Modeling of Soil Organic Carbon Fractions Using Visible–Near-Infrared Spectroscopy

Gustavo M. Vasques
Sabine Grunwald*
Soil and Water Science Dep.
Univ. of Florida
2169 McCarty Hall
P.O. Box 110290
Gainesville, FL 32611

James O. Sickman
Dep. of Environmental Sciences
Univ. of California
Riverside, CA 92521

There is a pressing need for rapid and cost-effective tools to estimate soil C across larger landscapes. Visible–near-infrared diffuse reflectance spectroscopy (VNIRS) offers comparable levels of accuracy to conventional laboratory methods for estimating various soil properties. We used VNIRS to estimate soil total organic C (TC) and four organic C fractions in 141 samples collected in the Santa Fe River watershed of Florida. The C fractions measured were (in order of decreasing potential residence time in soils): recalcitrant C (RC), hydrolyzable C (HC), hot-water-soluble C (SC), and mineralizable C (MC). Soil samples were scanned in the visible–near-infrared spectral range. Six preprocessing transformations were applied to the soil reflectance, and five multivariate techniques were tested to model soil TC and the organic C fractions: stepwise multiple linear regression (SMLR), principal components regression, partial least squares regression (PLSR), regression tree, and committee trees. Total organic C was estimated with the highest accuracy, obtaining a coefficient of determination using a validation set (R_v^2) of 0.86, followed by RC (R_v^2 = 0.82), both using PLSR. The SC fraction was modeled best by SMLR (R_v^2 = 0.70), while PLSR produced the best models of MC (R_v^2 = 0.65) and HC (R_v^2 = 0.40). The addition of TC as a predictor improved the VNIRS models of the soil organic C fractions. Our study indicates the suitability of VNIRS to quantify soil organic C pools with widely varying turnover times in soils, which are important in the context of C sequestration and climate change.

Abbreviations: CT, committee trees; HC, hydrolyzable organic carbon; LOG, log(1/reflectance) transformation; MC, mineralizable organic carbon; NGD, Norris gap derivative across a seven-band window; NRA, normalization by the range; PCR, principal components regression; PLSR, partial least squares regression; RC, recalcitrant organic carbon; $RMSE_c$, root mean square error of calibration; $RMSE_v$, root mean square error of validation; RPD, residual prediction deviation; RT, regression tree; SC, hot-water-soluble organic carbon; SFRW, Santa Fe River watershed; SGD, Savitzky–Golay first derivative using a first-order polynomial across a nine-band window; SMLR, stepwise multiple linear regression; TC, total organic carbon; VNIRS, visible–near-infrared diffuse reflectance spectroscopy.

Soil organic C sequestration has received much attention recently as the concentration of CO_2 rises in the atmosphere, intensifying climate change (Keeling et al., 1995; Carbon Dioxide Information Analysis Center, 2008; Grunwald, 2008). Long-term sequestration of C in soils typically involves the decomposition of biologically labile organic matter into more recalcitrant macromolecules through the process of humification (Quideau, 2006). Several discrete organic C pools can be identified on the basis of size, turnover rate, and ecosystem function. The smallest pool with the most rapid turnover rate is typically designated *labile* (e.g., residence time of days to years), and the larger pool, with the longest residence time (e.g., decades to thousands of years), is described as *recalcitrant* (McLauchlan and Hobbie, 2004; Rice, 2006). Thus, quantifying the different compartments of soil organic C improves our understanding of how and at what rate stable forms of C are being formed or lost in soils. Moreover, the dynamics of soil organic C across a landscape are strongly controlled by environmental determinants such as temperature and moisture, which are sensitive to climate change. Recent evidence suggests that soil C is being lost across wide regions in response to climate change (Bellamy et al., 2005). Better tools are needed to monitor changes in soil organic C across large regions and to provide inputs to process-based models of C dynamics (Parton et al., 1983; Coleman and Jenkinson, 1996).

Characterizing organic C pools across large regions is critical to understanding the dynamics of soil C in the context of climate change. Measurement of discrete soil organic C fractions is time consuming, however, and requires intensive field sampling and costly laboratory analyses. An alternative approach to estimate soil C in a cost-effective manner is to build soil spectral libraries using VNIRS and chemometric modeling (Shepherd and Walsh, 2002; Brown et al., 2006). Visible–near-infrared diffuse reflectance spectroscopy has been used in the last decades for the rapid characterization of various materials (McClure, 2003). Numerous soil C models derived from visible–near-infrared spectra have been presented and validated (Chang and Laird, 2002; Dunn et al., 2002; McCarty et al.,

Soil Sci. Soc. Am. J. 73:176-184
doi:10.2136/sssaj2008.0015
Received 19 Jan. 2008.
*Corresponding author (sabgru@ufl.edu).

677 S. Segoe Rd. Madison WI 53711 USA

2002; Shepherd and Walsh, 2002; Islam et al., 2003; Brown et al., 2005). Although many soil properties have been investigated using VNIRS, less attention has been given to VNIRS modeling of soil physical and chemical organic C fractions.

Our objective was to estimate TC and four soil C fractions using VNIRS. These fractions are, in order of decreasing stability and residence time in soils: RC, HC, SC, and MC.

MATERIALS AND METHODS

Field and Laboratory Measurements

Soil samples were collected in the Santa Fe River watershed (SFRW), a 3585-km^2 watershed in north-central Florida. A total of 141 soil samples were collected from the surface to a depth of 30 cm and sieved through a 2-mm mesh. The soil samples proportionally represent all soil orders and land uses that occur in the watershed. The stratified random sampling design we used was presented in Vasques et al. (2008). Thirty-six percent of the samples occurred in Ultisols, 28% in Spodosols, and 22% in Entisols. Other soil orders accounted for the remaining 14% of the samples, and included Alfisols (11%), Mollisols (2%), and Inceptisols (1%). Major land uses where the samples were collected were pine plantations (28%), improved pasture (15%), upland forest (14%), and wetlands (13%). The remaining land uses included urban (11%), agriculture (10%), and rangelands (9%).

The soil samples were dried for 12 h at 45°C and scanned using a QualitySpec Pro spectroradiometer (Analytical Spectral Devices, Boulder, CO). The instrument collects diffuse reflected light in the wavelength range of 350 to 2500 nm, with 10 co-added scans averaged at 1-nm intervals. An average spectral curve was calculated based on four scans of each soil sample, rotated by an angle of 90°.

The soil samples were sieved through a 2-mm mesh and ball milled before chemical analysis of the organic C fractions. In our study area, organic C represents >98% of the total soil C (N.B. Comerford, personal communication, 2005; Guo et al., 2006); therefore, soils were not pretreated with acid to remove carbonates. All analytical methods used to measure organic C concentrations in the samples are well documented in the literature and are in routine use. Total organic C was determined by high-temperature combustion on a FlashEA 1112 Elemental Analyzer (Thermo Electron Corp., Waltham, MA). Recalcitrant organic C was measured on the elemental analyzer using soil samples that were refluxed with 6 mol L^{-1} HCl for 16 h, following the methods of Paul et al. (2001) and McLauchlan and Hobbie (2004). Hydrolyzable organic C was computed as the difference between TC and RC. Soluble organic C was extracted using hot water, according to Sparling et al. (1998) and Gregorich et al. (2003), and measured on a Shimadzu TOC 5050 Analyzer (Shimadzu Scientific Instruments Inc., Columbia, MD) using Pt-catalyzed combustion and nondispersive infrared detection of CO_2. The extracts were centrifuged, decanted, and filtered using a 0.2-μm filtration membrane before the determination of SC.

The soil organic C mineralization rate was estimated based on soil respiration inside an incubation chamber. Before incubation, 1 g of soil was wetted daily to 100% water holding capacity for the first 5 d in 12-mL vials and preincubated in an open system at 35°C in the dark. After the preincubation, the vials were filled with CO_2–free air, sealed with rubber septa, and incubated at 35°C in the dark. The first measurement of the CO_2 concentration was taken after 3 d of incubation (the eighth day) using a CO_2 coulometer (UIC Inc., Joliet, IL) with CO_2–free air used as a purge and carrier gas. The subsequent CO_2 concentration measurements were taken on a weekly basis until the 36th day of incubation. After the first incubation period of 3 d, during which CO_2 release was relatively high, mineralization rates became constant during the remainder of the incubation period. The rate of CO_2 release was modeled from the eighth until the 36th day of incubation using linear regression (R^2 = 0.98, data not shown). Mineralizable organic C was calculated by integrating these measurements between the 15th and 29th day of incubation.

The TC and C fractions covered a great variety of different soils and land uses within the watershed. Our aim was to represent this environmental variability in the spectral data set; thus, high and low values (e.g., high TC in wetlands, low TC in Entisols) were not excluded from the data set.

All soil organic C properties were positively skewed, and had a lognormal frequency distribution. Thus, the VNIRS models were developed based on $\log_{10}$–transformed values that approximated a Gaussian distribution.

Pretreatment of Soil Spectra and Multivariate Methods

The collected soil spectral curves were composed of 2151 reflectance measurements (bands) for each sample. The average soil spectral curves, obtained from the four rotations, were smoothed across the spectral bands (wavelengths) using a Savitzky–Golay smoothing algorithm (Savitzky and Golay, 1964) with a third-order polynomial across a nine-band window, and then averaged (pooled) across 10-nm intervals to match the spectral resolution of the spectroradiometer in the near-infrared region (Analytical Spectral Devices, 2008). This resulted in the reduction of the soil spectra to 214 reflectance values.

We compared six preprocessing transformations to prepare the soil spectra for analysis, and five multivariate methods to develop the predictive models. The six preprocessing transformations were assembled to represent an array of different types of pretreatments that can be used to transform spectral data, and included smoothing, standardization, normalization, and derivation routines. They were selected based on a comprehensive comparative analysis described by Vasques et al. (2008). The six preprocessing transformations tested were: Savitzky–Golay smoothing across a nine-band window, followed by averaging across a 10-band window (SAV); log(1/reflectance) (LOG); normalization by the range (NRA); Norris gap derivative across a seven-band window (NGD); Savitzky–Golay first derivative using a first-order polynomial across a nine-band window (SGD); and standard normal variate transformation (SNV). Savitzky–Golay smoothing across a nine-band window, followed by averaging across a 10-band window, was used as a standard preparation of the soil spectral curves; the SAV-transformed curves served as input to all the other preprocessing transformations tested. All preprocessing transformations were implemented in the Unscrambler 9.5 software (CAMO Software Inc., Woodbridge, NJ).

The five multivariate methods we compared to model soil organic C fractions using VNIRS were stepwise multiple linear regression (SMLR), principal components regression (PCR), partial least squares regression (PLSR), regression tree (RT), and committee trees (CT). All the models were developed using a calibration data set comprising 102 measurements that were randomly selected from the whole data set. A validation set of 39 measurements was used to evaluate the accuracy of the different multivariate methods and preprocessing transformations. The coefficient of determination (R^2) was used to compare the models, but other error statistics were provided, including the RMSE and the residual prediction deviation (RPD; Williams, 1987):

$$R^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 \Big/ \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \quad [1]$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2 \Big/ n} \quad [2]$$

$$\mathrm{RPD} = \mathrm{SD}_{\mathrm{val}} \Big/ \mathrm{RMSE}_v \sqrt{n/(n-1)} \quad [3]$$

where $\hat{y}$ is the predicted value; $\bar{y}$ is the mean of the observed values; y is the observed value; n is the number of predicted or observed values with $i = 1, 2, \ldots, n$; $\mathrm{SD}_{\mathrm{val}}$ is the standard deviation of the validation set; and $\mathrm{RMSE}_{\mathrm{v}}$ is the root mean square error of validation.

Stepwise multiple linear regressions used a cutoff F threshold of 0.05 to include and exclude variables from the models. Principal components regression and PLSR were developed using full (leave-one-out) cross-validation on the calibration set. The optimum number of principal components (PCs) was chosen based on the RMSE of cross-validation ($\mathrm{RMSE}_{\mathrm{cv}}$; Martens and Næs, 1989), up to a maximum of 20 PCs. Stepwise regression was performed in SPSS 11.0 (SPSS Inc., Chicago, IL), while PCR and PLSR were developed in the Unscrambler 9.5.

Regression trees (Breiman et al., 1984) and CT (Breiman, 1996) are nonparametric data mining techniques that have recently been incorporated in soil science (Fidêncio et al., 2002; Shepherd and Walsh, 2002; Brown et al., 2006). They have no assumptions about the distribution of the data and can identify nonlinear relationships in the data, offering an alternative to linear methods to analyze soil properties. The CT models were generated by bagging, using a committee of 100 trees, with a maximum number of sample redraws of three. The trees were pruned based on the least squares error. The estimated values were calculated by averaging multiple versions of predictors generated by bootstrapping, using 10-fold cross-validation (Breiman, 1996). Regression tree and CT were implemented in CART 5.0 (Salford Systems, San Diego, CA).

To improve the estimations of soil organic C fractions, TC was added as an additional predictor along with the visible–near-infrared reflectance values, and PLSR was used to derive the models. Complementarily, a simple linear regression model was derived for each soil organic C fraction using TC as the only explanatory variable.

Table 1. Descriptive statistics of measured soil organic C properties.

Statistic	Property			$\log_{10}$ of property		
	Whole set	Calibration	Validation	Whole set	Calibration	Validation
	mg kg^{-1}			log(mg kg^{-1})		
	Total organic C					
Mean	14,828	16,625	10,128	4.0327	4.0631	3.9532
SE	1,852	2,518	909	0.0242	0.0305	0.0335
Median	10,529	10,885	9,438	4.0224	4.0368	3.9749
SD	21,993	25,427	5,675	0.2879	0.3082	0.2093
CV	148.32	152.94	56.03	7.14	7.59	5.29
Skewness	6.35	5.51	2.01	1.33	1.31	0.42
Kurtosis	46.74	34.53	4.95	3.98	3.58	0.29
Range	199,318	199,318	25,870	1.8788	1.8788	0.9213
Min.	2,670	2,670	3,523	3.4265	3.4265	3.5469
Max.	201,988	201,988	29,393	5.3053	5.3053	4.4682
	Recalcitrant organic C					
Mean	11,122	12,634	7,165	3.8704	3.9026	3.7862
SE	1,616	2,200	761	0.0276	0.0348	0.0381
Median	7,382	7,730	6,387	3.8682	3.8882	3.8053
SD	19,194	22,223	4,751	0.3277	0.3517	0.2381
CV	172.58	175.90	66.31	8.47	9.01	6.29
Skewness	6.64	5.75	2.24	1.03	0.96	0.50
Kurtosis	51.49	38.03	5.98	3.28	2.97	0.30
Range	180,587	180,587	23,069	2.1986	2.1986	1.0508
Min.	1,150	1,150	2,253	3.0609	3.0609	3.3527
Max.	181,738	181,738	25,322	5.2594	5.2594	4.4035
	Hydrolyzable organic C					
Mean	3,707	3,991	2,963	3.4619	3.4900	3.3884
SE	277	369	240	0.0275	0.0305	0.0582
Median	2,892	2,921	2,749	3.4612	3.4655	3.4392
SD	3,292	3,725	1,502	0.3261	0.3079	0.3635
CV	88.80	93.34	50.69	9.42	8.82	10.73
Skewness	4.58	4.19	1.05	−1.41	−0.18	−3.43
Kurtosis	30.09	23.98	2.57	8.51	2.03	16.65
Range	29,362	29,143	8,022	2.9037	2.0607	2.3417
Min.	37	256	37	1.5646	2.4076	1.5646
Max.	29,399	29,399	8,059	4.4683	4.4683	3.9063
	Hot-water-soluble organic C					
Mean	809	869	655	2.8287	2.8465	2.7824
SE	69	93	44	0.0196	0.0248	0.0274
Median	664	697	563	2.8218	2.8431	2.7501
SD	818	942	272	0.2324	0.2504	0.1710
CV	101.11	108.40	41.53	8.22	8.80	6.15
Skewness	7.57	6.73	0.95	1.01	0.97	0.33
Kurtosis	72.65	55.86	0.19	3.31	3.07	−0.88
Range	8,774	8,774	1,048	1.6104	1.6104	0.6218
Min.	221	221	329	2.3436	2.3436	2.5170
Max.	8,995	8,995	1,377	3.9540	3.9540	3.1388
	Mineralizable organic C					
Mean	111	120	86	1.9450	1.9737	1.8699
SE	9	12	8	0.0231	0.0283	0.0368
Median	90	94	71	1.9564	1.9744	1.8519
SD	107	120	51	0.2745	0.2856	0.2299
CV	96.39	99.94	59.80	14.12	14.47	12.29
Skewness	5.39	5.01	1.62	0.50	0.44	0.35
Kurtosis	41.36	33.98	2.22	1.28	1.38	0.06
Range	1,018	1,018	205	1.7614	1.7614	0.9667
Min.	18	18	25	1.2541	1.2541	1.3936
Max.	1,036	1,036	229	3.0154	3.0154	2.3603

RESULTS AND DISCUSSION

Descriptive Statistics

The descriptive statistics of all soil organic C properties measured, both in the original units and in log units, are shown in Table 1. Total organic C varied from 2670 to 201,988 mg kg^{-1}, with a mean of 14,828 mg kg^{-1} and median of 10,529 mg kg^{-1}. Except for HC, the range of values in the validation set was encompassed by the range of the calibration set. The mean and median values were smaller in the validation sets for all soil organic C properties due to the presence of C-rich wetland soils in the calibration set, which accounted for the extremely high values.

The SC and MC fractions represent the most labile C fractions in the soil, i.e., the most readily available to heterotrophic microorganisms. The SC fraction is composed of water-soluble organic compounds, including simple organic molecules. The HC fraction is less labile than SC, but still can be used by organisms utilizing enzymes and hydrolytic mechanisms to acquire soil nutrients. The remaining fraction, RC, is the most stable soil C pool and is composed of complex organic molecules of low decomposability by microorganisms (e.g., humic and fulvic acids) (Rice, 2006). This fraction is associated with the long-term accumulation of soil organic matter and is thus the most important fraction from a C sequestration perspective.

There was an inverse relationship between lability of the organic C fractions and their relative sizes in the SFRW soils (Table 1). On average, MC was the smallest organic C fraction, followed by SC, then HC; RC was the largest fraction. The average MC was 111 mg kg^{-1}, and the minimum and maximum MC values were 18 and 1036 mg kg^{-1}, respectively. During the MC incubations, soil C showed a steady rate of mineralization between the 15th and 29th days of incubation, with an average of 7.91 mg kg^{-1} d^{-1}. This steady C mineralization rate was confirmed until the 36th day of incubation, the day when the experiment was terminated.

The labile HC and SC fractions accounted for an average of 25 and 5% of TC, respectively. The HC concentrations ranged from 37 to 29,399 mg kg^{-1}, with an average of 3707 mg kg^{-1}. The SC concentrations varied from 221 to 8995 mg kg^{-1}, with an average of 809 mg kg^{-1}.

The RC fraction accounted for an average of 75% of the organic C in the samples. It ranged from 1150 to 181,738 mg kg^{-1}, with an average of 11,122 mg kg^{-1}. As the most stable organic C fraction, RC is composed mainly of complex humic substances with relatively high lignin contents and C/N ratios (Bouchard and Cochran, 2006; Rice, 2006). Recalcitrant organic C in the region of study possibly originates mainly from the humification of pine (*Pinus* sp.) litter and residues of pasture and crops. Saturated soil conditions also foster the accumulation of stable C forms due to reduced oxidation of organic materials (Bouchard and Cochran, 2006).

All soil organic C properties were highly variable across the SFRW, with coefficients of variation of at least 89% for the whole sample sets. The sampling design covered an extensive variety of soil types and land uses, from Histosols to Entisols, and from wetlands to urban areas, which explains the large range of values for the soil organic C properties.

The Pearson's correlations between the soil organic C properties were all significant at the 99% confidence level (Table 2). All correlation coefficients were >0.50, with the highest values between TC, RC, and SC. Mineralizable organic C had good linear correlation with all properties except HC, and HC had the lowest correlations with the other soil organic C properties. High significant linear correlations between TC and the soil organic C fractions justified the addition of TC as an auxiliary predictor in the VNIRS models of the soil organic C fractions.

Visible–Near-Infrared Spectroscopy Models of Soil Organic Carbon Properties

Among the five soil organic C properties investigated, TC was estimated with the greatest accuracy. The maximum coefficient of determination of validation (R_v^2) obtained for TC, 0.86, was from the PLSR model after LOG transformation, and the LOG-PLSR model also had the highest residual prediction deviation of the models tested (RPD = 2.64) (Table 3). Chang et al. (2001) categorized the accuracy and stability of their spectroscopy models based on the RPD values. Values >2.0 were considered stable and accurate predictive models; RPD values between 1.4 and 2.0 indicated fair models that could be improved by more accurate predictive techniques; RPD values <1.4 indicated poor predictive capacity. The TC models for the Santa Fe soils had comparable accuracy to models developed using VNIRS produced elsewhere (Chang et al., 2001; Chang and Laird, 2002; McCarty et al., 2002). We could not find VNIRS models of discrete soil C fractions in the literature; thus we could not compare our results for these properties.

Table 2. Pearson's correlation coefficients between the measured soil organic C properties†.

	LogTC	LogRC	LogHC	LogSC	LogMC
LogTC	1.00	0.98**	0.69**	0.90**	0.79**
LogRC	0.98**	1.00	0.54**	0.87**	0.78**
LogHC	0.69**	0.54**	1.00	0.64**	0.53**
LogSC	0.90**	0.87**	0.64**	1.00	0.79**
LogMC	0.79**	0.78**	0.53**	0.79**	1.00

** Correlation is significant at the 0.01 level.

† LogHC, $\log_{10}$ of hydrolyzable organic C; LogMC, $\log_{10}$ of mineralizable organic C; LogRC, $\log_{10}$ of recalcitrant organic C; LogSC, $\log_{10}$ of hot-water-soluble organic C; LogTC, $\log_{10}$ of total organic C.

Since the VNIRS models were validated and some of them produced RPD values >2.0, this indicates that the models are reliable and offer good generalization potential. In the case of HC, SC, and MC, RPD values <2.0 indicate that there is room for improvement of these models (Chang et al., 2001). The models obtained for each soil organic C property by the different multivariate methods, associated with their respective best preprocessing transformations among the six preprocessing transformations tested, are summarized in Table 3.

Usually, preprocessing transformations of spectral data improve the accuracy of regression models. Some studies have reported improvements of the regression models by using first and second derivatives (Dunn et al., 2002), normalization of the data (McCarty et al., 2002), and scatter corrections (Kooistra et al., 2003), while others found better results with untransformed reflectance data (Kooistra et al., 2001). In this study, the preferred preprocessing transformation associated with the different multivariate methods varied according to the soil organic C property investigated. Only NRA and SGD were not selected as the best preprocessing transformation for any of the multivariate methods or soil organic C properties investigated.

When considering the calibration quality, all soil organic C properties were estimated with high accuracy, even HC, whose CT model had a coefficient of determination of calibration (R_c^2) of 0.89. In terms of calibration, CT provided the best results, with R_c^2 varying from 0.87 for the MC model to 0.93 for the TC and RC models. When validated using the independent validation set, however, CT models performed poorly and were only better than the RT models. One reason for the poor performance of RT and CT is that these models are data mining techniques that require large data sets for robust model predictions. The 102 measurements contained in the calibration may have been too limiting to produce models that had

Table 3. Summary statistics of the models obtained for each soil organic C property by the different multivariate methods associated with their respective best preprocessing transformations.

Multivariate method†	Preprocessing transformation‡	Number of predictors or factors§	Calibration¶		Validation#		
			R_c^2	RMSE$_c$	R_v^2	RMSE$_v$	RPD
Log$_{10}$ of total organic C (log mg kg^{-1})							
SMLR	SNV	6	0.82	0.132	0.77	0.102	2.02
PCR	LOG	16	0.90	0.098	0.79	0.095	2.17
PLSR	LOG	12	0.93	0.082	0.86	0.078	2.64
RT	SAV	9	0.92	0.085	0.68	0.131	1.58
CT	NGD	214	0.93	0.087	0.72	0.129	1.60
Log$_{10}$ of recalcitrant organic C (log mg kg^{-1})							
SMLR	LOG	4	0.84	0.140	0.73	0.124	1.90
PCR	SNV	13	0.80	0.157	0.72	0.125	1.88
PLSR	SAV	11	0.90	0.109	0.82	0.108	2.17
RT	SAV	5	0.86	0.133	0.55	0.181	1.30
CT	SAV	214	0.93	0.096	0.69	0.142	1.65
Log$_{10}$ of hydrolyzable organic C (log mg kg^{-1})							
SMLR	SAV	2	0.47	0.223	0.23	0.315	1.14
PCR	NGD	8	0.48	0.222	0.40	0.283	1.27
PLSR	SAV	5	0.49	0.218	0.40	0.285	1.26
RT	SAV	7	0.72	0.163	0.16	0.338	1.06
CT	SNV	214	0.89	0.120	0.23	0.316	1.13
Log$_{10}$ of hot-water-soluble organic C (log mg kg^{-1})							
SMLR	SNV	7	0.88	0.087	0.70	0.095	1.77
PCR	SNV	12	0.78	0.118	0.65	0.101	1.67
PLSR	SNV	6	0.81	0.110	0.69	0.100	1.68
RT	SAV	4	0.80	0.113	0.44	0.146	1.16
CT	SAV	214	0.92	0.072	0.52	0.135	1.25
Log$_{10}$ of mineralizable organic C (log mg kg^{-1})							
SMLR	SAV	1	0.56	0.188	0.54	0.157	1.44
PCR	SNV	10	0.59	0.182	0.65	0.141	1.61
PLSR	SNV	6	0.69	0.159	0.65	0.137	1.66
RT	SAV	4	0.55	0.191	0.53	0.161	1.41
CT	LOG	214	0.87	0.106	0.51	0.164	1.38

† SMLR, stepwise multiple linear regression; PCR, principal components regression; PLSR, partial least squares regression; RT, regression tree; CT, committee trees.

‡ SNV, standard normal variate transformation; LOG, log(1/reflectance); SAV, Savitzky–Golay smoothing across a nine-band window, followed by averaging across a 10-band window; NGD, Norris gap derivative across a seven-band window.

§ Predictors refers to the number of reflectance bands used by the SMLR, RT, and CT models; factors refers to the number of principal components or partial least squares factors used by the PCR or PLSR models, respectively. Note that PCR, PLSR, and CT use all the available reflectance bands to calibrate the models, but in PCR and PLSR, these reflectance bands are first converted to factors, then the factors are used as predictors in the models.

¶ R_c^2, coefficient of determination of calibration; RMSE$_c$, root mean square error of calibration.

R_v^2, coefficient of determination of validation; RMSE$_v$, root mean square error of validation; RPD, residual prediction deviation.

good generalization capacity, i.e., that had high R_v^2 and RPD values and low RMSE$_v$ values.

In validation mode, except for HC, all models of soil organic C properties were robust, explaining at least 65% of the variability of the validation set in the case of MC, and up to 86% of the variability of the validation set for the TC model. Biplots of estimated and measured soil organic C concentrations, derived using the validation set, are presented in Fig. 1. The plots show trends of underestimation of high values and overestimation of low values for HC, SC, and MC. The estimated TC and RC values, however, closely approximated the 1:1 line and show little bias. This indicates that the TC and RC models can be reasonably generalized to estimate total organic C as well as the most stable organic C fraction. It is worth noting that the low HC value of 1.5646 in Fig. 1c in the validation set was lower than HC values found in the calibration set. Since this low range of HC values was not covered in the calibration set, it led to an extrapolation in validation mode. This explains the overestimation of this HC value.

Among the multivariate methods tested, PLSR provided the best validation results. Similar to PCR, PLSR is a robust statistical method that uses all the available reflectance data to build the models. Both methods can deal with collinearity and are fairly robust to nonlinearity and data outliers. The main advantage of PLSR relative to PCR is that it takes into account the variability of the target variable when calculating the factors, which is not the case with PCR (Martens and Næs, 1989).

The PLSR models generated for the different soil organic C properties selected a minimum of five factors (RC model) and a maximum of 12 (TC model). The number of partial least squares (PLS) factors was chosen based on the RMSE$_{cv}^2$. Figure 2 shows the cumulative percentage of explained variance by the number of PLS factors for the LOG-PLSR TC model. According to Fig. 2, since the TC model was developed using 12 PLS factors, it explained virtually 100% of the variability of both dependent (TC) and independent (reflectance data) variables. Thus, the PLSR model reduced the number of predictors from 214 reflectance bands to 12 factors while keeping almost all of the variability information contained in the 214 bands.

The PLSR coefficients used in the models of the five soil organic C properties are shown in Fig. 3. In both TC and RC models (Fig. 3a and 3b), important wavelengths concentrated around 400, 1000, 1400, 1900, and 2100 nm, and after 2200 nm. Since RC represented the greatest part of the TC, the RC and TC models were closely related and were sensitive to similar spectral regions. Models of HC, SC, and MC (Fig. 3c, 3d, and 3e) had important wavelengths approximately in the same regions, as shown by the relatively large coefficients around 1400 nm and between 2050 and 2400 nm. The SC and MC models also had important wavelengths close to 1900 nm, which is the region of absorbance features of OH and water.

Stepwise multiple linear regression and PCR were the second best multivariate methods. Stepwise multiple linear regression is a relatively rapid and easy technique to analyze multivariate data and requires that linear relationships exist between the target and predictor variables. Therefore, SMLR usually selects those predictors that have the strongest linear correlations with the target variable, which will reflect the highest predictive capacity.

Since SMLR did not use all the reflectance bands in the models, it suggests that the predictive information contained in the soil spectral curves is actually concentrated in a subset of important wavelengths. Based on this observation, one would expect that the different multivariate methods would consistently select the same spectral regions in the models. This is confirmed in Fig. 4, which shows the most important wavelengths used to estimate TC by four multivariate methods associated with their respective best preprocessing transformations of soil spectra. Stepwise multiple linear regression, PCR, and PLSR captured approximately the same regions of absorbance features of the main constituents of soil organic matter. The main absorbance regions selected in the models were, with the respective associated soil organic constituents in parentheses: ~400 nm (chromophorous groups), ~960 nm (organic pigments), ~1400 and 1900 nm (OH groups, including water), ~2000 to 2200 nm (CH and NH groups), and ~2,200 to 2400 nm (CH groups) (Goddu and Delker, 1960; Gaffey et al., 1993; Siesler et al., 2002; Analytical Spectral Devices, 2003).

Since the effect of moisture content as well as particle size was standardized by sieving and oven drying of the soil samples, one can expect that the reflectance values, especially at 1400 and 1900 nm, actually translate the interaction of soil organic matter with water and soil particles, and not the presence of water per se or differences in particle size.

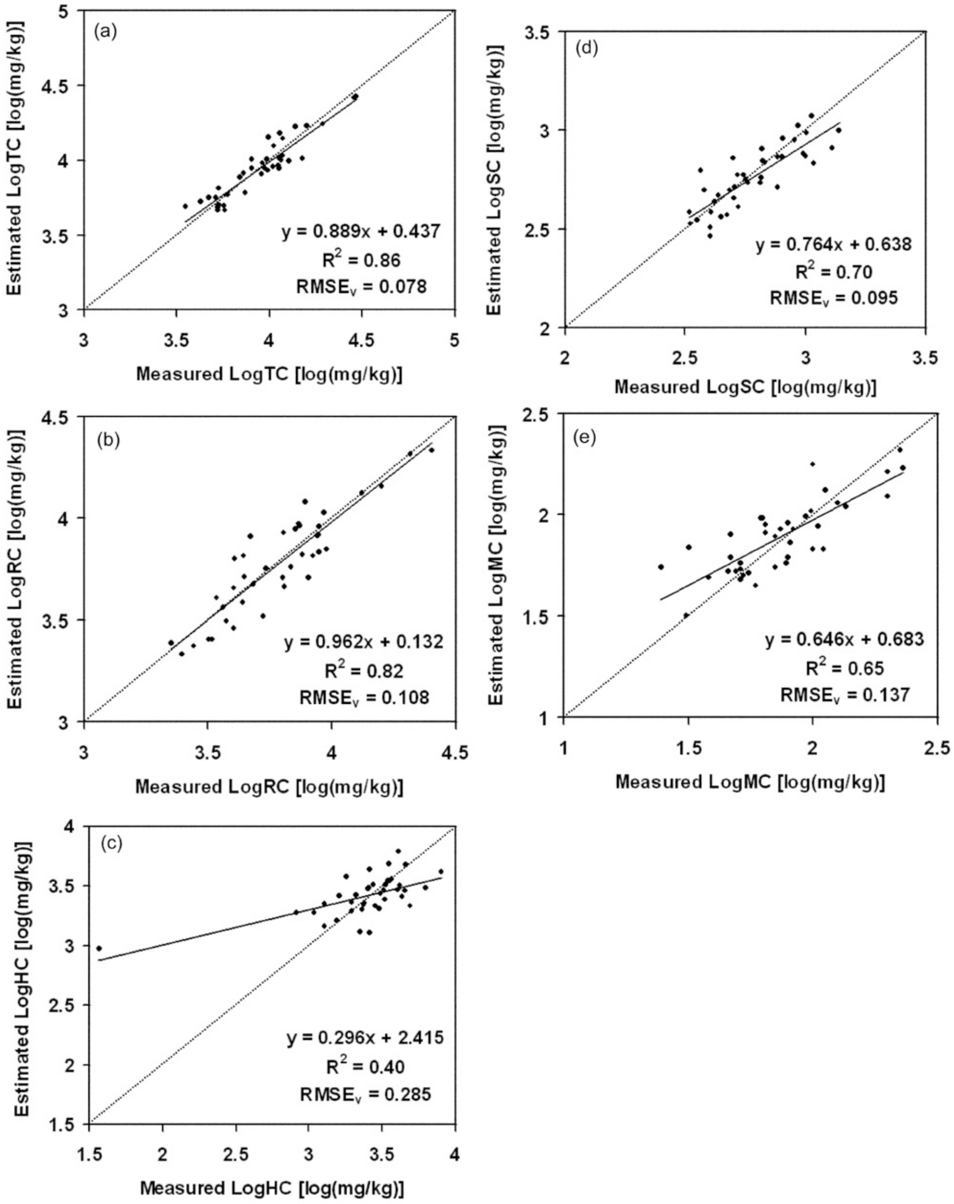

Fig. 1. Estimated vs. measured values in the validation of the best visible–near-infrared spectroscopy models of the soil organic C properties: (a) total organic C (LogTC) estimated by partial least squares regression (PLSR) using log(1/reflectance) transformation; (b) recalcitrant organic C (LogRC) estimated by PLSR using Savitzky–Golay smoothing across a nine-band window, followed by averaging across a 10-band window (SAV) transformation; (c) hydrolyzable organic C (LogHC) estimated by PLSR using SAV transformation; (d) hot-water-soluble organic C (LogSC) estimated by stepwise multiple linear regression using standard normal variate (SNV) transformation; and (e) mineralizable organic C (LogMC) estimated by PLSR using SNV transformation.

Regression trees use a different approach to select the most informative predictors in a model. Tree-based models partition the data, separating the target variable recursively into more homogeneous classes. The wavelengths selected by the RT model were mainly in the visible part of the spectrum, which suggests that RT estimated TC mainly based on the color of the soil. Because RT estimates the target variables as discrete values, or classes, RT was not as suitable for estimating TC and soil organic C fractions, and had the worst performance among the multivariate methods tested for all the soil organic C properties investigated.

As for TC, the best models of the soil organic C fractions (HC, RC, and SC) also consistently captured the regions of absorbance features (including overtones and combinations of the fundamental vibrations) of important chemical groups related to soil organic matter. This confirms that our VNIRS models were sensitive to soil organic components, and were not a mere consequence of loading the models with multiple predictors.

The models developed using TC as an auxiliary explanatory variable are presented in Table 4. When simple linear regression was used to estimate soil organic C fractions as a function of TC alone, the R_v^2 varied from 0.30 for the HC model to 0.91 for the RC model, whereas the $RMSE_v$ varied from 0.069 to 0.304. Recalcitrant organic C was estimated

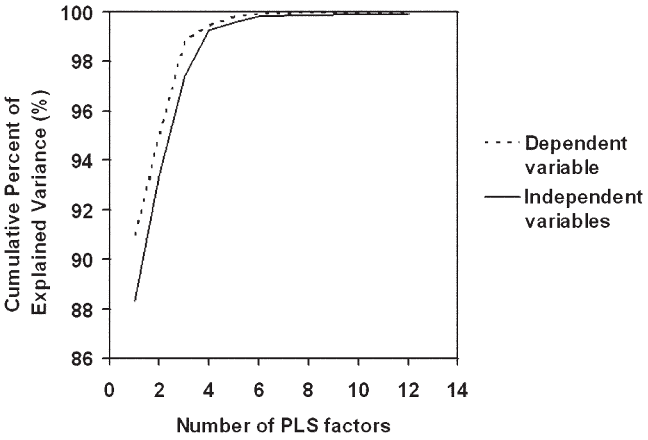

Fig. 2. Cumulative percentage of explained variance as a function of the number of partial least squares (PLS) factors for the total organic C model estimated by PLS regression using log(1/reflectance) transformation.

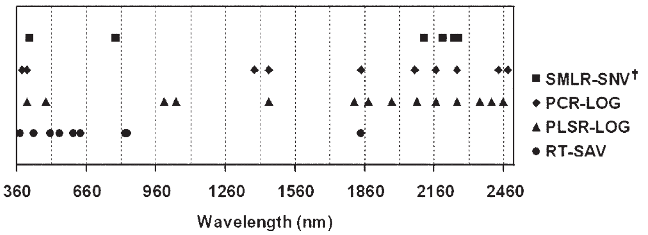

Fig. 4. Important wavelengths used in the total organic C models produced by four multivariate methods, associated with their best preprocessing transformations: stepwise multiple linear regression using standard normal variate transformation (SMLR-SNV), principle components regression using log(1/reflectance) transformation (PCR-LOG), partial least squares regression using log(1/reflectance) transformation (PLSR-LOG), and regression tree using Savitzky–Golay smoothing across a nine-band window, followed by averaging across a 10-band window transformation (RT-SAV).

with high accuracy, probably because it constituted the major part of the TC and thus had the most relevant chemical constituents of soil organic matter with absorbance features in the visible–near-infrared region.

When TC was added as an additional predictor in the PLSR models of the soil organic C fractions, all models improved except the HC model. The RC model showed a substantial response to the addition of TC, with the R_v^2 improving from 0.82 to 0.91. The SC model showed improvement of the R_v^2 from 0.69 to 0.81, whereas the MC model had improvement of the R_v^2 from 0.65 to 0.73. The RPD of all soil organic C fraction models improved with the addition of TC as a predictor, including the one from the HC model.

The improvement of the PLSR models with the addition of TC as a predictive variable was expected, since all soil organic C fractions were highly correlated with TC (Table 2).

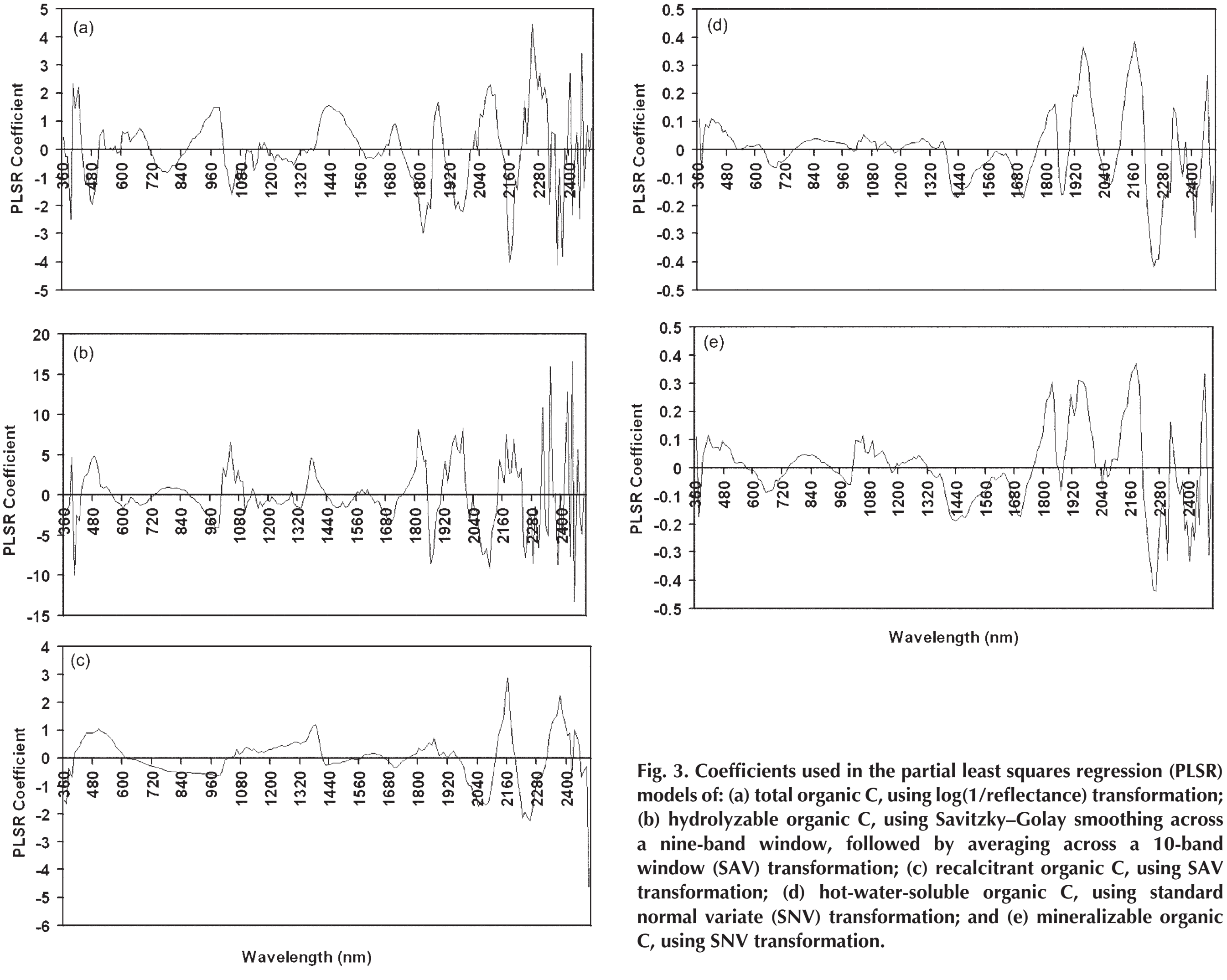

Fig. 3. Coefficients used in the partial least squares regression (PLSR) models of: (a) total organic C, using log(1/reflectance) transformation; (b) hydrolyzable organic C, using Savitzky–Golay smoothing across a nine-band window, followed by averaging across a 10-band window (SAV) transformation; (c) recalcitrant organic C, using SAV transformation; (d) hot-water-soluble organic C, using standard normal variate (SNV) transformation; and (e) mineralizable organic C, using SNV transformation.

Determination of total soil organic C in the laboratory is relatively easy and cheap, making it a very good auxiliary variable to VNIRS models of soil C fractions. More improvement could be achieved, especially in the HC models, if other properties correlated to HC were also included in the models. Ideally, these properties would be cheaply and easily measurable.

Table 4. Summary statistics of the models obtained for each soil organic C fraction by simple linear regression and by partial least squares regression (PLSR) using both soil reflectance transformed by Savitzky–Golay smoothing across a nine-band window, followed by averaging across a 10-band window (SAV) and $\log_{10}$ of total organic C (LogTC) as predictors.

Soil organic C fractions†	Number of predictors or factors‡	Calibration§		Validation¶		
		R_c^2	RMSE$_c$	R_v^2	RMSE$_v$	RPD
Simple linear regression using LogTC as predictor						
LogRC	1	0.96	0.071	0.91	0.069	3.40
LogHC	1	0.55	0.206	0.30	0.304	1.18
LogSC	1	0.83	0.104	0.75	0.086	1.97
LogMC	1	0.64	0.170	0.47	0.166	1.36
PLSR using SAV-transformed soil reflectance and LogTC as predictors						
LogRC	2	0.96	0.071	0.91	0.072	3.25
LogHC	7	0.60	0.194	0.36	0.280	1.28
LogSC	4	0.87	0.091	0.81	0.075	2.26
LogMC	4	0.73	0.149	0.73	0.122	1.86

† LogRC, $\log_{10}$ of recalcitrant organic C; LogHC, $\log_{10}$ of hydrolyzable organic C; LogSC, $\log_{10}$ of hot-water-soluble organic C; LogMC, $\log_{10}$ of mineralizable organic C.

‡ Predictors refers to the number of predictors used by the simple linear regression models, only LogTC in this case; factors refers to the number of partial least squares factors used by the PLSR models.

§ R_c^2, coefficient of determination of calibration; RMSE$_c$, root mean square error of calibration.

¶ R_v^2, coefficient of determination of validation; RMSE$_v$, root mean square error of validation; RPD, residual prediction deviation.

CONCLUSIONS

Our modeling study indicated that, besides TC, other ecologically relevant organic C fractions in soils can be estimated from soil spectra in the visible–near-infrared range. Total soil organic C and the most stable and largest pool of C in the soils (RC) were estimated with high accuracy. In contrast, the organic C pool with moderate lability, HC, was difficult to model using soil spectra. Visible–near-infrared spectroscopy models of the smallest and most labile fractions of soil organic C, SC and MC, had intermediate predictive power.

Separate validation of the models provided evidence to support the use of the VNIRS models developed in this study to estimate soil total organic C and soil organic C fractions in soils with similar characteristics in Florida. Soil C fractions are important components of the soil C cycle and are essential inputs in various process-based soil C modeling systems.

Given that soil C fractions are more costly and laborious properties to measure, VNIRS models offered good estimates of these properties in a cost-effective way. The VNIRS technique offers the possibility for measuring important soil organic C pools across large areas and may be useful for monitoring changes in C sequestration and storage in the context of climate change. Furthermore, the addition of TC as an explanatory variable improved the residual prediction deviation of the VNIRS models of all the soil organic C fractions analyzed. Soil total organic C is relatively cheap to measure, justifying its inclusion in the VNIRS models of soil organic C fractions.

Our results demonstrate the effectiveness of using VNIRS to estimate ecologically relevant soil organic C fractions in a mixed-use landscape in north-central Florida. Further research is needed to validate the models developed in this study in other places in Florida and in the southeastern United States.

ACKNOWLEDGMENTS

We would like to thank Chunhao Xu, Mi-Youn Ahn, and Yu Wang for the laboratory analyses, Wade Hurt for the taxonomic soil classification, Nicholas B. Comerford for support, and Carolyn Olsen for funding through the Cooperative Ecosystem Service Unit, NRCS, USDA. Funding was also provided by the large-scale project “Geo-temporal estimation and visualization of nitrogen and other soil properties in the mixed-use Santa Fe River Watershed,” Nutrient Science for Improved Watershed Management Program, USDA.

REFERENCES

Analytical Spectral Devices. 2003. QualitySpec Pro manual. ASD Doc. 600510, Rev. 1. ASD, Boulder, CO.

Analytical Spectral Devices. 2008. Product specifications: QualitySpec Pro. Available at www.asdi.com/products_specifications-QSP.asp (verified10 Nov. 2008). ASD, Boulder, CO.

Bellamy, P.H., P.J. Loveland, R.I. Bradley, R.M. Lark, and G.J.D. Kirk. 2005. Carbon losses from all soils across England and Wales 1978–2003. Nature 437:245–248.

Bouchard, V., and M. Cochran. 2006. Wetland and carbon sequestration. p. 1887–1890. *In* R. Lal (ed.) Encyclopedia of soil science. Vol. 2. CRC Press, Boca Raton, FL.

Breiman, L. 1996. Bagging predictors. Mach. Learn. 24:123–140.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classification and regression trees. Wadsworth Stat./Probab. Ser. Wadsworth Int., Belmont, CA.

Brown, D.J., R.S. Bricklemyer, and P.R. Miller. 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma 129:251–267.

Brown, D.J., K.D. Shepherd, M.G. Walsh, M. Dewayne Mays, and T.G. Reinsch. 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132:273–290.

Carbon Dioxide Information Analysis Center. 2008. Atmospheric CO_2 records from sites in the SIO air sampling network. Available at cdiac.esd.ornl.gov/trends/co2/sio-keel.html (verified 15 Nov. 2008). CDIAC, Oak Ridge, TN.

Chang, C., and D.A. Laird. 2002. Near-infrared reflectance spectroscopic analysis of soil C and N. Soil Sci. 167:110–116.

Chang, C., D.A. Laird, M.J. Mausbach, and C.R. Hurburgh, Jr. 2001. Near-infrared reflectance spectroscopy: Principal components regression analysis of soil properties. Soil Sci. Soc. Am. J. 65:480–490.

Coleman, K., and D.S. Jenkinson. 1996. RothC-26.3: A model for the turnover of carbon in soil. p. 237–246. *In* D.S. Powlson et al. (ed.) Evaluation of soil organic matter models using existing, long-term datasets. NATO ASI Ser. I, 38. Springer-Verlag, Heidelberg, Germany.

Dunn, B.W., H.G. Beecher, G.D. Batten, and S. Ciavarella. 2002. 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.

Fidêncio, P.H., R.J. Poppi, and J.C. Andrade. 2002. Determination of organic matter in soils using radial basis function networks and near infrared spectroscopy. Anal. Chim. Acta 453:125–134.

Gaffey, S.J., L.A. McFadden, D. Nash, and C.M. Pieters. 1993. Ultraviolet, visible, and near-infrared reflectance spectroscopy: Laboratory spectra of geologic materials. p. 43–77. *In* C.M. Pieters and P.E. Englert (ed.) Remote geochemical analysis: Elemental and mineralogical composition. Topics in Remote Sens. Ser. 4. Cambridge Univ. Press, Cambridge, UK.

Goddu, R.F., and D.A. Delker. 1960. Spectra–structure correlations for the near-infrared region. Anal. Chem. 32:140–141.

Gregorich, E.G., M.H. Beare, U. Stoklas, and P. St-Georges. 2003. Biodegradability of soluble organic matter in maize-cropping soils. Geoderma 113:237–252.

Grunwald, S. 2008. Role of soils to sequester carbon. *In* S. Mulkey et al. (ed.) Opportunities for greenhouse gas reduction through forestry and agriculture in Florida. School of Nat. Resour. and Environ., Univ. of Florida, Gainesville.

Guo, Y., R. Amundson, P. Gong, and Q. Yu. 2006. Quantity and spatial variability of soil carbon in the conterminous United States. Soil Sci. Soc. Am. J. 70:590–600.

Islam, K., B. Singh, and A. McBratney. 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. Aust. J. Soil Res. 41:1101–1114.

Keeling, C.D., T.P. Whorf, W. Wahlen, and J. van der Plicht. 1995. Interannual extremes in the rate of rise of atmospheric carbon dioxide since 1980. Nature 375:666–670.

Kooistra, L., J. Wanders, G.F. Epema, R.S.E.W. Leuven, R. Wehrens, and L.M.C. Buydens. 2003. The potential of field spectroscopy for the assessment of sediment properties in river floodplains. Anal. Chim. Acta 484:189–200.

Kooistra, L., R. Wehrens, R.S.E.W. Leuven, and L.M.C. Buydens. 2001. Possibilities of visible–near-infrared spectroscopy for the assessment of soil contamination in river floodplains. Anal. Chim. Acta 446:97–105.

Martens, H., and T. Næs. 1989. Multivariate calibration. John Wiley & Sons, Chichester, UK.

McCarty, G.W., J.B. Reeves III, V.B. Reeves, R.F. Follett, and J.M. Kimble. 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Sci. Soc. Am. J. 66:640–646.

McClure, W.F. 2003. 204 years of near infrared technology: 1800–2003. J. Near Infrared Spectrosc. 11:487–518.

McLauchlan, K.K., and S.E. Hobbie. 2004. Comparison of labile organic matter fractionation techniques. Soil Sci. Soc. Am. J. 68:1616–1625.

Parton, W.J., D.W. Anderson, C.V. Cole, and J.W.B. Stewart. 1983. Simulation of soil organic matter formation and mineralization in semiarid agroecosystems. p. 533–550. *In* R.R. Lowrance et al. (ed.) Nutrient cycling in agricultural ecosystems. Spec. Publ. 23. Univ. of Georgia Press, Athens.

Paul, E.A., S.J. Morris, and S. Bohm. 2001. Determination of soil C pool sizes and turnover rates: Biophysical fractionation and tracers. p. 193–206. *In* R. Lal et al. (ed.) Assessment methods for soil carbon. CRC Press, Boca Raton, FL.

Quideau, S.A. 2006. Organic matter accumulation. p. 1172–1175. *In* R. Lal (ed.) Encyclopedia of soil science. Vol. 2. CRC Press, Boca Raton, FL.

Rice, C.W. 2006. Organic matter and nutrient dynamics. p. 1180–1183. *In* R. Lal (ed.) Encyclopedia of soil science. Vol. 2. CRC Press, Boca Raton, FL.

Savitzky, A., and M.J.E. Golay. 1964. Smoothing and differentiation of data by simplified least-squares procedures. Anal. Chem. 36:1627–1639.

Shepherd, K.D., and M.G. Walsh. 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66:988–998.

Siesler, H.W., Y. Ozaki, S. Kawata, and H.M. Heise (ed.). 2002. Near-infrared spectroscopy: Principles, instruments, applications. Wiley-VCH, Weinheim, Germany.

Sparling, G., M. Vojvodić-Vuković, and L.A. Schipper. 1998. Hot-water-soluble C as a simple measure of labile soil organic matter: The relationship with microbial biomass C. Soil Biol. Biochem. 30:1469–1472.

Vasques, G.M., S. Grunwald, and J.O. Sickman. 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146:14–25.

Williams, P.C. 1987. Variables affecting near-infrared reflectance spectroscopic analysis. p. 143–167. *In* P. Williams and K. Norris (ed.) Near-infrared technology in the agricultural and food industries. Am. Assoc. of Cereal Chemists, St. Paul, MN.