Environmental Pollution 190 (2014) 10–18

Contents lists available at ScienceDirect

Environmental Pollution

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

Analysis of petroleum contaminated soils by spectral modeling and pure response profile recovery of *n*-hexane

Somsubhra Chakraborty [a], David C. Weindorf [b,*], Bin Li [c], Md. Nasim Ali [a], K. Majumdar [d], D.P. Ray [e]

[a] Ramakrishna Mission Vivekananda University, Kolkata 700103, India
[b] Department of Plant and Soil Science, Texas Tech University, Box 42122, Lubbock, TX 79409, USA
[c] Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803, USA
[d] Soil Testing Laboratory, Kalimpong 734301, India
[e] National Institute of Research on Jute and Allied Fibre Technology, Kolkata 700040, India

ARTICLE INFO

Article history:
Received 20 January 2014
Received in revised form
1 March 2014
Accepted 3 March 2014

Keywords:
n-hexane
Penalized spline
Petroleum
Random forest
Standard normal variate
Visible near infrared diffuse reflectance spectroscopy

ABSTRACT

This pilot study compared penalized spline regression (PSR) and random forest (RF) regression using visible and near-infrared diffuse reflectance spectroscopy (VisNIR DRS) derived spectra of 164 petroleum contaminated soils after two different spectral pretreatments [first derivative (FD) and standard normal variate (SNV) followed by detrending] for rapid quantification of soil petroleum contamination. Additionally, a new analytical approach was proposed for the recovery of the pure spectral and concentration profiles of *n*-hexane present in the unresolved mixture of petroleum contaminated soils using multivariate curve resolution alternating least squares (MCR-ALS). The PSR model using FD spectra ($r^2 = 0.87$, RMSE = 0.580 $\log_{10}$ mg kg^{-1}, and residual prediction deviation = 2.78) outperformed all other models tested. Quantitative results obtained by MCR-ALS for *n*-hexane in presence of interferences ($r^2 = 0.65$ and RMSE 0.261 $\log_{10}$ mg kg^{-1}) were comparable to those obtained using FD (PSR) model. Furthermore, MCR ALS was able to recover pure spectra of *n*-hexane.

1. Introduction

While petroleum provides abundant energy, as well as economic and manufacturing resources, its extraction, refinement and transportation also present innumerable opportunities for spillage. On April 20, 2010, the largest accidental marine oil spill in the history of the petroleum industry occurred following a sea-floor oil spill gusher from the Deepwater Horizon drilling rig explosion in the Gulf of Mexico south of Louisiana. Although major cleanup efforts have taken place and the Deepwater Horizon oil leak has been stopped, the costs associated with lost jobs, contaminated food and water, cleanup, restoration, and environmental damage has not been fully determined (Camilli et al., 2010) and may not be for many years. Quick quantification of total petroleum hydrocarbon (TPH) and polycyclic aromatic hydrocarbons (PAHs) in contaminated soils can reduce the costs involved with their management, and devise effective contingency planning through early prediction.

Wide scale gas chromatography based quantifications of spilled petroleum hydrocarbon in contaminated soils are time consuming, lacks field-portability, sometimes show high variability (an order of magnitude) in TPH results across commercial laboratories (Malley et al., 1999), and need rigorous field sampling (Dent and Young, 1981). These techniques are exceedingly laborious and often require extensive laboratory preparation (e.g., drying, grinding, acid digestion, extraction, clean-up, and quantification) and analysis which makes widespread characterization of spatial and temporal levels of contaminants prohibitively expensive. Visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) is a promising hyperspectral scanning technology that has become popular for rapidly quantifying and identifying several soil properties at a time in the laboratory and in situ with minimum or no sample pretreatments (Chang et al., 2001; Brown et al., 2006; Viscarra Rossel et al., 2006; Vasques et al., 2009). Chakraborty et al. (2010, 2012a, 2012b) demonstrated the effective use of VisNIR DRS for rapidly quantifying hydrocarbon contamination in soils

Abbreviations: FD, first derivative; MCR-ALS, multivariate curve resolution alternating least squares; PSR, penalized spline regression; RF, random forest; SNV, standard normal variate; VisNIR DRS, visible near-infrared diffuse reflectance spectroscopy.

* Corresponding author.
E-mail address: david.weindorf@ttu.edu (D.C. Weindorf).

http://dx.doi.org/10.1016/j.envpol.2014.03.005
0269-7491/

of Southern and Central Louisiana. Essentially, the C–H bond in hydrocarbons causes diagnostic absorption bands (primarily overtones and combinations) in the VisNIR region (Aske et al., 2001), making them excellent for both qualitative and quantitative analysis of contaminated soils. Similar studies by other researchers have independently confirmed that reflectance spectroscopy is suitable for on-site quantification of TPH and PAH (Graham, 1998; Forrester et al., 2010; Schwartz et al., 2012; Okparanma et al., 2014).

Since it is hard to obtain totally selective signals for every analyte of interest in a multicomponent complex sample like petroleum contaminated soil, physical separation of the analyte of interest by chromatography and/or other wet chemistry method or their mathematical resolution using chemometric methods is a critical preliminary step for their quantification (Naes et al., 2002). Given that conventional PLS-based sensors are complicated for field use (Ge et al., 2007), the overarching goal of this study was to identify some other options to conventional VisNIR-PLS models, which would aid in designing a practical sensor configuration for a field person. In this work, we compared two different state-of-the-art regression algorithms: penalized spline regression (PSR) and random forest (RF) using VisNIR DRS spectra of petroleum contaminated soils after two different spectral preprocessing techniques [first derivative of reflectance spectra and standard normal variate transformation (SNV) followed by detrending (DT)] for rapid quantification of soil TPH.

Additionally, for further recovery of the compositional and pure spectral profiles of constituents present in the unresolved mixture of petroleum contaminated soils when no prior information is available about the nature and composition of these mixtures, we used multivariate curve resolution alternating least squares (MCR-ALS) method. MCR–ALS, a soft modeling method, decomposes the bilinear data set obtained from the two or higher way array of spectroscopic data, into the pure component spectra and the relative concentration profiles (Kumar and Mishra, 2012). This method has been shown to provide an improved resolution compared to other methods and to allow quantitative determinations in the analysis of complex mixtures using spectroscopic means (Azzouz and Tauler, 2008). Thus far, MCR-ALS has been used to study complex industrial processes, multiequilibria systems using spectroscopic titrations (fluorescence, UV–vis absorption, etc.) and many others (Saurina et al., 1995; Saurina and Hernández-Cassou, 2001). Terrestrial oil spill pollution does not receive enough attention because it commonly involves small scales. Measurement of oil spill contamination in soil is also complicated by mixed spectral signatures consisting of contaminants, recovering vegetation, dead vegetation, and the signal from soils (Hese and Schmullius, 2008). In this study, we used MCR-ALS for both quantitative and qualitative determination of *n*-hexane present in petroleum contaminated soils using VisNIR reflectance spectra. We further introduced a *correlation constraint* to establish an alternating least squares (ALS) multivariate model (Antunes et al., 2002). The question is: How useful would such a model be when there is a larger contribution of unknown physical contributions and chemical interferents in the measured VisNIR spectra of the analyzed soil samples?

Thus, the objectives of this study were to: (i) compare two different multivariate data-mining tools with two different spectral pretreatments for characterizing petroleum contaminated soils and (ii) to examine whether MCR–ALS analysis can be used for the simultaneous quantification and spectral profile recovery of *n*-hexane in petroleum contaminated soils. To the best of our knowledge this study is the first attempt to adopt VisNIR DRS in combination with MCR-ALS for the determination of petroleum hydrocarbons in contaminated soils. It is our hope, through this paper, to raise awareness of a methodology that could be adapted by many of the reclamation agencies and organizations involved, directly benefiting people, wildlife, and the environment. We also hope this study can provide some rudimentary, but important, insights on the future utility of VisNIR DRS sensors for real time soil petroleum contamination monitoring.

2. Materials and methods

2.1. Soil sample collection

Initially, a total of 46 samples (0–15 cm) were collected from multiple sites [Sabine (SAB), Sonat (SON), Winn Dixie (WD), Mississippi River (MR), and Alpine (ALP)] located in different parishes within Central and Southern Louisiana in August, 2008 (Table 1, Fig. 1). The sampling strategy was developed with information provided by the Louisiana Oil Spill Coordinators Office (LOSCO) to ensure maximum TPH variability within the soil samples collected. Soil samples included both contaminated and similar soils with no-known petroleum contamination. Samples were placed on ice for transport to the laboratory and refrigerated at 4 °C in the laboratory. Another 118 surface (0–15 cm) soil samples were collected from a crude oil well blowout site (near the Sonat site, referred to as SV) located in Kisatchie National Forest in Vernon Parish, Louisiana across an 80 ha area densely vegetated by trees, shrubs, and grasses. The sampling locations at the blowout site represented four soil series: Caddo silt loam (fine-silty, siliceous, active, thermic Typic Glossaqualf), Guyton silt loam (fine-silty, siliceous, active, thermic Typic Glossaqualf), Malbis fine sandy loam (fine-loamy, siliceous, subactive, thermic Plinthic Paleudult), and Ruston fine sandy loam (fine-loamy, siliceous, semiactive, thermic Typic Paleudult) (Table 1). All soil samples were sealed in air-tight glass jars to prevent hydrocarbon volatilization and preserve field-moisture status. Readers are referred to Chakraborty et al. (2010) and Chakraborty et al. (2012b) for detailed information regarding soil sampling. Additionally, we collected 20 sand samples from beaches impacted by the Deepwater Horizon oil spill near Pensacola, Florida in the summer of 2010. However, due to the compositional difference from soil matrix, sand samples were not included in multivariate analysis and only used for qualitative spectral exploration.

2.2. Spectral scanning

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

Table 1
Location, soil series, and classification of soils evaluated for petroleum contamination using VisNIR DRS in Louisiana, USA.

Site	Parish	Soil series	Classification[a]	Samples
Alpine	Jefferson	Barbary	Very-fine, smectitic, nonacid, hyperthermic Typic Hydraquents	12
Mississippi River 1	Plaquemine	Carville	Coarse-silty, mixed, superactive, calcareous, hyperthermic Fluventic Endoaquepts	2
Mississippi River 2	Saint Charles	Cancienne	Fine-silty, mixed, superactive, nonacid, hyperthermic Fluvaquentic Epiaquepts	2
Sabine	Cameron	Creole	Fine, smectitic, nonacid, hyperthermic Typic Hydraquents	8
Sonat	Vernon	Ruston	Fine-loamy, silicious, semiactive, thermic Typic Paleudults	12
Winn Dixie	East Baton Rouge	N/A[b]	Udarents	10
SV	Vernon	i. Caddo	Fine-silty, siliceous, active, thermic Typic Glossaqualfs	32
		ii. Guyton	Fine-silty, siliceous, active, thermic Typic Glossaqualfs	28
		iii Malbis	Fine-loamy, siliceous, subactive, thermic Plinthic Paleudults	34
		iv. Ruston	Fine-loamy, siliceous, semiactive, thermic Typic Paleudults	24

[a] Soil Survey Staff, 2005.
[b] Not applicable.

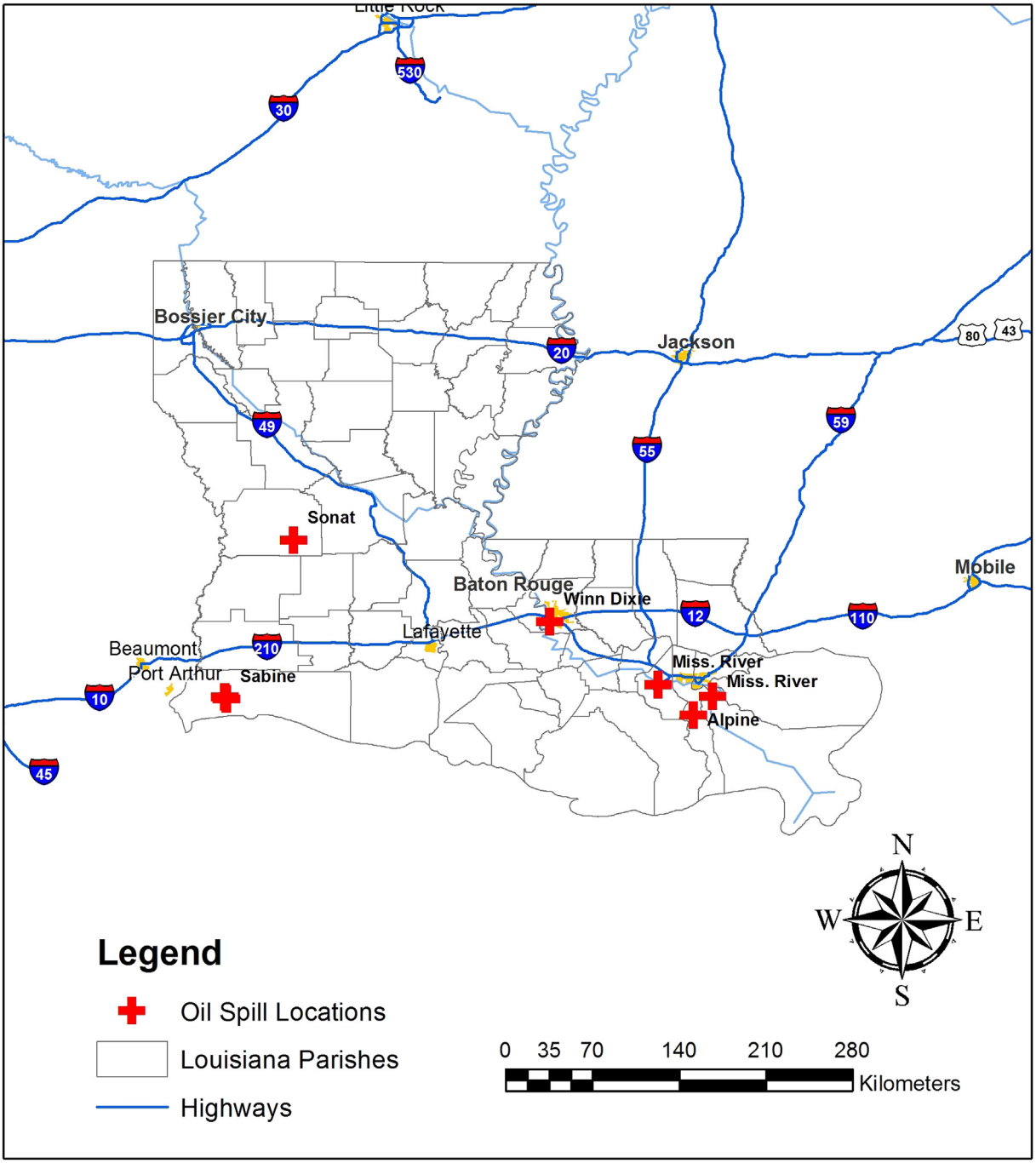

Fig. 1. Soil sampling location in Louisiana, USA. A total of 164 soil samples were collected and used in regression analysis.

2500 nm, respectively) was used to scan soil samples with a contact probe. The contact probe had a circular viewing area (20 mm diameter) and its own halogen light source. Each sample was scanned four times with a 90° rotation between scans to obtain an average spectral curve. Prior to scanning, 2 h of warming was allowed to stabilize instrument temperature and account for dark current drift. White reference scans (with a Spectralon™ panel having 99% reflectance) were taken every four samples. The spectroscopic reflectance measurement for each soil sample was then obtained by averaging the four raw scans.

2.3. Spectral preprocessing

We used both derivative spectroscopy and SNV-DT to preprocess soil spectra preceding multivariate analysis. Derivative spectra remove the baseline shift arising from detector inconsistencies, albedo, and sample handling (Demetriades-Shah et al., 1990). Raw reflectance spectra were processed via a statistical analysis software package, R version 2.11.0 using custom 'R' routines (R Development Core Team, 2008). These routines involved (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, and (iv) 1st-derivatives (FD) at 10-nm intervals. The resulting FD spectra were extracted and individually combined with the laboratory measured TPH. The processed data was used to build prediction models using PSR and RF algorithms. Barnes et al. (1989) proposed SNV-DT to remove multiplicative interferences of scatter and particle size and to explain the difference in baseline shift and curvilinearity in diffuse reflectance spectra. SNV, which is also known as *z*-transformation or centering and scaling (Otto, 1998), normalizes each spectrum ρ to zero mean and unit variance by subtracting the mean of this spectrum ρ' and dividing the difference by its standard deviation σ_ρ (Eq. (1)):

$$\mathrm{SNV} = (\rho - \rho')/\sigma_\rho \quad (1)$$

This is followed by a detrending step which is a 2nd-order polynomial fit to the SNV transformed spectrum and subtracted from it to correct for wavelength-dependent scattering effects (Buddenbaum and Steffens, 2012). Like FD spectra, SNV-DT spectra of soil samples were also used for PSR and RF modeling. For MCR-ALS, the reflectance spectra were transformed to absorbance (log 1/reflectance).

2.4. Total petroleum hydrocarbon and n-Hexane analysis

The petroleum in soil samples were extracted using method 5520 D Soxhlet extraction in a commercial laboratory (Clesceri et al., 1998), and TPH was quantified by method 5520 F (Clesceri et al., 1998). Briefly, 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.

Volatile *n*-hexane in contaminated soil samples was quantified using supercritical fluid extraction with a sorbent trap as described by Yang et al. (1995). Briefly, 10 g of soil sample was used to fill the HP 7680T supercritical fluid extractor's stainless steel extraction cell and was extracted at 340 atm and 80 °C with SFC-grade CO_2 at a flow-rate of 1.5 ml/min for 30 min. Collection traps were filled with 1 ml of trapping material (Porapak Q, 60/80 mesh) and extraction was continued for 30 min. 1, 3, 5-triisopropylbenzene was used as the internal standard. Extracts were analyzed by an HP 5890 GC coupled with an FID detector. Both TPH and n-hexane analysis were done in triplicate.

2.5. Multivariate total petroleum hydrocarbon predictive models

It is known that classical least squares modeling approaches usually fail on high-dimensional multivariate calibration problems because the size of the regressors is larger than the sample size. Unlike most other approaches, the PSR model makes use of the ordered structure among the regressors (Eilers and Marx, 1996) and is well-suited for ill-posed problems (the dimensionality is much larger than the sample size) such as signal regression problems. In this study, the cubic B-spline was used (using R version 2.14.1) as the base function with 100 equally spaced knots. The order of the penalty was set to the default value of three. The data was randomly split into training (120 samples, ~73% of the whole data) and test sets (44 samples, 27%). The optimal value for the penalty-tuning parameter was selected by minimizing the LOOCV error on the training set. Since, the original TPH values were widely and non-normally ($1.22–3.74 \times 10^9$ mg kg^{-1}) distributed, Box–Cox transformation (Box and Cox, 1964) was applied to the original TPH data using $\lambda = 0$ ($\log_{10}$-transformed) to bring the data to a Gaussian distribution. The $\log_{10}$-transformed TPH values were used for both PSR and RF models as the dependent variable. In this study, the random forest package was used in R to build the random forest model (Breiman, 2001). The number of trees in random forest was set to the default value of 500. The coefficient of determination (r^2), bias (systematic error), and RPD were used as rubrics for evaluating the quality of RF and PSR in real-world situations.

2.6. Multivariate curve resolution alternating least squares

We used MCR-ALS to quantify *n*-hexane present in petroleum contaminated soils via VisNIR DRS spectra using MATLAB 2009a (The Mathworks, MA, USA). The algorithm started with a data matrix, **M** which contained the different individual spectra measured for the different soil samples in the rows, while the columns represented absorbance values at each spectral wavelength. Subsequently, principal component analysis (PCA) was done to estimate the possible number of components. Comparable to Beer–Lambert law (Skoog et al., 1995), the aim of MCR is to establish a bilinear relation between the experimental data, the concentrations, and the pure spectra as follows (Azzouz and Tauler, 2008) (Eq. (2)):

$$\mathbf{M} = \mathrm{CS}^{\mathrm{T}} + \mathrm{E} \tag{2}$$

where **M** (*A*, *B*) denotes the matrix of experimental data, of dimensions *A* soil samples (spectra) by *B* wavelengths; **C** (*A*, *K*) represents the matrix of concentration profiles of the different *K* analytes present in the contaminated soil samples; $\mathbf{S}^{\mathrm{T}}$ (*K*, *B*) stands for the spectra matrix, whose *K* rows contain the pure spectra associated with the *K* species present in the samples; **E** (*A*, *B*) is the matrix associated to the experimental error. See Azzouz and Tauler (2008) for details on the resolution of experimental spectral data matrix **M**. We followed purest variable based initial estimation of spectral or concentration profile for each species to start the iterative ALS process (Windig and Stephenson, 1992; Windig, 1992; Azzouz and Tauler, 2008). The unconstrained least squares solutions for estimating spectral profiles and concentration profiles followed Eqs. (3) and (4), respectively (Azzouz and Tauler, 2008).

$$\mathbf{C} = \mathbf{M}\left(\mathrm{S}^{\mathrm{T}}\right)^{+} \tag{3}$$

$$\mathrm{S}^{\mathrm{T}} = \mathbf{C}^{+}\mathbf{M} \tag{4}$$

Note that, $(\mathbf{S}^{\mathrm{T}})^{+}$ and $\mathbf{C}^{+}$ are the pseudoinverse matrices of the spectra matrix $\mathbf{S}^{\mathrm{T}}$ and **C** matrix, respectively. Initial estimates needed to start the ALS procedure described by these two abovementioned equations were obtained by SIMPLISMA (SIMPLe-to-use Interactive Self Modeling Analysis) algorithm (Windig and Stephenson, 1992). A series of constraints like non-negativity concentration constraint, non-negativity spectra constraint, and correlation constraint were imposed. For details on these constraints, see Sajda (2006) and Kumar and Mishra (2012).

2.7. Principal component analysis

PCA was applied for qualitative VisNIR discrimination of the prepared samples according to the variable locations for both derivative spectra and SNV-DT spectra. The cumulative proportion of variance explained by the leading principal components (PC) was used to extract optimum PCs. Furthermore, pairwise scatterplots of the first three PCs were produced to provide visual assessment on how different groups were separated in the PC space. PCA was performed using R version 2.11.0 (function: prcomp).

3. Results and discussions

In the present study, $\log_{10}$-transformed TPH contents of 164 analyzed soil samples ranged from 0.05 to 9.57 $\log_{10}$ mg kg^{-1}. These values were used as the dependent variable for subsequent RF and PSR modeling (Table 2). Usually, scientists collect soil and other environmental samples from the petroleum spill site and subsequently analyze for TPH using several standard methods. However, these methods neither measure the entire hydrocarbons in a sample, nor do they measure identical subsets of hydrocarbons when methods are compared (Baugh and Lovergreen, 1990). Consequently, reported TPH is a method-dependent entity in all media (McKenna et al., 1995). Moreover, owing to varying magnitudes of volatility and dissolution ability of petroleum constituents, the results of TPH analyses are believed to be controlled by elapsed time since spill as well as the fuel type. Considering the abovementioned limitations, we measured *n*-hexane as an alternative to TPH. Notably, *n*-hexane was positively but non-significantly correlated with TPH content ($\rho = 0.8$). Maximum average TPH content was obtained from the MR site (74,131 mg kg^{-1}) followed by the SON (22,387 mg kg^{-1}), WD (6606 mg kg^{-1}), SV (4850 mg kg^{-1}), ALP (4265.8 mg kg^{-1}), and SAB (4073.8 mg kg^{-1}) sites. Conversely, maximum average n-hexane concentration was achieved from the SAB site (213.8 mg kg^{-1}) followed by the MR (85 mg kg^{-1}), ALP (65 mg kg^{-1}), SON (63.5 mg kg^{-1}), WD (52.4 mg kg^{-1}), and SV (42.6 mg kg^{-1}) sites. However, we refrained from comparing and interpreting the spatial variability of the TPH and *n*-hexane within and among sites due to an unequal number of samples for each site within the experimental design. Indeed, almost 72% of the samples ($n = 118$) were collected from SV site. Researchers have reported that the collection of a sufficient number of petroleum contaminated samples and crude oil with different compositions and quality indices is not an easy task and sometimes involves legal proceedings (Balabin and Safieva, 2007; Chakraborty

Table 2
Descriptive statistics of measured TPH and n-hexane for 164 soil samples scanned with VisNIR DRS.

Parameter	*n*	Max	Min	Mean	Standard deviation
			mg kg^{-1}		
TPH	164	3.74×10^9	1.12	23,485,340	2.92×10^8
n-hexane	164	378	2.51	76.98	69.28
			$\log_{10}$ mg kg^{-1}		
TPH ($\log_{10}$ mg kg^{-1})	164	9.57	0.05	3.90	1.44
n-hexane ($\log_{10}$ mg kg^{-1})	164	2.57	0.04	1.70	0.43

et al., 2012a). We acknowledge that the unevenness in the number of soil samples somewhat constrain the global applicability of the dataset. Nevertheless, this research was intended to investigate the viability of different spectral analysis techniques, and ascertain which techniques show the most promise for future investigations. Low molecular weight alkanes like *n*-hexane tends to volatilize following a spill and do not have a great tendency to leach and migrate into subsurface soils or groundwater due to their hydrophobicity and affinity for organic carbon (CRC, 1986). Thus, we feel that other indicators like benzene, toluene, ethylbenzene, and xylene (BTEX) should be individually analyzed and combined with VisNIR spectra for subsequent multivariate analysis. The testing of crude oil composition (BTEX and PAHs) was, however, beyond the scope of this study.

Spectral data from five randomly selected beach sand samples are displayed in Fig. 2 to show the appearance of the spectral features in the contaminated samples. In general, the spectra of similar types of samples (contaminated and clean sand) were similar while increasing crude oil concentration caused higher absorbance, reflecting less light than non-contaminated or reference samples (Weindorf et al., 2011). Changes in the reflectance pattern in the visible region (350–700 nm) possibly appeared from color associated changes when crude oil was mixed with clean sand. The same trend was observed by researchers in case of other soil matrices (Mouazen et al., 2005; Okparanma and Mouazen, 2013). Sand contaminated with heavy oil and clean sand exhibited the highest and least absorbance, respectively, with intermediate reflectance for sand containing tar balls. Interestingly, in the reflectance curves of the contaminated sands, spectral absorption minima of crude oil was appeared around 1714 and 1758 nm in the first overtone region of the NIR band, possibly arising from C–H stretching modes of terminal CH_3 and saturated CH_2 groups linked to TPH (Okparanma et al., 2014). Although the locations of abovementioned signatures were a bit shifted from the exact anticipated positions (1712 nm, 1752 nm, respectively), it was natural in the sense that real molecules do not behave totally harmonically. These signatures, however, were absent in case of clean sand samples. Overall, VisNIR DRS appeared to be successful in identifying the variability in crude oil content in a sand matrix owing to differential absorbance.

The upper and lower panels of Fig. 3 show the pairwise PC plots among the first three leading components on the FD data and SNV-DT data, respectively. Based on these two panels, the PC plots for both spectral pretreatments showed a similar pattern and appeared like mirror images of each other. Moreover, all plots exhibited clear location-clustered structure. This could be explained by the spectral heterogeneity which resulted partly from soil matrix variability (from Typic Hydraquents to Typic Paleudults) and partly from variable crude oil composition (Chakraborty et al., 2010). For example, the first leading component (PC1) separated the SV samples from the ALP, SAB, MR, and WD samples. Samples from SON were mixed between the two groups since SON samples belonged to the Ruston fine sandy loam series from where some of the SV samples were collected later. Only SV samples tended to have very low to very high scores along PC2, showing greater spectral diversity. The first three leading PCs constituted over 97% of the spectral variation. No obvious outlier among the samples was seen. Therefore, it can be concluded that this hyperspectral sensing demonstrates both the power of unsupervised PCA as a means of

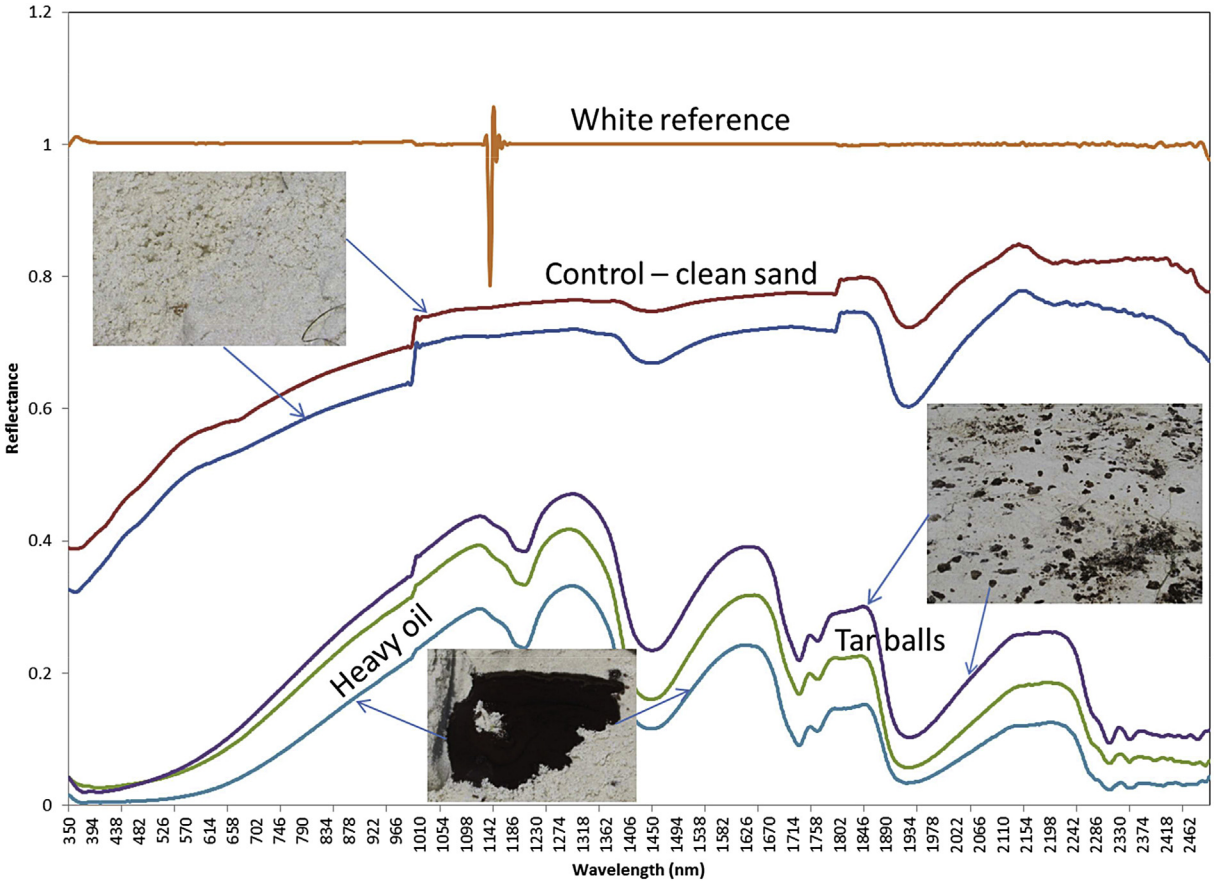

Fig. 2. VisNIR diffuse reflectance spectra of five sand samples after the 2010 Deepwater Horizon spill, Pensacola beach, Florida, USA.

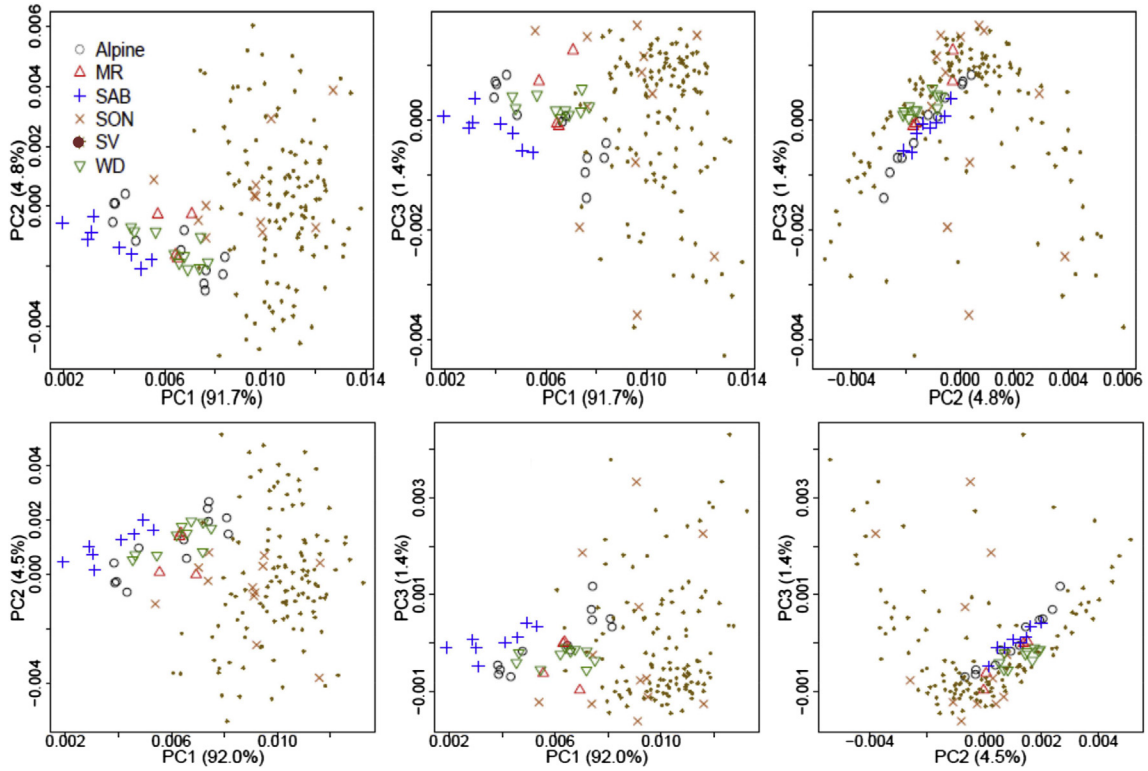

Fig. 3. The upper and lower panels exhibit pairwise PC plots of the first three components for FD spectra and SNV-DT spectra, respectively.

discriminating VisNIR reflectance spectra of oil contaminated soils and a tendency for regional similarity of contaminated soils. Interestingly, considering site wide average *n*-hexane content, a subtle distribution of samples was observed on PC1 of the FD spectra (Fig. 3, upper left plot). The separation of the VisNIR profiles of the SAB and SV sites with the highest (213.8 mg kg^{-1}) and lowest (42.6 mg kg^{-1}) average *n*-hexane contents, respectively was obvious.

Applied to the quantification of log_{10}-transfomed TPH, VisNIR DRS produced significant and strong correlations to traditional lab analysis for both PSR and RF models (Figs. 4 and 5). Fig. 4a shows the FD (PSR) predicted TPH contents against the lab measured values using the whole dataset. The overall r^2 (including both the training and test sets) was 90%. The non-unity regression line value (see Brown et al., 2006) was only 2%, suggesting that VisNIR-FD (PSR) did not substantially over or underestimate soil TPH. The summary error statistics on the 27% test dataset for both spectral preprocessing is presented in Table 3. Moreover, Fig. 4b shows the estimated regression coefficients together with the 90% confidence interval band for the FD (PSR) model. Two spikes on the coefficient plot with different signs were observed: one was ~1200 nm and the other was ~2200 nm which could contain the 2298-nm (stretch + bend) signature correspond to combination and overtone bands primarily of saturated CH_2 or CH_3 of crude oil as reported by Mullins et al. (1992).

In this study, the number of trees in RF was set to the default value of 500. The data splitting scheme was the same as the one described in the PSR model section to obtain a fair comparison. The training/out-of-bag (OOB)/test errors (in RMSE) on different values of *mtry* (random subset), which is the size of the candidate subset for each splitting, are illustrated in Fig 5a. The vertical dashed line is the optimal *mtry* value that minimizes the OOB error. It was

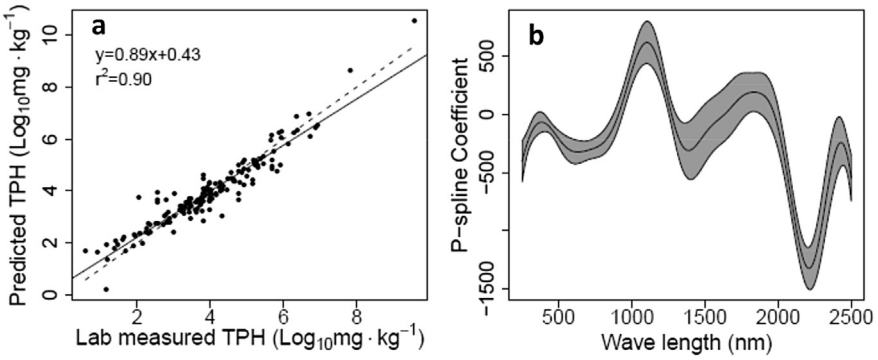

Fig. 4. Figures showing a) FD (PSR) predicted vs. lab measured TPH contents using the whole dataset (with dotted 1:1 line) and b) the fitted penalized splines (P-splines) coefficient curve at each waveband. The gray-shaded area is the 90% confidence interval.

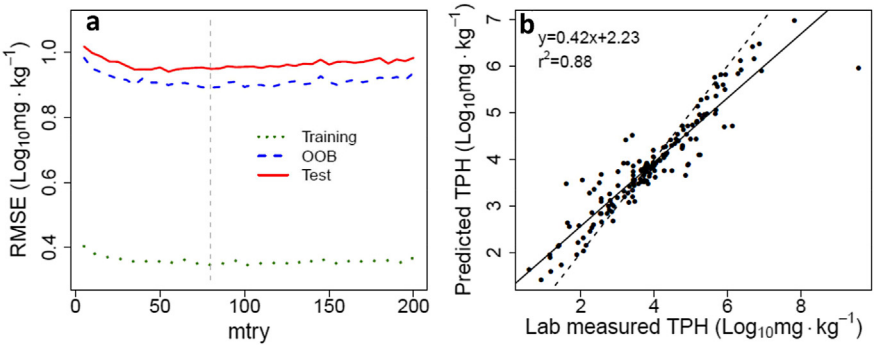

Fig. 5. Figures showing a) changes in RMSE ($\log_{10}$ mg kg^{-1}) on different values of *mtry* for FD (RF) and b) FD (RF) predicted vs. lab measured TPH contents using the whole dataset (with dotted 1:1 line).

apparent that the prediction performance was quite insensitive to the *mtry* value. Moreover, the OOB error was very close to the test error. Hence OOB error was used as an estimate of the prediction performance for the model on a new dataset, while the training error was substantially below the test error. On the other hand, Fig. 6 illustrates the RMSE on different values of the tuning parameter λ (on $\log_{10}$ scale) for FD (PSR) model. The vertical dash line is the optimal value for λ which minimizes the LOOCV error (blue dash line) (in the web version). It was clear that the selected tuning parameter value was very close to the optimal value that minimized on the test set (red line). Conversely, the training error monotonically decreased with the size of λ. Furthermore, Fig. 5b shows the FD (RF) predicted vs. lab measured TPH values using the whole dataset. The overall r^2 (including both the training and test sets) was 88%. Nevertheless, while comparing the prediction performances of both models based upon the 27% test set, the FD (PSR) model outperformed all other models tested, providing the highest coefficient of determination (0.87) along with the highest RPD (2.78) and lowest RMSE (0.528 $\log_{10}$ mg kg^{-1}), indicating the robustness and accuracy of the model (Table 3). Accuracy and stability of different multivariate models were evaluated according to the RPD-based rubrics by Chang et al. (2001). Summarily, considering regression models using both the complete dataset and test set, the FD (PSR) model outperformed all other models. The influence of the two different spectral pretreatments on the VisNIR prediction of soil TPH was rather small. Even if researchers have reported the advantages of spectral pretreatments before subsequent modeling (Ben-Dor et al., 1997), in this study those transformations were not very useful. While spectral pretreatments might have emphasized the information contained in the spectra, the powerful PSR algorithm did not seem to depend on that. Another reason of little impact of spectral pretreatment could be the very broad absorption features which appeared in the VisNIR region (Kooistra et al., 2001).

Since VisNIR DRS is sensitive to changes in the matrix material scanned, especially moisture (as it relates to O–H bonding and color) (Bishop et al., 1994; Zhu et al., 2010), bringing the soil samples to standard water content (field capacity) before scanning was found to be crucial for obtaining consistent results. Notwithstanding that under laboratory controlled conditions soil can be scanned under uniform moisture content, maintaining unique water content in the field is not easy. Thus, in this study we relied on findings of several researchers where it was reported that NIR spectra calibrated using field soil samples can be used for field prediction of soil properties (Christy, 2008; Kusumo et al., 2008; Bricklemyer and Brown, 2010). Scientists suggested that incorporating an extensive range of moisture content in the calibration set could reasonably handle the issue of moisture variation (Minasny et al., 2011).

In this work, we compared MCR-ALS and PSR with cross-validation in the analysis of contaminated samples using VisNIR DRS for generating both qualitative and quantitative information of the *n*-hexane present in the samples. The FD data was used for PSR while no data pretreatment was done for MCR-ALS. Model prediction performance indicated that multivariate generalization capacity worsened for *n*-hexane compared to TPH. However, FD (PSR) slightly outperformed MCR-ALS with a higher coefficient of determination (0.70) and better RMSE (0.241 $\log_{10}$ mg kg^{-1}), while the latter produced an r^2 of 0.65 and an RMSE of 0.261 $\log_{10}$ mg kg^{-1} (Fig. 7). Interestingly, this trend again upheld the idea that cross-validation models developed by pre-processed spectra had better quality than raw reflectance models (Okparanma et al., 2014). For both algorithms, we can see that the prediction deteriorates significantly and also has a bias, a tendency to under-predict with

Table 3
Summary statistics for PSR and RF models using a 27% test set with FD and SNV-DT spectra for predicting soil TPH.

Model	r^2	RPD[a]	RMSE ($\log_{10}$ mg kg^{-1})[b]	Bias ($\log_{10}$ mg kg^{-1})
FD (PSR)	0.87	2.78	0.528	0.009126
SNV-DT (PSR)	0.80	2.21	0.664	0.02886
FD (RF)	0.58	1.56	0.954	−0.09446
SNV-DT (RF)	0.58	1.57	0.948	−0.0009141

[a] RPD, Ratio of standard deviation and RMSE.
[b] RMSE, Root mean squared error of prediction.

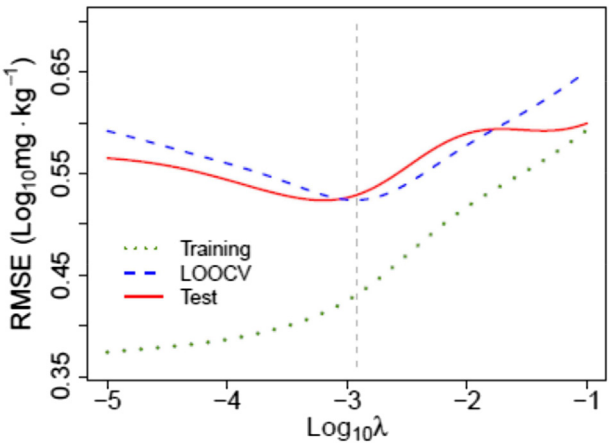

Fig. 6. Changes in RMSE ($\log_{10}$ mg kg^{-1}) on different values of the tuning parameter λ (on $\log_{10}$ scale) for the FD (PSR) model.

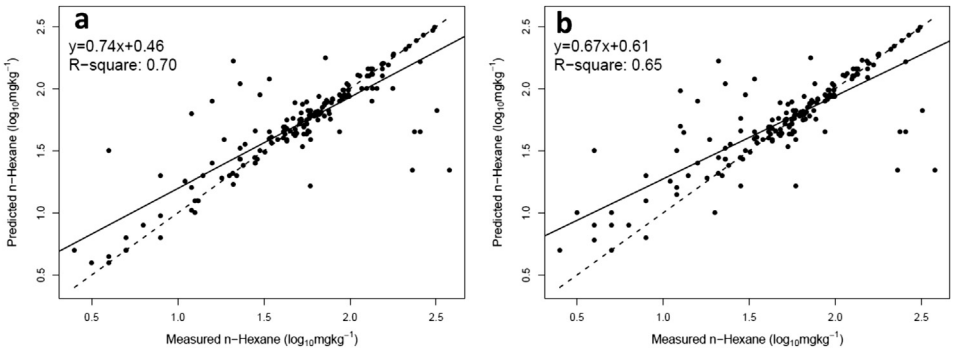

Fig. 7. Figures showing a) FD (PSR) predicted vs. lab measured *n*-hexane contents (with dotted 1:1 line) and b) MCR-ALS predicted vs. lab measured *n*-hexane contents for the using the whole dataset.

increasing n-hexane content. Apparently, achieving a high calibration performance was difficult perhaps due to highly volatile nature of *n*-hexane and because of the substantial contribution of unknown physicochemical interferents (e.g., PAHs, soil texture, soil organic matter) in the measured VisNIR spectra of the analyzed soil samples. We believe that an intensive discussion on these practical difficulties would require the study of a larger number of samples with a better control of such interferents. Nonetheless, *n*-hexane determination using MCR-ALS showed some potential and was close to the optimal one achieved through FD (PSR). While the *n*-hexane prediction accuracy using both algorithms was not as high as that obtained for other constituents of petroleum contaminated soils (Okparanma et al., 2014), the results were encouraging, considering the magnitude of difficulties inherent to MCR-ALS to correctly resolve and quantify components contributing insignificantly to the measured spectra and also to the fact that neither sample nor spectral pretreatment was performed using MCR-ALS. While comparing the VisNIR spectra of n-hexane as resolved by MCR-ALS and reported elsewhere (Analytical Spectral Devices, 2001), the former produced very similar curves and clearly identified the characteristics bands at ~1150 nm, 1450 nm, and ~1850 nm (Fig. 8). However, one must use caution these plots are interpreted, since these bands were broad and perhaps overlapping. In general, the presence of octane in hexane (Fig. 8b) perhaps caused a deviation in the form of the bands located below 1100 nm. Note that the peaks from the study all appeared to be shifted relative to the literature peak pattern. While the patterns were very similar, the first peak from the study appeared to be shifted a bit to the left, while the second and third peaks were shifted to the right. This was expected since real molecules do not behave totally harmonically (Bishop et al., 1994), suggesting that interpretations of soil VisNIR spectra can be difficult particularly when the target species is present in only small amounts in the soil. Although MCR-ALS was able to calculate pure spectra of all chemical species in the contaminated soil, we presented only n-hexane since detailed compositional characterization with reference methods was beyond the scope. Summarily, in this work, it was shown that the MCR–ALS can be used for the simultaneous extraction of the pure synchronous profile at various wavelengths from the VisNIR DRS. However, it is also true that the obtained results may not be as accurate as in the synthetic mixtures.

4. Conclusions

A fast and convenient soil analytical technique is needed for post-spill soil quality assessment and formulating an effective restoration plan. VisNIR DRS is a simple and non-destructive analytical method that can be used to predict several soil properties simultaneously. This study utilized a total of 164 petroleum contaminated soil samples from central and southern Louisiana for scanning via VisNIR spectrometer and correlating diffuse reflectance data and soil TPH values. The results showed that the derivative spectroscopy was a better pretreatment as compared to the SNV-DT pretreatment. The PSR model using FD spectra outperformed all models tested, producing 87% of the variability of the independent validation set. Unsupervised PCA also qualitatively separated contaminated soils location-wise. Visual interpretations from PC plots revealed subtle separation of samples based on average site wise n-hexane contents. Hence, VisNIR DRS showed the potential to be used as a post-spill rapid soil contamination monitoring tool. Furthermore, the predictive ability of MCR-ALS for a specific analyte (*n*-hexane, in this study) in the presence of interferences in crude oil contaminated soil samples showed potentiality and produced comparable results ($r^2 = 0.65$ and RMSE 0.261

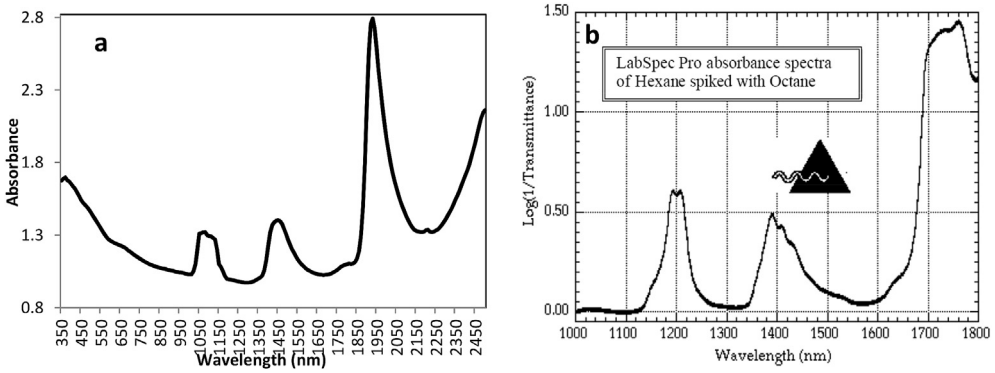

Fig. 8. Plots exhibiting a) pure spectra of n-hexane estimated by MCR-ALS and b) spectra of hexane taken from the literature (Analytical Spectral Devices, 2001).

$\log_{10}$ mg kg^{-1}) to the FD (PSR) model ($r^2 = 0.70$ and RMSE 0.241 $\log_{10}$ mg kg^{-1}) along with simultaneous recovery of pure spectra. In this present feasibility study, we have presented a preliminary contribution to this problem and further intensive research is recommended to confirm the results here obtained.

Acknowledgments

The authors wish to gratefully acknowledge financial assistance from the Louisiana Applied Oil Spill Research Program (LAOSRP).

References

Aske, N., Kallevik, H., Sjoblom, 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.
Analytical Spectral Devices, 2001. Using NIR Spectroscopy to Determine Hexane Purity. Application Note 1013.01E.
Antunes, M.C., Simao, J.E.J., Duarte, A.C., Tauler, R., 2002. Multivariate curve resolution of overlapping voltammetric peaks: quantitative analysis of binary and quaternary metal mixtures. Analyst 127, 809–817.
Azzouz, T., Tauler, R., 2008. Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples. Talanta 74, 1201–1210.
Balabin, R.M., Safieva, R.Z., 2007. Capabilities of near infrared spectroscopy for the determination of petroleum macromolecule content in aromatic solutions. J. Near Infrared Spectrosc. 15, 343–349.
Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777.
Baugh, A.L., Lovergreen, J.R., 1990. Differentiation of crude oils and redefined petroleum products in soil. In: Kostecki, P.T., Calabese, E.J. (Eds.), Petroleum Contaminated Soils, vol. 3. Lewis Publishers, Boca Raton, FL, pp. 141–163.
Ben-Dor, E., Inbar, Y., Chen, Y., 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. Remote Sens. Environ. 61, 1–15.
Bishop, J.L., Pieters, C.M., Edwards, J.O., 1994. Infrared spectroscopic analyses on the nature of water in montmorillonite. Clays Clay Miner. 42, 702–716.
Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. Roy. Stat. Soc. Ser. B. Methodol. 26, 211–252.
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.
Buddenbaum, H., Steffens, M., 2012. The effects of spectral pre-treatments on chemometric analyses of soil profiles using laboratory imaging spectroscopy. Appl. Environ. Soil. Sci.. http://dx.doi.org/10.1155/2012/274903.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Bricklemyer, R.S., Brown, D.J., 2010. On-the-go VisNIR: potential and limitations for mapping soil clay and organic carbon. Comput. Electron. Agric. 70, 209–216.
Camilli, R., Reddy, C.M., Yoerger, D.R., Van Mooy, B.A.S., Jakuba, M.V., Kinsey, J.C., McIntyre, C.P., Silva, S.P., Maloney, J.P., 2010. Tracking hydrocarbon plume transport and biodegradation at Deepwater Horizon. Science 330, 201–204.
Chakraborty, S., Weindorf, D.C., Zhu, Y., Li, B., Morgan, C.L.S., Ge, Y., Galbraith, J., 2012a. Spectral reflectance variability from soil physicochemical properties in oil contaminated soils. Geoderma 177–178, 80–89.
Chakraborty, S., Weindorf, D.C., Zhu, Y., Li, B., Morgan, C.L.S., Ge, Y., Galbraith, J., 2012b. Assessing spatial variability of soil petroleum contamination using visible near-infrared diffuse reflectance spectroscopy. J. Environ. Monit. 14, 2886–2892.
Chakraborty, S., Weindorf, D.C., Morgan, C.L.S., Ge, Y., Galbraith, J.M., Li, B., Kahlon, C.S., 2010. Rapid identification of oil-contaminated soils using visible near-infrared diffuse reflectance spectroscopy. J. Environ. Qual. 39, 1378–1387.
Chang, C., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near infrared reflectance spectroscopy: principal components regression analysis of soil properties. Soil. Sci. Soc. Am. J. 65, 480–490.
Christy, C.D., 2008. Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. Comput. Electron. Agric. 61, 10–19.
Clesceri, L.S., Greenberg, A.E., Eaton, A.D. (Eds.), 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.
CRC, 1986. Handbook of Chemistry and Physics, 67th ed. CRC Press, Boca Raton, FL.
Demetriades-Shah, T.H., Steven, M.D., Clark, J.A., 1990. High-resolution derivative spectra in remote sensing. Remote Sens. Environ. 33, 55–64.
Dent, A., Young, A., 1981. Soil Survey and Land Evaluation. George Allen & Unwin Publ., Boston, MA.
Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-spline and penalties (with comments and rejoinder). Stat. Sci. 11, 89–121.
Forrester, S., Janik, L., McLaughlin, M., 2010. An infrared spectroscopic test for total petroleum hydrocarbon (TPH) contamination in soils. In: Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia, August 1–6, 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. Trans. ASABE 50, 303–311.
Graham, K.N., 1998. Evaluation of Analytical Methodologies for Diesel Fuel Contaminants in Soil (M.Sc. thesis). The University of Manitoba, Canada (Unpublished results).
Hese, S., Schmullius, C., 2008. Object oriented oil spill contamination mapping in west Siberia with quickbird data. Lect. Notes Geoinf. Cartogr., 367–382.
Kooistra, L., Wehrens, R., Leuven, R.S.E.W., Buydens, L.M.C., 2001. Possibilities of visible-near-infrared spectroscopy for the assessment of soil contamination in river floodplains. Anal. Chim. Acta 446, 97–105.
Kumar, K., Mishra, A.K., 2012. Application of 'multivariate curve resolution alternating least square (MCR–ALS)' analysis to extract pure component synchronous fluorescence spectra at various wavelength offsets from total synchronous fluorescence spectroscopy (TSFS) data set of dilute aqueous solutions of fluorophores. Chemom. Intel. Lab. Syst. 116, 78–86.
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. Aust. J. Soil. Res. 46, 623–635.
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. Sediment. Contam. 8, 481–489.
McKenna, E.A., Youngren, S.H., Baker, S.R., Schroeder, J.R., Piccin, T.B., Weisman, W.H., Long, C., 1995. Evaluation of the total petroleum hydrocarbon (TPH) standard for JP-4 jet fuel. J. Soil. Contam. 4, 355–407.
Minasny, B., McBratney, A., Bellon-maurel, V., Roger, J.M., Gobrecht, A., Ferrand, L., Joalland, S., 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. Geoderma 167–168, 118–124.
Mouazen, A.M., De Baerdemaeker, J., Ramon, H., 2005. Towards development of online soil moisture content sensor using a fibre-type NIR spectrophotometer. Soil. Till. Res. 80, 171–183.
Mullins, O.C., Mitra-Kirtley, S., Zhu, Y., 1992. The electronic absorption edge of petroleum. Appl. Spectrosc. 46, 1405–1411.
Naes, T., Isaksson, T., Fearn, T., Davies, T., 2002. A User Friendly Guide to Multivariate Calibration and Classification. NIR Publications, Chichester, UK.
Okparanma, R.N., Mouazen, A.M., 2013. Combined effects of oil concentration, clay and moisture contents on diffuse reflectance spectra of diesel-contaminated soils. Water Air Soil. Pollut. 224 (5), 1539–1556.
Okparanma, R.N., Coulon, F., Mouazen, A.M., 2014. Analysis of petroleum-contaminated soils by diffuse reflectance spectroscopy and sequential ultrasonic solvent extraction-gas chromatography. Environ. Pollut. 184, 298–305.
Otto, O., 1998. Statistics and Computer Application in Analytical Chemistry. Wiley-VCH, Weinheim, Germany.
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 6 January 2014).
Sajda, P., 2006. Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng. 8, 537–565.
Saurina, J., Hernández-Cassou, S., Tauler, R., 1995. Multivariate curve resolution applied to continuous-flow spectrophotometric titrations. Reaction between amino acids and 1, 2-naphthoquinone-4-sulfonic acid. Anal. Chem. 67, 3722–3726.
Saurina, J., Hernández-Cassou, S., 2001. Quantitative determinations in conventional flow injection analysis based on different chemometric calibration strategies: a review. Anal. Chim. Acta 438, 335–352.
Schwartz, G., Ben-Dor, E., Eshel, G., 2012. Quantitative analysis of total petroleum hydrocarbons in soils: comparison between reflectance spectroscopy and solvent extraction by 3 certified laboratories. Appl. Environ. Soil. Sci. 2012, 1–11.
Skoog, D.A., West, D.M., Holler, F.J., 1995. Fundamentals of analytical Chemistry, seventh ed. Sanders Coll. Pub.
Soil Survey Staff, 2005. Official soil series descriptions. Available online at: http://soils.usda.gov/technical/classification/osd/index.html. NRCS, Washington, DC. (Verified 26 Feb. 2014).
Vasques, G.M., Grunwald, S., Sickman, J.O., 2009. Modeling of soil organic carbon fractions using visible-near-infrared spectroscopy. Soil. Sci. Soc. Am. J. 73, 176–184.
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.
Weindorf, D.C., Chakraborty, S., Zhu, Y., Galbraith, J., Ge, Y., 2011. New Technologies in Field Soil Survey. Applied Industrial Optics: Spectroscopy, Imaging, and Metrology [AIO]. July 10–14, Toronto, Canada.
Windig, W., 1992. Self-modeling mixture analysis of spectral data with continuous concentration profiles. Chemom. Intel. Lab. Syst. 16, 1–16.
Windig, W., Stephenson, D.A., 1992. Self-modeling mixture analysis of second-derivative near infrared spectral data using the SIMPLISMA approach. Anal. Chem. 62, 2735–2742.
Yang, Y., Hawthrone, S.B., Miller, D., 1995. Comparison of sorbent and solvent trapping after supercritical fluid extraction of volatile petroleum hydrocarbons from soil. J. Chromatogr. 699, 265–276.
Zhu, Y., Weindorf, D.C., Chakraborty, S., Haggard, B., Bakr, N., 2010. Characterizing surface soil water with field portable diffuse reflectance spectroscopy. J. Hydrol. 391, 133–140.