy (MIR) has a better predictive accuracy for soil organic carbon than near infrared (NIR) derived models (Pirie et al., 2005; McCarty et al., 2002; Reeves, 2010; Viscarra Rossel et al., 2006c). From a review of both NIR and MIR spectroscopic techniques for soil carbon studies, Bellon-Maurel and McBratney (2011) highlighted that MIR predictive models for soil carbon have prediction errors generally 10 – 40% lower than NIR.

Few studies have compared the predictive accuracies of VNIR and MIR derived regression models across a range of soil types, an important research gap addressed in this study. McDowell *et al*. (2012), compared partial least square (PLSR) regression models generated from VNIR and from MIR spectra and concluded that both models had comparable prediction qualities (R^2: 0.95, 0.94 ; RMSE: 2.80%, 3.08%; RPD: 4.25, 3.91 for VNIR and MIR respectively) although VNIR PLSR models gave slightly better predictions. However, Knox *et al*., (2015) evaluated the potential of VNIR, MIR and a combined VNIR–MIR spectral region to estimate and predict soil C, and reported that the MIR spectral region produced slightly better results than the VNIR spectral region (R^2 : 0.84, 0.92 ; RMSE: 0.41, 0.30 log g .kg^{-1}; RPD: 2.4, 3.4 for VNIR and MIR respectively).

The MIR spectra is characterised by well-defined features, compared to the broad overlapping features in the VNIR spectra, and may be well suited for prediction purposes due to its higher specificity. It has, however, been suggested that the visible near infrared interacts better with soil organic carbon than the infrared region alone and, therefore, gives a better prediction when used in calibrating models for carbon (ViscarraRossel *et al*., 2006). Lastly, the organic carbon content of soils influences soil colour, an important factor which may improve predictions for SOC in the visible region (Udelhoven et al., 2003), thereby supporting the need to compare the predictive capabilities of MIR to VNIR for the analysis of soil carbon.

Thus, the objectives of this study were as follows:

1. Characterise soil organic carbon using VNIR diffuse reflectance spectra
2. Investigate the effect of soil spatial variability on the viability of VNIR diffuse reflectance spectra to characterise soil organic carbon

3. Compare the predictive capabilities of MIR and VNIR diffuse reflectance datasets for prediction soil organic carbon.

4.2. Methodology

4.2.1. Soil Collection and Processing

Three groups of soils were used for this study; these were selected to represent different degrees of spatial variation that may be expected in soil analysis. The groupings were:

1. Wisley farm soils: 80 agricultural topsoil samples collected from the Royal Horticultural Society experimental plots at Wisley (51°19'24.34"N; 0°28'27.81"W). These were collected from 80 experimental beds consisting of eight replicate beds which are being amended with different amounts of fertility amendment in an ongoing fertility assessment.

2. South-west England soils: 53 topsoil samples collected within south-west England at sites with differing land use and land management activities (Figure 4.1). Multiple samples were collected in six of the sampling locations to accommodate the different land uses.

3. England soils: 87 topsoil samples collected at different locations spread across England and Wales

Soils were collected in plastic containers or ziplock bags and refrigerated at 4^0C in the laboratory until processing. They were air-dried and then sieved to 2mm. 10g of homogenised soil samples were then ball milled using the Tema mill which reduces the sample to a fine powder (80% less than 150 microns) (Thompson, 2009) and stored in ziplock bags.

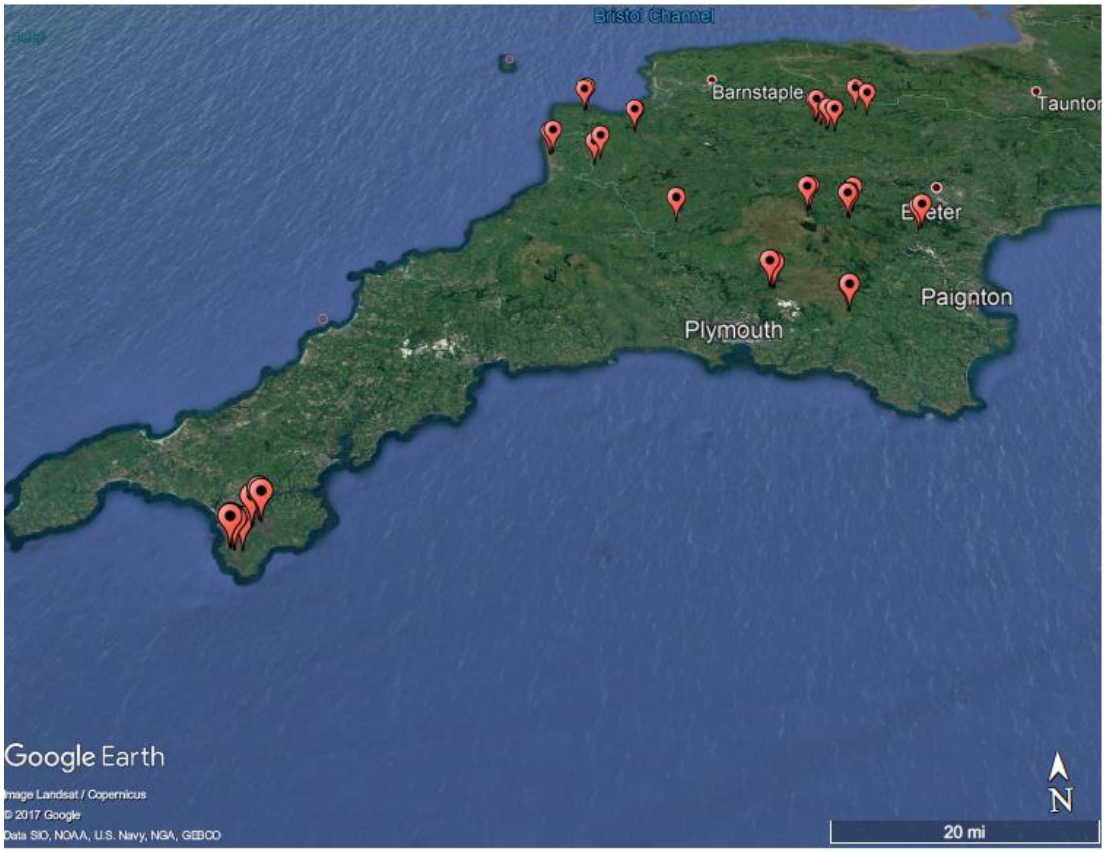


Figure 4.1: Location of soil collection area for South west soils with markers indicating collection sites. This map was created using the **Google Earth** software, August 4, 2017.

4.2.2. Soil Chemical Analysis

Total organic carbon content was determined in 10mg of ball milled soil subsamples using a Flash 2000 organic elemental analyser(Thermo Scientific Flash 2000) with a reproducibility of 0.07% calculated using the equation below (Fearn, 2008):

$$SE = \sqrt{\frac{\sum_{i=0}^{n} E_i^2}{2n}}$$

Where *SE* is the standard error of the instrument, E_i is the difference between measurements and the estimated true value of a certified reference material.

Prior to analysis using the Flash 2000 organic elemental analyzer, a simple test for carbonates was done by checking the pH of soil samples. When soil samples possess pH less than 7.4, it is concluded that the samples contain none to insignificant quantities of carbonates (Schumacher, 2002). Otherwise, samples are initially dried at 105^0C overnight and treated with a combination of H_2SO_4 and $FeSO_4$ to remove carbonates. All soils within this study had pH less than 7.4 Therefore:

$$\%\, Total\ carbon \cong \%\, Total\ organic\ carbon$$

4.2.3. Spectral Measurement

4.2.3.1. Visible, Near-Infrared Diffuse Reflectance Spectroscopy

Diffuse reflectance spectra of soil samples were obtained in a dark room with a GER3700 VNIR spectrophotometer (350 - 2500nm) coupled with a quartz-halogen balanced daylight source. The

spectrophotometer has one Si array (350 - 1050 nm) and two Peltier-cooled InGaAs detectors (1050 - 1900 nm and 1900 - 2500 nm). Spectral sampling interval of the instrument was 3nm at 350 - 1050 nm, 7nm at 1050 - 1900 nm and 9.5nm at 1900 - 2500 nm. The light source projected at 45^0 to the detector.

A Spectralon white reference panel was scanned before each measurement to convert radiance measurements of the soils to calibrated reflectance. Scans were taken at nadir, with illumination at 45°, from the soil samples tightly packed in borosilicate petri dishes (figure 4.2). Replicate measurements were collected in four illumination directions by rotating the petri dish at an angle of 90^0, to increase precision of measurements. Each scan was an average of 16 internal scans. The 64 replicate scans were then averaged to produce a single spectrum for each sample.

4.2.3.2 . *Mid infrared diffuse reflectance spectroscopy*

Mid infrared spectra of soils were collected from South west soils that had been reduced to fine powders (about 80% less than 150 microns) by a Tema mill using a Nicolet iS10 FT-IR spectrometer (Thermo Fisher Scientific Inc., Madison, WI, USA). Spectral acquisition was performed by diamond attenuated total reflectance (MIR-ATR) spectroscopy over the spectral range 2500 –16660 nm (4000–650 cm^{-1} wavenumbers) with spectral resolution of 4 cm^{-1} and 16 scans per replicate.

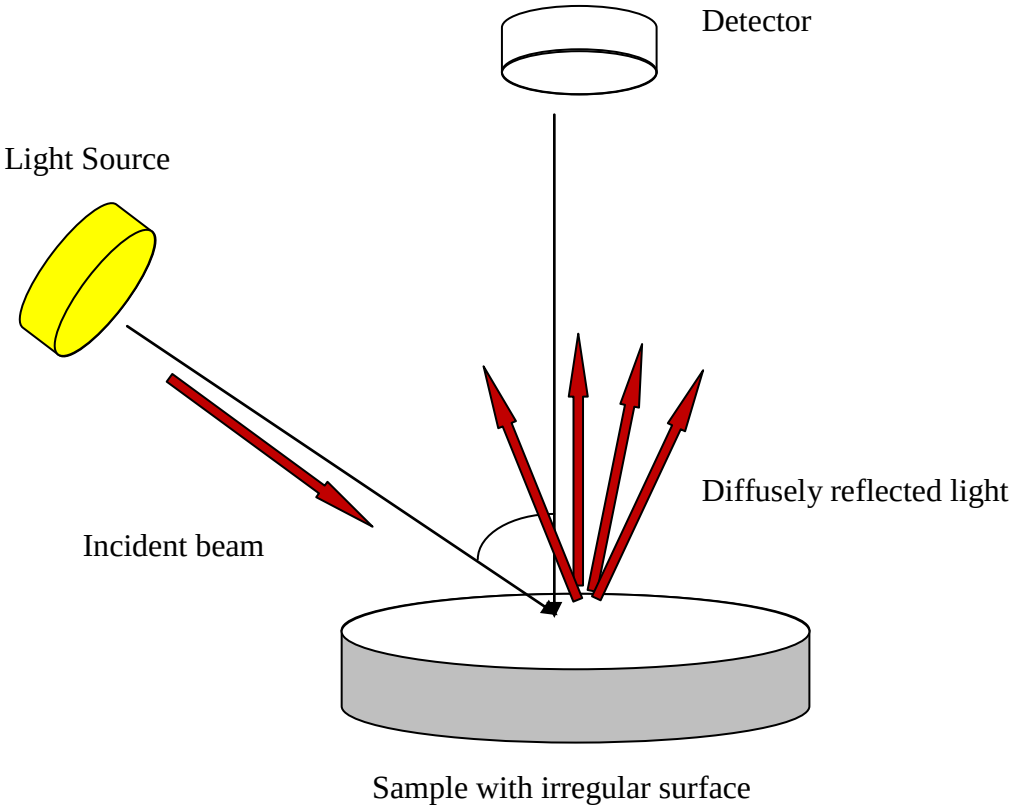


Figure 4.2: Schematic diagram of scanning arrangement

4.2.4. Data Analysis and Model Development

All spectral pre-treatment and multivariate calibration/validations were carried out using the Unscrambler-X 10.3 analytical software developed by CAMO.

PCA analysis was done on data sets to observe groupings within soil samples and to identify the presence of outliers. Possible outliers were identified as samples with very high leverages and residuals. A spectral range for the data set was calculated (mean ± 2 SD) and each individual spectrum compared to it. If more than 75% of the spectrum fell outside of this range, the sample is considered a possible outlier and excluded from further modelling.

Prior to modelling, a small section (1905nm – 1928nm) of each spectrum was removed, due to an offset in reflectances between the detectors. The spectra is then averaged (i.e. every 5 wavelength points) to give 120 predictor variables for model calibration. Due to the scattering effects and pathlength variations inherent in diffuse spectra, four common spectral pre-processing techniques were applied to spectra and compared to spectra that had only been averaged. Techniques tested include $Log_{10}(1/R)$ transformation, standard normal variate coupled with the detrending (SNV-DT) which reduces scattering effects and linearizes the spectra, Savitzky-Golay smoothing of the spectra (SG), and a combination of Savitzy-Golay and derivatization which smoothes data, enhances resolution, boosts weak signals and eliminates linear baseline drift between spectra.

4.2.5. Model Calibration/ Validation

A total of four partial least square regression (PLSR) models were developed for the relationship between soil spectra and the measured soil organic carbon concentrations using each of the data sets along with a fourth combination of all soil spectra. 76% of each of the data sets were selected for model calibration while the remaining 24% were used for model validation. This approach to model validation is recommended as a robust procedure by Varmuza and Filzmoser (2009). Calibration and validation samples were carefully selected to be similar to each other, both in terms of the range organic carbon content and the land management practice at sites of soil collection.

Predictive accuracy and stability of PLSR models were evaluated using the coefficient of determination R^2, relative percent difference (RPD) values, root means square error (RMSE) values, number of latent factors for the optimum model and the percentage prediction error (% PE).

R^2: This is the square of Pearson Product Moment correlation coefficient. It measures the relationship between the modelled and the measured data RMSE: This is a measure of the average deviation between modelled and measured data. $RMSE = \sqrt{\frac{\sum(TPHp - TPHm)^2}{n}}$ where *TPHp* are the modelled values, *TPHm* are the laboratory measured values and *n* is the number of validation samples.

% prediction error: Due to the dependence of RMSE on individual data sets, this statistics measures the prediction error as a percentage of the highest data point within the validation data set. It allows for the comparison of the error term between data sets.

$$\% PE = \frac{RMSE}{Xh_v} \times 100$$

Where $Xh_v$is the largest measured data point within the validation set.

RPD: This is a measure of predictive accuracy.

$$RPD = {}^{SD}/_{RMSE}$$

where SD is the standard deviation of the predicted values. According to Viscarra Rossel et al. (2006) and Chang et al. (2001), very poor models show RPD < 1.0; poor models: 1.0 ≤ RPD ≤ 1.4; fair models: 1.4 ≤ RPD ≤ 1.8; good models: 1.8 ≤ RPD ≤ 2.0 and very good models: 2.0 ≤ RPD ≤ 2.5; and excellent models have RPD > 2.5.

4.3. Results and Discussion

4.3.1. Laboratory analysis

Table 4.1 shows the distribution of samples selected for model calibration and validation. Original soil carbon data was $\log_{10}$– transformed to normalize its distribution (Figure 4.3). Consequently, all PLS models were developed based on Log_{10} – transformed soil carbon data that approximated a Gaussian distribution after stabilizing the variance.

Table 4.1: Descriptive statistics of soil organic carbon (%) in the data sets used for partial least squares regression (PLSR)

	Wisley			South west England			Across England		
	Range	Mean ± SD	Skewness	Range	Mean ± SD	Skewness	Range	Mean ± SD	Skewness
Total	2.38 - 18.80	7.88 ± 4.14	0.77	1.87 - 30.90	7.08 ± 5.18	1.72	1.19 - 11.37	4.33 ± 1.74	1.67
Calibration	2.38 - 18.80	8.13 ± 4.25	0.76	1.87 - 30.90	6.99 ± 5.15	1.81	1.19 - 11.37	3.28 ± 1.90	1.63
Validation	2.86 - 15.52	7.13 ± 3.77	0.70	2.50 - 24.02	7.37 ± 5.45	1.54	2.38 - 6.24	4.13 ± 1.09	1.10

Where SD means standard deviation

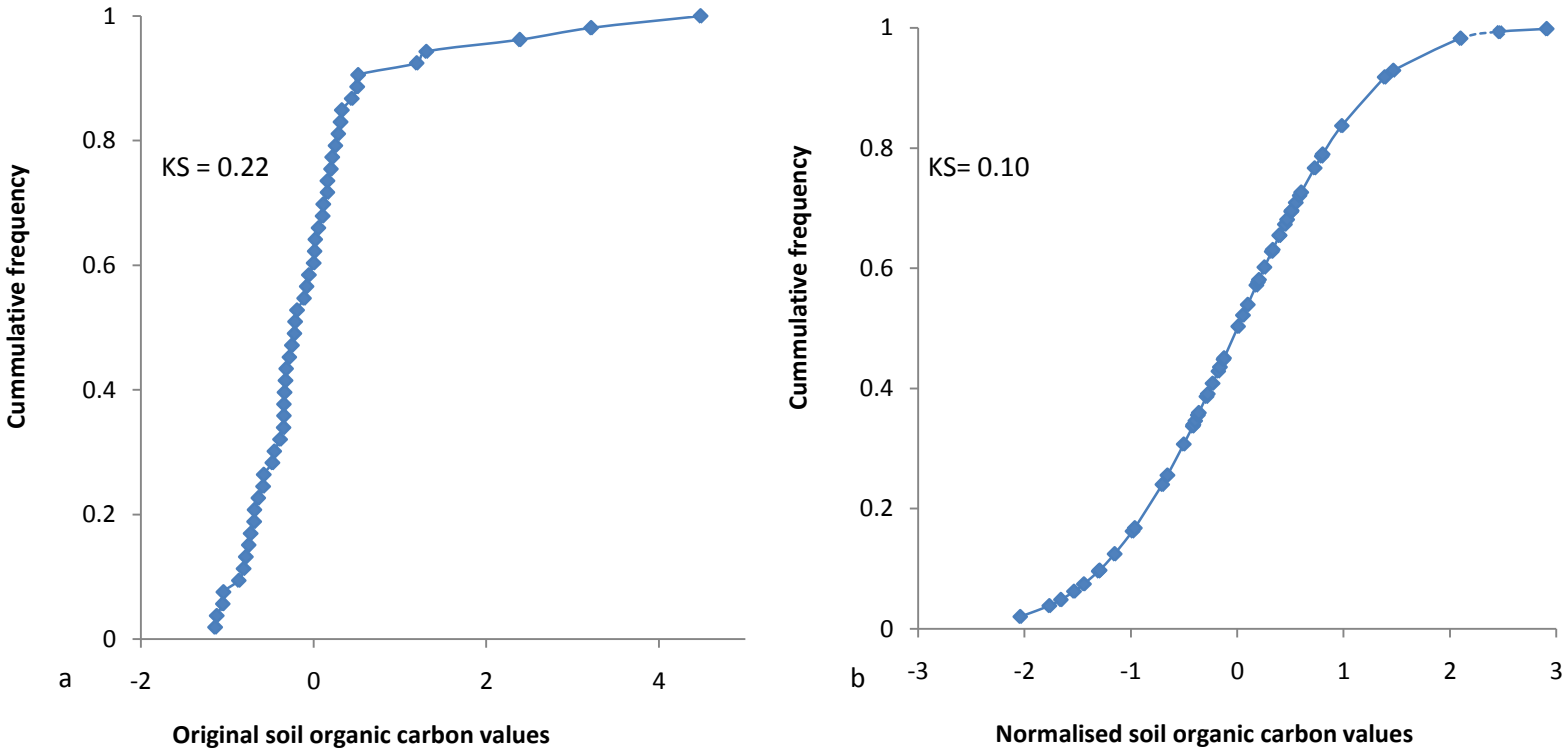


Figure 4.3: Cumulative fraction plot of (a) Original soil organic carbondata (b) Log_{10}-transformed soil organic carbon data. A reduction in the Kolmogorov-Smirnov statistic for log_{10}-transformed data shows that the Log_{10}- transformation improved the original organic carbon data.

4.3.2. Qualitative Description of Spectra

Figure 4.4 illustrates the shape and variations observed within the VNIR spectra of soils. To compare soil spectra from a common baseline and allow for identification of absorption features that can be linked to soil organic carbon content, the continuum removed technique was applied to the spectra. The continuum removal technique involves fitting a maxima curve over a spectrum to form a continuum or Hull. A mathematical function is then fitted to the spectrum to represent absorptions due to processes other than those of interest. These effects can then be removed by dividing or subtracting the measured spectra from the continuum function (Green and Craig, 1985)

VNIR spectra of all soil samples follow a similar shape with absorption minima at ~1400–1450 nm and ~1910–1930 nm characteristic of -OH and H_2O (Fortes and Dematte, 2006). The steep slope observed at ~400 - 900 nm has been associated with iron and iron oxide minerals such as goethite and hematite (Vaughan, 1996) but can also contain overtones of C-H, N-H absorptions of alkyls, aromatics and amines (Viscarra Rossel and Behrens, 2000). Slight absorptions around 1626 – 1750 can be linked to C-H bonds (Stuart, 2004) while absorption features between ~2100 and 2500 nm may be attributed to C-O absorptions of polysaccharides, C-H_2, C-H_3, NH_3 absorptions of alkyls and aromatics and Si-OH and –OH absorptions in minerals (Clark et al., 1990).

Assigning specific functional groups to absorption features within the VNIR spectrum may be difficult due to the numerous organic and inorganic molecules that absorb within this region. However, the continuum removed spectral reflectances in figure 3; show a broad but clear absorption feature at ~ 510 – 950 nm that can be linked to increasing soil carbon content. In general, soils with higher carbon content were observed to be darker in tone and had lower

reflectances across all wavelengths of measurement and this may be responsible for the clear separation of soil samples within the visible region. Other soil properties that may influence overall reflectance of soils include moisture content, particle size, type of organic matter in soil (Stoner and Baumgardner,1981) and cation exchange capacity (Brady, 1986).

Other absorption features that may be associated with soil carbon identified within the spectra include ~1070 – 1250 nm, 1600 – 1800 nm, ~2000 – 2200 nm and ~2230 – 2400 nm. These absorption features have been assigned as combination bands, first and second overtones of C–H stretch fundamentals of amines, carboxylic acids, amides, aliphatics, phenolics and polysaccharides within the mid infrared region (Viscarra Rossel& Behrens, 2010). They have also been proposed as being particularly significant for calibrating regression models for soil organic carbon and total nitrogen (Malley et al., 2000; Martin et al., 2002; Stenberg, 2010).

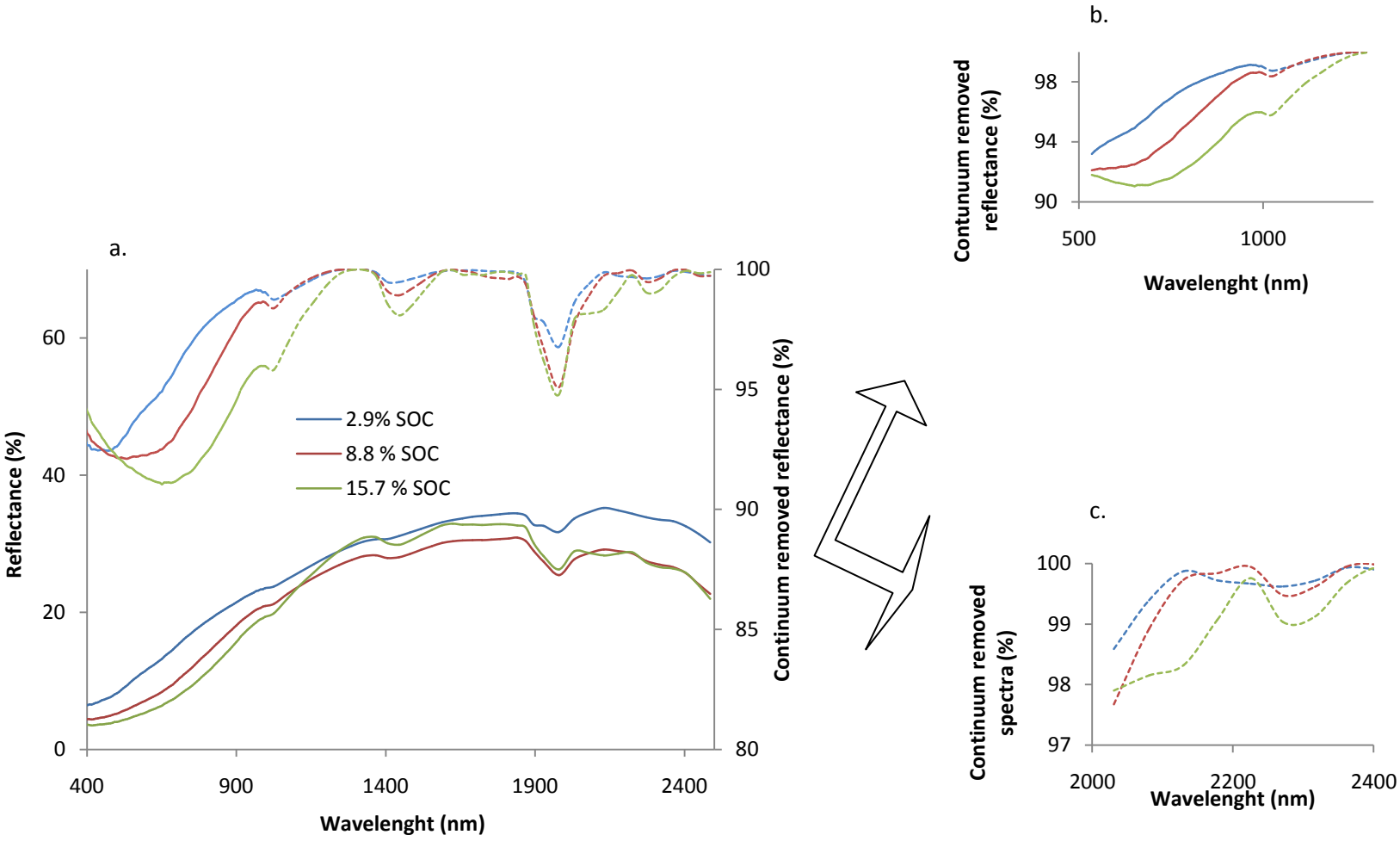


Figure 4.4: (a) Mean reflectance (solid lines, left scale) and continuum removed (dotted lines, right scale) visible, near-infrared diffuse reflectance spectra of three soil samples with different soil organic carbon content. (b) and (c) indicate wavelength regions that may be important to organic carbon modelling

4.3.3. Performance of Partial Least Square - VNIR DR models

The best model for the Wisley soils was achieved using the standard normal variate – detrending transformed spectra, while Log (1/R) transformation improved the Southwest data best. The wider England data were best modelled with a first derivative coupled with Savitzy-Golay smoothing, though the model was unsatisfactory for carbon prediction (calibration $R^2 < 0.25$) while a combination of all three data sets was best left unprocessed. This is suggestive that there may not be a "best" spectral data pre-processing technique and soil spectra should be carefully studied to identify which technique best suits the spectra/soil property under investigation to avoid unintentional downgrading/ screening out of minute absorption features that could be important to the modelling of the soil feature under investigation.

Model quality statistics for the best models from the four data sets are presented in Table 4.3. The PLSR model of Wisley soils performed excellently with a validation R^2 of 0.94 and a RPD value of 3.9. The optimum model had a percentage error of prediction of 5.5%. The PLSR model of South west soils also had a good predictive performance with a validation R^2 of 0.90 and a RPD value of 3.1 and a percentage error of prediction of 5.4%. However, PLSR models developed from the England soils had a very poor calibration and prediction performance ($R^2 < 0.25$, RPD = 0.7) and are therefore not acceptable for organic carbon prediction. A combination of all three data sets had model quality statistics with a validation R^2 of 0.52, RPD value of 1.4 and a percentage error of prediction of 12.2%.

When all calibration models were compared, R^2 values were observed to decrease and both RMSE and RPD values increased with increase in the geographical size of the soil sampling locations (Figure 4.5). Figure 4.6 illustrates the predicted vs measured values for the validation set for PLSR models for the groups. Variation in soil matrix increases with sampling area and the

low performance of PLSR models derived from the England soil spectra can be attributed to the higher geological variability expected in soils collected from such a larger area. This is also observed by Udelhoven et al. (2003) who compared predictions of SOM in two regions of Germany and observed that those based on a data set from the Eifel region (with a higher geological variability) performed less than those based on data from the Hunsruck region. It may therefore, be an important strategy to develop local or regional prediction models for soil organic carbon from geographical areas with a homogeneous geological characteristics rather than developing universal or large scale models.

Table 4.3: Performance of PLSR – VNIR DR models

Data set	Pre-processing	Optimum factor	Calibration		validation			
			R^2	RMSE (Log_{10} %SOC)	R^2	RMSE (Log_{10} %SOC)	RPD	% PE
Wisley	SNV-DT	1	0.63	0.19	0.94	0.07	3.9	5.5
South west England	Log (1/R)	5	0.74	0.12	0.90	0.07	3.1	5.4
Across England	SG – 1^{st} D	3	0.17*	0.27	-	-	-	-
All soils	Raw	6	0.42	0.21	0.52	0.17	1.4	12.2

Raw= averaged raw spectra; $Log_{10}(1/R)$ = Absorbance, where R is the reflectance value; SG = Savitzky–Golay filter with 5 window smoothing ; SG-1^{st} D = first derivative of 5windowsmoothingSavitzky–Golay filter ; SNV-DT = standard normal variate coupled with detrend correction.
*: Poor calibration performance (R^2<0.25)

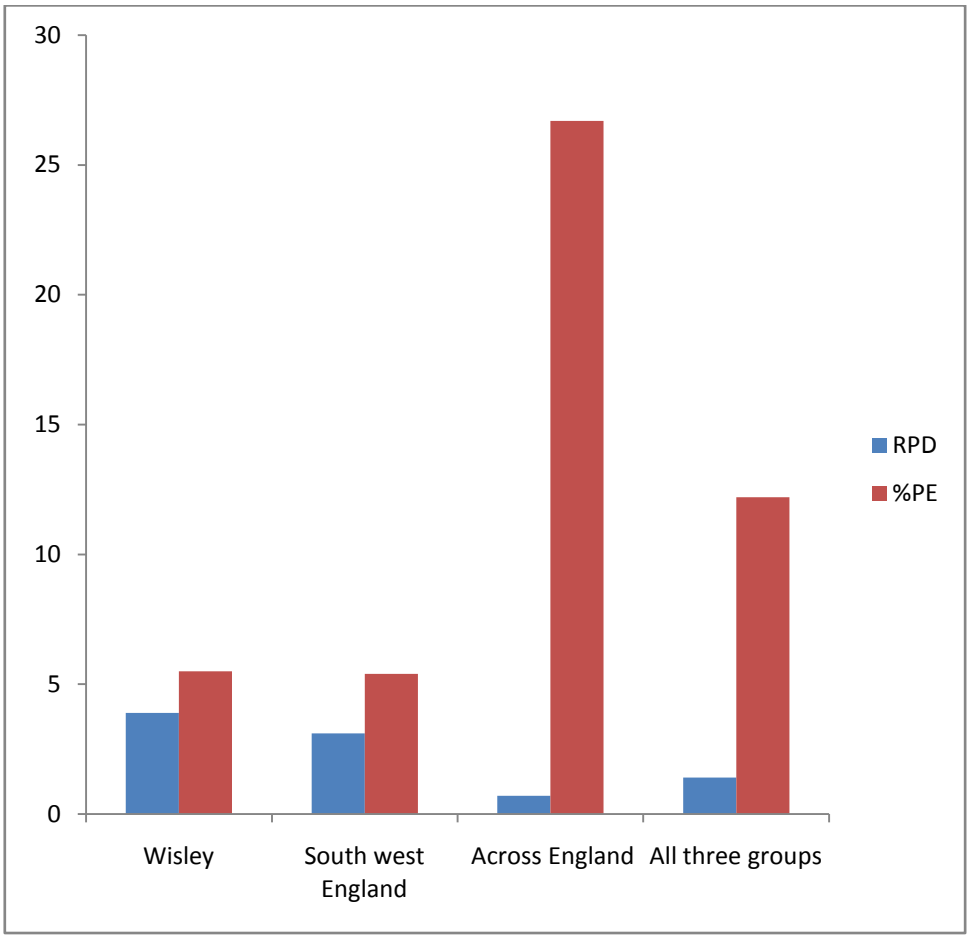


Figure 4.5: Relative percent difference (RPD) and Percentage prediction error (%PE) for best VNIR-PLSR models of soil groups

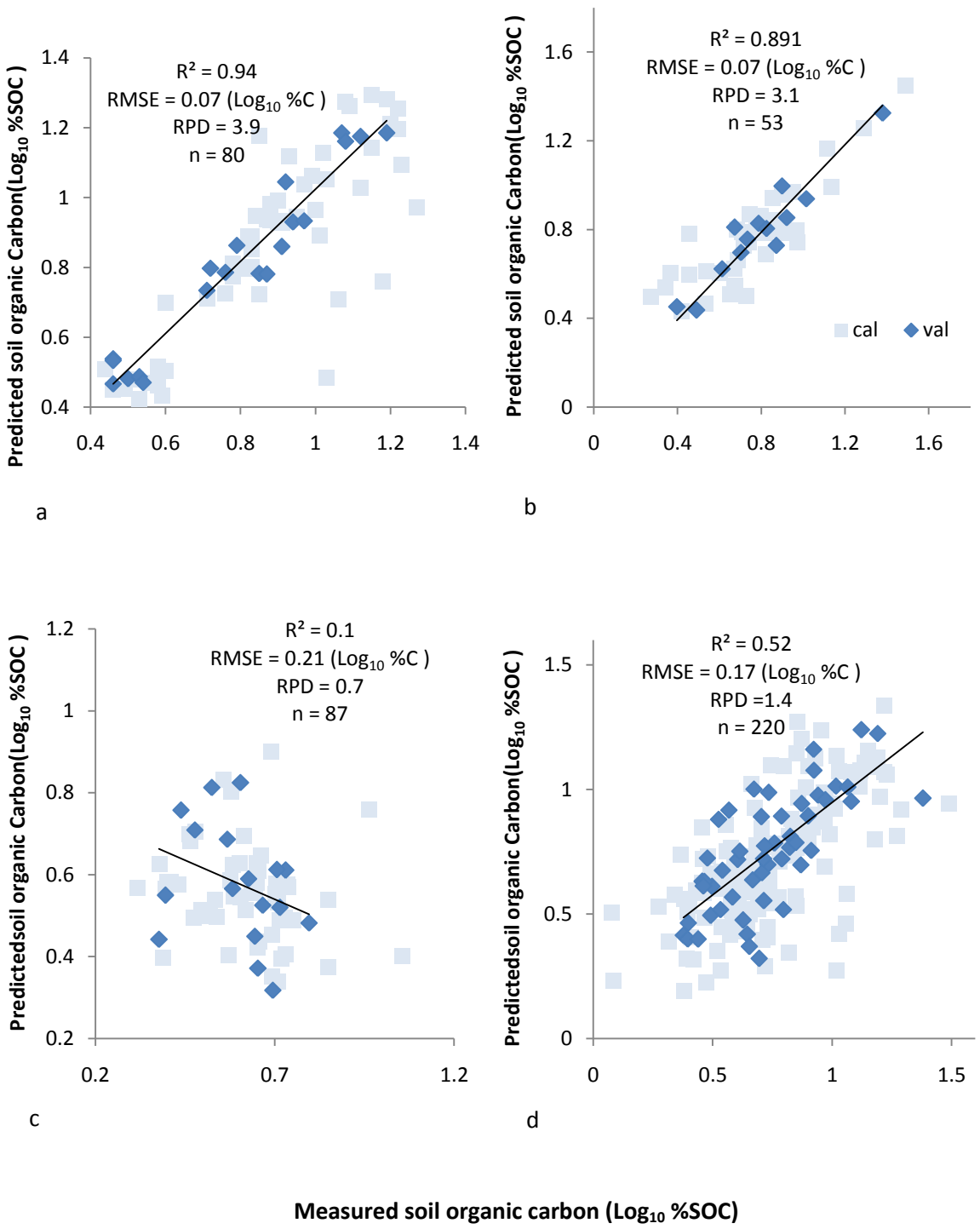


Figure 4.6: Predicted vs measured values of the validation data sets in VNIR-PLSR models of (a) Wisley farm soils (b) southwest England soils (c) across England soils (d) a combination of all three groups. n = number of samples

4.3.4. Important Wavelength in VNIR-PLSR Models

Figure 4.7 shows the line plots of regression coefficients versus wavelengths derived from PLSR analysis for the four data sets. The regression coefficients are approximations of model parameters resulting from linear combinations of the predictors, and the absolute values of the coefficients indicate the relative importance of the wavelength on the basis of explained wavelength-variance within the model. Wavelength variables with large coefficients play an important role in the model, with a positive coefficient showing a positive relationship to the response and a negative coefficient shows a negative relationship (CAMO, 2012). It is important to note that the numbers of important wavelengths within the models is observed to decrease with increase in the geographical area from which soils were collected. This implies that co-variation of soil carbon with other soil properties having direct spectral responses in the VNIR is stronger in soils with similar geological composition, but reduces in data sets obtained from soils from a wider geological area or wider range of land use types.

Important wavelengths within all three acceptable models include ~440 – 500 nm and ~630 -770 nm, implying that the brightness of soils could be important factor in the visible near infrared region for prediction of organic C content. Other important wavelength regions include ~950 – 1000nm, ~1800 – 1900nm, ~2000 - 2200nm and ~2300 – 2490nm. Typical O-H peaks at ~1400 and 1900nm were observed not to be significant. This is in line with the observations of Charkraborty et al. (2010) and Waiser et al. (2007), where water within soil samples did not affect the prediction accuracies of the validation models.

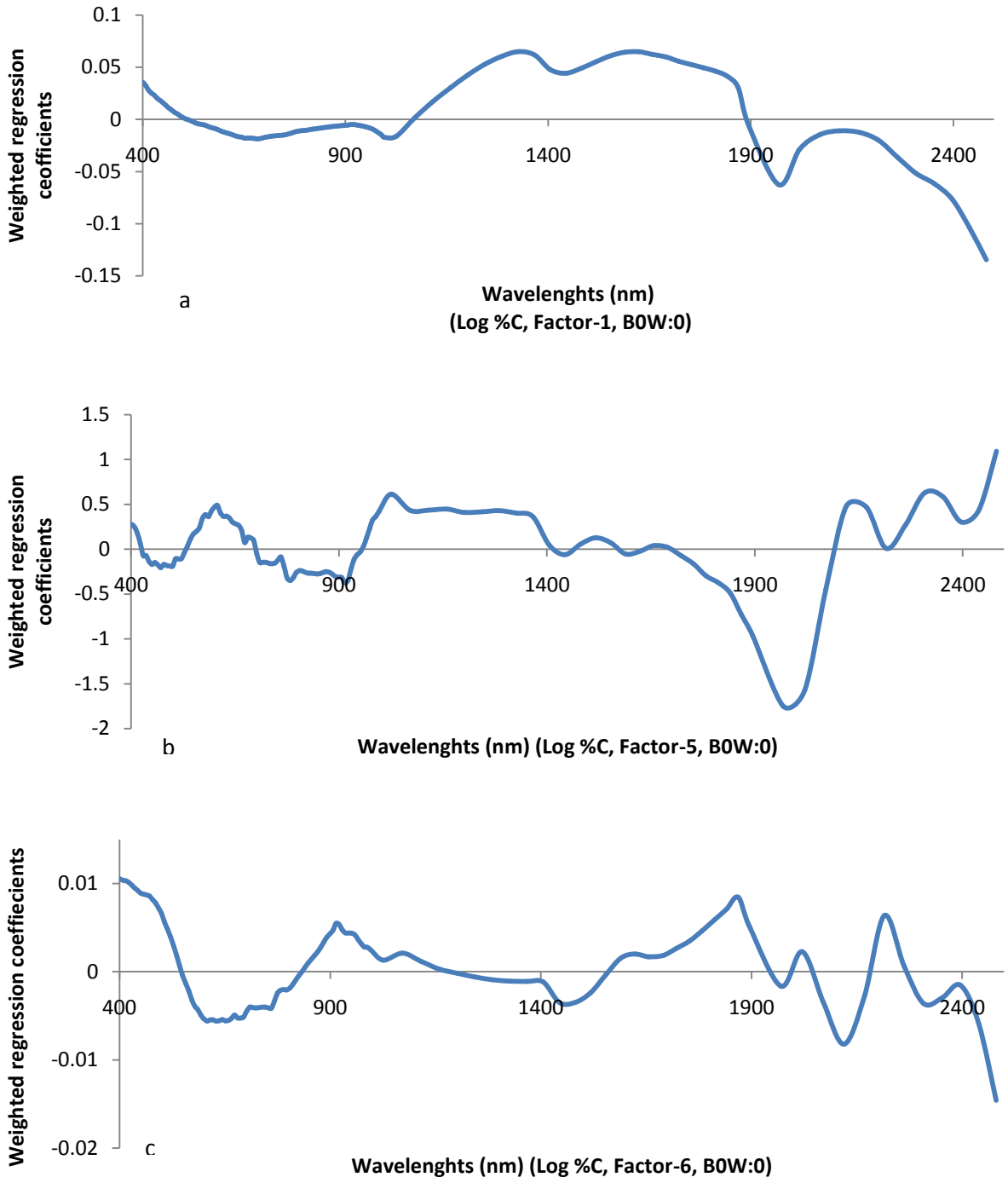
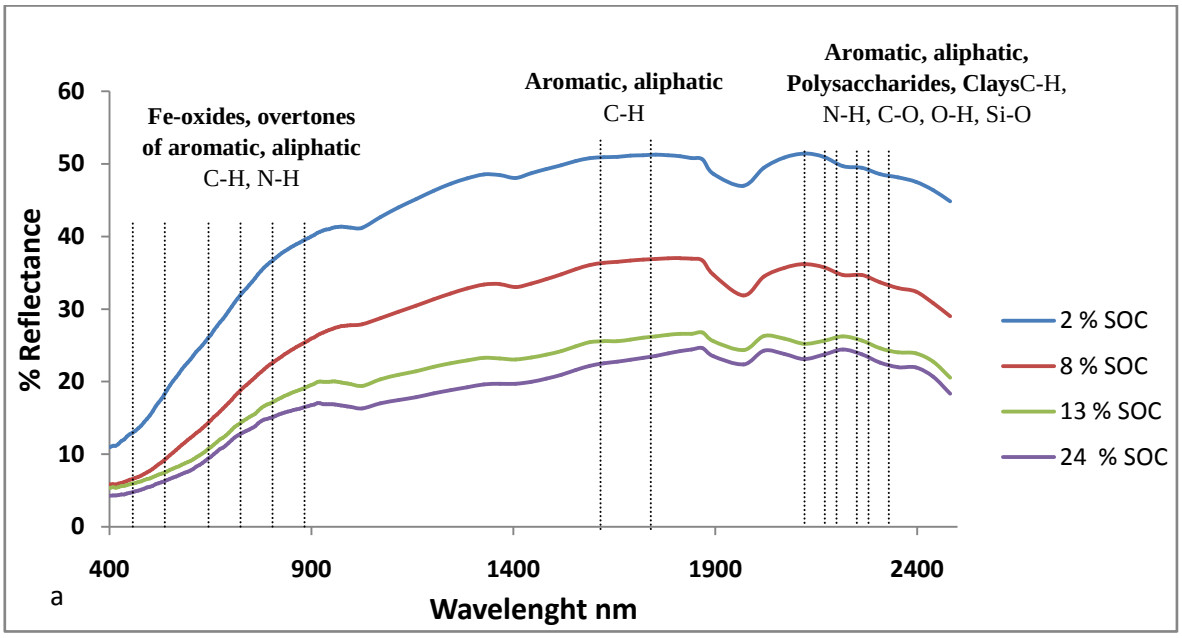


Figure 4.7: Regression coefficients of the best VNIR-PLSR model of (a) Wisley farm soils (b) southwest England soils (c) a combination of all three groups. The magnitude of the regression coefficient at each wavelength is proportional to the importance of the wavelength variable in the model.

4.3.5. Comparison between Mid Infrared (MIR) and Visible Near- Infrared (VNIR) Diffuse Reflectance Spectroscopy

MIR spectra of soil samples covered a spectral range of 2500 –16660 nm (4000 – 600 cm^{-1} wavenumbers) and contained better defined features than the VNIR spectra (Figure 4.8). A prominent feature present in all the spectra is the steep absorption at ~7812 – 9852nm which may be associated with clay minerals such as smectite, illite, and kaolinite, and minimal amounts of silica (Byrappa and Suresh Kumar, 2007). Absorptions around 5680 - 6200 nm can be due to C-H and C-O bonds of organic compounds such as amides, carboxyl, and/or silicate minerals such as quartz and kaolinite as they all have overlapping features in this region (Nguyen et al., 1991; Stuart, 2004). Other features identified include absorptions at ~3500 – 3850nm and at ~ 3070 3360nm that are likely associated with C-H bonds in alkyl groups (Gaffey et al., 1993; Stuart, 2004). The MIR spectra compared with the VNIR spectra of soils used in this study can be seen to be less affected by shifting baselines and clearly separable at 2760 – 3730 nm and at 5540 – 8470 nm.

Both the VNIR and MIR partial least square regression models performed excellently, with validation R^2 values of 0.90 and0.90; RMSE values of 0.07 and 0.08Log_{10} %SOC and RPD values of 3.12 and 3.09 for VNIR and MIR respectively (table 4.4). This result showed that the VNIR and MIR PLSR models are similar, even though validation parameters were slightly are slightly better for the VNIR model. Figure 4.9 compares the measured and predicted values plots for the validation set for both models.



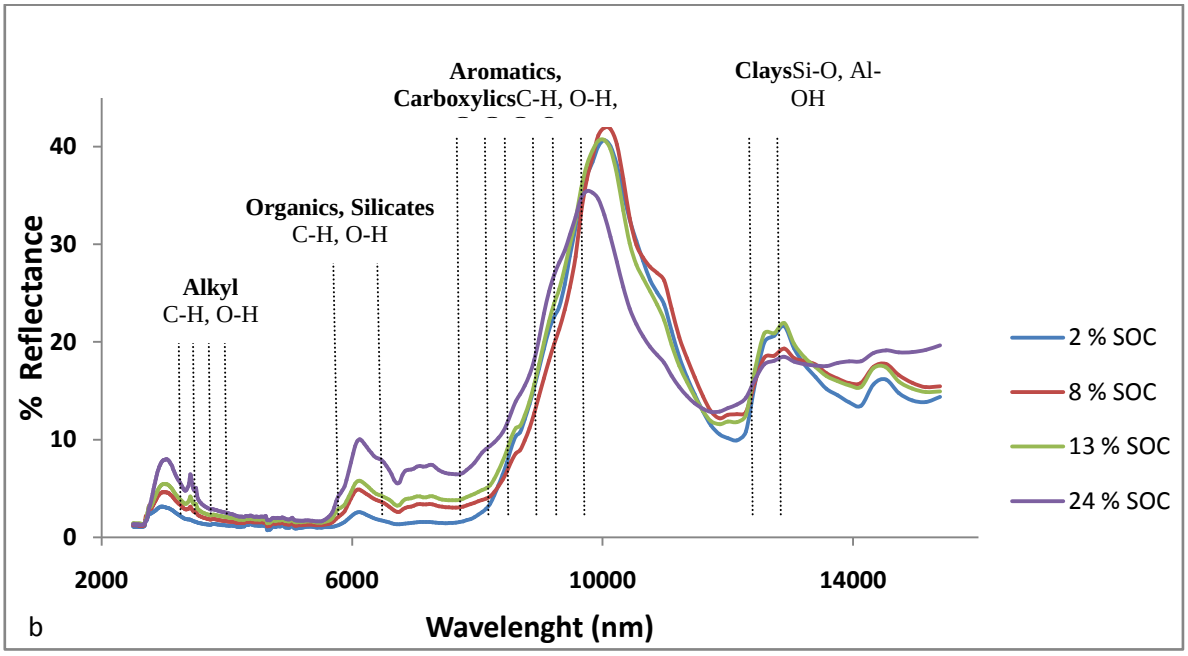


Figure 4.8: a) Visible, near-infrared and b) Mid-infrared diffuse reflectance spectra of Southwest soil samples used for this study showing absorbing functional groups.

Table 4.4: Calibration and validation results of PLSR models to predict total soil carbon from VNIR and MIR diffuse reflectance spectra.

	Optimum factors	Calibration		Validation		
		R^2	RMSE Log_{10} %SOC	R^2	RMSE Log_{10} %SOC	RPD
VNIR	5	0.74	0.12	0.90	0.07	3.12
MIR	2	0.89	0.08	0.90	0.08	3.09

The PLSR model built from VNIR diffuse reflectance spectra for this study compares favourably with results of models developed for carbon by previous researchers. Studies by Mcdowell *et al.* (2012); McCarty et al.(2010); Mouazen et al. (2010); Vasques et al. (2009), Vasques et al.(2010) and ViscarraRossel and Behrens (2010) have reported similar R^2 values. Lower values were also reported by Knox et al. (2015) (0.86/ 2.60), Sarkhot et al. (2011) (0.85 /2.59) and Vasques et al. (2008): (0.79 / 2.14) for R^2 and RPD respectively. A higher R^2 value (0.97) was produced in the by Reeves et al.(2002).

RMSE values of these studies have, however, varied considerably with some lower and others higher than the values obtained within this study. The range and composition of organic carbon in soils used for the various studies have also varied widely e.g. <1 - ~56% in Mcdowell et al. (2015), ~2 – 980% in McCarty et al. (2002), ~5 – 496% in Sarkhot et al. (2011), ~0 – 57.4% in Vasques et al. (2008) and ~2 – 31% within this study. RMSE values are unique to each calibration/validation and most likely will reflect the range of values used in the modelling, contributing to the differences in RMSE values across studies.

Validation R^2/ RPD values for MIR PLSR model within this study compares with those from Mcdowell *et al*. (2012) (R^2 = 0.94; RPD = 3.91), Knox et al. (2015) (R^2 = 0.95 ; RPD = 4.6) , Reeves et al. (2001) (R^2 = 0.93), McCarty et al. (2002) (R^2 = 0.95), Reeves et al. (2002) (R^2 = = 0.98) and McCarty et al. (2010) (R^2 = 0.94).

The relative importance of the variables (wavelengths) within the MIR-PLSR prediction model is illustrated using the weighted regression coefficient chart in Figure 4.10. Important wavelengths within the VNIR PLSR model include 2500 – 2760 nm, ~3430 nm, 5780 – 8070 nm and ~10000 – 11430 nm.

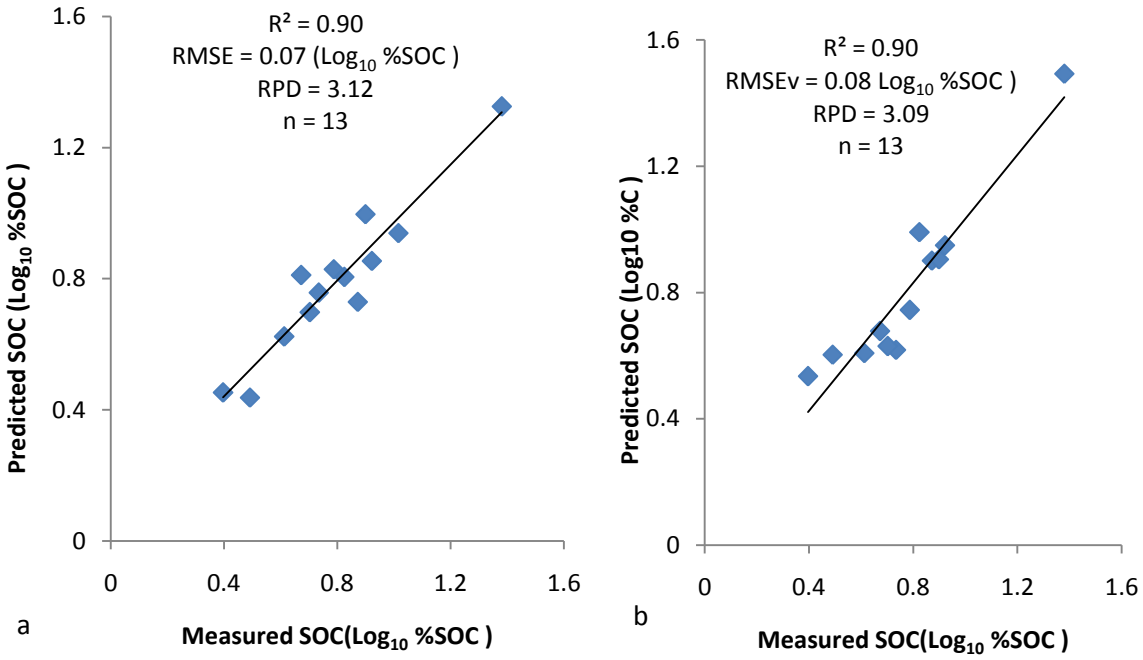


Figure 4.9: Measured vs. Predicted total soil organic carbon values for the validation set of PLSR models from a) visible, near-infrared and b) mid-infrared diffuse reflectance spectra. n = number of South West England samples.

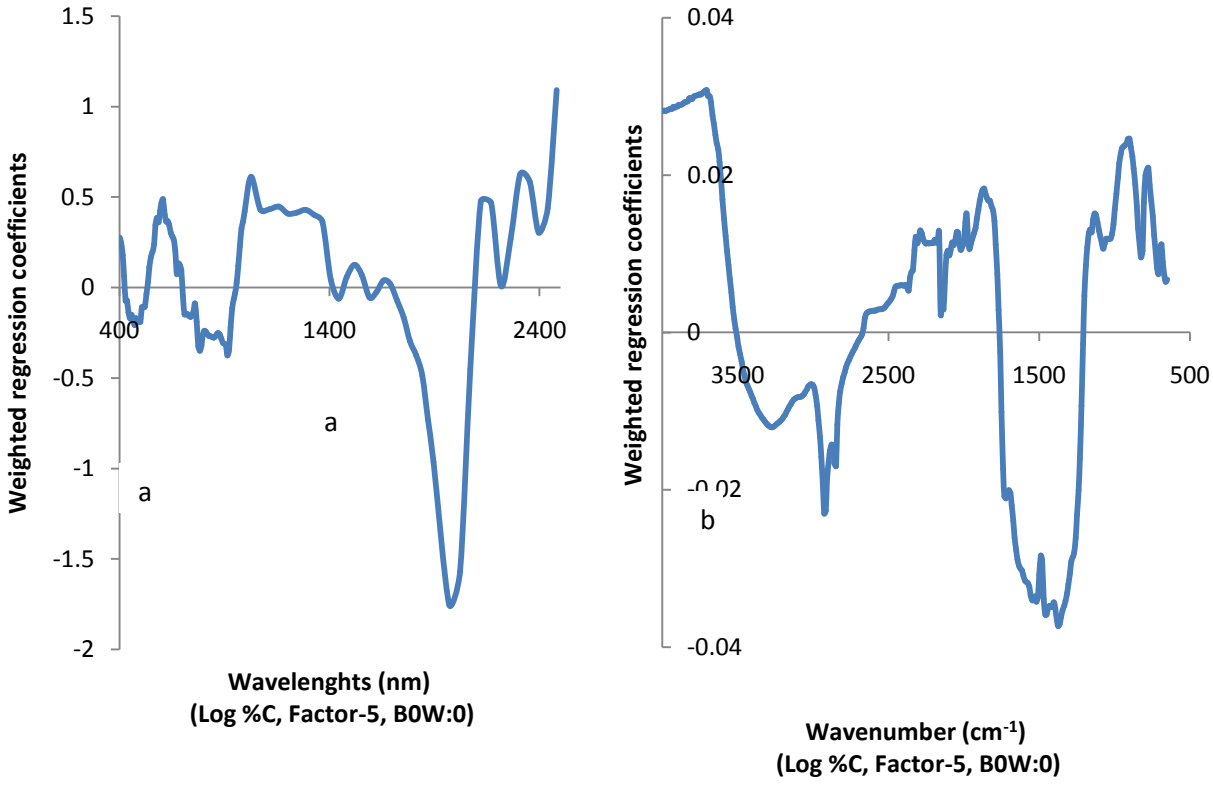


Figure 4.10: Important variables for the prediction of organic carbon in PLSR models of a) visible, near-infrared and b) mid-infrared diffuse reflectance spectra

Some studies have reported a better model for C prediction using the MIR wavelength range (Knox et al., 2015; McCarty et al., 2002; Reeves, 2010; Reeves et al., 2002; Viscarr aRossel et al., 2006). This study, however, demonstrates that the VNIR wavelength range can perform just as well or better than the MIR wavelength region. This is similar to the observation of Mcdowell et al. (2012).

The decision on which spectral range to use for model calibration/prediction depends on: (i) the accuracy of the predictions, (ii) the cost of the technology and (iii) the amount of sample preparation required and (iv) the degree of accuracies required from such analysis. Presently, MIR spectrometers are considerably more expensive and the technology requires more sample preparation compared to portable VNIR fibre optic spectrometers that are developed for field use, which require little to no sample preparation. Therefore, the potential increase in accuracies of MIR models reported by some studies must also be weighed against the greater utility of VNIR spectroscopy, which may be better suited for assessing spatial variability of SOC (Mcdowell *et al*., 2012), or on-the-go proximal soil sensing systems (Viscarra Rossel and McBratney, 1998b; Shibusawa et al., 2001) where the slight gain in prediction accuracy from using an MIR instead of an VNIR spectrometer may not be significant taking into account the higher cost of MIR technology.

The VNIR spectra used in this study were collected from 2 mm sieved soils. However, studies have shown that VNIR spectra of field intact soils are able to generate robust models, implying that measurements can be taken *insitu* with lesser effects of interferences from soil properties such as moisture and particle size distribution on the modelling. MIR spectra are, however, obtained from ball milled homogeneous soils, increasing the time and cost of sample collection and/or preparation.

4.4. Conclusion

This study shows that good prediction models for soil organic carbon can be generated from VNIR spectra of soils and the approach can be a viable rapid tool for characterising soil organic carbon contents. However, the quality of models reduced with an increase in the geographical size of the study area from which samples were collected, indicating that the best calibration models will be generated from spectral data derived from local soils with similar geology, and accurate predictions for soil organic carbon using large scale universal libraries may be hard to achieve. Hence, to improve this approach, more efforts should be directed towards the development of local spectral libraries that adequately capture soil spatial variation. This will be of particular utility to individual farmers aiming to improve their soil carbon. A comparison of the use of VNIR and MIR diffuse reflectance spectroscopy in analyzing organic carbon in soils shows comparable performances. The PLSR models developed for organic carbon using both VNIR and MIR spectra had excellent predictive accuracies, with R^2 values of 0.90%, RMSE values ≤ 0.07 Log_{10} %C and RPD values ≥ 3.09. The quality of models generated from both sets of spectra was comparable, although VNIR PLSR models slightly outperformed MIR PLSR models. This study demonstrates that using visible near infrared diffuse reflectance spectroscopy for prediction of organic carbon in soils is as efficient as using the mid infrared and can potentially be a more convenient alternative to MIR for soil organic carbon predictions.

References

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

Box, G.E.P. and Cox, D.R (1964). An analysis of transformations, *Journal of the Royal Statistical Society*, Series B (26): 211-252.

Brady, N.C. (1986). *Advances in Agronomy*. Academic Press.

Brown D.J, Shepherd K.D, Walsh M.G, Mays M.D and Reinsch T.G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* (132) 273–290.

Byrappa, K and Suresh Kumar, B.V. (2007).Characterization of Zeolites by Infrared Spectroscopy.*Asian Journal of Chemistry*(19) No. 6: 4933-4935.

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J. and Smaling, E.M.A. (2012). Building a near infraredspectral library for soil organic carbon estimation in the Limpopo National Park,Mozambique. *Geoderma* 183–184.

CAMO Software, *Interpreting PLS Plots*, The Unscrambler X, Version 10.2 User's Guide, CAMO Software AS, Oslo, Norway.

Chang, C., Laird, D.A., Mausbach, M.J. and Hurburgh Jr., C.R.(2001). Near-infrared reflectance spectroscopy-principal components regression analysis of soil properties. *Soil Sci. Soc. Am. J.*(65): 480–490.

Clark, R.N., King, T.V.V., Klejwa, M. and Swayze, G.A.(1990). High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research* (95): 12653–12680.

Craswell, E.T. and Lefroy, R.D.B. (2001). The role and function of organic matter in tropical soils. *Nutrient Cycling in Agroecosystems* (61): 7–18.

Fortes, C. and Dematte, J.A.M. (2006). Discrimination of sugarcane varieties using Landsat 7 ETM+ spectral data.*International Journal of Remote Sensing*(27): 1395-1412.

Gaffey, S.J., McFadden, L.A., Nash, D. and Pieters, C.M.(1993). *Ultraviolet, visible, and near infrared reflectance spectroscopy: laboratory spectra of geologic materials*. In: Pieters,C.M., Englert, P.A. (Eds.), Remote geochemical analysis: elemental and mineralogical composition. Cambridge University Press, Cambridge, UK, pp. 43–77.

Ge, Y., Morgan, C.L.S., Grunwald, S., Brown, D.J. and Sarkhot, D.V. (2011). Comparison of soil reflectance spectra and calibration models obtained using multiple spectrometers. *Geoderma* 161 (3–4), 202–211.

Green, A.A. and Craig, M.D. (1985). *Analysis of aircraft spectrometer data with logarithmic residuals*. JPL Publication 85-41: 111–119.

Joachim, H. and Jacques, M. (2007). *Imaging Spectrometry - a tool for environmental observations*. Springer Science & Business Media.

Knox N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B. and Harris, W.G. (2015). Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma*, 239–240

Lal, R., Kimble, J.M., Follett, R.F. and Stewart, B.S. (2001). *Methods of assessment of soil carbon*. CRC Press, Boca Raton, FL.

McBratney, A.B., Minasny, B.andViscarra-Rossel, R.A., (2006). Spectral soil analysis and inference systems: a powerful combination for solving the soil data crisis. *Geoderma* (131): 59–75.

McCarty, G.W., Reeves, J.B., Reeves, V.B., Follett, R.F. and Kimble, J.M. (2002). Mid-infrared

and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Science Society of America Journal* (66): 640–646.

McCarty, G.W., Reeves, J.B., Yost, R., Doraiswamy, P.C. and Doumbia, M. (2010). Evaluation of methods for measuring soil organic carbon in West African soils. *African Journal of Agricultural Research* 5 (16), 2169–2177.

McCarty, G.W. and Reeves, J.B. (2006). Comparison of IR and MIR diffuse reflectance spectroscopy for field-scale measurement of soil fertility parameters. *Soil Science* (171): 94–102.

McDowell, M.L., Bruland, G.L., Deenik, J.L., Grunwald, S. and Knox, N.M. (2012). Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. *Geoderma*312–320

Mouazen, A.M.,Kuang, B., De Baerdemaeker, J. and Ramon, H. (2010). Comparison amongprincipal 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): 23–30

Nelson, D.W. and Sommers, L.E. (1996). *Total carbon, organic carbon, and organic matter*.In:Sparks, D.L., Page, A.L., Lemke, P.A., Loeppert, R.H., Softanpour, P.N., Tabatabai, M.A.,Johnston, C.T., Sommers, M.E. (Eds.), Methods of soil analysis, Part 3. SSSA-ASA,Madison, WI, pp. 983–997.

Nguyen, T.T., Janik, L.J. and Raupach, M. (1991). Diffuse reflectance infrared fourier transform (DRIFT) spectroscopy in soil studies. *Australian journal of soil research* (29): 49-67.

Pirie, A., Singh, B. and Islam, K. (2005). Ultra-violet, visible, near-infrared, and mid infrared diffuse reflectance spectroscopic techniques to predict several soil properties. *Australian Journal of Soil Research* 43 (6), 713-721.

Reeves, J.B., McCarty, G.W. and Reeves, V.B. (2001). Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soils. *Journal of Agricultural and Food Chemistry* (49): 766–772.

Reeves, J.B., McCarty, G.W. and Mimmo, T. (2002). The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soils. *Environ. Pollut*. (116): 277–284.

Reeves, J.B. (2010). Near vs 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.

Sarkhot, D.V., Grunwald, S., Ge, Y. and Morgan, C.L.S. (2011). Comparison and detection of soilcarbon under Arundodonax and coastal bermuda grass using visible/near infrared diffuse reflectance spectroscopy. *Geoderma* 164, 22–32.

Schwartz. G., Eshel, G., Ben-Haim, M. and Ben-Dor, E. (2009). Reflectance spectroscopy as a rapid tool for qualitative mapping and classification of hydrocarbons soil contamination.

Shepherd, K.D. and Walsh, M.G.(2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal* (66), 988–998.

Stoner, E.R. and Baumgardner, M.F.(1981). Characteristic variations in reflectance of surfacesoils.*Soil Science Society of America Journal* (45): 1161–1165.

Stuart, B. (2004). *Infrared spectroscopy: fundamentals and applications*. John Wiley &Sons Ltd., West Sussex, UK.

Thompson, C. (2009). *Chemical analysis of contaminated land*. John Wiley and Sons, Pp 312.

Udelhoven, T., Emmerling, C. and Jarmer, T. (2003). Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: a feasibility study. *Plant Soil* 251 (2), 319–329.

Varmuza,K. And Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*.CRC Press

Vasques, G.M., Grunwald, S. and Sickman, J.O. (2008). Comparison of multivariate methods for inferential modelling of soil carbon using visible/ near-infrared spectra. *Geoderma* (146) 14–25.

Vasques, G.M., Grunwald, S. and Sickman, J.O. (2009). Modelling of soil organic carbon fractions using visible/ near-infrared spectroscopy. *Soil Sci. Soc. Am. J*. (73) 176–184.

Vasques, G.M., Grunwald, S. and Harris, W.G. (2010). Spectroscopic models of soil organic carbon in Florida, USA. *Journal of Environmental Quality* 39, 923–934.

Vaughan, R.(1996). A review of: "*Imaging Spectroscopy—a Tool for Environmental Observations*". Edited by J. Hill and J. Megier (Dordrecht: Kluwer, 1994).[Pp. 335.].Published online: 27 Apr 2007, *International journal of remote sensing*.

Viscarra Rossel, R.A. and Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 158 (1–2), 46–54.

Viscarra Rossel, R.A., McBratney, A.B. (1998b). Laboratory evaluation of a proximal sensing technique for simultaneous measurement of clay and water content. *Geoderma* 85: 19-39.

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

Watson, R.T., Noble, I.R., Bolin, B., Ravindranath, N.H., Verardo, D.J. and Dokken, D.J. (2000).Land *Use, Land-Use Change, and Forestry, A Special Report of the IPCC*. Cambridge University Press, Cambridge, UK, pp. 53–126.

Wetterlind, J., Stenberg, B. and Soderstrom, M. (2008b). The use of near infrared (NIR) spectroscopy to improve soil mapping at the farm scale. *Precision Agric*. (9): 57–69.

Chapter 5. Characterisation of Petroleum Hydrocarbon Contamination Using Visible Near Infrared Diffuse Reflectance Spectroscopy

5.1 Introduction

Petroleum hydrocarbon contamination of soils is almost an inevitable occurrence during exploration and transport of crude oil and its products. These hydrocarbon compounds have attendant adverse environmental and human health consequences as they are potential soil contaminants and neurotoxins for humans and animals (Petersen and Talcott, 2013). There is, therefore, a need for a rapid but robust characterization of the type and extent of hydrocarbon contamination.

Characterisation of petroleum hydrocarbon is conventionally done using highly sensitive and specific wet chemistry analytical procedures that involve extraction and consequent gravimetric or chromatographic techniques (Brassington *et al*., 2010). These procedures, however, have drawbacks such as being laborious, high costs (as high as $100per sample), long waiting periods for results, requirement of a skilled analyst to perform measurements and their inadequacy when high spatial and temporal resolutions of petroleum hydrocarbon contents are required (Chakraborty *et al*., 2012; 2015; Schwartz et al., 2012).There is no doubt that management and clean-up of petroleum hydrocarbon contaminated soils will benefit from a rapid and less stressful technique such as visible near infrared diffuse reflectance spectroscopy (VNIR DRS).

Within chapter 2 of this thesis, a principal component analysis of soil spectra from diesel contaminated soils from a single location were observed to be clearly distributed within the spectral space based on the extent of petroleum hydrocarbon contamination. The areas and depths of two identified absorption features (~1640 – 1760nm and ~2240 – 2340nm) were

observed to have high positive linear correlations with the measured extractable petroleum hydrocarbon content of soils, demonstrating that visible near infrared diffuse reflectance spectra, coupled with regression analysis, can be a reliable qualitative and quantitative tool for identifying petroleum hydrocarbon contamination. However, the utility of this approach to quantitatively characterise petroleum hydrocarbon contamination using visible near infrared spectral measurements in a broader range of soils from different locations and containing different oil contamination types (diesel, motor oil, crude oil sludge) also needs to be investigated.

The diffuse reflectance spectra of soils, the measurement on which this approach is based, is the totality of spectral responses from all spectrally active constituents present in soil. It is, therefore, influenced by a variety of soil components such as organic carbon and other spectrally active chemical components. The organic carbon content of soils is the most influential component affecting spectral characterisation of hydrocarbon contamination. These two components contain similar functional groups such as carbon-hydrogen bonds (C-H, $C-H_2$, $C-H_3$), hydroxyl groups (O-H), double and triple bonds of aliphatics and aromatics, carboxyl groups (C=O), ester groups (C-O-C), amino groups (N-H) that absorb at similar regions of the electromagnetic spectrum (Aske *et al*.,2001; Ben-Dor *et al*., 1999). The presence of either high or low organic carbon in soils has no significant effect on gas chromatographic analysis of petroleum hydrocarbons in soils, as most conventional GC-FID procedures incorporate a silica – gel clean-up process that minimises the presence of non-polar organic hydrocarbons in the final extract. However, spectral measurements of petroleum hydrocarbon contaminated soils can be affected by interferences with naturally occurring soil organic compounds leading to strong increases in absorbance at

certain bands (Paiga et al., 2012) and/or masking of absorption features associated with petroleum hydrocarbons.

Different petroleum products vary in characteristics and composition though they all contain mostly carbon, hydrogen, nitrogen, oxygen, sulphur and various types of metal. Soils also become darker, with the degree of darkness varying when contaminations from diesel and other dense petroleum hydrocarbon fractions, such as crude oil, oil sludges and motor oils, occur. The effect this has on the viability of the spectral approach to characterising petroleum hydrocarbon in soils is yet to be sufficiently studied.

Some efforts have been made within literature to investigate the potentials of using the reflectance spectra from contaminated soils to quantitatively predict petroleum hydrocarbon contamination with more emphasis on the use of infrared spectra (Forrester et al., 2010; 2012; Malley et al., 2010; Chakraborty et al. 2010; 2012; Paiga et al., 2012; Schwartz el al., 2012; Okparanma and Mouazen, 2013; 2014; Okparanma et al., 2014; Ng et al., 2017. However, there is less information on the use of the visible near infrared diffuse reflectance approach as well as theeffects soil constituents, such as organic carbon and petroleum hydrocarbon contamination type, have on the potentials of using VISNIR DRS as a rapid quantitative tool for petroleum hydrocarbons in contaminated soils.

The overall aim of this study is to investigate the potential of quantitatively predicting petroleum hydrocarbons in contaminated soils across a range of soil organic carbon content, using visible near infrared diffuse reflectance spectrometry. Specific objectives of this study include:

1. Assess the viability of quantitatively characterising petroleum hydrocarbon contamination in soils using visible near infrared diffuse spectroscopy (VNIR DRS).

2. Investigate the effect of soil natural organic carbon on the performance of VNIR DRS models developed for characterizing soil extractible total petroleum hydrocarbons (ETPH)
3. Assess the effect different types of fuels (diesel, motor oil, crude oil sludge) will have on the performance of VNIR DRS models developed for characterizing soil ETPH

5.2. Methodology

5.2.1 Soil Collection and Processing

To investigate the prediction accuracy of VNIR DRS in quantifying the amounts of extractible petroleum hydrocarbons (ETPH) in contaminated soils, two groups of soils were used. Group A (Wisley soils) consisted of 51 topsoil samples collected from the Royal Horticultural Society experimental plots at Wisley, Surrey, United Kingdom (51°19'24.34"N; 0°28'27.81"W). Group B (England soils) includes 56 topsoil samples collected at different sites across England and Wales to represent different possible variations in soil composition. Soil samples were air dried and then sieved to 2mm. They were then contaminated in the laboratory with different amounts of diesel fuel.

To assess the effect of organic carbon content on the performance of VNIR DRS models developed for characterising ETPH in contaminated soils, a total of 151 topsoil samples collected from the Royal Horticultural Society experimental plots at Wisley, Surrey were divided into three (C_1, C_2, C_3) and each category amended with varying levels of organic compost to give three levels of organic carbon content. Soil collection from one single location ensures that the three categories of soils are underlain by similar geological composition and are, therefore,

expected to differ with respect to organic carbon contents only. This ensures that variations between the compositions of other spectrally active components within soils are limited. Figure 1 illustrates the analytical overview of the experiment.

To assess the effect the type of petroleum hydrocarbon contamination has on the performance of VNIR DRS models developed for characterising ETPH in contaminated soils, 150 soils samples also collected from the Royal Horticultural Society experimental plots at Wisley, Surrey, with organic carbon content ranging between 12% – 15%. The soils were split in three categories (X, Y, Z), one contaminated with increasing quantities of diesel fuel, another contaminated with increasing quantities of motor oil and another was contaminated with increasing quantities of crude oil sludge (Figure 5.1).

Contamination in soils was thoroughly mixed by agitation with a spatula for 1 minute. Samples were then allowed to equilibrate for 3 days prior to spectral and laboratory analysis.

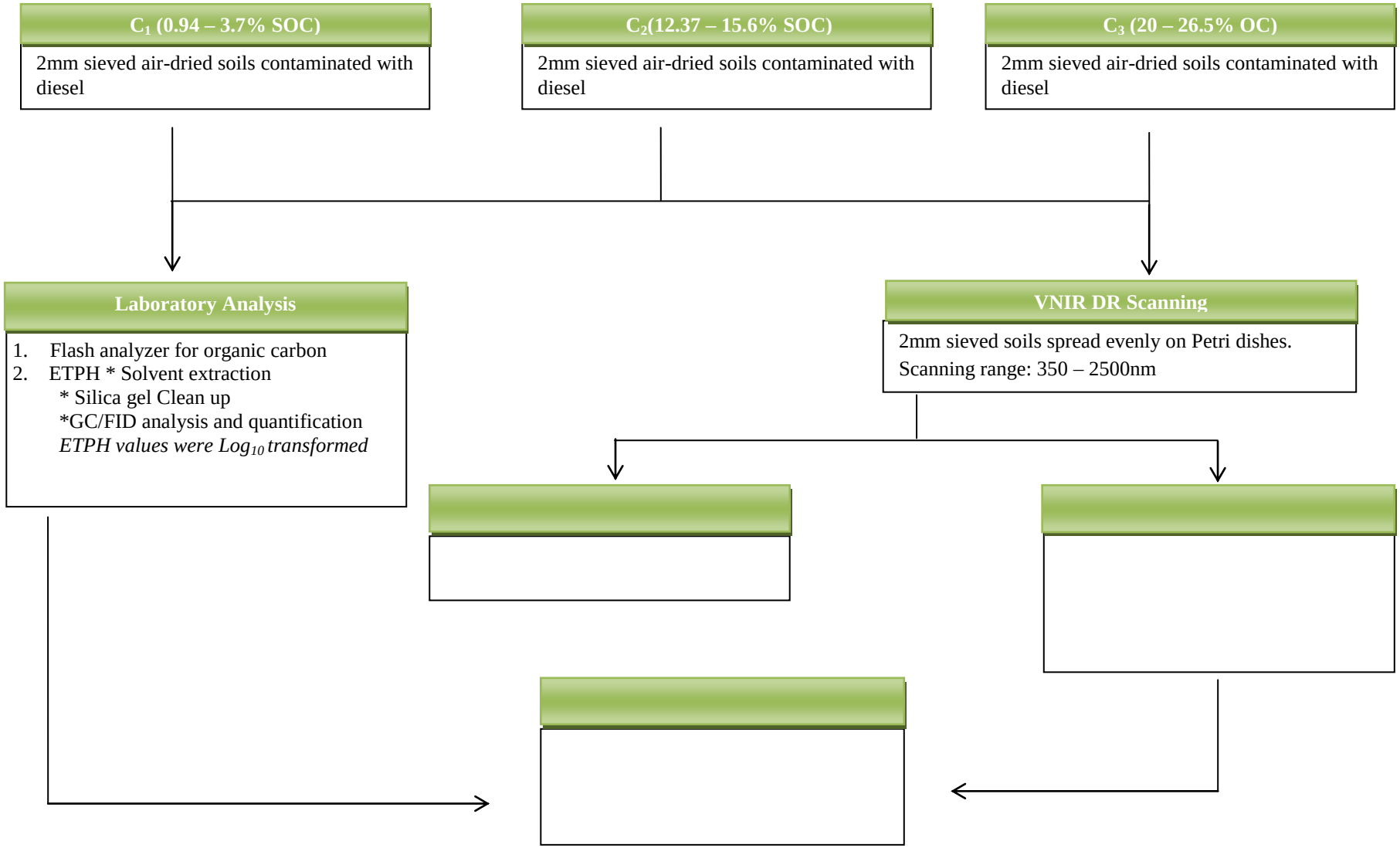


Figure 5.1: Experimental set - up to assess effect of soil organic carbon content on performance of VNIR models

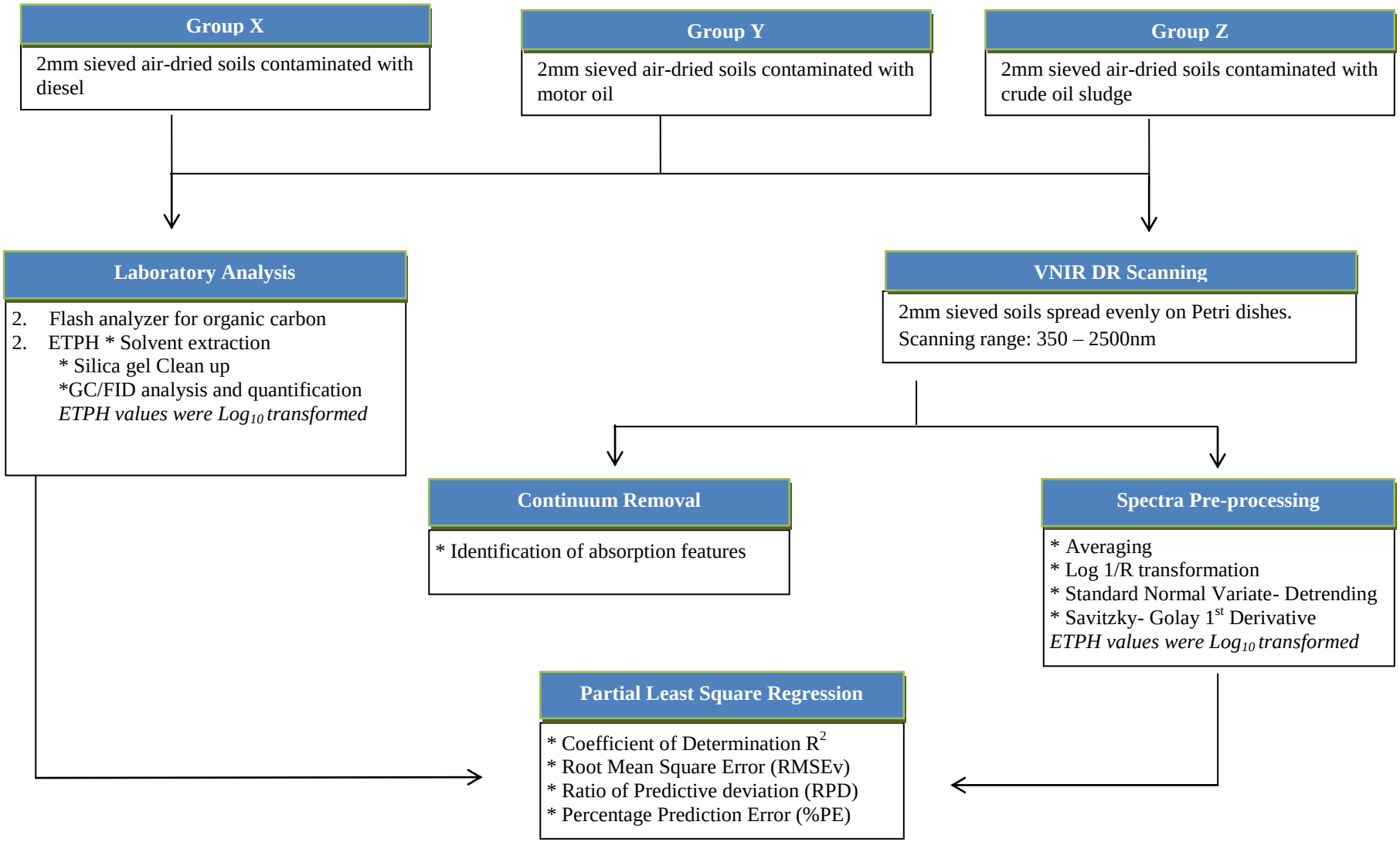


Figure 5.1: Experimental set-up to assess the effect the type of petroleum contamination on the performance of VNIR DRS models

5.2.2. Laboratory Analysis

In the laboratory, the extractible total petroleum hydrocarbon (ETPH) content of soils were determined by a solvent extraction, silica gel solid-phase clean-up/fractionation process coupled with gas chromatographic (GC/FID) quantification of target aliphatic and polynuclear aromatic hydrocarbon analytes in the extract (MADEPH, 2009).

Soils were extracted using a mixture of acetone and hexane (1:1). Extracts were cleaned and fractionated into aliphatic and aromatic fractions using 1g activated silica gel and 0.5g anhydrous sodium sulphate packed tight within SPE Cartridges. The fractions were concentrated to 1ml under a gentle flow of nitrogen. 14 aliphatic hydrocarbon (C_9-C_{36}) and 16 aromatic hydrocarbon (C_{11}-C_{22}) components were quantified in extracts using gas chromatography coupled with a flame ionization detector.

5.2.3. Spectral Analysis

Diffuse reflectance spectra of soil samples were obtained in a dark room with a GER3700 VNIR spectrophotometer (350 - 2500nm) coupled with a 1 kW quartz-halogen light source. The spectrophotometer has one Si array (350 - 1050 nm) and two Peltier-cooled InGaAs detectors (1050 - 1900 nm and 1900 - 2500 nm). The spectral sampling interval of the instrument was 3nm at 350 - 1050 nm, 7nm at 1050 - 1900 nm and 9.5nm at 1900 - 2500 nm. The angle of illumination was $45^0$and the spectrometer was viewing at nadir.

A white reference panel of Spectralon was scanned before each measurement to convert radiance values to percent reflectance. Scans were taken from the soil samples, tightly packed and levelled in borosilicate petri - dishes. Replicate measurements were collected at four positions by rotating the petri dish at an angle of 90^0 to account for variations in surface

scattering. Each scan was an average of 16 internal scans. The 64 replicate scans were then averaged to produce a single spectrum for each sample.

5.2.4. Spectra Pre-processing

All spectral pre-treatment and multivariate calibration/validations were carried out using the Unscrambler-X 10.3 analytical software developed by CAMO.

Prior to modelling, a small portion (1905nm – 1928nm) within the spectra was removed due to an offset in reflectances between the two SWIR detectors. The spectra is then averaged (i.e. every 5 wavelength points) to give 120 predictor variables for model calibration. Due to scattering effects and pathlength variations inherent in diffuse spectra, four common spectral pre-processing techniques were applied to spectra. Techniques tested included Log_{10} (1/R) transformation, standard normal variate coupled with the detrending (SNV-DT), Savitzky-Golay smoothing of the spectra (SG), and a combination of Savitzy-Golay and derivatization. These pre-processed spectra were individually combined with the laboratory measured ETPH contents and used to create prediction models using partial least squares (PLSR) regression

5.2.5. Partial Least Square: Model Calibration and Validation

PCA analysis was performed on data sets to determine how contaminated soils were distributed within the spectral space, observe groupings within soil samples and to identify the presence of outliers. Possible outliers were identified as samples with very high leverages and residuals.

Within this study, the original ETPH contents of samples were non-normally distributed. The ETPH data were then Log_{10}- transformed to make it more Normal. Therefore, PLS models were developed based onlog_{10}-transformed ETPH data that approximated a Gaussian distribution after stabilizing the variance.

Partial least square regression (PLSR) models were developed to quantify ETPH in soils and to assess the effects both different levels of soil organic carbon and different types of petroleum hydrocarbon contamination would have on the VNIR DRS prediction of ETPH. Using an already established data split pattern, 76% of each of the data set was selected for model calibration while the remaining 24% were used for model validation. This approach to model validation is recommended as a robust procedure by Varmuza and Filzmoser (2009). The validation samples have been carefully selected to be similar to the calibration set in terms of the range ETPH content and the locations of collection.

Models with as much as seven latent factors were considered and the optimum model was determined by choosing the number of factors with the least root mean square error of prediction ($RMSE_p$). Increasing the number of latent factors can further reduce the RMSE. However, this can result to over fitting the model such that the model equation is data dependent, leading to poor predictions. Other model quality statistics: coefficient of determination (R^2), relative percent difference (RPD) values and the percentage prediction error (% PE) were evaluated and used to compare the predictive accuracy and stability of models.

$$RMSEp = \sqrt{\frac{\sum(TPHpred - TPHmeas)^2}{n}} \quad -----(1)$$

$$\%\, PE = \frac{RMSEp}{Xh_v} \times 100 \quad ----------(2)$$

$$RPD = {}^{SD}/_{RMSEp} \quad ---------- (3)$$

Where n is the number of validation samples, SD is the standard deviation of the predicted validation values and $Xh_v$is the largest measured data point within the validation set.

According to ViscarraRossel *et al*. (2006) and Chang *et al*. (2001), very poor models show RPD < 1.0; poor models: 1.0 ≤ RPD ≤ 1.4; fair models: 1.4 ≤ RPD ≤ 1.8; good models : 1.8 ≤ RPD ≤ 2.0 and very good models : 2.0 ≤ RPD ≤ 2.5; and excellent models have RPD > 2.5.

5.3. Results and Discussion

Figure 5.2 illustrates the shape and variations observed within the VNIR spectra of diesel contaminated soils. To identify specific absorption features associated with petroleum hydrocarbons, the continuum removal technique was applied to the spectra to allow for a comparison of spectra from a common baseline (Green and Craig, 1985). In all spectra studied, two absorption features at ~1640 – 1760nm and ~2240 – 2440nm, whose areas had previously been observed to be positively correlated with increasing ETPH concentrations, were clearly identified.

Spectral absorption minima of hydrocarbon based oil are apparent around 1647, 1712 and 1759nm in the first overtone region of the near infrared region (Okparanma and Mouazen, 2013). Absorptions around 1647nm could indicate a C-H stretch of an aromatic C-H likely linked to PAH while absorptions around 1712nm and 1759nmcould indicate C-H stretch of terminal CH_3 and CH_2 of saturated hydrocarbons (Okparanma and Mouazen, 2013).

Even though there was a general overall increase in soil reflectances with increasing diesel contamination, variations in continuum removed reflectances, with negative correlations with increasing diesel contamination, are also observed at ~400 - 610nm and at ~1895 - 2020nm. *Absorption* in the visible *range ~400 - 600nm* are indicative of the presence of iron-containing minerals such as goethite and hematite (Vaughan, 1996) and can also be indicative of carbonyl, carboxyl and hydroxyl bond absorptions of humic acids (Shoji et al., 1994). Soil reflectances around this spectral range decreases with increasing soil organic matter (SOM) content, especially if the SOM content is bigger than 2% (Stoner and Baumgardner, 1981; Henderson et al., 1992). Increasing diesel contamination in soil samples results in a lower organic matter content per gram of soil sample, hence the negative relationship observed in soil spectra. Similarly, absorptions at ~1895 – 2020nm are characteristic of soil moisture whose concentrations in the soil samples diminish with increasing diesel fuel contamination.

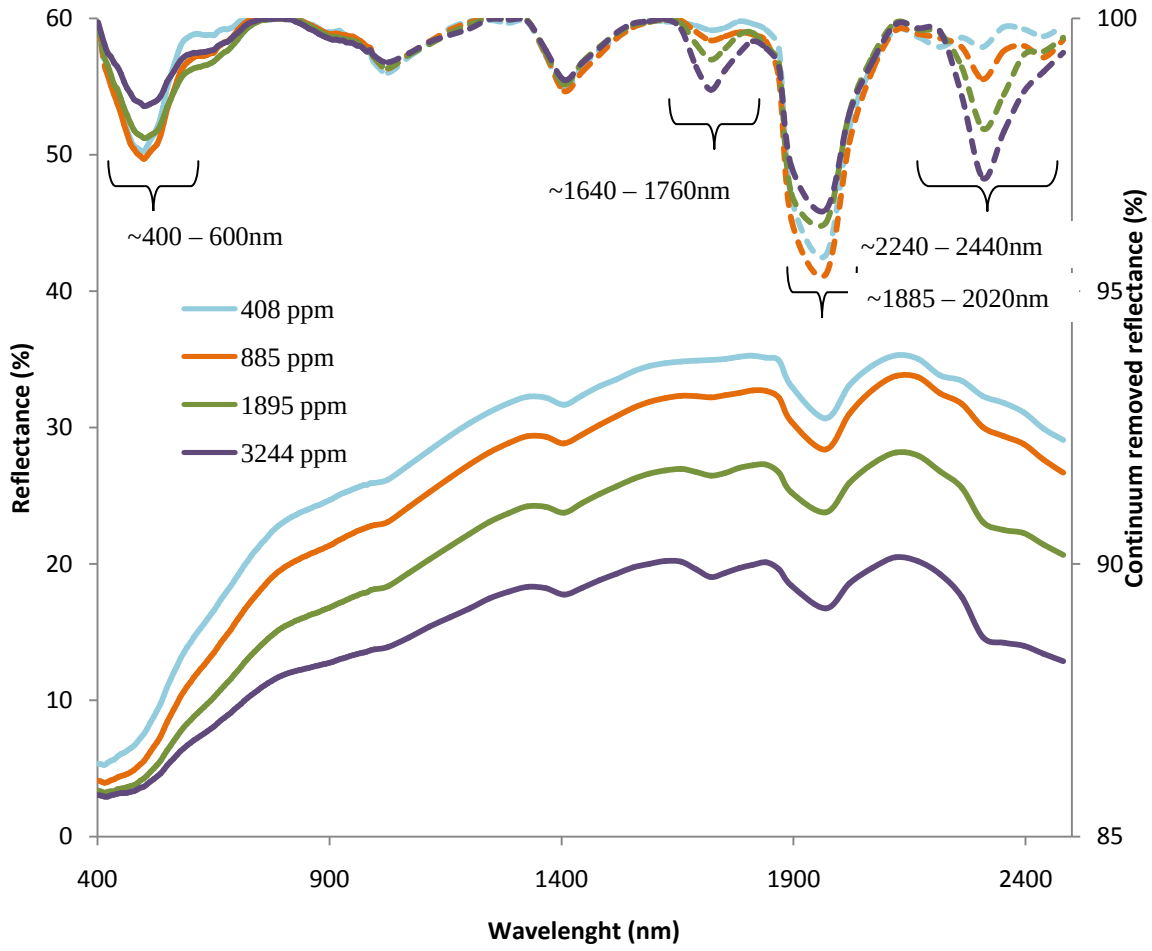


Figure 5.2: Mean reflectance (left scale) and continuum removed (right scale) visible, near-infrared diffuse reflectance spectra of four diesel contaminated soil samples. Drop lines indicate two wavelength regions where reflectances have strong correlated with ETPH concentrations

5.3.1. Characterisation of ETPH using VNIR DRS

The calibration data sets were carefully chosen to be as similar as possible to the validation set, both in terms of the ETPH concentrations and variations observed within the spectra (Table 5.1). Partial least square regression models were developed for both data sets after spectral pre-processing using the pre-processing techniques that best improves the data. The prediction accuracies of the models were tested using the validation sets and compared using validation R^2, RPD and %PE as main criteria (table 5.2).

Table 5.1: Descriptive statistics of ETPH concentrations (ppm) within data sets used for partial least squares regression (PLSR)

	Wisley		England	
	Range	Mean ± SD	Range	Mean ± SD
Total Samples	1168 - 8952	2718 ± 1800	2530 - 8582	4400 ± 1360
Calibration samples (76%)	1168 - 8952	3190 ± 1858	2530 - 8582	4370 ± 1440
Validation samples (24%)	1382 - 5895	2760 ± 1400	2649 - 6206	4470 ± 1100

Table 5.2: Performance of ETPH – VNIR partial least square regression models of diesel contaminated soils

Data set	Pre-processing method	Optimum factor	Calibration		Validation			
			R^2	RMSE (Log_{10}ppm)	R^2	RMSE (Log_{10}ppm)	RPD	% PE
Wisley	SNV-DT	2	0.87	0.09	0.97	0.06	4.5	1.5
England	SG-1st D	3	0.46	2.86	0.72	2.79	1.4	9.8

SG-1st D = first derivative of 5windowsmoothing Savitzky – Golay filter; SNV-DT = standard normal variate coupled with detrend correction.

The model generated from the Wisley soils was superior to that developed using the England soils with a higher validation R^2 of 0.97, an RPD value of 4.5 and a %PE of 1.5. The lower performance of England soils compared to the Wisley soils can be attributed to a wider geological variation of soils within this group. Diffuse reflectance spectra from all soils are a combination of responses from all spectrally active components within the soil. Therefore, certain spectrally active components within some of the England soils, not present in others, could serve as interferences, reducing the relationships between overall soil spectra and measured ETPH in the models. Wisley soils were obtained from a less diverse geographical location and compositions within the soil samples would be expected to be geologically similar. This implies that the quality of prediction models developed for characterizing ETPH may reduce with an increase in the variations within soils used for model calibrations/validation. According to Stenberg et al., (2010), predictions of soil properties using VNIR DRS may be affected by soil's spatial variation such that field or farm-scale calibration could be better than regional calibration models.

It is worth noting that PLSR models developed from the England soils also showed promising model quality statistics with a validation R^2 of 0.72, an RPD value of 1.4 and a %PE of 9.8. A RPD value of 1.4 indicates that there is room for model enhancement (Chang et al., 2001). Similar model statistics for total petroleum hydrocarbons predictions in soils (validation R^2: 0.64 and RPD: 1.70; validation R^2: 0.68 and RPD: 1.76; validation R^2: 0.77–0.89 and RPD: 1.86 – 3.12) were reported by Chakraborty *et al*., (2010), Malley *et al*., (1999) and Okparanma and Mouazen (2014) respectively. Forrester *et al*., (2012) also reported an R^2 of 0.93 using a partial least squares (PLS) cross-validation approach for infrared spectroscopic identification of TPH in soils. Figure 5.3 illustrates the predicted *vs* measured values for the validation set for PLSR models for both groups.

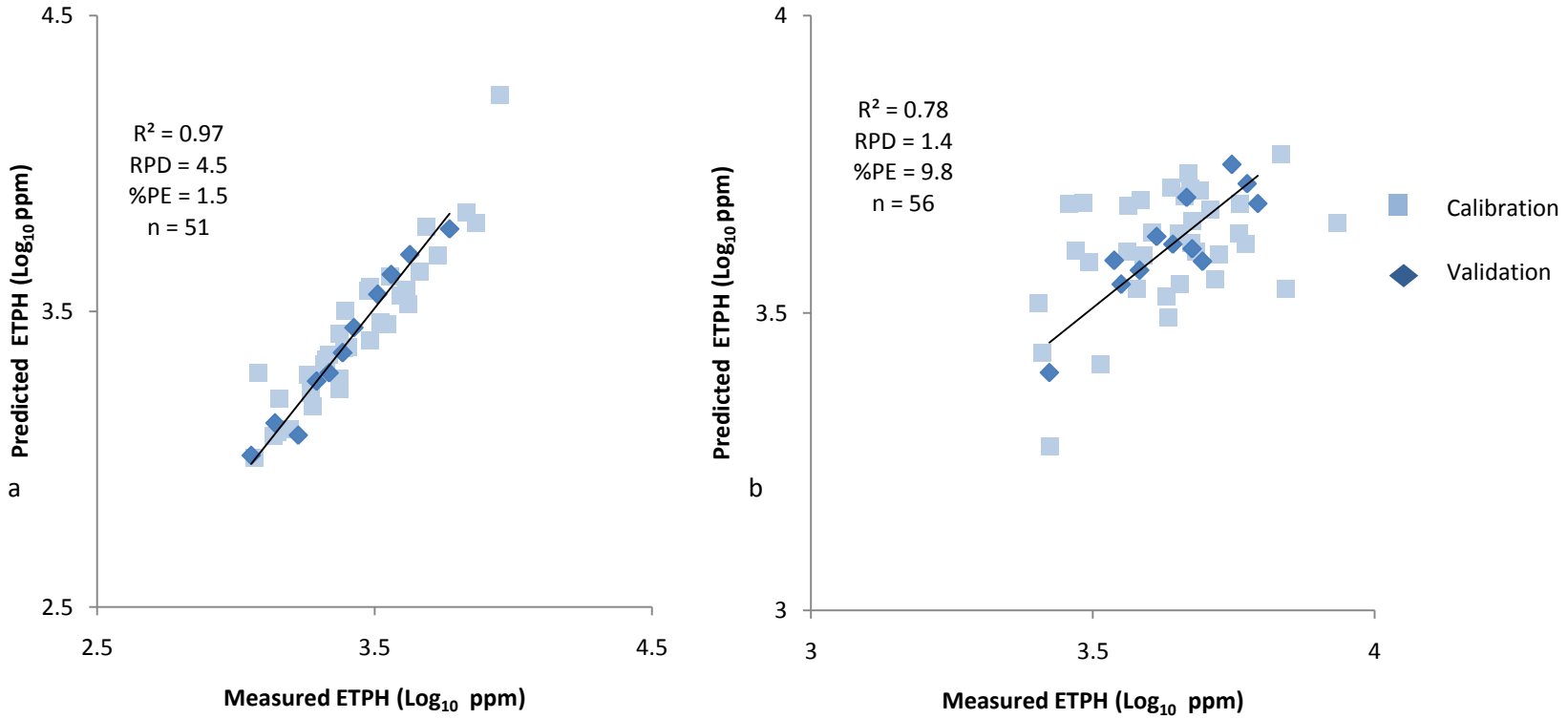


Figure 5.3: Predicted vs measured values of the validation data sets in ETPH-VNIR partial least square regression models of (a) Wisley soils (b) England soils

The relative importance of wavelength variables within the regression models is shown within the regression coefficients chart (Figure 5.4).The magnitude of the regression coefficient at each wavelength is proportional to the importance of the wavelength variable in the model. Important wavelengths are observed around ~1150 – 1550nm, ~1650 – 1750nm, ~1800 – 2100nm and ~2200 – 2449nm. Absorptions around ~1650 – 1750nm and ~2200 – 2449nm have been associated with C-H stretch and bend of crude oil signatures (Mullins *et al*., 1992) and have been identified within continuum removed visible to near infrared diffuse reflectance spectra of contaminated soil samples.

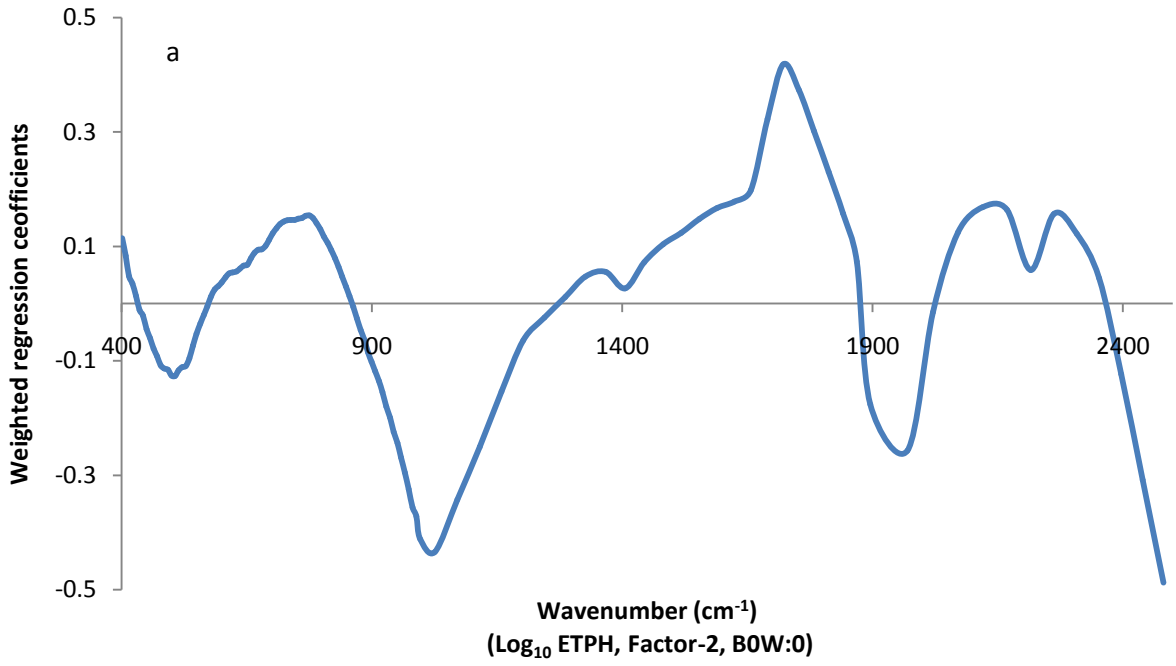


Figure 5.4: Regression coefficients of the ETPH – VNIR partial least square regression model of Wisley soils. The magnitude of the regression coefficient at each wavelength is proportional to the importance of the wavelength variable in the model.

5.3.2. Assessing Effect of Soil Organic Carbon on the Performance of VNIR DRS Models Developed for Characterizing Soil ETPH

The scores plot from a principal component analysis of spectra from the three categories of soils clearly differentiated between soils based on both the organic carbon content (C_1: 0.94 – 3.7 %SOC; C_2: 12.4 – 15.6 %SOC; C_3: 20 – 26.5 %SOC) as well as the levels of petroleum hydrocarbon contamination (Figure 5.5). This indicates that soil spectra are able to capture inherent information within soils and the VNIR DRS method could be a very important qualitative technique for distinguishing soils with respect to organic carbon content and levels of petroleum hydrocarbon contamination.

The first principal component (PC1) accounts for a very large proportion (92%) of the variance whereas principal component 2 (PC2) accounts for only about 6% of the variance within soil spectra. PC1 most likely describes the hydrocarbon contamination of soils while PC2 describes the organic carbon content of soils as can be observed with positioning of soils with lower SOC content (category C_1) from soils with higher SOC (C_2 and C_3)in a separate region of PC2.

Soil categories C_1 and C_2 are clustered within the scores plot whereas C_3 is spread along PC1. All these soils had been collected from the Royal Horticultural Society experimental plots at Wisley, Surrey. The first group of soils (C_1) were collected from plots with no amendments added to them. C_2soils were collected from plots which had been amended with organic amendments such as peat, horse manure, composted bracken and mushroom compost added over a period of six years. C_3 soils were soils amended with organic compost purchased from a local store made from 100% coco coir. The wider spread of C_3 soils compared to C_1 and C_2 on the scores plot may be attributed to the difference in organic matter composition.

Table 5.3 summarizes the statistics for PLSR models from the three soil categories and all categories combined. PLSR models developed from all four data sets gave very good prediction performances with RPD values greater than 2. According to Chang et al. (2001), the quality of spectroscopic models should be based on their RPD values, with stable and accurate predictive models having RPD values greater than 1.8. When all four models were compared, coefficient of determination (validation R^2) seems to reduce while the percentage prediction error (%PE) increased with larger range of soil organic carbon within the groups. Absorptions due to soil organic carbon are as a result of stretching and bending of C-H, N-H, S-H and O-H bonds of functional groups. These bonds are also responsible for the spectral absorptions associated with petroleum hydrocarbons (Okparanma and Mouazen, 2013). Therefore, a higher organic carbon content could potentially increase the overall spectral absorption by a soil sample. This, in itself, may not have a considerable effect on the robustness of models calibrated for ETPH when all soils used with the calibration have similar organic carbon contents. It could, however, reduce model quality when soils with a wider range of organic carbon content are used for model calibrations (Figure 5.6). Despite this, the model developed from a combination of all three data sets can still be classified as 'good' based on the classification of Chang et al., (2001). A similar observation was made by Paiga et al. (2012) in an assessment of the effect of soil organic matter on total petroleum quantification using infrared spectra. Paiga et al., 2012, concluded that, though a significant matrix effect was observed using infrared methods, experiments with soil with high organic matter content also provided satisfactory average recovery (around 94 %) comparable with those obtained by gas chromatography.

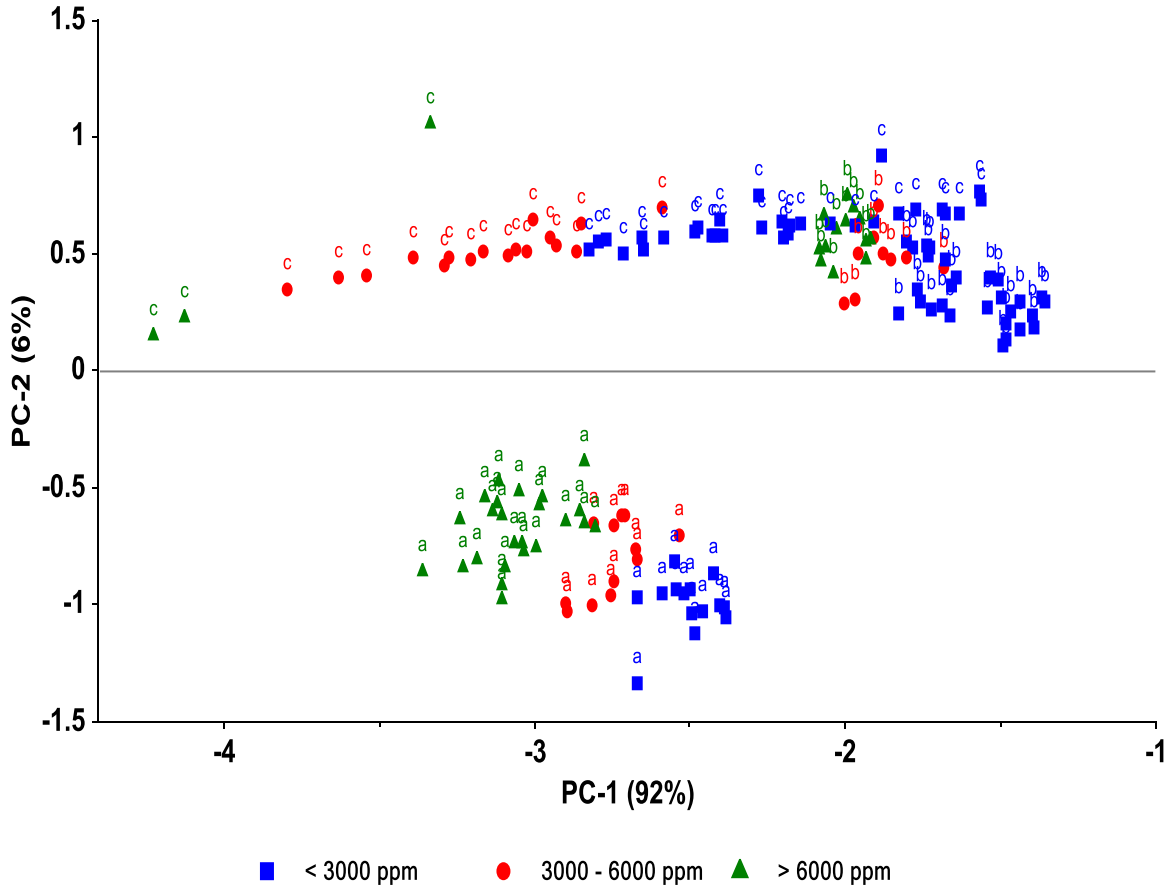


Figure 5.5: Scores plot of a principal component analysis showing qualitative discrimination between and within soil categories. Where symbols a= C_1; b=C_2; c= C_3

Table 5.3: Effect of organic carbon content on performance of ETPH – VNIR partial least square regression models of diesel contaminated soils

Data set	Range of OC (%)	Pre-processing	Optimum factor	Calibration		validation			
				R^2	RMSEc (Log_{10}ppm)	R^2	RMSEv (Log_{10}ppm)	RPD	% PE
C_1 (51)	0.94 – 3.7	SNV-DT	2	0.87	0.09	0.97	0.06	4.5	1.5
C_2 (50)	12.37 – 15.6	SNV-DT	1	0.94	0.11	0.97	0.08	5.5	2.1
C_3 (50)	20 – 26.5	Log 1/R	2	0.89	0.12	0.91	0.11	3.3	2.7
All three categories	0.94 – 26.5	SNV-DT	3	0.85	0.16	0.90	0.16	2.7	4.0

Where OC = Organic carbon content; $Log_{10}(1/R)$ = Absorbance, where R is the reflectance value; Savitzky–Golay filter ; SNV-DT = standard normal variate coupled with detrend correction.

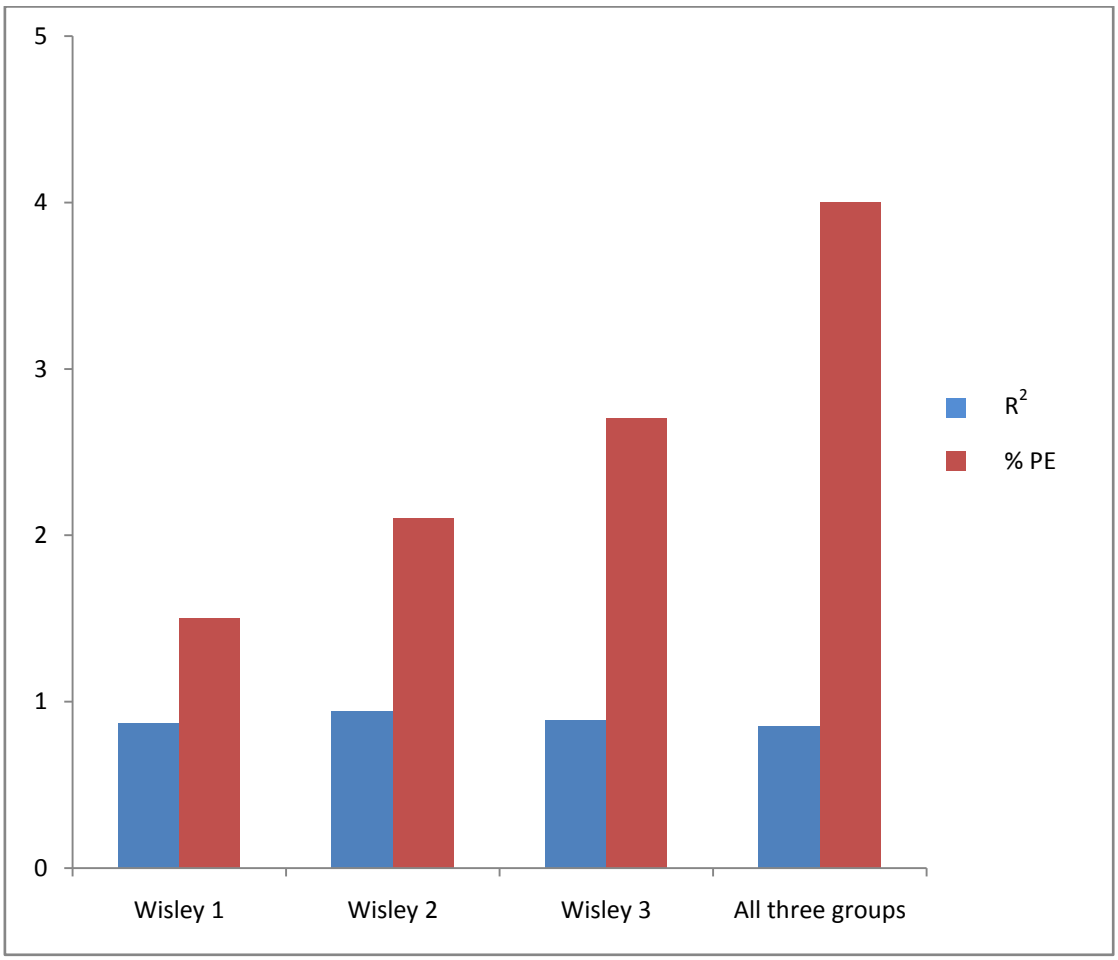


Figure 5.6: Effect of organic carbon content on validation R^2 and percentage prediction error (%PE) of ETPH – VNIR partial least square regression models of diesel contaminated soils

5.3.3. Assessing the Effect of the Type of Petroleum Hydrocarbon Contamination on the Performance of VNIR DRS Models Developed for Characterizing Soil ETPH

Figure 5.7 illustrates the scores plot of a principal component analysis of soil spectra derived from three categories of soils, X, Y and Z (from same location) contaminated with diesel, motor oil and crude oil sludge respectively. Soil organic carbon content within the three groups range within 12.4% - 15.6%. PCA analysis show soils separated mostly on the basis of the extent of petroleum hydrocarbon contamination and less on the basis of type of petroleum hydrocarbon contamination as opposed to the clear separation on soils within the spectral space impacted by organic carbon content.

Visual discrimination between spectra of different oils, such as diesel and motor oils, is difficult as they produce very similar absorption features (Rusak et al., 2003) except by careful comparison of absorption features from oils as they mature over similar time, which can provide distinguishing characteristics (Allen and Krekeler, 2010) . Absorptions within spectra of petroleum hydrocarbon contaminated soils originate from combinations and overtones of C-H, N-H, O-H, and S-H stretching and bending vibrations of saturated and aromatic functional groups within the soils (Aske et al., 2001). It is, therefore, expected that the amounts of these absorbing bonds or functional groups should be solely responsible for the absolute absorptions observed within soil spectra. However, upon contamination, both the motor oil and crude oil sludge imparted a darker tone to the soils than observed in the diesel contaminated soils. As measurements within the visible region is sensitive to color variations, the dark tone imparted to soils by motor oil and crude oil sludge may likely be responsible for the observed slight discrimination among the groups within the PCA scores plot, rather than the type of petroleum hydrocarbon contamination.

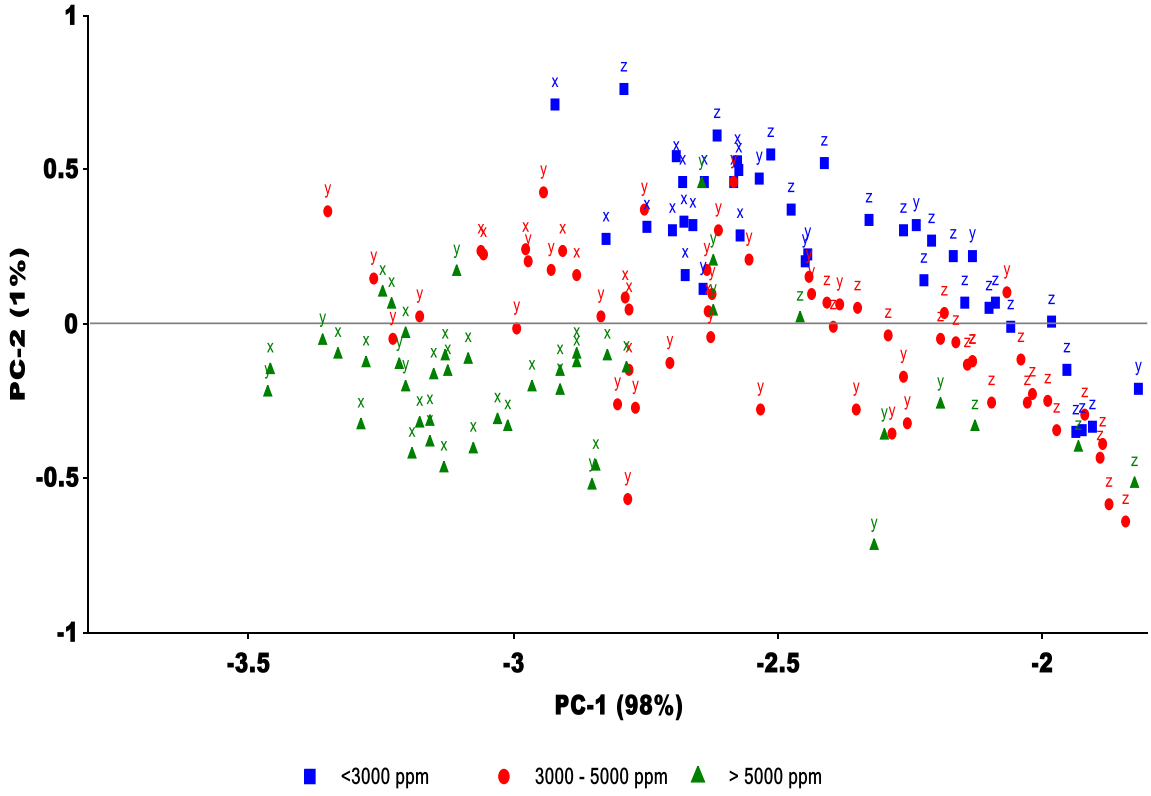


Figure 5.7: Scores plot of a principal component analysis showing spectral distribution of soils. Where symbols indicate contaminating oil: X=contaminated with diesel; Y=contaminated with motor oil; Z= contaminated with crude oil sludge.

Partial least square regression models derived from all four data sets gave good quality statistics with validation R^2 greater than 0.80 and RPD values greater than 1.8 (Table 5.4). Comparing the models, PLSR model of diesel contaminated soils performed the best while PLSR model of motor oil contaminated soils performed better than PLSR model of oil sludge contaminated soils. The validation R^2 and RPD values were observed to decrease while the percentage prediction error increased as the soils darken due to contamination (Figure 5.7).

Table 5.4: Effect of type of contaminating material on performance of ETPH – VNIR partial least square regression models of petroleum hydrocarbon contaminated soils

Data set (12.37 – 15.6 % OC)	Contamination material	Pre-processing	Optimum factor	Calibration		validation			
				R^2	RMSEc (Log_{10}ppm)	R^2	RMSEv (Log_{10}ppm)	RPD	% PE
Soil X	Diesel	SNV-DT	1	0.94	0.11	0.97	0.08	5.5	2.1
Soil Y	Motor oil	SNV-DT	2	0.85	0.15	0.80	0.18	2.3	3.8
Soil Z	Crude oil sludge	SNV-DT	2	0.80	0.17	0.83	0.17	1.8	4.5
All three groups		Log_{10} (1/R)	2	0.83	0.17	0.74	0.21	2.0	5.3

Where OC = Organic carbon content; $Log_{10}(1/R)$ = Absorbance, where R is the reflectance value; Savitzky–Golay filter ; SNV-DT = standard normal variate coupled with detrend correction.

5.4 Conclusion

Quantitative predictions of extractible petroleum hydrocarbon content in soils were achieved using a combination of visible near infrared spectra and partial least square regression modeling. Quality of models was found to be affected by the spatial variations within soil samples used for model calibration/validation, though PLSR model developed from contaminated soils with a highly variable geological composition gave reasonable model quality statistics (RPD = 1.4).

A higher organic carbon (OC) content in soils used for model calibration/validation was observed to lower the quality statistics of PLSR models developed for ETPH, even though all three models developed within the OC range (0.94 – 26.5% OC) gave quite good prediction statistics. Likewise, models developed for spectra derived from soils that had been contaminated with dark petroleum hydrocarbon products (e.g. crude oil sludge and motor oil) had lower model quality statistics compared with those derived from clear products such as diesel that do not darken soils as much.

Considering the high costs and analytical time associated with current wet chemistry procedures for analysis of petroleum hydrocarbons in soils, the prospect of using this relatively cheaper and faster procedure of VIS NIR spectroscopy is particularly promising. Additional research with larger data sets and real-time field contaminated soils is suggested.

References

Allen, C. and Krekeler M. (2010). Reflectance spectra of crude oils and refined petroleum products on a variety of common substrates. Active and Passive Signatures, edited by G. Charmaine Gilbreath, Chadwick T. Hawley, Proc. of SPIE Vol. 7687.

Aske, N., Kallevik, H. and Sjoblom, J. (2001). Determination of saturate, aromatic, resin, and asphaltenic(SARA) components in crude oils by means of infrared and near-infrared spectroscopy. *Energy and Fuels* 15 (5):1304–1312

Brassington, K.J., Pollard, S.J. and Coulon, F. (2010). *Weathered hydrocarbon wastes: a risk management primer*. In Handbook of hydrocarbon and lipid microbiology. K.N. Timmis Edition. Pp.2488–2499, Springer, Berlin, Germany

Chakraborty, S.,Weindorf, D., Morgan, C., Ge, Y., Galbraith J., Li, B. and Kahlon, C.(2010). Rapid identification of oil-contaminated soils using visible near-infrared diffuse reflectance spectroscopy. *J. Environ. Qual*. 39:1378–1387

Chakraborty, S., Weindorf, D., Zhu, Y., Li, B., Morgan, C., Ge, Y., and Galbraith, J. (2012). Spectral reflectance variability from soil physicochemical properties in oil contaminated soils. *Geoderma* 177–178

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.

Chang, C., Laird, D.A., Mausbach, M.J. and Hurburgh, C. (2001). Near-infrared reflectance spectroscopy: principal components regression analyses of soil properties. *Soil Sci. Soc. Am. J*. 65:480–490

Forrester, S., Janik, L. and 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, Pp. 13–16.

Green, A.A. and Craig, M.D. (1985). *Analysis of aircraft spectrometer data with logarithmic residuals*. JPL Publication 85-41: 111–119.

Henderson, T.L., Baumgardner,M.F., Franzmeier, D,P., Stott, D.E, and Coster,D.C. (1992). High dimensional reflectance analysis of soil organic matter. *Soil Sci. Sot. Am. J*. 56 (3): 865-872,

MADEPH, (2009). Recommended reasonable confidence protocols quality assurance and quality control requirements for extractable petroleum hydrocarbons. State of Connecticut. Department of environmental protection.

Malley, D.F., Hunter, K.N. and Webster, G.R. (1999). Analysis of diesel fuel contamination in soils by near-infrared reflectance spectrometry and solid phase micro extraction-gas chromatography. *J. Soil Contam.* 8:481–489

Okparanma, R.N. and Mouazen, A.M. (2013). Determination of total petroleum hydrocarbon (TPH) and polycyclic aromatic hydrocarbon (PAH) in soils: a review of spectroscopic and non- spectroscopic techniques. *Applied Spectroscopy Reviews* 48(6): 458–486

Okparanma, R.N., Coulon, F. and Mouazen, A.M. (2014). Analysis of petroleum-contaminated soils by diffuse reflectance spectroscopy and sequential ultrasonic solvent extraction–gas chromatography. *Environmental Pollution* 184: 298–305

Paiga, P., Mendes, L., Albergaria, J. and Delerue-Matos, C. (2012). Determination of total petroleum hydrocarbons in soil from different locations using infrared spectrophotometry and gas chromatography. *Chemical Papers* 66 (8) 711–721

Peterson, M.E and Talcott, P.A. (2013). *Small Animal Toxicology.* Elsevier Health Sciences. 928pp

Rusak, D and Brown, L. and Martin, S. (2003). Classification of vegetable oils by principal component analysis of FTIR Spectra. *J Chem Ed.chem* 80 (5): 541 - 543

Shoji, S., Nanzyo, M. and Dahlgren, R. (1994). Volcanic Ash Soils: Genesis, Properties and Utilization. Developments in Soil Science. Elsevier Pp 287

Stenberg, B., ViscarraRossel R.A., Mouazen, A.M. and Wetterlind. J. (2010).*Visible and near infrared spectroscopy in soil science*. In Donald L. Sparks, editon: Advances in Agronomy, Academic Press, 163-215

Stoner, E.R. and Baumgardner. M.F. (1981).Characteristic variations in reflectance of surfaces soils.*Soil Sci. Sot. Am. J.* 45 (6): 1161-1165

Viscarra Rossel, R.A., Walvoort, D.J., McBratney, A.B., Janik, L.J. and 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.

Chapter 6. Conclusion

This study investigated the potential of visible near infrared diffuse reflectance spectroscopy (VNIR DRS) for soil analysis, focusing on soil carbon and extractable petroleum hydrocarbons. This area of research has been characterised by differences in the levels of accuracies reported by previous studies of both soil organic carbon and total petroleum hydrocarbons, including the results obtained within this study. However, this study, in common with most previous studies, shows that diffuse reflectance spectra contain information that can be used to qualitatively discriminate between soils, based on chemical composition. Additionally, combining the spectra with data obtained using alternative analytical procedures in a multivariate regression has demonstrated considerable potential for quantitative characterisation of soil organic carbon and petroleum hydrocarbon contamination in soils. It is, however, important to ensure that reference procedures are accurate and reproducible, as the reliability of regression models rests solely on the quality of data used to populate them.

As the overall aim of multivariate calibrations is to generate robust predictions, they require appropriate regression methods and adequate validation. Within this study, a 76/24 dataset split pattern is identified as an optimal split, ensuring that significant proportions of datasets are used for both model calibration and testing. Little difference was observed between the predictive performances of the support vector machine regression (SVMR) (R^2 = 0.94) and the partial least square regression (PLSR) (R^2 = 0.96). The multiple linear regression model was found to be the least robust (R^2 = 0.90), despite calibrating with wavelength variables ~1150 – 1550nm, ~1650 – 1750nm, ~1800 – 2100nm and ~2200 – 2449nm where absorptions were found to havehigh correlation (R^2 > 0.70) with measured extractible total petroleum hydrocarbon contents of contaminated soil. This implies that modeling soil

parameter using visible near infrared diffuse reflectance spectroscopy may be better done by incorporating all information provided by several spectral features using regression tools that can accommodate multi-collinearity such as the SVMR and the PLSR, to ensure that subtle information are not discarded and the correlations between absorptions at all wavelengths are taken into consideration.

The prediction performances of VNIR DRS models (for both soil carbon and extractable total petroleum hydrocarbons) were observed to reduce with an increase in the geographical size of soil collection sites. VNIR DRS models derived from soils obtained from the University of Reading farm in Wisley exhibited better predictions than models derived from soils collected from a wider area in Southwest England, themselves also performing better than models derived from soils collected across England. This is an indication that constructing robust universal models using large scale universal libraries equipped with VNIR DR spectra of soils from large geographical areas such as the size of England may be hard to achieve and the best calibration models will likely be generated from spectral data derived from local soils with similar geology such as a large farm.

A comparison of the predictive accuracies of visible near infrared diffuse reflectance spectroscopy and mid infrared diffuse reflectance spectroscopy (MIR DRS) showed little difference in performances for soil carbon analysis. This is a positive indication that soil carbon can be predicted with similar accuracies from VNIR spectra, which requires less sample preparation and considerably less costly analytical equipment than MIR DRS.

In the study using agricultural soils contaminated with petroleum hydrocarbon products in the laboratory, the performance of VNIR DRS for characterizing extractible total petroleum hydrocarbons (ETPH) in soils was observed to be affected by organic carbon content of soils. Higher model performances were observed at low organic carbon content (SOC 0.94 – 3.7%,

R^2 = 0.97; SOC 12.4 – 15.6%, R^2 = 0.97; SOC 20 – 26.5% R^2 = 0.91), though all partial least square models developed for the organic carbon range studied (0.94 – 26.5%) had good prediction qualities (RPD > 2). Similarly, the quality of models were also affected by the type of petroleum hydrocarbon product that had been used to contaminate the soils, with models developed for spectra derived from soils that had been contaminated with dark toned petroleum hydrocarbon products (e.g. crude oil sludge) having lower model quality statistics compared with that derived from clear light toned products (e.g. diesel).

Future work

This study suggests additional research should address the following points:

1. Modelling with larger sample sets of soils and a wider range of soil carbon and extractible total petroleum hydrocarbon contents, from larger geological ranges than used in this study.
2. Use of real time on-field spectral measurements with VNIR DR incorporating actual field contaminated soils for petroleum hydrocarbons. This will allow for a study of how atmospheric moisture, size and shape of soil aggregates, management/ farming practices and other factors that can influence soil reflectance contribute to the robustness of VNIR DRS models.
3. An improvement of data selection procedures using algorithms that consider both the spectral data and the reference measurements for selecting a calibration set for regression modelling
4. An assessment of other non-linear regression methods such as the random forests and artificial neural networks which have higher capacities to accommodate the non-linear characteristics of soil spectra for VNIR DR characterization of soil carbon and extractible total petroleum hydrocarbons.

5. Development of soil spectral libraries at the regional scale

In conclusion, considering the high costs and analytical time associated with conventional procedures for soil analysis, the prospect of using VNIR DRS offers a promising, cheaper and faster approach. However, despite some research into its application, the potential of this approach has not been fully exploited.

Appendix i

Weighted regression coefficients of VNIR-PLSR models for the prediction of soil organic carbon (SOC). The magnitude of the regression coefficient at each wavelength is proportional to the importance of the wavelength variable in each model. Each column is independent of other columns.

	Soil group			
Wavelength variables (nm)	Wisley	South west England	England	All three groups
401.38	0.0352	0.2770	0.0000	0.0105
408.23	0.0319	0.2460	0.0000	0.0104
415.12	0.0277	0.1796	0.0177	0.0103
422.05	0.0252	0.0630	0.0169	0.0101
429.01	0.0230	-0.0731	0.0190	0.0098
436.00	0.0201	-0.0698	0.0220	0.0095
443.01	0.0181	-0.1368	0.0252	0.0092
450.05	0.0159	-0.1715	0.0286	0.0089
457.11	0.0132	-0.1502	0.0286	0.0088
464.19	0.0110	-0.1743	0.0280	0.0087
471.28	0.0088	-0.2071	0.0266	0.0086
478.39	0.0067	-0.1696	0.0205	0.0082
485.51	0.0051	-0.1823	0.0161	0.0078
492.64	0.0035	-0.1883	0.0107	0.0072
499.78	0.0016	-0.1895	0.0054	0.0066
506.93	0.0003	-0.1079	0.0014	0.0056
514.09	-0.0005	-0.1084	-0.0026	0.0048
521.26	-0.0017	-0.1059	-0.0045	0.0039
528.44	-0.0031	-0.0374	-0.0038	0.0029
535.62	-0.0041	0.0324	-0.0016	0.0019
542.80	-0.0046	0.1113	0.0005	0.0007
550.00	-0.0052	0.1690	0.0022	-0.0004
557.19	-0.0061	0.1987	0.0052	-0.0014
564.40	-0.0072	0.2390	0.0102	-0.0022
571.60	-0.0082	0.3430	0.0158	-0.0032
578.82	-0.0090	0.3856	0.0208	-0.0039
586.04	-0.0100	0.3617	0.0251	-0.0043
593.26	-0.0112	0.4257	0.0293	-0.0049
600.49	-0.0123	0.4667	0.0329	-0.0053
607.72	-0.0131	0.4869	0.0345	-0.0056
614.96	-0.0140	0.4045	0.0340	-0.0055
622.20	-0.0150	0.3642	0.0339	-0.0055
629.45	-0.0159	0.3690	0.0343	-0.0056
636.70	-0.0166	0.3476	0.0324	-0.0056
643.96	-0.0171	0.3022	0.0339	-0.0054
651.22	-0.0181	0.2842	0.0341	-0.0056
658.49	-0.0180	0.2658	0.0337	-0.0055

Wavelength variables (nm)	Wisley	South west England	England	All three groups
665.76	-0.0182	0.2065	0.0323	-0.0053
673.03	-0.0182	0.0744	0.0287	-0.0049
680.31	-0.0186	0.1327	0.0295	-0.0053
687.60	-0.0186	0.1243	0.0317	-0.0053
694.89	-0.0181	0.0969	0.0328	-0.0051
702.18	-0.0174	-0.0483	0.0309	-0.0044
709.47	-0.0167	-0.1484	0.0283	-0.0040
716.77	-0.0162	-0.1470	0.0251	-0.0041
724.07	-0.0159	-0.1462	0.0227	-0.0041
731.38	-0.0156	-0.1564	0.0218	-0.0041
738.68	-0.0154	-0.1622	0.0223	-0.0040
745.99	-0.0150	-0.1537	0.0242	-0.0040
753.30	-0.0146	-0.1218	0.0261	-0.0041
760.61	-0.0139	-0.0881	0.0277	-0.0041
767.92	-0.0130	-0.1897	0.0286	-0.0033
775.23	-0.0121	-0.3241	0.0293	-0.0024
782.54	-0.0114	-0.3496	0.0296	-0.0021
789.85	-0.0110	-0.3201	0.0302	-0.0020
797.16	-0.0107	-0.2611	0.0310	-0.0021
804.47	-0.0104	-0.2391	0.0320	-0.0018
811.78	-0.0100	-0.2442	0.0331	-0.0014
819.08	-0.0095	-0.2588	0.0332	-0.0009
826.39	-0.0091	-0.2685	0.0332	-0.0004
833.70	-0.0086	-0.2680	0.0333	0.0000
841.00	-0.0083	-0.2733	0.0333	0.0005
848.31	-0.0079	-0.2765	0.0330	0.0010
855.61	-0.0075	-0.2707	0.0323	0.0014
862.92	-0.0072	-0.2583	0.0309	0.0018
870.22	-0.0068	-0.2507	0.0284	0.0022
877.53	-0.0065	-0.2594	0.0268	0.0027
884.85	-0.0063	-0.2820	0.0244	0.0033
892.17	-0.0060	-0.3063	0.0234	0.0039
899.49	-0.0058	-0.3114	0.0237	0.0043
906.82	-0.0055	-0.3160	0.0216	0.0047
914.16	-0.0050	-0.3745	0.0201	0.0055
921.52	-0.0051	-0.3469	0.0171	0.0054
928.89	-0.0055	-0.2199	0.0150	0.0048
936.27	-0.0059	-0.1148	0.0145	0.0044
943.68	-0.0065	-0.0627	0.0128	0.0044
951.11	-0.0072	-0.0308	0.0094	0.0044
958.57	-0.0081	0.0275	0.0057	0.0042
966.06	-0.0093	0.1199	0.0039	0.0037
973.59	-0.0108	0.2294	0.0024	0.0032
981.16	-0.0127	0.3267	0.0016	0.0028
988.78	-0.0147	0.3642	-0.0073	0.0028
996.46	-0.0173	0.4222	-0.0078	0.0025
1025.89	-0.0163	0.6133	-0.0058	0.0013

Wavelength variables (nm)	Wisley	South west England	England	All three groups
1071.13	0.0012	0.4314	-0.0024	0.0021
1115.63	0.0170	0.4336	0.0055	0.0012
1159.38	0.0304	0.4473	0.0086	0.0003
1202.37	0.0426	0.4089	0.0224	-0.0002
1244.60	0.0529	0.4151	0.0351	-0.0006
1286.07	0.0605	0.4297	0.0522	-0.0009
1326.76	0.0648	0.4025	0.0721	-0.0011
1366.68	0.0616	0.3626	0.0605	-0.0011
1405.81	0.0476	0.0445	0.0397	-0.0012
1444.16	0.0441	-0.0619	0.0152	-0.0035
1481.71	0.0489	0.0592	0.0093	-0.0035
1518.47	0.0551	0.1269	0.0301	-0.0024
1554.42	0.0610	0.0701	0.0371	-0.0004
1589.56	0.0643	-0.0545	0.0355	0.0015
1623.89	0.0648	-0.0248	0.0272	0.0020
1657.40	0.0622	0.0396	0.0214	0.0017
1690.08	0.0598	0.0205	0.0232	0.0018
1721.93	0.0557	-0.0695	0.0253	0.0026
1752.95	0.0524	-0.1651	0.0304	0.0035
1783.13	0.0495	-0.2987	0.0272	0.0046
1812.46	0.0463	-0.3744	0.0231	0.0059
1840.94	0.0415	-0.4851	0.0123	0.0071
1868.56	0.0309	-0.7323	0.0069	0.0085
1895.32	-0.0063	-0.9624	0.0324	0.0052
1969.60	-0.0628	-1.7527	0.0413	-0.0016
2020.73	-0.0287	-1.5710	0.0475	0.0023
2070.94	-0.0147	-0.4521	0.0575	-0.0035
2120.23	-0.0110	0.4773	0.0695	-0.0082
2168.59	-0.0126	0.4696	0.0485	-0.0028
2216.04	-0.0202	0.0092	0.0010	0.0064
2262.56	-0.0366	0.2669	-0.0258	0.0007
2308.17	-0.0518	0.6263	-0.0148	-0.0036
2352.85	-0.0616	0.5842	-0.0028	-0.0030
2396.61	-0.0764	0.3030	-0.0696	-0.0015
2439.45	-0.1037	0.4383	0.0000	-0.0059
2481.37	-0.1348	1.0917	0.0000	-0.0146

Appendix ii

Weighted regression coefficients of VNIR-PLSR model for the prediction of extractible total petroleum hydrocarbons (ETPH) in soils. The magnitude of the regression coefficient at each wavelength is proportional to the importance of the wavelength variable in the model.

Wavelenghts	Regression coefficients
401.38	0.1303
408.23	0.0940
415.12	0.0540
422.05	0.0376
429.01	0.0160
436.00	-0.0090
443.01	-0.0221
450.05	-0.0467
457.11	-0.0632
464.19	-0.0841
471.28	-0.0990
478.39	-0.1139
485.51	-0.1203
492.64	-0.1246
499.78	-0.1346
506.93	-0.1347
514.09	-0.1248
521.26	-0.1188
528.44	-0.1154
535.62	-0.1012
542.80	-0.0786
550.00	-0.0566
557.19	-0.0370
564.40	-0.0214
571.60	-0.0046
578.82	0.0134
586.04	0.0259
593.26	0.0326
600.49	0.0395
607.72	0.0489
614.96	0.0555
622.20	0.0568
629.45	0.0587
636.70	0.0634
643.96	0.0685
651.22	0.0699
658.49	0.0820
665.76	0.0915
673.03	0.0970
680.31	0.0977

Wavelenghts	Regression coefficients
687.6	0.1040
694.89	0.1151
702.18	0.1272
709.47	0.1361
716.77	0.1438
724.07	0.1478
731.38	0.1499
738.68	0.1506
745.99	0.1506
753.30	0.1524
760.61	0.1542
767.92	0.1580
775.23	0.1587
782.54	0.1534
789.85	0.1438
797.16	0.1326
804.47	0.1199
811.78	0.1080
819.08	0.0945
826.39	0.0794
833.70	0.0642
841.00	0.0469
848.31	0.0287
855.61	0.0106
862.92	-0.0088
870.22	-0.0281
877.53	-0.0486
884.85	-0.0667
892.17	-0.0885
899.49	-0.1061
906.82	-0.1219
914.16	-0.1394
921.52	-0.1607
928.89	-0.1856
936.27	-0.2054
943.68	-0.2300
951.11	-0.2506
958.57	-0.2741
966.06	-0.3002
973.59	-0.3304
981.16	-0.3603
988.78	-0.3818
996.46	-0.4176
1025.89	-0.4373
1071.13	-0.3398
1115.63	-0.2461
1159.38	-0.1469
1202.37	-0.0561

Wavelenghts	Regression coefficients
1244.60	-0.0156
1286.07	0.0221
1326.76	0.0557
1366.68	0.0608
1405.81	0.0300
1444.16	0.0750
1481.71	0.1056
1518.47	0.1276
1554.42	0.1524
1589.56	0.1728
1623.89	0.1794
1657.40	0.2003
1690.08	0.3201
1721.93	0.4140
1752.95	0.3697
1783.13	0.2972
1812.46	0.2245
1840.94	0.1530
1868.56	0.0625
1895.32	-0.1762
1969.60	-0.2729
2020.73	-0.0251
2070.94	0.1169
2120.23	0.1629
2168.59	0.1607
2216.04	0.0466
2262.56	0.1508
2308.17	0.1374
2352.85	0.0582
2396.61	-0.1107
2439.45	-0.3003
2481.37	-0.4777