Prediction of Soil Fertility Properties from a Globally Distributed Soil Mid-Infrared Spectral Library

Thomas Terhoeven-Urselmans*
Tor-Gunnar Vagen
World Agroforestry Center (ICRAF)
PO Box 30677
00100 Nairobi, Kenya

Otto Spaargaren
International Soil Reference and Information Centre (ISRIC)
World Soil Information
Akkermaalsbos 12
6708 WB Wageningen, the Netherlands

Keith D. Shepherd
World Agroforestry Center (ICRAF)
PO Box 30677
00100 Nairobi, Kenya

Globally applicable calibrations to predict standard soil properties based on infrared spectra may increase the use of this reliable technique. The objective of this study was to evaluate the ability of mid-infrared diffuse reflectance spectroscopy (4000–602 cm^{-1}) to predict chemical and textural properties for a globally distributed soil spectral library. We scanned 971 soil samples selected from the International Soil Reference and Information Centre database. A high-throughput diffuse reflectance accessory was used with optics that exclude specular reflectance as a potential source of error. Archived data on soil chemical and physical properties were calibrated to first derivative spectra using partial least-squares regression. Good predictions for the spatially independent validation set were achieved for pH value, organic C content, and cation exchange capacity (CEC) (n = 291, r^2 of linear regression of predicted against measured values $\geq$0.75 and ratio of standard deviation of measured values to root mean square error of prediction (RPD) $\geq$2.0). The root mean square errors of prediction (RMSEP) were 0.75 pH units, 9.1 g organic C kg^{-1} and 5.5 $cmol_c$ CEC kg^{-1}. Satisfactory predictions (r^2 = 0.65–0.75, RPD = 1.4–2.0) were obtained for exchangeable Mg concentration and clay content. The respective RMSEPs were 4.3 $cmol_c$ kg^{-1} and 126 g kg^{-1}. Poorer predictions (r^2 = 0.61 and 0.64) were achieved for sand and exchangeable Ca contents. Although RMSEP values are large relative to laboratory analytical errors, our results suggest a marked potential for the global spectral library as a tool for advice on land management, such as the classification of new samples into basic soil fertility classes based on organic C and clay contents, CEC, and pH. Further research is needed to test the stability of this global calibration on new data sets.

Abbreviations: CEC, cation exchange capacity; MIR, mid-infrared; NIR, near-infrared; PRESS, prediction error sum of squares; RMSEC, root mean square error of calibration; RMSEP, root mean square error of prediction; RPD, ratio of performance deviation; VIP, variable importance of projection.

Soil chemical and physical information is needed to give advice on land management. This is especially true in developing countries, where soil diagnostic surveillance systems have been proposed to overcome data shortages (Shepherd and Walsh, 2007). Mid-infrared (MIR) diffuse reflectance spectroscopy is a reliable and fast soil analytical tool (Janik et al., 1998) that could form a basis for diagnostic surveillance systems. Soil properties are predicted either by direct absorption of the light associated with functional groups (properties such as organic C, total N, or clay composition; Van der Marel and Beutelspacher, 1976) or by correlation to such properties and the mineral composition of the soil (properties such as cation exchange capacity [CEC] and soil texture). New samples can be predicted only if they fall within the property range of the calibration set (Naes et al., 2002). In many situations, a rapid and approximate estimate of soil chemical and physical properties is adequate, and resources for an elaborate analysis may not be available. A global calibration may meet this purpose.

Some studies have tested soil infrared spectroscopy on diverse data sets at the regional scale. Reeves and Smith (2009), working with a North American library of 720 samples, came to the conclusion that neither MIR nor near-infrared (NIR) spectra yielded suitable calibrations even for organic C. They attributed the poor performance to the extreme sample diversity in parent material, land

Soil Sci. Soc. Am. J. 74:1792–1799
Published online 9 Aug. 2010
doi:10.2136/sssaj2009.0218
Received 10 Jun. 2009.
*Corresponding author (t.urselmans@cgiar.org)

use, and climate in their data set. Shepherd and Walsh (2002) obtained reasonable calibrations for key soil fertility properties for a range of African soils using visible–NIR spectra. Brown et al. (2006), using a sample set of about 4000 visible–NIR spectra of U.S. soils, which included 416 samples from 36 different countries outside the United States, obtained good predictions for organic C content, CEC, and clay content (r^2 = 0.83–0.91). Madari et al. (2005), using 1135 soil samples from the Brazilian National Soil Collection, obtained good predictions for total C using MIR and NIR spectra (r^2 = 0.95 and 0.93, respectively). Their predictions for MIR spectra in most cases outperformed the predictions for NIR spectra. However, MIR diffuse reflectance spectroscopy has, to our knowledge, not yet been applied to a spectral library with coverage beyond a regional scale. Thus, the objective of this study was to evaluate the ability of MIR diffuse reflectance spectroscopy to predict chemical and textural properties for a globally distributed soil set.

MATERIALS AND METHODS

Soil Characterization

The library soils are part of the global soil collection from the International Soil Reference and Information Centre (2009). The spectral library consists of 4438 soil samples from 754 soil profiles (56 countries); however, only those samples with complete data on soil fertility variables were used in this analysis (n = 971). The library includes 219 soil profiles, with between 1 and 11 soil layers. The average depth of the soil layers ranges from 1 to 635 cm with a median of 51 cm and mean of 64 cm (e.g., a soil layer from 0–7 cm with an average depth of 3.5 cm). The library includes soils from 18 countries (Fig. 1): Australia (five profiles, 28 samples), Botswana (seven profiles, 36 samples), Brazil (25 profiles, 117 samples), Cameroon (one profile, 7 samples), China (43 profiles, 163 samples), Colombia (11 profiles, 42 samples), Costa Rica (11 profiles, 41 samples), Cote d'Ivoire (seven profiles, 17 samples), Cuba (20 profiles, 102 samples), Ecuador (20 profiles, 90 samples), Finland (five profiles, 15 samples), France (nine profiles, 27 samples), Gabon (six profiles, 11 samples), Germany (15 profiles, 68 samples), Ghana (two profiles, 18 samples), Greece (seven profiles, 35 samples), Hungary (17 profiles, 108 samples), and Spain (eight profiles, 46 samples).

The samples represent a wide variety of landscape positions, parent materials, and land uses. The landscape positions were: no record (247), flat (246), middle slope (171), upper slope (143), lower slope (67), crest (35), depression (25), slope (24), and open depression (13). The parent materials were: no record (378), limestone (64), calcareous (41), fine-grained basic igneous (40), sedimentary rock (33), mixed lithology and composition (31), claystone (28), igneous rock (27), sandstone (27), unconsolidated pyroclastic rocks (26), coarse-grained acid igneous (25), fine-grained intermediate igneous (24), ejecta ash (20), highly weathered material (19), basic gneiss (17), igneous and metamorphic (16), metamorphic rock (16), mixed lithology (16), acidic ash (15), acidic gneiss (14), basalt (12), gneiss (12), conglomerate (10), coarse-grained basic igneous (8), interbedded sedimentary rock (8), sandstone–shale (8), pyroclastic, consolidated (5), acidic tuff (4), fine-grained ultrabasic igneous (4), shale (4), shale–siltstone (4), siltstone (4), arkosic sandstone (2), dolomitic limestone (2), limestone–shale (2), marl (2), noncalcareous sandstone (2), and metamorphic–sedimentary (1). The land use of the samples was: arable farming (461), seminatural grassland, grazed (150), no record (96), (semi-) natural vegetation (91), cultivated pasture (45), mixed farming (53), afforestation (34), woodland, grazed (21), shrubland, grazed (14), fallow (3), and nonagricultural land (3). Soil types were classified according to the U.S. Soil Taxonomy. The soils belong to the orders Inceptisol (253), Alfisol (181), Mollisol (113), Entisol (80), Vertisol (78), Ultisol (71), Oxisol (67), Spodosol (5), Aridisol (4), and without classification (119).

Soil property analysis was done according to Van Reeuwijk (2002). Soil samples were air dried, the clods crushed, and the resulting sample material sieved through a 2-mm sieve before further analysis. The soil pH

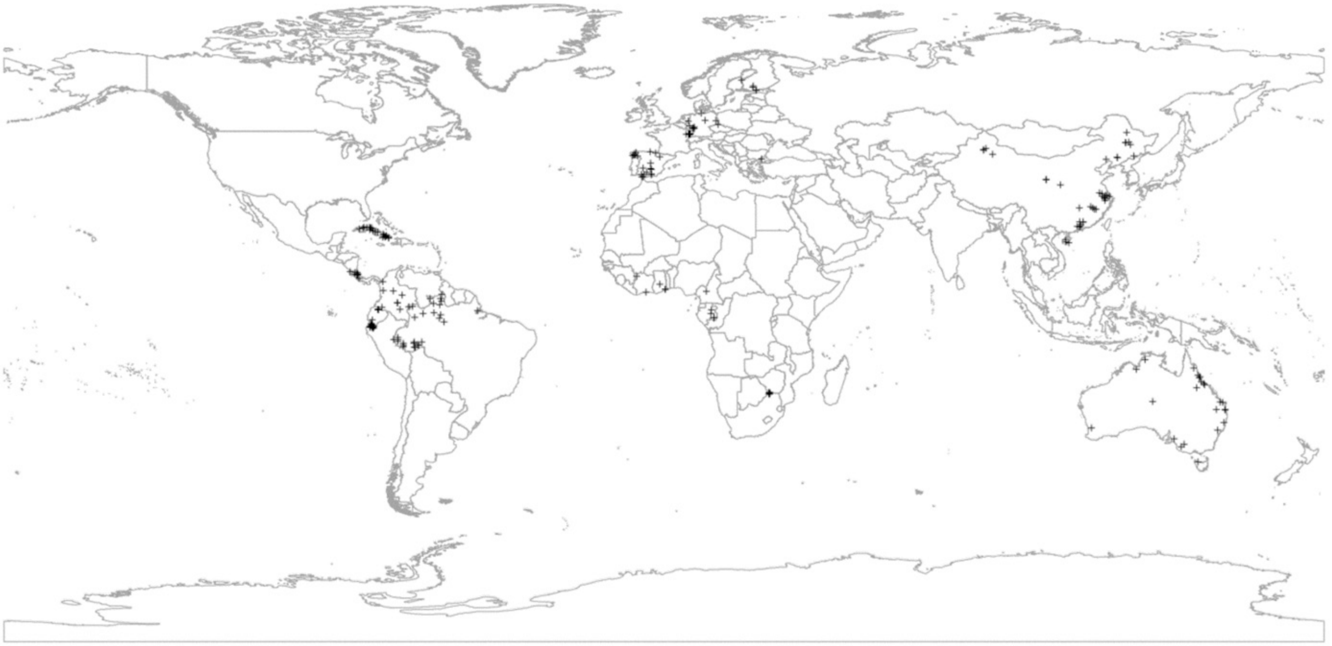

Fig. 1. Global distribution of the soil profiles (n = 219) in the sample library.

was determined by shaking soil together with deionized water for 2 h at a soil/liquid ratio of 1:2.5. Organic C content was determined using the Walkley–Black procedure (Heanes, 1984). This involves a wet combustion of the organic matter with a mixture of potassium dichromate and H_2SO_4 at about 125°C. The CEC and exchangeable Ca and Mg were determined using the NH_4OAc method, which included the use of NaOAc instead of NH_4OAc for samples with a pH value >7. Exchangeable Ca and Mg were measured by flame atomic absorption spectrophotometry and CEC by flame emission spectrophotometry. Soil organic matter and carbonates were removed before soil textural analysis using H_2O_2 and HCl, respectively. The sand fraction was separated first by wet sieving. Clay fractions were determined by the pipette method.

There are several sources of possible error associated with soil sample preparation and analysis: (i) samples for spectral measurement were resampled from the original uncrushed field samples and then crushed to pass a 2-mm sieve, which could give rise to subsampling errors, (ii) wet chemical analyses were done in different batches (one to two profiles per batch) between 1991 and 2003, (iii) there was a time lag of 2 to 14 yr between wet chemical analysis and spectral acquisition, which could lead to sample aging effects, (iv) a wetting procedure and redrying was done during subsampling, and (v) the Walkley–Black method underestimates the soil organic C content, compared with dry combustion, to a different degree depending on soil type (Moody and Cong, 2008).

Spectral Measurements

Soil MIR diffuse reflectance spectra were recorded for all samples in the year 2005 using a Fourier-transform MIR spectrometer (FT-IR; Tensor 27 with high-throughput screening extension unit with robotic arm [Twister Microplate Handler], Bruker Optics, Karlsruhe, Germany; illustrated in Shepherd and Walsh, 2007). The detector was a liquid N_2–cooled HgCdTe detector. The measured wavebands ranged from 4000 to 602 cm^{-1} with a resolution of 4 cm^{-1} and zero filling of 2, which resulted in 1763 data points at a waveband distance of about 2 cm^{-1}. A special feature of the instrument optics is that specular reflectance, which can distort the shape of MIR spectra strongly (Reeves et al., 2006), is shielded. This is, to our knowledge, the only optical setup that combines a high-throughput measurement (1000 samples d^{-1}) with the exclusion of specular reflectance.

Air-dried samples were finely ground to powder (approximately <100 μm), using an agate pestle and mortar. The samples were loaded into Al microtiter plates (A752-96, Bruker Optics, Karlsruhe) using a microspatula to fill the 6-mm-diameter wells and level the soil, taking care to avoid spillage into neighboring wells. Background measurements of the first empty well were taken before each single measurement to account for changes in temperature and air humidity. Aluminum is suitable as a reference material because it does not absorb infrared light. The bottoms of the Al wells are roughened to minimize specular reflectance. Soil samples were loaded into four replicate wells, each scanned 32 times, and the four spectra averaged to account for within-sample variability and differences in particle size and packing density.

Statistical Methods

All calculations and statistical analysis were done using R software version 2.7.1 (R Development Core Team, 2008). Box–Cox transformation (Box and Cox, 1964) was applied before statistical analysis of the soil properties (except pH) to obtain approximately normally distributed values, using the *box.cox.powers* and *bc* functions in the *car* package (Fox, 2009). The *locpoly* function in the *KernSmooth* package was applied for computing spectral derivatives based on Savitzky–Golay smoothing filters (Wand and Ripley, 2008). Principal components analysis was done using the *prcomp* function in the *stats* package (R Development Core Team, 2008), and partial least-squares regression was conducted using the *mvr* function in the *pls* package (Mevik and Wehrens, 2007). The relationships among soil properties were analyzed using the correlation coefficient (r) for pairwise associations.

Absorption spectra were preprocessed using a first-derivative transformation with a smoothing interval of 21 data points. Selection of calibration and validation samples was done following a procedure adapted from Kennard and Stone (1969) applied to the score values of the first eight principal components of the first-derivative MIR spectra. The function was written in R. The adaptations were (i) that no samples from the same soil profiles were allowed to be split between the calibration and validation sets, and (ii) that the two extreme score values of each principal component were chosen for the calibration set and the next extreme score values chosen as starting samples for the validation set. To avoid the risk of spatial correlation between calibration and validation sets, all the samples from a single profile were assigned to either the calibration or the validation set. The validation samples were chosen in a stepwise procedure by maximizing the Euclidean distances to the objects already included in the validation set until the number of validation samples was reached. The number of principal components (eight) was selected so that the increase in cumulative explained variance within the next three components was <4% (Table 1). About 70% of the MIR spectra (n = 679) of the whole sample set was chosen as the calibration set and the remaining about 30% (n = 292) as the validation set.

Partial least-squares regression was performed on the calibration set with transformed reference values (except pH) and first-derivative MIR spectra (we combined the used functions and others in an R package called *soil.spec*, which is available at www.cran.r-project.org/ under the Packages link [verified 26 June 2010]). Cross-validation (10-fold) was done to determine the optimal number of principal components for

Table 1. Cumulative proportion of explained variance of the first 12 principal components (PCs) for first derivative mid-infrared (MIR) spectra. First derivative spectra were zero centered and scaled to have unit variance prior to principal component analysis.

Principal component	MIR
PC 1	0.24
PC 2	0.43
PC 3	0.58
PC 4	0.69
PC 5	0.75
PC 6	0.80
PC 7	0.84
PC 8	0.87
PC 9	0.88
PC 10	0.90
PC 11	0.91

Table 2. Descriptive statistics of the soil properties (*n* = 971).

Soil property	Percentile					SD
	0th	25th	50th	75th	100th	
Organic C content, g kg^{-1}	0.2	2.2	4.9	13.0	159	17.1
pH	3.0	5.5	6.6	8.1	10.5	1.5
Exchangeable Ca content, cmol$_c$ kg^{-1}	0.2	2.4	9.7	24.5	102	17.5
Exchangeable Mg content, cmol$_c$ kg^{-1}	0.1	0.7	1.8	4.9	68.0	5.8
Cation exchange capacity, cmol$_c$ kg^{-1}	0.3	7.8	13.8	23.8	77.6	14.7
Clay content, g kg^{-1}	2.0	172	306	473	918	210
Sand content, g kg^{-1}	1.0	74.5	235	467	985	260

the first 20 principal components. The prediction error sum of squares (PRESS) was calculated as

$$\mathrm{PRESS}=\sum_{i=1}^{N}\left(y_i-x_i\right)^2 \qquad [1]$$

where *y* is the predicted reference value, *x* is the measured reference value, and *N* is the number of samples. The smallest number of principal components was chosen such that the ratio of PRESS divided by the minimum PRESS value was not significantly >1 (the *F* distribution was chosen for significance determination with α = 0.975; Haaland and Thomas, 1988). No spectra were removed as calibration outliers (the removal of samples having a Mahalanobis distance >12 did not improve the prediction performance of the validation set). The derived model was applied to the validation set for predictions. Predicted values (of the calibration and validation sets) were back-transformed before linear regression of the predicted against measured values. Predicted and back-transformed values for sand content >1000 g kg^{-1} were set to 1000 g kg^{-1} (23 samples), the theoretical maximum. One predicted value in the calibration set and two in the validation set for clay content were excluded from further analysis because they could not be back-transformed due to the inability of power transformations such as Box–Cox to compute for negative values below the original range. The prediction performance was evaluated using the coefficient of determination (r^2) of the linear regression of predicted against measured values, the root mean square errors of calibration (RMSEC):

$$\mathrm{RMSEC}=\sqrt{\frac{\sum_{i=1}^{N}\left(y_i-x_i\right)^2}{N-A-1}} \qquad [2]$$

where *A* is the number of principal components used in the model, and the root mean square errors of prediction (RMSEP):

$$\mathrm{RMSEP}=\sqrt{\frac{\sum_{i=1}^{N}\left(y_i-x_i\right)^2}{N}} \qquad [3]$$

for the calibration and validation set, respectively, and the ratio of performance deviation (RPD), which is the ratio of the standard deviation of the measured soil properties and the RMSEC or RMSEP. Good predictions for such a diverse data set are regarded as having an $r^2 \geq 0.75$ and an RPD ≥ 2 (Shepherd and Walsh, 2002; Chang et al., 2001). Satisfactory predictions have an r^2 from 0.65 to 0.75 and RPD from 1.4 to 2.0. Predictions below those values are considered to be poor, although this interpretation depends on the objective. In addition, the bias was calculated as

$$\mathrm{bias}=\frac{\sum_{i=1}^{N}\left(y_i-x_i\right)}{N} \qquad [4]$$

Identification of important wavebands was done as follows: the combination of partial least-squares regression (PLSR) coefficients **b** and the variable importance of projection (VIP) was used (Viscarra Rossel et al., 2008). The VIP was calculated by

$$\mathrm{VIP}_k(A)=K\sum_{A}w_{Ak}^2\left(\frac{\mathrm{SSY}_A}{\mathrm{SSY}_t}\right) \qquad [5]$$

where $\mathrm{VIP}_k(A)$ is the importance of the *k*th predictor variable based on a model with *A* factors, w_{Ak} is the corresponding loading weight of the *k*th variable in the *A*th PLSR factor, SSY_A is the explained sum of squares of the predicted *y* by a PLSR model with *A* factors, SSY_t is the total sum of squares of *y*, and *K* is the total number of predictor variables. Thresholds were introduced for the determination of important wavebands. The thresholds were 1 for the VIP and one SD of **b** for **b**. A waveband was considered to be important when both thresholds were satisfied.

RESULTS AND DISCUSSION

Global Spectral Library Assessment

The selection of sample sites for building a global soil spectral library is not trivial and is strongly dependent on the objective. For the objective of assessing soil suitability for agricultural production, it is important that the site selection captures as much variation as possible in (i) soil properties, (ii) climatic zones, and (iii) mineral composition. The selection of calibration and validation sets should ensure that (i) calibration samples are evenly distributed in spectral space, and (ii) spatially correlated samples are separated to avoid pseudo-replication.

The sample library satisfied many of the above criteria for achieving a global soil library. There was wide, globally distributed coverage of agricultural regions, but samples were concentrated in Latin America, China, Europe, and western Africa (Fig. 1) and there were no samples from North America, North- and East Africa, and Southeast Asia. Nevertheless, the soil properties spanned a wide range in terms of variability and should be representative of agricultural soils in these regions (Table 2).

Figure 2 shows 10 raw spectra, selected to illustrate the variability in the spectral library. There is considerable variation in absorbance across the full spectral range. In general terms, MIR spectra can be divided into four regions (e.g., Shepherd and Walsh, 2007): (i) fingerprint (e.g., O–Si–O stretching and bending) from 1500 to 600 cm^{-1}, (ii) double bond (e.g., C=O, C=C, and C=N) from 2000 to 1500 cm^{-1}, (iii) triple bond (e.g., C≡C, C≡N) from 2500 to 2000 cm^{-1}, and (iv) *X*–H stretching (e.g., O–H stretching) from 4000 to 2500 cm^{-1}. The strong absorbance in the 3600 to 3800 cm^{-1} region is due to hydroxyl stretching vibrations associated with clay minerals. Carbonates produce absorption at 2600 to 2500 cm^{-1} with little interference from other minerals (Nguyen et al., 1991), clearly visible in one of the example spectra in Fig. 2. Soil organic matter produces features across the entire spectral range, for example contributing to the broad absorption features near 3400, 1600, and 1400 cm^{-1} and due to absorption by aromatic structures, alkyls, carbohydrates, carboxylic acid, cellulose, lignin, C=C skeletal structures,

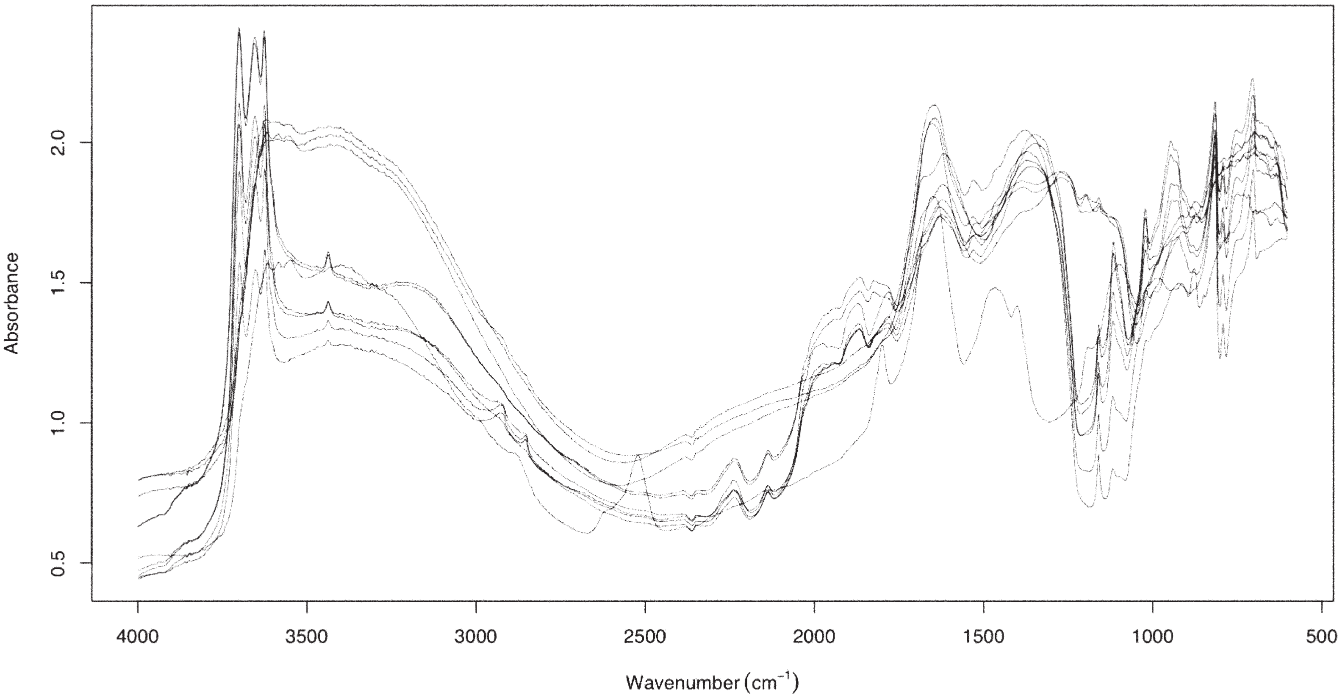

Fig. 2. Ten selected raw spectra that cover the spectral variability of the sample library.

ketones, and phenolics (Janik et al., 2007). Principal component analysis of the first-derivative spectra demonstrates the very complex structure of the spectral library: the first five principal components accounted for only 75% of the explained variance (Table 1). We had ensured that the validation sets were spatially independent from the calibration sets by assigning 157 out of the 219 soil profiles to the calibration set and the remaining 62 profiles to the validation set.

Predictions of Soil Properties from Spectra

Soil organic C was predicted well for the validation set (r^2 = 0.77 and RPD = 2.0, Table 3; Fig. 3). The RMSEP was 9.1 g kg^{-1} organic C. Viscarra Rossel et al. (2006) reported similar accuracy (r^2 = 0.73) for a much less diverse validation set of 118 samples from an 18-ha agricultural field in Australia. For a global set of visible–NIR spectra using boosted regression trees, Brown et al. (2006) reported for organic C an RMSD (which is equivalent to our RMSEP) of 9.0 g kg^{-1}, compared with our RMSEP value of 9.1 g kg^{-1}. It is difficult, however to compare the prediction performance of visible–NIR vs. MIR across data sets due to differences in data structure and errors associated with laboratory reference values. Better predictions might be possible for our data set if it was more balanced in the range 50 to 150 g kg^{-1} organic C. Important wavebands for organic C predictions occurred across the full spectral range, from 746 to 663, 1047 to 1041, 1139 to 1105, 1274 to 1265, 1573 to 1560, 1820 to 1781, 2547 to 2524, 2624 to 2611, 2985 to 2923, 3370 to 3353, 3683 to 3666, and 3783 to 3733 cm^{-1} (Fig. 4).

Soil pH was predicted well (r^2 = 0.80 and RPD = 2.21; Table 3; Fig. 3). The RMSEP was 0.75 pH units. This result is as good as that obtained by McCarty and Reeves (2006) for MIR analysis of 544 soil samples (272 locations) in one field (r^2 = 0.8). Our result is slightly better than that obtained by Janik and Skjemstad (1995), who reported an r^2 value of 0.72 for 291 Australian soils. They were able to increase the prediction accuracy up to an r^2 value of 0.85 by using the locally linear approximation method,

Table 3. Calibration (*n* = 679) and validation (*n* = 292) statistics of soil properties for first derivative mid-infrared spectra using partial least-squares regression. The coefficient of determination (r^2), bias, root mean square errors of calibration and prediction (RMSEC and RMSEP, respectively) and the ratio of measured property SD to RMSEC or RMSEP (RPD) are given for back-transformed data. No outliers were excluded.

Data set	Soil property	Number of principal components	r^2	bias	RMSEC or RMSEP	RPD
Calibration	organic C content, g kg^{-1}	8	0.77	0.00	9.40	1.75
	pH	13	0.81	0.00	0.63	2.29
	exchangeable Ca content, cmol$_c$ kg^{-1}	11	0.78	−0.95	8.52	2.11
	exchangeable Mg content, cmol$_c$ kg^{-1}	10	0.54	−0.46	3.63	1.42
	cation exchange capacity, cmol$_c$ kg^{-1}	9	0.83	−0.59	6.40	2.38
	clay content, g kg^{-1}	9	0.83	−5.00	82.3	2.40
	sand content, g kg^{-1}	15	0.72	−25.4	138.7	1.81
Validation	organic C content, g kg^{-1}		0.77	−1.50	9.10	2.04
	pH		0.80	0.10	0.75	2.21
	exchangeable Ca content, cmol$_c$ kg^{-1}		0.61	−0.72	10.2	1.55
	exchangeable Mg content, cmol$_c$ kg^{-1}		0.68	−1.03	4.33	1.64
	cation exchange capacity, cmol$_c$ kg^{-1}		0.82	−0.53	5.53	2.32
	clay content, g kg^{-1}		0.73	−33.6	126.2	1.85
	sand content, g kg^{-1}		0.64	−3.40	174.3	1.59

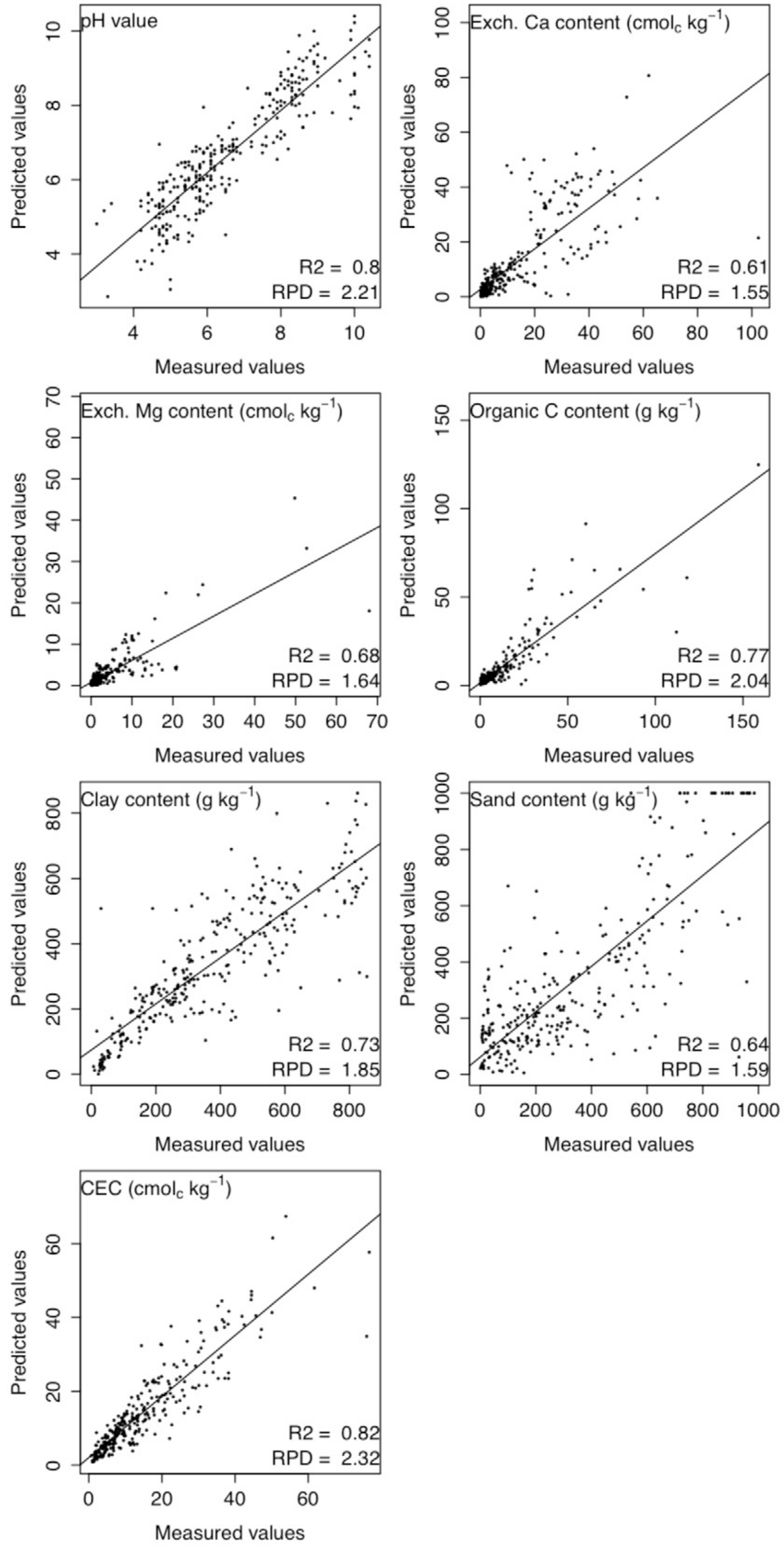

Fig. 3. Linear regressions for the validation set (*n* = 292) of predicted against measured soil property values (CEC, cation exchange capacity; R^2, coefficient of determination; RPD, ratio of standard deviation of soil measured reference values to root mean square error of prediction).

which breaks up the whole range into three subsets to deal with nonlinearity. Our model underestimated pH values <4 and especially >10 (Fig. 3); however, the samples with a pH > 10 belonged to only two profiles in Hungary and were therefore not well represented in the calibration. The result is surprisingly good considering that pH is indirectly predicted from soil organic matter and mineralogy. Correlation analysis supported this (Table 4): pH values in the calibration set were strongly correlated with exchangeable Ca and Mg (r = 0.74 and 0.54, respectively) although not with CEC or organic C (r = 0.28 and −0.21, respectively; Table 4). Important wavebands were 651 to 649, 835 to 798, 948 to 933, 1178 to 1151, 1398 to 1334, 1400, 1571 to 1554, 1604 to 1602, 1743 to 1706, 2534 to 2532, 3675 to 3660, 3694, 3725 to 3721, and 3783 to 3775 cm^{-1} (Fig. 4). These bands are mainly concentrated in the parts of the spectrum that contain mineral features: the fingerprint and *X*–H stretching regions. The quite stable predictions for this diverse data set are promising for obtaining good predictions for unknown samples.

Validation predictions were satisfactory for exchangeable Mg but were poorer for exchangeable Ca (r^2 = 0.68 and 0.61 and RPD = 1.64 and 1.55, respectively; Table 3), whereas CEC was predicted well (r^2 = 0.82 and RPD = 2.32). Important wavebands for CEC were 605 to 601, 869 to 836, 1004 to 946, 1126 to 1039, 1193 to 1149, 1245 to 1207, 1619 to 1594, 1754 to 1722, 3516 to 3507, and 3637 to 3550 cm^{-1}. Exchangeable Ca predictions were satisfactory based on the RPD, but the prediction values showed high scatter at high values. The calibration may be useful for predicting low exchangeable Ca levels. Exchangeable Ca extracted at natural soil pH usually predicts well from spectral data in tropical soils (Shepherd and Walsh, 2002), and the poorer prediction obtained here may be due to the extraction method, which used a different extractant above than below soil pH 7, and the inclusion of calcareous soils. Pirie et al. (2005) showed similar prediction accuracy for these three soil properties for 415 southeastern Australian samples. It is surprising, however, that their predictions were not stronger, because they allowed the possibility of spatial correlation when choosing their calibration and validation sets; they had up to 25 samples from one location. Our models for exchangeable Ca and Mg were not as stable as those for CEC due to some samples having high property values (Fig. 3). Different runs of the cross-validation procedure led to a varying number of partial least-squares components in the range of one to three.

Predictions for particle size were satisfactory for clay (r^2 = 0.73 and RPD = 1.85) but poorer for sand (r^2 = 0.64). Our results are broadly similar to those of previous researchers (McCarty and Reeves, 2006; Pirie et al., 2005). The model for sand content tended to overestimate. Predicted values in the calibration and validation sets were >1000 g kg^{-1} for 23 sand samples, which were then set to 1000 g kg^{-1} (Fig. 3). This occurred only for soils comprising >540 g kg^{-1} sand. Mid-infrared responds to quartz in the fingerprint region of the spectrum and we would expect that sand predictions would perform better after exclusion of specular reflection. Unlike the clay calibration, however, where the important wavebands were concentrated in the mineral regions of the spectrum, the important wavebands for sand were distributed across the entire spectrum, suggesting that prediction was indirect.

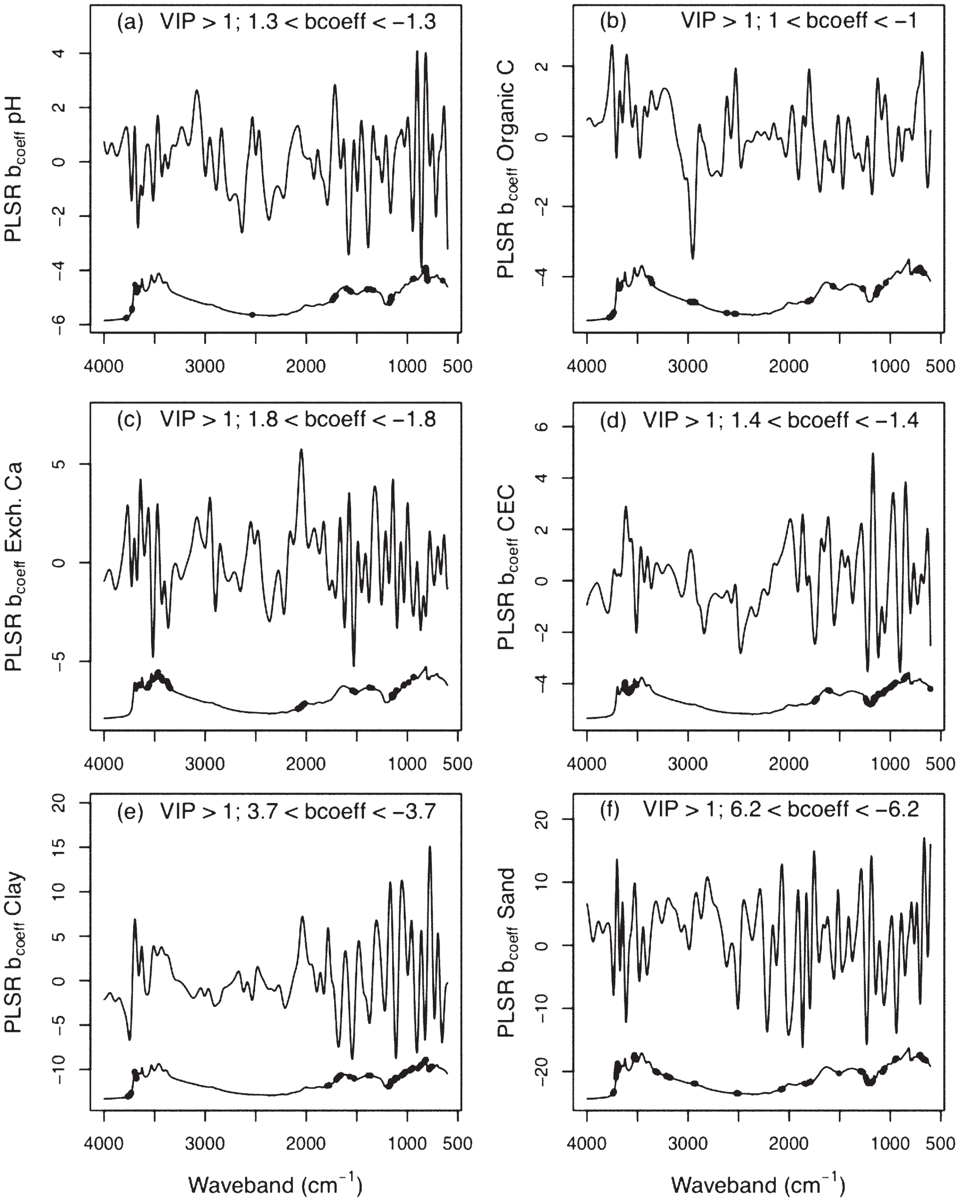

Fig. 4. Partial least-squares regression (PLSR) coefficients b and spectra with important wavebands highlighted. The wavebands were identified by a combined use of b and the variable importance for projection (VIP). Representations are given for (a) pH, (b) organic C content, (c) exchangeable Ca content, (d) cation exchange capacity (CEC), (e) clay content, and (f) sand content. Thresholds for b and VIP are given for each soil property.

CONCLUSIONS

We have shown that basic soil quality variables can be predicted from MIR spectra for globally diverse soils. Prediction accuracies for soil organic C content, CEC, and pH were 9.1 g kg^{-1}, 5.53 $cmol_c$ kg^{-1} and 0.75 pH units, respectively. The prediction accuracy is comparable to the one published study for a global soil spectral library based on visible–NIR spectra. As expected, the average prediction performance was lower than that reported for many local or regional studies. The prediction performance is good enough for several applications, however, and in data-sparse situations that are common in Africa. For example, the calibration equations could be used to classify unknown samples into basic soil fertility classes based on soil organic C content, clay content, pH, and CEC. Because these are empirical calibrations and depend on associations among soil properties, however, predictions of unknown samples should always be treated with caution and local calibrations will usually perform better. For new sample sets, including

Table 4. Correlation coefficients *r* for Box–Cox transformed soil properties (except pH) of the whole data set (*n* = 971, upper triangle) and of the calibration set (*n* = 679, lower triangle).

Property	Organic C	pH	Exchangeable Ca	Exchangeable Mg	CEC†	Clay	Sand
Organic C	–	−0.21	−0.03	−0.07	0.32	0.06	−0.11
pH	−0.21	–	0.77	0.58	0.23	−0.10	−0.14
Exchangeable Ca	−0.03	0.74	–	0.75	0.54	0.09	−0.31
Exchangeable Mg	−0.10	0.54	0.76	–	0.61	0.21	−0.27
CEC†	0.24	0.28	0.60	0.65	–	0.44	−0.44
Clay	0.03	−0.03	0.20	0.32	0.46	–	−0.58
Sand	−0.05	−0.20	−0.35	−0.28	−0.40	−0.51	–

† Cation exchange capacity.

diverse local samples in the global calibration is expected to increase the prediction accuracy and confidence in the results. Calibrations might have been stronger if soil reference analyses had been done in one laboratory in a short period of time and at the same time as the spectral measurements. Future work should expand global soil spectral calibration libraries using a small number of well-coordinated reference laboratories to reduce sources of variation in reference data or find ways to build these errors into the calibrations. The spectrometer system we used allowed us to combine the exclusion of specular reflectance with a powerful high-throughput scanning unit.

ACKNOWLEDGMENTS

We thank Dr. Ir. L.P. van Reeuwijk for his valuable comments on the manuscript. Moreover, we thank Andrew Sila for database management and Elvis Weullow and Dickens Ateku for technical support. Funded by the Federal Ministry for Economic Cooperation and Development, Germany.

REFERENCES

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc., Ser. B 26:211–246.

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

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

Fox, J. 2009. car: Companion to applied regression. R package version 1.2–8. Available at cran.r-project.org/web/packages/car/index.html (verified 20 June 2010). Inst. for Stat. and Math., Wirtschafts Univ., Vienna.

Haaland, D.M., and E.V. Thomas. 1988. Partial least-squares methods for spectral analysis: 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 60:1193–1202.

Heanes, D.L. 1984. Determination of total organic-C in soils by an improved chromic acid digestion and spectrophotometric procedure. Commun. Soil Sci. Plant Anal. 15:1191–1213.

International Soil Reference and Information Centre. 2009. ISRIC soil information system (ISIS). Available at www.isric.org/UK/About+Soils/Soil+data/Geographic+data/Global/ISIS.htm (verified 20 June 2010). ISRIC, Wageningen, the Netherlands.

Janik, L.J., R.H. Merry, and J.O. Skjemstad. 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? Aust. J. Exp. Agric. 38:681–696.

Janik, L.J., and J.O. Skjemstad. 1995. Characterization and analysis of soils using mid-infrared partial least-squares: II. Correlations with some laboratory data. Aust. J. Soil Res. 33:637–650.

Janik, L.J., J.O. Skemstad, K.D. Shepherd, and L.R. Spouncer. 2007. The prediction of soil carbon fractions using mid-infrared-partial least square analysis. Aust. J. Soil Res. 45:73–81.

Kennard, R., and L. Stone. 1969. Computer aided design of experiments. Technometrics 11:137–148.

Madari, B.E., J.B. Reeves, M.R. Coelho, P.L.O.A. Machado, and H. De Polli. 2005. Mid- and near-infrared spectroscopic determination of carbon in a diverse set of soils from the Brazilian national soil collection. Spectrosc. Lett. 38:721–740.

McCarty, G.W., and I.J.B. Reeves. 2006. Comparison of near infrared and mid infrared diffuse reflectance spectroscopy for field-scale measurement of soil fertility parameters. Soil Sci. 171:94–102.

Mevik, B.H., and R. Wehrens. 2007. The pls package: Principal component and partial least squares regression in R. J. Stat. Softw. 18:1–24.

Moody, P.W., and P.T. Cong. 2008. Soil constraints and management package (SCAMP): Guidelines for sustainable management of tropical upland soils. ACIAR Monogr. 130. Aust. Ctr. for Int. Agric. Res., Canberra, ACT, Australia.

Naes, T., T. Isaksson, T. Fearn, and T. Davies. 2002. A user-friendly guide to multivariate calibration and classification. NIR Publ., Chichester, UK.

Nguyen, T.T., L.J. Janik, and M. Raupach. 1991. Diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy in soil studies. Aust. J. Soil Res. 29:49–67.

Pirie, A., S. Balwant, and I. Kamrunnahar. 2005. Ultra-violet, visible, near-infrared, and mid-infrared diffuse reflectance spectroscopic techniques to predict several soil properties. Aust. J. Soil Res. 43:713–721.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Reeves, J.B., R.F. Follett, G.W. McCarty, and J.M. Kimble. 2006. Can near or mid-infrared diffuse reflectance spectroscopy be used to determine soil carbon pools? Commun. Soil Sci. Plant Anal. 37:2307–2325.

Reeves, J.B., and D.B. Smith. 2009. The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America. Appl. Geochem. 24:1472–1481.

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.

Shepherd, K.D., and M.G. Walsh. 2007. Infrared spectroscopy: Enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. J. Near Infrared Spectrosc. 15:1–19.

Van der Marel, H., and H. Beutelspacher. 1976. Atlas of infrared spectroscopy of clay minerals and their admixtures. Elsevier, Amsterdam.

Van Reeuwijk, L.P. 2002. Procedures for soil analysis. 6th ed. Tech. Pap. 9. Int. Soil Ref. and Inf. Ctr., Wageningen, the Netherlands.

Viscarra Rossel, R.A., Y.S. Jeon, I.O.A. Odeh, and A.B. McBratney. 2008. Using a legacy soil sample to develop a mid-IR spectral library. Aust. J. Soil Res. 46:1–16.

Viscarra Rossel, R.A., D.J.J. Walvoort, A.B. McBratney, L.J. Janik, and J.O. Skemstad. 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131:59–75.

Wand, M., and B. Ripley. 2008. KernSmooth: Functions for kernel smoothing for Wand & Jones (1995). R package version 2.22–22. Available at cran.r-project.org/web/packages/KernSmooth/KernSmooth.pdf (verified 20 June 2010). Inst. for Stat. and Math., Wirtschafts Univ., Vienna.