Available online at www.sciencedirect.com
ScienceDirect
Geoderma 140 (2007) 444–453
GEODERMA
www.elsevier.com/locate/geoderma

Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed

David J. Brown *

Land Resources and Environmental Sciences, Montana State University — Bozeman, USA

Abstract

Combining global soil-spectral libraries with local calibration samples has the potential to provide improved visible and near-infrared (VNIR, 400–2500 nm) diffuse reflectance spectroscopy (DRS) soil characterization predictions than with either global or local calibrations alone. In this study, a geographically diverse "global" soil-spectral library with 4184 samples was augmented with up to 418 "local" calibration soil samples distributed across a 2nd-order Ugandan watershed to predict the amount of clay-size material (CLAY), soil organic carbon (SOC) and proportion of expansible 2:1 clays (termed "montmorillonite" or MT in the global library). Stochastic gradient boosted regression trees (BRT) were employed for model construction, with a variety of calibration and validation schemes tested. Using the global library combined with 13- and 14-fold cross-validation by local profile for CLAY and SOC, respectively, yielded dambo/upland RMSD values of 89/68 g kg^{-1} for CLAY (N=429/410) and 4.2/2.6 g kg^{-1} for SOC (N=272/105). These results were obtained despite the challenge of combining spectral libraries constructed using different spectroradiometers and laboratory reference measurements (total combustion *vs.* Walkley–Black, hydrometer *vs.* pipette). Using *only* the global library, a VNIR-derived index of MT content was significantly correlated with the square root of X-ray diffraction (XRD) MT peak intensity for local dambo soils (r^2=0.52, N=59, p<0.0001), an acceptable result given the semi-quantitative nature of the reference XRD method. Though VNIR predictions did not approach laboratory precision, for soil-landscape modeling VNIR characterization worked remarkably well for clay mineralogy, was adequate for mapping dambo "depth to 35% clay", and was insufficiently accurate for SOC mapping.

Keywords: Diffuse reflectance spectroscopy; Proximal soil sensing; VNIR; Soil-landscape modeling; Boosted regression trees; Clay mineralogy; Soil organic carbon; Dambo

1. Introduction

Further advances in quantitative soil-landscape modeling (McKenzie et al., 2000), precision agriculture (Rossel and McBratney, 1998) and global soil organic carbon (SOC) monitoring (Post et al., 2001) require the development of techniques for rapid, inexpensive soil characterization. Recent research has suggested that proximal visible and near-infrared (VNIR, 0.4–2.5 μm[1]) diffuse reflectance spectroscopy (DRS) could provide rapid, inexpensive predictions of soil physical, chemical and biological properties (Ben-Dor and Banin, 1995; Dunn et al., 2002; McCarty et al., 2002; Shepherd and Walsh, 2002; Islam et al., 2003; Brown et al., in press). The advantages of VNIR-DRS for these applications include: (i) rapid scans, <1 s; (ii) a relatively large scanning area of ~3 cm^2; (iii) the ability to scan soils without fine grinding; (iv) minimal specular reflectance in the VNIR region minimal specular reflectance in the VNIR region (Olinger and Griffiths, 1993) and (v) light-weight, portable scanners that can be used in the field or laboratory.

Researchers have suggested that the construction of large, global soil-spectral libraries could facilitate the wider use of VNIR-DRS by reducing the number of calibration samples required for local applications (McCarty et al., 2002; Shepherd and Walsh, 2002; Brown et al., in press). *For a 2nd-order watershed in central Uganda*, I (i) test this hypothesis, using a global soil-spectral library (Brown et al., in press) to augment local calibration samples for soil organic carbon (SOC), clay content (CLAY), and clay mineralogy estimation; (ii) explore the impact of VNIR soil property prediction error on the construction of simple soil-landscape models for this catchment.

* Tel.: +1 406 994 3724; fax: +1 406 994 3933.
E-mail address: djbrown@montana.edu.
[1] In remote sensing terminology, 0.4–2.5 μm includes the visible (VIS), near-infrared (NIR), and short-wave infrared (SWIR) regions.

0016-7061/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.geoderma.2007.04.021

Adding a degree of "realistic" difficulty, different spectro-radiometers and laboratory reference methods were used for global and local soil analysis.

1.1. Fundamentals of VNIR-DRS

Spectral signatures of materials are defined by their reflectance, or absorbance, as a function of wavelength. Under controlled conditions, the signatures are due to electronic transitions of atoms and vibrational stretching and bending of structural groups of atoms that form molecules and crystals. The fundamental vibrations of most soil materials can be found in the mid-infrared (MIR) region, with weaker and broader overtones and combinations found in the near-infrared (NIR) region. For example, the C–H stretch fundamental absorption feature can be found at ~3.4 μm in the MIR, with overtones at ~1.7, 1.15, and 0.85 μm in NIR (Workman and Springsteen, 1998; Weyer and Lo, 2002). Similarly, clay minerals have diagnostic overtone and combination absorption features in the NIR region: the O–H stretch 1st overtone at ~1.4 μm; the O–H stretch, H_2O bend combination at ~1.9 μm; the O–H stretch, metal–OH bend combinations at ~2.2–2.3 μm; and many minor absorption features (Hunt, 1977; Clark, 1999). The secondary Fe-oxyhydroxides hematite and goethite are also easily identified in the VNIR region, with broad electronic absorptions at higher energy NIR wavelengths (0.7–1.0 μm) as well as in the Vis region (0.4–0.7 μm) giving rise to the distinctive red and yellow colors (Scheinost et al., 1998; Scheinost and Schwertmann, 1999). Given that clays and to a lesser extent organic matter have well-recognized diffuse reflectance absorption features in the VNIR region related to their basic chemistry and mineralogy, there is a reason to believe that combining local samples with a global soil-spectral library could improve on local calibration samples alone for the local prediction of SOC, CLAY and clay mineralogy.

1.2. Data-mining for model calibration

Boosted regression trees (BRT) have been proposed as an ideal data-mining or pattern-recognition tool for VNIR-DRS soil characterization (Brown et al., in press). The primary advantages of boosting include (i) the ability to include a large number of weak relationships in a predictive model; (ii) insensitivity to outliers in the calibration dataset; (iii) no need for uniform data transformations; and (iv) relative immunity to "overfitting" (Freund and Schapire, 1997; Freund and Schapire, 2000; Friedman et al., 2000a,b; Ridgeway, 2000; Friedman and Meulman, 2003).

Following Friedman (2001), boosted models can be expressed in the general form:

$$F\left(x;\{\beta_m,a_m\}_0^M\right)=\sum_{m=0}^{M}\beta_m h(x;a_m) \quad (1)$$

where $h(x;a)$ represents a simple classification function or "base learner" with parameters a and input variables x, m represents the model step, and β is a weighting coefficient. The base learner (*e.g.* a simple regression tree) is applied sequentially to reweighted calibration datasets such that observations with larger residuals receive proportionally greater weights in subsequent iterations. The final classification is computed with a weighted vote as shown in Eq. (1). Friedman (2001, 2002) has developed an approach to fitting additive models of the form shown in Eq. (1), termed a Gradient Boosting Machine. With this approach, a numerical solution is found that involves sequentially fitting the base learner (using least squares) to "pseudo"-residuals computed from the gradient of a differentiable, prescribed loss function (lack of fit)—with respect to the predicted value for each calibration observation for the current step. In a further development termed "stochastic gradient boosting", Friedman (2002) has found that the use of "bagging" or random subsampling from the calibration set in conjunction with boosting improves on boosting alone. The commercial Treenet® software package applies stochastic gradient boosting to classification and auto-regression trees (CART) for stochastic boosted regression tree modeling, and proved superior to partial least squares (PLS) regression in modeling a large, diverse soil-spectral library (Brown et al., in press).

2. Site description

A previously characterized catchment located approximately 100 km north of Lake Victoria (Fig. 1) was selected for this study (Brown et al., 2003, 2004a,b). Soils in this area are formed from Pre-Cambrian granitic gneiss and associated saprolite, with upland areas underlain by dismantled ferricrete. There is a well-defined catenary soil color gradient associated

Fig. 1. Map of Uganda showing location of research site approximately 100 km north of Kampala.

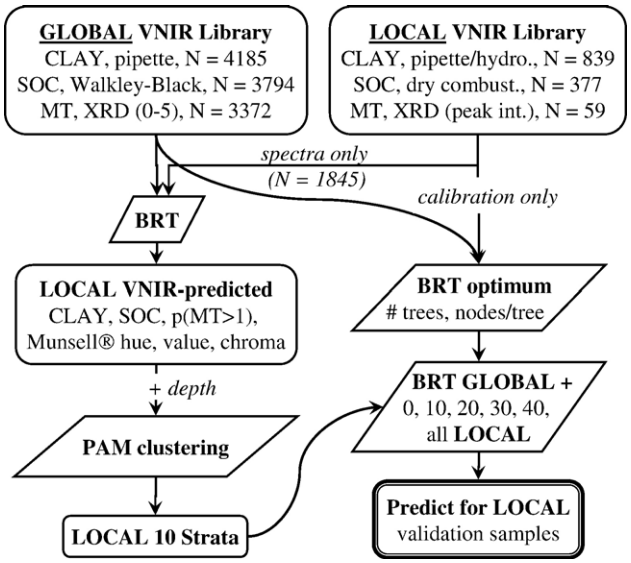

Fig. 2. Schematic VNIR spectroscopy modeling diagram highlighting the basic dimensions of the global and local soil-spectral libraries, local sample stratification, and calibration–validation procedure.

with local elevation above the valley floor due to groundwater-driven soil saturation. Dambos—grassy, seasonally saturated, gently sloping, channelless valley floors—occupy the lowest catena positions, with dark sandy clays found in the lowest seepage zones (bottoms), grey loamy sand over sandy clays on the flats, and yellow sands on the sloping dambo margins. Upland soils are comprised of more weathered, kaolinitic, red sandy clay loams to clays.

3. Methods

3.1. Global-local soil spectroscopy

The schematic diagram provided in Fig. 2 shows the spectral libraries and sequential modeling steps employed in this study.

3.1.1. Global soil-spectral library

Selection of 4184 samples for the global soil-spectral library (3768 diverse surface and subsoil samples from all U.S. states and territories, and an additional 416 samples from 36 different countries in Africa, Asia, the Americas and Europe) from the US National Soil Survey Center Soil Survey Laboratory (NSSC-SSL) archives in Lincoln, NE is described fully in a previous publication (Brown et al., in press). SOC was determined for 3794 samples using the Walkley–Black chemical oxidation method (Walkley and Black, 1934), clay content was measured with the pipette method (Gee and Bauder, 1986), and clay mineralogy was estimated using X-ray diffraction (XRD) applied to oriented clay fractions with peak intensity converted to an ordinal scale of relative mineral content (0–5) for kaolinite and 2:1 expansible smectites—termed "montmorillonite" in the Soil Survey database (Soil Survey Staff, 1996).

Air-dry (48 h at ~43 °C), crushed and sieved (<2 mm) soil samples were scanned using an ASD "Fieldspec Pro FR" VNIR spectroradiometer (Analytical Spectral Devices "ASD" FieldSpecPro® Spectroradiometer, Boulder, CO, 2–10 nm effective resolution, 0.35–2.5 μm). Soil was scanned from below using an ASD high-intensity source probe and white light source with Duraplan® borosilicate optical-glass Petri dishes to hold samples and a Spectralon® panel for white referencing. Two composite scans (consisting of 15 internally averaged scans of 100 ms each) were obtained for each sample, with a 90° sample rotation between scans. To avoid albedo effects due to differences in particle-size and moisture, the first derivatives of reflectance at 10 nm intervals were extracted from cubic smoothing splines following the procedures described in Brown et al. (in press).

3.1.2. Local soil-spectral library

At each of the 193 locations marked on Fig. 3, soil was sampled with a bucket auger (8.26 cm or 5.59 cm diameter) at 10-cm depth intervals to 270 cm (less if blocked by a gravel and/or ferricrete layer). Samples beginning at 0, 20, 50, 90, 140, 200 and 260 cm depths, and additional samples across stone lines or textural gradients were bagged for laboratory analysis. Material was air-dried, hand crushed with a mortar and pestle, and passed through a 2-mm sieve.

A total of 1845 local soil samples (<2 mm air dry) were scanned and processed following the procedures described above for the global library. Though both the global and local samples were scanned using the same protocol and with ASD instruments having identical resolution and range, there were some important differences in scanning instrumentation. The local samples were scanned (i) with an earlier version of the FieldSpecPro, having slightly different spectral ranges on the three internal detectors; and (ii) using a battery-powered, portable, lower energy white light source (leading to lower S/N performance characteristics).

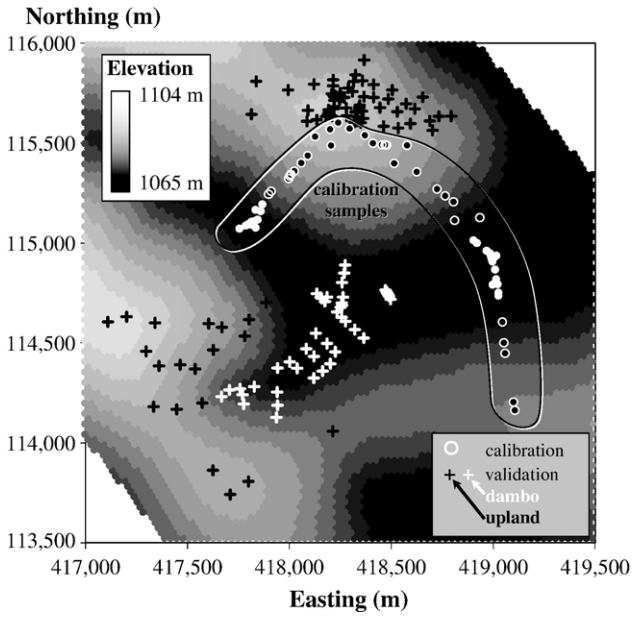

Fig. 3. Map of study catchment showing profile sampling locations, calibration *vs.* validation profiles, and poorly drained dambo *vs.* red upland soils.

The local library was stratified in three ways: (i) geographically distinct calibration *vs.* validation samples (Fig. 3) to ensure independent validation (Brown et al., 2005); (ii) well-drained upland *vs.* poorly drained dambo soils; (iii) and (ii) by preliminary estimations of composition and depth. Raw reflectance values for all local samples were converted to CIE *x*, *y*, and *z* color values using a 2° observer table. Using the Munsell Conversion® software version 6.2, the average CIE *x*, *y*, and *z* values were then converted to Munsell hue, value and chroma. For each profile, Munsell hue and chroma at 50–60 cm depth were plotted against local elevation above the valley floor (EASy) computed in a previous study (Brown et al., 2004a). Based upon the visual identification of a break between red and yellow soils, 74 dambo soils were defined by EASy <7 m.

For all 1845 local samples, montmorillonite (MT) content (ordinal, 0–5), CLAY and SOC were estimated using the global soil-spectral library and stochastic gradient boosted regression tree (BRT) models previously constructed using Treenet® (Salford Systems, San Diego, CA, USA) commercial software (Brown et al., in press). From the Treenet® MT class probability outputs, p(MT) >1 (greater than trace amounts) was computed. Local samples were then assigned to 10 strata using partitioning around mediods (S-Plus) applied to a principal components transformation of (i) sample depth; (ii) Munsell® hue, value and chroma; and (iii) VNIR predictions of SOC, CLAY and p (MT>1) using the global soil-spectral library *only*.

Characterization data was available for a subset of the 1845 samples used in previous studies (Brown et al., 2003, 2004a,b; Mahan and Brown, accepted for publication). Texture was determined using the hydrometer method (Gee and Bauder, 1986) for dambo surface soils and upland soil samples at depths of 0, 20 and 50 cm. Texture was determined using the pipette method for all remaining dambo samples in the upper and lower transects shown in Fig. 3, as well as for 200 samples selected randomly from 200 strata derived from an analysis VNIR 1st derivative spectra (Brown et al., 2003). SOC was measured by dry combustion (LECO CNS-2000®) less inorganic C (Sherrod et al., 2002) for the same 200 samples, all surface samples and for the top ~100 cm of soils in the upper and lower dambo transects (Fig. 3). SOC (dry combustion) and CLAY (pipette) were also determined for additional samples where soil material was available.

For X-ray diffraction (XRD) analysis, dambo samples were stratified by ordinal MT level with 20 samples randomly selected as evenly across classes 0–3 (maximum for local samples) as possible with a maximum of one sample per profile. An additional 39 samples were selected to characterize the clay mineralogy of representative profiles in the highest and lowest dambo transects. The clay fractions from these 59 samples were saturated with Mg, prepared and plated with and without glycerin solvation on porous ceramic tiles under suction (Whittig and Allardice, 1986). Prepared samples were scanned with a Scintag Pad V X-ray diffractometer with Cu-alpha source with entrance and exit slit settings of 2 and 3 mm and 1 and 0.5 mm respectively, at a scan rate of 2.5° min^{-1}. Peak counts per second (CPS) were recorded for indicative montmorillonite and kaolinite peaks with glycerin solvation (kaolinite and montmorillonite were the predominant clay minerals identified).

3.1.3. VNIR modeling

Stochastic gradient boosted regression trees (BRT), as implemented in the commercial Treenet® software package, were used for all VNIR modeling. To account for the stochastic nature of Treenet®, each model was fit three times with a random record-order shuffling between iterations. Since the ordinal classification scheme for clay mineralogy in the global library could not be replicated at Montana State University, the previously developed global model (Brown et al., in press) was applied to local samples *without* local calibration samples (300 trees, 10 nodes/tree). For SOC and CLAY, the optimum number of tree and nodes/tree was estimated heuristically by applying BRT to the global library plus all local calibration samples and 100–1000 trees with 6, 8 and 10 nodes/tree (6-fold cross-validated for local samples only). Using these optimum model dimensions, BRT calibrations were then constructed using: (i) the global library only; (ii) global plus a stratified random selection of 10, 20, 30, 40 and all local samples (using the 10 strata described previously); and (iii) all local calibration samples.

These BRT calibrations were then applied to the local validation sample spectra and the following statistics were computed to evaluate the quality of model fits following Gauch et al. (2003):

$$\mathrm{MSD} = \sum\nolimits_n \left(Y_{\mathrm{pred}} - Y_{\mathrm{meas}}\right)^2 / N \tag{2}$$

$$\mathrm{RMSD} = \sqrt{\mathrm{MSD}} \tag{3}$$

$$\mathrm{Bias} = \sum\nolimits_n \left(Y_{\mathrm{pred}} - Y_{\mathrm{meas}}\right) / N \tag{4}$$

$$\mathrm{SB} = \mathrm{Bias}^2 \tag{5}$$

$$\mathrm{NU} = (1-b)^2 \times \mathrm{var}\left(Y_{\mathrm{meas}}\right) \tag{6}$$

$$\mathrm{LC} = \left(1 - r^2\right) \times \mathrm{var}\left(Y_{\mathrm{pred}}\right) \tag{7}$$

where b and r^2 are the slope and coefficient of determination respectively from the least squares regression of Y_{pred} on Y_{meas}. The Mean Squared Deviation (MSD) is partitioned into three independent components describing lack of accuracy due to bias (SB), non-unity regression line (NU), and lack of correlation (LC), with MSD=SB+NU+LC. Standard chemometric statistics were also computed, including SEP (validation standard error of prediction, RMSD $\times \sqrt{N/(N-1)}$) and RPD (standard deviation of validation samples/SEP) (Islam et al., 2003).

3.2. Soil-landscape modeling

For the whole catchment VNIR prediction (necessary for soil-landscape modeling), I used 13-fold and 14-fold cross-validation for CLAY and SOC respectively, with entire profiles randomly assigned to cross-validation clusters. To enforce spatial dispersion for SOC cross-validation, one sample per

Table 1
Summary of clay content data (CLAY) for soil-spectral libraries and partitions

	CLAY (g kg^{-1})				
	Global	Dambo calibration	Dambo validation	Upland calibration	Upland validation
N	4184	300	129	118	292
Quartiles					
Min	1	40	50	220	220
1st	112	198	180	343	340
Median	230	310	330	420	400
3rd	367	440	430	480	460
Max	912	740	650	610	740

profile was randomly selected for inclusion in model calibration (140 samples total) while all available samples (206) were used for model validation.

Profile indices were computed for MT (VNIR only), CLAY and SOC (VNIR and laboratory values) to facilitate landscape modeling. The continuous p(MT>1) was interpolated and averaged linearly from 50–100 cm for each profile, yielding an index of subsoil MT content (not a true probability). Similarly, average SOC from 0 to 30 cm depths was computed for all profiles using VNIR and for laboratory data where 0–10 and 20–30 cm SOC values were available. Since no upland soils had both 0–10 and 20–30 cm data available, SOC values for thirteen 0–10 cm and twelve 20–30 cm upland samples were averaged first within depth, then combined to derive a single composite upland profile average 0–30 cm value, with a maximum EASy value of 10 m.

As the dominant texture profile found in depositional dambo soils was sand over clay, a "depth-to-clay" was computed by moving down each profile until two consecutive samples with CLAY ≥ 35% were encountered. Since gravel layers were encountered in some profiles prior to reaching the clayey saprolite and in a few cases no clay was encountered within the 270 cm sampled, "depth-to-clay" was computed as the minimum of (i) depth to 35% clay; (ii) depth to 5 g gravel (>4 mm diameter) per 100 g soil; and (iii) lowest depth sampled. For VNIR, 35% clay was encountered first for 100 of 119 upland and 58 of 74 dambo profiles. For laboratory-characterized profiles, 35% clay was encountered first for 63 of 74 upland and 24 of 32 dambo profiles.

The MT, CLAY and SOC profile indices described above were regressed against local elevation above the valley floor (EASy, truncated at 10 m) computed as part of a previous study (Brown et al., 2004a). Standard linear least squares regression (LM), generalized least squares (GLS) regression, and generalized non-linear least squares regression (GNLS) were employed as appropriate for each modeled relationship (all statistical modeling in S-Plus 6.0®).

4. Results and discussion

4.1. Data summary

Significant amounts of kaolinite were detected in all samples analyzed using both XRD and VNIR. Smectites were the other major clay mineralogy constituent, or montmorillonite (MT) following the NSSC-SSL archive terminology. (Micas and halloysite were detected in a few samples using XRD). Overall variability in clay mineral composition was due largely to the variability in relative MT content.

A summary of CLAY data for various soil-spectral libraries and partitions (dambo *vs.* upland, calibration *vs.* validation) is provided in Table 1. As can be seen, the distribution of calibration and validation values for both dambo and upland soils are very well matched. The global library CLAY values extend beyond local samples on both ends, ranging from 0 to 91% clay.

A summary of SOC data for various soil-spectral libraries and partitions (dambo *vs.* upland, calibration *vs.* validation) is provided in Table 2. As can be seen, the distribution of calibration and validation SOC values for both dambo and upland soils are very well matched. The global library SOC values extend beyond the local samples, particularly on the upper end with the upper quartile of the global SOC values greater than all local SOC values.

4.2. Clay mineralogy

An examination of XRD peak intensity *vs.* VNIR-estimated ordinal level of MT content shows a strong correspondence (Table 3). Not only does the average peak intensity increase with increasing VNIR ordinal level, but with a few minor exceptions the quartiles also consistently increase with increasing levels of MT predicted by VNIR modeling. Converting VNIR ordinal data to a continuous index of MT content (the BRT-estimated probability that MT class is greater

Table 2
Summary of soil organic C (SOC) data for soil-spectral libraries and partitions

	SOC (g kg^{-1})				
	Global	Dambo calibration	Dambo validation	Upland calibration	Upland validation
N	3794	178	94	28	77
Quartiles					
Min	0.0	0.3	0.3	1.1	1.0
1st	1.9	2.7	1.9	2.1	3.1
Median	4.7	5.1	4.8	3.2	4.9
3rd	12.3	10.4	11.7	6.5	7.0
Max	536.8	46.8	57.7	15.0	18.6

Table 3
Dambo soil montmorillonite (MT) XRD peak intensity (counts per second or CPS) as a function of VNIR-estimated peak intensity ordinal class using a global soil-spectral library with 3372 samples (Brown et al., in press), boosted regression trees (300 trees, 10 nodes/tree) and no local calibration samples

VNIR ordinal MT class	X-ray diffraction							
	N (59)	Avg. Peak (CPS)	MT not detected	Quartiles (MT peak CPS)				
				Min	1st	Median	3rd	Max
0	27	132	13	0	0	74	207	513
1	4	294	1	0	246	333	381	508
2	14	403	1	0	225	386	546	1039
3	14	857	1	0	632	967	1204	1476

than 1) yielded a linear relationship between p(MT>1) and the square root of XRD peak intensity ($r^2=0.52$, $N=59$, $p<0.0001$ for both model and regression coefficient.)

Two samples had *no* XRD-detected MT but *high* VNIR p [MT>1] values of 0.77 and 0.88. Both were located in dambo bottom subsoils where MT is normally found. X-ray diffraction (XRD) analyses revealed the presence of hydrated halloysite (10 Å) in one of these samples, suggesting that hydrated minerals might be confused with MT in VNIR modeling. However, VNIR might be a more sensitive technique for some materials as it is very difficult to detect trace amounts of clay minerals using XRD (Whittig and Allardice, 1986). XRD peak intensities are only loosely correlated with clay mineral content, yielding semi-quantitative estimates (Whittig and Allardice, 1986). Though VNIR might also be semi-quantitative, it has the advantage of being rapid and inexpensive relative to the very time consuming and expensive procedures involved in XRD clay mineral determination. For many applications, VNIR might best be employed to screen a large number of samples with only a few representative samples submitted for detailed XRD analysis.

4.3. Clay content

The optimum model size for clay prediction—using all available samples from (i) the global soil-spectral library (4184); (ii) local dambo profiles (300); and (iii) local upland profiles (118)—was found to be 700 trees with 8 nodes/tree. However, there was a little difference in predictive accuracy for models with more or less trees (>400) and 6–10 nodes. The predictions for 10 nodes/tree showed slightly higher MSD, due largely to higher non-unity error contributions, suggesting that these more complex models might lead to slight overfitting.

Clay predictions for the holdout samples improved with increasing numbers of local calibration samples (Fig. 4). Validation RMSD declined with the addition of 10–40 local samples to the global library (127 to 105 g kg^{-1} dambo, 103 to 87 g kg^{-1} upland), but predictions improved even more when all 418 available local samples were included in the model (94 g kg^{-1} dambo, 75 g kg^{-1} upland). When only local samples were used for model calibration, validation RMSD for dambo soils increased to 103 g kg^{-1}, largely due to an increase in the square root of SB+NU that increased from 65 to 75 g kg^{-1}. This suggests that without the inclusion of a global dataset, BRT modeling can overfit local patterns, leading to the construction of highly local models. For upland soils, however, RMSD declined slightly to 71 g kg^{-1} using only the local 418 local samples for calibration.

The whole catchment 13-fold cross-validation (by profile) results for CLAY (Table 4, Fig. 5) are comparable to those obtained with the geographically independent holdout validation (Fig. 4). There is very little model bias, but a significant

Table 4
Cross-validation prediction results for clay (randomly assigned to 13 cross-validation clusters by profile) and SOC (randomly assigned to 14 cross-validation clusters by profile, with one sample per profile, 140 samples total, randomly selected for calibration)

Target	Clay (g kg^{-1})			SOC (g kg^{-1})		
	All	Dambo	Upland	All	Dambo	Upland
N	839	429	410	377	272	105
SEP	79	89	68	3.8	4.2	2.6
RMSD	79	89	68	3.8	4.2	2.6
MSD	6292	7936	4573	14.7	17.8	6.6
SB	0	2	1	0.0	0.0	0.0
NU	2007	2306	2394	3.6	5.0	0.4
LC	4285	5627	2177	11.1	12.8	6.2
Bias	0.2	1.5	−1.1	0.1	0.1	0.2
r^2	0.64	0.66	0.43	0.78	0.79	0.55
Slope	0.66	0.69	0.45	0.77	0.76	0.82
RPD	1.7	1.7	1.3	2.1	2.2	1.3

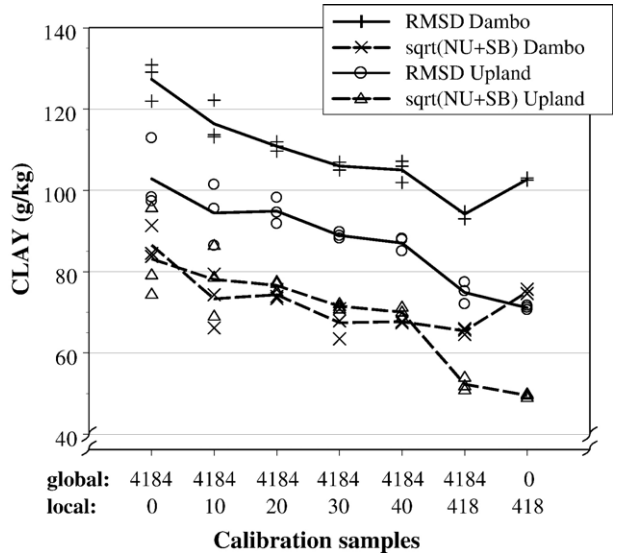

Fig. 4. Geographically independent holdout validation SOC RMSD and square root of NU+SB for various global–local calibration schemes. Three stochastic realizations are provided for each scenario, with the lines fit to average values.

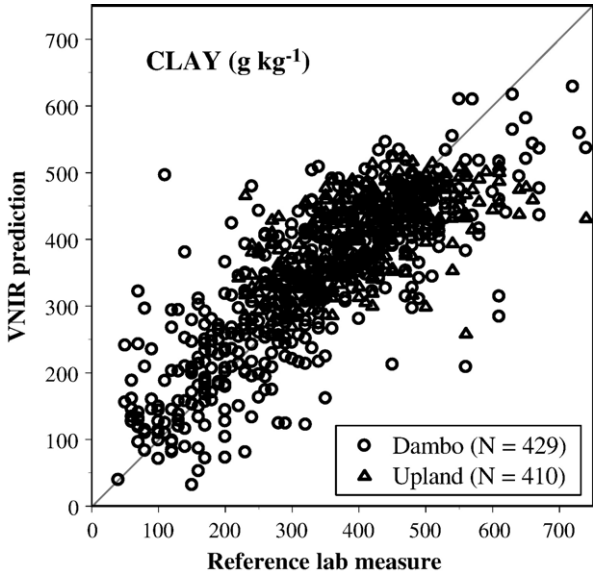

Fig. 5. VNIR predicted *vs.* laboratory-measured CLAY (pipette method) for 13-fold cross-validation with entire profiles randomly assigned to the 13 clusters. Complete prediction statistics are provided in Table 4.

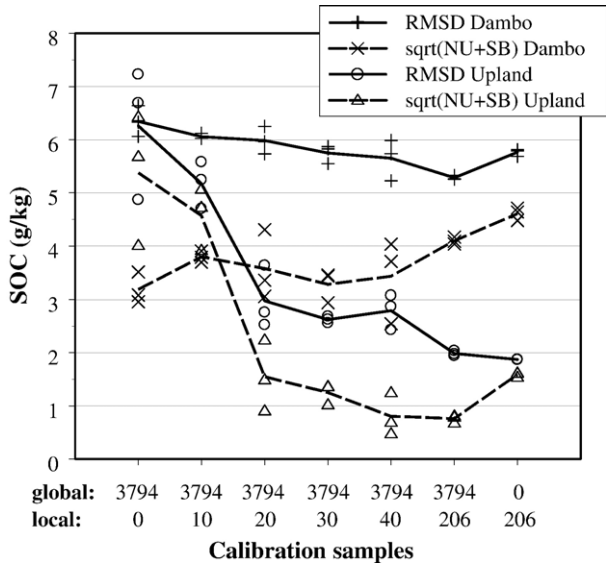

Fig. 6. Geographically independent holdout validation CLAY RMSD and square root of NU+SB for various global–local calibration schemes. Three stochastic realizations are provided for each scenario, with the lines fit to average values.

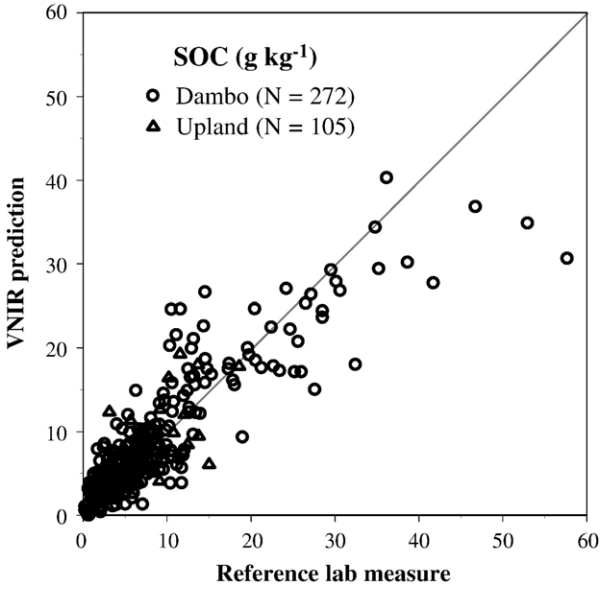

Fig. 7. VNIR predicted *vs.* laboratory-measured SOC (total combustion) for 14-fold cross-validation with entire profiles randomly assigned to 14 clusters and one sample randomly selected from each profile for calibration. Complete prediction statistics are provided in Table 4.

non-unity contribution to model error with regression slopes of 0.43–0.66. A visual examination of the scatterplot in Fig. 5 shows a "hockey-stick" pattern with notable underestimations at higher clay values. However, looking *only* at soils with 0–50% clay, the fit is improved with a regression slope of 0.76 (N=741). This bodes well for potential VNIR use in soil-landscape modeling. Given pure random error (no bias or systematic trend) and enough samples (high quantity), an accurate mean clay content can be estimated despite relatively high prediction error (low quality).

4.4. Soil organic C (SOC)

The optimum model size for SOC prediction—using all available calibration samples from (i) the global soil-spectral library (4184); (ii) local dambo profiles (178); and (iii) local upland profiles (28)—was found to be 800 trees with 8 nodes/tree. Similar results were obtained for 500–1000 trees and 6–8 nodes/tree. The MSD values for 10 node/tree models, however, were consistently higher largely due to greater bias and non-unity contributions (SB and NU). As with the CLAY models, it would seem possible to overfit VNIR boosted regression tree models, particularly with higher levels of predictor interaction included (more nodes/tree).

SOC predictions for holdout samples improved with increasing numbers of local calibration samples added to the global library (Fig. 6). Upland soil validation RMSD dropped dramatically from 6.3 to 3.0 g kg^{-1} dry soil with the addition of 20 local samples to the model calibration dataset, mainly through the elimination of a large prediction bias. With the addition of all 206 local samples, SOC validation RMSD dropped to 2.0 g kg^{-1} for upland samples. These dramatic gains could be due to the relative lack of highly weathered soils in the global spectral library (Brown et al., in press). For the dambo soils, however, improvements in SOC prediction were incremental with validation RMSD dropping from 6.4 to 5.7 g kg^{-1} and 5.3 g kg^{-1} with the addition of 40 and 206 local samples, respectively, to the global library. Moreover, the square root of NU+SB increased from 3.2 to 4.1 g kg^{-1} with the addition of 206 local samples.

When only the 206 available local samples were used for SOC model calibration, validation RMSD increased to 5.8 g kg^{-1} for

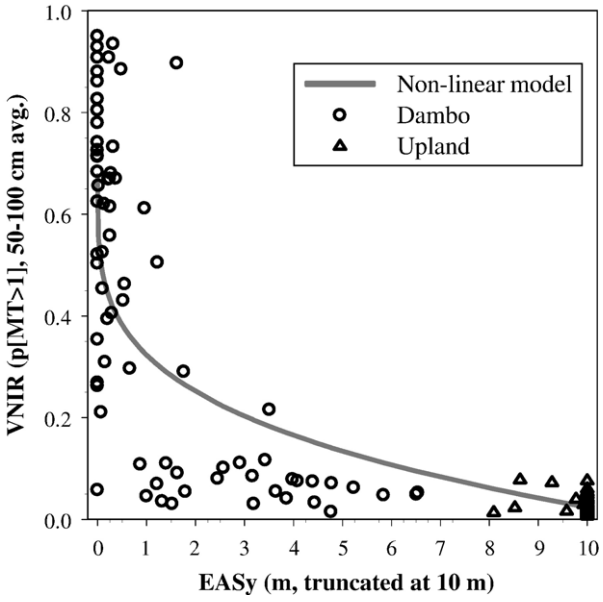

Fig. 8. Plot of a VNIR-derived index of subsoil montmorillonite (MT) content (50–100 cm linear average of p[MT>1], where 1=trace) against local elevation above the valley floor (EASy). Generalized non-linear least squares (GNLS) model fit with residual error estimated by group (upland *vs.* dambo), all model coefficients significant at $p<0.0001$ and whole model AIC=−660. Non-linear model parameterization is provided in Table 5.

Table 5
Models of soil profile attributes (probability of subsoil MT > trace, and depth to 35% clay) as function of local elevation above the valley floor (EASy), using linear least squares (LM), generalized least squares (GLS), and generalized non-linear least squares (GNLS) models; with all models and predictors significant at $p < 0.0001$

No.	Model	Type	N	r^2	Residual s.e.	Range	Nugget/Sill	Residual corr. structure
1	$p[MT > 1]_{50-100\,cm} = 0.68 - \left(\frac{EASy}{50}\right)^{0.27}$	GNLS	190	–	0.23	–	–	$\frac{\sigma_{dambo}}{\sigma_{upland}} = \frac{1}{0.06}$
2	$Clay.depth_{VNIR} = -3.4 + 1.00 \times Clay.depth_{LAB}$	LM	30	0.84	29	–	–	–
3	$Clay.depth_{LAB} = 51 + 41 \times EASy$	LM	30	0.74	36	–	–	–
4	$Clay.depth_{VNIR} = 79 + 32 \times EASy$	GLS	69	–	64	121	0.08	spherical

dambo soils and decreased very slightly to 1.9 g kg^{-1} for upland soils (relative to all global plus 206 local samples). For both, the square root of NU + SB increased substantially from 4.1 g kg^{-1} to 4.6 g kg^{-1} and 0.8 g kg^{-1} to 1.6 g kg^{-1} for dambo and NU contributions for upland soils also increased with the absence of the global library, though validation RMSD decreased very slightly. These results suggest that the addition of a diverse, global soil-spectral library to local calibration samples leads to the construction of more robust models and reduces overfitting.

For dambo samples, the whole catchment 14-fold cross-validation results for SOC (Table 4, Fig. 7) are greatly improved over those obtained with the geographically independent holdout validation shown in Fig. 6 (validation RMSD of 4.2 *vs.* 5.3 g kg^{-1}, global library included in both calibrations). The main difference lies in lower SB + NU values for the former (validation square root of SB + NU = 2.2 *vs.* 4.1 g kg^{-1}). For upland soils, the RMSD is higher with the cross-validation model (2.6 *vs.* 2.0 g kg^{-1}) but the square root of SB + NU is slightly lower (0.6 *vs.* 0.8 g kg^{-1}). For the SOC cross-validation model, only one local sample was selected from each profile for calibration (though all 206 samples were used in model validation). This guarantees a degree of geographic distribution and appears to result in improved robustness. As with CLAY prediction, the VNIR model underpredicts for high levels of SOC (Fig. 7), but generally fits the 1:1 line elsewhere with little bias (bias = 0.1 g kg^{-1}, regression slope = 0.77).

4.5. Soil-landscape modeling

A plot of subsoil montmorillonite content (50–100 cm average VNIR p[MT > 1]) against local elevation for the study catchment (190 profiles) reveals a clear, negative, exponential relationship (Fig. 8, Table 5, model 1). Montmorillonite is a secondary clay mineral that forms in generally wet, alkaline, Si-rich landscape positions (Wilson, 1999). The abundance of MT in the lower, geochemically enriched dambo positions is to be expected. The lack of fit in the reported non-linear model (Table 5, model 1), particularly for the lower elevation dambo profiles, could be due in part to the lack of precision (30-m horizontal resolution) for the elevation model employed. These results indicate that using currently available models and spectral libraries, VNIR diffuse reflectance spectroscopy provides a viable tool (perhaps the only viable tool) for quantifying smectitic in a large number of soil samples, such as for soil-landscape modeling applications.

As shown in Fig. 9, a similar relationship with local elevation (EASy) can be observed for average 0–30 cm depth SOC. However, the VNIR model systematically overestimates SOC content for the sandy margin soils (EASy = 3–7 m) and underestimates SOC for the lower dambo soils. As a result, the soil-landscape relationship between local elevation and SOC content that can be observed with the total combustion SOC data is obscured with VNIR-estimated SOC. Using presently available calibration samples, VNIR modeling does not provide sufficiently accurate SOC predictions for soil-landscape modeling in this study area.

Can VNIR-estimated clay content be used to model depth-to-clay (35%) for dambos? There was a very high correlation ($r^2 = 0.84$, $N = 30$, $p < 0.0001$) between laboratory-measured and VNIR measured depth-to-clay (Table 5, model 2) for depositional soils. Similar linear model coefficients were also obtained when depth-to-clay (laboratory and VNIR) was regressed against EASy (Table 5, models 3 and 4). The VNIR-estimated depth-to-clay data show more scatter (Fig. 10) and a slightly lower regression slope, but this appears to be largely due to the inclusion of more profiles in the model at a wider range of catchment positions. The use of VNIR for clay estimation allows for more soils to be analyzed and therefore the construction of more complex and sophisticated soil-landscape models. For a more complete evaluation of VNIR applications to soil-landscape modeling, systematic studies are needed

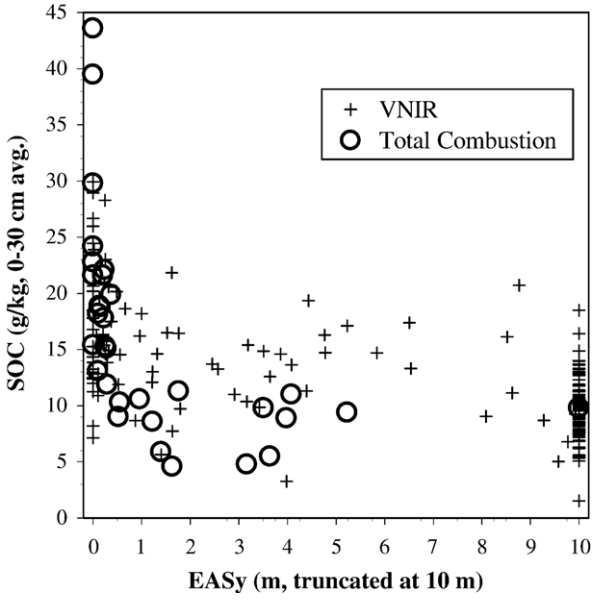

Fig. 9. Plot of average surface (0–30 cm) SOC against local elevation above the valley floor (EASy) for both profiles with available total combustion data ($N = 32$) and VNIR cross-validation estimated SOC ($N = 193$).

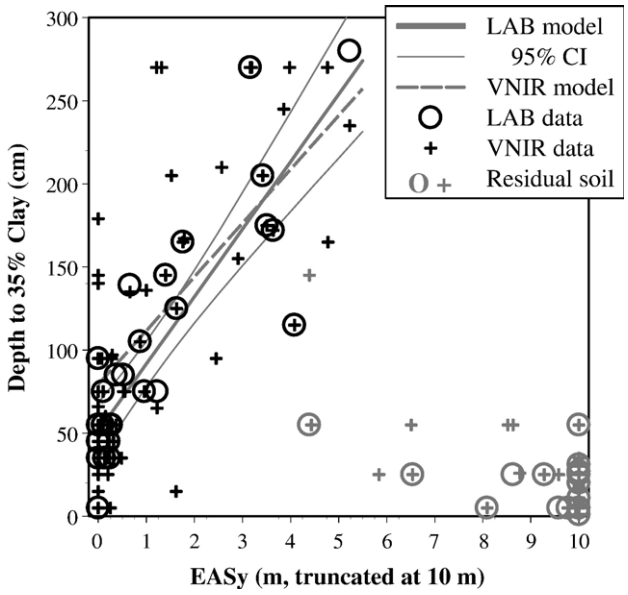

Fig. 10. Plot of depth to 35% clay against local elevation above the valley floor (EASy) for both profiles with available laboratory reference CLAY (pipette dambo and hydrometer upland) and VNIR cross-validation estimated CLAY. Only depositional dambo profiles (black plot symbols) were included in soil-landscape modeling, including 30 profiles with laboratory data and 69 with VNIR-estimated depth-to-clay. While laboratory- and VNIR-derived depth-to-clay values were highly correlated ($r^2=0.84$, $p<0.0001$, $N=30$), the greater number of VNIR-characterized profiles at more diverse catchment positions introduced a more complex soil-landscape relationship. Complete modeling results are provided in Table 5.

where laboratory reference and VNIR data are available for all samples in the study area.

5. Conclusions

In previous work, Brown et al. (in press) hypothesized that combining a global soil-spectral library with local calibration samples could yield better VNIR predictions of soil properties than either global or local calibration alone. The results of this study generally support that assertion. For both well-drained upland and poorly drained dambo soils, validation RMSD values for CLAY and SOC declined as 40 diverse local samples were added to the global library, and declined further as all calibration samples were included (206 for SOC, 418 for CLAY). For SOC, most of the predictive improvement can with the addition of just 40 local calibration samples whereas CLAY predictions improved markedly with all calibration samples included.

For dambo soils, the combination of all local and global calibration samples yielded lower validation RMSD values for both CLAY and SOC relative to the local calibration samples alone. For upland soils, the reverse was observed with slightly worse (higher) validation RMSD values for the combined global-local calibration. Since SOC and CLAY variability was much greater for dambo soils, the overall validation RMSD values were lower for global-local *vs.* local only calibrations. The major advantage of combining global and local samples, however, might lie in the ability to construct viable calibrations with relatively few local samples.

Using the global soil-spectral library combined with 13- and 14-fold cross-validation by local profile for CLAY and SOC, respectively, yielded dambo/upland RMSD values of 89/68 g kg^{-1} for CLAY ($N=429/410$) and 4.2/2.6 g kg^{-1} for SOC ($N=272/105$). These results do not approach laboratory precision, but might well suffice for applications where many approximate measurements are more valuable than a few precise analyses. For estimation of the relative proportion of expansible 2:1 clays in the clay-size fraction ("montmorillonite" or MT in the global library), VNIR was used to characterize local samples using *only* the global soil-spectral library (3372 samples) with *no* local calibration samples. Reflectance-estimated MT content (expressed as the probability of more than trace MT) was significantly correlated with the square root of XRD MT peak intensity for dambo soils ($r^2=0.52$, $N=59$, $p<0.0001$), an acceptable result given the semi-quantitative nature of the reference XRD method.

For soil-landscape modeling in central Uganda, VNIR worked best for clay mineralogy and worst for SOC. Reflectance (VNIR) SOC models yielded better predictions (regression slope, r^2 and bias) than CLAY models. However, more precision is also required for SOC analysis whereas "ballpark" clay content is often sufficient for many soil-landscape modeling applications. For example, while VNIR was only moderately successful in predicting dambo clay content relative to the reference method ($r^2=0.66$, regression slope=0.69, RMSD=89 g kg^{-1}, $N=429$), *profile* laboratory- and VNIR-estimated depth to 35% clay were highly correlated ($r^2=0.84$, $N=30$, $p<0.0001$). Consequently, VNIR performed adequately for modeling dambo "depth-to-clay" against local elevation (EASy) while VNIR SOC estimates were insufficiently accurate for landscape modeling in this study—with sandy margin SOC over predicted and clayey bottom soil SOC under predicted. Reflectance-estimated subsoil (50–100 cm) montmorillonite content showed a strong and geochemically explainable exponential relationship with local elevation.

Visible and near-infrared (VNIR) diffuse reflectance spectroscopy might be best suited for rapid, *in situ*, semi-quantitative soil characterization. It is noteworthy that the results reported in this paper were obtained despite using different spectroradiometers for global and local sample VNIR analyses, different laboratory reference methods for SOC (total combustion *vs.* Walkley–Black), and a mixture of hydrometer and pipette results for local CLAY determination. When combining soil-spectral libraries for real applications, different spectrometers and mixed laboratory analysis techniques will likely be the norm rather than the exception. Further research is needed to evaluate the potential of combining a large, global soil-spectral library (constructed using dry, sieved soils) with *in situ* soil scans to improve *in situ* VNIR soil characterization.

Acknowledgements

Staff at the U.S. National Soil Survey Center Soil Survey Laboratory provided invaluable assistance in the construction of the global soil-spectral library. Keith Shepherd and Marcus Walsh at ICRAF in Nairobi generously lent us their spectroradiometer

for analysis of Uganda samples, and assisted with the preprocessing of the spectral data. David Semugabe, Wadda Grace, and Edith Mbabazi were invaluable assistants in the field. The soils group at the Kawanda Agricultural Research Institute (KARI), and the Uganda National Council for Science and Technology (UNCST) provided institutional support within Uganda. Funding from a NSF Graduate Research Fellowship, a Fulbright Student Grant, and an NSF international postdoctoral fellowship (Award No. 0202582) supported this work.

References

Ben-Dor, E., Banin, A., 1995. Near infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Sci. Soc. Am. J. 59, 364–372.

Brown, D.J., Helmke, P.A., Clayton, M.K., 2003. Robust geochemical indices for redox and weathering on a granitic laterite landscape in central Uganda. Geochim. Et Cosmochim. Acta 67 (15), 2711–2723.

Brown, D.J., Clayton, M.K., McSweeney, K., 2004a. Potential terrain controls on soil color, texture contrast and grain-size deposition for the original catena landscape in Uganda. Geoderma 122 (1), 51–72.

Brown, D.J., McSweeney, K., Helmke, P.A., 2004b. Statistical, geochemical, and morphological analyses of stone line formation in Uganda. Geomorphology 62 (3–4), 217–237.

Brown, D.J., Bricklemyer, R.S., Miller, P.R., 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma 129 (3–4), 251–267.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D. and Reinsch, T.G. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma, (in press).

Clark, R.N., 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy. In: Rencz, N. (Ed.), Remote Sensing for the Earth Sciences: Manual of Remote Sensing. John Wiley & Sons, New York, pp. 3–52.

Dunn, B.W., Beecher, H.G., Batten, G.D., Ciavarella, S., 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 (5), 607–614.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139.

Freund, Y., Schapire, R.E., 2000. Additive logistic regression: a statistical view of boosting — discussion. Ann. Stat. 28 (2), 391–393.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat., 29 (5), 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38 (4), 367–378.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22 (9), 1365–1381.

Friedman, J., Hastie, T., Tibshirani, R., 2000a. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28 (2), 337–374.

Friedman, J., Hastie, T., Tibshirani, R., 2000b. Additive logistic regression: a statistical view of boosting — rejoinder. Ann. Stat. 28 (2), 400–407.

Gauch, H.G., Hwang, J.T.G., Fick, G.W., 2003. Model evaluation by comparison of model-based predictions and measured values. Agron. J. 95 (6), 1442–1446.

Gee, G.W., Bauder, J.W., 1986. Particle-size analysis. In: Klute, A. (Ed.), Methods of Soil Analysis: Part I. Physical and Minealogical Methods. Soil Science Society of America, Madison, WI, pp. 383–411.

Hunt, G.R., 1977. Spectral signatures of particulate minerals in visible and near IR. Geophysics 42 (3), 501–513.

Islam, K., Singh, B., McBratney, A., 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. Aust. J. Soil Res. 41 (6), 1101–1114.

Mahan, S.A. and Brown, D.J., An optical age chronology of Late Quaternary extreme fluvial events recorded in Ugandan dambo soils. Quaternary Geochronology, (accepted for publication).

McCarty, G.W., Reeves, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Sci. Soc. Am. J. 66 (2), 640–646.

McKenzie, N.J., Cresswell, H.P., Ryan, P.J., Grundy, M., 2000. Contemporary land resource survey requires improvements in direct soil measurement. Commun. Soil Sci. Plant Anal. 31 (11–14), 1553–1569.

Olinger, J.M., Griffiths, P.R., 1993. Effects of sample dilution and particle-size morphology on diffuse reflection spectra of carbohydrate systems in the near-infrared and midinfrared. 1. Single analytes. Appl. Spectrosc. 47 (6), 687–694.

Post, W.M., Izaurralde, R.C., Mann, L.K., Bliss, N., 2001. Monitoring and verifying changes of organic carbon in soil. Clim. Change 51 (1), 73–99.

Ridgeway, G., 2000. Additive logistic regression: a statistical view of boosting — discussion. Ann. Stat. 28 (2), 393–400.

Rossel, R.A.V., McBratney, A.B., 1998. Soil chemical analytical accuracy and costs: implications from precision agriculture. Aust. J. Exp. Agric. 38 (7), 765–775.

Scheinost, A.C., Schwertmann, U., 1999. Color identification of iron oxides and hydroxysulfates: use and limitations. Soil Sci. Soc. Am. J. 63 (5), 1463–1471.

Scheinost, A.C., Chavernas, A., Barron, V., Torrent, J., 1998. Use and limitations of second-derivative diffuse reflectance spectroscopy in the visible to near-infrared range to identify and quantify Fe oxide minerals in soils. Clay Clay Min. 46 (5), 528–536.

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66 (3), 988–998.

Sherrod, L.A., Dunn, G., Peterson, G.A., Kolberg, R.L., 2002. Inorganic carbon analysis by modified pressure-calcimeter method. Soil Sci. Soc. Am. J. 66, 299–305.

Soil Survey Staff, 1996. Soil survey laboratory methods manual. Soil Survey Investigations Report No. 42, version 3.0, USDA-NRCS National Soil Survey Center, Washington, D.C.

Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Sci. 37, 29–38.

Weyer, L., Lo, S.-C., 2002. Spectra–structure correlations in the near-infrared. In: Chalmers, J.M., Griffiths, P.R. (Eds.), Handbook of Vibrational Spectroscopy, vol. 3. John Wiley, New York, pp. 1817–1837.

Whittig, L.D. and Allardice, W.R., 1986. X-ray diffraction techniques. In: Klute, A. (Editor), Methods of Soil Analysis. Part 1. ASA and SSSA, Madison, WI.

Wilson, M.J., 1999. The origin and formation of clay minerals in soils: past, present and future perspectives. Clay Miner. 34 (1), 7–25.

Workman, J., Springsteen, A.W. (Eds.), 1998. Applied Spectroscopy, a Compact Reference for Practitioners. Academic Press, San Diego, CA. 539 pp.