Journal of Environmental Monitoring

Cite this: DOI: 10.1039/c2em30330b

www.rsc.org/jem

PAPER

Assessing spatial variability of soil petroleum contamination using visible near-infrared diffuse reflectance spectroscopy

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

Received 27th April 2012, Accepted 7th September 2012
DOI: 10.1039/c2em30330b

Visible near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) is a rapid, non-destructive method for sensing the presence and amount of total petroleum hydrocarbon (TPH) contamination in soil. This study demonstrates the feasibility of VisNIR DRS to be used in the field to proximally sense and then map the areal extent of TPH contamination in soil. More specifically, we evaluated whether a combination of two methods, penalized spline regression and geostatistics could provide an efficient approach to assess spatial variability of soil TPH using VisNIR DRS data from soils collected from an 80 ha crude oil spill in central Louisiana, USA. Initially, a penalized spline model was calibrated to predict TPH contamination in soil by combining lab TPH values of 46 contaminated and uncontaminated soil samples and the first-derivative of VisNIR reflectance spectra of these samples. The r^2, RMSE, and bias of the calibrated penalized spline model were 0.81, 0.289 $\log_{10}$ mg kg^{-1}, and 0.010 $\log_{10}$ mg kg^{-1}, respectively. Subsequently, the penalized spline model was used to predict soil TPH content for 128 soil samples collected over the 80 ha study site. When assessed with a randomly chosen validation subset ($n = 10$) from the 128 samples, the penalized spline model performed satisfactorily ($r^2 = 0.70$; residual prediction deviation = 2.0). The same validation subset was used to assess point kriging interpolation after the remaining 118 predictions were used to produce an experimental semivariogram and map. The experimental semivariogram was fitted with an exponential model which revealed strong spatial dependence among soil TPH [$r^2 = 0.76$, nugget = 0.001 ($\log_{10}$ mg kg^{-1})^2, and sill 1.044 ($\log_{10}$ mg kg^{-1})^2]. Kriging interpolation adequately interpolated TPH with r^2 and RMSE values of 0.88 and 0.312 $\log_{10}$ mg kg^{-1}, respectively. Furthermore, in the kriged map, TPH distribution matched with the expected TPH variability of the study site. Since the combined use of VisNIR prediction and geostatistics was promising to identify the spatial patterns of TPH contamination in soils, future research is warranted to evaluate the approach for mapping spatial variability of petroleum contaminated soils.

[a]*IRDM Faculty Center, Ramakrishna Mission Vivekananda University, Kolkata 700103, India*
[b]*Louisiana State University Agricultural Center, Baton Rouge, LA, USA. E-mail: dweindorf@agcenter.LSU.edu*
[c]*Department of Experimental Statistics, Louisiana State University, LA, USA*
[d]*Texas Agrilife Research, College Station, TX, USA*
[e]*Virginia Tech, Blacksburg, VA, USA*

Introduction

Petroleum contamination of soil is a widespread problem that occurs frequently with adverse environmental and human health consequences.[1–3] Accidental release of crude oil and refined oil products from oil drilling rigs (such as the BP Deepwater Horizon), automobiles, immense oil tanker accidents (such as Exxon Valdez, Erika, and Prestige), and pipeline and storage

Environmental impact

Hydrocarbon contamination is commonplace in soils due to spillage and production mishaps. However, quantification of total petroleum hydrocarbon (TPH) contaminant levels is cumbersome, often requiring extensive soil sampling and laboratory analysis. This paper presents a case study using visible near infrared diffuse reflectance spectroscopy for rapid, on-site quantification of TPH. Coupled with global positioning system georeferencing, kriging interpolation was employed to produce spatial variability maps of TPH which could be used to more precisely target remediation efforts and document the temporal success of remediation.

tank leakages, endanger local and regional ecological systems.[4] The extent of environmental contamination by petroleum spillage depends on the capability of soil to filter, retain, biodegrade, and release petroleum.[5] Vapor pressure and solubility of crude oil and other organic fractions (*n*, iso, and cycloparaffins, napthene, and aromatics) also influence the dynamics of petroleum distribution in soil. Remediation specialists are constantly challenged by the need to measure spatial variation of total petroleum hydrocarbons (TPH) within and across a spillage area for site specific remediation practices.[6] Soil TPH contamination maps are generated using a large number of soil samples and traditional, laboratory-based soil chemical analyses. However, such soil analyses are laborious, costly, time consuming, and inadequate when high spatial and temporal resolution of TPH content are warranted.[7] Consequently, there is a persistent need for the development of innovative, low-cost, and reproducible analytical package for mapping spatial variability of petroleum contaminated soils.

Visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) can be used for detecting and mapping inland oil spills. This proximal optical sensor has already demonstrated its potential as a viable alternative to the laborious field sampling and expensive lab analysis for on-site quantification of TPH. Chakraborty *et al.* reported the feasibility of VisNIR DRS for predicting soil TPH with a validation r^2 of 0.64, RMSE of 0.341 $\log_{10}$ mg kg^{-1}, and residual prediction deviation (RPD) of 1.70.[8] Forrester *et al.* used a partial least squares (PLS) cross-validation approach for infrared spectroscopic identification of TPH with an r^2 of 0.81 and RMSE of 4653 mg kg^{-1}.[9] Moreover, Chakraborty *et al.* evaluated three types of clustering analysis (linear discriminant analysis, support vector machines, and random forest) and three multivariate regression methods (stepwise multiple linear regression, MLR; partial least squares regression, PLSR; and penalized spline) for pattern recognition and developing TPH predictive models.[10] Using VisNIR DRS for TPH measurement provides multiple benefits over traditional sampling/labwork: (1) results are returned to the investigator, on-site instantly, (2) the process is non-destructive allowing for sample preservation for future comparisons, and (3) minimization or elimination of traditional laboratory analyses saves considerable money over long periods of deployment.

While VisNIR DRS-based TPH predictions have been shown to be feasible for future use, spatial dependence among soil TPH contents has not received much consideration because of the current expense of measuring TPH in soils. In case of a field campaign that has a high spatial resolution of soil sampling, the spatial autocorrelation among soil TPH contents in a geographic space is likely and the spatial dependence can be used to create a map of contamination. A successful combination of geostatistics with VisNIR DRS could identify spatial correlation among soil TPH contents, faster than traditional soil physicochemical analysis. In precision agriculture, several researchers have proposed the combination of VisNIR DRS and multivariate geostatistics for improved spatial prediction of soil properties.[11,12] Hengl *et al.* reported a basic framework for spatial variability mapping of soil properties based on hybrid regression-kriging.[13]

Because of rapidity in prediction and incorporation of geostatistics, VisNIR spectroscopy could greatly enhance the spatial variability mapping of soil petroleum contamination. We combined two techniques: penalized spline regression and geostatistics. A penalized spline model to predict TPH contamination in soils was created and used to predict soil TPH content over a particular spill location. Predictions from the model were used to produce an experimental semivariogram and point kriging map. The objective of this study was to combine a soil TPH prediction model based on VisNIR spectroscopy with point kriging to identify the spatial distribution of TPH contamination at an 80 ha spill site. Through this objective, we demonstrate the feasibility of using VisNIR spectroscopy to proximally sense and map soil TPH contamination by testing the accuracy of that TPH prediction model for the site, modeling the spatial distribution of TPH contamination at the site, and estimating the accuracy of the TPH contamination map with 10 TPH measurements from the study site.

Materials and methods

Study area and soil sampling

The field chosen for mapping was a crude oil well blowout site located in Kisatchie National Forest in Vernon Parish, central Louisiana, USA (30° 59′ 23″ N, 93° 1′ 48″ W) (Fig. 1). One hundred and twenty-eight surface (0–15 cm) soil samples were collected within an 80 ha area that is densely vegetated by trees, shrubs, and grasses. The soils and sampling locations at the site are represented by four soil series: Caddo silt loam (fine-silty, siliceous, active, thermic 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).[14] A sampling scheme was designed in ArcGIS 9.3 by combining both grid and random sampling.[15] Sampling points were uploaded into a hand-held GPS receiver and geo-located in the field for sampling (location error approx. ±6 meters). All soil samples were sealed in air-tight plastic bags to prevent hydrocarbon volatilization and preserve field-moisture status before VisNIR scanning.

Calibration dataset for VisNIR prediction

Forty-six soil samples (including both contaminated and uncontaminated samples) were collected from six locations, each from a different parish within southern and central Louisiana;[8]

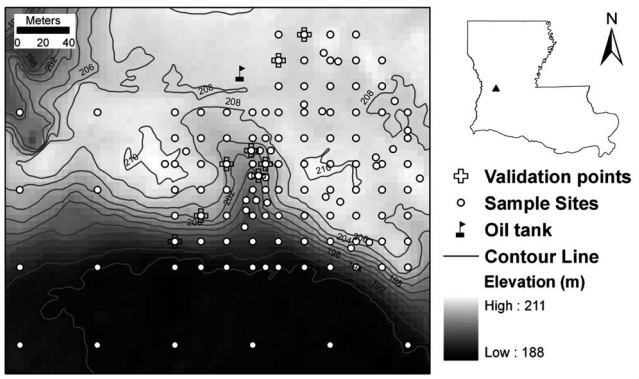

Fig. 1 The location, field boundary of the study site, and locations of collected soil samples in Louisiana, USA.

one of the six locations was the Kisatchie National Forest site. The sampling scheme was developed with the prior knowledge of oil spill locations supplied by the Louisiana Oil Spill Coordinators Office (LOSCO) to guarantee maximum TPH variability within the soil samples collected. The original TPH contents of the samples were widely and non-normally distributed from 44.3 to 48 188 mg kg^{-1} of soil. Crude oil was the main source of TPH. Soil textures at the six locations varied from clay to sandy loam. Twelve samples from the Ruston fine sandy loam were collected near, but not at the same location used in the validation dataset of this paper.

VisNIR spectroscopy and laboratory analyses

An AgriSpec VisNIR portable spectroradiometer (Analytical Spectral Devices, Boulder, CO) with a spectral range of 350 to 2500 nm (2 nm sampling resolution and a spectral resolution of 3 and 10 nm wavelengths from 350 to 1000 nm and 1000 to 2500 nm, respectively) was used to scan field-moist soil samples (both calibration samples and study site samples) with a contact probe. The contact probe had a circular viewing area (20 mm diameter) and its own halogen light source. Each sample was scanned four times with a 90° rotation between scans to obtain an average spectral curve. A spectralon panel with 99% reflectance was used every four samples to optimize and white reference the spectroradiometer to offset dark current and temperature effects. The spectroscopic reflectance measurement for each soil sample was then obtained by averaging the four raw scans.

A statistical analysis software package, R version 2.11.0 was used to preprocess raw reflectance spectra.[16] Based on a comparative analysis by Chakraborty *et al.*, only the reflectance and the first-derivative of reflectance spectra on 10 nm intervals were extracted using custom 'R' routines.[8,17] Spectroscopic reflectance splines and first derivative spectra of one contaminated (48 188 mg kg^{-1}) and one uncontaminated sample (44.3 mg kg^{-1}) from the calibration dataset are presented in Fig. 2. In general, mean spectral reflectance and first derivative spectra decreased as TPH concentration increased, as expected.[18] Note that the apparent spectral discrimination between contaminated and uncontaminated samples in the first derivative spectra was less than the mean reflectance spectra. However, derivative spectroscopy was preferred as it allows the detection and positive identification of trace levels of a component of interest (TPH in this paper) in the presence of a strongly absorptive matrix and corrects baseline shifts resulting from detector instabilities and faulty sample handling in the mean reflectance spectra, thereby increasing quantitative accuracy.

A validation subset with 10 samples was randomly selected from the 128 study site samples. In a commercial laboratory, TPH was measured gravimetrically for both the 46 calibration samples and 10 validation subset samples, following the method of Clesceri *et al.*[8,19]

Penalized spline model

The penalized spline calibration model was developed using the first-derivative of the reflectance spectra of the 46 contaminated and uncontaminated soils collected in six Louisiana parishes. Penalized spline is more stable and flexible than parametric

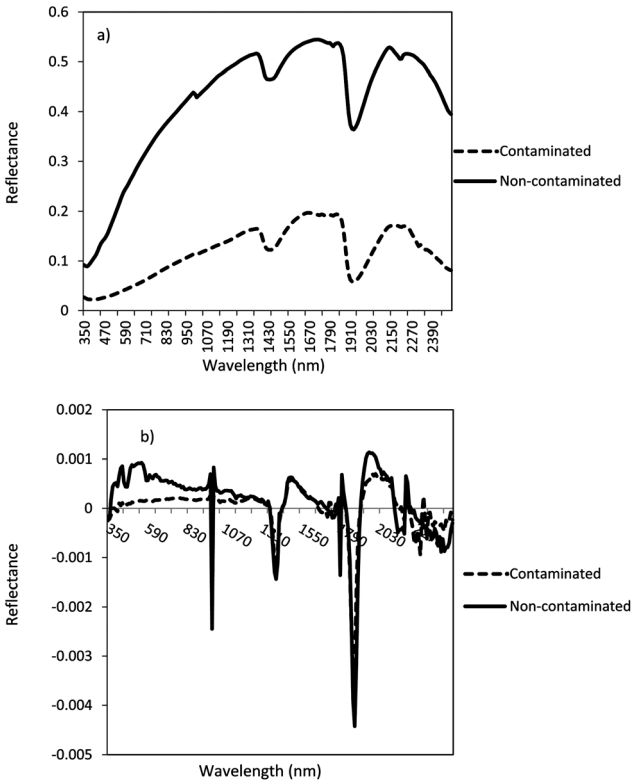

Fig. 2 (a) Reflectance and (b) first-derivative spectra for one petroleum contaminated (48 188 mg kg^{-1}) and one uncontaminated (44.3 mg kg^{-1}) soil sample from the calibration dataset.

principle components regression and partial least squares regression as the shape of the functional relationship amongst covariates and the dependent variable (TPH, in this study) is governed by the data.[20,21] Penalized spline 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.

Although a penalized spline model can handle both parametric and non-parametric data, transformation on the response variable is necessary because the TPH content of the samples is widely and non-normally distributed from 44.3 to 48 188 mg kg^{-1}. Therefore without transformation, the model results were highly affected by outliers. Vasques *et al.* also transformed on the response variable even though they had a large dataset for non-parametric models.[22] In the present study, the Box–Cox transformation was applied to the original TPH data using lambda = 0 (log_{10}-transformed) to bring the data to a more normal distribution (Fig. 3).[23] Thus, the penalized spline model was developed based on log_{10}-transformed data that approximated a Gaussian distribution after stabilizing the variance. As such, the remaining penalized spline model and kriging interpolation reported here all show log_{10}-transformed TPH.

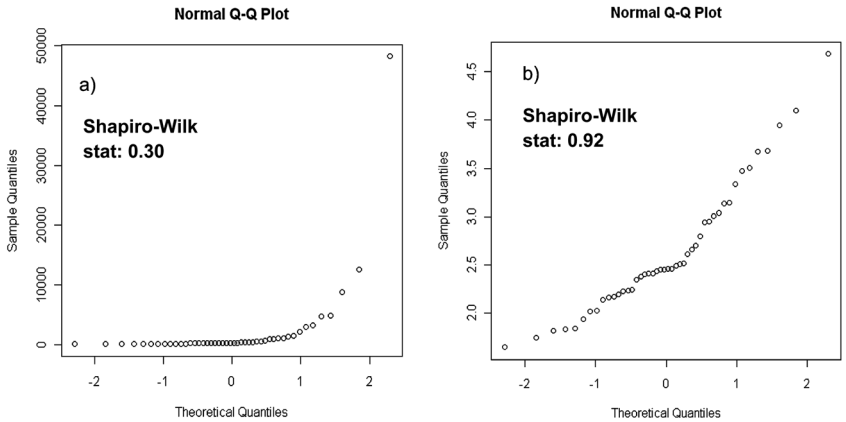

Fig. 3 Normal Q–Q plots of the (a) original ($\lambda = 1$) and (b) $\log_{10}$-transformed ($\lambda = 0$) total petroleum hydrocarbon (TPH) contents of the soil samples collected from six different parishes in Louisiana, USA and used to calibrate the penalized spline model to predict TPH. An increase in the Shapiro–Wilk statistic for log-transformed data revealed that the Box–Cox transformation made the original TPH data more normal.

For the penalized spline model, the cubic B-spline was used (using R version 2.11.0) as the base function with 100 equally spaced knots. The order of the penalty was set to the default value of three. The optimal value for the penalty-tuning parameter was selected by minimizing the leave-one-out cross-validation error. This penalized spline model was used to predict the TPH contents of the 128 study site samples, which include the subset of 10 that had TPH measurements. The ten validation samples were used to assess the performance of both the penalized spline predictive model and kriging exercise. The remaining 118 VisNIR-based penalized spline predictions were used for point kriging TPH soil contamination values across the study site.

Geostatistical analyses and kriging

A variogram and subsequently a kriged map of TPH were created using the 118 VisNIR-based penalized spline predictions. According to the usual method of moments, the experimental semivariogram of soil TPH at the Kisatchie National Forest site was calculated using,[24,25]

$$\gamma(h) = \frac{1}{2 \times n(h)} \sum_{i=1}^{n(h)} \left\{ z(s_i) - z(s_i + h) \right\}^2 \quad (1)$$

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

In the present study, point kriging was used to produce a TPH contour map. Note that the exponential model, as shown in eqn (2), best fit the experimental semivariogram for TPH in this study.

$$\gamma(h) = C_0 + C\left[1 - e^{-\left(h/A_0\right)}\right] \quad \text{when} \quad |h| > 0 \quad (2)$$

The objective criteria for measuring the goodness-of-fit for point kriging were the coefficient of determination (r^2) and RMSE. Calculated RMSE was derived with eqn (3):[12]

$$\text{RMSE} = \frac{1}{\bar{Z}_{li}} \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[Z_{pi} - Z_{li} \right]^2} \quad (3)$$

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

Results and discussion

The calibration set of soil samples used for penalized spline model building had the following range in soil properties, salinity (0–2.5 dS m^{-1}), pH (5.2–7.8), clay content (160–600 g kg^{-1}), and organic matter (9.3–130.5 g kg^{-1}) (see Chakraborty *et al.* for details[8]). No association between clay, organic matter, and TPH was found (both F-test and randomization test p-values were insignificant at the 0.05 and 0.10 significance level).

Calibration ($n = 46$) and validation ($n = 10$) datasets both had different mean TPH values (2.60 and 2.21 $\log_{10}$ mg kg^{-1}, respectively) but similar standard deviations (0.66 and 0.67 $\log_{10}$ mg kg^{-1}, respectively). The calibration statistics, using full cross

validation, resulted in a r^2 of 0.81, an RMSE of 0.289 $\log_{10}$ mg kg^{-1}, a bias of 0.010 $\log_{10}$ mg kg^{-1}, and an RPD of 1.77. The calibration results were especially encouraging given an RPD of 1.77, which indicates the calibration model is stable and has some predictability.[26] This calibration was created without using any soils from the actual contaminated site; therefore, the true performance of the model was difficult to evaluate on the calibration statistics alone. In the soil VisNIR literature, it is well established that the accuracy of VisNIR-based prediction of a soil constituent is closely related to the likeness of the calibration set to that of the validation of test set that is to be predicted.[27–29] This relationship is especially true when using intact, field-moist soil samples.[28]

The prediction of soil TPH of the $n = 10$ validation data was reasonably satisfactory with a r^2 of 0.70 and an RPD of 2.0. Other validation statistics, such as the RMSE and bias were 0.409 and 0.235 $\log_{10}$ mg kg^{-1}, respectively. The left panel of Fig. 4 shows the actual *versus* predicted TPH concentration using the penalized spline prediction model. The right panel of Fig. 4 shows the fitted coefficient curve from the penalized spline. It was apparent that the fitted coefficient curve was smooth across the spectrum, indicating the stability of the model. The grey-shaded band shows the 95% confidence interval for the coefficients and can be used to discover the region that has a coefficient significantly different from zero, and the impact of this region on the response. Based on the foregoing results, it can be concluded that soil TPH at this study site was reasonably predicted by the penalized spline model.

Fig. 5 shows the experimental semivariogram and the fitted, exponential model for the predicted TPH values of the 118 collected samples. At 200 m, the semivariance appears stable, *i.e.* the sill is not changing, indicating second-order stationarity of $\log_{10}$-transformed TPH. As shown, the isotropic experimental semivariogram was best fitted by an exponential model with a r^2 of 0.76. In addition, the effective range, nugget, sill, and nugget-to-sill ratio were 52 m, 0.001 ($\log_{10}$ mg kg^{-1})^2, 1.044 ($\log_{10}$ mg kg^{-1})^2, and 0.001, respectively, implying a strong spatial dependence of the TPH values according to the classification criteria reported by Cambardella *et al.*[30] Moreover, a nugget close to zero revealed that all variance of TPH was reasonably well explained, at the sampling distance used in this experiment.

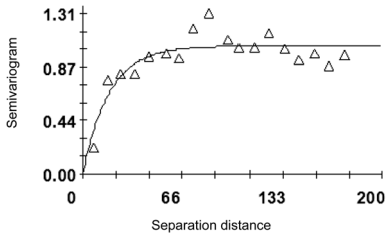

Fig. 5 Experimental semivariogram and fitted theoretical model of $\log_{10}$-transformed total petroleum hydrocarbon (TPH). As shown, the isotropic experimental semivariogram was best fitted by an exponential model with effective range, r^2, nugget, sill, and nugget-to-sill ratio of 52 m, 0.76, 0.001 ($\log_{10}$ mg kg^{-1})^2, 1.044 ($\log_{10}$ mg kg^{-1})^2, and 0.001, respectively.

Using the semivariogram of the predicted TPH values, a distribution map was developed using point kriging interpolation in ArcGIS 9.3 (Fig. 6). The prediction capability of point kriging (using penalized spline predictions) in terms of r^2 (0.88) was higher than simple penalized spline predictions in the validation subset, and had a smaller RMSE (0.312 $\log_{10}$ mg kg^{-1}) (Fig. 7). This slight improvement in our validation shows that the interpolation did not decrease predictability. Perhaps the experimental semivariogram which was fitted in a trial-and-error mode was optimal in reflecting the spatial pattern of TPH.

Furthermore, as shown in Fig. 6, the TPH distribution matched well with the topography of the study site (Fig. 1). The highest TPH values were found around the area where the oil spill occurred. In the valley (the middle section of the study site), significantly elevated TPH levels can also be identified readily as expected. After the occurrence, the spilled oil accumulated in the surrounding area and then naturally moved down to the area with lower elevations along with surface runoff and sediment. It should be noted that a low level of TPH was predicted in the valley and the elevated TPH values were also limited to within the valley. This implies that the spilled oil had a fairly low mobility and the impacts can be limited within a small area if proper remediation was implemented. Overall, the successful combination of VisNIR predictions and point kriging helped to improve TPH prediction accuracy. The interpolated TPH predictions matched the topography and the expectation well. The prediction based on VisNIR DRS could provide a fast way

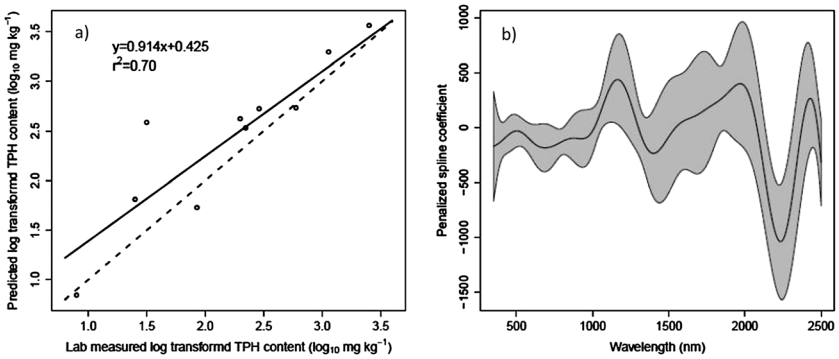

Fig. 4 The left panel shows actual *versus* predicted total petroleum hydrocarbon (TPH) ($\log_{10}$ mg kg^{-1}) using penalized splines. The dotted line is the 1 : 1 line. The right panel shows the fitted penalized splines (P-splines) coefficient curve at each waveband. The grey-shaded area is the 95% confidence interval.

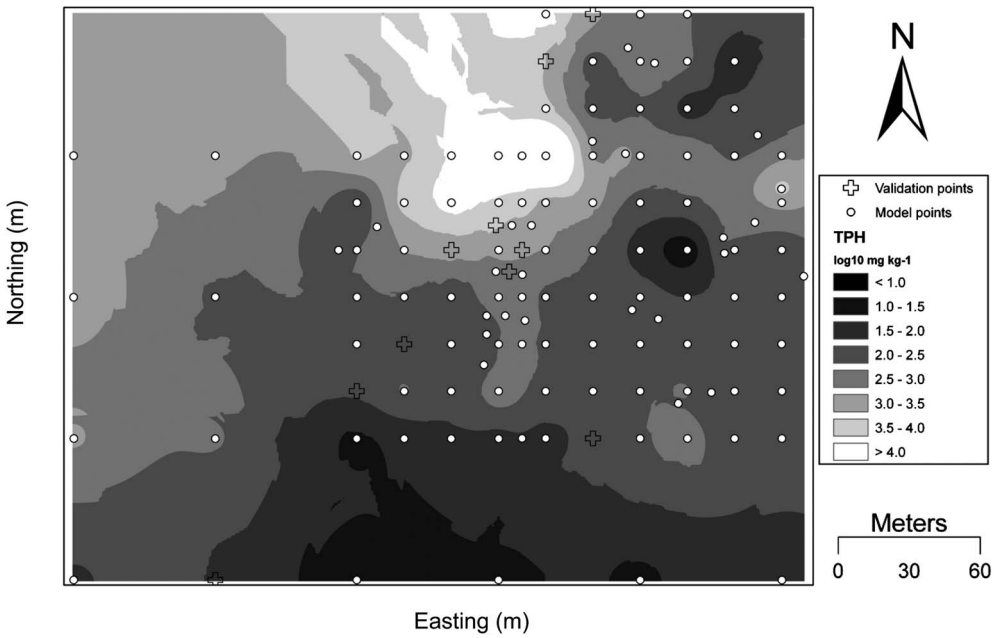

Fig. 6 Kriging map for $\log_{10}$-transformed total petroleum hydrocarbon (TPH) contamination of soils at the sampling site in Louisiana, USA.

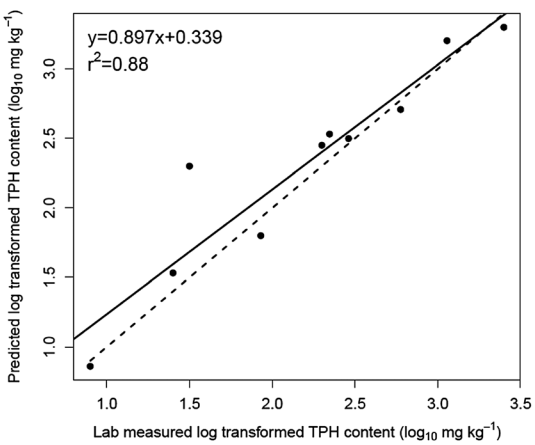

Fig. 7 Lab measured *versus* predicted (kriging interpolated) total petroleum hydrocarbon (TPH) ($\log_{10}$ mg kg^{-1}) for the validation subset ($n = 10$). The dotted line is the 1 : 1 line.

to understand the spatial distribution of TPH contamination, to estimate the impacted area, and finally to expedite the process of decision making after contamination occurred.

Data with a lognormal distribution, which is characterized by a positive skewness, pose a potential problem in kriging estimation. Journel revealed that experimental semivariograms are highly affected by lognormal data and there are only two solutions: trimming off high values or data transformation.[31] Yamamoto and Furuie reported that logarithmic data transformation is always a better solution than data trimming in geostatistics.[32] However, kriging approximations in the transformed domain need to be back-transformed into the original scale to obtain an unbiased result, after correcting the smoothing effect.[33] In this present study, kriging estimates have not been back transformed. When we developed the penalized spline calibration model with lognormal data, we tried to fit the linear relationship in eqn (4):

$$Y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p + e \quad (4)$$

where Y stands for the abundance of the response of interest (TPH) at the $\log_{10}$ space, and x_1, x_2, …, x_p are the spectral variables; p and b are the number of spectral variables and penalized spline coefficients for spectral variables, respectively. We attempted to identify the optimal fitting at the $\log_{10}$ space, *i.e.*, e (error) is i.i.d. (independently and identically distributed) normal at the $\log_{10}$ space. However, if back transformation is applied, it approximates the relationship (not mathematically strict but close) given in eqn (5):

$$\exp(Y) = \exp(b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p) \times \exp(e) \quad (5)$$

The error e would not be optimal at the original (non $\log_{10}$-transformed) space in the sense $\log_{10}$ Y was used instead of Y. It is noteworthy that this error is not additive, but rather multiplicative. To obtain the optimal solution (in the sense of ordinary least squares) of the original space, it is necessary to start with the non-transformed variable. However, that would have produced useless predictions, affected by outliers. Therefore, we conclude that the statistical relationship is not as simple as just doing the forward transformation then transformation back while the mathematical formulation indicates a far more complicated relationship. Improvements in non-parametric modeling could offset the outlier effect and handle the lognormal, limited data without transformation. Subsequently, VisNIR model predicted results could be transformed and interpolated with the option of back-transformation.

With VisNIR DRS, regression-kriging utilizes cheaply and quickly obtainable reflectance spectra of the target component (with spatial reference *via* GPS) as auxiliary co-variables instead

of landform attributes (such as slope, curvature, aspect, and elevation) derived from a digital elevation model. Nonetheless, a combination of VisNIR spectroscopy and point kriging appears to be a reliable and efficient strategy for determining the spatial patterns of TPH contamination in soils, providing information for unvisited locations.

Conclusions

In this pilot study, the VisNIR predicted TPH results were incorporated into point kriging to identify spatial patterns of soil TPH contamination. A penalized spline model was developed with full cross-validation to predict TPH contamination in soils using lab TPH data and VisNIR spectra (first-derivative only) from contaminated and uncontaminated soils from central and southern Louisiana. This penalized spline model was used to predict soil TPH content for 128 soil samples collected over an 80 ha crude oil spill location. Independent validation ($n = 10$) results showed that a penalized spline model could use VisNIR spectra to predict TPH with an r^2 of 0.70 and an RMSE of 0.409 $\log_{10}$ mg kg^{-1}. That same validation dataset was used to validate kriging interpolation of 118 sampling locations. The exponential semi-variogram model showed strong spatial dependence among soil TPH samples. Validation of the point kriging resulted in r^2 and RMSE values of 0.88 and 0.312 $\log_{10}$ mg kg^{-1}, respectively. In the kriging map, TPH distribution matched well with the topography of the study site. Overall, this study suggested that the combined use of VisNIR prediction and geostatistics have the potential to identify the spatial patterns of TPH contamination in soil quickly on site, reducing the need for expensive laboratory analyses.

Acknowledgements

The authors wish to thankfully acknowledge financial assistance from the Louisiana Applied Oil Spill Research Program (LAOSRP). The location of suitable sampling sites was provided by the Louisiana Oil Spill Coordinator's Office (LOSCO).

References

1 J. D. MacEwen and E. H. Vernot, *Toxic Hazards Research Unit Annual Technical Report: 1985*, Aerospace Medical Research Laboratory, 1985.
2 M. S. Hutcheson, D. Pedersen, N. D. Anastas, J. Fitzgerald and D. Silveman, *Regul. Toxicol. Pharmacol.*, 1996, **24**(1), 85–101.
3 P. Bofetta, N. Jourenkova and P. Gustavson, *Cancer, Causes Control*, 1997, **8**(3), 444–472.
4 E. J. Calabrese and P. T. Kostecki, *Soils Contaminated by Petroleum*, John Wiley & Sons, 1988.
5 P. Fine, E. R. Graber and B. Yaron, *Soil Technol.*, 1997, **10**, 133–153.
6 G. M. Cole, *Assessment and Remediation of Petroleum Contaminated Sites*, Lewis Publishers, 1994.
7 M. Odlare, K. Svensson and M. Pell, *Geoderma*, 2005, **126**, 193–202.
8 S. Chakraborty, D. C. Weindorf, C. L. S. Morgan, Y. Ge, J. Galbraith, B. Li and C. S. Kahlon, *J. Environ. Qual.*, 2010, **39**, 1378–1387.
9 S. Forrester, L. Janik and M. McLaughlin, *Proceedings 19th World Congress of Soil Science*, 2010.
10 S. Chakraborty, D. C. Weindorf, Y. Zhu, B. Li, C. L. S. Morgan, Y. Ge and J. Galbraith, *Geoderma*, 2012, **177–178**, 80–89.
11 A. V. Bilgili, F. Akbas and H. M. van Es, *Precis. Agric.*, 2011, **12**, 395–420.
12 Y. Ge, J. A. Thomasson, C. L. Morgan and S. W. Searcy, *Trans. ASABE*, 2007, **50**, 1081–1092.
13 T. Hengl, G. B. M. Heuvelink and A. Stein, *Geoderma*, 2004, **120**, 75–93.
14 Soil Survey Staff, *Official Soil Series Descriptions*, http://soils.usda.gov/technical/classification/osd/index.html, 2009.
15 Environmental Systems Research Institute, *ArcGIS Desktop: Release 9.3*, 2008.
16 R Development Core Team, *R: A Language and Environment for Statistical Computing*, 2008.
17 D. J. Brown, K. D. Shepherd, M. G. Walsh, M. D. Mays and T. G. Reinsch, *Geoderma*, 2006, **132**, 273–290.
18 B. Hoerig, F. Kuehn, F. Oschuetz and F. Lehnann, *Int. J. Remote Sens.*, 2001, **8**, 1413–1422.
19 L. S. Clesceri, A. E. Greenberg and A. D. Eaton, *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association, American Water Work Association, and Water Environment Federation, 20th edn, 1998.
20 B. D. Marx and P. H. C. Eilers, *Technometrics*, 1999, **41**, 1–13.
21 C. M. Crainiceanu, D. Ruppert and M. P. Wand, *J. Statistical Software*, 2005, **14**, 1–24.
22 G. M. Vasques, S. Grunwald and J. O. Sickman, *Soil Sci. Soc. Am. J.*, 2009, **73**, 176–184.
23 G. E. P. Box and D. R. Cox, *J. Roy. Stat. Soc. B*, 1964, **26**, 211–252.
24 G. Matheron, *Les Variables Regionalises et Leur Estimation*, Masson, 1965.
25 R. Webster and M. A. Oliver, *Geostatistics for Environmental Scientists*, Wiley, 2001.
26 C. Chang, D. A. Laird, M. J. Mausbach and C. R. Hurburgh, *Soil Sci. Soc. Am. J.*, 2001, **65**, 480–490.
27 D. J. Brown, R. S. Bricklemeyer and P. R. Miller, *Geoderma*, 2005, **129**, 251–267.
28 T. H. Waiser, C. L. S. Morgan, D. J. Brown and C. T. Hallmark, *Soil Sci. Soc. Am. J.*, 2007, **71**, 389–396.
29 C. L. S. Morgan, T. H. Waiser, D. J. Brown and C. T. Hallmark, *Geoderma*, 2009, **151**, 249–256.
30 C. A. Cambardella, T. B. Moorman, J. M. Novak, T. B. Parkin, D. L. Karlen, R. F. Turco and A. E. Konopka, *Soil Sci. Soc. Am. J.*, 1994, **58**, 1501–1511.
31 A. G. Journel, *Math. Geol.*, 1983, **15**, 445–468.
32 J. K. Yamamoto and R. D. A. Furuie, *Geociências*, 2010, **29**(1), 5–19.
33 J. K. Yamamoto, *Comput. Geosci.*, 2007, **11**, 219–234.