2011

Rapid identification of oil contaminated soils using visible near infrared diffuse reflectance spectroscopy

Somsubhra Chakraborty
Louisiana State University and Agricultural and Mechanical College

RAPID IDENTIFICATION OF OIL CONTAMINATED SOILS USING VISIBLE NEAR INFRARED DIFFUSE REFLECTANCE SPECTROSCOPY

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The School of Plant, Environmental & Soil Sciences

by
Somsubhra Chakraborty
B.S., Bidhan Chandra Krishi Viswavidyalaya, 2006
M.S., Punjab Agricultural University, 2008
August 2011

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Weindorf, and my committee members, Drs. Morgan, Galbraith, Li, and Selim for their constant support and guidance during the course of my research.

I would also like to thank my dean representative Dr. Tsai, my friends, fellow graduate students, Dr. Zhu, faculty, and staff at Louisiana State University for their guidance and continual support. I gratefully acknowledge financial assistance from the Louisiana Applied Oil Spill Research Program (LAOSRP). I appreciate the laboratory work which Ms. Noura Bakr completed on ICP. I would also like to thank all the donors of scholarships that I have received during my time at Louisiana State University. I am grateful to the ES&H Consulting and Training Group and Stone Energy Corporation for supplying crude oil for my research.

Finally, thanks to my parents and fiancée for their patience and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Initially, 46 petroleum contaminated and non-contaminated soil samples were collected and scanned using visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) at three combinations of moisture content and pretreatment. The VisNIR spectra of soil samples were used to predict total petroleum hydrocarbon (TPH) content using partial least squares (PLS) regression and boosted regression tree (BRT) models. The field-moist intact scan proved best for predicting TPH content with a validation r^2 of 0.64 and relative percent difference (RPD) of 1.70. Those 46 samples were used to calibrate a penalized spline (PS) model. Subsequently, the PS model was used to predict soil TPH content for 128 soil samples collected over an 80 ha study site. An exponential semivariogram using PS predictions revealed strong spatial dependence among soil TPH [r^2 = 0.76, range = 52 m, nugget = 0.001 $(\log_{10}$ mg $kg^{-1})^2$, and sill 1.044 $(\log_{10}$ mg $kg^{-1})^2$]. An ordinary block kriging map produced from the data showed that TPH distribution matched the expected TPH variability of the study site. Another study used DRS to measure reflectance patterns of 68 artificially constructed samples with different clay content, organic carbon levels, petroleum types, and different levels of contamination per type. Both first derivative of reflectance and discrete wavelet transformations were used to preprocess the spectra. Principal component analysis (PCA) was applied for qualitative VisNIR discrimination of variable soil types, organic carbon levels, petroleum types, and concentration levels. Soil types were separated with 100% accuracy, and organic carbon levels were separated with 96% accuracy by linear discriminant analysis. The support vector machine produced 82% classification accuracy for organic carbon levels by repeated random splitting of the whole dataset. However, spectral absorptions for each petroleum hydrocarbon overlapped with each other and could not be separated with any classification scheme when contaminations were mixed. Wavelet-based multiple linear regression performed best for predicting petroleum amount

with the highest residual prediction deviation (RPD) of 3.97. While using the first derivative of reflectance spectra, PS regression performed better (RPD = 3.3) than the PLS (RPD= 2.5) model. Specific calibrations considering additional soil physicochemical variability are recommended to produce improved predictions.

CHAPTER 1

INTRODUCTION

1.1. Oil Spills and Visible Near-infrared Diffuse Reflectance Spectroscopy

The term "petroleum" originated from the Latin word "petra" meaning rock and "oleum" meaning oil. However, in the present day world, petroleum is used synonymously with crude oil and natural gas. Crude oil which is the raw, unprocessed oil, is mostly mined from deep geologic deposits (both inland and off-shore). Crude oil and natural gas are the sources of large number of consumer products ranges from gasoline and kerosene to asphalt and chemical agents have played a crucial role in energy sector and numerous petrochemical industries (Chung et al., 1999). The United States ranks third in crude oil production. However, crude oil alone does not satisfy the total U.S. petroleum demand. Therefore, high internal petroleum demand has led the U.S. to be a major consumer of crude oil and refined petroleum products[1] (Fig. 1) (USEIA, 2010). During 2009, while net imports and domestic petroleum as shares of total U.S. need accounted for 49% and 51%, respectively; daily consumption of petroleum products was 18.8 million barrels (Fig. 2). Daily petroleum production (including crude oil, natural gas plant liquid, and other oils) in U.S. during 2009 was 7.27 million barrels day^{-1}. This massive dependency on petroleum and petroleum products inevitably brings about frequent spillage accidents during petroleum extraction, refinement, and transportation (Fig. 3) (USCG, 1999). Moreover, natural disasters like hurricanes Katrina, Rita, and Wilma in 2005 can also cause oil spill accidents (Burgherr, 2006). The disruption of fragile marine and coastal environments due to oil spills has become a sensitive environmental issue during recent years. Off shore oil spills breach the food chain by poisoning benthic, neritic, and pelagic organisms. The Amoco Cadiz, Exxon Valdez, and Deepwater Horizon oil spills have stressed the need for effective contingency planning and

[1]This dissertation follows the style of "Soil Science Society of America Journal".

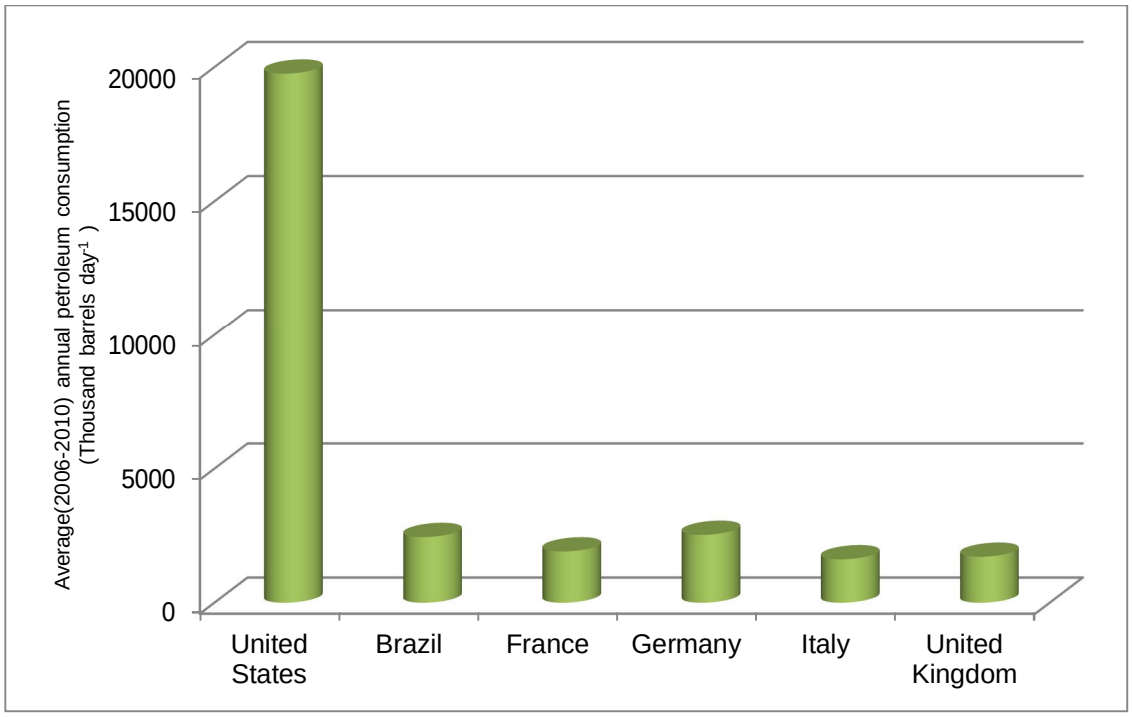

Fig. 1. Average (2006-2010) annual petroleum consumption of different countries. (USEIA, 2010).

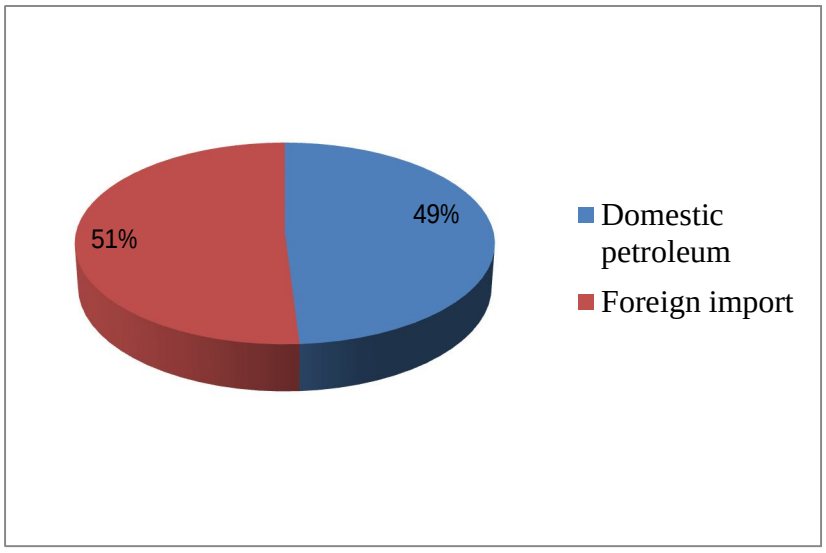

Fig. 2. Net foreign import and domestic petroleum as shares of total U.S demands in 2009. (USEIA, 2010).

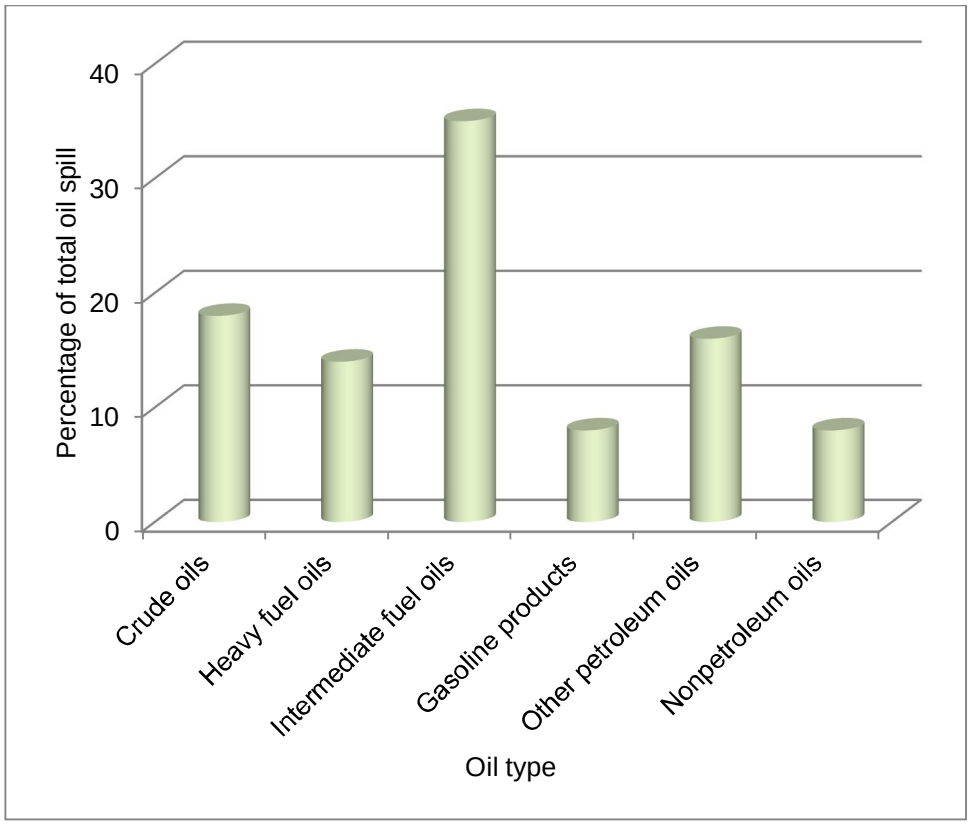

Fig. 3. Reported oil spills by type (annual average from 1994 to 1999) (USCG, 1999).

monitoring systems to reduce the impact of devastating oil spills. However, massive oil spills from offshore rig blowout and tanker accidents only characterize a small portion of total annual oil spillage. According to the Emergency Response Notification System (ERNS), the inland oil spills from pipelines, storage tanks, fixed industrial facilities, railways, automobiles, and barges account for the major part of total oil spill (Stalcup et al., 1997). Pipelines are a major source of oil spill in U.S as they have spilled more oil than tankers and barges combined since 1985 (Etkin, 2001). The damages caused by oil spills ranges from human health hazards to environmental damage. Polycyclic aromatic hydrocarbons (PAH), which mainly occur in petroleum, are potentially carcinogenic to human health. Inland oil spill causes human exposure to PAHs by inhalation, ingestion, and skin contact (Bofetta et al., 1997). Pérez-Cadahía et al. (2007) reported that PAHs form covalent links with macromolecular DNA and initiate the process of chemical carcinogenesis. According to International Agency for Research on Cancer (1983) small amounts of PAHs can induce malignant tumors that affect mostly the skin and other epithelial tissues of humans. Moreover, oil spill results in considerable loss of wetlands. Therefore, there is an extreme need for reliable monitoring approaches for effective response to inland oil spills.

According to Clark (1999), spectroscopy deals with light as a function of wavelength that has been emitted, reflected or scattered from a solid, liquid, or gas. Spectral measurement can measure the quantity of reflected or transmitted light from an object (Workman and Shenk, 2004). Since soil is heterogeneous in nature, the standard physicochemical evaluations of soil attributes are both expensive and time-consuming (Chang et al., 2001). However, application of spectroscopy in the visible to near-infrared range (350-2500 nm) in combination with regression analysis (multilinear, principal component regression, and partial least squares regression) offers

a non-destructive means for measurement of different soil components (Meyer, 1989; Sudduth and Hummel, 1993; Reeves et al., 2001; McCarty et al., 2002).

Spectral preprocessing involving transformations, averaging, splicing, and smoothing is a prerequisite for a robust regression model. Transformation of lognormal data is essential to maintain the assumption of normality when models with linear framework like partial least squares regression and principal component regression are considered. Derivative spectroscopy is a useful tool for simplifying a large hyperspectral database, thereby removing the inherent noise of the spectral data. Tsai and Philpot (1998) reported that the first and second derivatives of the reflectance spectra significantly enhance the absorption features by reducing the albedo effect. They created a modular program to execute derivative analysis by means of either a convolution (Savitzky-Golay) or finite divided difference algorithm. Earlier, the simplified least-square-fit convolution for smoothing and computing first and higher order derivatives from the average reflectance was revealed by Savitzsky and Golay (1964). According to Williams (1987), smoothing helps in removing the noises which arise from atmospheric influences and light intensity fluctuations.

By utilizing the visible to near infrared range (350-2500 nm) of the electromagnetic spectrum, radiation reflected from the soil surface can be modeled against total petroleum hydrocarbon (TPH) content of the petroleum contaminated soil, determined by traditional wet chemical analysis. Furthermore, the constructed statistical model could be successfully utilized to quantify the TPH from unknown soil samples. Scientists generally use TPH, a mixture of different hydrocarbons as an indicator of petroleum contaminated soils. Visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) permits rapid and cost-effective quantification of TPH as compared to the traditional costly and time consuming chemical analyses (Schwartz et

al., 2009). Cloutis (1989) reported two characteristic absorption peaks (1730 nm and 2310 nm) as spectral signatures of hydrocarbon-bearing substances. Schneider et al. (1995) assembled a near-infrared fiber-optic chemical sensor for remote detection of organics in soils. Stallard et al. (1996) investigated motor oil contaminated sandy loam soil with near-infrared reflectance spectroscopy and indicated the potentiality of this technology for other types of soil matrices. However, Malley et al. (1999) reported low correlation (r=0.68) while dealing with field collected samples. To date, information on the prediction accuracy of VisNIR DRS when quantifying TPH in soil samples is lacking. Moreover, none of the previous studies explicitly revealed the effects of different types of petroleum fractions, soil organic carbon, and soil textures on spectral reflectance variability of petroleum contaminated soils.

The overall objectives of this research project were to: 1) determine the prediction accuracy of VisNIR DRS in quantifying the amount of hydrocarbons in contaminated soils; 2) compare the accuracies of partial least squares regression and boosted regression trees in predicting TPH in contaminated soils; 3) investigate whether the combination of VisNIR spectroscopy and geostatistics has the potential to identify the spatial distribution of TPH contamination; and 4) examine the effect of variable soil texture, organic carbon, and oil types on VisNIR reflectance patterns of petroleum contaminated soils. We used an ASD AgriSpec VisNIR portable spectroradiometer (Analytical Spectral Devices, Boulder, CO) with a spectral range of 350 to 2,500 nm [a) ultra violet / VISNIR (350-965 nm), b) short wave infrared 1 (966-1,755 nm), and c) short wave infrared 2 (1,756-2,500 nm)] (Fig 4.). The results of this pilot research will help to determine the potentiality and limitations of using a VisNIR DRS for quantifying petroleum contaminated soils.

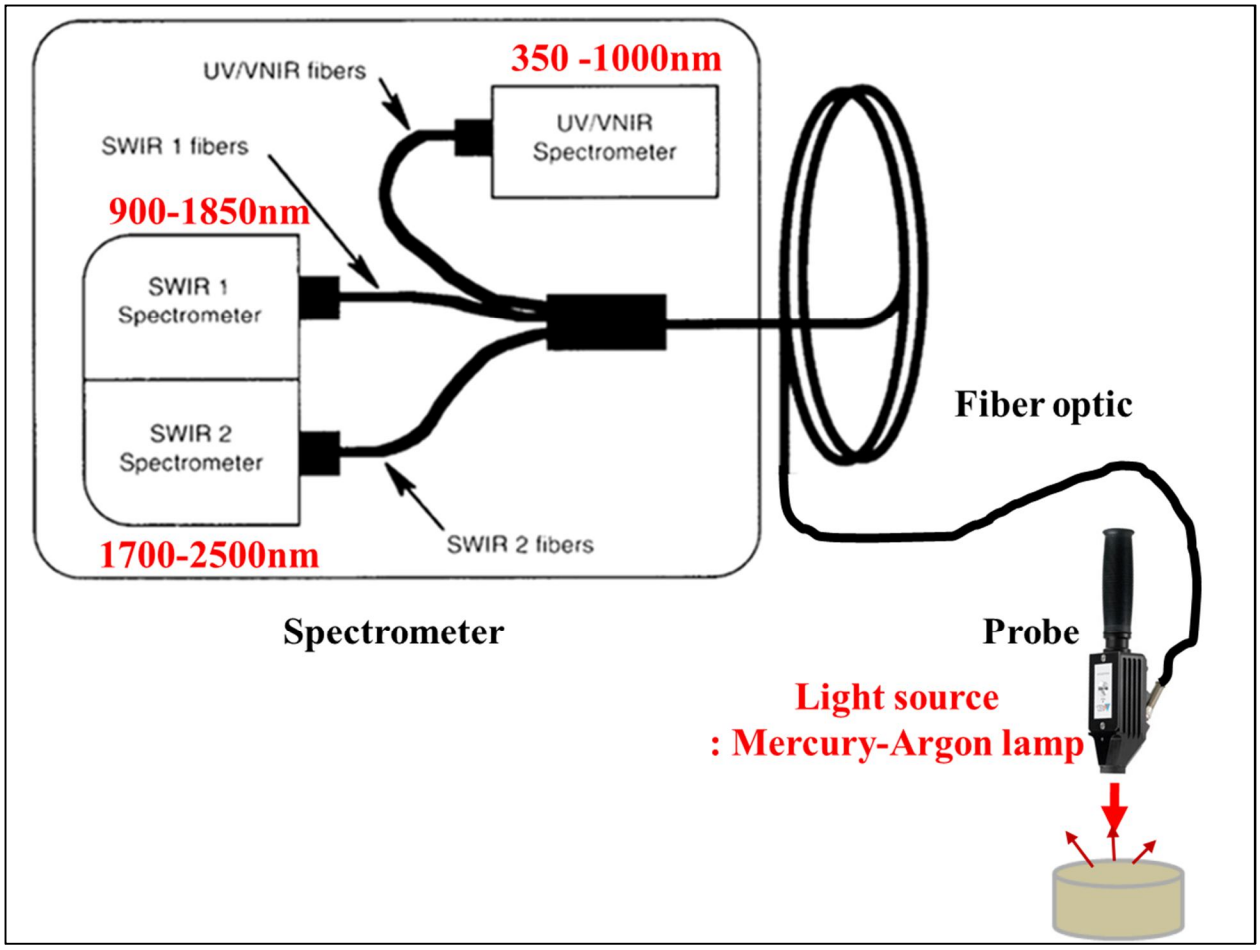

Fig. 4. Inner diagram of an ASD AgriSpec® portable visible and near-infrared diffuse reflectance spectrometer. (D. Hatchell, personal communication, 2009)

1.2. References

Bofetta, P., N. Jourenkova, and P. Gustavson. 1997. Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons. Cancer Causes and Control 8(3):444-472.

Burgherr, P. 2006. In-depth analysis of accidental oil spills from tankers in the context of global oil spill trends from all sources. J. Hazard. Mater. 140: 245-256.

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

Chung, H., H. Choi, and M. Ku. 1999. Rapid identification of petroleum products by near-infrared spectroscopy. Bull. Korean Chem. Soc. 20:1021-1025.

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

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

Etlkin, D.S. 2001. Analysis of oil spill trends in the United States and world wide. p. 1291-1300. *In* Proc. Int. Oil Spill Conf. 2001. American Petroleum Institute, Washington, DC.

International Agency for Research on Cancer. 1983. IARC monographs on evaluation of polynuclear aromatic compounds, part 1, chemical, environmental, and experimental data. I.A.R.C. monographs on the evaluation of carcinogenic risks to humans, vol. 32. International Agency for Research on Cancer, Lyon, France.

Malley, D.F., K.N. Hunter, G. R. Barrie Webster. 1999. Analysis of diesel fuel contamination in soils by near-infrared reflectance spectrometry and solid phase microextraction-gas chromatography. Soil Sediment Contam. 8(4):481–489.

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

Meyer, J.H. 1989. Rapid simultaneous rating of soil texture, organic matter, total nitrogen and nitrogen mineralization potential by near infra-red reflectance. S. Afr. Tydskr. Plant Grond. 6:59–63.

Pérez-Cadahía B, A. Lafuente, T. Cabaleiro, E. Pásaro, J. Méndez, and B. Laffon. 2007. Initial study on the effects of Prestige oil on human health. Environ. Int. 33: 176–185.

Reeves, J.B., III , G.W. McCarty, and V.B. Reeves. 2001. Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soil. J. Agric. Food Chem. 49:766–772.

Savitzky, A., and M.J.E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36:1627–1639.

Schwartz. G, G. Eshel, M. Ben-Haim, and E. Ben-Dor. 2009. Reflectance spectroscopy as a rapid tool for qualitative mapping and classification of hydrocarbons soil contamination. Available at http://www.earsel6th.tau.ac.il/3080%20Schwartz.pdf (verified 2 June. 2011).

Schneider, I., G. Nau, T.V.V. King, and I. Aggarwal. 1995. Fiber-optic near infrared reflectance sensor for detection of organics in soils. Photonics Technology Letters, IEEE. 7(1): 87-89.

Stalcup, D., G. Yoshika, E. Mantus, and B. Kaiman. 1997. Characteristics of oil spills: inland versus coastal. *In* Proc. Int. Oil Spill Conf., Fort Lauderdale, FL. 7-10 Apr.

Stallard, B.R., M.J. Garcia, and S. Kaushik. 1996. Near-IR reflectance spectroscopy for the determination of motor oil contamination in sandy loam. Appl. Spect. 50:334-338.

Sudduth, K. A., and J. W. Hummel. 1993. Portable, near–infrared spectrophotometer for rapid soil analysis. Trans. ASAE 36:185–193.

Tsai F., and W. Philpot. 1998. Derivative analysis of hyperspectral data. Remote Sens. Environ. 66:41-51.

U.S. Coast Guard. 1999. Volume of spills by source, oil spill compendium data tables. Available at http://www.bts.gov/publications/transportation_statistics_annual_report/2000/chapter6/oil_spills_fig1.html (verified 1 June, 2011).

U.S. Energy Information Administration. 2010. How dependent are we on foreign oil. Available at http://www.eia.gov/energy_in_brief/foreign_oil_dependence.cfm (verified 1 June, 2011). USEIA, Washington, DC.

Williams, P.C. 1987. Variables affecting near-infrared reflectance spectroscopy analysis. p. 143-167. *In* P.C. Williams and K. Norris (ed.) Near-infrared technology in the agricultural and food industries. Am. Assoc. Cereal Chem., St. Paul, MN.

Workman, J., Jr. and J. Shenk. 2004. Understanding and using the near-infrared spectrum as an analytical method. p. 3-10. *In* C.A. Roberts, J. Workman, Jr., and J.B. Reeves III (ed.) Near-infrared spectroscopy in agriculture. ASA, CSSA, and SSSA, Madison, WI.

CHAPTER 2

RAPID IDENTIFICATION OF OIL CONTAMINATED SOILS USING VISIBLE NEAR-INFRARED DIFFUSE REFLECTANCE SPECTROSCOPY

2.1. Synopsis

In the United States, petroleum extraction, refinement, and transportation present countless opportunities for spillage mishaps. A method for rapid field appraisal and mapping of petroleum hydrocarbon contaminated soils for environmental cleanup purposes would be useful. Visible near-infrared (VisNIR, 350–2,500 nm) diffuse reflectance spectroscopy (DRS) is a rapid, non-destructive, proximal-sensing technique that has proven adept at quantifying soil properties *in-situ*. The objective of this study was to determine the prediction accuracy of VisNIR DRS in quantifying petroleum hydrocarbons in contaminated soils. Forty-six soil samples (including both contaminated and reference samples) were collected from six different parishes in Louisiana. Each soil sample was scanned using VisNIR DRS at three combinations of moisture content and pretreatment: a) field-moist intact aggregates, b) air-dried intact aggregates, c) and air-dried ground soil (sieved through a 2-mm sieve). The VisNIR spectra of soil samples were used to predict total petroleum hydrocarbon (TPH) content in the soil using partial least squares (PLS) regression and boosted regression tree (BRT) models. Each model was validated with 30% of the samples that were randomly selected and not used in the calibration model. The field-moist intact scan proved best for predicting TPH content with a validation r^2 of 0.64 and relative percent difference (RPD) of 1.70. Because VisNIR DRS is promising for rapidly predicting soil petroleum hydrocarbon content, future research is warranted to evaluate the methodology for identifying petroleum contaminated soils.[2]

[2] Reprinted by permission of "Journal of Environmental Quality".

2.2. Introduction

While petroleum provides abundant energy, economic, and manufacturing resources for the United States, its extraction, refinement, and transportation also present innumerable opportunities for spillage accidents or operational losses. Given that petroleum hydrocarbon is a potential soil and water contaminant and neurotoxin for humans and animals (Schwartz et al., 2009), long term exposure could increase the risk of lung, skin, and bladder cancer (Hutcheson et al., 1996; Boffetta et al., 1997). The protection and enhancement of the nation's natural resource base and environment require the development of innovative, low cost, and reproducible analytical tools to assess the spatial and temporal variability of soil and soil contamination. So far, researchers have established several spectroscopic techniques to identify specific petroleum properties including the application of nuclear magnetic resonance or near-infrared spectroscopy for predicting octane numbers of gasoline compounds, along with the quantification of petroleum contaminants based upon combinations and overtones of C-H, N-H, O-H, and S-H bonds (Kelly et al., 1989; Dorbon et al., 1990; Stallard et al., 1996; Lee and Chung, 1998). Crude oil signatures originate mainly from combinations or overtones of C-H stretching vibrations of saturated CH_2 and CH_3 groups in addition to methylenic, oleifinic, or aromatic C-H functional groups (Aske et al., 2001). The introduction of Urbach tail edge detection technology (Mullins et al., 1992) has established distinctive spectral signatures for most crude oils in the near-infrared region (2,298 nm [stretch+bend]; 1,725 nm [two-stretch]; 1,388 nm [two-stretch+bend]; 1,190 nm [three-stretch]; 1,020 nm [three-stretch+bend]; 917 nm [four-stretch]). Chung et al. (1999) reported 95% (for light gas oil) and 99 % (for light straight-run, kerosene, gasoline, and diesel) near-infrared prediction accuracy, while principal component analysis (PCA) combined with Mahalanobis distance could be used to segregate unique spectral signatures for each of those

fractions. Moreover, internal research from Analytical Spectral Devices (ASD) (Boulder, CO) has clearly reported unique spectral reflectance signatures for crude oil, hexane, and diesel fuel (D. Hatchell, personal communication, 2007). However, the inherent complexity of petroleum composition has made it impossible to screen out any particular spectroscopic technique for the whole range of petroleum spectral signatures. The task of identifying a specific petroleum signature becomes more complex when petroleum products are mixed with another heterogeneous mixture like soil (Wang and Fingas, 1997).

Visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) is a scanning technology that has recently become popular for rapidly quantifying and identifying soil properties in the laboratory and on-site (*in-situ*). Stoner and Baumgardner (1981) reported close association between soil parameters and their spectral reflectance curve forms. Krishnan et al. (1980) utilized spectral reflectance in the VisNIR range to select optimal wavelengths for predicting percent organic matter content in soil. Moreover, simultaneous predictions of total organic carbon, total nitrogen, and moisture content of air-dried soils were carried out utilizing reflectance at three wavelengths in the form of log (1/R) (Dalal and Henry, 1986). In the laboratory, VisNIR DRS has been used to quantify soil electrical conductivity, pH, organic carbon, particle size, mineralogy, cation exchange capacity, nutrients, lime requirement, and clay mineralogy, both rapidly and non-destructively (Henderson et al., 1992; Thomasson et al., 2001; Shepherd and Walsh, 2002; Brown et al., 2006; Madari et al., 2006; Viscarra Rossel et al., 2006a; Viscarra Rossel et al., 2006b; Vasques et al., 2009).

This proximal soil sensing technology, which is well suited for rapid scanning, has been used with portable equipment, on-site, to quantify soil organic and inorganic carbon, and clay content (Sudduth and Hummel, 1993; Ge et al., 2007; Waiser et al., 2007; Morgan et al., 2009).

Several soil spectral libraries were created using a wide array of soils considering their physicochemical and biological properties (Ben-Dor et al., 1999; Malley et al., 2000; Chang et al., 2001). However, only a few studies have been reported so far on VisNIR DRS characterization of oil contaminated soils. Malley et al. (1999) reported linear regression relationships between NIR-predicted total petroleum hydrocarbon (TPH) concentrations and reference data. Additionally, VisNIR DRS has been used to show unique reflectance patterns for bitumen (a heavy, tar-like hydrocarbon used in making asphalt) in a sand-clay-water matrix under field conditions in Alberta, Canada (Analytical Spectral Devices, 2007). A portable version of the ASD spectrometer has become a useful tool for mapping the spatial extent (vertical and horizontal) of oil spills.

The standard gravimetric lab method (Clesceri et al., 1998) is time consuming and costly (~USD 50 per sample). If reliable models that estimate contamination concentrations could be developed and validated for on-site VisNIR spectroscopy, oil and hydrocarbon contamination in soils could be rapidly mapped, minimizing time consuming laboratory measurements. Conversely, if VisNIR DRS cannot be used on-site, soil samples could be collected, air-dried and scanned in a matter of hours under lab conditions. Either way, a tool for rapid identification, mapping, and quantification of oil and hydrocarbon spills in soils could be obtained. Therefore, the overall goal of this study was the successful combination of spectrometry and chemometry to investigate the usefulness of VisNIR DRS for predicting petroleum hydrocarbons in contaminated soils.

Application of the technology and methods tested in this study could be used for rapidly and inexpensively identifying concentrated areas of contamination requiring remediation prior to rebuilding. Furthermore, contamination might be recognized in areas where it may have gone

undetected. Hence, the specific objectives of this research were the following: 1) determine the prediction accuracy of VisNIR DRS in quantifying the amount of hydrocarbons in contaminated soils; and 2) compare the accuracies of partial least squares regression and boosted regression trees in predicting TPH in contaminated soils.

2.3. Materials and Methods

2.3.1. Soil Samples

Forty-six soil samples (including both contaminated and reference or uncontaminated samples) were collected from six sites each located in a different parish within southern and central Louisiana (Table 1). The sampling scheme was carefully developed in accordance with the prior knowledge of oil spill in locations provided by Louisiana Oil Spill Coordinators Office (LOSCO) to ensure maximum TPH variability within the soil samples collected. Areas of known oil contamination or spillage were identified by visible evidence or odor of petroleum and sampled first. Subsequently, nearby areas of similar soil series with no known contamination were identified and sampled. The samples were collected to a depth of 15 cm and placed in air-tight glass bottles to prevent hydrocarbon volatilization and preserve field moisture status. Samples were placed on ice for transport to the lab and refrigerated at 5^{o}C in the lab. The official soil series description of each sampling site showed a wide variation in soil properties between sites (Table 1).

2.3.2. VisNIR DRS Scanning

The collected soil samples were scanned with an ASD AgriSpec VisNIR portable spectroradiometer (Analytical Spectral Devices, Boulder, CO) with a spectral range of 350 to 2,500 nm [a) ultra violet / VISNIR (350-965 nm), b) short wave infrared 1 (966-1,755 nm), and

c) short wave infrared 2 (1,756-2,500 nm)]. The spectroradiometer had a 2-nm sampling resolution and a spectral resolution of 3 nm and 10 nm wavelengths from 350–1,000 nm and 1,000–2,500 nm, respectively. For field-moist, intact aggregates scanning, each sample was spread evenly on a plastic dish and scanned with a contact probe, having a 2-cm diameter circular viewing area and built-in halogen light source. Each sample was scanned four times with the contact probe at different locations within a sample to obtain multiple sample spectra for averaging purposes. Each individual scan was an average of 10 internal scans over a time of 1.5 seconds. The detector was white referenced using a white spectralon panel with 99% reflectance, ensuring that fluctuating downwelling irradiance could not saturate the detector. Moreover, white referencing removes dark current and ambient temperature humidity variation effects. After scanning, the samples were again bottled and sent to a commercial lab for TPH analysis. Following TPH analysis, the samples were air dried, equally divided into two parts (weight basis), and placed into separate air-tight plastic bags. For each air-dried sample, one part was left as intact aggregates and the other part was ground to pass a 2-mm sieve to produce air-dried ground soil for scanning. Thirty (30) grams of each sample was spread evenly in a borosilicate optical-glass petri dish and scanned from below four times with a muglight (high intensity source probe with a halogen light source), connected to the ASD AgriSpec. Between each of the four scans, the sample was rotated 90^{o}.

2.3.3. Laboratory Analysis

In the commercial lab, petroleum in soil samples was extracted using method 5520 D Soxhlet extraction (Clesceri et al., 1998) and TPH was quantified by method 5520 F (Clesceri et al., 1998). In the Soxhlet extraction, petroleum was extracted at a rate of 20 cycles h^{-1} for 4 h using *n*-hexane or solvent mixture (80% *n*-hexane/20% Methyl tert-butyl ether, v/v).

Table 1. Location, soil series, and classification of soils evaluated for petroleum contamination using visible and near infrared diffuse reflectance spectroscopy in Louisiana, USA.

Site	Parish	Soil series	Classification†	Contaminated samples	Non-contaminated samples
Alpine	Jefferson	Barbary	Very-fine, smectitic, nonacid, hyperthermic Typic Hydraquents	6	6
Mississippi River 1	Plaquemine	Carville	Coarse-silty, mixed, superactive, calcareous, hyperthermic Fluventic Endoaquepts	1	1
Mississippi River 2	Saint Charles	Cancienne	Fine-silty, mixed, superactive, nonacid, hyperthermic Fluvaquentic Epiaquepts	1	1
Sabine	Cameron	Creole	Fine, smectitic, nonacid, hyperthermic Typic Hydraquents	4	4
Sonat	Vernon	Ruston	Fine-loamy, silicious, semiactive, thermic Typic Paleudults	6	6
Winn Dixie	East Baton Rouge	N/A	Udarents	5	5

† Soil Survey Staff, 2005.

For gravimetric determination of TPH (method 5520 F), the extracted oil was redissolved in *n*-hexane and an appropriate amount of silica gel was added. The solution was stirred with a magnetic stirrer for 5 minutes and filtered through a filter paper pre-moistened with solvent.

Other laboratory analyses of each soil sample consisted of standard physical and chemical soil analyses, including particle size analysis by modified hydrometer method with 24 h and 40 s clay and sand determinations (Gee and Bauder, 1986), respectively; saturated paste pH (Soil Survey Staff, 2004), salinity (Soil Survey Staff, 2004), and total organic carbon by dry oxidation (Nelson and Sommers, 1996). All samples were subjected to Mehlich III extraction (Mehlich, 1984) and ion concentrations in the extracted solution were quantified by inductively coupled argon plasma (ICAP) analysis (Soltanpour et al., 1996) with a SPECTRO CIROS CCD (SPECTRO Analytical Instruments Inc, NJ, USA). X-ray diffraction analysis was conducted for bulk soil mineralogy confirmation (Wittig and Allardice, 1986) on selected representative samples. Siemens Diffrac AT V3.1 software was used to run the Siemens D5000 X ray diffractometer (Bruker AXS Inc., Madison, WI, USA). The MacDiff 4.0.0 program, a Macintosh shareware application, was used to interpret each representative sample using the International Centre for Diffraction Data's Powder Diffraction File. Estimates of quantitative mineral abundance (% weight basis) were obtained with XRDFIL, a computer application based on the technique described by Cook et al. (1975), except that the total clay mineral peak intensity factor was changed to 20.

2.3.4. Spectral Preprocessing

Spectral data was processed in 'R' (R Development Core Team, 2004) using custom 'R' routines (Brown et al., 2006). These routines included: (i) a parabolic splice to correct for "gaps" between detectors; (ii) averaging replicate spectra; and (iii) fitting a weighted (inverse

measurement variance) smoothing spline to each spectra with direct extraction of smoothed reflectance; (iv) first derivatives at 10-nm intervals; and subsequently (v) second derivatives at 10-nm intervals. The resulting 10-nm average reflectance, first derivative, and second derivative spectra were individually combined with the laboratory measured TPH contents. These processed data were used to create prediction models using partial least squares (PLS) regression and boosted regression tree (BRT) analyses.

2.3.5. Partial Least Squares: Model Calibration and Validation

Partial least squares regression was used to develop TPH prediction models through spectral decomposition. This regression technique produces robust statistical models by utilizing all available soil reflectance data (Vasques et al., 2009).

In the present study, the original TPH contents of the samples were widely and non-normally distributed from 44.3 to 48,188 mg $kg^{-1}$of soil. Therefore, the Box-Cox transformation (Box and Cox, 1964) was applied to the original TPH data and the original data (λ=1) was $\log_{10}$-transformed (λ=0) to make it more normal (Fig. 5.). Thus, PLS models were developed based on $\log_{10}$-transformed data that approximated a Gaussian distribution after stabilizing the variance.

A total of nine models were developed using the PLS algorithm with Unscrambler 9.0 (CAMO Software, Woodbridge, NJ) to identify the effects of different levels of soil processing on VisNIR DRS prediction of TPH. In response to the variability of TPH distribution, 70% (32) of the samples were randomly selected to build the calibration or training dataset and the remaining 30% (14) were used for the validation or testing dataset. The same split for the calibration, or training dataset, was used for all scans with leave-one-out cross-validation for model creation and selection for the number of latent factors. Models with as many as ten factors

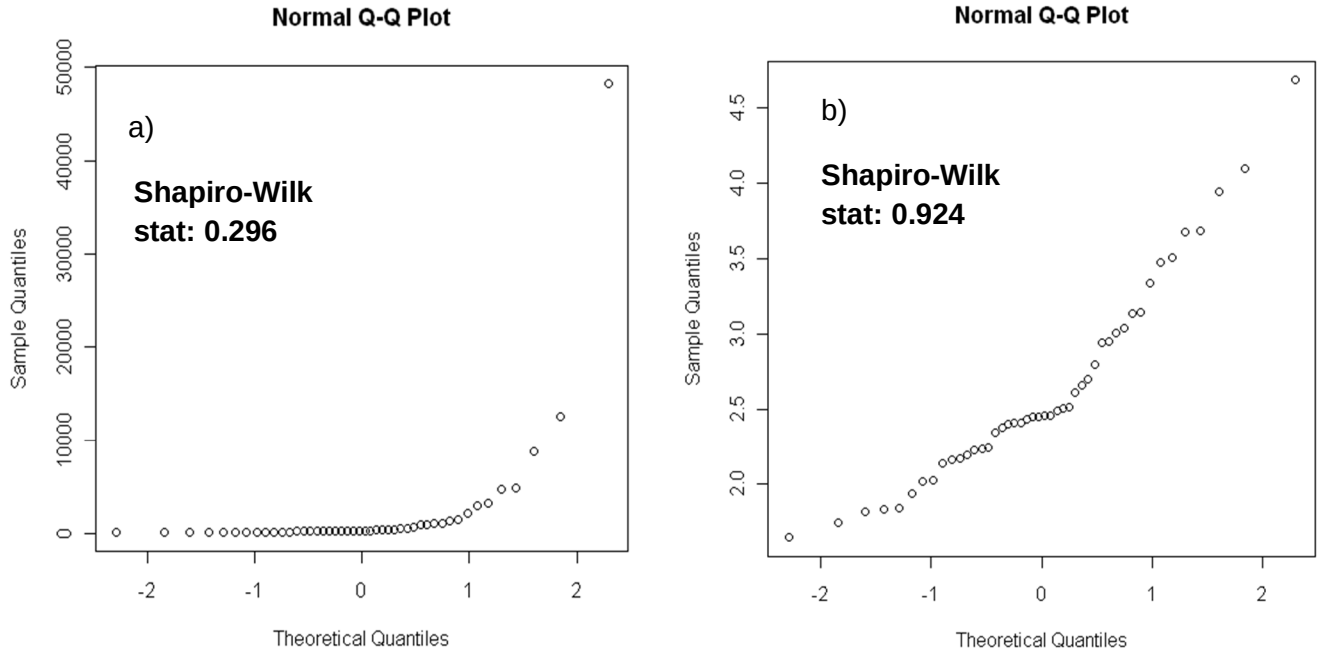

Fig. 5. a) Original (λ=1) and b) log-transformed (λ=0) total petroleum hydrocarbon (TPH) contents of the soil samples collected from six different parishes in Louisiana, USA. Increase in Shapiro-wilk statistic for log-transformed data revealed that the Box-Cox transformation made the original TPH data more normal.

were considered, and the optimal model was determined by choosing the number of factors with the first local minimum in root mean square error of cross-validation ($RMSE_{cv}$). The significant wavelengths in the first derivative model were plotted to identify what portions of the spectra were important for TPH predictions. The significant wavelengths ($p<0.05$) were selected by 'R' based on Tukey's jackknife variance estimate. The coefficient of determination (r^2), root mean square error of prediction ($RMSE_p$), relative percent difference (RPD), and bias were calculated for each model using the validation data. The statistical formulae of the aforementioned indicators followed Gauch et al. (2003), Brown et al. (2005), and Chang et al. (2005) in the following equations:

$$RMSE_p = \sqrt{\sum_n \frac{\left(TPH_{pred} - TPH_{meas}\right)^2}{n}} \quad [1]$$

$$RPD = SD / RMSE_p \quad [2]$$

$$Bias = \sum_n \frac{\left(TPH_{pred} - TPH_{meas}\right)}{n} \quad [3]$$

Where, SD is the standard deviation of measured TPH of the validation data, and n is the number of validation data.

2.3.6. Boosted Regression Tree Analysis

Following Friedman's Gradient Boosting Machine (Friedman, 2001), boosted regression tree (BRT) analysis was employed as well. These models have the ability to partition the data, creating more homogeneous classes by separating the target variables recursively (Vasques et al., 2009). The analysis was carried out by Treenet®2.0 (Steinberg et al., 2002)(Salford Systems, San

Diego, CA, USA), a multiple additive regression trees (MART) based program. The same datasets used in PLS (70% calibration, 30% validation) were used for boosted regression tree with a maximum of 12 branches per node, to identify the higher order interactions. Log_{10}-transformed data was used in boosting mode to have a fair comparison with the PLS models. The Huber-M loss criterion (Huber, 1981), which encompasses the best properties of least absolute deviation and least square deviation, was utilized. Initially, the maximum number of trees to be grown was set to 200. The number of trees was increased (>200) manually in two conditions: a) up to a point when $RMSE_p$ value stopped decreasing and b) when the optimal number of trees was close enough to the maximum numbers of trees specified beforehand.

2.3.7. Principal Component Analysis

Principal component analysis was performed in 'R' (R Development Core Team, 2004) to determine the ability of VisNIR-DRS to distinguish contaminated versus non contaminated soils qualitatively. The first 15 principal components (PC) of field-moist intact first derivative spectra were used to produce a "Screeplot", which was used to choose the number of PCs in the following supervised classification. Fisher's Linear Discriminant Analysis (LDA) was used, assuming equal prior probability for each group. Additionally, pairwise scatterplots of the first 3 PCs were produced to generate the ideas on how contaminated and reference soils were separated from each other in the spectral space.

2.4. Results and Discussion

Forty six soil samples were analyzed for TPH and used as dependent variable for the PLS and BRT analyses. Calibration (n=32) and validation (n=14) datasets were selected randomly; however, both had similar means (2.62 and 2.66 log_{10} mg kg^{-1}) along with similar standard

deviations (0.72 and 0.58 $\log_{10}$ mg kg^{-1}), respectively. The similarity among the validation and calibration data indicated that validation models should not be skewed. Among other soil properties, soil salinity varied from 0 to 2.54 dS m^{-1}. The highest salinity was identified in the soils of coastal areas. Substantial variability was observed for soil pH (5.20 to 7.85), clay content (160 to 600 g kg^{-1}), organic matter (9.3 to 130.5 g kg^{-1}), and bulk mineral concentrations (% weight basis) from site to site (Table 2). The Sabine site samples had the highest salinity, which was further supported by the elemental extraction analysis (7,758 mg kg^{-1} Na, on average). Extractable element concentrations differed between sites, as expected. No significant relationship was identified between organic matter, clay, and TPH content (both F-test and randomization test p-values were 0.11 using 0.05 or 0.10 as significance level).

2.4.1. Partial Least Squares Regression Models

Using the PLS regression algorithm for VisNIR DRS analysis, calibration models were developed using reflectance, first, and second derivatives. The calibration quality was evaluated by calibration r^2. Despite the widespread use of the first derivative of reflectance spectra for VisNIR models to predict soil properties (Reeves et al., 1999; Brown et al., 2006; Waiser et al., 2007); reflectance and first-derivative-based calibration models of field moist intact scans performed similarly, while the second derivative model was unsatisfactory (calibration r^2 < 0.15) (Table 3). Although the field-moist intact first derivative model exhibited a slightly better $RMSE_{cv}$ than the field-moist intact reflectance model, the main advantage of the first derivative was fewer latent factors (5) as compared to the reflectance-based model (eight latent factors) to prevent over-fitting. Results indicated a continuous reduction of latent factors as the model changed from reflectance to first and second derivatives. This reduction of principal components

Table 2. Soil pH, quantitative mineral abundance (% weight basis), clay (g kg^{-1}), and organic matter (g kg^{-1}) of soils evaluated for petroleum contamination using visible and near infrared diffuse reflectance spectroscopy in Louisiana, USA.

	pH	Minerals					Clay	Org. matter
Site		Quartz	K-feldspar	Plagioclase	Anhydrite	Clay minerals		
		..... ……%.....					……g kg^{-1}……	
Alpine	7.66	87.7	4.3	6.0	-	1.7	224.9	9.3
Mississippi River 1	7.85	82.1	3.8	6.0	-	8.0	206.5	47.9
Mississippi River 2	7.20	73.0	-	-	-	6.0	160.0	43.0
Sabine	6.46	39.8	2.3	3.5	1.0	53.3	600.0	130.5
Sonat	5.20	97.8	0.7	0.1	-	1.3	229.7	20.7
Winn Dixie	7.01	72.3	3.6	6.5	-	17.4	335.2	126.6

Table 3. Calibration and validation statistics for partial least square regression models of soils evaluated for petroleum contamination using visible and near infrared diffuse reflectance spectroscopy in Louisiana, USA.

Model	Latent factors	Calibration r^2	$RMSE_{cv}$ † (log_{10} mg kg^{-1})	Validation r^2	$RMSE_p$‡ (log_{10} mg kg^{-1})	RPD§	Bias (log_{10} mg kg^{-1})
			Field-moist intact				
Reflectance	8	0.79	0.323	0.64	0.353	1.64	-0.101
First derivative	5	0.81	0.311	0.64	0.341	1.70	-0.054
Second derivative	-	Unsatisfactory ¶	Unsatisfactory	-	-	-	-
			Air-dried intact				
Reflectance	5	0.57	0.436	0.63	0.216	1.94	-0.07
First derivative	4	0.64	0.393	0.57	0.335	1.25	-0.20
Second derivative	-	Unsatisfactory	Unsatisfactory	-	-	-	-
			Air-dried ground				
Reflectance	5	0.75	0.346	0.48	0.429	1.35	-0.14
First derivative	5	0.81	0.303	0.42	0.547	1.06	0.15
Second derivative	4	0.79	0.312	-	-	-	-

†$RMSE_{cv}$: root mean square error of cross-validation.
‡ $RMSE_p$: root mean square error of prediction.
§ RPD: relative percent difference.
¶ Model performance was unsatisfactory based upon very low calibration r^2 (<0.15) .

(latent factors using PLS) could be due to the use of higher degree spectrally preprocessed (first and second derivatives) data to refrain from viewing geometry effects (Demetriades-Shah et al., 1990). Brown et al. (2006) reported the advantage of using the PLS regression to surmount the inherent dimensionality of spectral data. When all calibration models were compared, the field-moist intact and air-dried ground models outperformed the air-dried intact models. Prediction accuracies of the aforementioned calibration models were evaluated incorporating the separate validation sets where only the reflectance and first derivative were taken into consideration (Fig. 6). According to Chang et al. (2001), accuracy and stability of spectroscopic models should be based on their RPD statistics. Stable and accurate predictive models showed an RPD >2.0; fair models with potential for prediction improvement had an RPD value between 1.4 and 2.0; while models with RPD values <1.4 were categorized as poor predictive models. Therefore, the present study considered validation r^2 and RPD as the main criteria for comparing model performances, but other error statistics were provided, including the $RMSE_p$ and bias (Table 3). The field-moist intact first derivative model was superior to the more biased field-moist intact reflectance model, with a slightly higher RPD value of 1.70 though the validation r^2 value for both was 0.64. However, both in terms of validation r^2 and RPD, the air-dried intact reflectance and air-dried ground reflectance models always performed better than their first derivative models. The air-dried ground scan showed the largest prediction error (0.547 $\log_{10}$ mg kg^{-1}) with dispersion about the validation subset. Moreover, in both reflectance and first derivative models, the air-dried ground scan showed a very low RPD (1.34 and 1.06, respectively). It is worth noting that the prediction accuracy of the air-dried intact reflectance model (validation r^2=0.63) was comparable to the field-moist intact scan models (validation r^2=0.64 in both reflectance and first derivative). Additionally, the air-dried intact reflectance model showed the highest RPD (1.94).

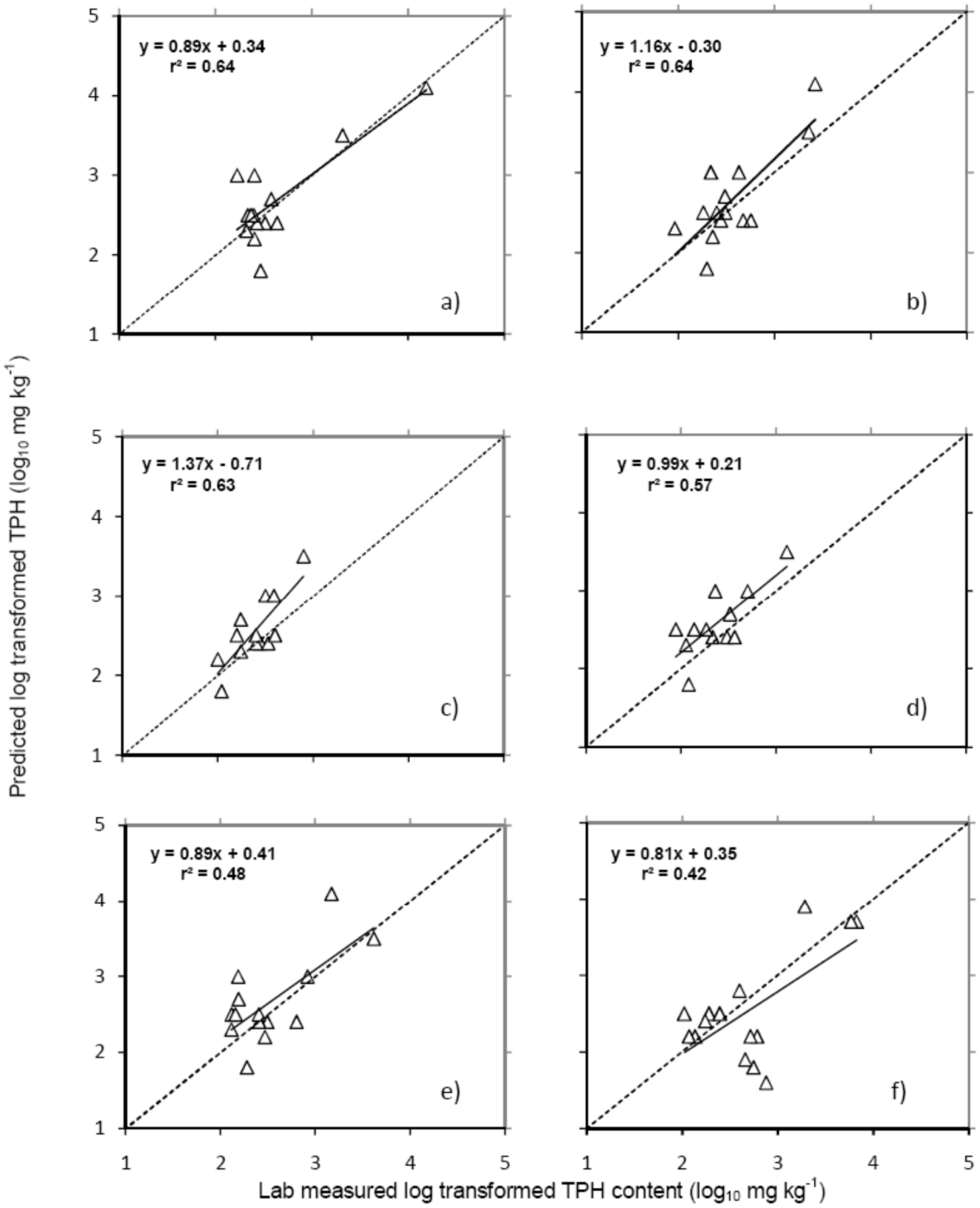

Fig. 6. Predicted vs. measured total petroleum hydrocarbon (TPH) content of the validation data set for a) field-moist intact reflectance, b) field-moist intact first derivative, c) air-dried intact reflectance, d) air-dried intact first derivative, e) air-dried ground reflectance, and f) air-dried ground first derivative models for soils from Louisiana, USA. The triangles represent validation samples. The dashed line and dark line represent the 1:1 line and the prediction trend, respectively.

2.4.2. Boosted Regression Tree Analysis

Model statistics for the BRT analysis, summarized in Table 4, showed much higher validation $RMSE_p$ as compared to the PLS models. Field-moist intact models included most optimal trees. Notably, the numbers of predictors increased as first derivative data were used. Perhaps the first derivative models used more predictors because of multiple interactions with linear and non-linear co-relations, as reported by Brown et al. (2006). In terms of validation r^2 and RPD, BRT did not perform as well as PLS. However, the field-moist intact first derivative model exhibited the highest predictability (RPD=1.49), which somewhat confirmed the PLS trend. Considering the calibration quality of field-moist intact scans, the first derivative calibration model generated a better calibration r^2 (0.85) than the reflectance model (0.38), whereas in the air-dried intact scan, the BRT reflectance model did not performed satisfactorily (calibration $r^2 < 0.15$). Improving BRT predictive performance by increasing the number of important predictors from reflectance to first derivative based models was consistent with prior knowledge of BRT model performance (Snelder et al., 2009). However, in case of field-moist intact scan, the reflectance model exhibited more optimum trees (507) as compared to the first derivative model (500). Given that tree based models require large data sets for robust model predictions (Vasques et al., 2009), the small dataset was most likely the crucial factor for BRT underperformance as compared to PLS models.

2.4.3. Underperformances of Air-dried Models

Results indicated that in both PLS and BRT, the first derivative model of the field-moist intact scan outperformed the air-dried intact and air-dried ground models, respectively (except in

Table 4. Calibration and validation statistics for boosted regression tree models of soils evaluated for petroleum contamination using visible and near infrared diffuse reflectance spectroscopy in Louisiana, USA.

Model	Important predictors	Calibration r^2	Validation r^2	$RMSE_p$† (log_{10} mg kg^{-1})	RPD‡	Optimal trees
			Field-moist intact			
Reflectance	11	0.38	0.42	0.420	1.38	507
First derivative	13	0.85	0.52	0.387	1.49	500
			Air-dried intact			
Reflectance	-	Unsatisfactory§	-	-	-	-
First derivative	7	0.68	0.45	0.589	0.98	179
			Air-dried ground			
Reflectance	5	0.41	0.39	0.437	1.32	75
First derivative	12	0.47	0.42	0.392	1.47	204

†$RMSE_p$: root mean square error of prediction.
‡RPD: relative percent difference.
§ Model performance was unsatisfactory based upon very low calibration r^2 (<0.15).

the air-dried intact reflectance PLS model). Soil reflectance is an integrated property which depends on various soil parameters like soil moisture, texture, and organic matter content (Morgan et al., 2009). The air-dried ground models were expected to perform better due to the removal of some water signals (due to air drying) that could mask the spectral signatures of other important predictors (soil properties). Additionally, smaller, more homogeneous particle sizes are known to produce higher absorption peaks because of more surface area for absorption.

However, to study the possible reasons for the air-dried model's underperformances, air-dried intact subsamples (10 samples) from the whole range of samples (46) were carefully selected so that each of them would represent a specific range of TPH. These subsamples were further analyzed for TPH in the same commercial lab using method 5520 D Soxhlet extraction and method 5520 F for quantification (Clesceri et al., 1998). Results confirm that TPH contents were significantly (sign test p-value: 0.001) lower in most of the subsamples, reanalyzed for TPH, as compared to the primary TPH contents as a result of drying (Fig. 7). Similar losses in TPH contents were reported as a result of varying degrees of weathering where volatilization, oxidation, reduction, and microbial metabolism were the prime factors (Whiteside, 1993; Malley et al., 1999).

The regression coefficients (black) of the first derivative PLS model of each scan and those that were significant (red, thick bar, $p < 0.05$) based on Tukey's jackknife variance estimate were plotted in Fig. 8. Notably, both the number and intensities of significant wavelengths changed in air-dried intact and air-dried ground scans as compared to field-moist intact scans. The change in numbers and intensities were apparent, specifically in the 1,600-1,850 and ~2,250-2,350 nm regions which could contain the 1,725 nm (two-stretch) and 2,298

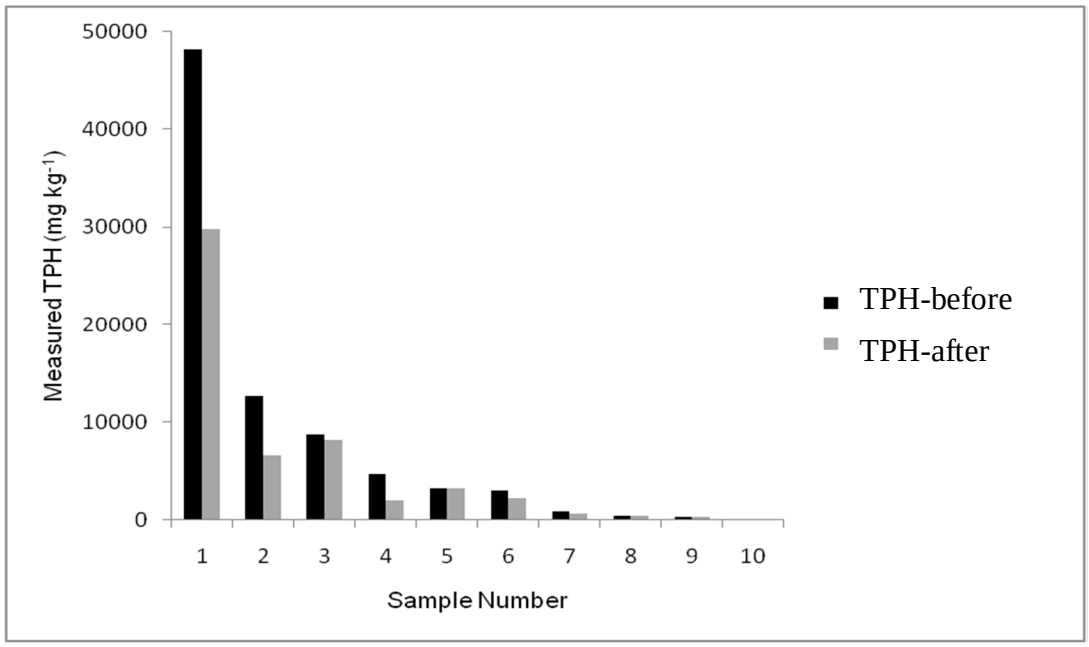

Fig. 7. Total petroleum hydrocarbon (TPH) contents (mg kg^{-1}) of 10 selected subsamples for soils from Louisiana. The black bars and gray bars represent TPH contents of subsamples before and after air drying, respectively.

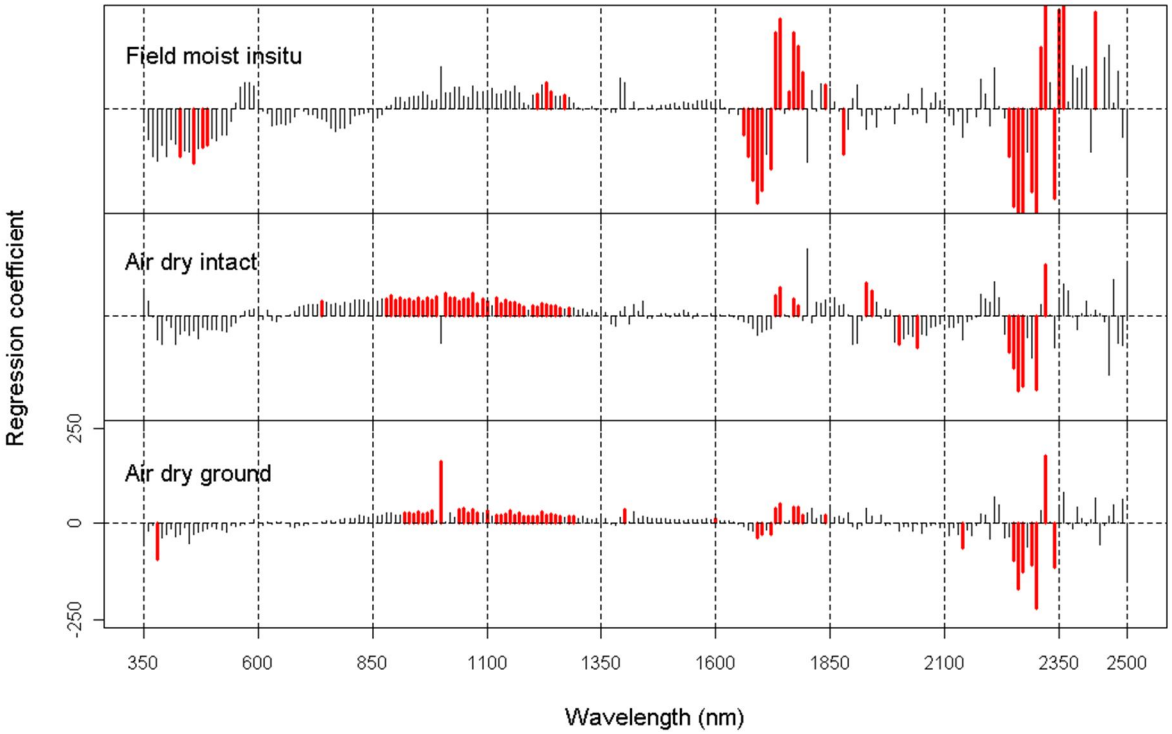

Fig. 8. Regression coefficients (black) of the first-derivative partial least squares model of each visible and near-infrared diffuse reflectance spectroscopy scan of contaminated soils from Louisiana. The magnitude of the regression coefficient at each wavelength is proportional to the height of the bar. Significant wavebands ($p < 0.05$) as indicated by Tukey's jackknife variance estimate procedure are shown as thick, red bars. All plots are on the same x axis. Values of all the y axes are not shown, but all y axes are on the same scale.

nm (stretch+bend) crude oil spectral signatures as reported by Mullins et al. (1992). Moreover, typical 1,450 and 1,940 nm spectral signatures for water were not highly significant. This trend was somewhat consistent with previous VisNIR DRS work by Waiser et al. (2007), where the amount of water in the soil samples did not alter the predication accuracy of the validation models.

Soil moisture loss due to air drying might have some effects on decreasing predictability (decreasing validation r^2 and RPD) in the first derivative models of field-moist intact to air-dried ground scans, but random loss of TPH in the air dried samples was likely the principal contributor for poor performance of the air-dried intact and air-dried ground models. Notably, due to random loss of TPH upon air drying, the air-dried model statistics weakened and the field-moist intact first derivative PLS model was selected as the best among all the models investigated.

While the RPD was not as high as that obtained for other constituents of soils (Malley, 1998), the results were encouraging, considering the effects of weathering processes on petroleum hydrocarbon. Moreover, it should be remembered that TPH does not have a fixed composition and is a term used to express a large family of several hundred chemical compounds originating from crude oil. Nonetheless, the corresponding RPD of 1.70 also indicated that there is sufficient scope for model enhancement (Chang et al., 2001). Malley et al. (1999) reported comparable statistics for near-infrared TPH predictions (validation r^2 of 0.68 and 0.72).

2.4.4. Principal Component Analysis

The "Screeplot" of the first 15 PCs of field-moist intact first derivative spectra was plotted in Fig 9. The first PC accounted for 61% of the variance, whereas the second and third

PCs accounted for 16% and 11% of the variance, respectively. Thus, the first 3 PCs accounted for 88% of the total variance. It was obvious that the selection of 3 or 6 PC scores for LDA were appropriate considering their percent variance in each PC. Pair-wise principal component plots (i.e. PC1 vs PC2, PC2 vs PC3, and PC1 vs PC3) of the field-moist intact first derivative spectra were plotted in Fig. 10. The circles and squares represent contaminated and non-contaminated samples, respectively.

The supervised classification results of contaminated versus non-contaminated soils using the Fisher's LDA method were presented in Table 5. The first 3 and 6 PC scores of the field-moist intact first derivative spectra were used as the explanatory variable. The classification results were quite promising. Using 3 PCs, the overall classification accuracy was 76% (35 out of 46 were correct); and when 6 PCs were used, the overall accuracy was 91%.

Thus, PCA results indicated that the soil spectra were highly correlated and a 3-d representation could capture the intrinsic data structure fairly well. Contaminated and un-contaminated samples could be reasonably separated by the first 3 PCs or first 6 PCs, which was an indication that the spectral method might be useful for distinguishing contamination qualitatively.

2.5. Conclusion

The present feasibility study with varying degrees of TPH contamination indicated that petroleum hydrocarbon could be predicted from the soil spectra in the visible-near-infrared range without any prior sample preparation. Among all models investigated, TPH was estimated by the field-moist intact first derivative PLS model with greatest accuracy. In validation mode, this model explained 64% of the variability of the validation set. Nevertheless, random loss of TPH

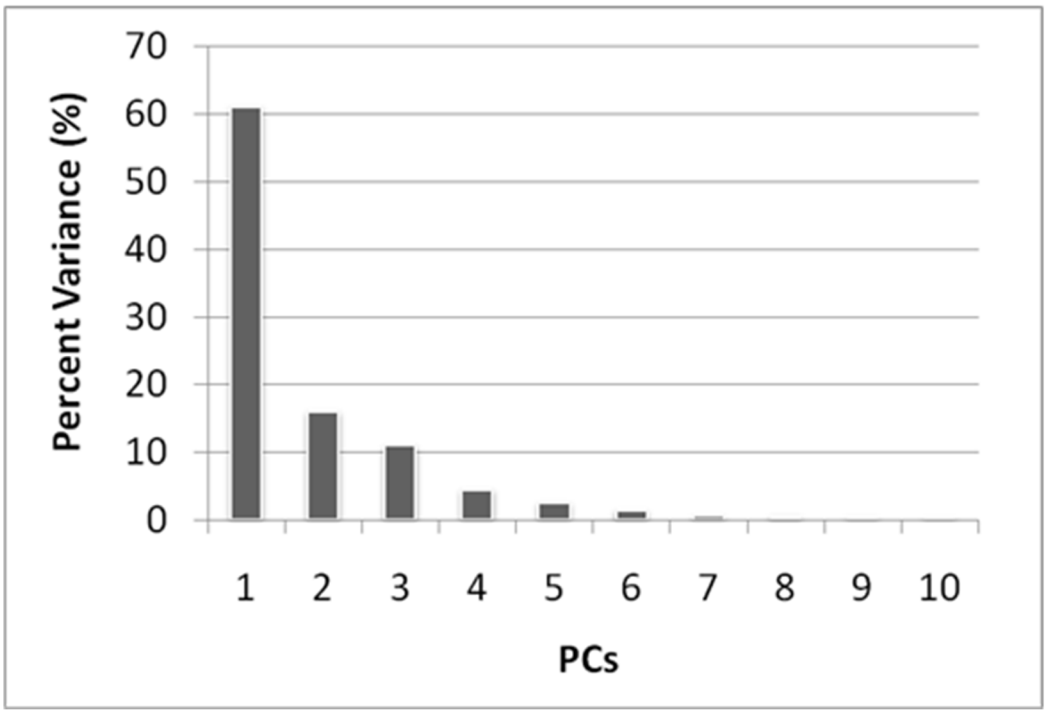

Fig. 9. “Screeplot” of the first 15 principal components (PCs) of field-moist intact first derivative spectra of soils evaluated for petroleum contamination using visible and near-infrared diffuse reflectance spectroscopy from Louisiana.

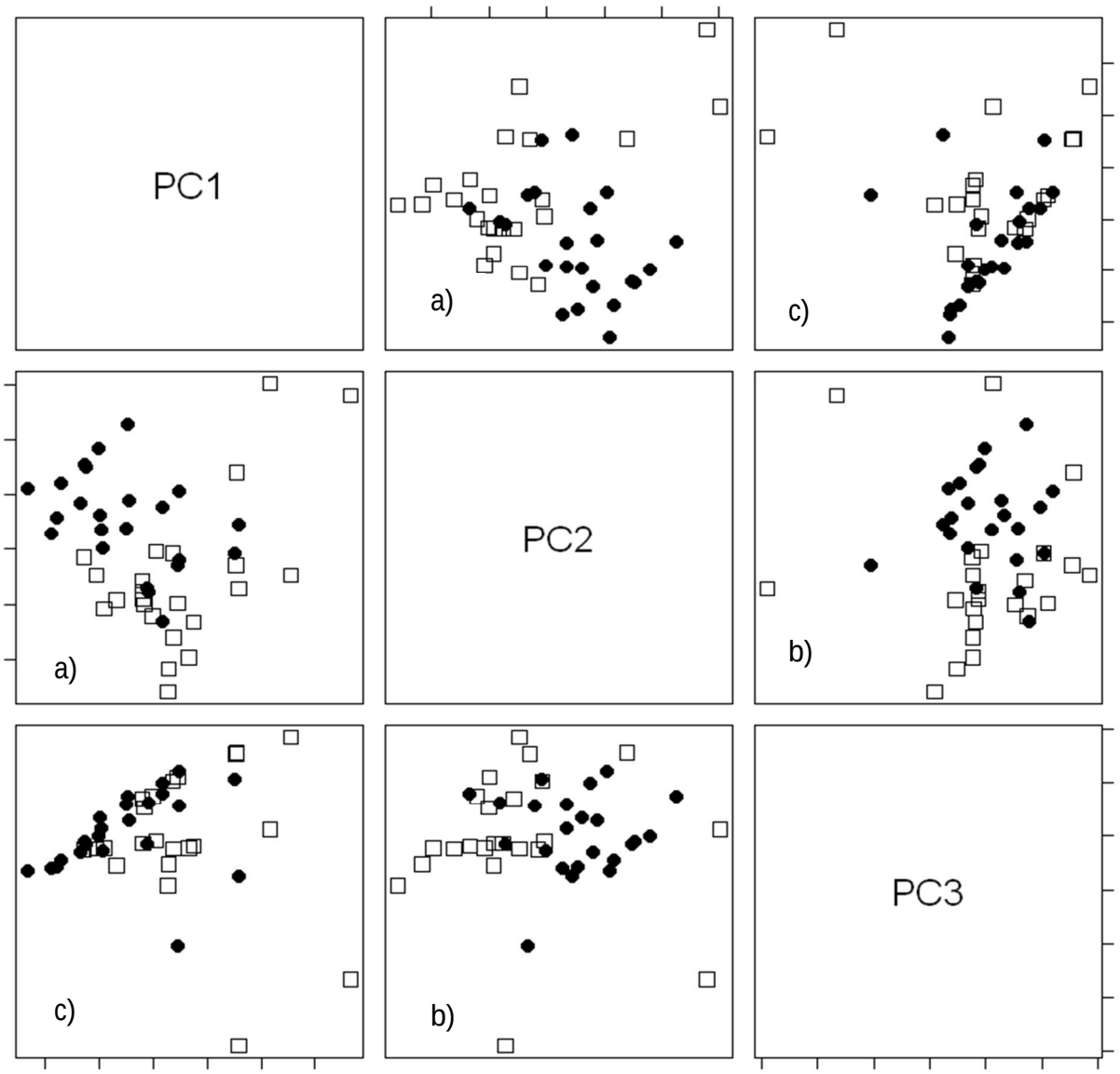

Fig. 10. Pairwise principle component (PC) plots for (a) PC1 vs. PC2, (b) PC2 vs. PC3, and (c) PC1 vs. PC3 of field-moist intact first-derivative spectra of soils evaluated for petroleum contamination using visible and near-infrared diffuse reflectance spectroscopy from Louisiana, USA. The circles and squares represent contaminated and noncontaminated samples, respectively.

Table 5. Classification result of contaminated versus non-contaminated soils using the Fisher's Linear Discriminant Analysis method for soils from Louisiana, USA. The first three and six principal component (PC) scores of the field-moist intact first derivative spectra were used as the explanatory variable.

		Using first three PCs			Using first six PCs		
		To group			To group		
		Contaminated	Non-contaminated	Sum	Contaminated	Non-contaminated	Sum
From group	Contaminated	16	7	23	20	3	23
	Non-contaminated	4	19	23	1	22	23
Sum		20	26	46	21	25	46
Overall accuracy				76%			91%

due to air drying was a major constraint responsible for the poor predictive abilities of air-dried models. Furthermore, the use of a small sample set in the BRT failed to produce robust models with good generalization capacity. It is noteworthy that no significant effect of variable water contents was observed.

A fair RPD value (1.70) for field-moist intact first derivative PLS model identified the scope for model improvement. In particular, continued research is recommended with a larger sample set along with other approaches like wavelet analysis, random forest, support vector, and spatial variability analysis evaluating VisNIR DRS prediction efficacy on a larger diversity of soils and a wider assortment of soil properties.

Summarily, provided that soil petroleum contamination is costly and time consuming to estimate, the prospect of using VisNIR DRS as a proximal soil sensor of petroleum contamination appears promising. If specific wavelengths related to various hydrocarbon signatures can be more precisely defined, remote sensing of hydrocarbon contamination plumes may be possible from airborne or satellite platforms.

2.6. References

Analytical Spectral Devices. 2007. Application note 1016.01E. Available at http://www.safeco.ir/en/documents/4.pdf (verified 5 April, 2010). Analytical Spectral Devices, Boulder, CO.

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

Bofetta, P., N. Jourenkova, and P. Gustavson. 1997. Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons. Cancer Causes and Control 8(3):444-472.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. J. Royal Statistical Soc. Series B. 26:211-252.

Ben-Dor, E., J.R. Irons, and G.F. Epema. 1999. Soil reflectance. p. 111–188. *In* N. Rencz (ed.) Remote sensing for the earth sciences: Manual of remote sensing. Vol. 3. John Wiley & Sons, New York.

Brown, D.J., R.S. Bricklemeyer, and P.R. Miller. 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VisNIR soil C prediction in Montana. Geoderma 129:251–267.

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

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

Chang, C.W., D.A. Laird, and C.R. Hurburgh, Jr. 2005. Influence of soil moisture on near-infrared reflectance spectroscopic measurement of soil properties. Soil Sci. 70:244–255.

Chung, H., H. Choi, and M. Ku. 1999. Rapid identification of petroleum products by near-infrared spectroscopy. Bull. Korean Chem. Soc. 20:1021-1025.

Clesceri, L.S., A.E. Greenberg, and A.D. Eaton. (ed.) 1998. Standard methods for the examination of water and wastewater. 20th ed. American Public Health Association, American Water Work Association, and Water Environment Federation, Washington, DC.

Cook, H.E., P.D. Johnson, J.C. Matti, and I. Zemmels. 1975. Methods of sample preparation and x-ray diffraction data analysis, x-ray mineralogy laboratory, Deep Sea Drilling Project, University of California, Riverside, *In* D.E. Hayes, L.A. Frakes, et al., Initial reports of Deep Sea Drilling Project. 28:999-1007.

Dalal, R.C., and R.J. Henry. 1986. Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance. Soil Sci. Soc. Am. J. 50:120–123.

Demetriades-Shah, T.H., M.D. Steven, and J.A. Clark. 1990. High-resolution derivative spectra in remote sensing. Remote Sens. Environ. 33(1):55-64.

Dorbon, M., J.P. Durand, and Y. Boscher. 1990. On-line octane-number analyser for reforming unit effluents. Principle of the analyser and test of a prototype. Anal. Chim. Acta. 238:149-160.

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat. 29:1189–1232.

Gauch, H.G., J.T.G. Hwang, and G.W. Fick. 2003. Model evaluation by comparison of model-based predictions and measured values. Agron. J. 95:1442–1446.

Ge, Y., C.L.S. Morgan, J.A. Thomasson, and T. Waiser. 2007. A new perspective to near infrared reflectance spectroscopy: A wavelet approach. Trans. ASABE. 50:303 - 311.

Gee, G.W., and J.W. Bauder. 1986. Particle size analysis. *In* A. Klute (ed.) Methods of soil analysis, Part 1. Physical and mineralogical methods 2nd ed. ASA and SSSA, Madison, WI.

Henderson, T.L., M.F. Baumgardner, D.P. Franzmeirer, D.E. Stott, and D.C. Coster. 1992. High dimensional reflectance analysis of soil organic matter. Soil Sci. Soc. Am. J. 56:865–872.

Huber,P.J. 1981. Robust statistics. Wiley, New York.

Hutcheson, M.S., D. Pedersen, N.D. Anastas, J. Fitzgerald, and D. Silveman. 1996. Beyond TPH: health based evaluation of petroleum hydrocarbon exposures. Regulatory Toxicology and Pharmacology 24(1):85-101.

Kelly, J.J., C.H. Barlow, T.M. Jinguji, and J.B. Callis. 1989. Prediction of Gasoline Octane Numbers from Near-Infrared Spectral Features in the Range 660-1215 nm. Anal. Chem. 61:313-320.

Krishnan, R., J.D. Alexander, B.J. Butler, and J.W. Hummel. 1980. Reflectance technique for predicting soil organic matter. Soil Sci. Soc. Am. J. 44:1282–1285.

Lee, J.S., and H. Chung. 1998. Rapid and nondestructive analysis of the ethylene content of propylene/ethylene copolymer by near-infrared spectroscopy. Vib. Spectrosc. 17:193-201.

Lohmannsroben, H., and L. Schober. 1999. Combination of laser-induced fluorescence and diffuse-reflectance spectroscopy for the in situ analysis of diesel-fuel-contaminated soils. Applied Optics 38:1404-1410.

Madari, B.E., J.B. Reeves, P.L.O.A. Machado, C.M. Guimaraes, E. Torres, and G. McCarty. 2006. Mid-and near-infrared spectroscopic assessment of soil compositional parameters and structural indices in two Ferralsols. Geoderma 136:245-259.

Malley, D.F. 1998. Near-infrared spectroscopy as a potential method for routine sediment analysis to improve rapidity and efficiency. Water Sci. Technol. 37:181–188.

Malley, D.F., K.N. Hunter, G. R. Barrie Webster. 1999. Analysis of diesel fuel contamination in soils by near-infrared reflectance spectrometry and solid phase microextraction-gas chromatography. Soil Sediment Contam. 8(4):481–489.

Malley, D.F., P.D. Martin, L.M. McClintock, L. Yesmin, R.G. Eliers, and P. Haluschak. 2000. Feasibility of analyzing archived Canadian prairie agricultural soils by near infrared reflectance spectroscopy. p. 579–585. *In* A.M.C. Davies and R. Giangiacomo (ed.) Near infrared spectroscopy: Proceedings of the 9th international conference. NIR Publications, Chichester, UK.

Morgan, C.L.S., T.H. Waiser, D.J. Brown, and C.T. Hallmark. 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. Geoderma 151:249-256.

Mehlich, A. 1984. Mehlich 3 soil extractant: A modification of Mehlich 2 extractant. Commun. Soil Sci. Plant Anal. 15:1409-1416.

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

Nelson, D.W., and L.E. Sommers. 1996. Total carbon, organic carbon and organic matter. *In* D.L. Sparks (ed.) Methods of soil analysis. Part 3. Chemical methods. ASA and SSSA, Madison, WI.

R Development Core Team. 2004. The R project for statistical computing . Available at www.r-project.org (verified 26 Nov. 2006). R Foundation for Statistical Compuing, Vienna.

Reeves, J.B., III, G.W. McCarty, and J.J. Meisinger. 1999. Near infrared reflectance spectroscopy for the analysis of agricultural soils. J. Near Infrared Spectrosc.7:179–193.

Schwartz. G, G. Eshel, M. Ben-Haim, and E. Ben-Dor. 2009. Reflectance spectroscopy as a rapid tool for qualitative mapping and classification of hydrocarbons soil contamination. Available at http://www.earsel6th.tau.ac.il/~earsel6/. /3080%20Schwartz.pdf (verified 19 Apr. 2010).

Shepherd, K.D., and M.G. Walsh. 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66:988–998.

Snelder, T.H., N. Lamouroux, J.R. Leathwick, H. Pella, E. Sauquet, and U. Shankar. 2009. Predictive mapping of the natural flow regimes of France. J. Hydrol. 373:57-67.

Soil Survey Staff. 2004. Soil survey laboratory methods manual (version 4.0). USDA-NRCS. US Gov. Print. Off. Washington, DC.

Soil Survey Staff. 2005. Official soil series descriptions. Available at soils.usda.gov/technical/classification/osd/index.html (verified 26 Nov. 2006). NRCS, Washington, DC.

Soltanpour, P.N., G.W. Johnson, S.M. Workman, J.B. Jones, and R.O. Miller. 1996. Inductively coupled plasma emission spectrometry and inductively coupled plasma-mass

spectrometry. *In* D.L. Sparks (ed.) Methods of soil analysis. Part 3. Chemical methods. SSSA, Madison, WI.

Stallard, B.R., M.J. Garcia, and S. Kaushik. 1996. Near-IR Reflectance spectroscopy for the determination of motor oil contamination in sandy loam. Appl. Spect. 50:334-338.

Steinberg, D., M. Golovnya, and D. Tolliver. 2002. TreeNet 2.0 user guide. Salford Syst., San Diego, CA.

Stoner, E. R., and M.F. Baumgardner. 1981. Characteristic variations in reflectance on surface soils. Soil Sci. Soc. Am. J. 45:1161-1165.

Sudduth, K.A., and J.W. Hummel. 1993. Soil organic matter, CEC and moisture sensing with a portable NIR spectrophotometer. Trans. ASAE 36:1571-1582.

Thomasson, J.A., R. Sui, M.S. Cox, and A. Al-Rajehy. 2001. Soil reflectance sensing for determining soil properties in precision agriculture. Trans. ASAE 44:1445-1453.

Vasques, G.M., S. Grunwald, and J.O. Sickman. 2009. Modeling of soil organic carbon fractions using visible-near-infrared spectroscopy. Soil Sci. Soc. Am. J. 73:176-184.

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

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

Waiser, T.H., C.L.S. Morgan, D.J. Brown, and C.T. Hallmark. 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Sci. Soc. Am. J. 71:389-396.

Wang, Z., and M. Fingas. 1997. Developments in the analysis of petroleum hydrocarbons in oils, petroleum products and oil-spill-related environmental samples by gas chromatography. J. Chromatogr. A.774:51-78.

Whiteside, S. E. 1993. Biodegradation studies of Saudi Arabian crude oil. p. 281-287. *In* Abstracts, Annual Technical Conference and Exhibition of the Society of Petroleum Engineers, Houston, TX. 3-6 Oct. 1993.

Whittig, L.D. and W.R. Allardice. 1986. X-ray diffraction techniques. *In* A. Klute (ed.) Methods of soil analysis. Part 1. Physical and mineralogical methods 2nd ed. ASA and SSSA, Madison, WI.

CHAPTER 3

ASSESSING SPATIAL VARIABILITY OF SOIL PETROLEUM CONTAMINATION USING VisNIR DRS

3.1. Synopsis

This study evaluated whether a combination of two methods, penalized spline regression and geostatistics could provide an efficient approach to assess spatial variability of soil TPH from visible near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS), using soil samples from an 80-ha crude oil spill location in central Louisiana. Initially, a penalized spline model was calibrated to predict TPH contamination in soils by combining lab TPH values of 46 contaminated and uncontaminated soil samples and the first-derivative of VisNIR reflectance spectra of these samples. The r^2, RMSE, and bias of the calibrated penalized spline model were 0.81, 0.289 $\log_{10}$ mg kg^{-1}, and 0.010 $\log_{10}$ mg kg^{-1}, respectively. Subsequently, the penalized spline model was used to predict soil TPH content for 128 soil samples collected over the 80-ha study site. When assessed with a randomly chosen validation subset (10 samples) from the 128 samples, the penalized spline model performed satisfactorily (r^2=0.70; residual prediction deviation=2.0). That same validation subset was used to assess ordinary block kriging interpolation after the remaining 118 predictions were used to produce an experimental semivariogram and map. The experimental semivariogram was fitted with an exponential model which revealed strong spatial dependence among soil TPH [r^2 = 0.76, nugget = 0.001 ($\log_{10}$ mg kg^{-1})^2, and sill 1.044 ($\log_{10}$ mg kg^{-1})^2]. In the kriged map, TPH distribution matched with the expected TPH variability of the study site. Since the combined use of VisNIR prediction and geostatistics was promising to identify the spatial patterns of TPH contamination in soils, future

research is warranted to evaluate the approach for mapping spatial variability of petroleum contaminated soils.

3.2. Introduction

Petroleum contamination of soil is a widespread problem that occurs frequently with adverse environmental and human health consequences (MacEwan and Vernot, 1985; Hutcheson et al., 1996; Boffetta et al., 1997). Accidental release of crude oil and refined oil products from oil drilling rigs (such as the BP Deepwater Horizon), automobiles, immense oil tanker accidents (such as Exxon Valdez, Erika, and Prestige), and pipeline and storage tank leakages, endanger local and regional ecological systems (Calabrese and Kostecki, 1988). The extent of environmental contamination by petroleum spillage depends on the capability of soil to filter, retain, biodegrade, and release petroleum (Fine et al., 1997). Besides, vapor pressure and solubility of crude oil and other fractions (*n, iso*, and cycloparaffins, napthene, and aromatics) also influence the dynamics of petroleum distribution in soil. Remediation specialists are constantly challenged by the need to measure spatial variation of total petroleum hydrocarbons (TPH) within and across a spillage area for site specific remediation practices (Cole, 1994). Soil TPH contamination maps are generated using a large number of soil samples and traditional soil chemical analyses. However, such soil analyses are laborious, costly, time consuming, and inadequate when high spatial and temporal resolution of TPH content are warranted (Odlare et al., 2005). Consequently, there is a persistent need for the development of innovative, low-cost, and reproducible analytical package for mapping spatial variability of petroleum contaminated soils. Hyperspectral visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) can be used for detecting and mapping inland oil spills. This proximal optical sensor has already demonstrated its potential as a viable alternative to the laborious field sampling and expensive

lab analysis for on-site quantification of TPH. Chakraborty et al. (2010) reported the feasibility of VisNIR DRS for predicting soil TPH with a validation r^2 of 0.64 and relative percent difference (RPD) of 1.70. Forrester et al. (2010) used a partial least squares (PLS) cross-validation approach for infrared spectroscopic identification of TPH with an r^2 of 0.81 and RMSE of 4653 mg kg^{-1}. Moreover, Chakraborty et al. (unpublished data, 2011) evaluated three types of classification analysis (linear discriminant analysis, support vector machines, and random forest) and three multivariate regression methods (stepwise multiple linear regression, MLR; partial least squares regression, PLSR; and penalized spline) for pattern recognition and developing TPH predictive models. The approach provides multiple benefits over traditional sampling/labwork: 1) results are returned to the investigator, on-site instantly, 2) the process is non-destructive allowing for sample preservation for future comparisons, and 3) minimization or elimination of traditional laboratory analyses saves considerable money over long periods of deployment.

While VisNIR DRS-TPH predictions have already shown potentiality for future use, spatial dependence among soil TPH contents has not received much consideration yet. One of the key presumptions in VisNIR DRS modeling is that the targeted component (soil TPH, in our case) should be independently distributed from each other to guarantee an optimal prediction model. Nevertheless, in case of high spatial resolution of soil sampling, the spatial autocorrelation among soil TPH contents in a geographic space is likely. Ignoring the issue of spatial dependency can render sub-optimal VisNIR-TPH predictive models. However, a successful combination of geostatistics with VisNIR DRS could identify spatial correlation among soil TPH contents, faster than traditional soil physicochemical analysis. In precision agriculture, several researchers have proposed the combination of VisNIR DRS and multivariate

geostatistics for improved spatial prediction of soil properties (Bilgili et al., 2010; Ge et al., 2007). Hengl et al. (2004) reported a basic framework for spatial variability mapping of soil properties based on hybrid regression-kriging. Geostatistics is a combination of variography and kriging. While variography aims to produce semivariograms for modeling spatial variance of data, kriging uses the modeled variance to estimate interpolated values between samples (Odlare et al., 2005).

Because of rapidity in prediction and incorporation of geostatistics, VisNIR spectroscopy could greatly enhance the spatial variability mapping of soil petroleum contamination. We combined two techniques: penalized spline regression and geostatistics. A penalized spline model to predict TPH contamination in soils was created and used to predict soil TPH content over a particular spill location. Predictions from the model were used to produce an experimental semivariogram and ordinary block kriging map. The objective of this study was to investigate whether the combination of VisNIR spectroscopy and ordinary block kriging has the potential to identify the spatial distribution of TPH contamination and test the accuracy of that map with the measured locations at the study site.

3.3. Materials and Methods

3.3.1. Study Area and Soil Sampling

The field chosen for mapping was a crude oil well blowout site located in Kisatchie National Forest in Vernon Parish, central Louisiana (30˚ 59'23" N, 93˚ 1' 48" W) (Fig. 11). One hundred and twenty-eight surface (0–15 cm) soil samples were collected within an 80 ha area that is densely vegetated by trees, shrubs, and grasses. The soils and sampling locations at the site are represented by four soil series: Caddo silt loam (fine-silty, siliceous, active, thermic

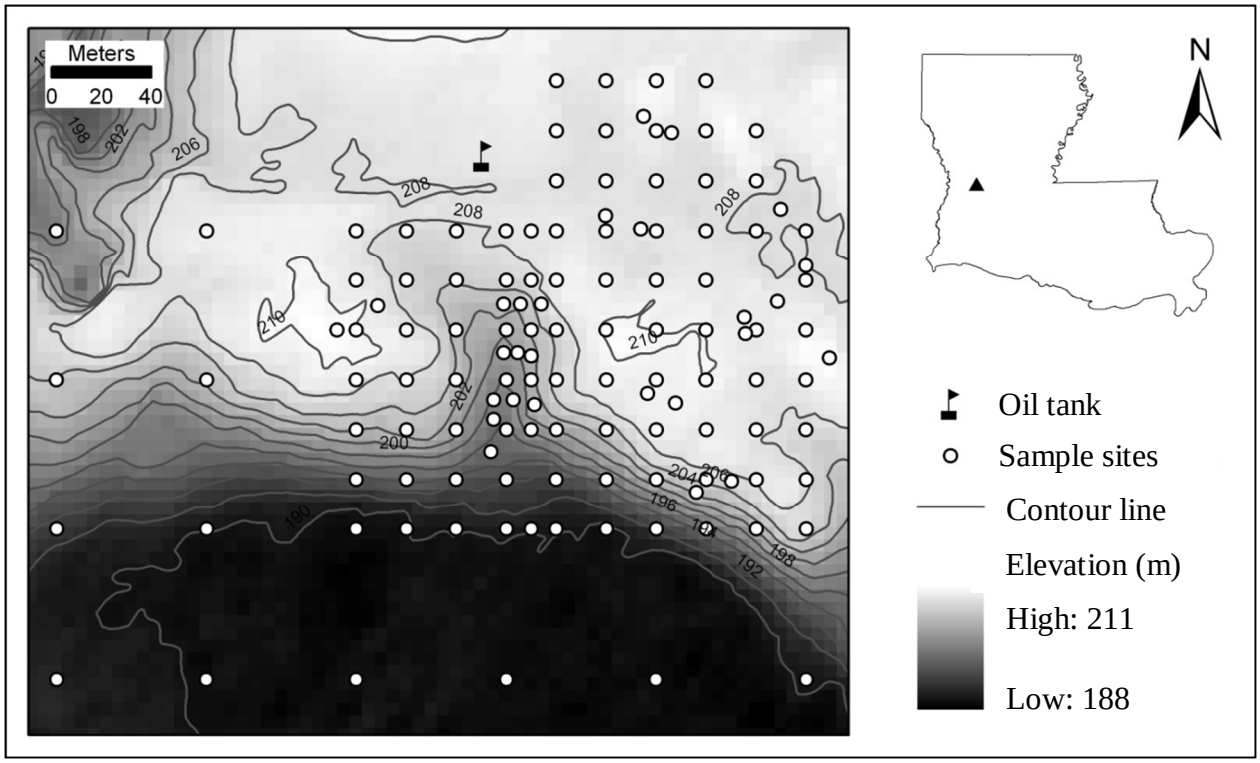

Fig. 11. The location, field boundary of the study site, and locations of collected soil samples in Louisiana, USA.

Typic Glossaqualf), Guyton silt loam (fine-silty, siliceous, active, thermic Typic Glossaqualf), Malbis fine sandy loam (fine-loamy, siliceous, subactive, thermic Plinthic Paleudult), and Ruston fine sandy loam (fine-loamy, siliceous, semiactive, thermic Typic Paleudult) (Soil Survey Staff, 2009). A sampling scheme was designed in ArcGIS 9.3 (Environmental Systems Research Institute, 2008) by combining both grid and random sampling. Sampling points were uploaded into a hand-held GPS receiver and geo-located in the field for sampling (location error approx. ± six meters). All soil samples were sealed in air-tight plastic bags to prevent hydrocarbon volatilization and preserve field-moisture status before scanning.

3.3.2. Calibration Data Set for VisNIR Prediction

Forty-six soil samples (including both contaminated and uncontaminated samples) were collected from six sites, each located in a different parish within southern and central Louisiana. The sampling scheme was developed with the prior knowledge of oil spill locations supplied by the Louisiana Oil Spill Coordinators Office (LOSCO) to guarantee maximum TPH variability within the soil samples collected. The original TPH contents of the samples were widely and non-normally distributed from 44.3 to 48,188 mg $kg^{-1}$of soil. Crude oil was the source of TPH. Soil texture mainly varied from clay to sandy loam. Twelve samples came from the Ruston fine sandy loam series. More detailed information on soil sample collection, soil sample preparation, and the official soil series description for sampling sites was reported by Chakraborty et al. (2010).

3.3.3. VisNIR Spectroscopy and Laboratory Analyses

An AgriSpec VisNIR portable spectroradiometer (Analytical Spectral Devices, Boulder, CO) with a spectral range of 350 to 2500 nm (2-nm sampling resolution and a spectral resolution of 3- and 10-nm wavelengths from 350 to 1000 nm and 1000 to 2500 nm, respectively) was used

to scan field–moist soil samples (both calibration samples and study site samples) with a contact probe. The contact probe had a circular viewing area (20-mm diameter) and its own halogen light source. Each sample was scanned four times with a 90° rotation between scans to obtain an average spectral curve. A spectralon panel with 99% reflectance was used every four samples to optimize and white reference the spectroradiometer to offset dark current and temperature effects. The spectroscopic reflectance measurement for each soil sample was then obtained by averaging the four raw scans.

A statistical analysis software package, R version 2.11.0 (R Development Core Team, 2008) was used to preprocess raw reflectance spectra. Based on a comparative analysis by Chakraborty et al. (2010), only the reflectance and the first-derivative of reflectance spectra on 10-nm intervals were extracted using custom 'R' routines (Brown et al., 2006). Spectroscopic reflectance splines and first derivative spectra of three soil samples are presented in Fig. 12.

A validation subset with 10 samples was randomly selected from the 128 study site samples. In a commercial laboratory, TPH was measured gravimetrically for both 46 calibration samples and 10 validation subset samples, following the method of Clesceri et al. (1998). Detailed procedures for TPH and other lab analyses were reported by Chakraborty et al. (2010).

3.3.4. Penalized Spline Model

The penalized spline calibration model was developed using the first-derivative of the reflectance spectra of the 46 contaminated and uncontaminated soils collect in six Louisiana parishes. Penalized spline is more stable and flexible than parametric PCR and PLSR as the shape of the functional relationship amongst covariates and the dependent variable (TPH, in this study) is governed by the data (Marx and Eilers, 1999; Crainiceanu et al., 2005). Penalized spline attempts to take advantage of the additional structure from the order of regressors. Namely, it

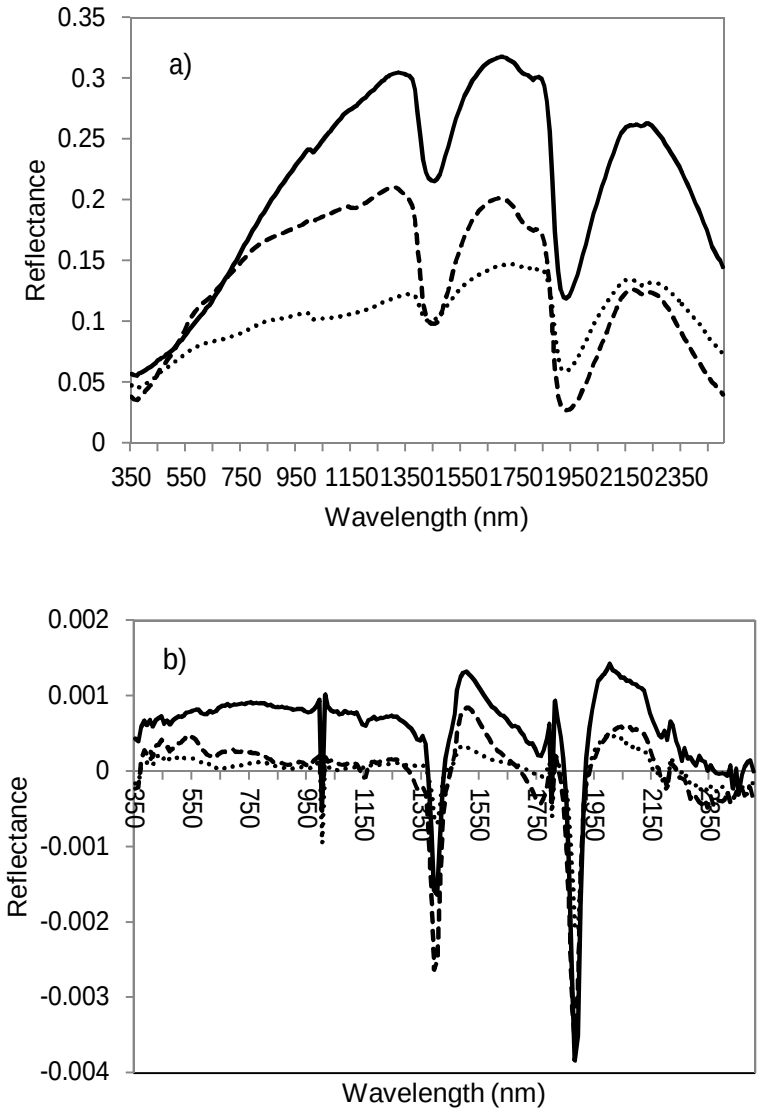

Fig. 12. a)Average reflectance spectra and b) first-derivative spectra for three randomly selected soil samples from Louisiana, USA.

forces the regression coefficients to be smooth (i.e. constraining the difference between the neighboring regression coefficients). The smoothness comes from a difference penalty on adjacent regression coefficients. This penalty is proportional to the size of the difference between neighborhood coefficients. Because of the additional constraint imposed by the difference penalty, penalized spline is well-suited for ill-posed problems (the dimensionality is much larger than the sample size) such as signal regression problems.

Although the penalized spline model can handle both parametric and non-parametric data, transformation on the response variable is necessary because the TPH content of the samples was widely and non-normally distributed from 44.3 to 48188 mg kg^{-1} of soil. Therefore without transformation, the model results were highly affected by outliers. Vasques et al. (2009) also transformed on the response variable even though they had a large dataset for non-parametric models.

In the present study, the Box-Cox transformation (Box and Cox, 1964) was applied to the original TPH data and the original data (lambda = 1) was $\log_{10}$–transformed (lambda = 0) to bring the data to a more normal distribution (Fig. 13). Thus, penalized spline model was developed based on $\log_{10}$–transformed data that approximated a Gaussian distribution after stabilizing the variance. As such, the remaining penalized spline model and kriging interpolation reported here all show $\log_{10}$–transformed TPH.

The cubic B-spline was used (using R version 2.11.0.) as the base function with 100 equally-spaced knots. The order of the penalty was set to the default value of three. The optimal value for the penalty-tuning parameter was selected by minimizing the leave-one-out cross-validation error. This penalized spline model was used to predict the TPH contents of the study site samples (128). The validation subset (n=10) was utilized to validate both penalized spline

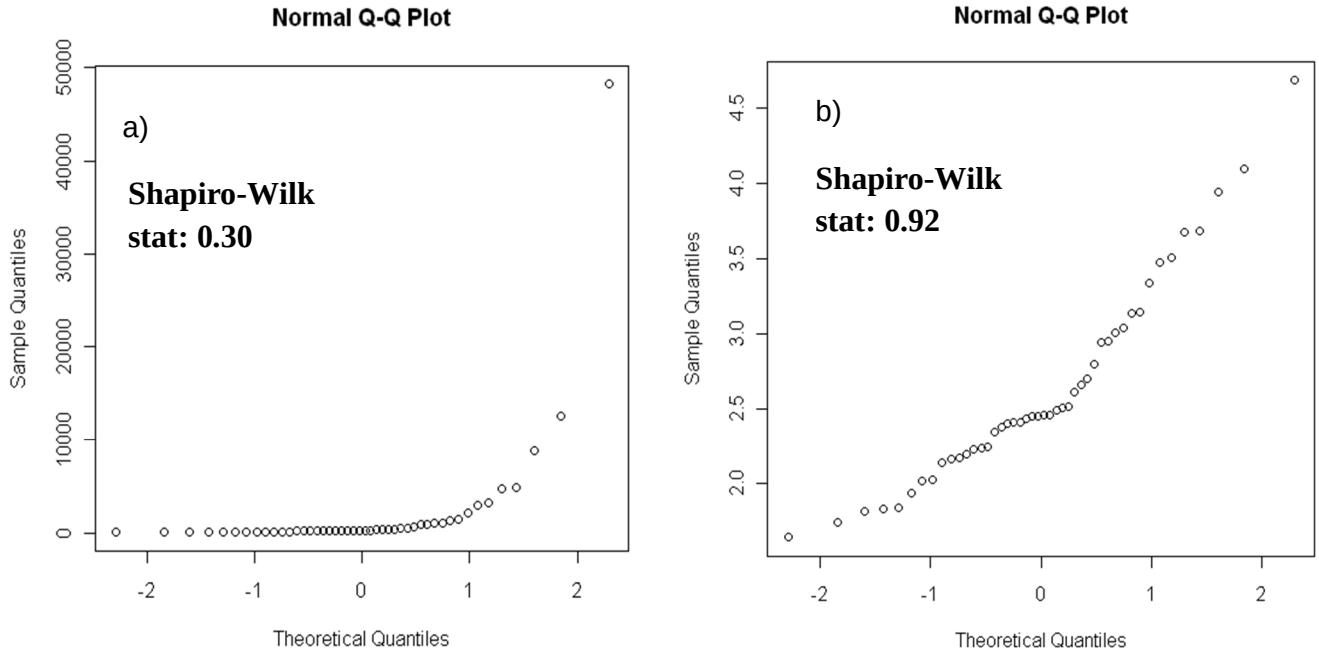

Fig. 13. Normal Q-Q plots of the a) Original (λ=1) and b) $\log_{10}$-transformed (λ=0) total petroleum hydrocarbon (TPH) contents of the soil samples collected from six different parishes in Louisiana, USA and used to calibrate the penalized spline model to predict TPH. Increase in Shapiro-wilk statistic for log-transformed data revealed that the Box-Cox transformation made the original TPH data more normal.

predictive model and kriging exercise. The rest 118 VisNIR-based penalized spline predictions were used for kriging.

3.3.5. Geostatistical Analyses and Kriging

A variogram and subsequently a kriged map of THP were created using the 118 VisNIR-based penalized spline predictions. According to the usual method of moments (Matheron, 1965; Webster and Oliver, 2001), the experimental semivariogram of soil TPH at the Kisatchie National Forest site was calculated using Eq. 1,

$$\gamma(h) = \frac{1}{2\times n(h)}\sum_{i=1}^{n(h)}\{z(s_i) - z(s_i + h)\}^2 \qquad [1]$$

where $\gamma(h)$ is the experimental semivariance at distance interval h; $n(h)$ is the number of observation pairs separated by the lag distance h (omnidirectional); and $z(s_i)$ and $z(s_i+h)$ denote soil TPH contents at spatial locations s_i and (s_i+h), respectively. Geostatistical package GS+ 9.0 (Gamma design Software, Plainwell, MI) was used to calculate the semivariogram of predicted TPH. A cross-validation approach was used to perform jackknife analysis in which every measured point in the data set was temporarily deleted from the data set and then estimated to provide an indication of the appropriateness of the semivariogram model. Interpolation by ordinary kriging was then conducted based on the parameters of the semivariogram (C_0 as the nugget, C_0+C as the sill, and A_0 as the range). The nugget was used to indicate the combination of measurement error and fine-scale spatial variability. The range indicates the distance in the field where TPH concentrations are no longer spatially correlated. The strength of the spatial structure was calculated in terms of sill: nugget ratio. The goodness of fit between experimental and theoretical semivariograms was measured in terms of r^2. Ordinary kriging is a geostatistical method which uses information from neighboring points to calculate a specific variable at a

target point. It relies on the spatial autocorrelation structure of the target variable to define the weighting values, and can be calculated using Eq. 2 (Cressie, 1990):

$$Z^*_{OK}(s_0) = \sum_{i=1}^{n} w_i \, z(s_i) \qquad [2]$$

where, $Z^*_{OK}(s_0)$ is the ordinary kriging estimate at an unvisited place(s_0), n is the number of samples in the search neighborhood, while w_i are the weights assigned to the i^{th} observation $z(s_i)$. In the present study, ordinary block kriging using 16 neighbors and lag class distance of 22.22 m for data interpolation was used to produce a TPH contour map. Ordinary block kriging is a straightforward extension of ordinary kriging which predicts the average value over a whole block. Note that the exponential model, as shown in Eq. 3, best fit the experimental semivariogram for TPH in this study.

$$\gamma(h) = C_0 + C\left[1 - e^{-\left(h/A_0\right)}\right] \quad \text{when } |h| > 0 \qquad [3]$$

The objective criterions for measuring the goodness of fit for ordinary block kriging were the coefficient of determination (r^2) and RMSE. Calculated RMSE was derived with Eq. 4 (Ge et al., 2007):

$$RMSE = \frac{1}{\bar{Z}_{li}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[Z_{pi} - Z_{li}\right]^2} \qquad [4]$$

where, Z_{pi} and Z_{li} were the predicted and laboratory-measured TPH values of i^{th} sample, $\bar{Z}_{li}$ was the mean of laboratory-measured TPH values, and N was the total number of sample in the validation subset (10 in the present study).Additionally, the kriged map was used to investigate if the VisNIR detected soil TPH variability could match the expected TPH variability of the study site, considering variable topography.

3.4. Results and Discussion

For the initial set of soil samples used for penalized spline model building, soil salinity (0-2.5 dS m^{-1}), soil pH (5.2-7.8), clay content (160-600 g kg^{-1}), organic matter (9.3-130.5 g kg^{-1}), and bulk mineral concentration (% weight basis) showed wide ranges of variation (see Chakraborty et al. 2010 Table 2 for details). Salinity was identified in the coastal soil samples. No association between clay, organic matter, and TPH was recognized (both *F*-test and randomization test *p*-values were 0.11 using 0.05 or 0.10 as significance level).

Calibration (n = 46) and validation (n = 10) datasets both had non-significantly (t test *p*-value= 0.08) different means (2.60 and 2.21 $\log_{10}$ mg kg^{-1}) as well as similar standard deviations (0.66 and 0.67 $\log_{10}$ mg kg^{-1}), respectively. The similarity among the validation and calibration data specified that validation model should not be skewed. The results of the TPH prediction model using penalized splines indicated that using VisNIR to predict TPH was reasonable. The calibration statistics, using full cross validation, resulted in a r^2 of 0.81, an RMSE of 0.289 $\log_{10}$ mg kg^{-1}, a bias of 0.010 $\log_{10}$ mg kg^{-1}, and an RPD of 1.77. This calibration was created without using any soils from the actual contaminated site; therefore, the true performance of the model was difficult to evaluate on the calibration statistics alone. In the soil VisNIR literature, it is well established that the accuracy of VisNIR-based prediction of a soil constituent is closely related to the likeness of the calibration set to that of the validation of test set that is to be predicted (Brown et al., 2005; Waiser et al., 2007; Morgan et al., 2009). This relationship is especially true when using intact, field-moist soil samples (Waiser et al., 2007). Furthermore, the prediction accuracy and stability of the penalized spline model (using the 10 field-specific validation subset) were evaluated according to the residual prediction deviation (RPD)-based guidelines by Chang et al. (2001). While the best prediction models are characterized by a RPD

of >2.0, fair models with potential for prediction improvement include RPD values of 1.4-2.0, while unreliable models have RPD values of <1.40. The prediction of soil TPH was reasonably satisfactory with a validation r^2 of 0.70 and an RPD of 2.0. The results were encouraging, considering the chemical complexity and decomposability of crude oil. Other validation statistics, such as the RMSE and bias were 0.409 and 0.235 $\log_{10}$ mg kg^{-1}, respectively. The left panel of Fig.14 shows the actual versus predicted TPH concentration using the penalized spline prediction model. The right panel of Fig. 14 shows the fitted coefficient curve from the penalized spline. The fitted coefficient curve was smooth across the spectrum, indicating the stability of the model. The grey-shaded band shows the 95% confidence interval for the coefficients and can be used to discover the region that has a coefficient significantly different from zero, and the impact of this region on the response. Based on the foregoing results, it can be concluded that TPH was reasonably predicted by the penalized spline model. Figure 15 shows the experimental semivariogram and the fitted, exponential model for the predicted TPH values of the 118 collected samples. The maximum separation distance was 200 m. Bounded semivariance was observed at 200 m, indicating second-order stationarity of $\log_{10}$-transformed TPH. No nested spatial effect was detected visually. As showed, the isotropic experimental semivariogram was best fitted by an exponential model with a r^2 of 0.76. In addition, the effective range, nugget, sill, and nugget-to-sill ratio were 52m, 0.001 ($\log_{10}$ mg kg^{-1})^2, 1.044 ($\log_{10}$ mg kg^{-1})^2, and 0.001, respectively, implying a strong spatial dependence of the TPH values according to the classification criteria reported by Cambardella et al. (1994). Moreover, a nugget close to zero revealed that all variance of TPH was reasonably well explained, at the sampling distance used in this experiment by the lag. Using the semivariogram of the predicted TPH values, a distribution map was developed using ordinary block kriging interpolation in GS+ (Fig.

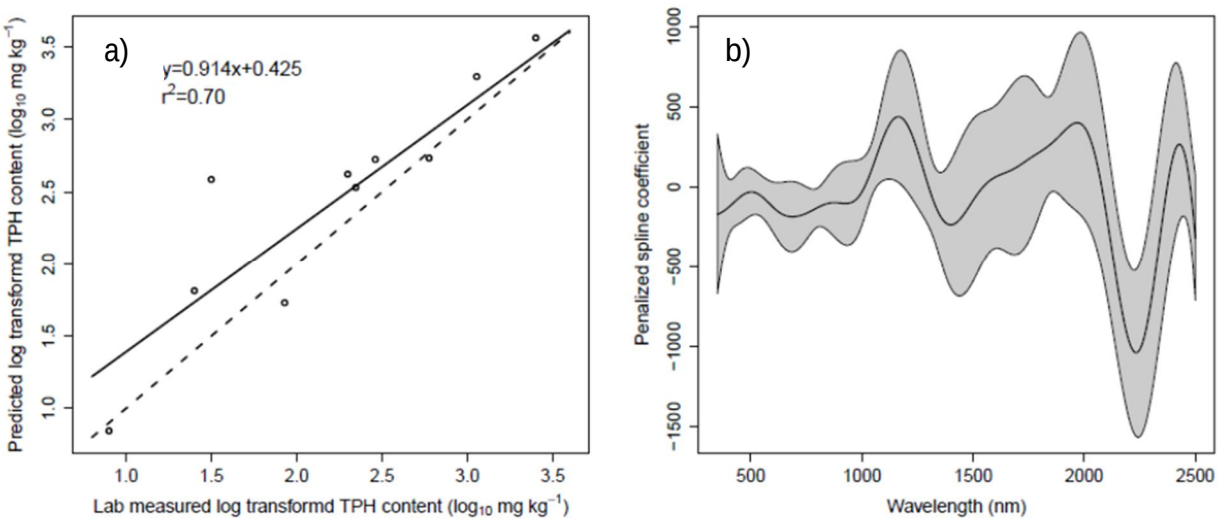

Fig. 14. a) Actual versus predicted total petroleum hydrocarbon (TPH) (log_{10} mg kg^{-1}) using penalized splines. The dotted line is the 1:1line and b) Fitted penalized splines coefficient curve at each waveband. The grey-shaded area is the 95% confidence interval.

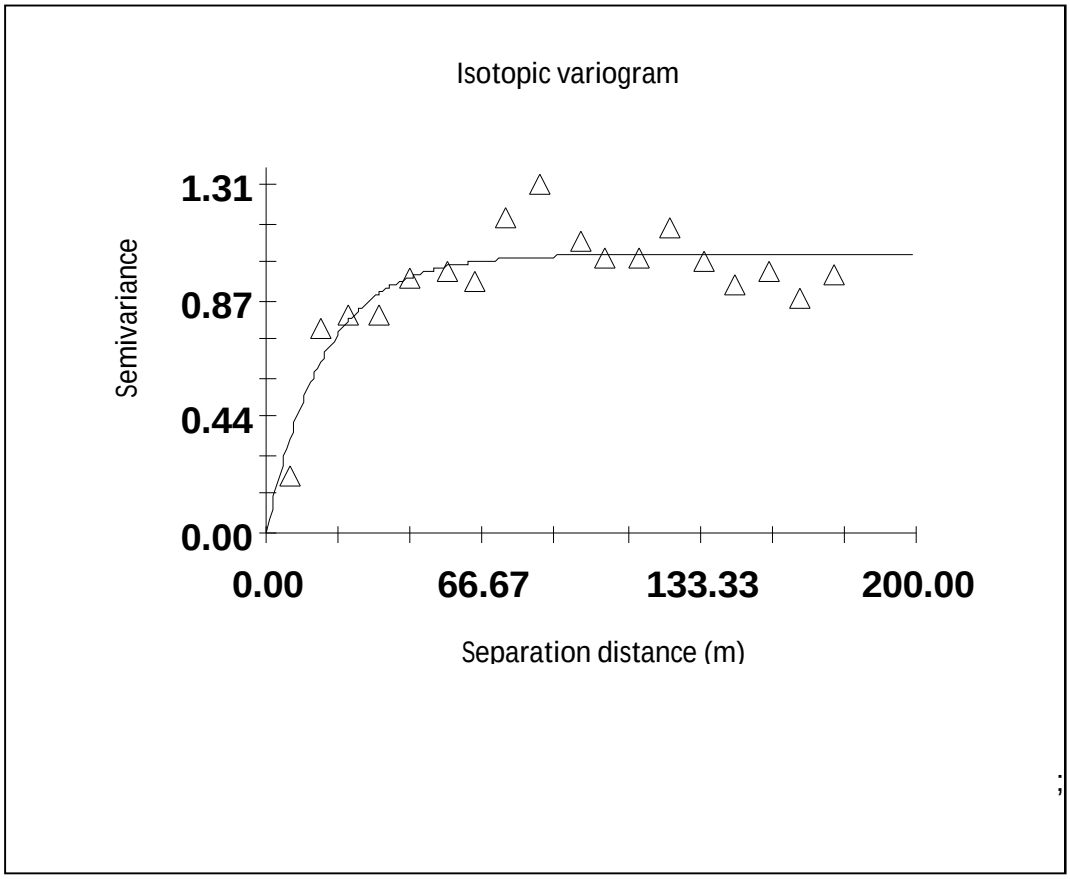

Fig 15. Experimental semivariogram and fitted theoretical model of $\log_{10}$-transformed total petroleum hydrocarbon (TPH). As showed, the isotropic experimental semivariogram was best fitted by an exponential model with r^2, range, nugget, sill, and nugget-to-sill ratio of 0.76, 52 m, 0.001 $(\log_{10}$ mg kg$^{-1})^2$, 1.044 $(\log_{10}$ mg kg$^{-1})^2$, and 0.001, respectively.

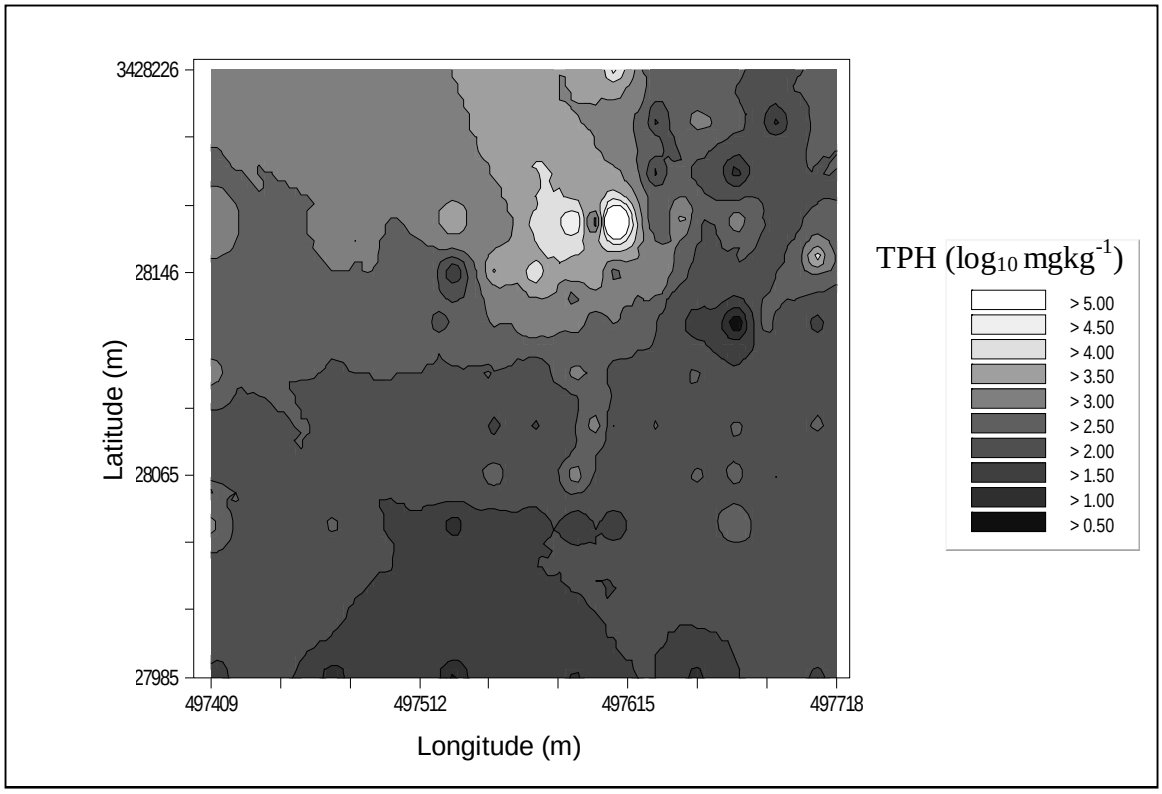

Fig 16. Kriging map for log_{10}-transformed total petroleum hydrocarbon (TPH).

16). The prediction capability of ordinary block kriging (using penalized spline predictions) in terms of r^2 (0.56) was less than simple penalized spline predictions in the validation subset, however, it showed a much smaller RMSE (0.307 $\log_{10}$ mg kg^{-1}) (Fig. 17). Perhaps the experimental semivariogram which was fitted in a trial-and–error mode was sub-optimal in reflecting the precise spatial pattern of TPH, as indicated by Ge et al. (2007). McBratney and Webster (1983) also revealed that the performance of kriging could be sufficiently enhanced by the use of a large dataset and optimal semivariogram. The 118 TPH predictions contained in the semivariogram may have been too limiting to generate kriging that had good generalization capability, i.e., that had high r^2 and low RMSE values. Nonetheless, as shown in Figure 16, the TPH distribution matched well with the topography of the study site (Fig. 11). The highest TPH values were found around the area where the oil spill occurred (Fig. 11). In the valley (the middle section of the study site), significantly elevated TPH level can also be identified readily as expected. After the occurrence, the spilled oil accumulated in the surrounding area and then naturally moved down to the area with lower elevations along with surface runoff and sediment. It should be noted that only a fairly low level of TPH was predicted in the valley and the elevated TPH values were also limited to within the valley. This implies that the spilled oil had a fairly low mobility and the impacts can be limited within a small area if proper remediation was implemented. Overall, the interpolated TPH predictions matched the topography and the expectation well. The prediction based on VisNIR DRS could provide a fast way to understand the spatial distribution of TPH contamination, to estimate the impacted area, and finally to expedite the process of decision making after contamination occurred.

There are some important issues that need to be explained. Data with a lognormal distribution which is characterized by a positive skewness, poses potential problem in kriging

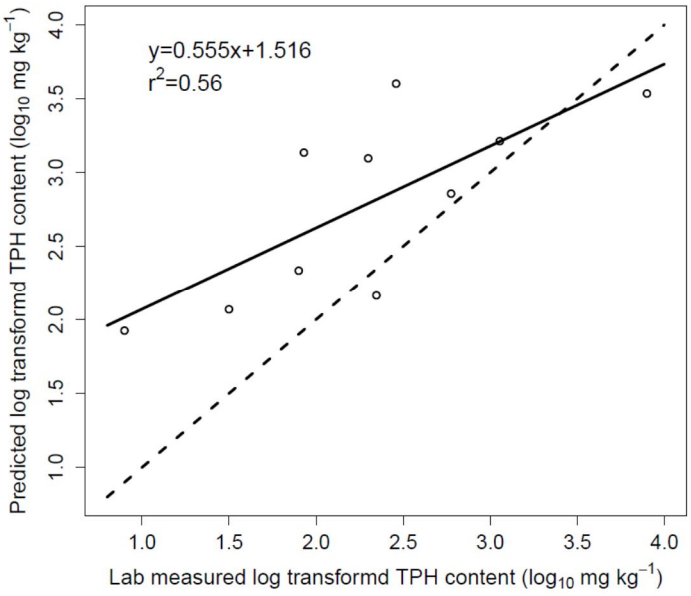

Fig. 17. Lab measured versus predicted (kriging interpolated) total petroleum hydrocarbon (TPH) (log_{10}mg kg^{-1}) for the validation subset (n=10). The dotted line is the 1:1line.

estimation. Journel (1983) revealed that experimental semivariograms are highly affected by lognormal data and there are only two solutions: trimming off high values or data transformation. Transformations can help validate assumptions of normality for kriging interpolation. Yamamoto and Furuie (2010) reported that logarithmic data transformation is always a better solution than data trimming in geostatistics. However, kriging approximations in the transformed domain need to be back-transformed into the original scale to obtain an unbiased result, after correcting the smoothing effect (Yamamoto, 2007). Note that in the present study, while using the penalized spline model predicted results for modeling the semivariogram, we did not back transform the kriging estimates. When we developed the penalized spline calibration model with lognormal data, we tried to fit the linear relationship in Eq. 5:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + e \qquad [5]$$

where Y stands for the abundance of the response of interest (TPH) at the $\log_{10}$ space, and x_1, x_2, …, x_p are the spectral variables; *p* and *b* are the number of spectral variables and penalized spline coefficients for spectral variables, respectively. We attempted to identify the optimal fitting at the $\log_{10}$ space, i.e., e (error) is i.i.d (independently and identically distributed) normal at the $\log_{10}$ space. However, if back transformation is applied, it approximates the relationship (not mathematically strict but close) given in Eq. 6:

$$\exp(Y) = \exp\left(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p\right) * \exp(e) \qquad [6]$$

The error e would not be optimal at the original (non $\log_{10}$-transformed) space in the sense $\log_{10}Y$ was used instead of Y. It is noteworthy that this error is not additive, but rather multiplicative. To obtain the optimal solution (in the sense of ordinary least squares) of the original space, it is necessary to start with the non-transformed variable. However, that would have produced useless predictions, affected by outliers. Therefore, it could be concluded that the

statistical relationship is not as simple as just doing the forward transformation then transformation back while the mathematical formulation indicates a far more complicated relationship. A solution to the aforementioned issue could be the further improvement of non-parametric modeling that could offset the outlier effect and handle the lognormal, limited data without transformation. Subsequently, VisNIR model predicted results could be transformed and used in kriging interpolation with the option of back-transformation. Further improvement could be made by using co-kriging or hybrid regression-kriging which is a combination of regression of spectral variable (reflectance) and geostatistical analysis of prediction residuals (Ge at al., 2007; Bilgili et al., 2010). With VisNIR DRS, regression-kriging utilizes cheaply and quickly obtainable reflectance spectra of the target component (with spatial reference via GPS) as auxiliary co-variables instead of landform attributes (such as slope, curvature, aspect, and elevation) derived from a digital elevation model. Nonetheless, a combination of VisNIR spectroscopy and ordinary kriging appears to be a reliable and efficient strategy for determining the spatial patterns of TPH contamination in soils, providing information for unvisited locations.

3.5. Conclusion

In this pilot study, the VisNIR predicted TPH results were incorporated into ordinary block kriging to identify spatial patterns of soil TPH contamination. A penalized spline model was developed with full cross-validation to predict TPH contamination in soils using lab TPH data and VisNIR spectra (First-derivative only) from contaminated and uncontaminated soils from central and southern Louisiana. This penalized spline model was used to predict soil TPH content for 128 soil samples collected over an 80 ha crude oil spill location. When validating with an independently chosen validation subset (n=10) from the aforementioned 128 samples, the penalized spline model performed satisfactorily with an r^2 of 0.70 and an RMSE of 0.409

$log_{10}mgkg^{-1}$. That same validation dataset was used to validate kriging interpolation after the remaining 118 predictions were used to produce an experimental semivariogram and kriging map. The experimental semivariogram was fitted with an exponential model which revealed strong spatial dependence among soil TPH contents. The prediction capability of ordinary block kriging using penalized spline predictions in terms of RMSE was 0.307 log_{10} $mgkg^{-1}$. In the kriging map, TPH distribution matched well with the topography of the study site. Overall, this study suggested that the combined use of VisNIR prediction and geostatistics have the potential to identify the spatial patterns of TPH contamination in soil quickly on site, reducing the need for expensive laboratory analyses.

3.6. References

Bilgili, A.V., F. Akbas, and H.M.V. Es. 2010. Combined use of hyperspectral VNIR reflectance spectroscopy and kriging to predict soil variables spatially. Precision Agric. Doi: 10.1007/s11119-010-9173-6.

Bofetta, P., N. Jourenkova, and P. Gustavson. 1997. Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons. Cancer Causes and Control 8(3):444-472.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. J. Royal Statistical Soc. Series B. 26:211-252.

Brown, D.J., R.S. Bricklemeyer, and P.R. Miller. 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VisNIR soil C prediction in Montana. Geoderma 129:251–267.

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

Calabrese, E.J., and P.T. Kostecki. 1988. Soils contaminated by petroleum. John Wiley & Sons, New York.

Cambardella, C.A., T.B. Moorman, J.M. Novak, T.B. Parkin, D.L. Karlen, R.F. Turco, and A.E. Konopka. 1994. Field-scale variability of soil properties in Central Iowa soils. Soil Sci. Soc. Am. J. 58:1501–1511.

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

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

Clesceri, L.S., A.E. Greenberg, and A.D. Eaton. (ed.) 1998. Standard methods for the examination of water and wastewater. 20th ed. American Public Health Association, American Water Work Association, and Water Environment Federation, Washington, DC.

Cole, G.M. 1994. Assessment and remediation of petroleum contaminated sites. Lewis Publishers Inc, Boca Raton, FL.

Crainiceanu, C.M., D. Ruppert, and M.P. Wand. 2005. Bayesian analysis for penalized spline regression using WinBUGS. Journal of Statistical Software 14:1-24.

Cressie, N. 1990. The origins of kriging. Mathematical Geol. 22:239-252.

Duffera, M., J.G. White, and R. Weisz. 2007. Spatial variability of Southeastern U.S. Coastal Plain soil physical properties: Implications for site-specific management. Geoderma 137:327-339.

Environmental Systems Research Institute. 2008. ArcGIS Desktop: Release 9.3. Redlands, CA.

Fine, P., E.R. Graber, and B. Yaron. 1997. Soil interactions with petroleum hydrocarbons: abiotic processes. Soil Tech. 10:133–153.

Forrester, S., L. Janik, and M. McLaughlin. 2010. An infrared spectroscopic test for total petroleum hydrocarbon (TPH) contamination in soils. Proc. 19th World Congress of Soil Science. 1-6 August. 2010. Brisbane, Australia.

Ge, Y., J. A. Thomasson, C.L. Morgan, and S.W. Searcy. 2007. VNIR diffuse reflectance spectroscopy for agricultural soil property determination based on regression-kriging. Trans. ASABE 50:1081–1092.

Hengl, T., G. B. M. Heuvelink, and A. Stein. 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120: 75-93.

Hutcheson, M.S., D. Pedersen, N.D. Anastas, J. Fitzgerald, and D. Silveman. 1996. Beyond TPH: health based evaluation of petroleum hydrocarbon exposures. Regulatory Toxicology and Pharmacology 24(1):85-101.

Journel, A.G. 1983. Nonparametric estimation of spatial distributions. Mathematical Geology 15:445-468.

MacEwen, J.D., and E.H. Vernot. 1985. Toxic hazards research unit annual technical report. AMRL-TR-81-126. Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio.

Malley, D.F. 1998. Near-infrared spectroscopy as a potential method for routine sediment analysis to improve rapidity and efficiency. Water Sci. Technol. 37:181–188.

Marx, B.D., and P.H.C. Eilers. 1999. Generalized linear regression on sampled signals and curves: a P-spline approach. Technometrics 41:1–13.

Matheron, G. 1965. Les variables regionalises et leur estimation. Masson, Paris, France.

McBratney, A.B., and Webster, R. 1983. Coregionalization and multiple sampling strategy. J of Soil Sci. 34: 249-263.

Morgan, C.L.S., T.H. Waiser, D.J. Brown, and C.T. Hallmark. 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. Geoderma 151:249-256.

Odlare, M., K. Svensson, and M. Pell. 2005. Near infrared spectroscopy for assessment of spatial soil variation in an agricultural field. Geoderma 126:193–202.

R Development Core Team. 2008. R: A language and environment for statistical computing. Available at http://www.R-project.org (verified 30 October 2010). R Foundation for Statistical Computing, Vienna, Austria.

Soil Survey Staff. 2009. Official soil series descriptions. Available at soils.usda.gov/technical/classification/osd/index.html (verified 26 April 2011). NRCS, Washington, DC.

Vasques, G.M., S. Grunwald, and J.O. Sickman. 2009. Modeling of soil organic carbon fractions using visible-near-infrared spectroscopy. Soil Sci. Soc. Am. J. 73:176-184.

Waiser, T.H., C.L.S. Morgan, D.J. Brown, and C.T. Hallmark. 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Sci. Soc. Am. J. 71:389-396.

Webster, R., and M.A. Oliver. 2001. Geostatistics for environmental scientists. Wiley, Chichester.

Yamamoto, J.K. 2007. On unbiased backtransform of lognormal kriging estimates. Computers and Geosciences 11: 219-234.

Yamamoto, J.K., and R.D.A. Furuie. 2010. A survey into estimation of lognormal data. Geociências 29(1): 5-19.

Zhu, Y., D.C. Weindorf, S. Chakraborty, B. Haggard, and N. Bakr. 2010. Determination of soil surface water content using diffuse reflectance spectroscopy. J. Hydrol. 391:133–140.

CHAPTER 4

SPECTRAL REFLECTANCE VARIABILITY FROM SOIL PHYSICOCHEMICAL PROPERTIES IN OIL CONTAMINATED SOILS

4.1. Synopsis

Oil spills occur across large landscapes in a variety of soils. Visible and near-infrared (VisNIR, 350–2500 nm) diffuse reflectance spectroscopy (DRS) is a rapid, cost-effective sensing method that has shown potential for characterizing petroleum contaminated soils. This study used DRS to measure reflectance patterns of 68 samples made by mixing samples from two soils with different clay content, three levels of organic carbon, three petroleum types and three or more levels of contamination per type. Both first derivative of reflectance and discrete wavelet transformations were used to preprocess the spectra. Three classification analyses (linear discriminant analysis, support vector machines, and random forest) and three multivariate regression methods (stepwise multiple linear regression, MLR; partial least squares regression, PLSR; and penalized spline) were used for pattern recognition and to develop the petroleum predictive models. Principal component analysis (PCA) was applied for qualitative VisNIR discrimination of variable soil types, organic carbon levels, petroleum types, and concentration levels. Soil types were separated with 100% accuracy (LDA), and levels of organic carbon were separated with 96% accuracy by linear discriminant analysis using the first nine principal components. The support vector machine produced 82% classification accuracy for organic carbon levels by repeated random splitting of the whole dataset. However, spectral absorptions for each petroleum hydrocarbon overlapped with each other and could not be separated with any classification scheme when contaminations were mixed. Wavelet-based MLR performed best for predicting petroleum amount with the highest residual prediction deviation (RPD) of 3.97. While using the first derivative of reflectance spectra, penalized spline regression performed better

(RPD = 3.3) than PLSR (RPD= 2.5) model. Specific calibrations considering additional soil physicochemical variability and integrating wavelet-penalized spline are expected to produce useful spectral libraries for petroleum contaminated soils.

4.2. Introduction

Oil contaminated soils are problematic in many areas; both coastal and inland. While there is heightened media attention on the 2010 Deepwater Horizon oil spill in the Gulf of Mexico, smaller inland spills occur on a regular basis. These spills typically occur in the form of broken oil well service lines, leaking storage tanks or crumbling infrastructure, long term leakage, and underground gasoline storage tanks at local fuel stations. In some cases, agricultural soils are affected (where oil production is occurring concurrently with crop production), but in other instances, the contamination may take place in wildlife refuges or national parks. 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). To better understand contaminate transport, fate, and remediation, reliable methodologies for monitoring/measuring petroleum hydrocarbon contamination in soils are warranted.

Measurement of petroleum hydrocarbons in contaminated soils is time consuming and requires rigorous field sampling besides costly wet chemical analyses, making wide-scale quantitative assessment challenging (Dent and Young, 1981). Gas-chromatography based laboratory methods for total petroleum hydrocarbon (TPH) quantification lack field-portability (Forrester et al., 2010). Moreover, a lack of standardized methods has resulted in high variability (an order of magnitude) in TPH results across commercial laboratories (Graham, 1998; Malle and Fowlie, 1998; Malley et al., 1999). Hence, there is a pressing need for an innovative, rapid, environmentally responsible, and cost-effective sensing technology to identify petroleum

contaminated areas for remediation and to monitor restoration on an ongoing basis (Prince, 1993).

Optical sensors can differentiate and quantify spectrally alike (but unique) objects having subtle signature variations (Wetzel, 1983; Hyvarinen et al., 1992; Ge et al., 2007). Besides, advancements in both near-infrared (NIR) based proximal sensors with a fiber optic probe and chemometric analysis have extended near-infrared spectrometry (NIRS) to petroleum industries for identification of gasoline and middle distillate fuel properties (Westbrook, 1993; Workman, 1996; Current and Tilotta, 1997; Chung et al., 1999; Chung and Ku, 2000; Yoon et al., 2002; Balabin and Safieva, 2008). Synergistic arrangement of optical sensors for diverse regions of the electromagnetic spectrum is capable of identifying petroleum contamination in a targeted matrix.

Visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) might be a useful proximal sensing tool to identify soil petroleum contamination because the scanning is rapid and non-destructive, instruments are field portable, and costs are fixed. Recent evidence suggests that VisNIR DRS and chemometric modeling offer comparable levels of accuracy to standard physicochemical analysis of various soil properties (Ben-Dor and Banin, 1995; Reeves et al., 2000; Islam et al., 2003; Brown et al., 2005; Viscarra Rossel et al., 2006; Morgan et al., 2009; Vasques et al., 2009). To date, researchers have identified various spectral regions in VisNIR associated with soil clays and organic matter. Ben-Dor and Banin (1990) proved the usefulness of near-infrared reflectance spectroscopy in chemical characterization of clay minerals. Moreover, Waiser et al. (2007) concluded that VisNIR DRS could predict soil clay content with reasonable accuracy. While overtones of OH^-, SO_4^{2-}, and CO_3^{2-} groups and combination bands of H_2O and CO_2 are responsible for unique spectral signatures of common clay minerals; O-H, C-N, N-H, and C=O groups are active bonds for soil organic matter in the NIR region (Bowers and

Hanks, 1965; Hunt and Salisbury, 1970; Al-Abbas et al., 1972; Hunt, 1982; Malley et al., 2002; Brown et al., 2005). A number of studies have reported an increase in prediction accuracy when VisNIR-organic C models were created for small, homogenous areas (Lee et al., 2003; Chang et al., 2005). Concurrently, other researchers have observed decreased prediction accuracy for larger geographic areas (Dunn et al., 2002; Shepherd and Walsh, 2002; Brown et al., 2006; Kusumo et al; 2008). Nonetheless, less attention has been given to quantitative spectral analysis of petroleum contaminated soils with variable texture and organic carbon and remains a considerable task.

While researchers have proposed several calibration techniques to relate NIR spectra with measured soil properties, only a few studies have sought to quantitatively understand the effect of petroleum hydrocarbon on shortwave reflection. Malley et al. (1999) reported validation r^2 of 0.68 and 0.72 for NIR TPH predictions in diesel fuel contaminated soils. Forrester et al. (2010) used PLS cross-validation chemometric modeling for infrared spectroscopic identification of TPH. Chakraborty et al. (2010) used PLS regression and boosted regression tree modeling for identification of petroleum contaminated soils. However, there has been little effort on the development of dedicated spectral libraries for soil-petroleum contamination appraisal.

The goal of this communication is two-fold and is a continuation of the work by Chakraborty et al. (2010), which demonstrated the feasibility of VisNIR DRS for rapid and in-situ identification of petroleum hydrocarbon in soil, without prior sample preparation. Our primary goal is the further clarification of the relationship between soil petroleum hydrocarbon and reflectance measurements based on multivariate regression methods and classification techniques, in the context of variable soil texture and organic carbon levels. Furthermore, this

research investigates the possibility of linking specific wavebands to unique petroleum hydrocarbons.

The authors acknowledge that the limited number of samples (68) somewhat constrain the global applicability of the dataset. However, this research was intended to investigate the effect of soil variability on VisNIR-based TPH predictions in soil, investigate the viability of different spectral analysis techniques and ascertain which techniques show the most promise for future investigations.

The applicability of VisNIR technology and methods tested in this study is broad. Most NIR spectroscopic investigations of petroleum contaminated soils have had limited scope because of the limited variability of oil types, and/or because less importance was given to soil texture and organic carbon, which can be both spatially and temporally variable, and management dependent **(**Russell et al., 2005). Characterization of petroleum spectral patterns for variable amounts of soil organic carbon and variable soil texture might be more useful for creating a spectral library for large geographic areas. Combinations of ideal data-mining or pattern-detection tools for using VisNIR DRS to characterize petroleum contaminated soil is useful for understanding other potential applications of the technology. The present research envisions a VisNIR-DRS optical sensor located in a soil probe for in-situ characterization of both surface and subsurface petroleum contamination in soils. Hence, the specific objectives of this research were to: (i) examine the effect of variable soil texture, organic carbon, and oil types on VisNIR reflectance patterns of petroleum contaminated soils and, (ii) compare different spectral preprocessing and multivariate data-mining tools for characterizing petroleum contaminated soils and future development of VisNIR-based optical sensors.

4.3. Materials and Methods

4.3.1. Sample Preparation

Two soil samples (10-30 cm) with no known hydrocarbon contamination were collected from an active agricultural production field at the LSU AgCenter St. Gabriel Research Station, near Baton Rouge, Louisiana, USA, (30°16' 8" N, 91°6' 16" W). Soil A is a Commerce silt loam (Fine-silty, mixed, superactive, nonacid, thermic Fluvaquentic Endoaquept), and Soil B is a Schriever clay (Very-fine, smectitic, hyperthermic Chromic Epiaquert) (Soil Survey Staff, 2005). Soil samples were air-dried, ground, and passed through a 2-mm sieve. A gravimetric soil moisture subsample was used for oven-dry weight correction for laboratory analysis. Laboratory procedures included particle size analysis by pipette method with an error of ±1% clay (Steele and Bradfield, 1934; Kilmer and Alexander, 1949; Gee and Or, 2002) and saturated paste pH (Soil Survey Staff, 2004). Total carbon levels were determined by Dumas Method combustion using a TruSpec ® CN analyzer (LECO, St. Joseph, MI, USA) (Dumas, 1831; Wang et al., 2003). Inorganic C was measured using the modified pressure calcimeter method (Sherrod et al., 2002). Organic carbon was determined as the difference of total carbon and inorganic carbon. Natural organic carbon levels for Soil A and Soil B were both very low (≤0.5%). These soils were spiked with a mixture of natural muck (collected from a local swamp) and commercially available sphagnum so that Soils A and B were made to contain approximately 1%, 5%, and 10% organic carbon on a gravimetric basis. Before spiking the soils, the dried sphagnum was chopped and the spagnum and muck were sieved (2 mm). Each of the soil-organic matter mixtures were spiked with three types of petroleum and at three concentrations. The three grades of petroleum included crude oil, diesel, and used (Penzoil 10-30 weight) motor oil, and the three levels of concentration were 1000, 10000, and 30000 ppm. Additionally, an extra set of nine intermediate

levels of crude oil concentrations (4000-28000 ppm at 3000 ppm intervals) was created for Soil B, organic carbon content of 5 %, improve the capability to fit models and to test their results. Before spiking with organic material and petroleum, all soil samples were moistened to reach 7.5% moisture content, by weight. Each sample was thoroughly homogenized using a stainless steel spatula, stored in sealed glass jars capped with an aluminum lined cap, and refrigerated to prevent hydrocarbon volatilization.

4.3.2. VisNIR DRS Scanning

In the laboratory, the constructed samples were scanned using a field portable AgriSpec VisNIR spectroradiometer (Analytical Spectral Devices, CO, USA) with a spectral range of 350 to 2500 nm (ultraviolet/VisNIR [350–965 nm], short-wave infrared 1 [966–1,755 nm], and short-wave infrared 2 [1756–2500 nm]). The spectroradiometer had a 1-nm sampling interval and a spectral resolution of 3- and 10-nm wavelengths from 350 to1000 nm and 1000 to 2500 nm, respectively. About 30 g of each sample was placed into a Duraplan® borosilicate optical-glass Petri dish and scanned from below using a muglamp with a tungsten quartz halogen light source (Analytical Spectral Devices, CO, USA). Each sample was scanned four times with a 90° rotation between successive scans to obtain an average spectral curve. A spectralon panel with 99% reflectance was used every five samples to optimize and white reference the spectroradiometer.

4.3.3. Pre-treatment of Spectral Data

In the present study, we compared two techniques (1st derivative of reflectance and discrete wavelet transform) to preprocess the soil spectra prior to analysis. Three classification analysis techniques were utilized for pattern recognition, including linear discriminant analysis, support vector machines, and random forest. Moreover, three multivariate regression methods

(stepwise multiple linear regression, MLR; partial least squares regression, PLSR; and penalized spline) were compared to develop the petroleum predictive models. A statistical analysis software package, R version 2.11.0 (R Development Core Team, 2008) was used to preprocess raw reflectance spectra. Based on a comparative analysis described by Chakraborty et al. (2010), only the smooth reflectance and the first-derivative of reflectance spectra on 10-nm intervals were extracted using custom 'R' routines (Brown et al., 2006). From previous studies, it is apparent that first-derivative spectra can remove the baseline shift arising from detector inconsistencies, albedo, and sample handling, improving the accuracy of quantification **(**Demetriades-Shah et al., 1990).

4.3.4. Wavelet Analysis and Stepwise Multiple Linear Regression

In VisNIR spectroscopic analysis, wavelets have been proposed by several researchers to pre-treat spectral data and develop calibration models (Lark and Webster, 1999; Ge et al., 2007; Viscarra Rossel and Lark, 2010). Wavelet coefficients at higher scales have local support and correspond to fast varying, undesirable noise of individual bands in the spectral measurement; whereas, those at lower scales have wide support and correspond to slow, varying signal shifts involving many contiguous bands (e.g., instrument dark current shift due to ambient temperature change). For these reasons, wavelets are regarded as a useful tool for VisNIR spectral data pretreatment and model calibration. By discarding wavelet coefficients at high and low scales, the remaining coefficients capture the absorption information and give rise to a more informative calibration model compared to PLSR techniques. A spectroscopy-specific example of wavelet transformation can be seen in Ge et al. (2007).

The Haar wavelet system was used to process spectral data and the filter bank algorithm was implemented to dyadically decompose each soil spectrum (original noise corrupted, 1-nm

interval) from the highest scale (Scale 11, representing the raw spectrum itself) to lowest (scale 0, representing the average of the spectrum). The wavelet coefficients at scales 7, 6, and 5 which had bandwidths of 128, 64, and 32 nm, respectively, were extracted. Among them, six wavelet coefficients were selected (by stepwise multiple linear regression) for VisNIR model development. The wavelet decomposition was performed using the Wavelet Toolbox in MatLab R20009a (The MathWorks, MA, USA) while the stepwise MLR model was built in R version 2.11.0.

4.3.5. Principal Component Analysis

Principal component analysis (PCA) was applied for qualitative VisNIR discrimination of the prepared samples according to the variable soil types, organic carbon levels, oil grades, and oil concentrations. The cumulative proportion of variance explained by the leading principal components (PC) was used to extract optimum PCs. Fisher's linear discriminant analysis (LDA) was then applied on the selected leading PCs, assuming equal prior probability for each group. To assess classification results, kappa coefficients were computed (Thompson and Walter, 1988). Furthermore, pairwise scatterplots of the first three PCs were produced to provide visual assessment on how different groups were separated in the PC space. To test whether soil type and organic carbon content mask the signature of different oil types, pairwise scatterplots of the first nine PCs were produced for a particular soil type with a specific organic carbon content and multiple oil grades. PCA was performed using R version 2.11.0 (function: prcomp).

4.3.6. Support Vector Machine and Random Forest

Support vector machine (Boser et al., 1992; Vapnik, 1995) and random forest (Breiman, 2001) are two popular data mining methods, which were recently proposed for VisNIR modeling applications (Stum, 2010). From the geometric perspective, support vector machine is a margin-

based classifier. For a separable binary classification problem support vector machine chooses a hyperplane so that the distance from it to the nearest data point on each side is maximized. For non-separable data (VisNIR data), the soft-margin support vector machine chooses a hyperplane that splits two classes as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. A complex space with non-linear multivariate relationships is transformed into a higher dimensional, linear (inner product) space via the *kernel trick*, the SVM problem is solved in the linear dataspace, then back-transformed to the lower dimensional space for the result. A desirable property of support vector machine is that its solution only depends on a subset of training examples called support vectors. The support vector machine was performed by using the "e1071 package", an R interface to library for support vector machines (LIBSVM) (Chang and Lin, 2001). The radial basis kernel was used.

Random forest is an enhancement that aims to improve the performance of a single decision tree by fitting many trees (and thus the name 'forest') and combining them for prediction. The final prediction is based on majority votes over all the trees built. In random forest, the decision trees are different because of the following two factors: (1) at each tree node (splitting point), a best split is chosen from a random subset of the input variables rather than all of them and, (2) each tree is built based on a bootstrap sample of the observations. The random forest was performed in R using the "randomForest package" developed by Breiman and Cutler (Breiman, 2001). A total of 500 trees were generated for each random forest model.

4.3.7. Wavelet- Support Vector Machine Classifier

The discrete wavelet decomposition algorithm (Mallat, 1989) was applied to the spectral matrix (both reflectance and first-derivative of reflectance at 10-nm intervals) and the wavelet

coefficients to extract important features. Before the wavelet coefficients were decomposed, thresholding was applied to eliminate "unimportant" (not significantly different from zero mean) coefficients considered to be noise. Details of wavelet thresholding are elucidated by Donoho and Johnstone (1994). After wavelet decomposition, support vector machine was applied to the extracted features (the wavelet coefficients after thresholding). After thresholding, 108 wavelet coefficients were left (i.e. at least one non-zero value among all the samples).

4.3.8. Partial Least Squares Regression

Partial least squares regression was employed to help predicting petroleum content through spectral and concentration matrix decomposition using R version 2.11.0. Quantitative PLSR modeling can handle the complicated relationship between the predictors and responses resulting from multicolinearity of predictors, random linear baselines, and overlapping of major spectral components of predictors with that of the analytes (Wold et al., 2001). The whole dataset (68 samples) was used for training with leave-one-out cross-validation for model creation and selection for the number of latent factors (rotations of PCs for a different optimization criterion). Models with as many as nine factors were considered, and the optimal model was determined by choosing the number of latent factors with the first local minimum in root mean squared error of cross-validation (RMSEcv). The coefficient of determination (r^2), and residual prediction deviation (RPD) (the ratio of standard deviation to $RMSE_{CV}$) were used as rubrics for evaluating the quality of PLSR and other models in real-world situations.

4.3.9. Penalized Spline

In PLSR, the order of the regressor channels (wavelengths) is ignored. In other words, the same results will be obtained when the regressors are shuffled. Penalized spline (Eilers and Marx, 1996) attempts to take advantage of the additional structure from the order of regressors.

Namely, it forces the regression coefficients to be smooth (i.e. constraining the difference between the neighboring regression coefficients). The smoothness comes from a difference penalty on adjacent regression coefficients. This penalty is proportional to the size of the difference between neighborhood coefficients. Because of the additional constraint imposed by the difference penalty, penalized spline is well-suited for ill-posed problems (the dimensionality is much larger than the sample size) such as signal regression problems. A nice property of the penalized spline is that it is within the linear regression framework. Hence, it inherits all the statistical inferences of linear regression, such as confidence intervals. In addition, like linear regression, penalized spline can run the leave-one-out cross-validation by fitting the model on the entire dataset once, without recomputation of the regression model omitting each observation. For the details of penalized spline, we refer readers to Eilers and Marx (1996).

In the present study, the cubic B-spline was used (using R version 2.11.0) as the basis functions with 100 equally-spaced knots. The order of the penalty was set to the default value of three. The optimal value for the penalty-tuning parameter was selected by minimizing the leave-one-out cross-validation error.

4.4. Results and Discussion

Particle size analysis confirmed soil textures of soil A and B as silt loam (8.7% clay) and clay (47.1% clay), respectively. Both soils exhibited similar pHs (6.3 and 6.6, respectively). Average reflectance spectra for soil samples with 1% organic carbon and three concentrations of diesel (ppm or mg kg^{-1}) are shown in Fig. 18. In general, mean spectral reflectance decreased as diesel concentration increased, as expected (Hoerig et al., 2001). Note that, the specific absorption maximums of petroleum at 1730 (C-H stretch 1^{st} overtone band) and 2310 nm (C-H

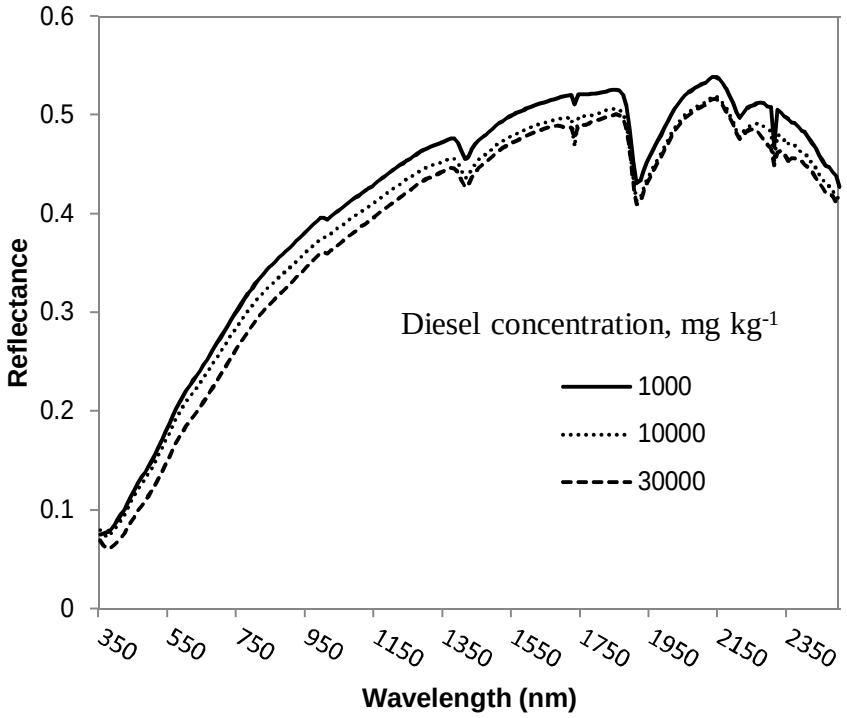

Fig. 18. Average reflectance spectra is shown for Soil A from Louisiana, USA with 1% organic carbon and different concentrations of diesel (ppm or mg kg^{-1}).

stretch combination band), as already exhibited by Cloutis (1989), were clearly identified by VisNIR DRS. Other researchers identified that the first overtone of the C-H band makes the most important contribution for analysis of oil systems (Balabin and Safieva, 2007). It is always desirable to use individual reflectance/absorption features while calibrating petroleum concentrations and spectral reflectance.

4.4.1. Classification

Eighty-eight percent of the spectral variance was explained by the first nine PCs. Despite the high dimensionality of the spectral data (215 channels from 350 to 2500 nm at 10-nm intervals), three quarters of the variation was primarily explained by the first five PCs (76%). Separate pairwise PC score plots for soil types and oil grades indicating organic carbon levels were used (Figs. 19 and 20, respectively) to discriminate reflectance spectra and identify classification patterns. Figure 19a, illustrates how the first PC separates the samples from Soil A and B with less differentiation by organic carbon content. Principal component two delineates the three quantities of organic carbon (Fig. 19a and b). Conversely, clear separations between contaminant oil types and concentrations were not delineated by first three PCs nor any of the first nine PCs.

Results of LDA classification closely followed results of visual PC plot inspections. Notably, for soil type classification, LDA was 100% accurate in classifying soil types; LDA correctly classified all but three samples by soil organic carbon content, but oil type was not discernable using LDA (Table 6).

Figure 21 shows pairwise scatterplots of the first nine PCs for soil B with 5% organic carbon and multiple oil types. It was challenging to test whether soil type and organic carbon

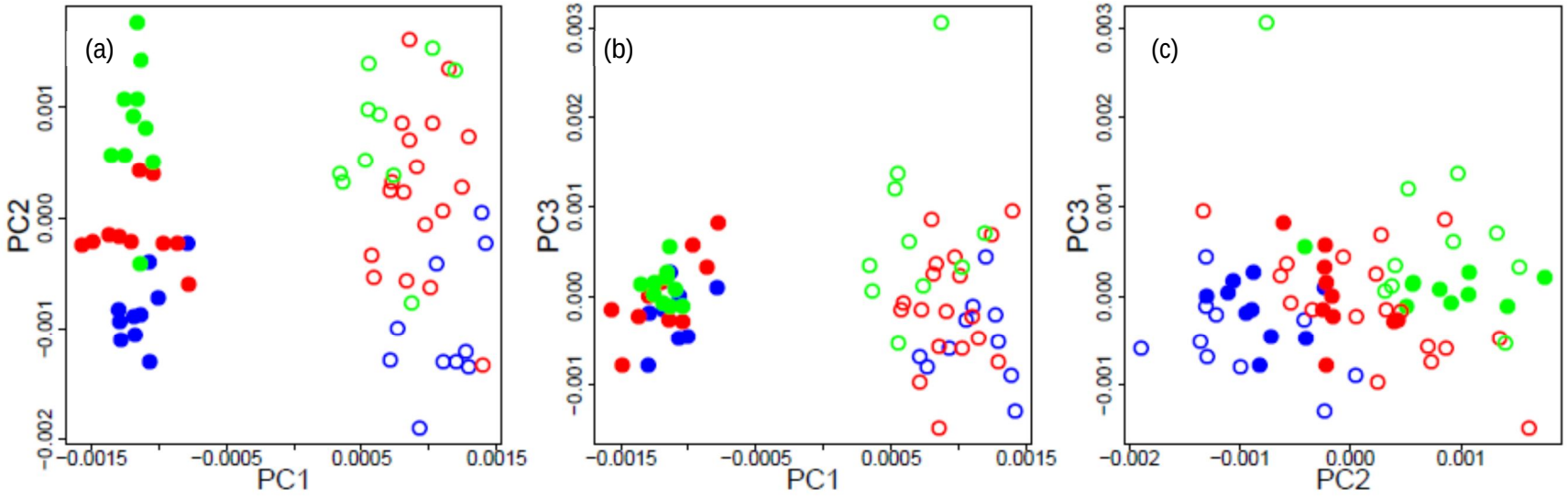

Fig. 19. Principal component (PC) plots for (a) PC1 vs. PC2, (b) PC1 vs. PC3, and (c) PC2 vs. PC3 of the first-derivative of VisNIR reflectance spectra. The solid circles and open circles represent Soil A and Soil B, respectively. Blue, red, and green represent soils with 1%, 5%, and 10% organic carbon, respectively.

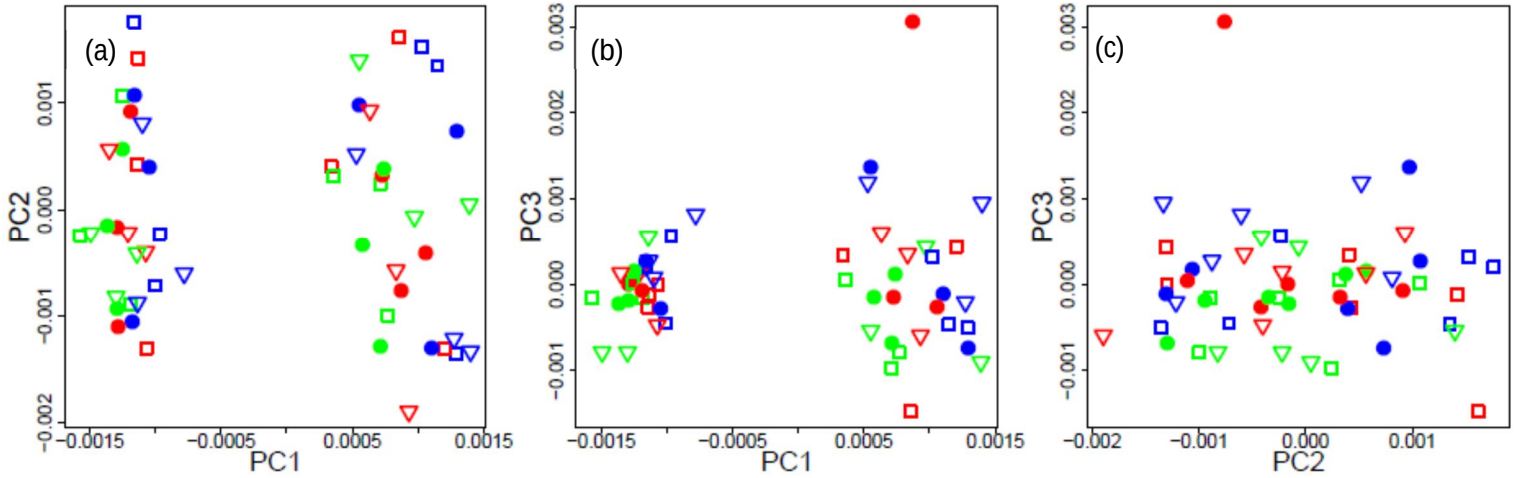

Fig. 20. Principal component (PC) plots for (a) PC1 vs. PC2, (b) PC1 vs. PC3, and (c) PC2 vs. PC3 using the first-derivative of VisNIR reflectance spectra. The circles, squares, and triangles represent soils with crude oil, diesel, and motor oil respectively. The blue, red, and green represent 1000, 10000, and 30000 oil concentrations (ppm or mg kg^{-1}), respectively.

Table 6. Results for classifying soil organic carbon levels and oil types using the Fisher's Linear Discriminant Analysis (LDA). The first nine principal components (PC) scores of the first-derivative spectra were used as the explanatory variable. Control type was excluded from oil types analysis. The weighted kappa coefficients are 0.96 and 0.16 for organic carbon and oil types, respectively.

.

	LDA-classified organic carbon					LDA-classified oil types			
	1%	5%	10%	Sum		Crude	Diesel	Motor oil	Sum
Low	19	1	0	20	Crude	14	7	5	26
Medium	0	28	0	28	Diesel	9	3	6	18
High	0	2	18	20	Motor oil	7	4	7	18
	19	31	18	68		30	14	18	62
				96%					40%

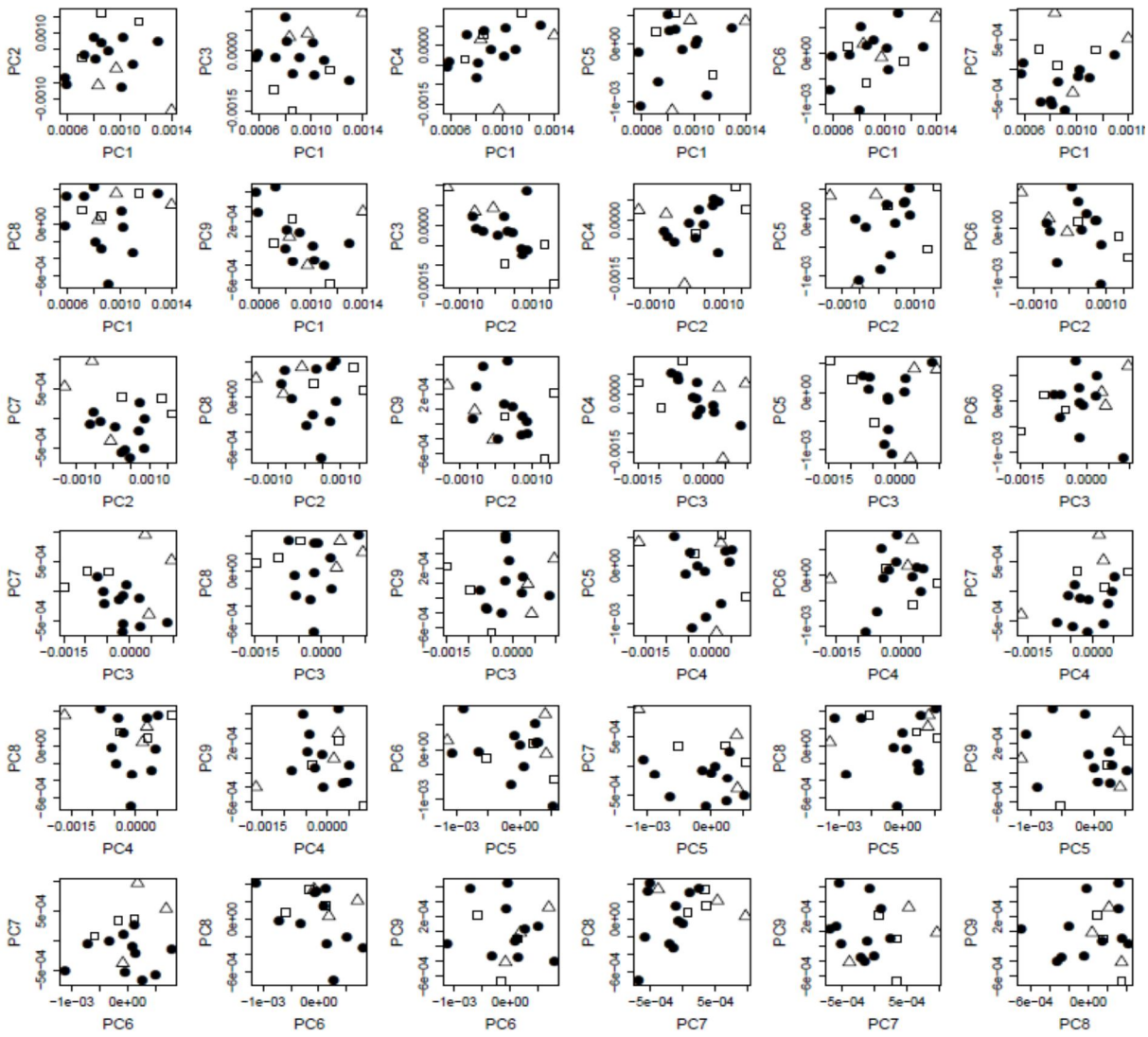

Fig 21. Principal component (PC) plots using the first-derivative of VisNIR reflectance spectra. The circles, squares, and triangles represent soil B with crude oil, diesel, and motor oil, respectively. All samples contain 5% organic carbon.

mask the oil type signatures due to the small sample size. However, the plots indicated some separation of three oil types (the motor oil & diesel are two ends and crude oil is in the middle) and implied that oil type signatures were not completely masked by soil type and organic carbon signatures. Nonetheless, no conclusion could be made unless comparing these results with more soil types with different organic carbon contents. To evaluate the prediction performance for the support vector machine, random forest, and wavelet-based support vector machine, the whole dataset was randomly split 50 times. For each split, a training set contained 48 samples and a test set contained 14 samples. The control (no petroleum contaminate added) samples were excluded. The models were trained on the training set, while the prediction was evaluated on the test set. The prediction performance of the support vector machine, random forest, and wavelet-based support vector machine was compared based on the percent misclassifications on the 50 test sets. The summary prediction performance on oil type, organic carbon level, and soil type is presented in Table 7. From the average misclassification rate (%) from 50 test sets, it was clear that the support vector machine, random forest, and wavelet-support vector machine had similar prediction performances. All methods separated the two soil types with little to no error. For organic carbon, the support vector machine performed slightly better (18%) than the random forest(18.5%), while the wavelet-support vector machine misclassified twice as often as the first two methods. For oil types, all three methods misclassified over 50% of the time. The misclassification rates for a full leave-one out cross validation, using the entire dataset, were much smaller than the 48-sample training set misclassification rate (Table 7). Particularly, the misclassification rate for classing oil types went to 0 to 23 %. A misclassification rate between 60 and 0 % is a large but realistic estimate of the ability for VisNIR spectroscopy to classify petroleum contamination type in soils. Clearly more samples in a training set and clearly defined

Table 7. Summary of classification performance on oil type, organic carbon content, and soil type for four classification methods.

	Soil type		Organic carbon content		Oil type	
	Average MR†	MR for whole data set (no split)	Average MR	MR for whole data set (no split)	Average MR	MR for whole data set (no split)
	----------%----------					
Support vector machine	0	0	18	1.6	67	23
Random forest	0	0	18.5	0	63	0
Wavelet(1^{st})-SVM‡	1.3	0	26	0	64	6.5
Wavelets (r)-SVM§	0	0	30	1.6	65	23

† MR, Misclassification rate.
‡ Wavelet(1^{st})-SVM, Wavelet decomposition on first-derivative of reflectance followed by Support Vector Machine.
§ Wavelets (r)-SVM, Wavelet decomposition on raw noise corrupted reflectance spectra followed by Support Vector Machine.

contamination types improves the probability of a correct classification. However these results are not encouraging especially if contamination types were mixed.

4.4.2. Multivariate Modeling

Three multivariate regression techniques were used to relate the VisNIR reflectance spectra to oil concentrations with leave-one-out cross validation. Accuracy and stability of different multivariate models were evaluated according to the RPD-based guidelines by Chang et al. (2001). The best prediction models are characterized by a RPD of >2.0 with r^2 of ~0.80-1.00, fair models with potential for prediction improvement include RPD values of 1.4-2.0, while unreliable models have RPD values of <1.40. These RPD values are most useful when the validation set is independent of the calibration set; however, with leave-on-out cross validation they are still useful indicators for describing the potential of the technology.

A plot of actual versus PLSR predicted oil concentration in soil samples is presented in Fig. 22a. The PLSR plot showes that the prediction method was less accurate at larger concentrations. In linear regression modeling (which includes PLSR), one of the assumptions is homogeneity of variance, also known as homoscedasticity assumption. However, it was evident that the error variance was not constant in case of PLSR (i.e. variance increases with the actual oil concentration). A trend in prediction residuals by soil type, organic carbon levels, and oil type was investigated (Fig. 23). Non apparent trends in the TPH prediction residuals were found by organic matter and soil type; however it does seem that overall motor oil residuals were higher than the residuals from the other oil types. The motor oil was used motor oil. Perhaps the PLSR model had difficulty with the spectral signatures of the impurities. The PLSR model used

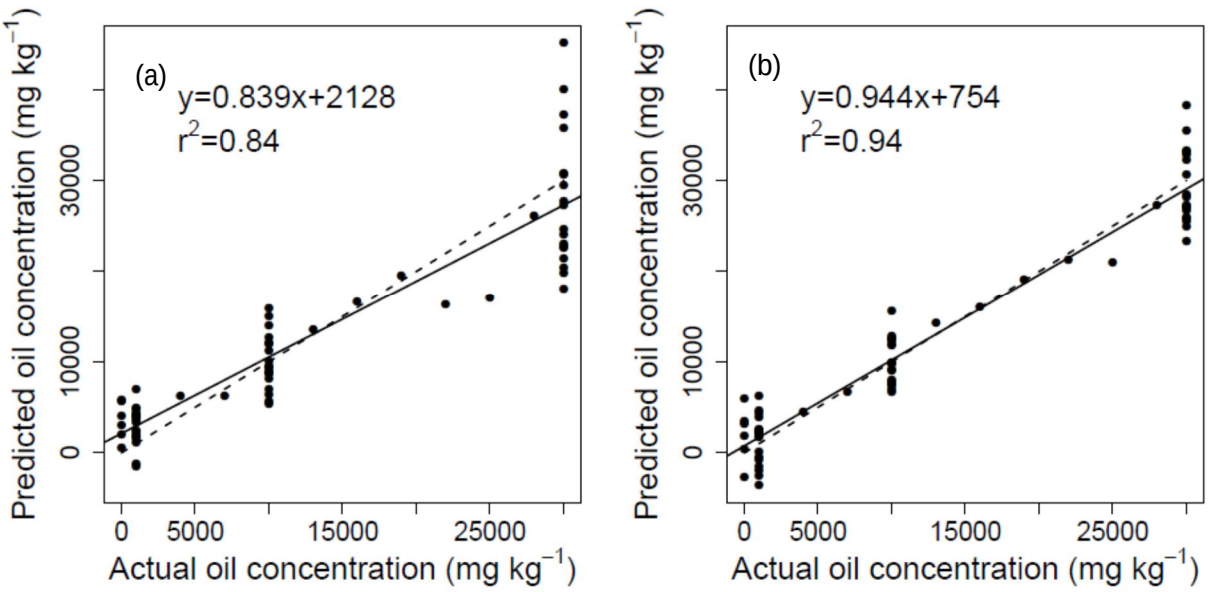

Fig. 22. Actual versus predicted oil concentration (mg kg^{-1}) using a) partial least squares regression (PLSR) and b) wavelet coefficients from the reflectance, and stepwise multiple linear regression (MLR). The solid line is the regression line, and the dashed line is a 1:1 line.

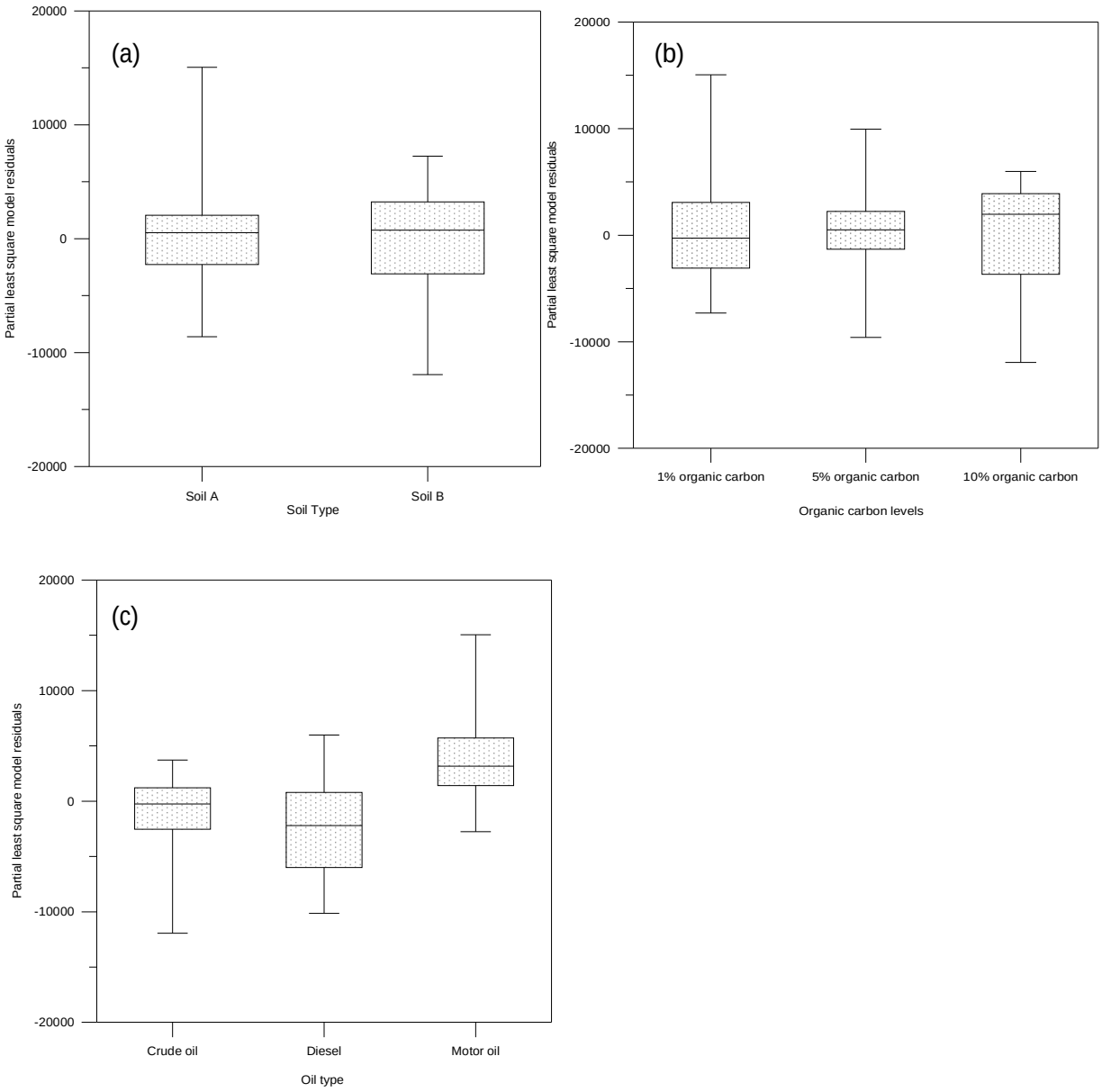

Fig. 23. Plots showing partial least squares model prediction residuals vs. a) soil type, b) organic carbon levels, and c) oil type.

three latent factors. The number of latent factors in the present study was less than the values reported by other oil related research (Aske et al., 2001; Balabin and Safieva, 2007). Notably, the abovementioned oil related research used petroleum macromolecules in the model systems without the influence of heterogeneous soil matrix.

Predictions of oil concentration using wavelet-MLR more closely approximated the 1:1 line and had less bias (33 mg kg^{-1}) (Fig. 22b). The wavelet model, developed by using wavelet coefficients from the reflectance spectra with leave-one-out cross validated stepwise MLR performed the best of the three models. Most importantly the residuals from the wavelet-MLR model were homoscedastic, which lends more credibility to this model. Prediction accuracy and model fit of the penalized splines method were better than the PLS, but had a lower RPD than wavelet-MLR (Fig. 24 and Table 8). The fitted coefficient curve was smooth across the spectrum, indicating stability of the model. The grey-shaded band shows the 95% confidence interval for the coefficients and can be used to discover the region that has a coefficient significantly different from zero, and the impact of this region on the response. For example, the 1300-1400nm and 1550-1700nm regions are both away from zero. However, the former contributes a positive effect on the oil concentration while the latter has a negative effect.

Among the multivariate methods tested, the wavelet-MLR and penalized spline regression models performed better than PLSR model. The wavelet-MLR yielded the highest predictability (RPD=3.97), with the lowest $RMSE_{cv}$ (3010 mg kg^{-1}). Furthermore, the penalized spline model provided the highest coefficient of determination (0.98) along with a high RPD (3.32), indicating the robustness and accuracy of both wavelet-MLR and penalized spline models. Ge et al. (2007) concluded that the main advantage of dyadic discrete wavelet transformation over traditional PLSR and principal component regression based methods is the

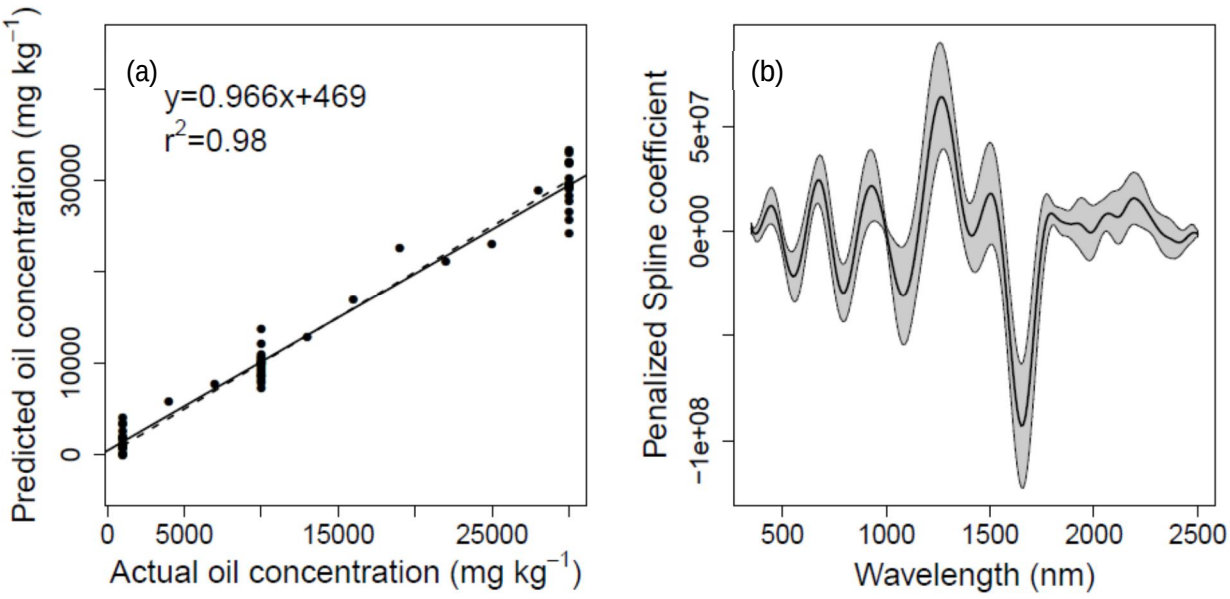

Fig. 24. (a) Actual versus predicted oil concentration (mg kg^{-1}) using penalized splines for soils from Louisiana, USA and (b) Fitted penalized splines coefficient curve with a grey-shaded area showing the 95% confidence interval at each waveband.

Table 8. Summary of oil contamination prediction performance using different multivariate models.

Model	r^2	$RMSE_{cv}$† (mg kg^{-1})	RPD‡	Bias (mg kg^{-1})
Partial least square regression	0.84	4791	2.50	68
Wavelet-multiple linear regression	0.94	3010	3.97	33
Penalized splines	0.98	3553	3.32	48

† $RMSE_{cv,}$ Root mean square error of cross-validation.
‡ RPD, residual prediction deviation.

use of fewer regressors, separated into different scales. Since in the present study, the neighboring channels were highly correlated, we believe that the effect of neighboring channels (through the regression coefficients) were also highly correlated (i.e. the regression coefficient curve is smooth.). It is noteworthy that the estimator from penalized spline was more stable than non-penalized method (PLSR) given that the neighboring regression coefficients were *hand-in-hand* connected, which was not true of PLSR. The order of the channels was ignored by PLSR. Summarily, the wavelet-MLR and penalized spline models reasonably predicted petroleum concentration.

4.4.3. VisNIR DRS as a Proximal Sensor for Petroleum Content

A possible explanation for the high accuracy in separating soil types could be the fact that soil particle size (soil texture) affects the transmission of light and reflectance spectra, as indicated by Chang et al. (2001). The possible reasons for the insensitivity of DRS in separating oil types in VisNIR range when contaminations were mixed could be 1) the heterogeneity and opacity of the soil matrix in addition to light scattering effects (Ko et al., 2010) and 2) crude oils contain mixtures of heavy asphaltic crudes to light crudes that are similar to a diesel fuel (Mattson et al., 1977).

The study was intended for testing the capability of VisNIR viability instead of making a lab-grade predictive model. While working with crude oil and other petroleum products, researchers encounter a number of problems. The collection of a comprehensive range of refined products and crude oil with different compositions and quality indices is not an easy task (Balabin and Safieva, 2007). Moreover, standard chemical analyses (both HPLC and

gravimetric) are costly, and time-consuming. Therefore, construction of large set of artificial samples with actual oil and soil mixture is very challenging.

Testing the heterogeneity within a range of soil physicochemical properties (more textures, organic carbon levels, and soil color) was beyond the scope of this project and requires intensive studies before drawing stronger conclusions. Auxiliary soil properties that can be measured quickly and easily may improve petroleum predictive models when used along with the soil spectra. More improvement could be achieved by increasing sample number and mapping the discrete wavelet transform regressors into a schematic, two-dimensional waveband-scale tiling for a more systematic and straightforward representation of the wavelet-based model.

4.5. Conclusion

This exploratory study utilized 68 lab constructed samples for identifying the significant effects of soil texture and organic carbon on VisNIR reflectance patterns of petroleum contaminated soils. The variable moisture effect on VisNIR reflectance spectra was offset by maintaining uniform moisture to all samples. The first nine principal components elucidated 88% of the variance in the data and plots of sample scores were satisfactory to identify the clusters by soil types and organic carbon levels. However, PCA could not separate different oil types when contaminations were mixed. Visual interpretations from PC plots were quantitatively confirmed by LDA. Support vector machine performed slightly better than random forest for classifying organic carbon levels. Subtle separations for oil types were obtained from PC plots of soil B with 5% organic carbon, indicating the need for future controlled research.

This study also elucidated the need of a reliable spectral pre-treatment as an alternative to traditional methods. Among different preprocessing and multivariate models tested, wavelet

preprocessing performed best with the highest predictability (RPD=3.97). However, while dealing with first-derivative spectra, penalized spline regression performed better than PLSR model, considering the order of the regressors. Heteroscedasticity and systematic non-linearity of residuals worsened PLSR model predictions at higher oil concentrations. More intensive research is recommended considering other soil physicochemical variability and integrating wavelet-penalized spline for VisNIR characterization of petroleum contaminated soils. Summarily, the cost-effectiveness, alacrity, and portability of this technique make it a promising tool that would give soil and/or environmental scientists the ability to characterize oil spills at a much larger scale and for a larger geographic area by utilizing a specialized spectral library focused on contaminant hydrocarbons. The goal for our future research should be to develop a general model which can lead to reliable hydrocarbon predictions under divergent soil matrix conditions.

4.6. References

Al-Abbas, H.H., P.H. Swain, and M.F. Baumgardner. 1972. Relating organic matter and clay content to the multi spectral radiance of soils. Soil Sci. 14:477-485.

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

Balabin, R.M., and R.Z. Safieva. 2007. Capabilities of near infrared spectroscopy for the determination of petroleum macromolecule content in aromatic solutions. J. Near Infrared Spectrosc. 15:343-349.

Balabin, R.M., and R.Z. Safieva. 2008. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. Fuel 87:1096-1101.

Ben-Dor, E., and A. Banin. 1990. Near-infrared reflectance analysis of carbonate concentration in soils. Appl. Spectrosc. 44:1064–1069.

Ben-Dor, E., and A. Banin. 1995. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Sci. Soc. Am. J. 59:364–372.

Boser, E., M. Guyon, and V. Vapnik. 1992. A training algorithm for optimal margin classifiers. p. 144-152. In Proc. Fifth ACM Workshop on Computational Learning Theory. 1992. Pittsburgh, PA, USA.

Bowers, S.A., and R.J. Hanks. 1965. Reflection of radiant energy from soils. Soil Sci. 100:130-138.

Breiman, L. 2001. Random forests. Machine Learning 45:5-32.

Brown, D.J., R.S. Bricklemeyer, and P.R. Miller. 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VisNIR soil C prediction in Montana. Geoderma 129:251–267.

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

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

Chang, C., and C. Lin. 2001. LIBSVM: A library for support vector machines. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm (verified 30 October, 2010).

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

Chang, C.W., D.A. Laird, and C.R. Hurburgh. 2005. Influence of soil moisture on near infrared reflectance spectroscopic measurement of soil properties. Soil Sci. 170:244–255.

Chung, H., and M. Ku. 2000. Comparison of near-infrared, infrared, and Raman spectroscopy for the analysis of heavy petroleum products. Appl. Spectrosc. 54:239-245.

Chung, H., H. Choi, and M. Ku. 1999. Rapid identification of petroleum products by near-infrared spectroscopy. Bull. Korean Chem. Soc. 20:1021-1025.

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

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

Demetriades-Shah, T.H., M.D. Steven, and J.A. Clark. 1990. High-resolution derivative spectra in remote sensing. Remote Sens. Environ. 33:55–64.

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

Donoho, D.L., and I.M. Johnstone. 1994. Ideal spatial adaptation via wavelet shrinkage. Biometrika 81:425-455.

Dumas, J.B.A. 1831. Procedes de l'analyse organique. Ann. Chim. Phys. 247:198-213.

Dunn, B.W., H.G. Beecher, G.D. Batten, and S. Ciavarella. 2002. The potential of near infrared reflectance spectroscopy for soil analysis—a case study from the Riverine Plain of south-eastern Australia. Aust. J. Exp. Agric. 42:607–614.

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

Fine, P., E.R. Graber, and B. Yaron. 1997. Soil interactions with petroleum hydrocarbons: abiotic processes. Soil Tech. 10:133–153.

Forrester, S., L. Janik, and M. McLaughlin. 2010. An infrared spectroscopic test for total petroleum hydrocarbon (TPH) contamination in soils. Proc. 19th World Congress of Soil Science. 1-6 August. 2010. Brisbane, Australia.

Gee, G.W., and D. Or. 2002. Particle-size analysis. p. 255–293. *In* J.H. Dane and G.C. Topp (ed.) Methods of soil analysis. Part 4. SSSA Book Ser. 5. SSSA, Madison, WI.

Ge, Y., C.L.S. Morgan, J.A. Thomasson, and T. Waiser. 2007. A new perspective to near infrared reflectance spectroscopy: A wavelet approach. Trans. ASABE 50:303–311.

Graham, K.N. 1998. Evaluation of analytical methodologies for diesel fuel contaminants in soil. M.Sc. Thesis. University of Manitoba, Winnipeg, MB, Canada.

Hoerig, B., F. Kuehn, F. Oschuetz, and F. Lehmann. 2001. HyMap hyperspectral remote sensing to detect hydrocarbons. Int. J. Remote Sensing 8:1413-1422.

Hunt, G.R. 1982. Spectroscopic properties of rocks and minerals. p. 295–385. *In* R.S. Carmichael (ed.) Handbook of physical properties of rocks. Vol. 1. CRC Press, Boca Raton, FL.

Hunt, G.R., and J.W. Salisbury. 1970. Visible and near-infrared spectra of minerals and rocks: I. Silicate minerals. Mod. Geol. 1:283–300.

Hyvarinen, T., E. Herrala, J. Malinen, and P. Niemla. 1992. NIR analysers can be miniature, rugged and handheld. p. 1-6. *In* K. Hildrum, T. Isaksson, T. Naes, and A. Tandberg (ed.) Near infrared spectroscopy. Bridging the gap between data analysis and NIR applications. Ellis Horwood, London.

Islam, K., B. Singh, and A. McBratney. 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and nearinfrared reflectance spectroscopy. Aust. J. Soil Res. 41:1101– 1114.

Kilmer, V.H., and L.Z. Alexander. 1949. Methods for making mechanical analyses of soil. Soil Sci. 68:15–24.

Ko, E.J., K.W. Kim, K. Park, J.Y. Kim, J. Kim, S.Y. Hamm, J.H. Lee, and U. Wachsmuth. 2010. Spectroscopic interpretation of PAH-spectra in minerals and its possible application in soil monitoring. Sensors 10:3868-3881.

Kusumo, B.H., C.B. Hedley, M.J. Hedley, A. Hueni, M.P. Tuohy, and G.C. Arnold. 2008. The use of diffuse reflectance spectroscopy for in situ carbon and nitrogen analysis of pastoral soils. Aust. J. Soil Res. 46:623–635.

Lark, R.M., and R. Webster. 1999. Analysis and elucidation of soil variation using wavelets. Eur. J. Soil Sci. 50:185-206.

Lee, S.W., J.F. Sanchez, R.S. Mylavarapu, and J.S. Choe. 2003. Estimating chemical properties of Florida soils using spectral reflectance. Trans. ASAE 46:1443–1453.

Mallat, S.G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. PAMI 11:674–693.

Malle, H., and P. Fowlie. 1998. A Canadian interlaboratory comparison for analysis of petroleum hydrocarbons in soil. p. 321-322. *In* Proc. Second Biennial International Conference on Chemical Measurement and Monitoring of the Environment, EnviroAnalysis'98 Conference, 11-14 May 1998. Ottawa, Canada.

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

Malley, D.F., L. Yesmin, and R.G. Eilers. 2002. Rapid analysis of hog manure and manure - amended soils using near-infrared spectroscopy. Soil Sci. Soc. Am. J. 66:1677–1686.

Mattson, J.S., C.S. Mattson, M.J. Spencer, and F.W. Spencer. 1977. Classification of petroleum pollutants by linear discriminant function analysis of infrared spectral patterns. Anal. Chem. 49:500-502.

Morgan, C.L.S., T.H. Waiser, D.J. Brown, and C.T. Hallmark. 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. Geoderma 151:249-256.

Prince, R. C. 1993. Petroleum spill bioremediation in marine environments. Crit. Rev. Microbiol. 19:217-242.

R Development Core Team. 2008. R: A language and environment for statistical computing. Available at http://www.R-project.org (verified 30 October 2010). R Foundation for Statistical Computing, Vienna, Austria.

Reeves, J.B., G.W. McCarty, and J.J. Meisinger. 2000. Near infrared reflectance spectroscopy for the determination of biological activity in agricultural soils. J. Near Infrared Spectrosc. 8:161– 170.

Russell, A.E., D.A. Laird, T.B. Parkin, and A.P. Mallarino. 2005. Impact of nitrogen fertilization and cropping system on carbon sequestration in Midwestern mollisols. Soil Sci. Soc. Am. J. 69:413–422.

Shepherd, K.D., and M.G. Walsh. 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66:988–998.

Sherrod, L.A., G. Dunn, G.A. Peterson, and R.L. Kolberg. 2002. Inorganic C analysis by modified pressure-calcimeter method. Soil Sci. Soc. Am. J. 66:299–305.

Soil Survey Staff. 2004. Soil survey laboratory methods manual (version 4.0). USDA-NRCS. US Gov. Print. Off. Washington, DC.

Soil Survey Staff. 2005. Official soil series descriptions. Available at soils.usda.gov/technical/classification/osd/index.html (verified 11 Oct. 2010). NRCS, Washington, DC.

Steele, J.G., and R. Bradfield. 1934. The significance of size distribution in the clay fraction. Bull. Am. Soil Surv. Assoc. 15:88–93.

Stum, A.K. 2010. Random forests applied as a soil spatial predictive model in arid Utah. M.S. thesis Utah State Univ., Logan, USA.

Thompson, W.D., and S.D. Walter. 1988. A reappraisal of the kappa coefficient. J. Clin. Epidemiol. 41:949–958.

Vapnik, V. 1995. The nature of statistical learning theory. Springer, NY.

Vasques, G.M., S. Grunwald, and J.O. Sickman. 2009. Modeling of soil organic carbon fractions using visible-near-infrared spectroscopy. Soil Sci. Soc. Am. J. 73:176-184.

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

Viscarra Rossel RA, Lark RM (2010) Improved modelling of soil diffuse reflectance spectra using wavelets. Eur. J. Soil Sci. 60:453–464

Waiser, T.H., C.L.S. Morgan, D.J. Brown, and C.T. Hallmark. 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Sci. Soc. Am. J. 71:389–396.

Wang, Z., A.C. Chang, and W.D. Crowley. 2003. Assessing the soil quality of long-term reclaimed wastewater-irrigated cropland. Geoderma 114:261–278.

Westbrook, S. R. 1993. Army use of near-infrared spectroscopy to estimate selected properties of compression ignition fuels. *In* Proc. SAE International Congress and Exposition. 1–5 March. 1993. Detroit, Michigan, USA.

Wetzel, D.L. 1983. Near-infrared reflectance analysis: Sleeper among spectroscopic techniques. Anal. Chem. 55:1165A-1176A.

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

Workman, J.J. 1996. Interpretive spectroscopy for near infrared. Appl. Spec. Rev. 31:251–320.

Yoon, J., B. Lee, and C. Han. 2002. Calibration transfer of near-infrared spectra based on compression of wavelet coefficients. Chemom. Intell. Lab. Syst. 64:1–14.

CHAPTER 5

CONCLUSION

This pilot research showed that visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) is capable of predicting soil TPH level without prior sample preparation. Variable soil moisture content and soil heterogeneity did not influence the prediction accuracies of the VisNIR DRS- based PLSR predictive model. Drying and grinding of soil samples were responsible for soil TPH loss, and resulted in low prediction accuracy of the PLSR model. However, sample number was clearly an issue and indicated the need for more sample for good generalizing capacity of the non-parametric BRT model. A fair RPD value of 1.70 indicated the scope for future model improvement. Also, this research recommended the combined use of VisNIR prediction and geostatistics to potentially to identify the spatial patterns of TPH contamination in soil quickly on site, reducing the need for expensive laboratory analyses. While utilizing TPH predictions from a penalized spline model in combination with ordinary block kriging, TPH distribution satisfactorily matched the topography of the study site.

Another exploratory study utilized lab constructed samples for identifying the significant effects of soil types and organic carbon on VisNIR reflectance patterns of petroleum contaminated soils. High accuracy in separating soil types was achieved for soil particle size (soil texture) effects. Heterogeneity, opacity, and light scattering of soil matrix were responsible for the insensitivity of DRS in separating oil types in the VisNIR range when contaminations were mixed. Subtle separations for oil types were obtained from PC plots for a specific soil type and specific organic carbon content, indicating the need for future controlled research. This study also elucidated the value of wavelet analysis (RPD=3.97) as an alternative to traditional spectral

preprocessing methods. However, while dealing with first-derivative spectra, penalized spline regression performed better than the non-parametric PLSR model.

Although, the full abilities of VisNIR spectroscopy have not been fully realized to date and its utilization in the oil spill research is remains novel, this present research identified the needs for:

1. More intensive research with a larger sample set;
2. Samples representing larger geographic ranges, with a wider assortment of soil properties;
3. Integrating hybrid regression kriging for sound spatial variability mapping of soil TPH;
4. Improvement of non-parametric modeling for handling small datasets;
5. Integrating wavelet-penalized spline for VisNIR characterization of petroleum contaminated soils; and
6. Labeling specific wavelengths associated to various petroleum signatures for future remote sensing applications of VisNIR DRS based sensors.

Summarily, the cost-effectiveness, quickness, and transportability of this proximal sensor method make it an encouraging tool that could provide environmental scientists the capability to characterize inland oil spills at a much larger scale, with high spatial resolution, and for a larger geographic area by utilizing a specialized spectral library focused on contaminant hydrocarbons.

APPENDIX A

LABORATORY SAMPLE CONSTRUCTION SCHEME

Table 9. Laboratory sample construction scheme.

Soil A										
Organic Carbon	Diesel			Crude oil			Motor oil			Control
	Ppm									
	1000	10000	30000	1000	10000	30000	1000	10000	30000	
1 %	1	1	1	1	1	1	1	1	1	1
5 %	1	1	1	1	1	1	1	1	1	1
10 %	1	1	1	1	1	1	1	1	1	1

Soil B										
Organic Carbon	Diesel			Crude oil			Motor oil			Control
	Ppm									
	1000	10000	30000	1000	10000	30000	1000	10000	30000	
1 %	1	1	1	1	1	1	1	1	1	1
5 %	1	1	1		11†		1	1	1	1
10 %	1	1	1	1	1	1	1	1	1	1

† 2 samples (1000 and 30000 ppm) + 9 intermediate samples from (4000-28000 ppm at each 3000 ppm interval).

APPENDIX B

AVERAGE REFLECTANCE SPECTRA

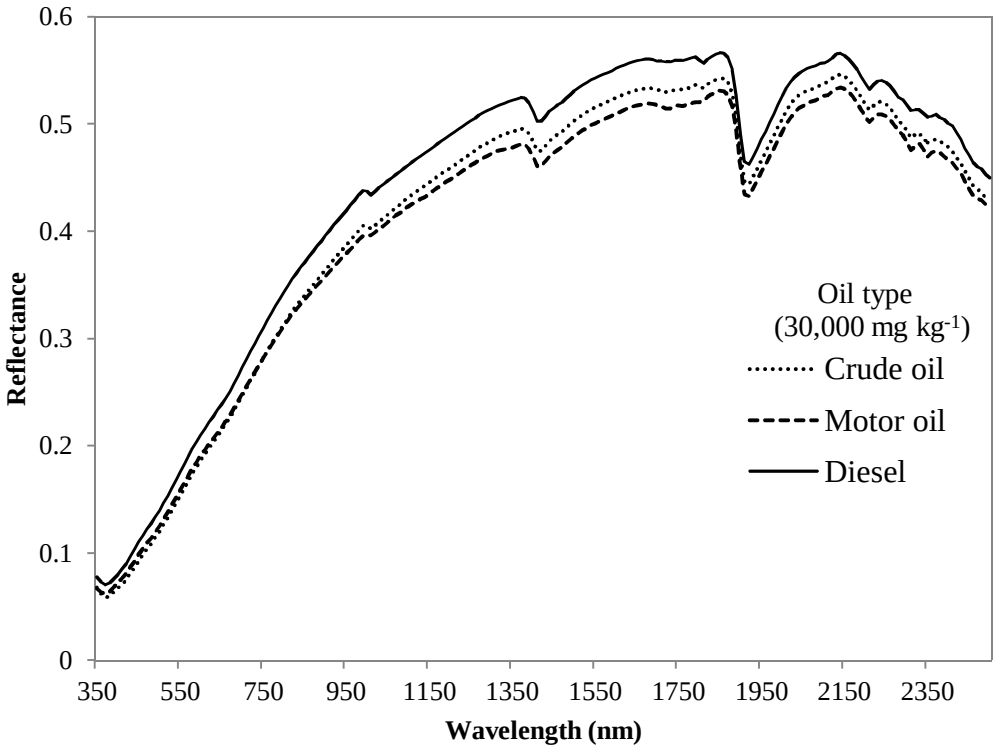

Fig 25. Average reflectance spectra are shown for Soil A from Louisiana, USA with 10% organic carbon and different types of oils (in 30,000 ppm or mg kg^{-1}).

APPENDIX C

PRINCIPAL COMPONENT REGRESSION

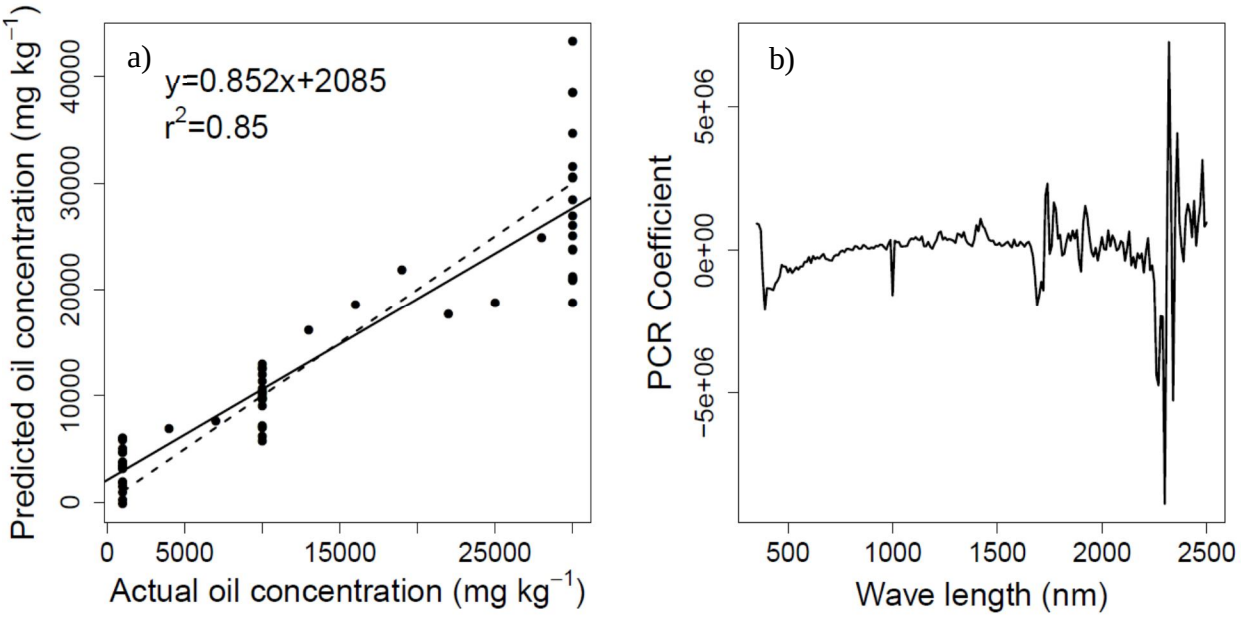

Fig. 26. (a) The actual versus predicted oil concentration (mg kg^{-1}) using principal component regression (PCR) for soils from Louisiana, USA. The solid line is the fitted linear regression, while the dashed line is the 1:1 line and (b) Regression coefficients for each 10-nm interval.

APPENDIX D

PRINCIPAL COMPONENTS

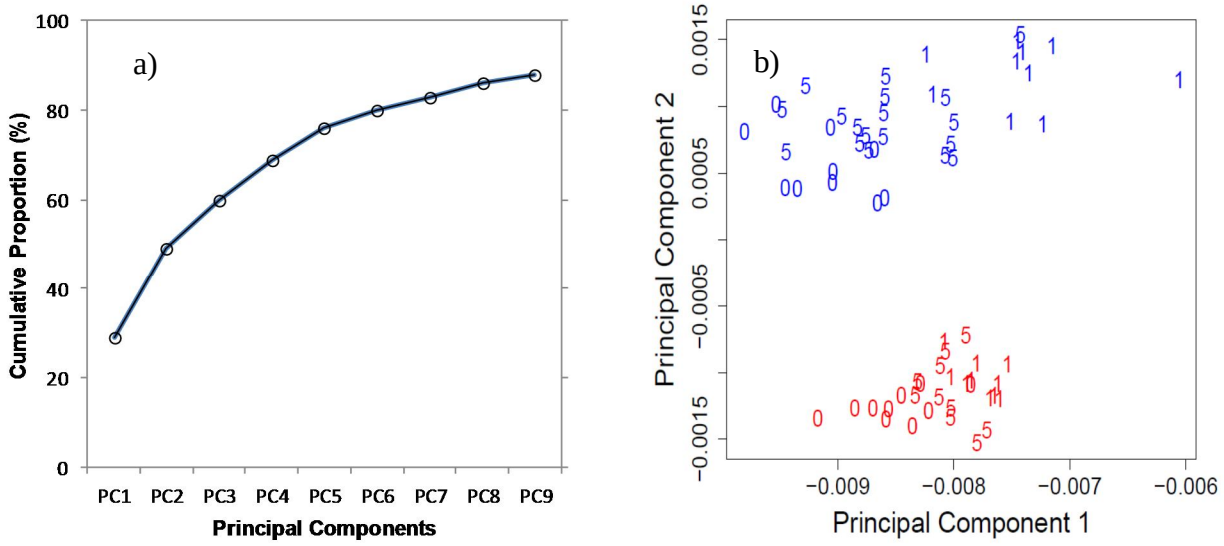

Fig. 27. (a) The cumulative proportion of variance explained by the first nine principal components of first-derivative of the reflectance spectra for the soils evaluated for petroleum contamination using visible and near infrared diffuse reflectance spectroscopy from Louisiana, USA, and (b) A scatter plot of the first two principal component score. Soil A is red; Soil B is blue and organic matter is 1, 5, and 0 for 1, 5, and 10%, respectively.

APPENDIX E

PAIRWISE SCORE PLOTS FOR SOILS

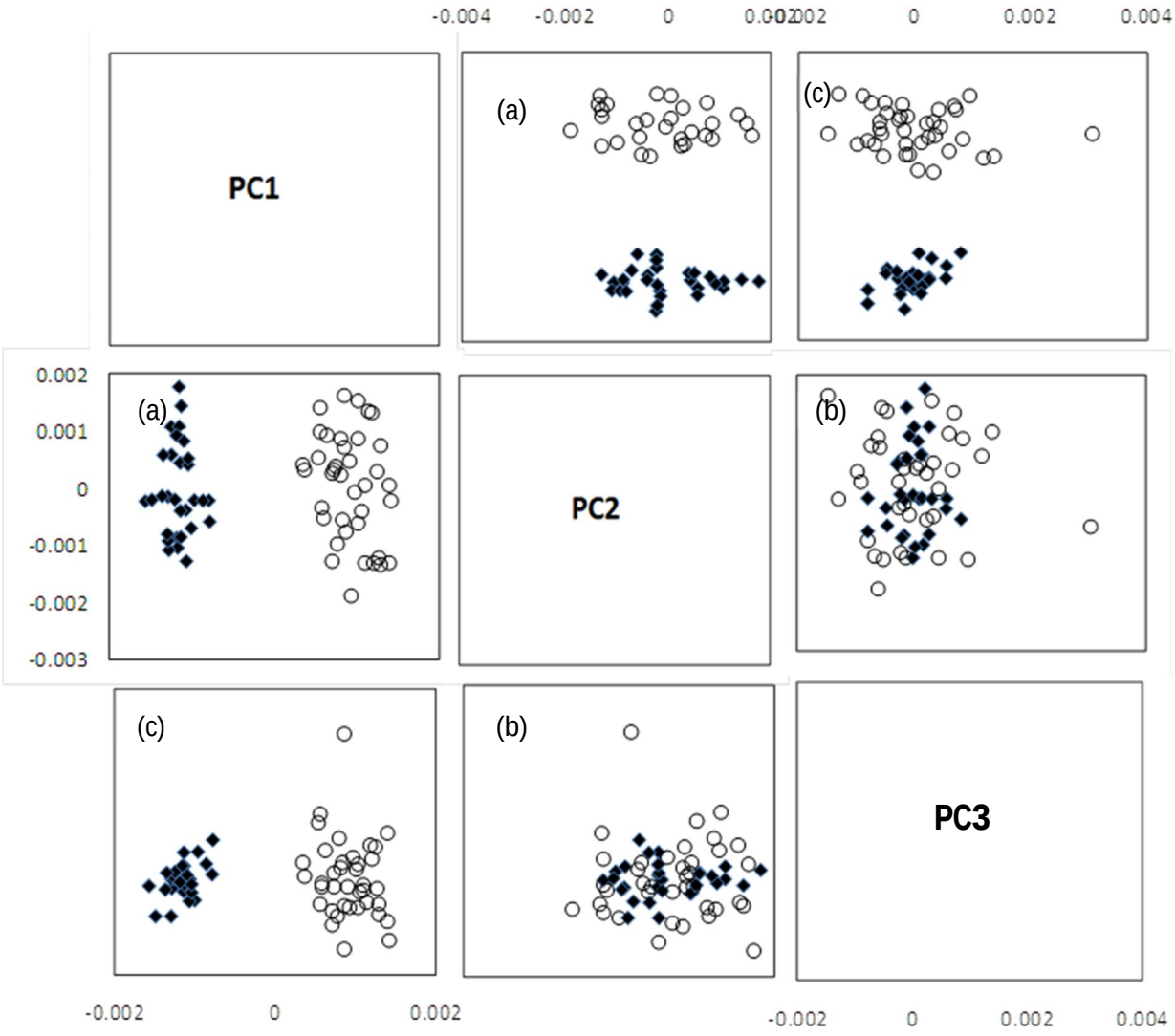

Fig. 28. Pairwise principal component (PC) plots for (a) PC1 vs. PC2, (b) PC2 vs. PC3, and (c) PC1 vs. PC3 of the first-derivative of spectral reflectance for soils evaluated for petroleum contamination using visible and near-infrared diffuse reflectance spectroscopy from Louisiana, USA. The solid diamonds and open circles represent soil A and soil B, respectively.

APPENDIX F

PAIRWISE SCORE PLOTS FOR ORGANIC CARBON

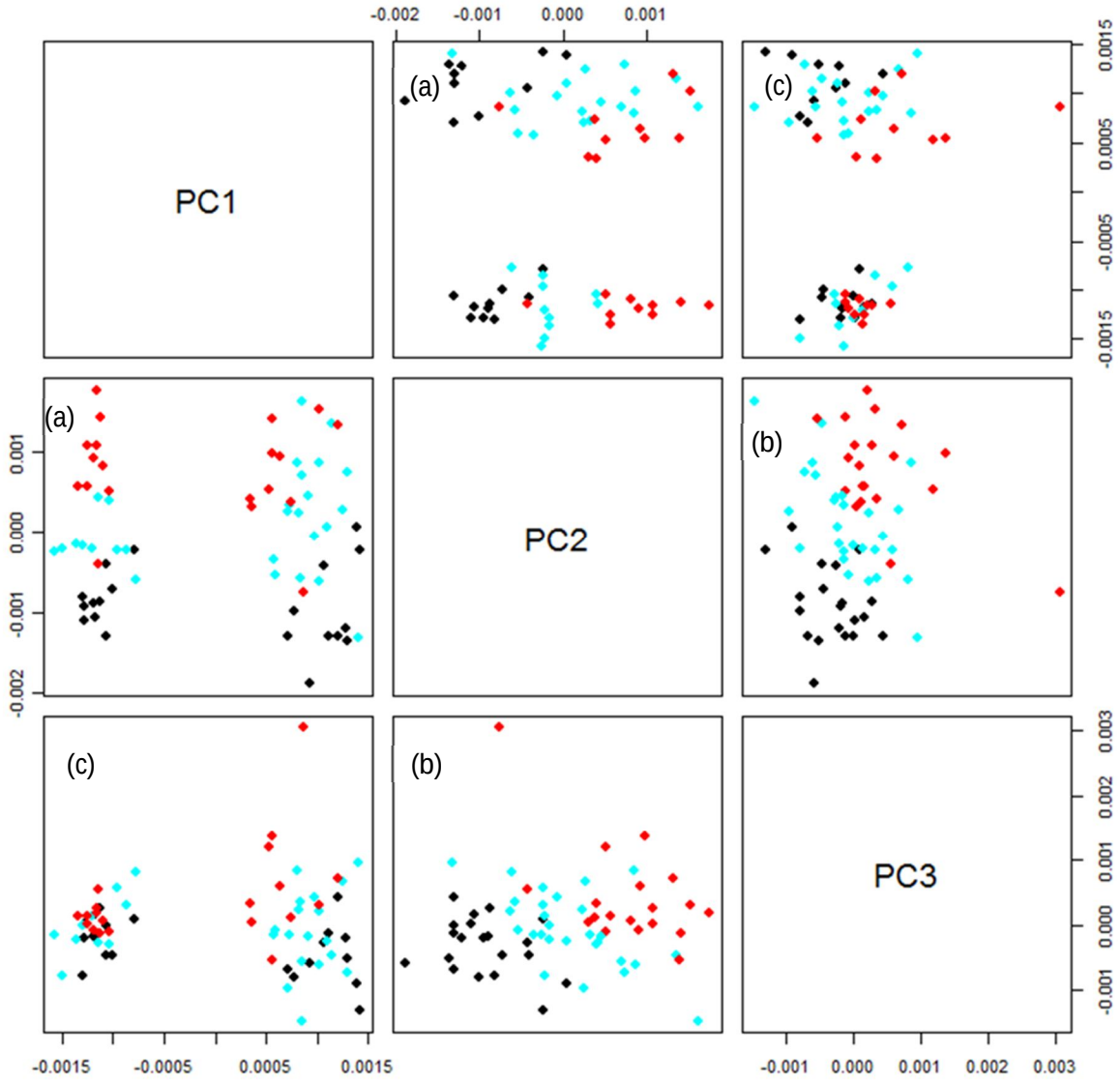

Fig. 29. Pairwise principal component (PC) plots for (a) PC1 vs. PC2, (b) PC2 vs. PC3, and (c) PC1 vs. PC3 using the first-derivative of soil reflectance spectra and evaluated for petroleum contamination using visible and near-infrared diffuse reflectance spectroscopy. The black, blue, and red diamonds represent soils with 1%, 5%, and 10% organic carbon, respectively.

APPENDIX G

PERMISSION TO REPRINT

FW: Request to Reprint Content Thursday, February 17, 2011 11:02 PM

From: "Ann Edahl" <aedahl@sciencesocieties.org>

To: "som_pau@yahoo.com" <som_pau@yahoo.com>

Dear Somsubhra Chakraborty,

Permission is hereby given to reproduce in your dissertation the JEQ article as outlined in your message below.

Please add a permission line, Used with permission, from Journal of Environmental Quality 39:1378-1387 (2010).

Sincerely,

Ann Edahl
Managing Editor
Journal of Environmental Quality
American Society of Agronomy * Crop Science Society of America * Soil Science Society of America
5585 Guilford Road
Madison, WI 53711

FAX: 608-273-2021
www.agronomy.org * www.crops.org * www.soils.org
"Fundamental for Life: Soil, Crop, & Environmental Sciences"
ASA, CSSA, and SSSA 2011 International Annual Meetings
SSSA 75th Anniversary | 1936-2011
October 16-19 | San Antonio, TX
www.acsmeetings.org

-----Original Message-----
From: SOMSUBHRA CHAKRABORTY [mailto:som_pau@yahoo.com]
Sent: Tuesday, February 15, 2011 12:40 PM
To: Meg Ipsen
Subject: Request to Reprint Content

VITA

Somsubhra Chakraborty was born in 1984, in West Bengal, India. He attended Bidhan Chandra Krishi Viswavidyalaya, India, and graduated in 2006 with a Bachelor of Science in Agriculture. After that, he went to Punjab Agricultural University, Ludhiana, India, and graduated in 2008 with Master of Science in Soil Science.

In August 2008 he was admitted into the doctoral program in the School of Plant, Environmental & Soil Sciences at Louisiana State University and Agricultural and Mechanical College. The title of his dissertation is — "Rapid Identification of Oil Contaminated Soils Using Visible Near Infrared Diffuse Reflectance Spectroscopy."

.