Geoderma 177–178 (2012) 80–89

Contents lists available at SciVerse ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Spectral reflectance variability from soil physicochemical properties in oil contaminated soils

Somsubhra Chakraborty [e,1], David C. Weindorf [a,*], Yuanda Zhu [a,2], Bin Li [b,3], Cristine L.S. Morgan [c], Yufeng Ge [c], John Galbraith [d]

[a] Louisiana State University Agricultural Center, Baton Rouge, LA, United States
[b] Department of Experimental Statistics, Louisiana State University, LA, United States
[c] Texas Agrilife Research, College Station, TX, United States
[d] Department of Crop and Soil Environmental Sciences, Virginia Tech, Blacksburg, VA, United States
[e] IRDM Faculty Center, Ramakrishna Mission Vivekananda University, Kolkata 700103, India

ARTICLE INFO

Article history:
Received 13 July 2011
Received in revised form 10 January 2012
Accepted 15 January 2012
Available online 14 March 2012

Keywords:
Diffuse reflectance spectroscopy
Partial least squares regression
Penalized spline
Petroleum hydrocarbon
Visible near-infrared
Wavelet

ABSTRACT

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 clustering 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 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 clustering 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.

Published by Elsevier B.V.

1. Introduction

Oil contaminated soils are problematic in many areas; both coastal and inland. Whilst 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.

Abbreviations: DRS, diffuse reflectance spectroscopy; LDA, linear discriminant analysis; LIBSVM, library for support vector machines; MLR, multiple linear regression; MR, misclassification rate; NIR, near-infrared; NIRS, near-infrared spectrometry; PC, principal component; PCA, principal component analysis; PLSR, partial least squares regression; RF, random forest; RMSD, root mean squared deviation; RMSEcv, root mean squared error of cross-validation; RPD, residual prediction deviation; SVM, support vector machine; TPH, total petroleum hydrocarbon; VisNIR, visible near-infrared.

* Corresponding author at: 307 M.B. Sturgis Hall, Louisiana State University Agricultural Center, Baton Rouge, LA 70803, United States. Tel.: +1 225 578 0396; fax: +1 225 578 1403.
E-mail address: DWeindorf@agcenter.lsu.edu (D.C. Weindorf).
[1] 301 M.B. Sturgis Hall, Louisiana State University Agricultural Center, Baton Rouge, LA 70803, United States.
[2] 307 M.B. Sturgis Hall, Louisiana State University Agricultural Center, Baton Rouge, LA 70803, United States.
[3] 61 Agriculture Administration Building, Louisiana State University, Baton Rouge, LA 70803, United States.

0016-7061/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.geoderma.2012.01.018

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 (Ge et al., 2007; Hyvarinen et al., 1992; Wetzel, 1983). 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 (Balabin and Safieva, 2008; Chung and Ku, 2000; Chung et al., 1999; Current and Tilotta, 1997; Westbrook, 1993; Workman, 1996; Yoon et al., 2002). 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; Brown et al., 2005; Islam et al., 2003; Morgan et al., 2009; Reeves et al., 2000; Vasques et al., 2009; Viscarra Rossel et al., 2006). 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 (Al-Abbas et al., 1972; Bowers and Hanks, 1965; Brown et al., 2005; Hunt, 1982; Hunt and Salisbury, 1970; Malley et al., 2002). A number of studies have reported an increase in prediction accuracy when VisNIR-organic C models were created for small, homogenous areas (Chang et al., 2005; Lee et al., 2003). Concurrently, other researchers have observed decreased prediction accuracy for larger geographic areas (Brown et al., 2006; Dunn et al., 2002; Kusumo et al., 2008; Shepherd and Walsh, 2002). 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 are 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.

2. Materials and methods

2.1. 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 (Gee and Or, 2002; Kilmer and Alexander, 1949; Steele and Bradfield, 1934) 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, 10,000, and 30,000 ppm. Additionally, an extra set of nine intermediate levels of crude oil concentrations (4000–28,000 ppm at 3000 ppm intervals) was created for Soil B, with an organic carbon content of 5% to improve the capability to fit models and 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.

2.2. 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 [966–1755 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.

2.3. 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 clustering 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).

2.4. 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 (Ge et al., 2007; Lark and Webster, 1999; 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.

2.5. 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).

2.6. 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.

2.7. 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).

2.8. 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.

2.9. 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.

3. 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. 1. 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 1st overtone band) and 2310 nm (C–H 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.

3.1. 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. 2 and 3, respectively) to discriminate reflectance spectra and identify clustering patterns. Fig. 2a, 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. 2a and b). Conversely, clear separations between contaminant oil types and concentrations were not delineated by first three PCs or 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 1).

Fig. 4 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 mask the oil type signatures due to the small sample size. However, the plots indicated

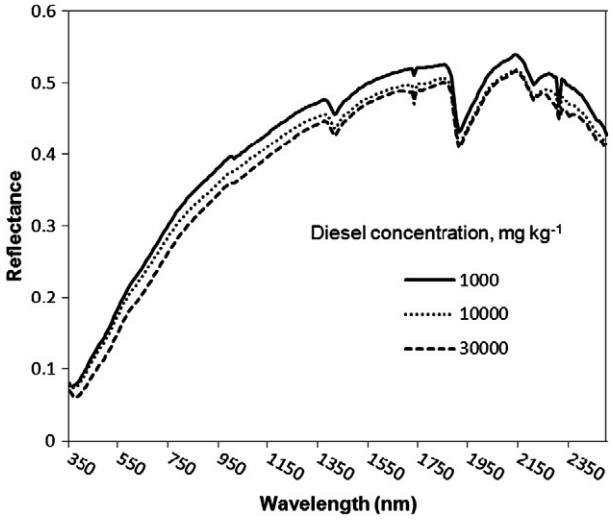

Fig. 1. Average reflectance spectra is shown for Soil A from Louisiana, USA with 1% organic carbon and different concentrations of diesel (ppm or mg kg^{-1}). Soil A is a Commerce silt loam and acidic in nature. Spectral absorption maximums of petroleum at 1730 nm and 2310 nm are apparent in mean spectral reflectance curve.

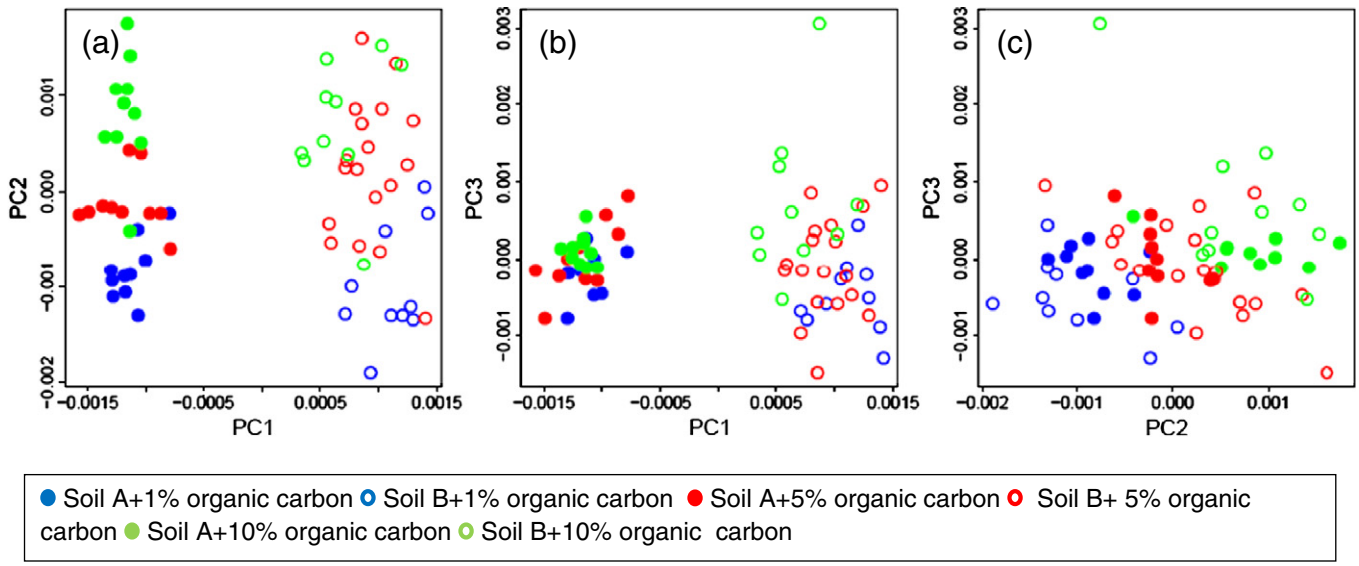

Fig. 2. Principal component (PC) plots for (a) PC1 vs. PC2, (b) PC1 vs. PC3, and (c) PC2 vs. PC3 of the first-derivative of VisNIR. The solid circles and open circles represent Soil A and Soil B, respectively. Blue, red, and green represent soils with 1%, 5%, and 10% organic carbon, respectively.

some separation of three oil types (the motor oil and 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 2. 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 than the random forest, 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 2). 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 contamination types

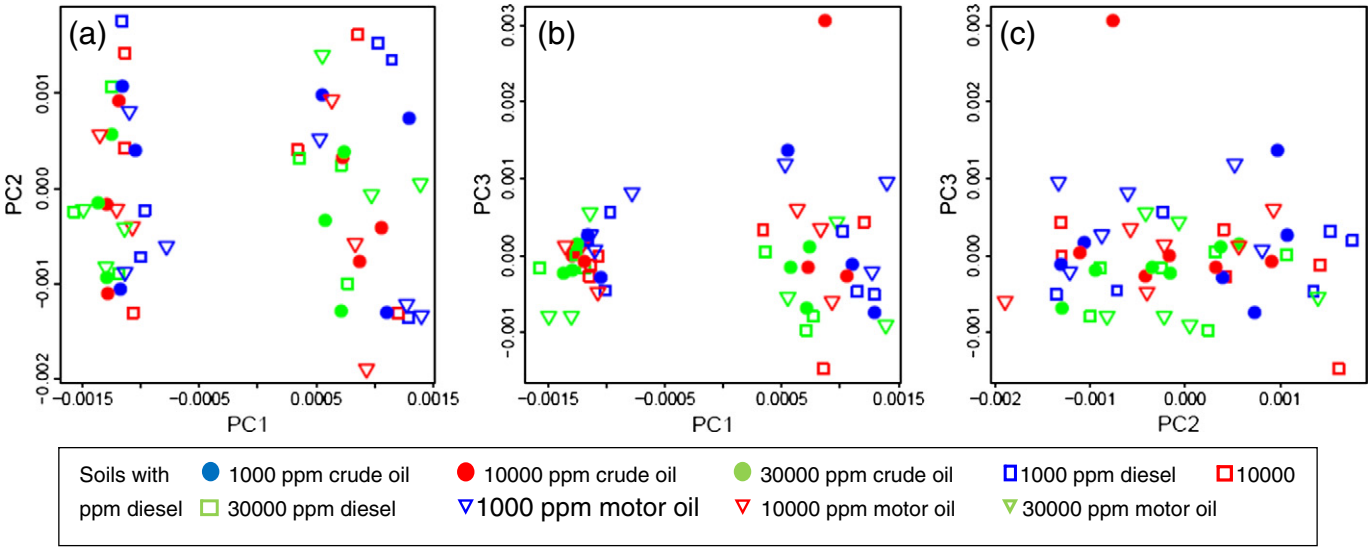

Fig. 3. Principal component (PC) plots for (a) PC1 vs. PC2, (b) PC1 vs. PC3, and (c) PC2 vs. PC3 using the first-derivative of VisNIR reflectance spectra. The circles, squares, and triangles represent soils with crude oil, diesel, and motor oil respectively. The blue, red, and green represent 1000, 10,000, and 30,000 oil concentrations (ppm or mg kg^{-1}), respectively.

Table 1
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
Sum	19	31	18	68		30	14	18	62
Overall accuracy				96%					40%

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

3.2. 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

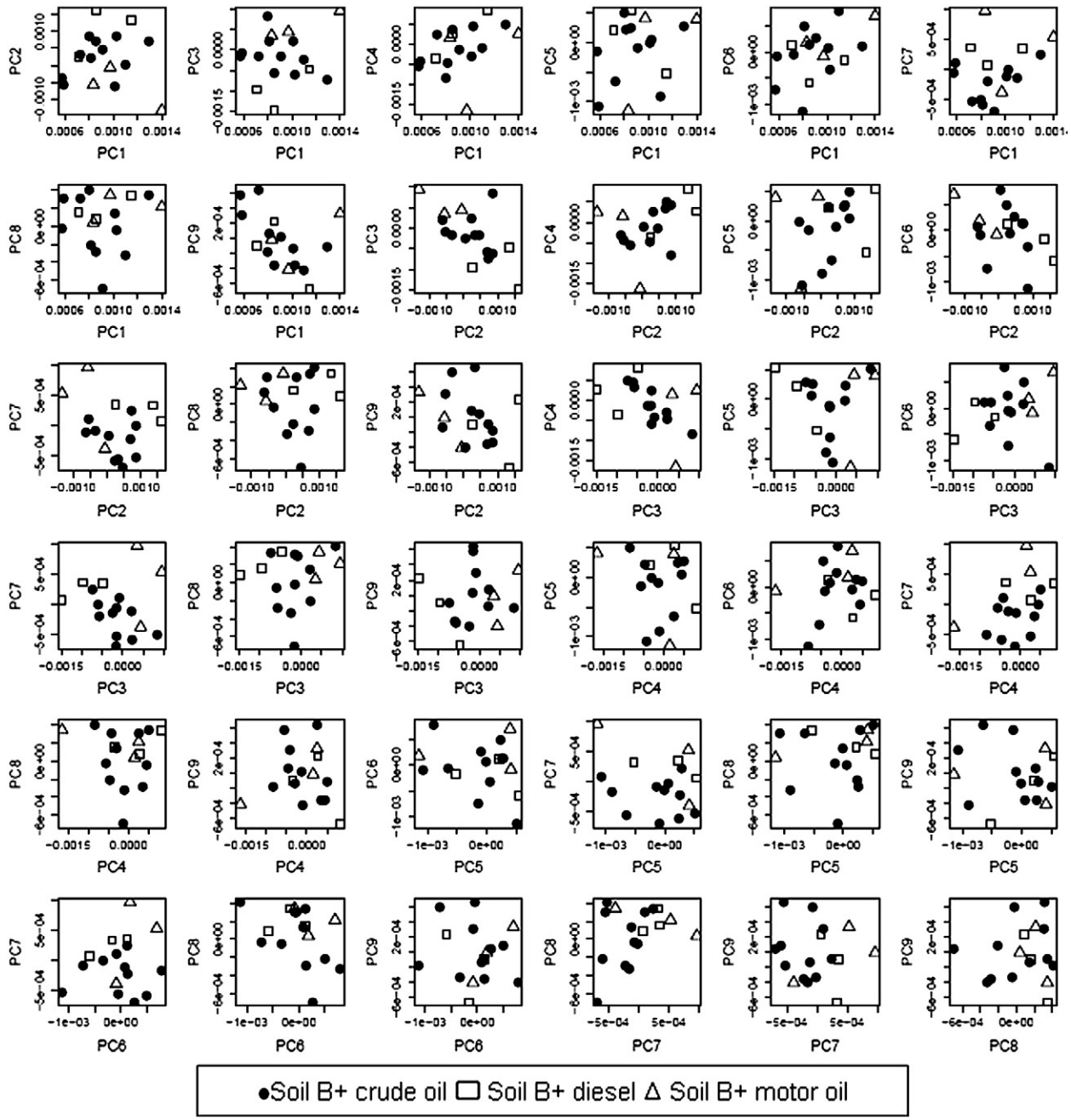

Fig. 4. Principal component (PC) plots using the first-derivative of VisNIR reflectance spectra. The circles, squares, and triangles represent soil B with crude oil, diesel, and motor oil, respectively. All samples contain 5% organic carbon.

Table 2
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[a]	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 (1st)-SVM[b]	1.3	0	26	0	64	6.5
Wavelets (r)-SVM [c]	0	0	30	1.6	65	23

[a] MR, Misclassification rate.
[b] Wavelet (1st)-SVM, Wavelet decomposition on first-derivative of reflectance followed by support vector machine.
[c] Wavelets (r)-SVM, Wavelet decomposition on raw noise corrupted reflectance spectra followed by support vector machine.

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. 5a. The PLSR plot shows 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 homoskedasticity 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. 6). 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 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. 5b). 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 homoskedastic, 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. 7 and Table 3). The fitted coefficient curve was smooth across the spectrum, indicating stability of the model. The gray-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 1300–1400 nm and 1550–1700 nm 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 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.

3.3. VisNIR DRS as a proximal sensor for petroleum content and some practical concerns

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

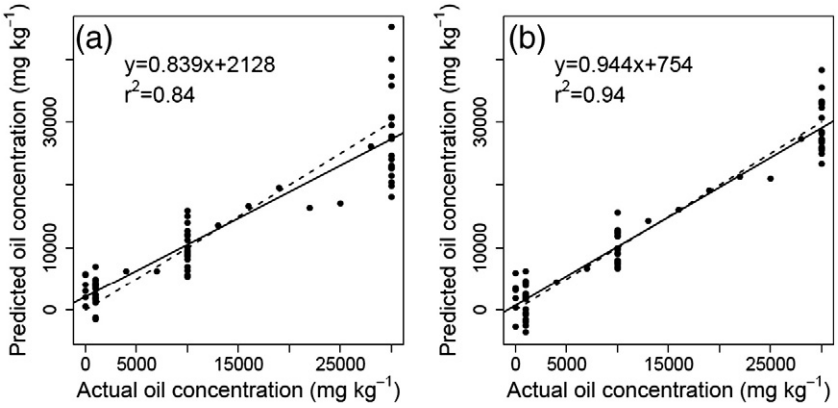

Fig. 5. Actual versus predicted oil concentration (mg kg^{-1}) using a) partial least squares regression (PLSR) and b) wavelet coefficients from the reflectance, and stepwise multiple linear regression (MLR). The solid line is the regression line, and the dashed line is a 1:1 line.

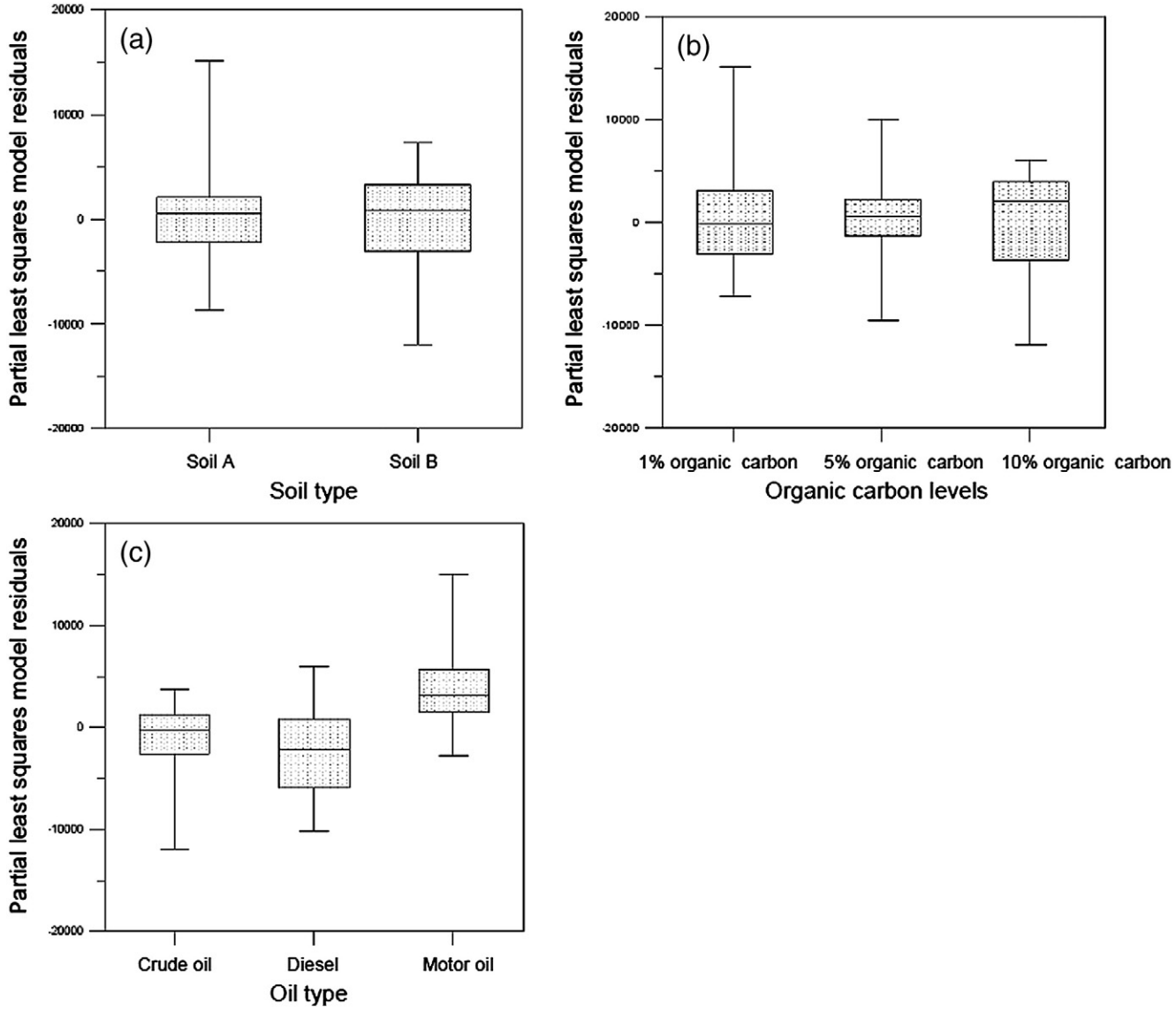

Fig. 6. Plots showing partial least squares model prediction residuals vs. a) soil type, b) organic carbon levels, and c) oil type.

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).

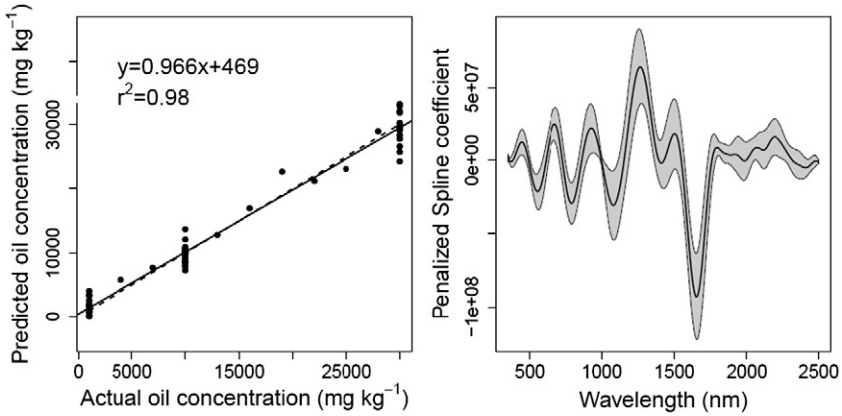

Fig. 7. The left panel shows actual versus predicted oil concentration (mg kg^{-1}) using penalized splines for soils from Louisiana, USA. The right panel shows the fitted penalized splines coefficient curve with a gray-shaded area showing the 95% confidence interval at each waveband.

Table 3
Summary of oil contamination prediction performance using different multivariate models.

Model	r^2	$RMSE_{cv}$[a] (mg kg^{-1})	RPD[b]	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

[a] $RMSE_{cv}$, Root mean square error of cross-validation.
[b] RPD, residual prediction deviation.

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. Conclusions

This exploratory study utilized 68 lab constructed samples for identifying the significant effects of soil type 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 pretreatment 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. Heteroskedasticity 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.

Acknowledgments

The authors wish to gratefully acknowledge financial assistance from the Louisiana Applied Oil Spill Research Program (LAOSRP). Crude oil was collected with the help of ES&H Consulting and Training Group and Stone Energy Corporation.

References

Al-Abbas, H.H., Swain, P.H., Baumgardner, M.F., 1972. Relating organic matter and clay content to the multi spectral radiance of soils. Soil Science 14, 477–485.
Aske, N., Kallevik, H., Sjöblom, J., 2001. Determination of saturate, aromatic, resin, and asphaltenic (SARA) components in crude oils by means of infrared and near-infrared spectroscopy. Energy & Fuels 15, 1304–1312.
Balabin, R.M., Safieva, R.Z., 2007. Capabilities of near infrared spectroscopy for the determination of petroleum macromolecule content in aromatic solutions. Journal of Near Infrared Spectroscopy 15, 343–349.
Balabin, R.M., Safieva, R.Z., 2008. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. Fuel 87, 1096–1101.
Ben-Dor, E., Banin, A., 1990. Near-infrared reflectance analysis of carbonate concentration in soils. Applied Spectroscopy 44, 1064–1069.
Ben-Dor, E., Banin, A., 1995. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Science Society of America Journal 59, 364–372.
Boser, E., Guyon, M., Vapnik, V., 1992. A training algorithm for optimal margin classifiers. Proc. Fifth ACM Workshop on Computational Learning Theory. Pittsburgh, PA, pp. 144–152.
Bowers, S.A., Hanks, R.J., 1965. Reflection of radiant energy from soils. Soil Science 100, 130–138.
Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.
Brown, D.J., Bricklemeyer, R.S., Miller, P.R., 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., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290.
Chakraborty, S., Weindorf, D.C., Morgan, C.L.S., Ge, Y., Galbraith, J., Li, B., Kahlon, C.S., 2010. Rapid identification of oil contaminated soils using visible near-infrared diffuse reflectance spectroscopy. Journal of Environmental Quality 39, 1378–1387.
Chang, C., Lin, C., 2001. LIBSVM: A library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm 2001(Verified 30 October, 2010).
Chang, C., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near infrared reflectance spectroscopy: principal components regression analysis of soil properties. Soil Science Society of America Journal 65, 480–490.
Chang, C., Laird, D.A., Hurburgh, C.R., 2005. Influence of soil moisture on near infrared reflectance spectroscopic measurement of soil properties. Soil Science 170, 244–255.
Chung, H., Ku, M., 2000. Comparison of near-infrared, infrared, and Raman spectroscopy for the analysis of heavy petroleum products. Applied Spectroscopy 54, 239–245.
Chung, H., Choi, H., Ku, M., 1999. Rapid identification of petroleum products by near-infrared spectroscopy. Bulletin of the Korean Chemical Society 20, 1021–1025.
Cloutis, E., 1989. Spectral reflectance properties of hydrocarbons: remote-sensing implications. Science 245, 165–168.
Current, R.W., Tilotta, D.C., 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. Journal of Chromatography A 785, 269–277.
Demetriades-Shah, T.H., Steven, M.D., Clark, J.A., 1990. High-resolution derivative spectra in remote sensing. Remote Sensing of Environment 33, 55–64.
Dent, A., Young, A., 1981. Soil Survey and Land Evaluation. George Allen & Unwin Publ., Boston, MA.
Donoho, D.L., Johnstone, I.M., 1994. Ideal spatial adaptation via wavelet shrinkage. Biometrika 81, 425–455.
Dumas, J.B.A., 1831. Procedes de l'analyse organique. Annales de Chimie Physique 247, 198–213.
Dunn, B.W., Beecher, H.G., Batten, G.D., Ciavarella, S., 2002. The potential of near infrared reflectance spectroscopy for soil analysis—a case study from the Riverine Plain of south-eastern Australia. Australian Journal of Experimental Agriculture 42, 607–614.
Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-spline and penalties (with comments and rejoinder). Statistical Science 11, 89–121.
Fine, P., Graber, E.R., Yaron, B., 1997. Soil interactions with petroleum hydrocarbons: abiotic processes. Soil Technology 10, 133–153.
Forrester, S., Janik, L., McLaughlin, M., 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, pp. 13–16.
Ge, Y., Morgan, C.L.S., Thomasson, J.A., Waiser, T., 2007. A new perspective to near infrared reflectance spectroscopy: a wavelet approach. Transactions of the ASABE 50, 303–311.

Gee, G.W., Or, D., 2002. Particle-size analysis. In: Dane, J.H., Topp, G.C. (Eds.), Methods of Soil Analysis. Part 4. SSSA Book Ser. 5. SSSA, Madison, WI, pp. 255–293.
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., Kuehn, F., Oschuetz, F., Lehmann, F., 2001. HyMap hyperspectral remote sensing to detect hydrocarbons. International Journal of Remote Sensing 8, 1413–1422.
Hunt, G.R., 1982. Spectroscopic properties of rocks and minerals. In: Carmichael, R.S. (Ed.), Handbook of Physical Properties of Rocks, Vol. 1. CRC Press, Boca Raton, FL, pp. 295–385.
Hunt, G.R., Salisbury, J.W., 1970. Visible and near-infrared spectra of minerals and rocks: I. Silicate minerals. Modern Geology 1, 283–300.
Hyvarinen, T., Herrala, E., Malinen, J., Niemla, P., 1992. NIR analysers can be miniature, rugged and handheld. In: Hildrum, K., Isaksson, T., Naes, T., Tandberg, A. (Eds.), Near Infrared Spectroscopy. Bridging the Gap Between Data Analysis and NIR Applications. Ellis Horwood, London, pp. 1–6.
Islam, K., Singh, B., McBratney, A., 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near infrared reflectance spectroscopy. Australian Journal of Soil Research 41, 1101–1114.
Kilmer, V.H., Alexander, L.Z., 1949. Methods for making mechanical analyses of soil. Soil Science 68, 15–24.
Ko, E.J., Kim, K.W., Park, K., Kim, J.Y., Kim, J., Hamm, S.Y., Lee, J.H., Wachsmuth, U., 2010. Spectroscopic interpretation of PAH-spectra in minerals and its possible application in soil monitoring. Sensors 10, 3868–3881.
Kusumo, B.H., Hedley, C.B., Hedley, M.J., Hueni, A., Tuohy, M.P., Arnold, G.C., 2008. The use of diffuse reflectance spectroscopy for in situ carbon and nitrogen analysis of pastoral soils. Australian Journal of Soil Research 46, 623–635.
Lark, R.M., Webster, R., 1999. Analysis and elucidation of soil variation using wavelets. European Journal of Soil Science 50, 185–206.
Lee, S.W., Sanchez, J.F., Mylavarapu, R.S., Choe, J.S., 2003. Estimating chemical properties of Florida soils using spectral reflectance. Transactions of ASAE 46, 1443–1453.
Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transsctions PAMI 11, 674–693.
Malle, H., Fowlie, P., 1998. A Canadian interlaboratory comparison for analysis of petroleum hydrocarbons in soil. Proc. Second Biennial International Conference on Chemical Measurement and Monitoring of the Environment, EnviroAnalysis'98 Conference, 11–14 May 1998. Ottawa, Canada, pp. 321–322.
Malley, D.F., Hunter, K.N., Barrie Webster, G.R., 1999. Analysis of diesel fuel contamination in soils by near-infrared reflectance spectrometry and solid phase microextraction–gas chromatography. Soil and Sediment Contamination 8, 481–489.
Malley, D.F., Yesmin, L., Eilers, R.G., 2002. Rapid analysis of hog manure and manure-amended soils using near-infrared spectroscopy. Soil Science Society of America Journal 66, 1677–1686.
Mattson, J.S., Mattson, C.S., Spencer, M.J., Spencer, F.W., 1977. Classification of petroleum pollutants by linear discriminant function analysis of infrared spectral patterns. Analytical Chemistry 49, 500–502.
Morgan, C.L.S., Waiser, T.H., Brown, D.J., Hallmark, C.T., 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. Critical Reviews in Microbiology 19, 217–242.
R Development Core Team, 2008. R: a language and environment for statistical computing. Avaliable online with updates at http://www.cran.r-project.org. R Foundation for Statistical Computing, Vienna, Austria. (Verified 30 October 2010).
Reeves, J.B., McCarty, G.W., Meisinger, J.J., 2000. Near infrared reflectance spectroscopy for the determination of biological activity in agricultural soils. Journal of Near Infrared Spectroscopy 8, 161–170.
Russell, A.E., Laird, D.A., Parkin, T.B., Mallarino, A.P., 2005. Impact of nitrogen fertilization and cropping system on carbon sequestration in Midwestern mollisols. Soil Science Society of America Journal 69, 413–422.
Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal 66, 988–998.
Sherrod, L.A., Dunn, G., Peterson, G.A., Kolberg, R.L., 2002. Inorganic C analysis by modified pressure-calcimeter method. Soil Science Society of America Journal 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. NRCS, Washington, DC. (Verified 11 Oct. 2010).
Steele, J.G., Bradfield, R., 1934. The significance of size distribution in the clay fraction. Soil Survey 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., Walter, S.D., 1988. A reappraisal of the kappa coefficient. Journal of Clinical Epidemiology 41, 949–958.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, NY.
Vasques, G.M., Grunwald, S., Sickman, J.O., 2009. Modeling of soil organic carbon fractions using visible-near-infrared spectroscopy. Soil Science Society of America Journal 73, 176–184.
Viscarra Rossel, R.A., Lark, R.M., 2010. Improved modelling of soil diffuse reflectance spectra using wavelets. European Journal of Soil Science 60, 453–464.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75.
Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T., 2007. *In situ* characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Science Society of America Journal 71, 389–396.
Wang, Z., Chang, A.C., Crowley, W.D., 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. 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. Analytical Chemistry 55, 1165A–1176A.
Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58, 109–130.
Workman, J.J., 1996. Interpretive spectroscopy for near infrared. Applied Spectroscopy Reviews 31, 251–320.
Yoon, J., Lee, B., Han, C., 2002. Calibration transfer of near-infrared spectra based on compression of wavelet coefficients. Chemometrics and Intelligent Laboratory Systems 64, 1–14.