Hindawi Publishing Corporation
Applied and Environmental Soil Science
Volume 2012, Article ID 294121, 14 pages
doi:10.1155/2012/294121

Research Article

Effects of Subsetting by Carbon Content, Soil Order, and Spectral Classification on Prediction of Soil Total Carbon with Diffuse Reflectance Spectroscopy

Meryl L. McDowell,[1] Gregory L. Bruland,[1, 2] Jonathan L. Deenik,[3] and Sabine Grunwald[4]

[1] *Natural Resources and Environmental Management Department, University of Hawaiʻi Mānoa, 1910 East-West Road, Sherman 101, Honolulu, HI 96822, USA*
[2] *Biology & Natural Resources Department, Principia College, 1 Maybeck Place, Elsah, IL 62028, USA*
[3] *Tropical Plant and Soil Sciences Department, University of Hawaiʻi Mānoa, 3190 Maile Way, Honolulu, HI 96822, USA*
[4] *Soil and Water Science Department, University of Florida, 2169 McCarty Hall, P.O. Box 110290, Gainesville, FL 32611-0290, USA*

Correspondence should be addressed to Meryl L. McDowell, mcdowell@hawaii.edu

Received 15 March 2012; Revised 20 August 2012; Accepted 14 October 2012

Academic Editor: Sabine Chabrillat

Subsetting of samples is a promising avenue of research for the continued improvement of prediction models for soil properties with diffuse reflectance spectroscopy. This study examined the effects of subsetting by soil total carbon (C_t) content, soil order, and spectral classification with k-means cluster analysis on visible/near-infrared and mid-infrared partial least squares models for C_t prediction. Our sample set was composed of various Hawaiian soils from primarily agricultural lands with C_t contents from <1% to 56%. Slight improvements in the coefficient of determination (R^2) and other standard model quality parameters were observed in the models for the subset of the high activity clay soil orders compared to the models of the full sample set. The other subset models explored did not exhibit improvement across all parameters. Models created from subsets consisting of only low C_t samples (e.g., $C_t < 10\%$) showed improvement in the root mean squared error (RMSE) and percent error of prediction for low C_t soil samples. These results provide a basis for future study of practical subsetting strategies for soil C_t prediction.

1. Introduction

Diffuse reflectance spectroscopy (DRS) and chemometric analysis have become popular subjects of research for their potential to predict soil carbon and other soil properties. This methodology could be beneficial for monitoring soil quality and temporal variation, as well as helping to facilitate digital soil mapping efforts. Both visible/near-infrared (VNIR) and mid-infrared (MIR) spectra show promise for the prediction of soil total carbon (C_t) and organic carbon, as well as organic matter, total N, total P, sand, silt, and clay fractions, cation exchange capacity, and pH (e.g., [1–8]). Particular attention has been given to soil carbon, which is an important indicator of soil fertility and biological activity and is crucial to carbon sequestration endeavors [9–12].

Partial least squares regression (PLSR) appears to be the most widely used chemometric method for developing prediction models from soil diffuse reflectance spectra. A sample set is commonly divided into two groups with the larger used for calibration and the smaller for validation to approximate true independent model validation, but no clear or consistent guidelines have been adopted for this process. Model results are known to vary with different groupings of samples for the calibration and validation sets. To address this issue, some studies have created multiple models, each with different random divisions of the sample set into calibration and validation sets, to reflect the range of possible results [13, 14].

Highly accurate prediction models are required for DRS to be an effective method for soil carbon determination in practical applications. Many statistically robust models have been developed (e.g., [5–8, 15]), but a single procedure is not necessarily the best for producing high quality models from different soils in different locations. Even models that have excellent correlation between soil spectra and properties could be improved. For instance, the robust PLSR models of McDowell et al. [8] have relatively large errors in C_t prediction at very low C_t values, which decreases the utility of the models in situations where low C_t soils or small changes in C_t are examined. Additional methods are being explored to produce the most robust and accurate DRS prediction models possible for different local and global soil datasets. One promising idea is to split the sample set into groups based on similar characteristics and to develop individual prediction models for each of these subsets. In studies of soils from Poland, Brazil, and Florida (USA), previous researchers have investigated subsetting by characteristics such as carbon content, soil order, soil texture, and spectral similarity with varied success for their particular sample sets [16–18].

The current work aimed to improve the prediction of C_t with VNIR and MIR DRS by creating attribute-specific chemometric models. Specifically, we investigated if predictions from a chemometric model built only from a subset of samples that are similar with respect to a particular characteristic (i.e., C_t) will provide better predictions than a comprehensive model built from a set of all possible samples. The study investigated the following three subsetting strategies: (1) soil C_t value; (2) soil order; (3) spectral classification with k-means cluster analysis. Each of the various subset models was compared against the original full sample set model to assess the magnitude of changes in the predictions. This study was built upon the research reported in McDowell et al. [8]. In that work the authors demonstrated the ability of DRS to predict C_t in Hawaiian soils. The success of different wavelength ranges (i.e., VNIR versus MIR) and chemometric methods was investigated, as well. Because these ideas have been previously explored in McDowell et al. [8], they will not be discussed further here.

2. Materials and Methods

2.1. Sample Collection and Preparation. The sample set for this study is composed of 307 soil samples collected across the five main Hawaiian Islands of Kauai, Oahu, Molokai, Maui, and Hawaii, illustrated in Figure 1. Two hundred and sixteen of these samples were collected from 1981 to 2007 and stored in the archive at the Natural Resources Conservation Service (NRCS) National Soil Survey Center in Lincoln, Nebraska, and the remaining 91 samples were newly collected in 2010. Within this full set of samples, 10 soil orders and more than 100 soil series are represented. Samples were predominantly from a variety of agricultural soils, hosting over 25 different crop types. The majority of samples are of surface soils (∼77%), and the remainder are of corresponding subsurface soil horizons from 17 of the collection sites. The soil samples were dried and sieved to retain the less than 2 mm fraction for VNIR DRS analysis. A portion of each sample was also ball-milled to less than 250 μm for MIR DRS analysis.

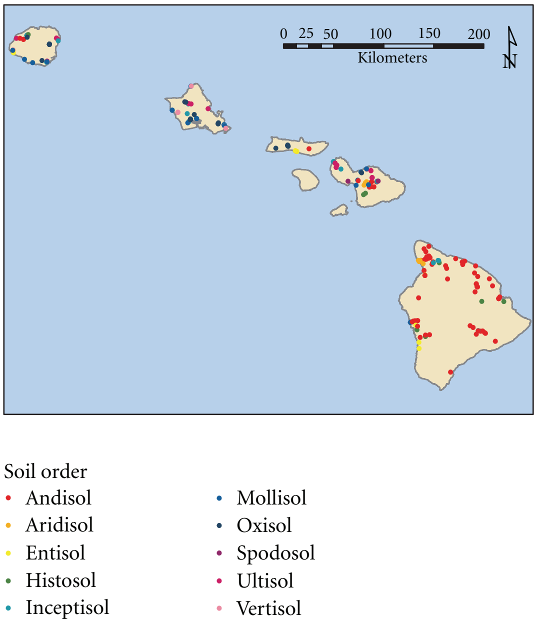


Figure 1: Distribution of soils sample collection sites throughout the Hawaiian Islands with symbol color indicating soil order.

2.2. Traditional Total Carbon Analysis. Dry combustion was used to measure the C_t of ball-milled soil samples. Several of the samples obtained from the NRCS archive were previously measured for C_t by dry combustion before storage. All remaining samples were analyzed at the Agricultural Diagnostic Services Center (ADSC) at the University of Hawaii Mānoa with an LECO CN2000 combustion gas analyzer [19]. A small portion of the previously measured NRCS archive samples were reanalyzed at ADSC to provide a cross-check of the values obtained from different laboratories. The C_t values of the full sample set range from <1% to 56% with a distribution weighted toward the lower C_t end.

2.3. Visible/Near-Infrared Diffuse Reflectance Spectroscopy. Visible/near-infrared diffuse reflectance spectra were collected from the 2 mm sieved soil samples with an Agrispec spectrometer and muglight light source (Analytical Spectral Devices, Inc., Boulder, CO, USA). The Agrispec has three detectors with a combined spectral range of 350 to 2500 nm, sampling interval of 1 nm, and spectral resolution from 3 nm (at 700 nm) to 10 nm (at 1400 nm). Each soil sample was measured three times, with the sample cup rotated 20° between each measurement. The three spectra were averaged to produce the final spectrum for each sample. A Spectralon (Labsphere, North Sutton, NH, USA) white reference was measured as a reference spectrum to begin each session and again every 30 minutes or less thereafter. A slight offset in reflectance between the range covered by the first and second detectors was observed in many spectra, and, therefore, we removed the narrow region of 990–1010 nm from the final spectra for analysis. The VNIR spectra of these soils commonly exhibit features associated with OH^-, H_2O, iron oxides, phyllosilicates, and organic molecules. For regression analysis the spectra were transformed using the pretreatment identified as most effective for this data set in McDowell et al. [8]. For the VNIR spectra, this optimal preprocessing transformation was mean normalization.

2.4. Mid-Infrared Diffuse Reflectance Spectroscopy. Mid-infrared diffuse reflectance spectra were collected from the ball-milled samples in neat form with a Scimitar 2000 FTIR spectrometer (Varian, Inc., now Agilent Technologies, Santa Clara, CA, USA) and diffuse reflectance infrared Fourier transform (DRIFT) accessory. The spectral range is 400 to 6000 cm^{-1}, with a sampling interval of 2 cm^{-1} and spectral resolution of 4 cm^{-1} (note: the range of our MIR spectra overlaps slightly with the range of our VNIR spectra.) Spectra were corrected for background atmospheric and instrument effects by the subtraction of the spectrum of KBr powder measured between every seven samples, but features in two narrow regions persisted. Therefore, we excluded the regions of 1350–1419 cm^{-1} and 2281–2449 cm^{-1} from the analysis. Features in the MIR spectra of these soils are attributable to OH^{-}, organic molecules, and a variety of silicate minerals. Based on the findings of McDowell et al. [8], before regression analysis the Savitzky-Golay 1st derivative transformation was applied to the MIR spectra as this was determined to be the most effective pretreatment for this data set.

2.5. Regression Analysis. Partial least squares regression (PLSR) was employed to develop the chemometric models for C_t prediction. Models were generated using the Unscrambler X Software package (CAMO Software Inc., Woodbridge, NJ, USA). The spectral range included in the analysis was decreased slightly by removing any high noise portions at the limit of the range; therefore, the VNIR spectra were restricted to the range of 425–2450 nm, and the MIR spectra were restricted to 489–5300 cm^{-1}. All spectra were mean centered for PLSR analysis. The optimal number of factors for regression was chosen individually for each model based on maximizing the explained variance but minimizing the possibility of over fitting. We considered several parameters when assessing the quality of models, including the coefficient of determination (R^2), root mean squared error (RMSE), residual prediction deviation (RPD) [20], and the ratio of performance to interquartile distance (RPIQ) [21]. We defined the RPD as the ratio of the standard deviation of the validation set to the standard error of prediction (RPD = SD/SEP) and the RPIQ as the ratio of the interquartile distance of the validation set to the standard error of prediction (RPIQ = IQ/SEP), where the interquartile distance is the difference between the third and first quartiles (IQ = Q3 − Q1). With respect to these general model quality parameters, the best model would have the highest R^2, RPD, and RPIQ, and the lowest RMSE. We also examined the success of the predictions for individual samples using the percent error, calculated as the absolute difference between the measured (i.e., by combustion) and predicted (i.e., by DRS) C_t values, divided by the measured value, and multiplied by 100.

2.6. Sample Subsetting. The motivation behind our selected subsetting strategies was to improve C_t prediction while still retaining the simplicity that makes DRS attractive. We focused on subsetting criteria that did not require additional highly detailed soil characterization, instead relying on general soil data and information within Soil Taxonomy.

2.6.1. C_t Content Subsets. A simple grouping of soils into low and high C_t was used for subsetting by C_t value. Preliminary work tested a variety of low C_t/high C_t divisions (e.g., 2, 4, 6, 8, and 10% C_t) iteratively. The initial results showed that a cutoff of 10% C_t was most promising and therefore was used for the final analysis. Additionally, a division at 10% allows for fairly easy assignment of unknown soils into low or high C_t groupings from C_t estimates based on general or readily available soil information.

To approximate independent validation, samples were randomly split into a group of 70% for model calibration and 30% for model validation. This random selection was repeated to produce 10 iterations of calibration/validation pairs from the full sample set. After this split, the samples from each iteration were divided into low C_t (<10%) and high C_t (>10%) subsets. Separate VNIR and MIR regression models were then developed from the low C_t and high C_t portions of each of the 10 iterations. For comparison, VNIR and MIR regression models from the full sample set using these same 10 calibration and validation divisions, but no separation by C_t value, also were created.

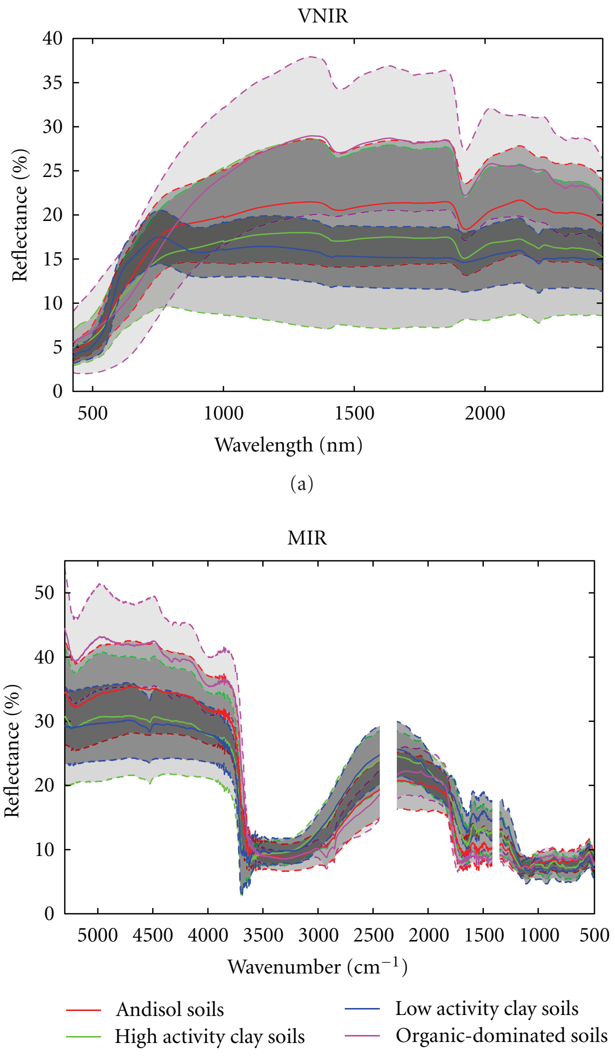
2.6.2. Soil Order Subsets. Four broad soil groups were created based on general similarity of soil order and number of samples available of that type. The allophane-dominated volcanic Andisol soils comprised one group ($n = 96$), the Aridisol, Entisol, Inceptisol, Mollisol, and Vertisol soils were combined to make a second group (high activity clay soils; $n = 101$), Oxisol and Ultisol soils made a third group (low activity clay soils; $n = 75$), and Histosol and Spodosol soils comprised the fourth group (organic-dominated soils; $n = 26$). These soil groupings are based upon information con tained in Soil Taxonomy allowing for the development of soil groups according to clay mineralogy and soil organic matter. Table 1 provides information on additional soil properties for each soil subset where available. The average spectra for each of these soil groups are shown in Figure 2. Nine soil samples from the NRCS archive had no recorded taxonomic classification and therefore were not included in these subsets.

The full sample set was randomly divided 10 times into a group of 70% of samples to be used for the calibration of the regression models and 30% of the samples to be used for validation. After this division, the samples of each of the ten iterations were grouped according to soil order as described above. Separate VNIR and MIR regression models were then developed for each soil group subset within each of the ten calibration/validation iterations. Because the number of low activity clay and organic-dominated soil samples was small (e.g., ≤80), full cross validation (i.e., leave-one-out cross validation) was used with the regression models for these

Table 1: Soil properties of selected samples for each soil grouping used in subsetting by soil order. Values listed in the table are the minimum and maximum for that specific subset with the mean in parentheses. Data is provided for samples from the Natural Resources Conservation Service (NRCS) archive where it is available. The compositional information (i.e., pH, texture, Al, Ca, and Fe) for the samples newly collected in 2010 has yet to be determined.

	Total carbon wt%	Organic carbon wt%	Clay wt%	Silt wt%	Sand wt%	pH	Total Al wt%	Total Ca wt%	Total Fe wt%
Andisol soils	0.24–51 (13.39)	0.39–55.59 (12.53)	0.3–59.8 (17.26)	4.7–81.3 (40.62)	2.4–94.9 (42.83)	3.7–8 (5.66)	1.58–13.89 (8.54)	0.025–4.80 (0.64)	7.33–22.63 (15.49)
High activity clay soils	0.21–53.63 (14.51)	0.3–14.65 (3.94)	0.2–66.7 (25.72)	10.8–93.2 (44.08)	0.4–88.6 (30.21)	3.3–8.3 (5.89)	10.95[a]	0.52[a]	10.13[a]
Low activity clay soils	0.15–10 (1.65)	0.2–3.58 (1.11)	7.6–88.7 (47.52)	10.4–69.5 (34.86)	0.75–69.8 (17.61)	4.5–7.3 (5.92)	7.66–9.61 (8.28)	0.049–0.16 (0.096)	13.43–27.03 (23.23)
Organic-dominated soils	5–55.29 (36.19)	2.62–54.98 (20.26)	4.4–67.6 (31.68)	11.5–45.7 (30.45)	1.3–84.1 (37.86)	3.3–5.8 (4.29)	Not available		

[a] Only one data point available.



(b)

Figure 2: Average (a) visible/near-infrared (VNIR) and (b) mid-infrared (MIR) diffuse reflectance spectra of soil groups used in subsetting by soil order. Dashed lines represent one standard deviation from the average.

two groups rather than committing 30% of those samples to validation as with the other subsets. Additional models were created from the 10 calibration/validation divisions of the full sample set with no separation of soil order for the comparison of results without subsetting. A full cross validation model of the full sample set was developed to be compared with the low activity clay and organic-dominated soil subsets' full cross validation models.

2.6.3. Spectral Classification Subsets. Our rationale behind grouping soil samples by spectral character is based on the assumption that this approach removes major spectral variation from consideration so that small-scale variation is used to produce a more refined C_t prediction model. Also, the division of soil samples into subsets created solely from spectral classification has the advantage of requiring no additional information about the soil.

The spectral classification subsets were created by k-means cluster analysis with Unscrambler X. Spectra were assigned to three cluster subsets based on the minimum Euclidean distance to cluster centers. Separate analyses were conducted for the VNIR and MIR spectra, resulting in different combinations of samples in their cluster subsets. The spectral range used for these cluster analyses was limited to the regions most relevant to carbon prediction as previously determined by the PLSR variable significance analysis by McDowell et al. [8]. Specifically, the ranges used were 600–750, 898–990, 1910–1938, 2070–2150, and 2288–2316 nm for the VNIR spectra and 1500–1870, 3650–3690, 4235–4260, 4305–4330, 4410–4455, and 5280–5245 cm^{-1} for the MIR spectra. Each cluster subset was randomly divided into a group of 70% for model calibration and a group of the remaining 30% for model validation, unless the number of samples in the cluster was small (e.g., ≤80), in which case samples were not divided and full cross validation was performed. The random division into calibration and validation groups was repeated nine more times to give 10 calibration/validation pairs for each of the VNIR and MIR cluster subsets. Separate C_t prediction models were created for each of the different cluster subsets. For comparison, we also developed 10 VNIR and 10 MIR models from the full sample set. The calibration and validation groups for these models were created by combining the respective calibration or validation groups from the three different cluster subset models. VNIR and MIR full cross validation models using the full sample set were also produced to compare with full cross validation models from small cluster subsets.

3. Results and Discussion

3.1. Modeling of C_t Content Subsets. The VNIR models subset by C_t content produced the results summarized in Table 2 and plotted in Figure 3(a). The range of results from the 10 random divisions of the samples into 70% calibration and 30% validation groups is given along with their mean value. The R^2, RPD, and RPIQ values for the low C_t subset were not as good as those produced using the full sample set, though the RMSE values were lower for the low C_t subset. The results for the high C_t models approached, but were not quite as good, as the results from the full sample set.

Results from the MIR C_t subset models are shown in Figure 3(b) and Table 3. The models produced by the low C_t subset were generally of lesser quality than those of the full sample set, with the exception of better RMSE values, a trend similar to the VNIR models. The high C_t models were comparable overall to the high quality models produced by using the full sample set.

From these results, it appears that a separate high C_t prediction model is not an improvement over a model utilizing the full C_t range of available samples for either the VNIR or MIR spectra from this data set. This statement may be true for a separate low C_t prediction model as well, but the benefit of a lower RMSE should also be considered.

Table 2: Detailed partial least squares regression model results for soil total carbon (C_t) prediction from the subsets of visible/near-infrared diffuse reflectance spectra based on C_t content. The range of values reflects the results of 10 random iterations of the models, and the number in parentheses is the mean. Detailed results are also given for full sample set models with no subsetting for comparison.

	Calibration			Validation				
	n^a	$R^{2,b}$	RMSE (%)^c	n	R^2	RMSE (%)	RPD^d	RPIQ^e
C_t < 10%	133–147	0.43–0.80 (0.64)	1.08–1.78 (1.46)	56–70	0.47–0.76 (0.61)	1.27–1.97 (1.59)	1.37–2.03 (1.63)	1.77–2.88 (2.12)
C_t > 10%	68–82	0.77–0.93 (0.86)	3.86–7.00 (5.33)	22–36	0.77–0.91 (0.84)	3.96–7.65 (5.87)	2.05–3.21 (2.55)	2.38–5.16 (4.02)
Full sample set	215	0.81–0.96 (0.91)	2.88–5.87 (4.06)	92	0.81–0.95 (0.91)	2.82–7.18 (4.24)	2.27–4.47 (3.46)	2.08–4.35 (3.19)

[a] Number of samples.
[b] Coefficient of determination.
[c] Root mean squared error.
[d] Residual prediction deviation.
[e] Ratio of performance to interquartile distance.

Table 3: Detailed partial least squares regression model results for soil total carbon (C_t) prediction from the subsets of mid-infrared diffuse reflectance spectra based on C_t content. The range of values reflects the results of 10 random iterations of the models, and the number in parentheses is the mean. Detailed results are also given for full sample set models with no subsetting for comparison.

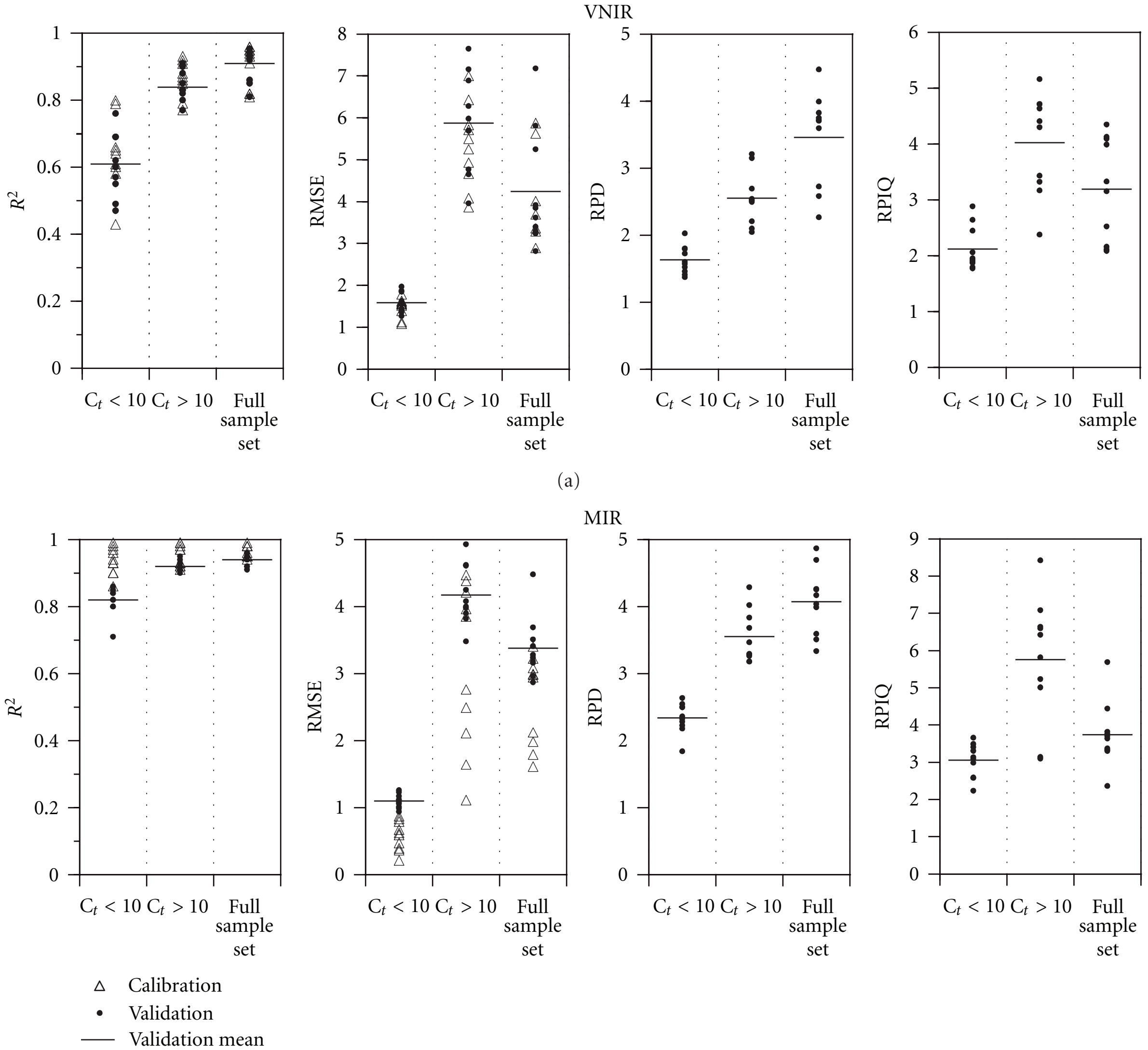
	Calibration			Validation				
	n^a	$R^{2,b}$	RMSE (%)^c	n	R^2	RMSE (%)	RPD^d	RPIQ^e
C_t < 10%	133–147	0.86–0.99 (0.94)	0.21–0.87 (0.58)	56–70	0.71–0.86 (0.82)	0.94–1.26 (1.10)	1.84–2.64 (2.34)	2.24–3.66 (3.05)
C_t > 10%	68–82	0.91–0.99 (0.95)	1.11–4.47 (3.10)	22–36	0.90–0.95 (0.92)	3.48–4.93 (4.17)	3.18–4.29 (3.55)	3.10–8.42 (5.75)
Full sample set	215	0.94–0.99 (0.96)	1.61–3.40 (2.61)	92	0.91–0.96 (0.94)	2.87–4.48 (3.38)	3.33–4.87 (4.07)	2.36–5.69 (3.74)

[a] Number of samples.
[b] Coefficient of determination.
[c] Root mean squared error.
[d] Residual prediction deviation.
[e] Ratio of performance to interquartile distance.

Results varied for previous studies examining the behavior of separate models based on carbon content. Madari et al. [16] found that limiting the C_t in their NIR and MIR calibration models to 0.4–99.10 g kg^{-1} and 0.4–39.90 g kg^{-1} decreased the not only R^2, but also the root mean squared deviation (RMSD) compared to the original NIR and MIR models (0.4–555 g kg^{-1} C_t); this behavior is similar to that observed in the low C_t models presented here. The study by Vasques et al. [18] developed separate VNIR organic carbon prediction models for their mineral and organic soil samples, which roughly correspond to division by carbon content in this case (mineral soils, 0.01–14.70% carbon; organic soils, 13.52–57.54% carbon). Compared to the original combined model, the R^2 improved for both of the subset models, but the RMSE decreased for the lower carbon mineral group and increased for the higher carbon organic group. The increase in R^2 values for the subset models differs from what is seen in our work and that of Madari et al. [16] and is an example of soils with different characteristics responding differently to the same treatment.

3.2. Modeling of Soil Order Subsets. The results of the VNIR models from the soil order subsets are given in Table 4 and Figure 4(a). The models from the Andisol subset did not perform as well as the models using the full sample set. The R^2, RMSE, and RPD values for the high activity clay subset were similar to those of the full sample set models, but the RPIQ values were generally slightly lower. The low activity clay and organic-dominated subsets were not validated with an independent validation set due to small sample numbers, and therefore their results may be overly optimistic. Compared to a full cross validation of a model created from the full sample set, the low activity clay subset model did not perform as well, except when considering the RMSE parameter, whereas the organic-dominated subset model is broadly similar.

Table 5 and Figure 4(b) show the results of the MIR soil order subset models. The models produced by the Andisol subset had no improvement on the models produced by the full sample set. Results for the high activity clay subset models were as good as or better than the full sample set model results, with the exception of lower RPIQ values. The



(a)

(b)

Figure 3: Visual assessment of partial least squares regression model results for soil total carbon (C_t) prediction from subsets of (a) visible/near-infrared (VNIR) and (b) mid-infrared (MIR) diffuse reflectance spectra based on C_t content. The parameters given are the coefficient of determination (R^2), root mean squared error (RMSE, %), residual prediction deviation (RPD), and the ratio of performance to interquartile distance (RPIQ). The range of values reflects the results of 10 random iterations of the models. Results are also shown for full sample set models with no subsetting for comparison.

overall performance of the low activity clay and organic-dominated subset models using full cross validation was not quite as good as the full cross validation model from the full sample set.

These results suggest that a separate prediction model for the high activity clay soil orders may have a slight advantage compared to a model with all available soil orders for both the VNIR and MIR spectra of this data set. Separate prediction models for the other soil order subsets do not appear to be as promising.

A study by Madari et al. [16] also investigated the benefits of subsetting their samples according to soil order. The authors produced separate models for the Histosols and Spodosols, the Ferralsols (classification according to the World Reference Base [22], approximately equivalent to most of the Oxisol soil order), and the Acrisols (classification according to the World Reference Base [22], consisting of many Ultisol suborders and some Oxisols). The results of these models varied. The Ferralsol and the Acrisol NIR and MIR models had lower R^2 than the original model and also lower RMSD; these two subsets included relatively low C_t (2–85.10 g kg^{-1} and 1.70–91.60 g kg^{-1}, resp.) compared to the full sample set (0.40–555 g kg^{-1}), so this lower R^2 and lower RMSD are a similar behavior to the low C_t subset models in the current study. The Histosol and Spodosol subset NIR and MIR models in Madari et al. [16] resulted in slightly higher R^2 values and much higher RMSD values. Our Histosol and Spodosol (i.e., organic-dominated soils) subset models did

Table 4: Detailed partial least squares regression model results for soil total carbon (C_t) prediction from the subsets of visible/near-infrared diffuse reflectance spectra based on soil order. The range of values reflects the results of 10 random iterations of the models, and the number in parentheses is the mean. Detailed results are also given for full sample set models with no subsetting for comparison. For models with full cross validation (i.e., leave-one-out cross validation), the same samples used to calibrate the model were used to validate the model.

	Calibration			Validation				
	n[a]	R^2[b]	RMSE (%)[c]	n	R^2	RMSE (%)	RPD[d]	RPIQ[e]
Andisol soils	64–71	0.62–0.86 (0.72)	2.71–7.75 (4.64)	25–32	0.37–0.93 (0.69)	3.38–7.48 (4.85)	1.01–3.80 (2.02)	1.29–3.38 (2.28)
High activity clay soils	67–72	0.86–0.98 (0.93)	2.38–5.17 (3.73)	29–34	0.74–0.98 (0.90)	2.19–6.31 (4.02)	1.89–7.74 (4.12)	0.71–3.03 (1.68)
Low activity clay soils	75	0.82	0.72	Full cross validation	0.74	0.90	1.93	1.82
Organic-dominated soils	26	0.96	3.35	Full cross validation	0.92	5.16	3.30	6.26
Full sample set	215	0.82–0.96 (0.92)	2.89–5.96 (3.89)	92	0.79–0.95 (0.91)	2.96–6.03 (4.02)	2.25–4.43 (3.58)	2.07–4.53 (3.42)
Full sample set	307	0.95	3.09	Full cross validation	0.94	3.39	4.09	3.80

[a] Number of samples.
[b] Coefficient of determination.
[c] Root mean squared error.
[d] Residual prediction deviation.
[e] Ratio of performance to interquartile distance.

Table 5: Detailed partial least squares regression model results for soil total carbon (C_t) prediction from the subsets of mid-infrared diffuse reflectance spectra based on soil order. The range of values reflects the results of 10 random iterations of the models, and the number in parentheses is the mean. Detailed results are also given for full sample set models with no subsetting for comparison. For models with full cross validation (i.e., leave-one-out cross validation), the same samples used to calibrate the model were used to validate the model.

	Calibration			Validation				
	n[a]	R^2[b]	RMSE (%)[c]	n	R^2	RMSE (%)	RPD[d]	RPIQ[e]
Andisol soils	64–71	0.84–0.96 (0.91)	1.92–3.02 (2.49)	25–32	0.41–0.92 (0.79)	2.99–6.94 (4.03)	1.12–3.60 (2.33)	1.87–4.09 (2.66)
High activity clay soils	67–72	0.96–0.99 (0.98)	0.96–2.71 (1.74)	29–34	0.95–0.99 (0.96)	1.70–3.60 (2.65)	4.34–9.81 (5.57)	0.92–4.38 (2.44)
Low activity clay soils	75	0.98	0.24	Full cross validation	0.79	0.80	2.10	2.01
Organic-dominated soils	26	0.97	2.9	Full cross validation	0.86	6.7	2.52	4.78
Full sample set	215	0.94–0.98 (0.96)	1.94–3.50 (2.78)	92	0.91–0.96 (0.94)	2.74–3.91 (3.39)	3.38–5.07 (4.07)	3.22–5.27 (3.89)
Full sample set	307	0.95	3.12	Full cross validation	0.94	3.52	3.94	3.68

[a] Number of samples.
[b] Coefficient of determination.
[c] Root mean squared error.
[d] Residual prediction deviation.
[e] Ratio of performance to interquartile distance.

not have significantly increased R^2 values, but the validation RMSE values were greater than the full sample set models' values.

Vasques et al. [18] developed separate organic carbon prediction VNIR models for each of the seven soil orders in their sample set consisting of soils from Florida, southeastern USA Compared to the original model containing all of these mineral soil samples, six of the seven soil order subset models resulted in improved R^2 values (Alfisols, Entisols, Inceptisols, Mollisols, Spodosols, and Ultisols). The RMSE values were also similar or better for these subsets. The Histosol subset model was the only one that did not improve in R^2 or RMSE. These results are somewhat different from those in this study, where only the high activity clay soils (i.e., Aridisols, Entisols, Inceptisols, Mollisols, and Vertisols) are suggested to provide an overall improvement on models including all available samples.

3.3. Modeling of Spectral Classification Subsets. The k-means cluster analysis of the VNIR spectra resulted in an unequal distribution of samples between the three clusters. The

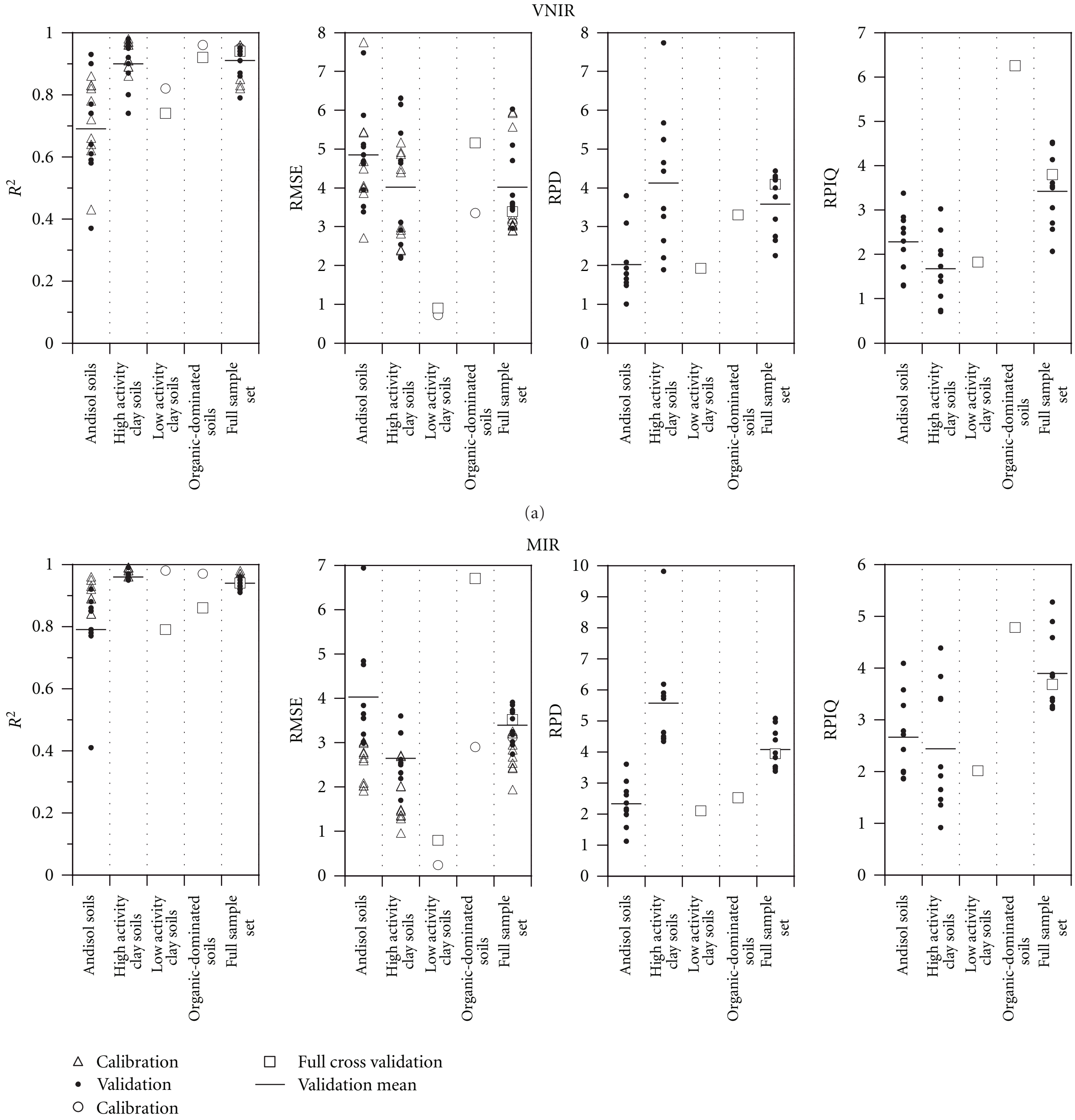
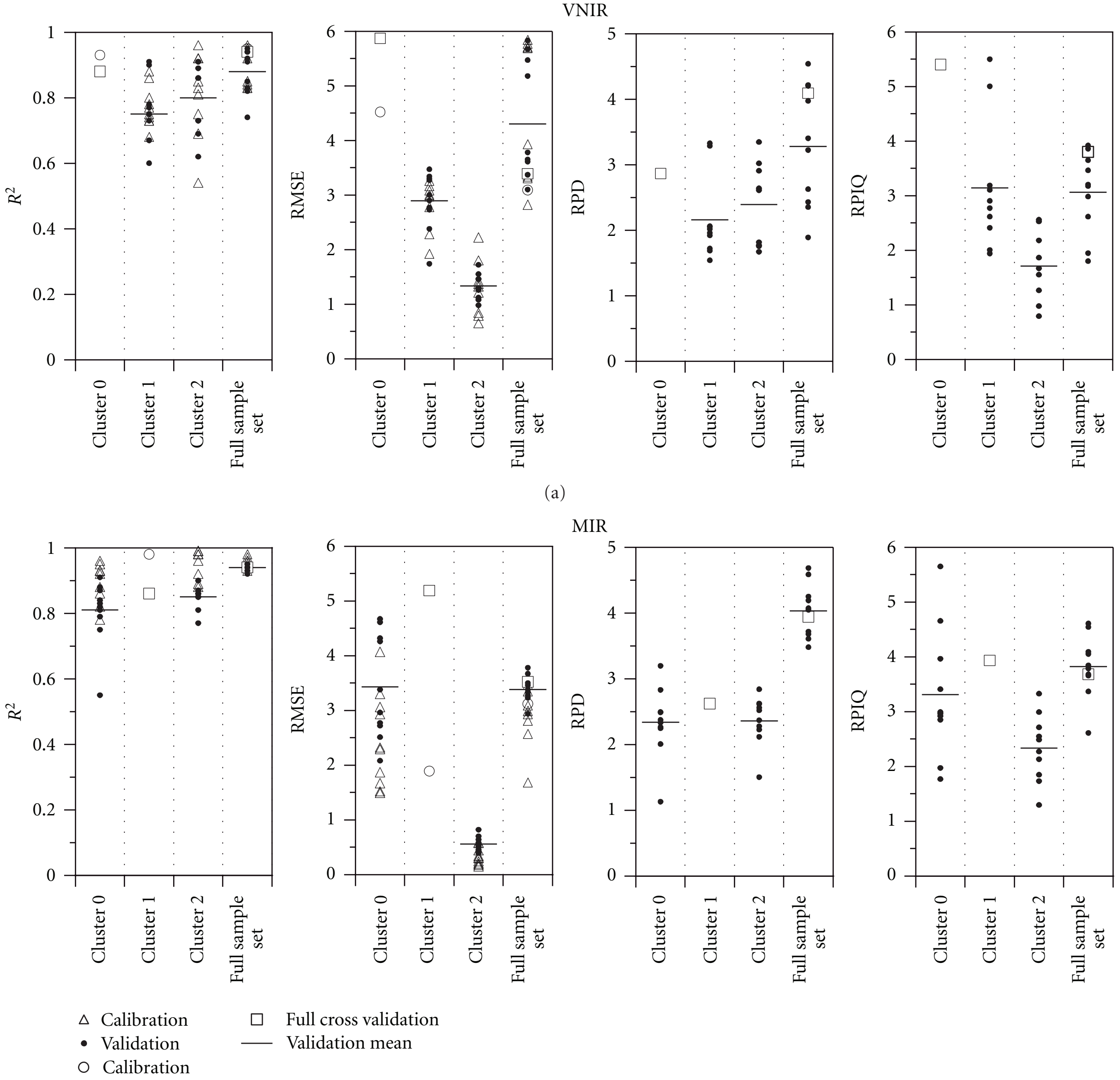


Figure 4: Visual assessment of partial least squares regression model results for soil total carbon (C_t) prediction from subsets of (a) visible/near-infrared (VNIR) and (b) mid-infrared (MIR) diffuse reflectance spectra based on soil order. The parameters given are the coefficient of determination (R^2), root mean squared error (RMSE, %), residual prediction deviation (RPD), and the ratio of performance to interquartile distance (RPIQ). The range of values reflects the results of 10 random iterations of the models. Results are also shown for full sample set models with no subsetting for comparison.

Cluster 0 subset consisted of only 78 samples (~3–56% C_t) and therefore all 78 samples were used in its model calibration and full cross validation. The Cluster 1 and Cluster 2 subsets contained 124 samples (~0–23% C_t) and 105 samples (~0–14% C_t), respectively, allowing for the independent validation of the models as initially planned. The results of the 10 VNIR C_t prediction models from each of the clusters are given in Table 6 and Figure 5(a). A comparison of the Cluster 0 subset model with a full cross validation model of the full sample set showed that the subset model was not quite as robust, though it did produce a higher RPIQ value. The Cluster 1 and Cluster 2 subset models' results generally



(a)

(b)

Figure 5: Visual assessment of partial least squares regression model results for soil total carbon (C_t) prediction from the subsets of (a) visible/near-infrared (VNIR) and (b) mid-infrared (MIR) diffuse reflectance spectra based on spectral classification with *k*-means cluster analysis. The parameters given are the coefficient of determination (R^2), root mean squared error (RMSE, %), residual prediction deviation (RPD), and the ratio of performance to interquartile distance (RPIQ). The range of values reflects the results of 10 random iterations of the models. Results are also shown for full sample set models with no subsetting for comparison.

had lower (i.e., better) RMSE values, but were otherwise not quite as robust as the full sample set models' results.

In the cluster analysis of the MIR spectra, the distribution of samples was heavily weighted toward the Cluster 0 (137 samples, ~0–52% C_t) and Cluster 2 (132 samples, ~0–11% C_t) subsets. The Cluster 1 subset contained only 38 samples (~15–56% C_t) and was validated with full cross validation instead of independent validation. Table 7 and Figure 5(b) present the results of the prediction models from the cluster subsets, as well as those from the full sample set models for comparison. The results for the Cluster 0 subset models are broadly similar to those of the full sample set models but overall they are not an improvement. Results from the full cross validation of Cluster 1 subset were slightly higher for calibration but much lower for validation than the full cross validation of the full sample set. In general, the Cluster 1 model is not as robust as the full sample set model. The overall performance of Cluster 2 subset models is not

Table 6: Detailed partial least squares regression model results for soil total carbon (C_t) prediction from the subsets of visible/near-infrared diffuse reflectance spectra based on spectral classification with k-means cluster analysis. The range of values reflects the results of 10 random iterations of the models, and the number in parentheses is the mean. Detailed results are also given for full sample set models with no subsetting for comparison. For models with full cross validation (i.e., leave-one-out cross validation), the same samples used to calibrate the model were used to validate the model.

	Calibration			Validation				
	n^a	$R^{2,b}$	RMSE (%)^c	n	R^2	RMSE (%)	RPD^d	RPIQ^e
Cluster 0	78	0.93	4.52	Full cross validation	0.88	5.87	2.86	5.40
Cluster 1	87	0.68–0.88 (0.77)	1.92–3.26 (2.86)	37	0.60–0.91 (0.75)	1.74–3.47 (2.89)	1.54–3.33 (2.16)	1.94–5.50 (3.14)
Cluster 2	73	0.54–0.96 (0.81)	0.65–2.22 (1.29)	32	0.62–0.91 (0.80)	0.98–1.72 (1.33)	1.67–3.34 (2.39)	0.79–2.56 (1.71)
Full sample set	215	0.83–0.96 (0.90)	2.82–5.84 (4.30)	92	0.74–0.95 (0.88)	3.10–5.83 (4.30)	1.89–4.54 (3.28)	1.80–3.92 (3.06)
Full sample set	307	0.95	3.09	Full cross validation	0.94	3.39	4.09	3.80

[a] Number of samples.
[b] Coefficient of determination.
[c] Root mean squared error.
[d] Residual prediction deviation.
[e] Ratio of performance to interquartile distance.

Table 7: Detailed partial least squares regression model results for soil total carbon (C_t) prediction from the subsets of mid-infrared diffuse reflectance spectra based on spectral classification with k-means cluster analysis. The range of values reflects the results of 10 random iterations of the models, and the number in parentheses is the mean. Detailed results are also given for full sample set models with no subsetting for comparison. For models with full cross validation (i.e., leave-one-out cross validation), the same samples used to calibrate the model were used to validate the model.

	Calibration			Validation				
	n^a	$R^{2,b}$	RMSE (%)^c	n	R^2	RMSE (%)	RPD^d	RPIQ^e
Cluster 0	96	0.78–0.96 (0.90)	1.49–4.07 (2.45)	41	0.55–0.91 (0.81)	2.08–4.67 (3.43)	1.13–3.20 (2.34)	1.77–5.65 (3.31)
Cluster 1	38	0.98	1.89	Full cross validation	0.86	5.19	2.62	3.93
Cluster 2	92	0.88–0.99 (0.95)	0.15–0.58 (0.33)	40	0.77–0.90 (0.85)	0.39–0.82 (0.56)	1.50–2.84 (2.36)	1.30–3.33 (2.33)
Full sample set	215	0.93–0.98 (0.95)	1.68–3.61 (2.98)	92	0.92–0.95 (0.94)	2.94–3.78 (3.38)	3.48–4.68 (4.03)	2.61–4.61 (3.82)
Full sample set	307	0.95	3.12	Full cross validation	0.94	3.52	3.94	3.68

[a] Number of samples.
[b] Coefficient of determination.
[c] Root mean squared error.
[d] Residual prediction deviation.
[e] Ratio of performance to interquartile distance.

quite as good as the full sample set models, but the limited C_t range of Cluster 2 subset is apparent from its much lower range of RMSE values.

For this sample set, the spectral classification by k-means clustering and separate prediction model for each cluster was not an obvious improvement over the original full VNIR or MIR models. The most noticeable difference is the lower RMSE for the subset models from clusters limited to low C_t values.

We have found one other study that investigated the effect of subsetting a sample set by spectral classification for the prediction of soil carbon. Cierniewski et al. [17] tested the effect of four different unsupervised classification algorithms (k-means, expectation-maximization, Ward's Euclidean distance, and Lance and Williams' Euclidean distance) on simple linear regression results from VNIR data. These clustering algorithms produced five or six clusters, and the number of samples per cluster ranged from four to 56. This is in contrast to the method of k-means cluster analysis used in our study, where we specified that three clusters be produced to decrease the probability of a very low number of samples in a cluster that would not be adequate for robust modeling. Cierniewski et al. [17] found that the majority of their cluster subsets had improved R^2 values compared to the original full sample set. An increase in R^2 was not observed for the spectral classification subsets in the current work. Instead, the most significant improvement was a lower RMSE for many of the cluster subset models. Because other parameters such as RMSE were not provided in Cierniewski et al. [17], it is difficult to determine if this behavior is an effect of their subsetting study.

3.4. Percent Error of Prediction. The subset models with improved RMSE values but an otherwise less-robust performance may still hold an advantage over the original full sample set model. If a more accurate prediction of the low C_t samples makes a significant contribution to the lowered RMSE, the model could be very helpful in addressing the issue of large errors at low C_t values. To evaluate the error at these low C_t values, the percent error of prediction was calculated for the samples with C_t values less than 10% and the average value was reported for each model (Figure 6). We use percent error rather than RMSE for comparing the subset models with the full sample set model to normalize the error of the predicted value with respect to its measured value.

The mean value of the average percent error for each of the ten iterations of the full sample set model is ∼160–200%, but the average percent error for a single model could be up to almost 400% (Figure 6). For example, with a measured value of 1% C_t, an error of 400% would be translated to a predicted value of 5% C_t. The MIR full sample set models have lower average percent error, with a mean average percent error of ∼135–150% and a maximum average percent error of ∼200%. Many of the low RMSE subset models have noticeably lower average percent errors. The low C_t VNIR and MIR models and the Cluster 2 MIR models appear to have the most significant improvement, with average percent errors of ∼80% or less. For a measured value of 1% C_t, a percent error of 80% would reduce the predicted value to 1.8% C_t. Clusters 1 and 2 VNIR models also show moderate improvement, with all average percent error results below ∼175%. The average percent error of the low activity clay soils full cross validation model is slightly lower than the full sample set model for both the VNIR and MIR data. The organic-dominated soils subset includes only two samples with C_t <10%, so a comparison of average percent error is not as reliable in this case.

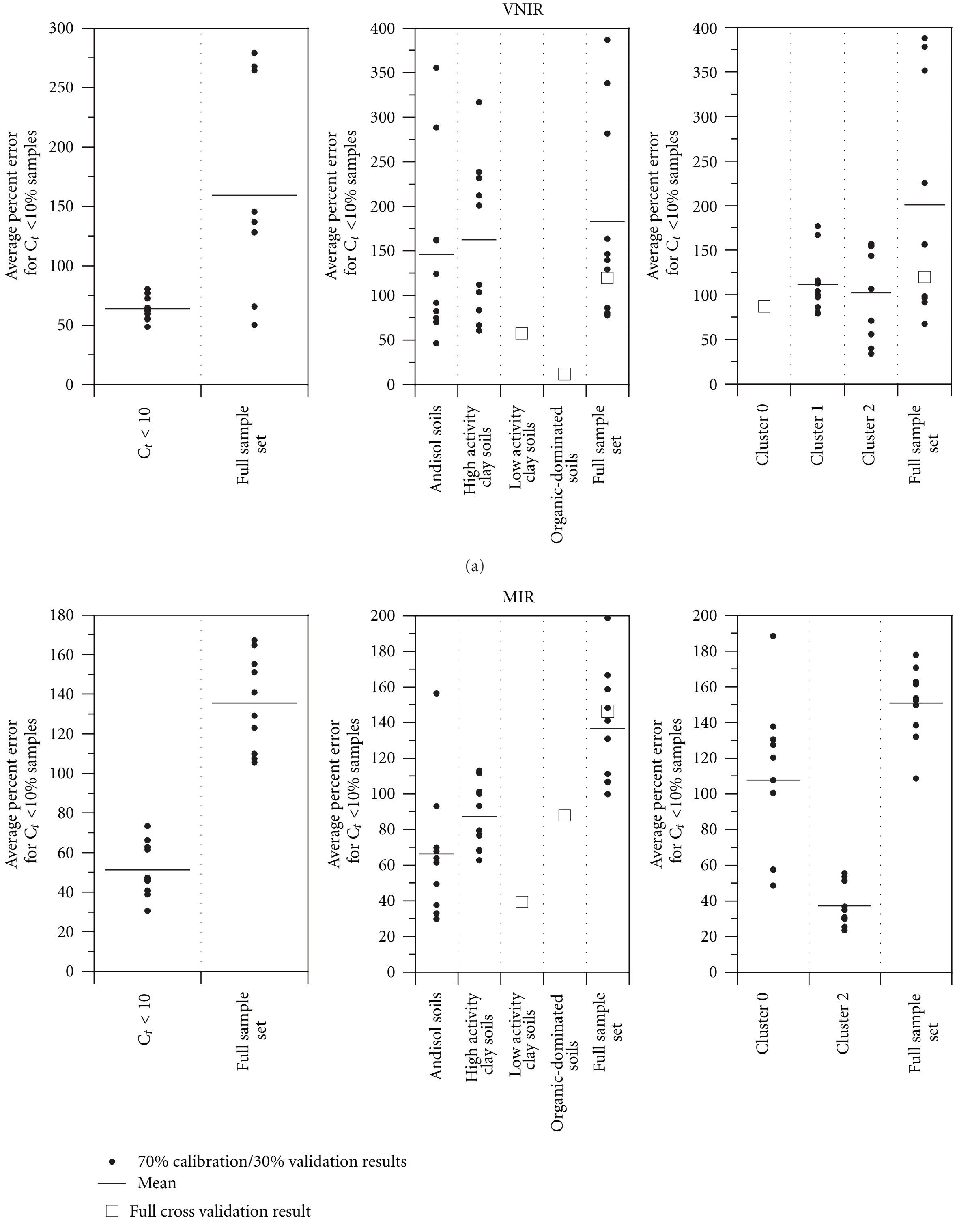
The subsets with the largest decreases in average percent error of prediction at low C_t content (i.e., $C_t < 10\%$) are the ones that included only low C_t samples in their models. The low C_t VNIR and MIR models contained samples with C_t values between ∼0 and 9.9% C_t, and the Cluster 2 MIR models had samples with C_t values between ∼0 and 11% C_t. These results suggest that a separate model for low C_t samples is beneficial for the accuracy of prediction for the samples in this range. This advantage is indicated by the RMSE of low C_t models, but may not be obvious from the R^2 parameter. The issue of relatively large errors of prediction for samples with very low C_t content has been understudied. To our knowledge there are no studies that have provided quantitative information addressing the degree of scatter observed for low C_t soils on most predicted versus measured plots.

3.5. Variation in Model Parameters. The ranges of PLSR model parameters produced by the 10 iterations of random calibration/validation set divisions in this study appear to be larger than the ranges of values encountered in previous studies where multiple PLSR model iterations were used. Brown et al. [13] reported results for five models produced from different random divisions of the sample set into 70% calibration and 30% validation groups. Values for organic carbon prediction from VNIR data ranged from 0.75 to 0.86 for R^2, 1.08 to 1.26 for RMSD, and 1.95 to 2.62 for RPD. Mouazen et al. [14] included three model iterations with random divisions into 90% calibration and 10% validation groups in their study. The exhaustive results are not reported, but visual estimation from plots of the mean and standard deviation for the R^2 and RMSE from the organic carbon prediction models suggests that the variation is similar to that in Brown et al. [13] or less. The greater range in model parameters observed in our study may be related to the testing of a greater number of iterations (i.e., 10 rather than five or three), or it could be related to a less obvious attribute, such as a greater variation in a spectral character within the sample set.

4. Summary and Conclusions

Our research has provided an introduction to the understudied idea of sample subsetting based on criteria that are simple and easily applied. This particular investigation of subsetting for C_t prediction had varied results with our Hawaiian soils sample set. Of all the different subset models created based on C_t content, soil order, and spectral classification, the subset of high activity clay soil orders was the only one to show improvement across all parameters (i.e., R^2, RMSE, RPD, and RPIQ) compared to the full sample set. Notably, one significant advantage was discovered; the subsets including only low C_t samples (e.g., C_t < 10% subset, MIR Cluster 2 subset) produced models with much lower RMSE values compared to the full sample set models, even though the other model parameters were not as robust. The lower RMSE for these models corresponds to a significant decrease in the percent error of predictions for low C_t samples, which could be very helpful for the analysis of soils with low C_t content or monitoring of small changes in C_t. Incorporation of a low C_t subset model in the future prediction of unknown soils C_t values could be done by first employing a model created with the full possible range of C_t values and then utilizing the separate low C_t subset model if the soil is predicted to have low C_t.

As seen from this study and previous studies, the effect of subsetting can have different results depending on the character of the sample set and the number of samples it includes. A small sample size may have limited the improvement possible by subsetting in the current work. In an effort to keep the size of subsets large enough for regression analysis, the subsetting may have been too coarse (e.g., too few subsets for C_t prediction by soil order and spectral classification). The types of subsetting strategies explored here may be most helpful for large datasets and should be tested with further research. Regardless of the strategy used to develop



(a)

(b)

Figure 6: Average percent error of the C_t <10% portion of the (a) visible/near-infrared (VNIR) and (b) mid-infrared (MIR) subset and full sample set models in this study. The range of values reflects the results of 10 random iterations of the models. The VNIR and MIR high C_t models and the MIR Cluster 1 models were not included because all samples had C_t >10%.

a model, our results suggest that multiple iterations of models with different calibration/validation groupings may help to produce a more complete picture of the overall model quality.

Acknowledgments

This research was supported by USDA CSREES TSTAR Project 2009-34135-20183 and UHM College of Tropical Agriculture and Human Resources (CTAHRs) Hatch Project HA-154. The authors thank J. Hempel, L. West, T. Reinsch, L. Arnold, and R. Nesser of the NRCS National Soil Survey Center in Lincoln, NE, USA, for help with access, sampling, and scanning of the archived samples; L. Muller and A. Quidez for help with scanning of samples at UHM; Drs. G. Uehara, R. Yost, and D. Beilman of UHM for the support of this project. They also appreciate the Hawaii landowners, managers, and extension agents that gave them access to their fields for collecting soil samples. These include from Kauai: R. Yamakawa and J. Gordines (CTAHR), S. Lupkes (BASF), and Grove Farms; from Oahu: R. Corrales, A. Umaki, and J. Grzebik (CTAHR), Hoa Aina, MAO Organic Farm, Nii Nursery, J. Antonio and M. Conway (Dole), C. and P. Reppun, L. Santo, T. Jones, and N. Dudley (HARC), and A. Sou (Aloun Farms); from Molokai: A. Arakaki (CTAHR), K. Duvchelle (NRCS), and R. Foster (Monsanto); from Maui: J. Powley and D. Oka (CTAHR), M. Nakahata and M. Ross (HC&S), T. Callender (Ulupono), and B. Abru.

References

[1] J. B. Reeves III, G. W. McCarty, and V. B. Reeves, "Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soils," *Journal of Agricultural and Food Chemistry*, vol. 49, no. 2, pp. 766–772, 2001.

[2] 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 Science Society of America Journal*, vol. 66, no. 2, pp. 640–646, 2002.

[3] K. D. Shepherd and M. G. Walsh, "Development of reflectance spectral libraries for characterization of soil properties," *Soil Science Society of America Journal*, vol. 66, no. 3, pp. 988–998, 2002.

[4] R. A. V. Rossel, D. J. J. Walvoort, A. B. McBratney, L. J. Janik, and J. O. Skjemstad, "Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties," *Geoderma*, vol. 131, no. 1-2, pp. 59–75, 2006.

[5] G. M. Vasques, S. Grunwald, and J. O. Sickman, "Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra," *Geoderma*, vol. 146, no. 1-2, pp. 14–25, 2008.

[6] G. M. Vasques, S. Grunwald, and J. O. Sickman, "Modeling of soil organic carbon fractions using visible—near-lnfrared spectroscopy," *Soil Science Society of America Journal*, vol. 73, no. 1, pp. 176–184, 2009.

[7] R. A. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, no. 1-2, pp. 46–54, 2010.

[8] M. L. McDowell, G. L. Bruland, J. L. Deenik, S. Grunwald, and N. M. Knox, "Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy," *Geoderma*, vol. 189-190, pp. 312–320, 2012.

[9] K. Paustian, O. Andrén, H. H. Janzen et al., "Agricultural soils as a sink to mitigate CO_2 emissions," *Soil Use and Management*, vol. 13, no. 4, pp. 230–244, 1997.

[10] H. Tiessen, E. Cuevas, and P. Chacon, "The role of soil organic matter in sustaining soil fertility," *Nature*, vol. 371, no. 6500, pp. 783–785, 1994.

[11] E. T. Craswell and R. D. B. Lefroy, "The role and function of organic matter in tropical soils," *Nutrient Cycling in Agroecosystems*, vol. 61, no. 1-2, pp. 7–18, 2001.

[12] R. Lal, "Soil carbon sequestration impacts on global climate change and food security," *Science*, vol. 304, no. 5677, pp. 1623–1627, 2004.

[13] D. J. Brown, R. S. Bricklemyer, and P. R. Miller, "Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana," *Geoderma*, vol. 129, no. 3-4, pp. 251–267, 2005.

[14] A. M. Mouazen, B. Kuang, J. De Baerdemaeker, and H. Ramon, "Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy," *Geoderma*, vol. 158, no. 1-2, pp. 23–31, 2010.

[15] D. V. Sarkhot, S. Grunwald, Y. Ge, and C. L. S. Morgan, "Comparison and detection of total and available soil carbon fractions using visible/near infrared diffuse reflectance spectroscopy," *Geoderma*, vol. 164, no. 1-2, pp. 22–32, 2011.

[16] B. E. Madari, J. B. Reeves, M. R. Coelho et al., "Mid- and near-infrared spectroscopic determination of carbon in a diverse set of soils from the Brazilian national soil collection," *Spectroscopy Letters*, vol. 38, no. 6, pp. 721–740, 2005.

[17] J. Cierniewski, C. Kaźmierowski, K. Kuśnierek et al., "Unsupervised clustering of soil spectral curves to obtain their stronger correlation with soil properties," in *Proceedings of the 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, (WHISPERS '10)*, Reykjavik, Iceland, June 2010.

[18] G. M. Vasques, S. Grunwald, and W. G. Harris, "Spectroscopic models of soil organic carbon in Florida, USA," *Journal of Environmental Quality*, vol. 39, no. 3, pp. 923–934, 2010.

[19] AOAC International, *Official Methods of Analysis of AOAC International*, AOAC International, Arlington, Va, USA, 16th edition, 1997.

[20] P. C. Williams, "Variables affecting near-infrared reflectance spectroscopic analysis," in *Near-Infrared Technology in the Agricultural and Food Industries*, P. Williams and K. Norris, Eds., pp. 143–167, American Association of Cereal Chemists, St. Paul, Minn, USA, 1987.

[21] V. Bellon-Maurel, E. Fernandez-Ahumada, B. Palagos, J. M. Roger, and A. McBratney, "Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy," *Trends in Analytical Chemistry*, vol. 29, no. 9, pp. 1073–1081, 2010.

[22] W. R. B. IUSS Working Group, "World reference base for soil resources," World Soil Resources report no. 103, FAO, Rome, Italy, 2006.