Rapid Identification of Oil-Contaminated Soils Using Visible Near-Infrared Diffuse Reflectance Spectroscopy

Somsubhra Chakraborty and David C. Weindorf* Louisiana State University AgCenter
Cristine L.S. Morgan and Yufeng Ge Texas Agrilife Research
John M. Galbraith Virginia Tech
Bin Li and Charanjit S. Kahlon Louisiana State University

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, 350–2500 nm) diffuse reflectance spectroscopy (DRS) is a rapid, nondestructive, 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: (i) field-moist intact aggregates, (ii) air-dried intact aggregates, (iii) 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 was promising for rapidly predicting soil petroleum hydrocarbon content, future research is warranted to evaluate the methodology for identifying petroleum contaminated soils.

Although 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 on 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 (2298 nm [stretch+bend]; 1725 nm [two-stretch]; 1388 nm [two-stretch+bend]; 1190 nm [three-stretch]; 1020 nm [three-stretch+bend]; 917 nm [four-stretch]). Chung et al. (1999) reported near-infrared prediction accuracies of 95% (for light gas oil) and 99% (for light straight-run which is a low boiling range and

J. Environ. Qual. 39:1378–1387 (2010)
doi:10.2134/jeq2010.0183
Published online 3 June 2010.
Freely available online through the author-supported open-access option.
Received 20 Apr. 2010.
*Corresponding author (dweindorf@agcenter.lsu.edu).

5585 Guilford Rd., Madison, WI 53711 USA

S. Chakraborty, 301 M.B. Sturgis Hall, Louisiana State Univ. AgCenter, Baton Rouge, LA 70803; D.C. Weindorf, 307 M.B. Sturgis Hall, Louisiana State Univ. AgCenter, Baton Rouge, LA 70803; C.L.S. Morgan, 545 Heep Center, Texas Agrilife Research, College Station, TX 77843; Y. Ge, 545 Heep Center, Texas Agrilife Research, College Station, TX 77843; J.M. Galbraith, Dep. of Crop and Soil Environmental Sciences, Virginia Tech, Blacksburg, VA 24061; B. Li, 61 Agric. Administration Building, Louisiana State Univ., Baton Rouge, LA 70803; C.S. Kahlon, Louisiana State Univ., Baton Rouge, LA 70803.
Assigned to Associate Editor Keith Goyne.

Abbreviations: ASD, Analytical Spectral Devices; BRT, boosted regression trees; DRS, diffuse reflectance spectroscopy; LDA, linear discriminant analysis; MART, multiple additive regression trees; PLS, partial least squares; PC, principal component; PCA, principal component analysis; $RMSE_{cv}$, root mean square error of cross-validation; $RMSE_p$, root mean square error of prediction; RPD, relative percent difference; TPH, total petroleum hydrocarbon; VisNIR, visible near-infrared.

low carbon number [C_5–C_7] petroleum product, kerosene, gasoline, and diesel), whereas principal component analysis (PCA) combined with Mahalanobis distance could be used to segregate unique spectral signatures for each of the aforementioned petroleum products. 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 such as 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 performed 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 nondestructively (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,b; 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). To date, few studies have reported on the use of VisNIR DRS to characterize 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, tarlike 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 laboratory method (Clesceri et al., 1998) is time consuming and costly (~US$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 laboratory 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 before rebuilding. Furthermore, contamination might be recognized in areas where it may have gone undetected. Hence, the specific objectives of this research were the following: (i) to determine the prediction accuracy of VisNIR DRS in quantifying the amount of hydrocarbons in contaminated soils and (ii) to compare the accuracies of partial least squares regression and boosted regression trees in predicting TPH in contaminated soils.

Materials and Methods

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 the 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 laboratory and refrigerated at 5°C in the laboratory. The official soil series description of each sampling site showed a wide variation in soil properties between sites (Table 1).

VisNIR DRS Scanning

The collected soil samples were scanned with an AgriSpec VisNIR portable spectroradiometer (Analytical Spectral Devices, Boulder, CO) 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 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. For field-moist scanning of intact aggregates, 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 s. 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

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

Site	Parish	Soil series	Classification†	Contaminated samples	Noncontaminated 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	NA‡	Udarents	5	5

† Soil Survey Staff (2005).

‡ NA, not applicable.

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 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 AgriSpec. Between each of the four scans, the sample was rotated 90°.

Laboratory Analysis

In the commercial laboratory, 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). 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 min and filtered through a filter paper premoistened with solvent and collected in a flask. The silica gel and filter paper were washed with 10 mL solvent, and combined with filtrate. Solvent was recovered by distillation from flask in a water bath at 85°C. The flask was cooled in a desiccator for at least 30 min and weighed.

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 CIROS CCD (SPECTRO Analytical Instruments Inc, NJ). X-ray diffraction analysis was conducted for bulk soil mineralogy confirmation (Whittig 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). 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.

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, (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.

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 ($\lambda = 1$) was $\log_{10}$–transformed ($\lambda = 0$) to make it more normal (Fig. 1). 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

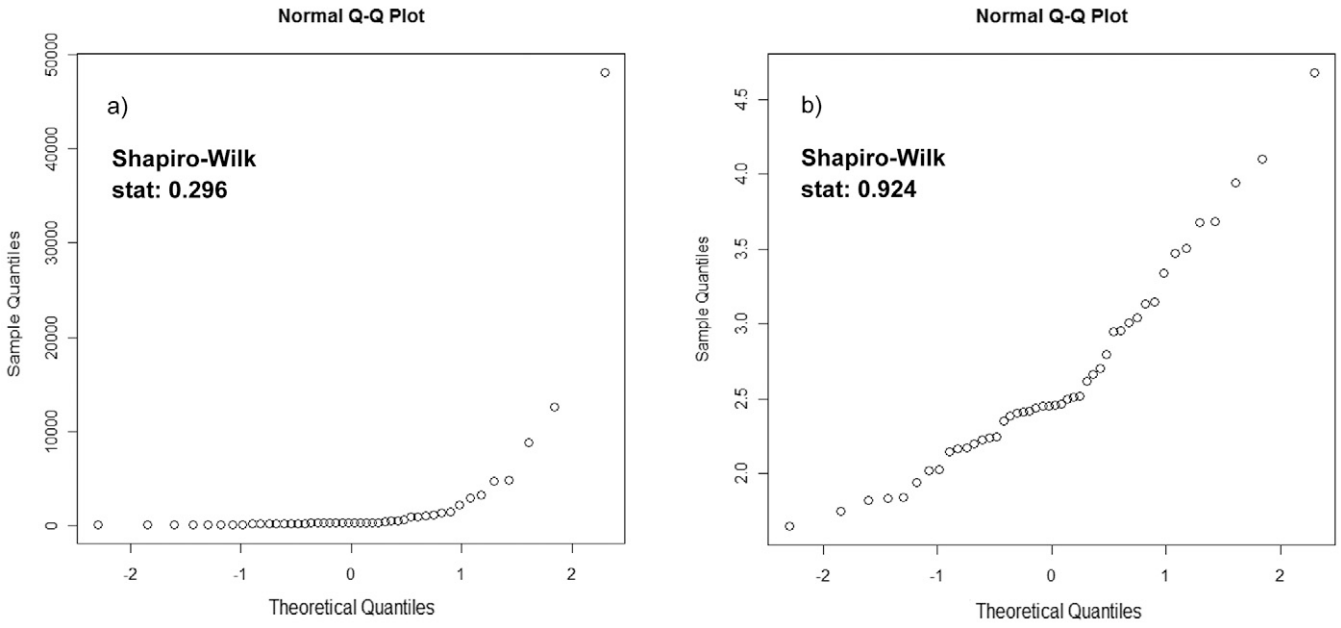

Fig. 1. (a) Original (λ = 1) and (b) log-transformed (λ = 0) total petroleum hydrocarbon contents of the soil samples collected from six different parishes in Louisiana. Increase in Shapiro–Wilk statistic for log-transformed data revealed that the Box-Cox transformation normalized the original TPH data.

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 10 factors 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.

Boosted Regression Tree Analysis

Following Friedman's Gradient Boosting Machine (Friedman, 2001), boosted regression tree (BRT) analysis was also used. 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 performed by Treenet 2.0 (Steinberg et al., 2002) (Salford Systems, San Diego, CA), a multiple additive regression trees (MART)–based program. The same datasets used in PLS (70% calibration, 30% validation) were used for BRT 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 used. Initially, the maximum number of trees to be grown was set to 200. The number of trees was increased (>200) manually in two conditions: (i) up to a point when $RMSE_p$ value stopped decreasing and (ii) when the optimal number of trees was close enough to the maximum numbers of trees specified beforehand.

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 noncontaminated 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 three PCs were produced to generate the ideas on how contaminated and reference soils were separated from each other in the spectral space.

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}) as well as 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}. 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 (7758 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).

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, whereas 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 did the field-moist intact reflectance model, the main advantage of the first derivative was fewer latent factors (five) compared with the reflectance-based model (eight latent factors) to prevent overfitting. Results indicated a continuous reduction of latent factors as the model changed from reflectance to first and second derivatives. This reduction of principal components (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 by incorporating the separate validation sets where only the reflectance and first derivative were taken into consideration (Fig. 2). According to Chang et al.

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.

Site	pH	Minerals					Clay	Organic matter
		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.

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 on very low calibration r^2 (<0.15).

(2001), accuracy and stability of spectroscopic models should be based on their RPD statistics. Stable and accurate predictive models showed a RPD > 2.0, fair models with potential for prediction improvement had a RPD value between 1.4 and 2.0, whereas 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 although 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 the 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 the reflectance and first-derivative models). Additionally, the air-dried intact reflectance model showed the highest RPD (1.94).

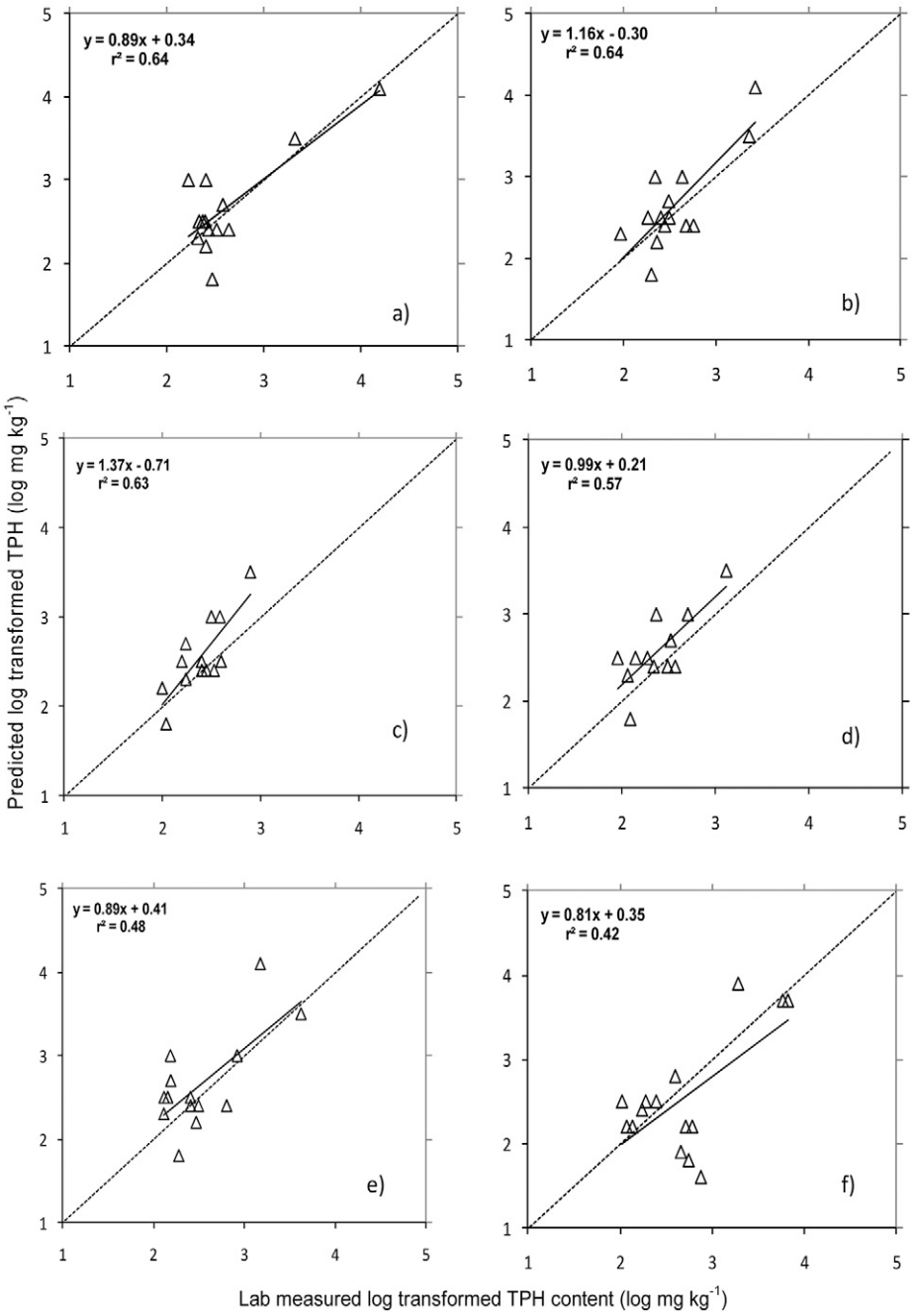

Fig. 2. Predicted vs. measured total petroleum hydrocarbon (TPH) content of the validation data set for (a) field-moist intact reflectance, b) field-moist intact first derivative, (c) air-dried intact reflectance, (d) air-dried intact first derivative, (e) air-dried ground reflectance, and (f) air-dried ground first derivative models for soils from Louisiana. The triangles represent validation samples. The dashed line and dark line represent the 1:1 line and the prediction trend, respectively.

Boosted Regression Tree Analysis

Model statistics for the BRT analysis, summarized in Table 4, showed much higher validation $RMSE_p$ compared with 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 nonlinear correlations, 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 did the reflectance model (0.38), whereas in the air-dried intact scan, the BRT reflectance model did not perform 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 the case of the field-moist intact scan, the reflectance model exhibited more optimum trees (507) compared with the first-derivative model (500). Given that tree-based models require large datasets for robust model predictions (Vasques et al.,

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.

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 on very low calibration r^2 (<0.15).

2009), the small dataset was most probably the crucial factor for BRT underperformance as compared to PLS models.

Underperformances of Air-Dried Models

Results indicated that for the PLS and BRT models (or techniques), the first-derivative model of the field-moist intact scan outperformed the air-dried intact and air-dried ground models, respectively (except in the air-dried intact reflectance PLS model). Soil reflectance is an integrated property that 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 greater surface area for absorption.

To study the possible reasons for the air-dried model's underperformance, air-dried intact subsamples (10 samples) from the whole range of samples (46) were carefully selected so that each would represent a specific range of TPH. These subsamples were further analyzed for TPH in the same commercial laboratory 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, compared with the primary TPH contents as a result of drying (Fig. 3). 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).

In Fig. 4, the regression coefficients (black bars) 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. Notably, the number and intensities of significant wavelengths changed in air-dried intact and air-dried ground scans compared with field-moist intact scans. The change in numbers and intensities were apparent, specifically in the 1600- to 1850- and ~2250- to 2350-nm regions, which could contain the 1725-nm (two-stretch) and 2298-nm (stretch+bend) crude oil spectral signatures as reported by Mullins et al. (1992). Moreover, typical 1450- and 1940-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 may 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

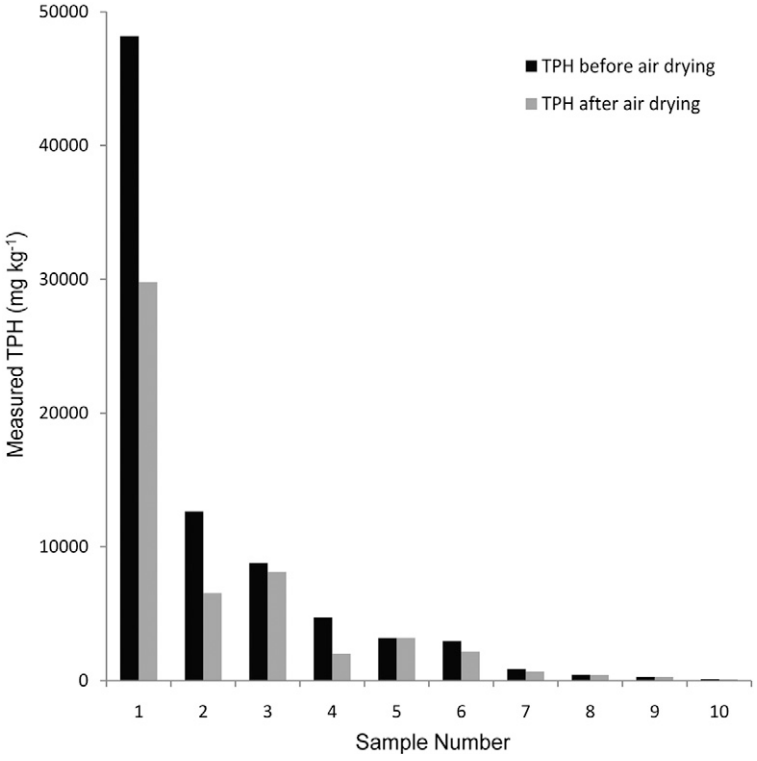

Fig. 3. Total petroleum hydrocarbon (TPH) contents (mg kg^{-1}) of 10 selected subsamples for soils from Louisiana. The black bars and gray bars represent TPH contents of subsamples before and after air drying, respectively.

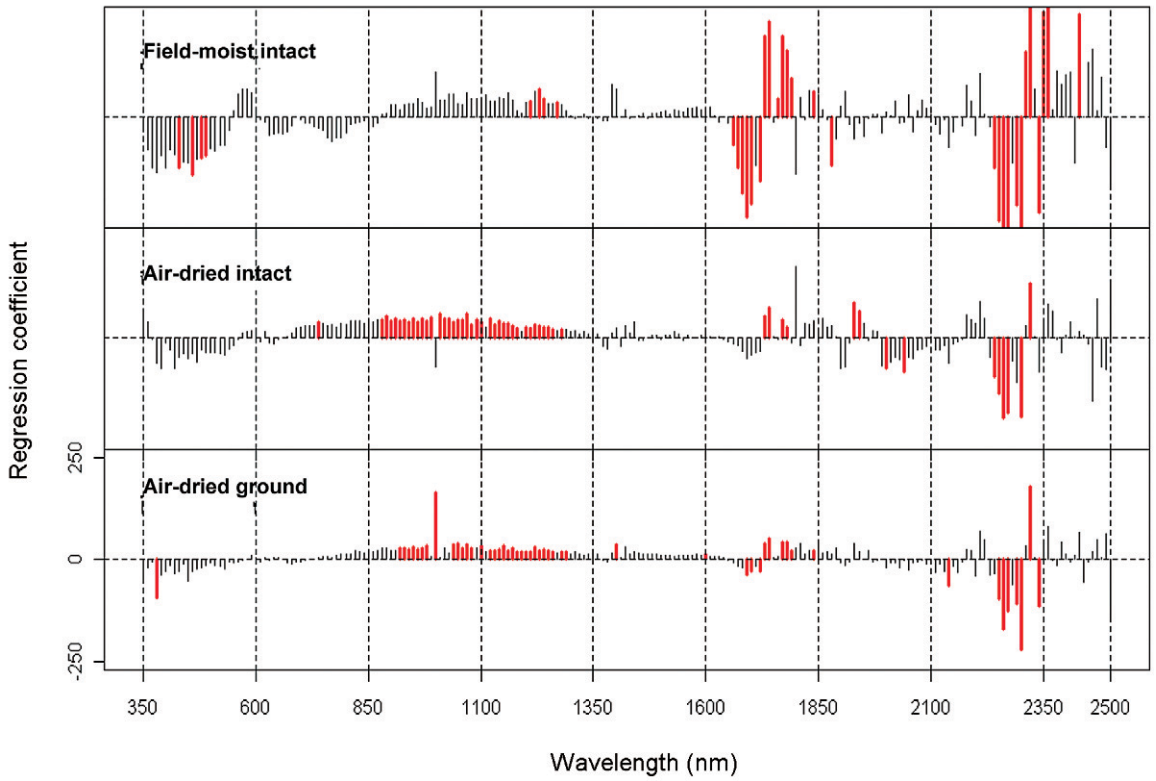

Fig. 4. Regression coefficients (black) of the first-derivative partial least squares model of each visible and near-infrared diffuse reflectance spectroscopy scan of contaminated soils from Louisiana. The magnitude of the regression coefficient at each wavelength is proportional to the height of the 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.

models. Notably, due to random loss of TPH on 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).

Principal Component Analysis

The "Screeplot" of the first 15 PCs of field-moist intact first derivative spectra was plotted in Fig. 5. 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 three PCs accounted for 88% of the total variance. It was obvious that the selection of three or six PC scores for LDA were appropriate considering their percent variance in each PC.

Pairwise 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. 6. The circles and squares represent contaminated and noncontaminated samples, respectively.

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

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

Conclusions

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 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 note-

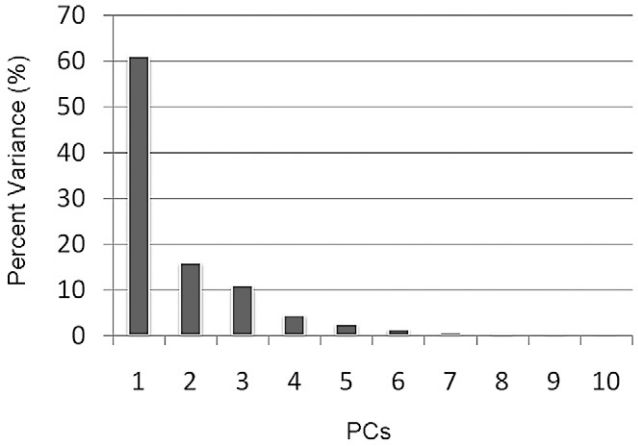

Fig. 5. "Screeplot" of the first 15 principal components (PCs) of field-moist intact first derivative spectra of soils evaluated for petroleum contamination using visible and near-infrared diffuse reflectance spectroscopy from Louisiana.

worthy 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 such as 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.

Acknowledgments

The authors wish to gratefully acknowledge financial assistance from the Louisiana Applied Oil Spill Research Program (LAOSRP). The location of suitable sampling sites was provided by the Louisiana Oil Spill Coordinator's Office (LOSCO). X-ray diffraction analysis of soil samples was carried out with the help of the Louisiana State University Department of Geology.

References

Analytical Spectral Devices. 2007. Application note 1016.01E. Available at http://www.safeco.ir/en/documents/4.pdf (verified 1 June 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.

Boffetta, P., N. Jourenkova, and P. Gustavson. 1997. Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons. Cancer Causes Control 8:444–472.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. J. Roy. Statist. Soc. Ser. B. Methodological 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

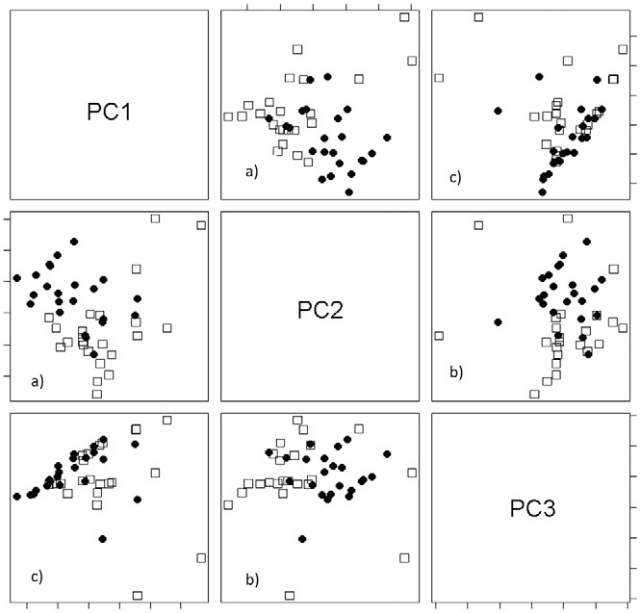

Fig. 6. Pairwise principle component (PC) plots for (a) PC1 vs. PC2, (b) PC2 vs. PC3, and (c) PC1 vs. PC3 of field-moist intact first-derivative spectra of soils evaluated for petroleum contamination using visible and near-infrared diffuse reflectance spectroscopy from Louisiana. The circles and squares represent contaminated and noncontaminated samples, respectively.

Table 5. Classification result of contaminated versus noncontaminated soils using the Fisher's linear discriminant analysis method for soils from Louisiana. 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	Noncontaminated	Sum	Contaminated	Noncontaminated	Sum
From group	Contaminated	16	7	23	20	3	23
	Noncontaminated	4	19	23	1	22	23
Sum		20	26	46	21	25	46
Overall accuracy				76%			91%

laboratory. Deep Sea Drilling Project 28:999–1007, doi:10.2973/dsdp.proc.28.app4.1975.

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. 33:55–64.

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. John Wiley & Sons, 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. Regul. Toxicol. Pharmacol. 24:85–101.

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.

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, 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., 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: Proc. of the 9th Int. Conf.. NIR Publications, Chichester, UK.

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

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.

Mullins, O.C., S. Mitra-Kirtley, and Y. Zhu. 1992. The electronic absorption edge of petroleum. Appl. Spectrosc. 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 http://www.r-project.org (verified 1 June 2010). R Foundation for Statistical Computing, 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/CD/PDF/earsel-PROCEEDINGS/3080%20Schwartz.pdf (verified 1 June 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. U.S. Gov. Print. Office, Washington, DC.

Soil Survey Staff. 2005. Official soil series descriptions. Available at http://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. Spectrosc. 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., R.N. McGlynn, and A.B. McBratney. 2006a. Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. Geoderma 137:70–82.

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

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. p. 331–362. *In* A. Klute (ed.) Methods of soil analysis. Part 1. Physical and mineralogical methods. 2nd ed. ASA and SSSA, Madison, WI.