Geoderma 183–184 (2012) 41–48

Contents lists available at SciVerse ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique

A.H. Cambule [a,*], D.G. Rossiter [b], J.J. Stoorvogel [c], E.M.A. Smaling [b]

[a] Universidade Eduardo Mondlane, Faculdade de Agronomia e Engenharia Florestal, C.P. 257, Maputo, Mozambique
[b] University of Twente, Faculty of Geoinformation Science and Earth Observation (ITC), PO Box 217, 7500 AE Enschede, The Netherlands
[c] Wageningen University, Land Dynamics Group, PO Box 47, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 21 September 2010
Received in revised form 24 January 2012
Accepted 6 March 2012
Available online 7 May 2012

Keywords:
Near-infrared spectroscopy
Spectral library
Soil organic carbon
Limpopo
Mozambique
Partial least-squares regression

ABSTRACT

Soil organic carbon (SOC) is a key soil property and particularly important for ecosystem functioning and the sustainable management of agricultural systems. Conventional laboratory analyses for the determination of SOC are expensive and slow. Laboratory spectroscopy in combination with chemometrics is claimed to be a rapid, cost-effective and non-destructive method for measuring SOC. The present study was carried out in Limpopo National Park (LNP) in Mozambique, a data- and access-limited area, with no previous soil spectral library. The question was whether a useful calibration model could be built with a limited number of samples. Across the major landscape units of the LNP, 129 composite topsoil samples were collected and analyzed for SOC, pH and particle sizes of the fine earth fraction. Samples were also scanned in a near-infrared (NIR) spectrometer. Partial least square regression (PLSR) was used on 1037 bands in the wavelength range 1.25–2.5 μm to relate the spectra and SOC concentration. Several models were built and compared by cross-validation. The best model was on a filtered first derivative of the multiplicative scatter corrected (MSC) spectra. It explained 83% of SOC variation and had a root mean square error of prediction (RMSEP) of 0.32% SOC, about 2.5 times the laboratory RMSE from duplicate samples (0.13% SOC). This uncertainty is a substantial proportion of the typical SOC concentrations in LNP landscapes (0.45–2.00%). The model was slightly improved (RMSEP 0.28% SOC) by adding clay percentage as a co-variable. All models had poorer performance at SOC concentrations above 2.0%, indicating a saturation effect. Despite the limitations of sample size and no pre-existing library, a locally-useful, although somewhat imprecise, calibration model could be built. This model is suitable for estimating SOC in further mapping exercises in the LNP.

1. Introduction

The increasing need to manage land sustainably has triggered the debate on soil quality, its definition and the indicators best reflecting it (Arshad and Martin, 2002). Some researchers have developed indicators based on selected specific combinations of soil characteristics to characterize soil quality (Yemefack et al., 2006) but still there is no consensus on how the indicators should be interpreted (Bouma, 2002). However, all indicators related to soil quality include soil organic carbon (SOC) as one of the most important properties (Arshad and Martin, 2002). Shukla et al. (2006) state that if only one soil attribute were to be used for monitoring soil quality changes, it should be SOC. The widely-used soil fertility-crop production model QUEFTS uses SOC, or total nitrogen as a proxy (assuming a stable C/N ratio), as the major yield-explaining variable (Janssen et al., 1990; Liu et al., 2006; Pathak et al., 2003; Smaling and Janssen, 1993). This comes as no surprise in strongly weathered tropical soils that largely rely on the organic fraction for their inherent soil fertility. SOC is also recognized as the best entry point for land degradation assessment (Gisladottir and Stocking, 2005).

Assessment of SOC over larger areas by field sampling and conventional laboratory analysis is expensive and slow. Laboratory spectroscopy is widely-applied in chemometrics (Geladi and Kowalski, 1986) and recently also to soil characterization (Brown et al., 2006; Shepherd and Walsh, 2002). It offers rapid and about 50% cheaper soil analysis (Cécillon et al., 2009a; O'Rourke and Holden, 2011), and, as an added benefit it is non-destructive, so samples can be analyzed repeatedly.

The most common form of spectroscopy for SOC determination is visible and near-infrared reflectance (VNIR, 0.4–3.0 μm) and mid-infrared (MIR, −3.0–30 μm) (Clark, 1999). Other authors indicate different spectral ranges for the same regions, *e.g.* Vis-NIR-SWIR to be 0.4–2.5 μm (Ben-Dor, 2002; Shepherd and Walsh, 2002). SOC produces a spectral signature, defined by the reflectance or absorbance of electromagnetic radiation as a function of wavelength. In the case of SOC, as with the absolute majority of absorbants, combination of

* Corresponding author. Tel.: +258 843285400.
E-mail address: armindo.cambule@uem.mz (A.H. Cambule).

0016-7061/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2012.03.011

bands and overtones of the fundamental spectral features is detected in the NIR region (Shepherd and Walsh, 2002).

Direct quantitative prediction from spectra is almost impossible because soil constituents interact in a complex way to produce a given spectrum. Therefore, quantification of the property of interest is done with multivariate statistical models (Cécillon et al., 2009b). Viscarra Rossel et al. (2006) demonstrated the potential of reflectance spectroscopy along with the chemometric methods applied to develop these multivariate statistical models to predict soil properties. Partial least-squares regression (PLSR) and principal components regression (PCR) were the multivariate methods most applied for SOC determination, while sample size varied from 68 to 674, resulting in a calibration R^2 of 0.86–0.96. However, only three of the reported 14 studies on SOC had a sample size below 150. Shepherd and Walsh (2002) indicate that as the sample size decreases, the predictive performance decreases gradually at large sample sizes but rapidly as sample size decreased between about 100 and 200 samples to a $R^2 < 0.7$ and even below 0.5 for sample sizes smaller than 100, implying more relative variation in the dataset.

Despite the reported problems with small sample sizes, there are many situations where it is impractical to obtain large sample sizes. Typical limitations are financial, access, and logistical (limited conventional laboratory facilities, limited access to spectrometers, limited trained technicians). Studies with such limitations have to make the best out of the limited data that can be collected and analyzed. However, local calibrations with small sample sizes may be possible if soil variation is limited within a specific study area (Brown, 2007).

Small sample sizes in a particular study are not a problem if there is a calibrated spectral library which includes soils similar to those collected in the new study (Shepherd and Walsh, 2002), but for many areas of the world, and for many soil types, such libraries do not exist.

Thus, the objective of this study was to test whether a locally-developed calibration model for SOC based on a limited number of samples can be developed within the context of a project with limited resources, in an area of limited access, and where no soil spectral library exists.

2. Material and methods

2.1. Study area

This study is part of the "Competing Claims on Natural Resources" project (Giller et al., 2008), centered on the trans-frontier national parks of the Mozambique–Zimbabwe–South Africa (RSA) border focused on the Limpopo National Park (LNP) of Mozambique with an area about 1.0 million ha (Stalmans et al., 2004). The park is located in a semi-arid climate, with annual precipitation between 380 (north) and 400 mm (south) and average max/min temperatures of 33/13 °C (south) and 35/15 °C (north) (INGC et al., 2003; Reddy, 1984).

The LNP is covered by savanna vegetation type made up of 15 plant communities, whose combinations resulted in eight major landscape units (Stalmans et al., 2004) (Fig. 1): Combretum/Mopane Rugged Veld (CMR), Limpopo Levubu Floodplains (LLF), Limpopo north (LN), Mixed Combretum/Mopane woodland (MCM), Mopane Shrubveld on Calcrete (MSC), Nwambia sandveld (NS), Pumbe Sandveld (PS), *Salvadora angustifolia* Floodpalins (SAF). Land use is primarily conservation area, with some villages (mostly planned to be relocated) practicing low-input, subsistence farming, with rainfed maize the primary crop and large herds of foraging domestic cattle. The characteristics of these landscape units in terms of geology (Manninen et al., 2008; Rutten et al., 2008), soils (FAO and Unesco, 1997) and land cover/use (Cenacarta, 1997) are presented in Table 1.

The park has only a few improved roads, and access is quite difficult, especially off-road, due to dense vegetation, rough ground, and large wild animals.

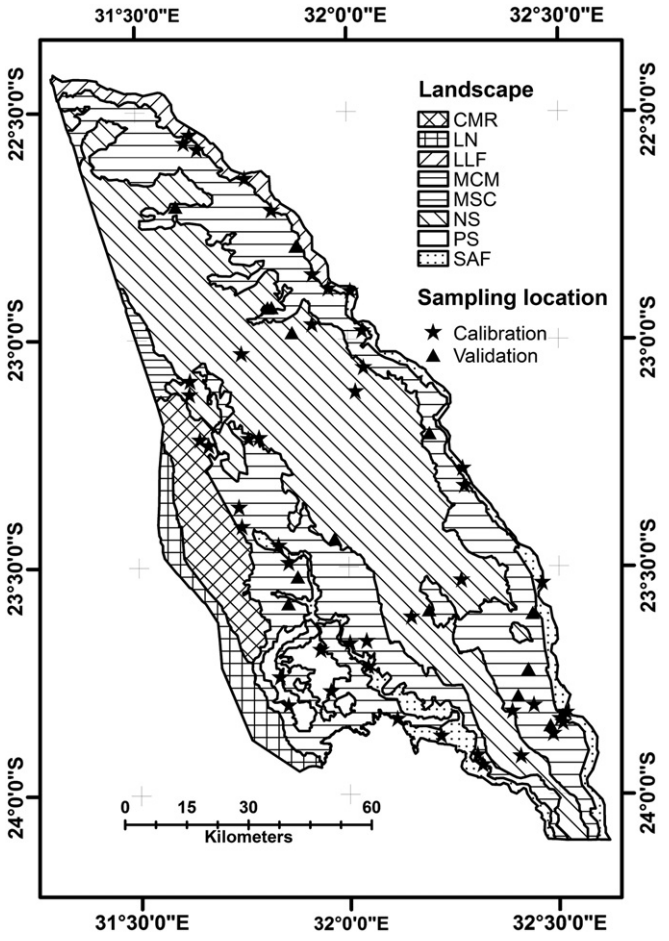

Fig. 1. The eight major landscape units of Limpopo National Park and sampling locations for model calibration (stars) and validation (triangles).
Redrawn from Stalmans et al. (2004), with permission from Koedeo.

2.2. Soil sampling and spectral acquisition

As part of a project to determine the spatial variability and map SOC in LNP by digital soil mapping (DSM) techniques, 412 soil samples (2/3 for calibration and 1/3 for validation) were collected following a stratified clustered random sampling for its high operational efficiency (De Gruijter et al., 2006). We refer to these as the "DSM calibration" and "DSM validation" samples. The DSM calibration clusters proportionally represented all accessible (<2.5 km from the road network) LNP landscape units and were randomized within each unit, whereas the DSM validation clusters were randomly collected across

Table 1
Main geological, soil and land cover occurring in the different landscape units in the study area.

Landscape	Geology	Major soil groupings	Vegetation/land cover
CMR	Rhyolite/basalt	Eutric leptosols	Grasslands
LLF	Fluvial sediments	Eutric fluvisols	Broadleaved deciduous trees
MSC	Sandstone/limestones	Calcaric cambisols	Broadleaved deciduous woodlands
NS	Aeolian sand	Arenosols/haplic luvisols	Open shrubland
PS	Aeolian sand	Ferralic arenosols	Open shrubland
SAF	Fluvial sediments	Eutric fluvisols	Open to closed shrubland (Riverine)
LN	Rhyolite/basalt	Eutric leptosols	Grasslands
MCM	Rhyolite/basalt	Eutric leptosols	Grasslands/broadleaved deciduous woodlands

the whole LNP. A DSM cluster was defined as two orthogonal and midway crossing transects of 720 and 360 m containing seven sampling points at 180 m intervals. Each sample was a composite of five sub-samples from the four corners of a 90 m square plus the center. Each sub-sample was from a (variable-thickness) field-identified A horizon, collected with a hand shovel after scooping out the upper 2–5 cm (to remove sticks, undecomposed leaves, etc.). Subsamples were thoroughly mixed in a bucket, then about a half a kg was collected in a plastic bag and sent to soil laboratory.

Samples were air-dried, gently crushed and passed through a 2 mm mesh sieve to collect the fine earth fraction. Samples were put in Petri dishes and then scanned in a Bruker FR-NIR MultiPurpose Analyser (MPA), (Bruker optic GmbH, Ettlingen, Germany) located in the Instituto de Investigação Agronómica de Moçambique (IIAM), Maputo. This instrument has built-in validation to perform instrument internal (operational and performance) quantification tests, and its spectrum is calibrated before each scan to an internal gold reference. Spectra were recorded from 0.8 to 2.6 μm at a spectral resolution of 1250 μm, with zero-filling factor of 2, resulting in an effective bandwidth of 3.86 μm. Each spectrum is an average of 64 scans. Spectra were further reduced to the range 1.25–2.5 μm as these bands contain most of relevant information.

2.3. Selection of soil samples for reference analysis

To form a subset of samples for reference analysis, a total of 129 samples were selected from both DSM calibration (104) and validation (25) sample sets as described in the previous section. The samples represent one third and one fourth of DSM calibration and validation sets, respectively. These proportions are commonly used in laboratory spectroscopy (Brown et al., 2005; Grinand et al., 2008). Reference samples from DSM calibration were used for model calibration (about four fifth) and those from DSM validation, used for model validation (about one fifth); note we refer to these as "model" calibration and validation, as opposed to the "DSM" calibration and validation sets from field sampling. To select a representative set covering the range of spectra and SOC contents, the spectra were compressed using principal component analysis (PCA), to summarize the data and examine its structure. The PCA scores were grouped by computing a K-means clustering in the Unscrambler 9.7 program (CAMO Software AS, Nedre Vollgate, Oslo, Norway). The number of groups was determined iteratively to minimize the sum of distances (SOD). Samples were randomly chosen from the different groups as suggested by Martens and Naes (1986) in order to enhance sample set diversity (Stenberg et al., 1995). Samples were then drawn from these groups, excluding any that met any of the following three conditions: (1) high residuals and low leverage, (2) both high residual and leverages or (3) high leverages and away from the PCA model trend, were considered outliers and not considered for laboratory analysis (Esbensen, 1994). Outliers as thus defined were automatically flagged based on the default threshold values in Unscrambler 9.7.

2.4. Laboratory analysis

The selected samples were analyzed in the soils laboratory of Eduardo Mondlane University, Maputo, for SOC and possible covariable predictors soil pH and particle size fractions, following standard ISRIC methods for soil laboratory analysis (van Reeuwijk, 2002). SOC was determined by the Walkley–Black method. Soil pH was measured potentiometrically in a supernatant suspension of 1:2.5 soil: liquid mixture (two determinations: in distilled water and 1 M KCl solution). Particle-size separates of the fine earth (<2 mm) fraction were determined after cementing agents were first removed by means of hydrogen peroxide, calgon and calcium chloride solution. The sand fraction (2 mm–50 μm) was washed onto a 50 μm sieve, after which silt (50 μm–2 μm) and clay (<2 μm) fractions were determined by hydrometer method. Twenty randomly-selected samples were analyzed in duplicate for quality control and to quantify laboratory precision.

2.5. Calibration and validation

Mathematical and statistical procedures were carried out in the R environment for statistical computing (Ihaka and Gentleman, 1996), Unscrambler (CAMO Software AS, Nedre Vollgate, Oslo, Norway), and ParLes (Viscarra Rossel, 2008). The "pls" package was used within R for multivariate calibration (Mevik and Wehrens, 2007).

PLSR was used to develop models based on spectra and reference laboratory data of the 129 selected soil samples. Models were evaluated in two ways: (1) "leave-one-out" cross-validation on models developed with all 129 samples and (2) true validation by splitting the sampling set into spectral calibration (104 samples) and validation (25 samples) sets. The former was used to search for the best pre-processing of the raw spectra and the latter to obtain realistic estimates of prediction accuracy.

Models were attempted with the original spectra, multiplicative scatter corrected (MSC) spectra, first derivatives of these; and all of these also after applying a Savitsky–Golay filter (2nd order polynomial covering 11 adjacent bands). MSC was applied since the original spectra showed additive effects which could result from differential scattering in the granular sample. The derivative transformation minimizes the effect of variation in sample grinding and optical set-up (Shepherd and Walsh, 2002). Transformations of the laboratory measurements were also attempted but did not improve results and therefore are not reported.

Model calibration accuracy was evaluated by means of the root-mean squared error of calibration (RMSE) of the cross-validation predictions, and R^2 (proportion of variation explained) of the SOC vector. Because the resulting model was intended to be used to map SOC across the LNP, we did not have the option of rejecting any observations as outliers. Prediction accuracy was assessed by the ratio of standard deviation (SD) to RMSE of cross-validation (RPD) and by the multiple R^2 (Chang et al., 2001; Waiser et al., 2007).

In an attempt to improve the predictive performance of the best PLSR model and following a suggestion by Fearn (2010), the proportion

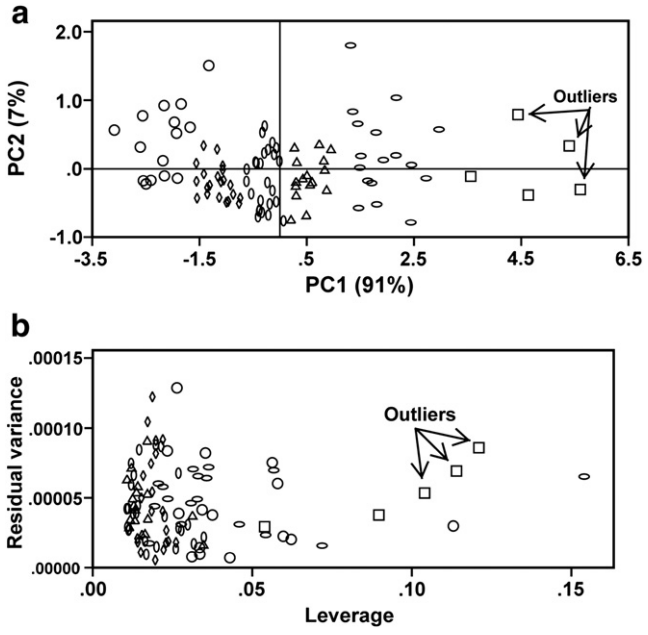

Fig. 2. Score plot of the first two principal components principal of spectra from DSM calibration set symbolized by their cluster (a) and sample influence plot (b) used to select samples for reference laboratory analysis.

Table 2
Summary statistics for 129 soil samples set submitted for reference laboratory analysis. Included is the correlation coefficient (r) of duplicate samples as well as the RMSE from 1:1 line.

Summary statistics (N = 129)	Soil property					
	pH water	pH KCl	SOC (%)	Clay (%)	Silt (%)	Sand (%)
Minimum	3.7	3.4	0.0	5.3	0.0	2.2
1st quartile	5.7	5.3	0.4	8.7	1.8	74.4
Median	6.7	6.2	0.7	13.9	4.6	81.3
Mean	6.5	6.1	0.9	14.4	7.4	78.3
3rd quartile	7.6	7.3	1.2	17.3	8.7	88.8
Maximum	9.0	8.1	2.7	47.3	50.5	93.5
SD	1.1	1.2	0.6	7.1	8.5	14.4
r (duplicates, n = 20)	0.96	0.99	0.97	0.89	0.93	0.98
RMSE (from 1:1 line)	0.32	0.17	0.13	2.6	3.2	2.9

of clay in the fine earth – a commonly used covariate for SOC spatial interpolation (McGrath and Zhang, 2003; Mutuo et al., 2006), – was added to the spectra information as a supplementary "band", and a new PLSR model built. Clay proportion may be helpful for PLSR modeling in cases where a collection of samples and its laboratory analysis results for clay (but not SOC) from past surveys is kept. For new surveys the laboratory costs of determining clay and SOC are comparable, so these models are not relevant. Clay variance was inflated 86 times to match the true dimensionality of spectra predictors so that it could be properly weighted, on the basis that previous PLSR model explained most of SOC variation with about 12 factors (as shown in the results, below); $12 \times 86 = 1037$, the approximate number of spectral bands.

3. Results and discussion

3.1. Sample selection for reference analysis

Fig. 2 shows the sample selection procedure for reference analysis from the DSM validation set. The first two principal components explained 98% (91 + 7%) of spectra variation; the SOD was achieved for 6 clusters. The influence plot was used to identify outliers. The same procedure was followed for the DSM calibration set, where the first two PC's explained 98% (95 + 3%) of spectra variation and the minimum SOD was achieved for 12 clusters. As per the sampling plan 25% (= 25) of DSM validation and 1/3 (= 104) of the DSM calibration sets were selected for reference analysis. Since the number of PCA score groups (intended to represent spectra variability) was small, more than one sample was selected from most of them. The small number of groups indicates the fairly homogeneous nature of the sample sets. This procedure has also been followed by other workers (Viscarra Rossel and Behrens, 2010) to select samples based on spectral variability. Whereas PCA score grouping enhances spectral diversity, it may also enhance the spatial autocorrelation between the selected samples due to possible coincidence of PCA score groups with the field sample clusters. This raises the possibility of false precision (Brown et al., 2005). However, RPD (see Section 3.5, below) suggests that this effect is minimal in our case.

3.2. Soil properties

The summary statistics of laboratory analysis (Table 2, Fig. 3) show a fairly wide range for SOC in this semi-arid environment, from below the detection limit to moderate values (2.7%), thus providing a good range for model calibration. Soils range from quite acid to alkaline, with a somewhat left-skewed distribution emphasizing the alkaline range. Most are coarse-textured. The empirical distributions of SOC, clay and silt appear positively skewed while that of the sand fraction is negatively skewed. Parametric correlations between duplicates were all linear and generally very good. Laboratory duplicate RMSE's (on the expected 1:1 line) were low, indicating good analytical precision. The moderate precision for particle sizes matches the expected precision of the hydrometer method. These RMSE set an upper limit on the precision of any calibration. Bivariate correlations between soil properties showed positive correlations between SOC and pH_{H20} (0.50), pH_{KCl} (0.49), clay (0.56) and a negative correlation between SOC sand (−0.65), all significant at 0.01 level. These results are expected: finer-textured soils retain moisture longer, and neutral to alkaline soils generally support more soil microorganisms and more vigorous vegetation (hence more leaf litter); both of these situations are conducive to higher levels of SOC.

3.3. Laboratory SOC vs landscape units

Mean SOC per landscape units is highest in CMR (2.00%), decreasing through PS (1.15%), MSC (0.95%), SAF (0.91%), LLF (0.60%) and NS (0.45%). Note that the laboratory RMSE for SOC (0.13%) is a significant proportion of the low-SOC landscape units. Duncan's multiple-range test shows that CMR is clearly separated from all other landscape units, NS and LLF are grouped at the lower end and cannot be separated from the grouping of MSC and SAF. This group cannot also be statistically separated from PS, due to the wide ranges of SOC for MSC, SAF and NS (Fig. 4). ANOVA shows that landscapes explain about 24% of the total SOC variation. Thus SOC is rather similar over most of the landscape, which should reduce prediction errors due to the small sample size. Separate analysis per landscape is in any case not possible because of the limited number of samples; this result shows that such an analysis would be unlikely to result in different models.

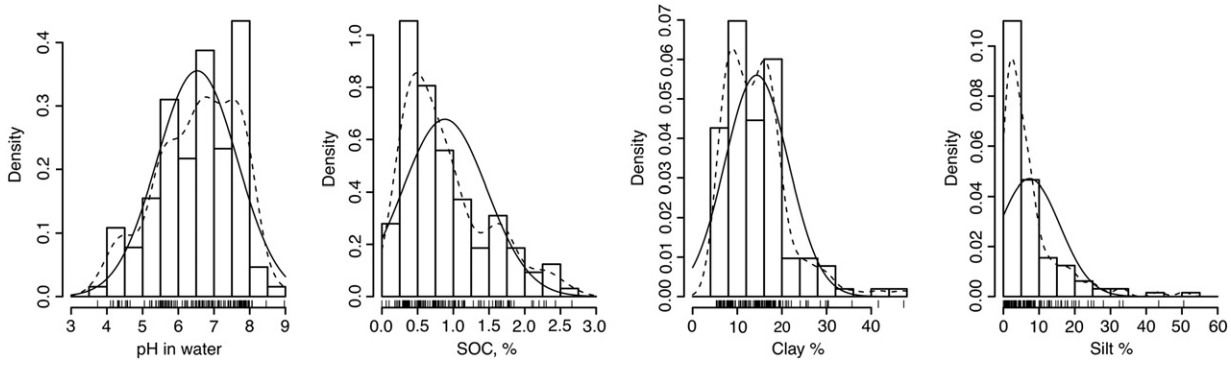

Fig. 3. Distribution of soil properties in laboratory samples; bars = histogram, dashed line = density and fine dashed line = normal fit, Soil fractions units in percentages. Rug marks along the x-axis show individual sample location.

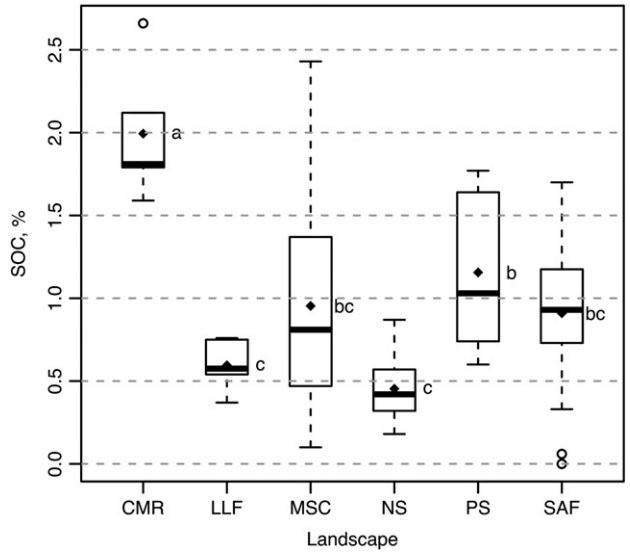

Fig. 4. Box-plots and Duncan's multiple range test (alfa = 0.05 and Df = 117) for the SOC per landscape unit.

3.4. Spectral features

The raw spectra (Fig. 5) generally showed the typical pattern of soil spectra, with three major absorption features around 1.37–1.46, 1.86–2.06 and 2.14–2.26 μm. The first absorption region (near 1.4 μm) is the first overtone of OH stretches (moisture adsorbed to the clay surfaces) and near 1.9 μm it is the combination of OH stretches and H – O – H bend in water molecules trapped in the crystal lattice (not present in for example well developed dried kaolinites). Near 2.2 μm it is OH-metal bend and OH stretch combinations where the metals can be Al or Fe or Mg substituting Si (Fe and Mg closer to 2.3 μm) (Clark et al., 1990). In addition, a number of spectra showed two noisy (or fluctuating) reflection regions around 1.34–1.39 and 1.79–1.92 μm. These ranges overlap with the first two absorption features. These raw spectra are similar to those found by other authors, *e.g.* Ben-Dor and Banin (1995) and Ben-Dor et al. (1999). The SOC component normally affects the overall positioning and shape of the spectrum (Shepherd and Walsh, 2007).

3.5. Prediction of SOC from NIR spectra

Since there is no *a priori* way to determine which spectra pre-processing methods result in the best predictive model (Ben-Dor and Banin, 1995), a number of spectra pre-processing methods were compared (Table 2). Pre-processed spectra showed peaks at around the same wavelength ranges as the raw spectra, regardless of pre-processing method.

The best PLSR model (Fig. 6, MSC smoothed and 1st derivative) for the prediction of SOC in the LNP was obtained with nine factors with a RMSEP of about 2.5 times that obtained from laboratory analysis on duplicate samples. The model also explained 99.5% of spectra variance. The median cross-validation residual was −0.0035%, interquartile range (IQR) −0.015 to +0.013%, but there were some very poorly-modeled points, at the extremes, −1.25 and +1.75% SOC. The loadings of the first two model factors explained 95.1% of the spectral variation. The other pre-processing methods (Fig. 6) resulted in PLSR models with RMSE slightly higher (0.36 to 0.32% SOC) and therefore lower SOC explained variation. In addition about half of these models suffered from non-linearity effects expressed in the form of "banana-like" trends, causing underprediction for the highest values. The 5% absolute extreme values of best model's regression coefficients (Fig. 8) show regions that were important for SOC predictions. These regions are in good agreement with those where assigned SOC spectral features are located.

However, these interpretations should be considered with caution given the fact that they are made on the base of pre-processed spectra (MSC smoothed 1st derivative spectra), which may not be as useful as if the analysis would have been carried out on the base of raw spectra. In derivative spectra following the derivatives, peaks will occur at maximum slopes of the original spectra and the original peaks will occur as crossing the zero line. Thus, in the derived spectrum each original peak will be represented by one positive and one negative peak.

The MSC minimized the amplification and offset effects of light scattering in the raw spectra, which resulted in PLSR calibration improvement. Shepherd and Walsh (2002) preferred the first derivative pre-processing technique to MSC, as the latter did not improve multivariate adaptive regression tree (MART) calibration. The first derivative is the most commonly applied transformation to minimize variation among samples caused by variation in grinding and optical set-up (Stenberg et al., 2010). MSC is not preferred by many authors because

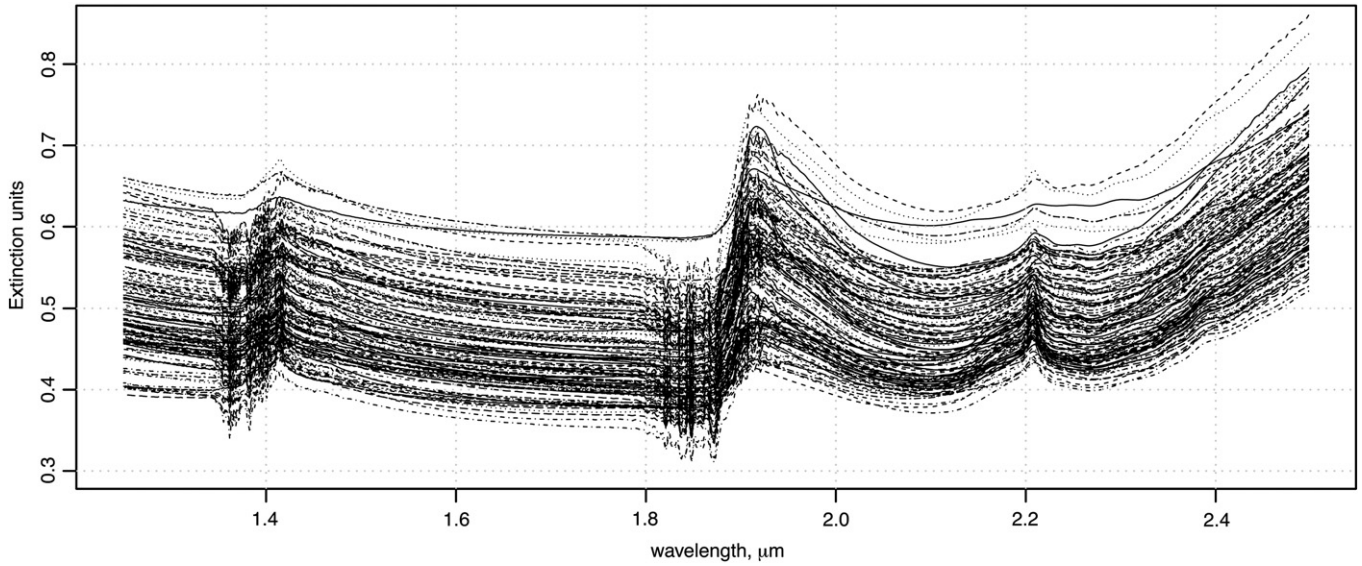

Fig. 5. Spectra of the soil samples showing three major absorption features, related to OH groups in both absorbed water (≈1.4 and 1.9 μm) and the crystal lattice water (≈2.2 μm), (Ben-Dor and Banin, 1995).

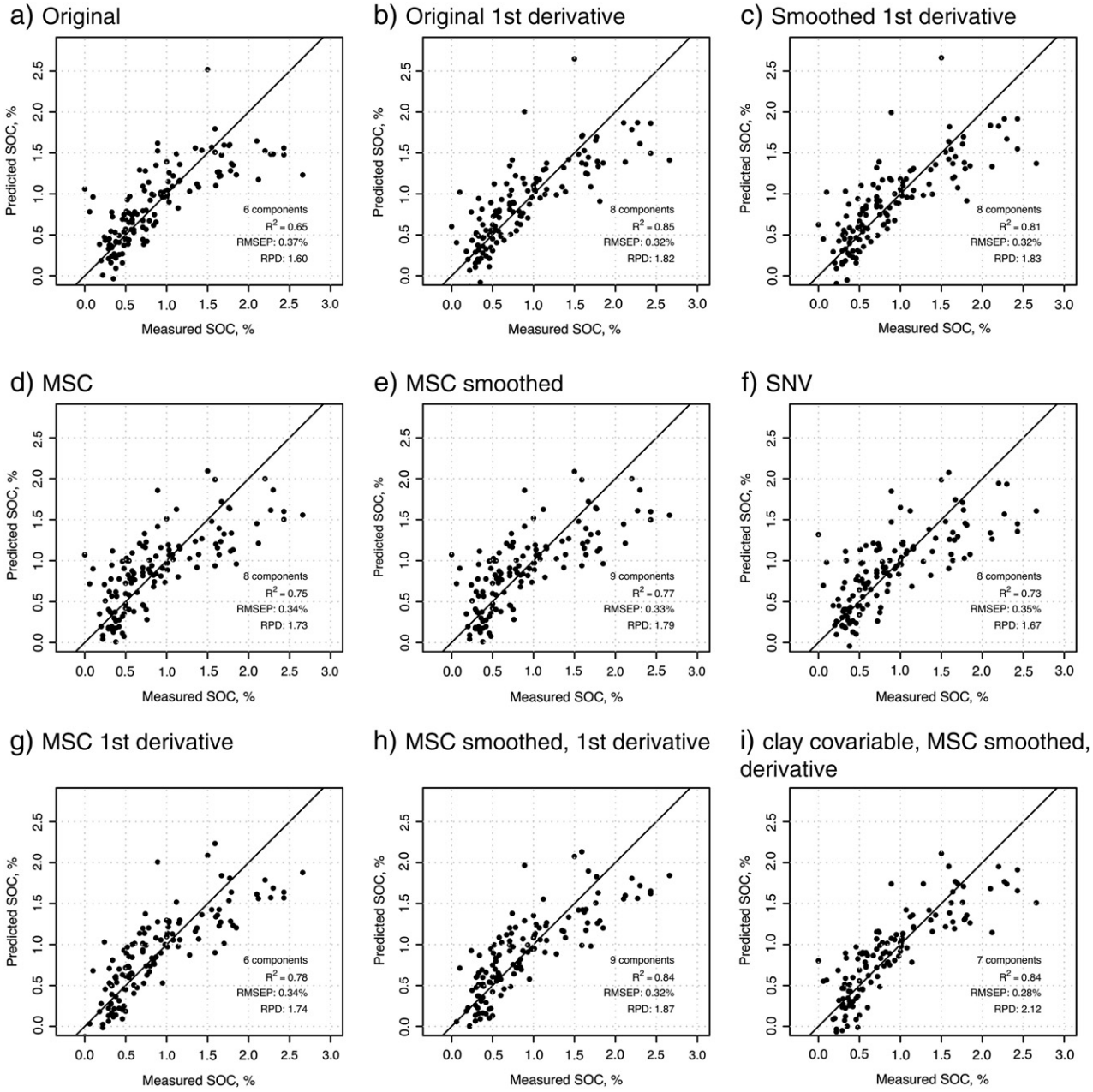

Fig. 6. SOC PLSR prediction models derived as a function of spectra pre-processing method; (a)–(h) which included a combination of raw (original)/smoothed spectra, multiplicative scatter correction (MSC), standard normal variate (SNV), and 1st derivative. (i) shows SOC PLSR model derived from inclusion of clay covariable in the predictor set.

it is difficult to locate an adequate spectral range to apply, raising the risk of affecting relevant spectral features for the component of interest (Esbensen, 1994).

Despite the acceptable model reliability, the proportion of RMSEP to the mean SOC of sample set is substantial, about 36%. Literature shows this proportion varies considerably. For example Fidêncio et al. (2002) determined SOM by radial basis function networks and NIR spectroscopy and found a proportion of RMSEP to mean SOC of between 9 and 108%, Brown et al. (2006) obtained a proportion around 265%, and Terhoeven-Urselmans et al. (2010) of about 190%. Better proportions were obtained by Shepherd and Walsh (2002), about 18%, Fystro (2002) about 20%, and Wetterlind et al. (2008) about 8%. Most of the high proportions are from studies covering large areas as does the present study, which suggests room for further improvement by spiking (Guerrero et al., 2010), *i.e.*, inclusion of a few local samples. Our results are between the local and large-area studies, hence our characterization of our results as "regional".

Although the best model found in the present study fitted well the 1:1 validation line, all eight observations with SOC concentrations above 2.0% were under-predicted. In addition there were three observations with moderate SOC concentrations but large negative residuals (over-predictions). The cause for these poor predictions was investigated by plotting the SOC against pH and clay proportion. Clay did not give an obvious explanation as it spanned a wide range for the poorly-predicted samples. The pH was in the range 6–7.5 for the underpredicted samples and around 8 for overpredicted ones. The pH range 7.8–8.4 is often indicative of carbonate presence (Schumacher, 2002). However we did not observe any effervescence after addition of HCl 10% to the samples. There was also no apparent relation between landscape units and poor predictions (Fig. 7).

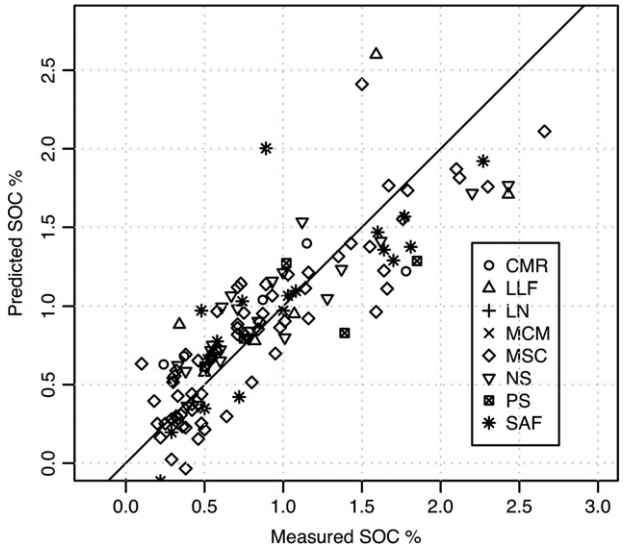

Fig. 7. The best PLSR prediction model showing the samples symbolized by the landscape unit (Stalmans et al., 2004) from where they were collected.

A scatterplot of SOC against clay or clay + silt revealed a fairly strong relation at lower values, which degraded above about 18% clay or 25% clay + silt. This disagrees with results reported by Stenberg (2010), who found that prediction of SOM could be substantially improved by removing the sandiest soils.

The wavelengths contributing most for the best model in the present study are near 1.4, 1.9 and 2.2 μm, which correspond to OH groups of soil moisture (first two) to the crystal lattice in soil clay minerals (last) (Ben-Dor and Banin, 1995) (Fig. 8). Although the latter do overlap with assigned wavelength for the determination of the alkaline-earth carbonates, calcite and dolomite by near infrared spectroscopy, it was not possible to identify them, possibly because carbonate content was not detected (far below the 10% weight basis threshold) and that samples were not pre-heated to 600 °C for 8 h in order to remove the strong absorption features of OH groups both in the organic matter and clay minerals, to enhance CO_3 features (Ben-Dor and Banin, 1990).

Ben-Dor and Banin (1995) identified the 1.4 and 1.9 μm bands as important for prediction of soil organic matter, while they are at the same time characteristic for OH and water molecules. This confirms the difficulty in identifying with confidence the spectral ranges characteristics for different compounds (Ben-Dor et al., 1999; Brown et al., 2006; Clark, 1999; Stenberg, 2010).

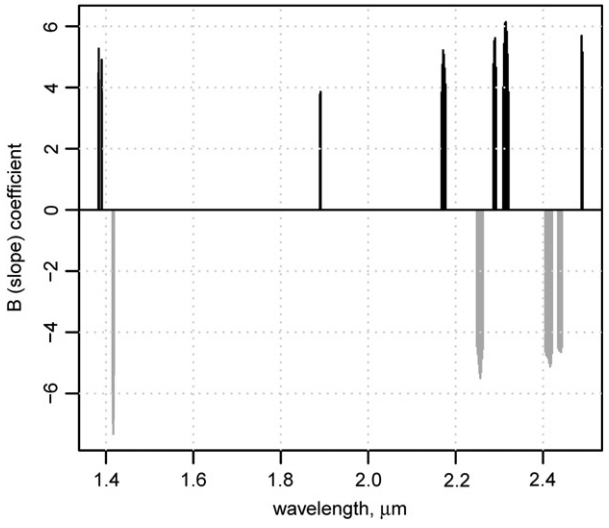

Fig. 8. The 5% extreme values of PLSR coefficients (2.5% positive and 2.5% negative) of best model, showing the most contributing wavelengths ranges for model predictions.

PLSR is a data compression method that summarizes most of variables' variance in a few factors and by so doing helps to reveal hidden patterns in the data (Esbensen, 1994). The analysis was performed here on the pre-processed spectra to help explain whether landscape units may have influence on the model prediction ability and therefore explain its poor performance for some of the samples. The score plot of the first three PLSR components (factors) did not reveal landscape-related pattern, except for the LLF (Limpopo Levubu Floodplains) landscape unit which did follow a specific pattern, but samples collected in this unit were not a problem for the prediction model. Thus there are groups of similar samples but these did not separate under- from over-predictions.

The normal probability plot of SOC residuals suggests that the PLSR model may still have some non-linearity, as the sample residuals at both ends slightly deviated from the tails of the normal distribution. All the under-predicted samples are located at the upper end of this plot while, surprisingly, the over-predicted ones do fall within the linear range of the plot.

3.6. Calibration subset models

The best model form (smoothed first derivative of MSC-corrected spectra) was fit to the 104-observation DSM calibration subsample. A nine-component PLSR model had an internal cross-validation RMSEP of 0.323% SOC, just a little worse than the model from the full set, 0.315%. Predictions from this model for the 25-observation spectral validation had errors from −0.50 to +0.65%SOC, with a median of −0.10% and inter-quartile range (IQR) from −0.22 to +0.27%; compared to cross-validation errors these are much lower extremes but wider IQR. The true validation RMSEP was 0.331%, just a bit higher than the full-set cross-validation RMSEP of 0.315%.

This shows that (1) cross-validation gives a realistic estimate of the true validation error, (2) the model built from DSM calibration spectra only is a little less accurate than that built from all spectra; (3) the 104/25 split fairly reflects model performance; (4) the DSM calibration and validation samples have similar characteristics.

3.7. Prediction of SOC from NIR spectra and clay

The PLSR model based on the NIR + clay (Fig. 6), following same spectra pre-treatment, shows some improvement compared to that based on the NIR spectra only. The best model now contained only seven factors (Fig. 6 (i)), explaining 100.0% of clay + spectra and 84.0% of SOC variances, with a RMSEP of 0.28% SOC, about 0.04% better than the model without the covariate and slightly above twice as much as that obtained for laboratory analysis on duplicate samples. Almost all clay + spectra variance is explained by the first factor, while this component explains about 32% of SOC. The remaining 52% of explained variance attained at the seventh factor of the model is generated by a cumulative <1% clay + spectra variance. This result is not surprising, given the generally good relation between clay and SOC in this sample set, and the strong diagnostic features in clay spectra.

This result agrees with that of Brown et al. (2006), who showed that the inclusion of sand fraction and soil pH as auxiliary predictors improved calibrations.

4. Conclusions

Using only 129 samples combined from the different landscape units of the LNP resulted in a fairly stable, effective NIR PLSR calibration model for SOC prediction in the target area. The model predicted fairly well irrespective of landscape unit. However, model performance was limited at higher SOC concentrations. The stable and effective

model here obtained from a limited number of samples shows that reasonable models can be built for areas of limited access, where a limited number of representative samples can be collected, as it is the case of LNP.

The addition of a moderately-correlated covariable (here, clay concentration) in the set of predictors slightly improved the precision (RMSE). This is of interest in the case where there are stored samples where particle-size has been analyzed in the lab; these samples may now be scanned and the developed predictive equations used to estimate SOC.

Despite the improvement of model accuracy by inclusion of clay, errors are still a substantial proportion of mean prediction. This suggests that caution must be considered when using spectroscopy to estimate SOC for mapping or monitoring low-SOC landscapes. While the model has a potential for SOC prediction in regional and baseline studies, it can be improved further for detailed ecological and farm-level studies within the LNP or in similar nearby soil landscapes by recalibrating the model after adding a few "local" samples (spiking).

Acknowledgments

The authors thankfully acknowledge funding from International Research and Education Fund (INREF) of the Wageningen University through the "Competing Claims on Natural Resources Programme" and from The International Institute for Geo-information Science and Earth Observation of the University of Twente. The training support in laboratory soil spectroscopy analysis from the soil team at ICRAF-Nairobi is highly appreciated.

References

Arshad, M.A., Martin, S., 2002. Identifying critical limits for soil quality indicators in agro-ecosystems. Agriculture, Ecosystems & Environment 88 (2), 153–160.
Ben-Dor, E., 2002. Quantitative remote sensing of soil properties. Advances in Agronomy. Academic Press, pp. 173–243.
Ben-Dor, E., Banin, A., 1990. Near-infrared reflectance analysis of carbonate concentration in soils. Applied Spectroscopy 44, 1064–1069.
Ben-Dor, E., Banin, A., 1995. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Science Society of America Journal 59, 364–372.
Ben-Dor, E., Irons, J.R., Epema, G.F., 1999. Soil reflectance. In: Rencz, A.N. (Ed.), Remote Sensing for Earth Sciences. Manual of Remote Sensing. John Wiley & Sons, Inc., Toronto, pp. 111–188.
Bouma, J., 2002. Land quality indicators of sustainable land management across scales. Agriculture, Ecosystems & Environment 88 (2), 129–136.
Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140 (4), 444–453.
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., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132 (3–4), 273–290.
Cécillon, L., et al., 2009a. Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). European Journal of Soil Science 60 (5), 770–784.
Cécillon, L., et al., 2009b. Predicting soil quality indices with near infrared analysis in a wildfire chronosequence. Science of the Total Environment 407 (3), 1200–1205.
Cenacarta, 1997. Carta de Uso e Cobertura da Terra. Ministério da Agricultura, Maputo.
Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh Jr., C.R., 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. Soil Science Society of America Journal 65 (2), 480–490.
Clark, R.N., 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy. In: Rencz, A.N. (Ed.), Remote Sensing for Earth Sciences. Manual of Remote Sensing. John Wiley & Sons, Inc., Toronto, pp. 3–58.
Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G.A., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. Journal of Geophysical Research 95 (B8), 12653–12680.
De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, Berlin.
Esbensen, K., 1994. Multivariate Data Analysis in Practice. CAMO Software AS, Oslo. 598 pp.
FAO and Unesco, 1997. Mapa dos Solos do Mundo - Legenda revista. FAO, Roma.
Fearn, T., 2010. Combining other predictors with NIR spectra. Chemometric Space 21 (2), 13–16.
Fidêncio, P.H., Poppi, R.J., de Andrade, J.C., 2002. Determination of organic matter in soils using radial basis function networks and near infrared spectroscopy. Analytica Chimica Acta 453 (1), 125–134.
Fystro, G., 2002. The prediction of C and N content and their potential mineralisation in heterogeneous soil samples using Vis-NIR spectroscopy and comparative methods. Plant and Soil 246 (2), 139–149.
Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Analytica Chimica Acta 185, 1–17.
Giller, K.E., et al., 2008. Competing claims on natural resources: what role for science? Ecology and Society 13 (2), 18.
Gisladottir, G., Stocking, M., 2005. Land degradation control and its global environmental benefits. Land Degradation and Development 16, 99–112.
Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143 (1–2), 180–190.
Guerrero, C., Zornoza, R., Gómez, I., Mataix-Beneyto, J., 2010. Spiking of NIR regional models using samples from target sites: effect of model size on prediction accuracy. Geoderma 158 (1–2), 66–77.
Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5 (3), 299–314.
INGC, UEM, Fews-NET MIND, 2003. Atlas for Disaster Preparedness and Response in the Limpopo Basin. Instituto Nacional de Gestao de Calamidades (INGC). 99 pp.
Janssen, B.H., et al., 1990. A system for quantitative evaluation of the fertility of tropical soils (QUEFTS). Geoderma 46 (4), 299–318.
Liu, D., et al., 2006. Spatial distribution of soil organic carbon and analysis of related factors in croplands of the black soil region, Northeast China. Agriculture, Ecosystems & Environment 113 (1–4), 73–81.
Manninen, T., et al., 2008. The Karoo volcanic rocks and related intrusions in southern and central Mozambique. Geological Survey of Finland Special Paper 48, 211–250.
Martens, H., Naes, T., 1986. Multivariate Calibration. John Wiley & Sons, Chichester.
McGrath, D., Zhang, C., 2003. Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Applied Geochemistry 18 (10), 1629–1639.
Mevik, B.-H., Wehrens, R., 2007. The pls package: principal component and partial least squares regression in R. Journal of Statistical Software 18 (2), 1–24.
Mutuo, P.K., Shepherd, K.D., Albrecht, A., Cadisch, G., 2006. Prediction of carbon mineralization rates from different soil physical fractions using diffuse reflectance spectroscopy. Soil Biology and Biochemistry 38 (7), 1658–1664.
O'Rourke, S.M., Holden, N.M., 2011. Optical sensing and chemometric analysis of soil organic carbon — a cost effective alternative to conventional laboratory methods? Soil Use and Management 27 (2), 143–155.
Pathak, H., et al., 2003. Modelling the quantitative evaluation of soil nutrient supply, nutrient use efficiency, and fertilizer requirements of wheat in India. Nutrient Cycling in Agroecosystems 65 (2), 105–113.
Reddy, S.J., 1984. General Climate of Mozambique. INIA, Maputo.
Rutten, R., Makitie, H., Vuori, S., Marques, J.M., 2008. Sedimentary rocks of the Mapai formation in the Massingir–Mapai region, Gaza province, Mozambique. Geological Survey of Finland Special Paper 48, 251–262.
Schumacher, B.A., 2002. Methods for Determination of Total Organic Carbon in Soils and Sediments. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-02/069 (NTIS PB2003-100822).
Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal 66 (3), 988–998.
Shepherd, K.D., Walsh, M.G., 2007. Infrared spectroscopy — enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. Journal of Near Infrared Spectroscopy 15 (1), 1–19.
Shukla, M.K., Lal, R., Ebinger, M., 2006. Determining soil quality indicators by factor analysis. Soil and Tillage Research 87 (2), 194–204.
Smaling, E.M.A., Janssen, B.H., 1993. Calibration of quefts, a model predicting nutrient uptake and yields from chemical soil fertility indices. Geoderma 59 (1–4), 21–44.
Stalmans, M., Gertenbach, W.P.D., Carvalho-Serfontein, F., 2004. Plant communities and landscapes of the Parque nacional do Limpopo, Mocambique. Koedoe 47 (2), 61–81.
Stenberg, B., 2010. Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon. Geoderma 158 (1–2), 15–22.
Stenberg, B., Nordkvist, E., Salomonsson, L., 1995. Use of near infrared reflectance spectra of soils for objective selection of samples. Soil Science 159 (2), 109–114.
Stenberg, B., Viscarra-Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. In: Donald, L.S. (Ed.), Advances in Agronomy. Academic Press, pp. 163–215.
Terhoeven-Urselmans, T., Vagen, T.-G., Spaargaren, O., Shepherd, K.D., 2010. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. Soil Science Society of America Journal 74 (5), 1792–1799.
van Reeuwijk, L.P., 2002. Procedures for soil analysis. ISRIC Technical Paper, p. 9.
Viscarra Rossel, R.A., 2008. ParLeS software for chemometric analysis of spectroscopic data. Chemometrics and Intelligent Laboratory Systems 90, 72–83.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1–2), 46–54.
Viscarra Rossel, R.A., McGlynn, R.N., McBratney, A.B., 2006. Determining the composition of mineral-organic mixes using UV–vis-NIR diffuse reflectance spectroscopy. Geoderma 137 (1–2), 70–82.
Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T., 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Science Society of America Journal 71 (2), 389–396.
Wetterlind, J., Stenberg, B., Söderström, M., 2008. The use of near infrared (NIR) spectroscopy to improve soil mapping at the farm scale. Precision Agriculture 9 (1), 57–69.
Yemefack, M., Jetten, V.G., Rossiter, D.G., 2006. Developing a minimum data set for characterizing soil dynamics in shifting cultivation systems. Soil and Tillage Research 86 (1), 84–98.