Spectral data mining for rapid measurement of organic matter in unsieved moist compost

Somsubhra Chakraborty,[1] David C. Weindorf,[2,*] Md. Nasim Ali,[1] Bin Li,[3]
Yufeng Ge,[4] and Jeremy L. Darilek[5]

[1]IRDM Faculty Centre, Ramakrishna Mission Vivekananda University, Kolkata 700103, India

[2]Louisiana State University Agricultural Center, Baton Rouge, Louisiana 70803, USA

[3]Department of Experimental Statistics, Louisiana State University, Baton Rouge, Louisiana 70803, USA

[4]Texas A&M University, 201 Scoates Hall, Mail Stop 2117, College Station, Texas 77843, USA

[5]Key Laboratory of Soil Environment and Pollution Remediation, Chinese Academy of Sciences, Nanjing 210008, China

*Corresponding author: Dweindorf@agcenter.lsu.edu

Received 10 July 2012; revised 23 December 2012; accepted 23 December 2012;
posted 2 January 2013 (Doc. ID 172213); published 30 January 2013

Fifty-five compost samples were collected and scanned as received by visible and near-IR (VisNIR, 350–2500 nm) diffuse reflectance spectroscopy. The raw reflectance and first-derivative spectra were used to predict $\log_{10}$-transformed organic matter (OM) using partial least squares (PLS) regression, penalized spline regression (PSR), and boosted regression trees (BRTs). Incorporating compost pH, moisture percentage, and electrical conductivity as auxiliary predictors along with reflectance, both PLS and PSR models showed comparable cross-validation r^2 and validation root-mean-square deviation (RMSD). The BRT–reflectance model exhibited best predictability (residual prediction deviation = 1.61, cross-validation $r^2 = 0.65$, and RMSD = $0.09\,\log_{10}\%$). These results proved that the VisNIR–BRT model, along with easy-to-measure auxiliary variables, has the potential to quantify compost OM with reasonable accuracy. © 2013 Optical Society of America

OCIS codes: 280.1415, 300.6340, 130.6010.

1. Introduction

Compost is an inherently variable product produced from a wide variety of organic source materials known as feedstocks. Worldwide, it serves as a means of recycling many types of green wastes for use as soil amendments and organic fertilizers. In the United States, "Test Methods for the Examination of Composting and Compost" (TMECC) provides the procedures and methods for compost analysis at certified labs as defined by the US Composting Council [1]. Organic matter (OM) is one of the major constituents in compost and plays an important role in the ability of compost to enrich soil and increase fertility in a wide variety of ways. OM percentage directly affects compost stability and market value. Currently, accurate quantitative analysis of OM must be conducted in laboratories using wet combustion [2] or dry combustion [3] methods, both of which determine the amount of organic C in a sample that is converted to OM percentage. Both methods must be performed in a laboratory setting and are laborious and time-consuming. The dry combustion method is very accurate and can be automated by commercially available instruments. Dry combustion requires air-drying, fine grinding the compost sample, and accurate quantification of compost moisture. Therefore, though the analysis is accurate and automated, preparation for the method is labor intensive.

1559-128X/13/040B82-11$15.00/0

Visible and near-IR diffuse reflectance spectroscopy (VisNIR DRS) is a tool that might eliminate the need to dry and fine-grind compost for OM analysis. In addition, VisNIR DRS is field portable and might remove the constraint of quantifying compost OM in the lab. It is a rapid, proximal sensing tool that has shown promise in an assortment of agronomic and waste management applications, including quantification of multiple soil properties, waste products, and environmental hazards [4–6].

The active bonds in OM in the VisNIR region (350–2500 nm) are the O–H, C–N, N–H, and C=O groups [7]. Models created from VisNIR DRS spectra have predicted organic C from air-dried, ground soil samples with root-mean-square deviations (RMSDs) ranging from 0.9 $\mathrm{g\,kg^{-1}}$ to 3.81 $\mathrm{g\,kg^{-1}}$ on small, localized areas [8–10]. Organic C models built using larger geographic areas had RMSDs ranging from 2.2 $\mathrm{g\,kg^{-1}}$ to 12.7 $\mathrm{g\,kg^{-1}}$ [11,12] and standard errors of prediction (SEPs) ranging from 1.34 $\mathrm{g\,kg^{-1}}$ to 4.4 $\mathrm{g\,kg^{-1}}$ [13–15]. Furthermore, *in situ* detection of soil organic C was done with an SEP of 2.3 $\mathrm{g\,kg^{-1}}$ on four test sites in Illinois [16]. A study revealed that OM absorbs IR light from 400 to 4000 $\mathrm{cm^{-1}}$ in the mid-IR range [17]. The spectra band 2930 $\mathrm{cm^{-1}}$ with a baseline from 3010 to 2800 $\mathrm{cm^{-1}}$ is caused by C–H stretching of the $\mathrm{CH_2}$ groups [18]. Another band used was 1590 $\mathrm{cm^{-1}}$ to detect elemental C in marine sediment, but this band did not appear without intensive grinding [19,20]. Some researchers have used wavelengths of 1744, 1870, and 2052 $\mathrm{cm^{-1}}$ to detect organic C [21], while others used 1736, 1766, and 2032 $\mathrm{cm^{-1}}$ [22].

Research at the Louisiana State University soils laboratory uses VisNIR DRS in field applications on agricultural soils. This tool is presently used to quantify soil C [11], soil organic C [17,19], and soil OM in marine sediments [23]. These works have identified useful techniques, methodologies, and limitations. It is conceivable that similar results could be obtained by using VisNIR DRS to quantify OM in compost. In fact, several studies have identified the use of VisNIR DRS on raw manure, which is a precursor of compost, and on compost itself with promising results [24,25]. Scientists used partial least squares (PLS) regression to calibrate their results on dairy manures and received very low RMSDs, showing that VisNIR DRS seems to be a good predictor of C and N in manure [26]. Moreover, a VisNIR–PLS model was viable for determining crude ash in handmade soil/manure mixes [27]. The accuracy of VisNIR DRS in quantifying crude ash content in feedyard manure was also explored [28]. Decent correlation values using multiple linear regression (MLR) statistics were obtained in [7] in determining the nutrient content of hog manure and manure-amended soils. Moreover, [29] used different mathematical pretreatments to analyze 11 different nutrients in manure. The best pretreatment method was based on the lowest SEP and the highest correlation value. They found that there was no overarching calibration data with universal application, but that VisNIR DRS was a potentially useful tool. Likewise, [30] used PLS regression to calibrate their findings on the nutrient content of pig manure. Their R (correlation coefficient) values and range: SEP ratios appeared to also support VisNIR DRS as a potential option for rapid analysis of manure. In addition, [31] used the MLR method and received high multiple correlation coefficients for C and N in compost made from tofu refuse. Scientists studied raw, stockpiled, and composted beef feedlot manure and found that VisNIR DRS was a viable rapid analysis tool for assessing nutrient availability, especially for C and N. This tool also appeared promising for the analysis of the composting process and temporal changes in nutrient content [32,33].

Because most of the studies done thus far apply directly to the analysis of manure rather than compost, more investigations on the true applicability of VisNIR DRS to compost nutrient analysis are warranted. Researchers have reported the potentiality of this technology in assessing microbial population, N content, C content, pH, electrical conductivity (EC), and OM content [32–35]. While correlating between predicted and measured values of compost ash percentage, [34] reported an r^2 of 0.85. Nevertheless, most of the studies were limited in monitoring the composting process rather than analyzing the material once it had reached a finished state before sale. Moreover, most VisNIR compost studies evaluated only dried and milled samples from specific feedstocks and composting methods. Hence, further study involving a sample set having a wide assortment of physical and chemical properties (such as moisture, pH, EC, OM percentage, etc.) is necessary. Since strong influence of moisture on soil reflectance in the shortwave-IR (1100–2500 nm) was identified by [36], the interaction between compost moisture and spectral assessment of compost quality needs further illustration too.

This article is a continuation of the work in [35], which demonstrated the feasibility of VisNIR DRS for rapid measurement of OM content in dried compost. In this study, we used a completely new dataset with following objectives: (1) incorporate boosted regression trees (BRTs) for predicting OM in composts with variable moisture content and compare the results to those of conventional VisNIR models and (2) test the applicability of the technology for *in situ* analysis by incorporating compost properties as auxiliary predictors along with spectral data. Since conventional PLS-based sensors are complicated for field use, the overall goal of this study was to identify some other alternatives to conventional VisNIR models, which would help in designing a realistic sensor configuration for a field person. Moreover, if VisNIR DRS proves to be a reliable method for the quantification of OM without any pretreatment, it could replace more time-consuming dry combustion analysis and aid in compost assessment *in situ*.

2. Materials and Methods

A. Compost Samples and Standard Laboratory Testing

Fifty-five dairy manure compost samples were collected from different producers in Erath County, Texas, USA. Thermophyllic composting methods prepared the various compost products in different windrows. The time of recharging was generally a few months. Each sample was hand-mixed well to ensure maximum homogeneity and maintained in plastic bags at 4°C. Subsequently, samples were tested according to standard TMECC [1] laboratory procedures. All samples were divided into four replicates, and each replicate was tested independently for OM (%). Moreover, moisture (%), EC (ds m^{-1}), and pH were evaluated to test the sample diversity, ensure the applicability of the technology to diverse compost products, and use them as auxiliary predictors along with reflectance spectra. Later the resulting data were averaged in a single dataset for individual parameters. The OM (%) of compost samples was analyzed following TMECC Method 0.50.7-A loss on ignition (LOI) using a Fisher Scientific Isotemp programmable forced-draft muffle furnace (Thermo Scientific Barnstead, Dubuque, IA). OM was computed according to [35]

$$\mathrm{OM} = (1 - \mathrm{Ash}_w / d_w) \times 100, \tag{1}$$

where OM is LOI organic matter (%), Ash_w is sample net weight (g) after ignition at 550°C, and d_w is sample net weight (g) after drying according to Method 03.09-A before ignition [1]. The analysis was run in triplicate to obtain an average for each sample. Method 04.11-A 1.5 (slurry) [1] was employed for determining compost pH and EC using an Orion 2-Star pH meter (Thermo Scientific, Waltham, MA) and model 4063CC digital salinity bridge (Traceable Calibration Control Company, Friendswood, TX), respectively. Moisture was assessed via Method 03.09-A [1]. Of the 55 samples, 10 samples were randomly selected and sent to a certified testing lab to validate the accuracy of our laboratory results. For most parameters, the lab-measured values fell inside the 95% confidence interval set by the certified lab.

B. Scanning with VisNIR DRS

In the laboratory, the 55 compost samples were scanned using a field portable ASD AgriSpec VisNIR spectrometer (Analytical Spectral Devices, CO, USA) with a spectral range of 350–2500 nm. The spectrometer had a 2 nm sampling interval and a spectral resolution of 3 and 10 nm wavelengths from 350 to 1000 nm and 1000–2500 nm, respectively. The samples were left intact without sieving to preserve the moist condition (as received). Samples were allowed to assume room temperature and then scanned with an ASD contact probe connected to the AgriSpec with a fiber-optic cable, having a 2 cm diameter circular viewing area and built-in halogen light source (Analytical Spectral Devices, CO, USA). The contact probe was inserted into the plastic bag that held the raw, unsieved compost, and full contact with the sample prevented outside interference. Each sample was scanned twice with a 45° rotation between scans to obtain an average spectral curve. Each individual scan was an average of 10 internal scans over a time of 1.5 s. The detector was white referenced (every five samples) using a white spectralon panel with 99% reflectance, ensuring that fluctuating downwelling irradiance could not saturate the detector.

C. Preprocessing of Spectral Data

Derivative spectroscopy was used to preprocess compost spectra for model development. Derivative spectra remove the baseline shift arising from detector inconsistencies, albedo, and sample handling [37]. If a spectrum is expressed as reflectance, R, as a function of wavelength, λ, the derivative spectra are calculated using

$$\text{Zero order,} \qquad R = f(\lambda) \tag{2}$$

$$\text{First order,} \qquad dR/d\lambda = f'(\lambda) \tag{3}$$

$$\text{Second order,} \qquad d^2R/d\lambda^2 = f''(\lambda). \tag{4}$$

Raw reflectance spectra were processed via a statistical analysis software package, R version 2.11.0 [38] using custom "R" routines [39]. 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, (iv) first derivatives at 10 nm intervals, and, subsequently, (v) second derivatives at 10 nm intervals. The resulting 10 nm average reflectance and first-derivative spectra were extracted and individually combined with the laboratory-measured OM. These processed data were used to build prediction models using PLS regression, penalized spline regression (PSR), and BRT algorithms.

D. Data Transformation and Principal Component Analysis

The original compost OM was widely and non normally (Shapiro–Wilk test, Lilliefors test, and Anderson–Darling test p-values were <0.05) distributed from 10% to 57.4%, with a few extreme and potentially influential values. Hence, the Box–Cox transformation [40] was applied to the original OM data using $\lambda = 0$ ($\log_{10}$-transformed) to bring the data to a more normal distribution after stabilizing the target variance, but without any monotonic transformation of predictor variables. Subsequently, eight models using three multivariate algorithms were compared for predicting compost OM ($\log_{10}\%$) using VisNIR spectra of 55 samples for two spectral pretreatments.

Principal component analysis (PCA) was used for dimensionality reduction. We further classified PCA scores into two clusters using k-means clustering [41]. Agglomerative hierarchical clustering was used with Ward's criterion [42] to select the number of clusters to help elucidate spectral features.

E. Partial Least Squares Regression Model

Chemometric PLS modeling has been successfully applied to VisNIR data through spectral decomposition [43]. Full-spectrum multivariate PLS combines the signal averaging advantages of PCA and classical least squares [44]. In the present study, two PLS models (both for reflectance and first derivative) with leave-one-out cross validation were built using Unscrambler 9.0 (CAMO Software, Woodbridge, NJ). Models with as many as 10 factors were considered, and the optimum model was determined by selecting the number of latent factors (rotations of principal components for a different optimization criterion) with the first local minimum in cross-validation RMSD (RMSDcv). Moreover, the significant wavelengths ($p < 0.05$) were plotted by "R" based on Tukey's jackknife variance estimate to identify what portions of the spectra were important for compost OM predictions for each spectral pretreatment.

F. Penalized Spline Model

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) [45]. It is well suited for ill-posed problems (dimensionality $\gg$ sample size) such as signal regression problems. In the present study, 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. The objective criteria for measuring both PLS and PSR prediction accuracy were cross-validation r^2, RMSDcv, residual prediction deviation (RPD), and model bias. Compost pH, EC, and moisture content were included as auxiliary predictors along with VisNIR spectra in both PLS and PSR to search for possible "improvement" of model predictability. In the case of PSR, two different models were built: a simple model utilizing only the VisNIR spectra (PSR) and another model incorporating auxiliary predictors + VisNIR spectra (we termed it "auxiliary PSR" or APSR).

G. Boosted Regression Tree Model

Friedman's gradient BRTs [46], also known as multiple additive regression trees, are a nonparametric data mining method that has recently been applied in soil science [39]. This model is very useful for selecting important variables and detecting their interactions, addressing missing values, minimizing the influence of a few ultrapowerful variables, and preventing overfitting. In BRTs, several hundred individual trees or stumps with a few terminal nodes contribute a small portion of the overall model, and the final model results summarize individual tree results as a whole, resulting in better prediction than a single model or assembly, such as bagging or traditional boosting. In the present study, BRTs were implemented using "auxiliary predictors + VisNIR spectra" in TreeNet (Salford Systems, San Diego, CA). To have a fair comparison with the PLS and PSR, we used OM ($\log_{10}$%) as the target with leave-one-out cross validation. A learn rate or "shrinkage" of 0.01 was used to control the rate at which the model is updated after each training step and to prevent overfitting. The subsample fraction was set to the default value of 0.5. Initially, 600 trees with six terminal nodes per tree were grown to capture higher order interactions. However, the number of trees was extended each time the "optimal" model was close to the maximum grown and when the RMSD value continued decreasing. The minimum number of training observations in terminal nodes was set to two. Tree pruning was based on least square error. Important predictors were selected based upon their corresponding scores, and two variable dependence plots were created to identify possible two-way interactions.

3. Results and Discussion

A. Compost Properties and Spectra

The summary statistics of all measured compost properties and their histograms are shown in Table 1 and Fig. 1, respectively. The OM was normally (Kolmogorov–Smirnov p-value = 0.2) distributed

Table 1. Descriptive Statistics of Measured Properties for 55 Compost Samples Analyzed with VisNIR DRS

								Correlation Matrix			
Variable	Mean	Std. Dev.	Min.	Max.	First Quartile	Median	Third Quartile	OM ($\log_{10}$ %)	Moisture (%)	pH	EC (dS m^{-1})
OM ($\log_{10}$ %)[a]	1.33	0.18	1.00	1.75	1.20	1.30	1.46	1.00	0.52[c]	−0.42[c]	0.41[c]
Moisture (%)	28.5	11.5	8.6	60.1	20.9	25.6	30.3	0.52[c]	1.00	−0.55[c]	0.19
pH[b]	8.6	0.5	7.3	9.6	8.3	8.7	9.1	−0.42[c]	−0.55[c]	1.00	−0.18
EC (dS m^{-1})[b]	3.4	2.6	0.3	11.0	1.7	2.8	4.0	0.41[c]	0.19	−0.18	1.00

[a] $\log_{10}$-transformed compost OM.

[b] Measured by Method 04.11-A 1.5 (slurry) [1].

[c] Statistically significant ($p < 0.05$) correlation coefficient.

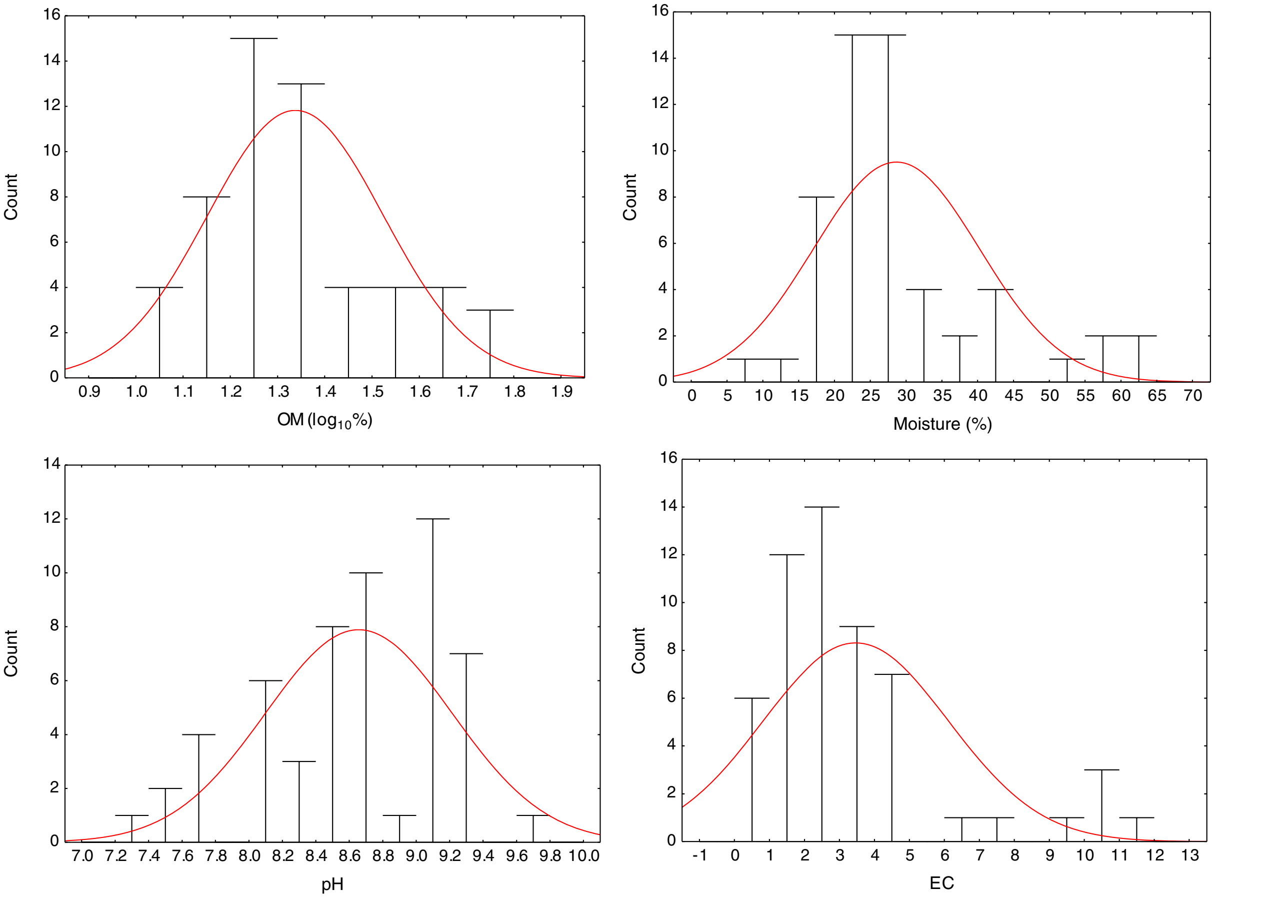

Fig. 1. (Color online) Histograms for measured compost properties used to calibrate PLS, penalized spline, and BRT models.

from 1.00 to 1.75 $\log_{10}\%$ with a mean of 1.33 $\log_{10}\%$ and standard deviation of 0.18 $\log_{10}\%$. Except for pH, considerable variability was observed for moisture content (8.60%–60.16%) and EC (0.36–11.09 $\mathrm{dS\,m^{-1}}$). OM was positively correlated with moisture ($\rho = 0.52$) and EC ($\rho = 0.41$) and negatively correlated with pH ($\rho = -0.42$). A significant negative correlation between moisture and pH ($\rho = -0.55$) was also observed.

For better interpretation of the compost spectral properties, the PCA scores were grouped into two classes using the k-means algorithm. Moreover, the average reflectance spectra across a 10-band window for two clusters are shown in Fig. 2(a). In general, the spectra of both clusters were similar, with very slight absorption features identified near 1730–1850 nm (methyl, $4\upsilon_1$) and 2137 nm (polysaccharides, $4\upsilon_1$), as previously identified with VisNIR DRS by [47]. Cluster 2 absorption near 877 nm was attributed to alkyl asymmetric–symmetric doublet ($4\upsilon_1$). The strong dips near 1412 and 1908 nm in the spectra of both clusters could also be suggestive of water ($3\upsilon_1$) or carboxylic acid ($3\upsilon_1$, $4\upsilon_1$). Note that one must use caution in how this region is interpreted, since these bands were broad and perhaps overlapping. A minor negative peak near 410 nm for soil OC confirmed previous research findings by [48]. We also qualitatively characterized compost reflectance spectra by interpreting negative and positive peaks associated with the component of interest and interfering components, respectively, at the specific wavelengths of the first three PLS loading weight vectors [Fig. 2(b)]. The aim was to identify the underlying correlation between spectral frequencies and compost OM through loading weight vectors. A moderate negative peak corresponding to the OC spectral signature was apparent at 410 nm, particularly for the second- and third-factor loadings. The first-factor loading weights showed a negative contribution for the whole VisNIR range, and the third-factor loading weights exhibited pronounced negative contributions for wave bands between 830 and 1430 nm from aromatics, with minor negative contributions for 1550–1800 nm for methyls. Positive contributions were found for 360–830 nm and >1870 nm. Conversely, the second-factor loading weights showed positive contributions for >830 nm, with minor interfering positive peaking to varying magnitudes at ~2030 nm (amides, $3\upsilon_1$) and 2275 nm (aliphatics, $3\upsilon_1$), with negative contributions for <830 nm. The shoulder at 2137 nm was due to polysaccharides, such as cellulose, which are part of the hard-to-decompose organic C. It is noteworthy that the lack of high-intensity spectral bands somewhat constrained the utility of qualitative analysis. That notwithstanding, it was obvious that even if fundamental vibration of organic molecules occurs

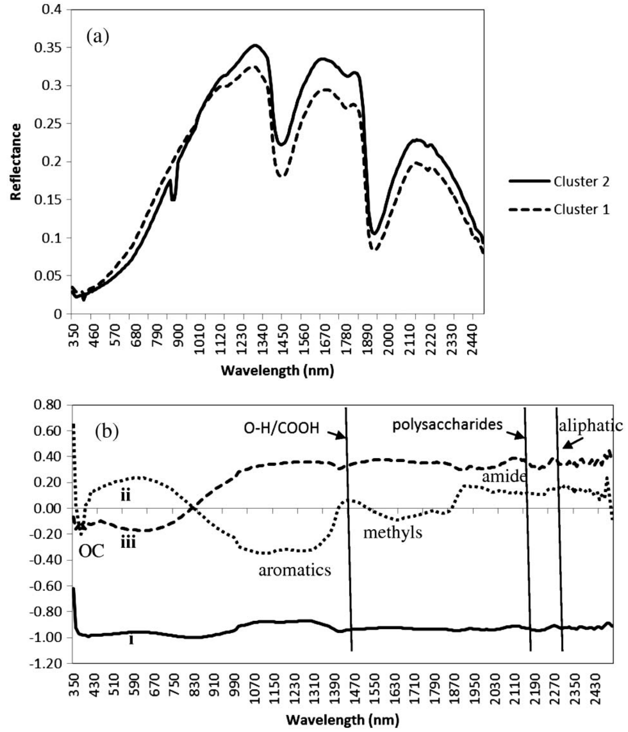

Fig. 2. Plots showing (a) VisNIR reflectance spectra of the two spectral clusters indicating spectral signatures of OM fractions and (b) the first three PLS regression factor loading weight vectors (i, ii, and iii) centered on zero for compost OM analyzed with VisNIR spectroscopy.

in the mid-IR region, relatively muted absorption features of their overtones and combination bands due to the stretching and bending of N–H, C–H, and C–O groups present in compost OM were identified by VisNIR DRS.

B. Model Predictions

Model stability and predictability were compared by a combination of both model generalization capacity (validation r^2, validation RMSD, and bias) and the RPD (ratio of standard deviation to RMSD) [49]. In the case of larger standard deviation, as compared to the estimation error (RMSD), the model predictability diminishes [50]. Given that the RPD-based guideline is best applicable when there is an independent validation set with leave-one-out cross validation, these values are still suitable indicators for describing the potential of the technology, particularly when considered with validation r^2 and supplementary error statistics like RMSD and bias.

The prediction accuracy and model parsimony for three different data mining algorithms are presented in Table 2. Among eight models tested, the first-derivative-based BRT model yielded the lowest validation r^2 (0.11) and was not acceptable for any applications. Predictions by reflectance-based PLS and APSR models were almost identical with cross-validation r^2 of 0.51 and 0.52 and RMSD of $0.13 \log_{10}\%$ and $0.12 \log_{10}\%$, respectively. Not shown, both PLS and APSR obtained satisfactory accuracy, with a coefficient of determination ranging from 0.7 to 0.8 using the whole dataset. For the PLS algorithm, the reflectance-based parsimonious model used four latent factors, whereas the first-derivative model used six latent factors. The significant regression coefficients (based on Tukey's jackknife variance estimate, $p < 0.05$) of the PLS–reflectance and PLS–first-derivative models are plotted in Fig. 3. Noticeably, both the number and intensity of significant wavelengths changed from reflectance to first-derivative models, specifically in the ~550–950, 1100–1400, 2100–2200, and 2300–2400 nm regions, which could contain the spectral signatures of aromatics [825 nm $(4\upsilon_1)$ and 1100 nm $(3\upsilon_1)$], amine [751 nm $(4\upsilon_1)$ and 1000 nm $(3\upsilon_1)$], alkyl asymmetric–symmetric doublets [853 nm $(4\upsilon_3)$, 877 nm $(4\upsilon_1)$, 1138 nm $(3\upsilon_3)$, and 1170 nm $(3\upsilon_1)$], polysaccharides [2137 nm $(4\upsilon_1)$], methyls [2307–2469 nm $(3\upsilon_1)$], carbohydrates [2381 nm $(4\upsilon_1)$], and water [940 nm $(2\upsilon_1 + \upsilon_3)$, 1135 nm $(\upsilon_1 + \upsilon_2 + \upsilon_3)$, and 1380 nm $(\upsilon_1 + \upsilon_3)$], as compiled by [47]. Thus, it was evident that higher spectral preprocessing smoothed out spectrally significant regions for model predictions. Notwithstanding that preprocessing transformations of the spectral data boost the accuracy of regression models, some researchers established better results with raw reflectance [50,51]. The remaining discussion of models for compost OM concerns the reflectance of the VisNIR spectra.

We had difficulty predicting OM based upon VisNIR spectra alone in the PSR models (both reflectance and first-derivative), which explained very

Table 2. Multivariate Model Results for 55 Compost Samples across the United States Evaluated for OM Using VisNIR DRS[a]

	PLS		PSR		APSR		BRT	
	Reflectance	First-Derivative	Reflectance	First-Derivative	Reflectance	First-Derivative	Reflectance	First-Derivative
Latent factors	4	6	—	—	—	—	—	—
Cross-validation r^2	0.55	0.47	0.40	0.36	0.52	0.48	0.65	0.11
RMSDcv ($\log_{10}\%$)	0.12	0.13	0.14	0.14	0.12	0.13	0.09	0.1
RPD	1.51	1.38	1.30	1.26	1.45	1.39	1.61	1.16
Bias ($10^{-15} \log_{10}\%$)	−0.23	−0.40	−0.62	0.49	0.11	0.24	−0.15	−0.7

[a] PLS, partial least squares; PSR, penalized spline regression; APSR, auxiliary penalized spline regression; BRT, boosted regression tree; *RMSDcv*, root mean squared deviation of cross validation; RPD, residual prediction of deviation.

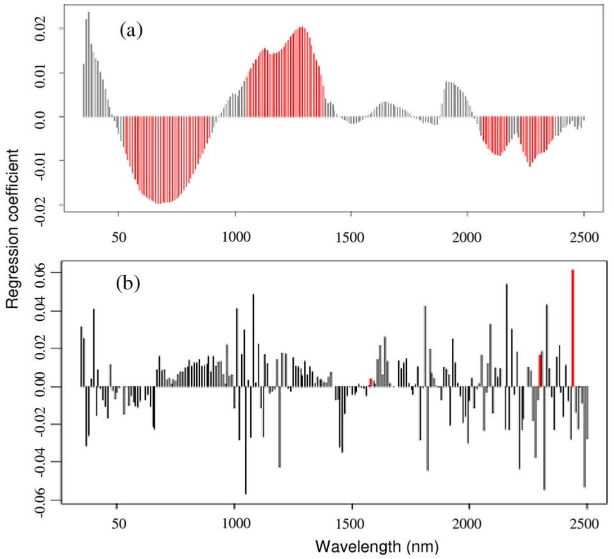

Fig. 3. (Color online) Regression coefficients (black) of the (a) reflectance and (b) first-derivative-based PLS model of compost samples. The magnitude of the regression coefficient at each wavelength is proportional to the height of the bar. Significant wave bands ($p < 0.05$) as indicated by Tukey's jackknife variance estimate procedure are shown in red. All plots are on the same x axis.

little of the variability. Plots of actual versus APSR- and PLS-predicted OM and fitted regression coefficient curves on the spectrum are presented in Fig. 4. In the APSR model, predictions of OM more closely approximated the 1:1 line and had less bias ($0.11 \times 10^{-15} \log_{10}\%$) than their PLS counterparts (Table 2). However, both model biases were very negligible. Thus, they accounted for a very trivial part of the overall lack of fit for cross validation. Note that the PLS model was less precise at both lower and higher concentrations and had a tendency to underestimate for high OM values, as specified by a regression slope value (<1) that was lower than that of APSR. However, the improvement was not impressive, since APSR also showed a lower regression slope value (<1) and subsequent underestimations of high OM values. Nonetheless, it should not be discounted as a feature of PLS and APSR, and a few of these underestimations could be due to the relative scarcity of estimated high OM values (a few samples with $>1.4 \log_{10}\%$ OM). Both PLS and APSR showed overestimation of low values too, with comparatively better predictions at middle-range values (1.2–$1.4 \log_{10}\%$). Several PLS predictions fell beyond 10% of the reference data and showed signs of residual heteroscedasticity, which was further confirmed

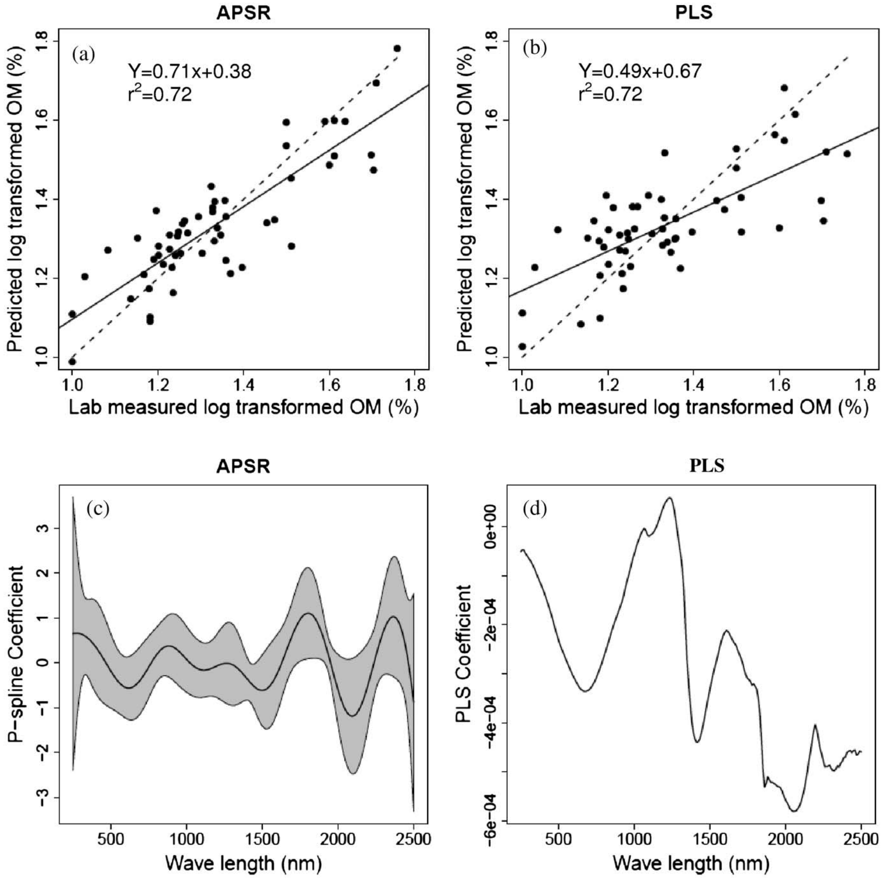

Fig. 4. Predicted versus measured OM ($\log_{10}\%$) for reflectance-based (a) auxiliary penalized spline (APSR) and (b) PLS models for 55 compost samples. The solid line is the regression line, and the dashed line is a 1:1 line. The fitted regression coefficient curves on the spectrum for reflectance-based (c) APSR and (d) PLS models are also shown.

when PLS residuals were plotted against the fitted values (not shown). Conversely, APSR model residuals were nearly homoscedastic, providing more credibility to this model. Moreover, while comparing the fitted coefficient curves, the APSR curve was smoother across the spectrum than PLS, indicating better stability in the former (Fig. 4). The gray-shaded band shows the 95% confidence interval for the coefficients and can be used to identify the region that has a coefficient significantly different from zero, and the impact of this region on the response. For example, the 400–600 and 1500–1700 nm regions were both away from zero. However, the former contributed a negative effect on the OM concentration, while the latter had a negative effect. The APSR estimator was more stable than that of nonpenalized PLS, since the APSR neighboring coefficients were *hand-in-hand* connected and PLS ignored the order of the regressor channels.

The BRT–reflectance model obtained the best fit among all the models, with the highest RPD of 1.61, highest cross-validation r^2 (0.65), and lowest cross-validation RMSD (0.09 $\log_{10}$%), indicating good generalization potential and room for further improvement by incorporating a larger dataset with parameterization. According to [10], RPD values between 1.4 and 2.0 indicate fair models that could be improved by more accurate predictive techniques. This result showed comparable accuracy to OM models developed using VisNIR produced elsewhere [13,52]. The optimal model was found to be 578 trees with 6 nodes per tree. Increasing the tree number (>600) and nodes per tree (6–9) did not radically change the model predictability; however, it did seem that using a complex model with 10 nodes per tree somewhat increased the MSD, leading to minor overfitting. The notable improvement in validation statistics in BRT as compared to PLS and APSR suggested the presence of a nonlinear and contingent relationship between VisNIR reflectance and compost composition. Such results were expected, given that BRT incorporates a complex, nonlinear relationship between target and predictor variables while PLS fits a linear relationship [39]. Cast in this light, Treenet produced two variable-dependence plots for OM with three auxiliary predictors (pH, EC, and moisture) (Fig. 5) to find any two-way interaction. Since optimum moisture content (30%–60%, wet weight basis) is essential for proper composting and may help in dissolving soluble salts [53], there might be a correlation between compost moisture and EC. This perhaps explains the interaction between compost moisture and EC for identifying compost OM. Laboratory estimation of compost pH, EC, and moisture content are quite straightforward and cheap, making these very practical auxiliary variables for incorporating into VisNIR models of compost OM and improving the overall prediction accuracy.

We also plotted the relative importance of BRT model predictor wavelengths, selected by Treenet (Fig. 6). The 100 most important wave bands were mainly concentrated around the 350–1450 nm region, with a pronounced presence at 350–550 nm (visible region), indicating the VisNIR sensitivity toward compost color. Interestingly, both PLS and BRT had important wavelengths roughly in the same regions, as shown by the relatively large BRT scores around 1000 (amine, $3\upsilon_1$), 1150, and 1250 nm. Although the locations of 1150 and 1250 nm were a bit shifted from the exact anticipated positions (1138 and 1170 nm, respectively), it was natural in the sense that real molecules do not behave totally harmonically [54]. The BRT model also had a significant wavelength near 1900 nm, which is the region of spectral signature of the OH of water or carboxylic acids. However, it was not feasible to select the precise spectral signature with confidence, due to the

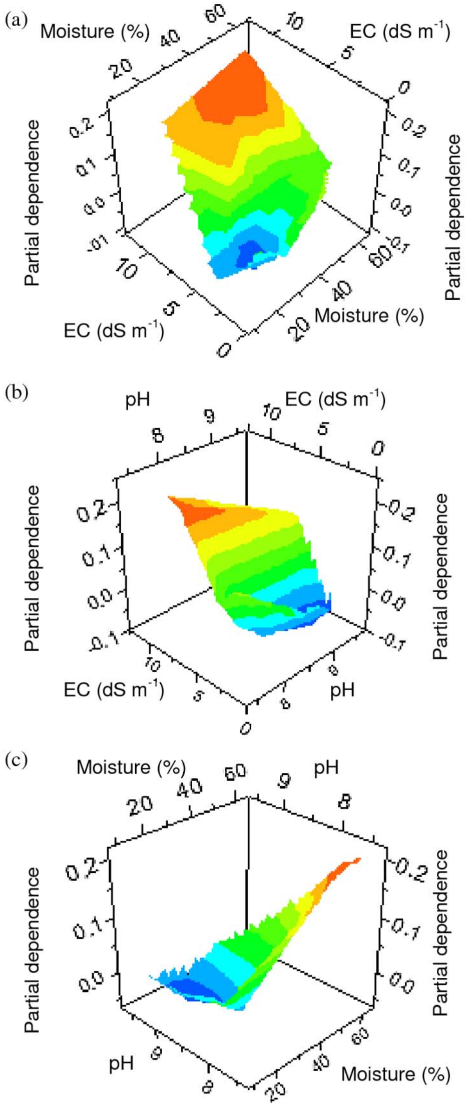

Fig. 5. (Color online) Two-variable dependence plots of compost OM ($\log_{10}$%) for (a) EC versus moisture, (b) EC versus pH, and (c) pH versus moisture. Blue and red shades (bottom and top of data shown) represent lowest and highest concentration of OM, respectively.

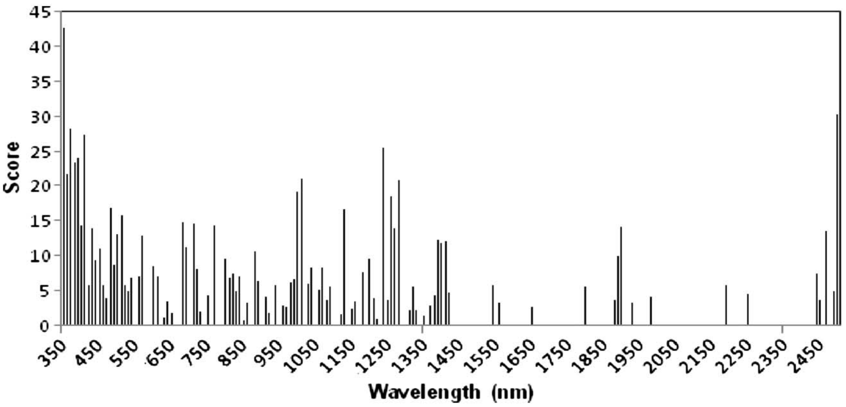

Fig. 6. Plot showing relative importance of BRT model predictor wavelengths. The magnitude of the score at each wavelength is proportional to the height of the bar.

large number of overlapping absorptions in the VisNIR range [55,56]. Testing the probable decrease in predictability due to the masking effect of the water signature was beyond the scope of this project and requires more investigations before drawing stronger conclusions. Moreover, another inherent problem for selecting the exact spectral signature by BRT might be spectral redundancy, as indicated by [39]. Treenet handles missing wave bands by substituting "surrogate splitters," backup rules that closely imitate the action of primary splitting rules (wave bands) without substantially affecting the prediction accuracy, creating diagnostic problems.

C. Practical Concern

Although the BRT model remained superior to PLS and APSR, it was by no means exhaustive and perhaps requires large data sets with a wide assortment of compost compositions before stronger conclusions can be drawn. The 55 values contained in the model may have been too limiting to produce the best empirical VisNIR compost characterization. Interestingly, [57] concluded that BRT also works very well for small datasets, since it helps in variable selection through recursive feature elimination where redundant predictors may degrade performance by increasing variance. Complex spectral interaction with background materials, as expected, was prohibitive in assigning precise wave bands and thus requires refined spectral preprocessing, such as discrete wavelet transformation and subsequent waveband-scale tiling so that the central wavelengths and scales of wavelet regressors become readily perceptible, facilitating physical interpretation of the model. Testing models for the prediction of specific organic functional groups was not performed, as wet chemistry determinations of fractions of OM (lignins, proteins, and other biopolymers) are themselves hampered by technical hurdles. Thus, a side-by-side comparison of spectral and analyte data is often not practical [58]. The preference of sample pretreatments is chiefly based on technology cost, prediction accuracy, and amount of samples. The fact that unprocessed compost samples yield an acceptable VisNIR–BRT predictive model would be more important when considering a VisNIR sensor system for *in situ* compost characterization. While other studies [27,28] showed better performances in air-dried samples, [39] found no discernible differences between accuracies of more (oven-dried) and less preprocessed (air-dried) models. Furthermore, [4,43] did not find any added advantage in more preprocessed models in other VisNIR applications. Note that most VisNIR compost studies utilized only dried, milled samples from specific feedstocks and composting methods. Most studies did not analyze compost as received. The difficulty is that diverse studies demonstrate diverse results, since the nature of the target function strongly controls the performance of the different prediction approaches. In precision agriculture, the development of electro-optical sensors based on PLS models is always problematic, since it involves hundreds of filter–detector pairs with variable central wavelengths and a constant bandwidth of 10 nm and an extra circuit block to combine these detectors' output signals into 18 synthetic signals, undoubtedly complicating the sensor design for field use [59]. Summarily, our study clearly identified the potential of the VisNIR–BRT model as a viable alternative to the VisNIR–PLS model for rapid and low-cost estimation of compost OM as an addition to the certified methods for compost analysis.

4. Conclusions

Using a rapid, cost-effective sensing method to characterize organic constituents in compost has many benefits. This pilot study utilized 55 compost samples and suggested an alternative approach to conventional models for estimating compost OM from spectral reflectance in the VisNIR range. OM is a quality parameter of compost and was estimated with reasonable accuracy by BRT (RPD = 1.61). In contrast, PLS (RPD = 1.51) and APSR (RPD = 1.45) had intermediate predictive power. However, it was difficult to get high prediction accuracy with first-derivative spectra. Since OM is a costly and laborious property to measure, the VisNIR model could offer a good estimate of it in a cost-effective way. Auxiliary compost properties that can be measured quickly and easily improved OM predictive models when used along with the spectra. More improvement could be achieved by increasing sample numbers or with an advanced spectral treatment, such as wavelet as an alternative to "black-box" first-derivative modeling. Clearly, more fundamental investigations as to how compost OM influences optical properties are warranted. Our study showed good potential as an impetus toward future VisNIR–BRT-based compost studies. Composts are very complex, and real-time compost OM characterization is expected to be complex as well. Our future research will be directed toward developing a general model, so that precise spectral features linked with compost OM can be identified and modeled as appropriate, reflecting divergent compost compositions.

The authors gratefully acknowledge Tarleton State University for access to the compost samples used in this study and Dr. C. Morgan of Texas Agrilife Research for use of the VisNIR DRS used in this study.

References

1. USDA-USCC, "Test methods for the evaluation of composts and composting" (CD ROM) (Composting Council Research and Education Foundation, 2002).
2. A. Walkley and I. A. Black, "An examination of the Degtjareff method for determining organic carbon in soils: effect of variations in digestion conditions and of inorganic soil constituents," Soil Sci. **63**, 251–264 (1934).
3. D. W. Nelson and L. E. Sommers, "Total carbon, organic carbon and organic matter," in *Methods of Soil Analysis, Part 3: Chemical Methods*, J. M. Bigham, ed. (ASA, 1996), pp. 961–1010.
4. S. Chakraborty, D. C. Weindorf, C. L. S. Morgan, Y. Ge, J. M. Galbraith, and C. S. Kahlon, "Rapid identification of oil-contaminated soils using visible near-infrared diffuse reflectance spectroscopy," J. Environ. Qual. **39**, 1378–1387 (2010).
5. D. C. Weindorf, J. P. Muir, and C. Landeros-Sánchez, "Organic compost and manufactured fertilizers: economics and ecology," in *Integrating Agriculture, Conservation, and Ecotourism: Examples from the Field (Issues In Agroecology—Present Status and Future Prospectus 1)*, W. B. Campbell and O. S. Lopez, eds. (Springer, 2011), pp. 27–53.
6. S. Chakraborty, D. C. Weindorf, Y. Zhu, C. L. S. Morgan, Y. Ge, and J. M. Galbraith, "Spectral reflectance variability from soil physicochemical properties in oil contaminated soils," Geoderma **177–178**, 80–89 (2012).
7. D. F. Malley, L. Yesmin, and R. G. Eilers, "Rapid analysis of hog manure and manure-amended soils using near-infrared spectroscopy," Soil Sci. Soc. Am. J. **66**, 1677–1686 (2002).
8. J. B. Reeves, G. W. McCarty, and J. J. Meisinger, "Near infrared reflectance spectroscopy for the analysis of agricultural soils," J. Near Infrared Spectrosc. **7**, 179–193 (1999).
9. W. S. Lee, J. F. Sanchez, R. S. Mylavarapu, and J. S. Choe, "Estimating chemical properties of Florida soils using spectral reflectance," Trans. ASAE **46**, 1443–1453 (2003).
10. C. Chang, D. A. Laird, and C. R. Hurburgh, "Influence of soil moisture on near-infrared reflectance spectroscopic measurement of soil properties," Soil Sci. **170**, 244–255 (2005).
11. G. W. McCarty, J. B. Reeves, V. B. Reeves, R. F. Follett, and J. M. Kimble, "Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement," Soil Sci. Soc. Am. J. **66**, 640–646 (2002).
12. K. D. Shepherd and M. G. Walsh, "Development of reflectance spectral libraries for characterization of soil properties," Soil Sci. Soc. Am. J. **66**, 988–998 (2002).
13. E. Ben-Dor and A. Banin, "Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties," Soil Sci. Soc. Am. J. **59**, 364–372 (1995).
14. B. W. Dunn, H. G. Beecher, G. D. Batten, and S. Ciavarella, "The potential of near-infrared reflectance spectroscopy for soil analysis—a case study from the Riverine Plain of south-eastern Australia," Aust. J. Exp. Agric. **42**, 607–614 (2002).
15. K. Islam, B. Singh, and A. McBratney, "Simultaneous estimation of several soil properties by ultraviolet, visible, and near-infrared reflectance spectroscopy," Aust. J. Soil Res. **41**, 1101–1114 (2003).
16. K. A. Sudduth and J. W. Hummel, "Soil organic matter, CEC, and moisture sensing with a portable NIR spectrophotometer," Trans. ASAE **36**, 1571–1582 (1993).
17. M. Kang, "Quantification of soil organic carbon using mid- and near-DRIFT spectroscopy," Master's thesis (Texas A&M University, 2002).
18. S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna, "Inspect: fast and accurate identification of post-translationally modified peptides from tandem mass spectra," Anal. Chem. **77**, 4626–4639 (2005).
19. J. W. T. Tung, "Determination of metal components in marine sediments using energy dispersive x-ray fluorescence (ED-XRF) spectrometry," Ann. Chim. **94**, 837–846 (2004).
20. D. M. Smith, J. J. Griffin, and E. D. Goldberg, "Spectrometric method for the quantitative determination of elemental carbon," Anal. Chem. **47**, 233–238 (1975).
21. R. C. Dalal and R. J. Henry, "Simultaneous determination of moisture, organic carbon and total nitrogen by near infrared reflectance spectrophotometry," Soil Sci. Soc. Am. J. **50**, 120–123 (1986).
22. M. J. Morra, M. H. Hall, and L. L. Freeborn, "Carbon and nitrogen analysis of soil fractions using near-infrared reflectance spectroscopy," Soil Sci. Soc. Am. J. **55**, 288–291 (1991).
23. L. Tremblay and J. Gagné, "Fast quantification of humic substances and organic matter by direct analysis of sediments using DRIFT spectroscopy," Anal. Chem. **74**, 2985–2993 (2002).
24. E. K. Kemsley, H. S. Tapp, A. J. Scarlett, S. J. Miles, R. Hammond, and R. H. Wilson, "Comparison of spectroscopic techniques for the determination of Kjeldahl and ammoniacal nitrogen content of farmyard manure," J. Agric. Food Chem. **49**, 603–609 (2001).
25. W. Saeys, A. M. Mouazen, and H. Ramon, "Potential for on-site and on-line analysis of hog manure using visual and near-infrared reflectance spectroscopy," in *Precision Livestock Farming'05*, S. Cox, ed. (Wageningen Academic, 2005), pp. 131–138.
26. J. B. Reeves and J. S. Van Kessel, "Determination of ammonium-*N*, moisture, total *C* and total *N* in dairy manures using a near infrared fibre-optic spectrometer," J. Near Infrared Spectrosc. **8**, 151–160 (2000).
27. S. L. Preece, C. L. S. Morgan, B. W. Auvermann, K. Wilke, and K. Heflin, "Determination of ash content in solid cattle manure with visible near-infrared diffuse reflectance spectroscopy," Trans. ASABE **52**, 609–614 (2009).
28. S. L. P. Sakirkin, C. L. S. Morgan, and B. W. Auvermann, "Effects of sample processing on ash content determination in solid cattle manure with visible/near-infrared spectroscopy," Trans. ASABE **53**: 421–428 (2010).
29. W. Ye, J. C. Lorimor, C. Hurburgh, H. Zhang, and J. Hattery, "Application of near-infrared reflectance spectroscopy for determination of nutrient contents in liquid and solid manures," Trans. ASAE **48**, 1911–1918 (2005).
30. Z. Yang, L. Han, and X. Fan, "Rapidly estimating nutrient contents of fattening pig manure from floor scrapings by near infrared reflectance spectroscopy," J. Near Infrared Spectrosc. **14**, 261–268 (2006).
31. K. Suehara, Y. Nakano, and T. Yano, "Simultaneous measurement of the carbon and nitrogen content of compost using near-infrared spectroscopy," J. Near Infrared Spectrosc. **9**, 35–41 (2001).
32. D. F. Malley, C. McClure, P. D. Martin, K. Buckley, and W. P. McCaughe, "Compositional analysis of cattle manure during

composting using a field-portable near-infrared spectrometer," Commun. Soil Sci. Plant Anal. **36**, 455–475 (2005).

33. E. Ben-Dor, Y. Inbar, and Y. Chen, "The reflectance spectra of organic matter in the visible near infrared and short wave infrared region (400–2500 nm) during a control decomposition process," Remote Sens. Environ. **61**, 1–15 (1997).
34. H. S. S. Sharma, M. Kilpatrick, G. Lyons, S. Sturgeon, J. Archer, S. Moore, L. Cheung, and K. Finegan, "Visible and near-infrared calibrations for quality assessment of fresh phase I and II mushroom (Agaricus bisporus) compost," Appl. Spectrosc. **59**, 1399–1405 (2005).
35. A. L. McWhirt, D. C. Weindorf, S. Chakraborty, and B. Li, "Visible near infrared diffuse reflectance spectroscopy (VisNIR DRS) for rapid measurement of organic matter in compost," Waste Manag. Res. **30**, 1049–1058 (2012).
36. Y. Zhu, D. C. Weindorf, S. Chakraborty, B. Haggard, S. Johnson, and N. Bakr, "Characterizing surface soil water with field portable diffuse reflectance spectroscopy," J. Hydrol. **391**, 133–140 (2010).
37. T. H. Demetriades-Shah, M. D. Steven, and J. A. Clark, "High-resolution derivative spectra in remote sensing," Remote Sens. Environ. **33**, 55–64 (1990).
38. R Development Core Team, "R: a language and environment for statistical computing," http://www.cran.r-project.org (2008).
39. D. J. Brown, K. D. Shepherd, M. G. Walsh, M. D. Mays, and T. G. Reinsch, "Global soil characterization with VNIR diffuse reflectance spectroscopy," Geoderma **132**, 273–290 (2006).
40. G. E. P. Box and D. R. Cox, "An analysis of transformations," J. R. Stat. Soc. Ser. B **26**, 211–252 (1964).
41. H. Steinhaus, "Sur la division des corp materiels en parties," Bull. Acad. Pol. Sci. **4**, 801–804 (1956) (in French).
42. J. H. Ward, "Hierarchical grouping to optimize an objective function," J. Am. Stat. Assoc. **48**, 236–244.(1963).
43. T. H. Waiser, C. L. S. Morgan, D. J. Brown, and C. T. Hallmark, "In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy," Soil Sci. Soc. Am. J. **71**, 389–396 (2007).
44. D. M. Haaland and E. V. Thomas, "Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information," Anal. Chem. **60**, 1193–1202 (1988).
45. P. H. C. Eilers and B. D. Marx, "Generalized linear additive smooth structures," J. Comput. Graph. Statist. **11**, 758–783 (2002).
46. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Technical report (Department of Statistics, Stanford University, 1999).
47. R. A. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," Geoderma **158**, 46–54 (2010).
48. G. A. Shonk, L. D. Gaultney, D. G. Schulze, and G. E. Van Scoyoc, "Spectroscopic sensing of soil organic matter content," Trans. ASAE **34**, 1978–1984 (1991).
49. P. C. Williams, "Commercial near-infrared reflectance analyzers," in *Near-infrared Technology in the Agricultural and Food Industries*, P. C. Williams and K. H. Norris, eds. (American Association of Cereal Chemists, 1987), pp. 107–136.
50. B. Minasny and A. B. McBratney, "Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy," Chemom. Intell. Lab. Syst. **94**, 72–79 (2008).
51. L. Kooistra, R. Wehrens, R. S. E. W. Leuven, and L. M. C. Buydens, "Possibilities of VNIR spectroscopy for the assessment of soil contamination in river floodplains," Anal. Chim. Acta **446**, 97–105 (2001).
52. S. Shibusawa, S. W. I. Anom, S. Sato, A. Sasao, and S. Hirako, "Soil mapping using the real-time soil spectrophotometer," in *ECPA 2001, Third European Conference on Precision Agriculture*, G. Grenier and S. Blackmore, eds. (Agro Montpellier, 2001), Vol. **1**, pp. 497– 508.
53. American Association for State Highway and Transportation Officials, "Standard specifications for compost for erosion/sediment control," http://compostingcouncil.org/admin/wp-content/plugins/wp-pdfupload/pdf/32/AASHTO-Specifications.pdf.
54. J. L. Bishop, M. D. Lane, M. D. Dyar, and A. J. Brown, "Reflectance and emission spectroscopy study of four groups of phyllosilicates: smectites, kaolinite–serpentines, chlorites and micas," Clay Miner. **43**, 35–54 (2008).
55. G. R. Hunt, "Spectral signatures of particulate minerals in visible and near IR," Geophysics **42**, 501–513 (1977).
56. R. N. Clark, "Spectroscopy of rocks and minerals, and principles of spectroscopy," in *Remote Sensing for the Earth Sciences: Manual of Remote Sensing*, N. Rencz, ed. (Wiley, 1999), pp. 3–52.
57. J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," J. Anim. Ecol. **77**, 802–813 (2008).
58. F. J. Calderon, J. B. Reeves, H. P. Collins, and E. A. Paul, "Chemical differences in soil organic matter fractions determined by diffuse-reflectance mid-infrared spectroscopy," Soil Sci. Soc. Am. J. **75**, 568–579 (2011).
59. Y. Ge, C. L. S. Morgan, J. A. Thomasson, and T. Waiser, "A new perspective to near-infrared reflectance spectroscopy: a wavelet approach," Trans. ASABE **50**, 303–311 (2007).