European Journal of Soil Science

European Journal of Soil Science, September 2014, **65**, 718–729 doi: 10.1111/ejss.12165

Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques

S. R. Araújo[a], J. Wetterlind[b], J. A. M. Demattê[a] & B. Stenberg[b]

[a]*Department of Soil Science, University of São Paulo, Av. Pádua Dias 11, 13418-260 Piracicaba, Brazil , and* [b]*Department of Soil and Environment, Swedish University of Agricultural Sciences, Gråbrödragatan 19, PO box 234, 532 23 Skara, Sweden*

Summary

Effective agricultural planning requires basic soil information. In recent decades visible near-infrared diffuse reflectance spectroscopy (vis-NIR) has been shown to be a viable alternative for rapidly analysing soil properties. We studied 7172 samples of seven different soil types collected from several regions of Brazil and varying in organic matter (OM) (0.2–10.3%) and clay content (0.2–99.0%). The aim was to explore the possibility of enhancing the performance of vis-NIR data in predicting organic matter and clay content in this library by dividing it into smaller sub-libraries on the basis of their vis-NIR spectra. We used partial least square regression (PLSR) models on the sub-libraries and compared the results with PLSR and two non-linear calibration techniques, boosted regression trees (BT) and support vector machines (SVM) applied to the whole library. The whole library calibrations for clay performed well (*ME* (modelling efficiency) > 0.82; RMSE (root mean squared error) < 10.9%), reflecting the influence of the direct spectral responses of this property in the vis-NIR range. Calibrations for OM were reasonably good, especially in view of the very small variation in this property (*ME* > 0.60; RMSE < 0.55%). The best results were, however, found when dividing the large library into smaller subsets by using variation in the mean-normalized or first derivative spectra. This divided the global data set into clusters that were more uniform in mineralogy, regardless of geographical origin, and improved predictive performance. The best clustering method improved the RMSE in the validation to 8.6% clay and 0.47% OM, which corresponds to a 21% and 15% reduction, respectively, as compared with whole library PLSR. For the whole library, SVM performed almost equally well, reducing RMSE to 8.9% clay and 0.48% OM.

Introduction

The efficient use of soils in agriculture requires a good understanding of their chemical, physical, mineralogical and biological characteristics. Soil texture and organic matter (OM) are two important properties of soils. In tropical soils, which are typically characterized by very small OM contents, the clay content is regarded as an important fertility factor because of its positive effect on nutrient supply, structure and water retention. Clay content, together with OM, directly affects the porosity, plasticity and erodibility of soils. Both soil properties also affect soil fertility by their capacity for binding plant nutrients and water. Organic matter also plays an important role in the mineralization of nitrogen to plant-available forms.

Correspondence: B. Stenberg. E-mail: bo.stenberg@slu.se

Received 4 October 2013; revised version accepted 9 May 2014

The accuracy of spatial maps of soil attributes is correlated positively with the density of soil observations. However, methods used to determine texture and organic matter content in conventional soil laboratories in Brazil and elsewhere are expensive and time-consuming. Moreover, such analyses can generate wastes rich in sodium or chromium which may pose hazards to the environment. There is thus a need for more efficient methods to reduce the number of soil chemical analyses and generate high-resolution soil property maps over large areas at a reasonable cost. Visible and near-infrared (vis-NIR) diffuse reflectance spectroscopy (400–2500 nm) has received increasing attention over the last two decades as a promising technique for soil analysis (Stenberg *et al.*, 2010).

The absorption of vis-NIR light occurs because of overtones and combinations of fundamental molecular effects in the mid-infrared region and is associated with soil moisture, organic materials and mineralogy. Because the clay fraction as analysed by traditional methods consists mainly of minerals, vis–NIR spectra can be of

value for predicting clay content (Stenberg *et al.*, 2010). Organic matter content can be related directly to the absorption of vis-NIR spectra through a number of functional groups such as the carboxyls, hydroxyls and amines (Viscarra Rossel & Behrens, 2010). Nevertheless, the degree to which vis-NIR spectral data are capable of predicting OM content is variable between reports (Stenberg *et al.*, 2010).

Viscarra Rossel & Behrens (2010) pointed out that for spectroscopic techniques to be effective for analysing soils over large areas, there must be a wide range of spectroscopic data from different soil types with varying organic and inorganic components. Thus a large number of samples are required to cover the relevant variation and the cost of building predictive models has to be considered. According to Sankey *et al.* (2008), global to regional calibrations are more cost-effective, but they may not provide sufficient accuracy. These authors studied soil samples collected at three temperate sites in Montana, USA, and obtained better predictions for some sites using the global library augmented with local samples than by using the local samples only. However, Wetterlind & Stenberg (2010) found that for a range of soil properties at farms in Sweden, local calibrations with only 25 calibration samples out-performed both the national library (396 samples) and subsets of the national library consisting of the 50 samples most similar to each farm. According to Udelhoven *et al.* (2003), OM (calculated as $1.72 \times$ soil organic carbon) predictions can be improved by stratifying samples according to geological conditions and deriving individual PLSR calibrations for each region. Stenberg (2010) did not find an accurate soil organic carbon (SOC) calibration model by using large datasets representative of Swedish agricultural soils, but these results were substantially improved when sandy soils were removed from the dataset. Similar negative effects of large sand contents were observed by Stevens *et al.* (2013), who studied a broad variety of soils from Europe. It is often suggested that libraries containing smaller soil variations at the field scale would result in better OM predictions than more general ones collected over larger geographical areas (Kuang & Mouazen, 2011). Local geographical datasets have been stated to be necessary for quantifying soil attributes (Demattê & Garcia, 1999). However, Stenberg *et al.* (2010), reviewing published predictions, found that variation in the texture or SOC variables themselves accounted for the majority of the variation in model accuracy for these properties and the size of the geographical area had a smaller influence. Thus, attempts to improve the prediction accuracy of a large heterogeneous spectral library may benefit from dividing the library into smaller sub-libraries with soils of greater similarities, regardless of the geographical origin of the samples. Dividing a global library into smaller models based on the variation in the spectra caused by clay minerals and SOC (because they have the largest influence on soil vis-NIR spectra: Stenberg *et al.*, 2010) is therefore one potential strategy for improving predictions. McDowell *et al.* (2012) compared the sub-division of a Hawaiian (five main islands) dataset into clusters based on total carbon (C) content, soil order and spectral features, for the prediction of total C. None of the sub-division strategies were better than a full sample set strategy. This could, however, be a result of the small number of samples in many of the calibration clusters because the full set comprised only 307 samples.

It is known that the soil vis-NIR spectra are largely non-specific, consisting of weak, broad and overlapping absorption bands. For this reason, information needs to be mathematically extracted from the spectra in order to correlate them with soil properties, and multivariate statistics are often used to calibrate soil prediction models. Partial least square regression (PLSR) is one of the most commonly used techniques to analyse data of this nature. Vasques *et al.* (2008) compared different techniques, such as stepwise multiple linear regression, principal component regression, regression trees, committee trees and PLSR, to analyse spectral information related to organic carbon and concluded that PLSR performed better than the other methods. Interest in using non-linear data mining calibration techniques is increasing, because relationships between soil characteristics are rarely linear in nature, especially in libraries containing a wide variety of soils. When dealing with a heterogeneous sample set in which soil composition may vary considerably, the precision of linear regression techniques decreases because of the non-linear nature of the relationship between spectral data and the dependent variable. Brown (2007) suggests the use of boosted regression trees and Kovačević *et al.* (2009) suggest the use of support vector machines as a solution for handling the calibration of large heterogeneous sample populations.

Although soil spectroscopy has the potential to simplify soil analysis and mapping, vis-NIR spectroscopy has not yet developed sufficiently for practical applications, and there is less information for tropical soils than for those in temperate regions.

Our study aims firstly to explore the possibility of enhancing predictions of OM and clay content in a large Brazilian soil spectral library by dividing it into smaller sub-libraries based on their vis-NIR spectra. In the process, we also tested the effect of three different pretreatments of the spectra; continuum removal, first derivative and mean normalization prior to division of the library. The second aim is to compare the total predictive performance of the sub-models with global models using PLSR, BT and SVM techniques. The comparison of clustering with BT and SVM techniques with a large variable dataset to reduce problems with heterogeneity and non-linearity has to our knowledge not yet been done.

Materials and methods

Spectral library

For this study we used 7172 samples in the soil spectral library made available by the Remote Sensing Laboratory at the Soil Science Department, University of São Paulo (co-author J.A.M. Demattê). In total, chemical and spectral analyses were carried out for 5750 auger samples collected at depths of 0–20, 40–60 and 80–100 cm and for 1440 samples collected from 360 soil profiles representing four Brazilian states (Goiás, Minas Gerais, Mato Grosso do Sul and São Paulo). The soils in this spectral library are diverse and represent several groups of soils, including

Ferrasols, Nitisols, Acrisols, Planosols, Gleysols, Arenosol and Cambisols (IUSS, 2006).

The samples were air-dried and ground to a particle size of <2 mm before being submitted to chemical and spectral analyses. Sand (2–0.05 mm), silt (0.05–0.002 mm) and clay (< 0.002 mm) contents were determined by the densimeter-sedimentation method, using 0.1 M calcium hexametaphosphate and 0.1 M sodium hydroxide as dispersing agents (Gee & Bauder, 1986). Organic matter content was determined applying the factor 1.72 to organic carbon as determined colorimetrically after oxidation of 1 ml air-dry soil with $K_2Cr_2O_7 + H_2SO_4$. The excess dichromate was titrated with $(NH_4)2Fe(SO_4)_2.6H_2O$ (van Raij *et al.*, 2001).

Vis-NIR measurements

The spectral reflectance of soils was measured in the vis-NIR (350–2500 nm) range, with a spectral resolution of 3 nm (from 350 to 1000 nm) and 10 nm (from 1000 to 2500 nm), using a FieldSpec Pro FR spectroradiometer (Analytical Spectral Devices, Boulder, Colorado, USA). The spectrum acquisition software interpolated reflectance data to a sampling interval of 1 nm. Approximately 15 cm^3 of each soil sample was placed in a Petri dish. A fibre-optic cable connected to the vis-NIR sensor was placed vertically at 8 cm from the sample, and we measured the reflected light in an area of approximately 15 cm^2 in the centre of the sample. The light source was a 50 W halogen bulb with the beam non-collimated to the target plan, positioned at 35 cm from the sample at a zenith angle of 30°. As a reference standard, a white plate covered with barium sulphate ($BaSO_4$) was used. Each spectrum was averaged from 100 readings over 10 s. All spectral measurements were carried out in a dark room to avoid interference from stray light. Before further analyses, soil spectra were reduced by averaging three successive wavelengths. For further analyses we excluded the noisiest parts at the edges of the spectrum and only considered the spectral range from 366 to 2484 nm.

Model development

Prior to any model development the global spectral library was randomly divided into a calibration set (CS) with 5169 samples and a validation set (VS) with 2003 samples, corresponding to one-third of the profiles. Layers of the same soil profile were kept together to ensure independence between CS and VS. Figure 1 shows the summary statistics for clay and organic matter content in the CS and VS. For subsequent analyses we restricted the datasets to OM content less than 6% as there were very few samples with more than 6% OM. In total, eight and five samples were removed from the CS and VS, resulting in 5161 and 1998 samples, respectively.

The general approach in model development was that two major lines of calibration procedures were performed and compared. One involved simple calibrations on the calibration set as a whole (global models), and the other one involved calibrations that were performed cluster by cluster after the calibration set had been divided into spectrally similar clusters (clustered models). For all calibrations the first derivative using a second-order polynomial Savitzky-Golay smoothing over 11 points (Savitzky & Golay, 1964) was applied as spectral preprocessing on the absorbance spectra (absorbance = 1/(log reflectance)). The first derivative was the best method in an initial screening test also including multiplicative scatter correction, moving average, median filters, standard normal variate and mean normalization.

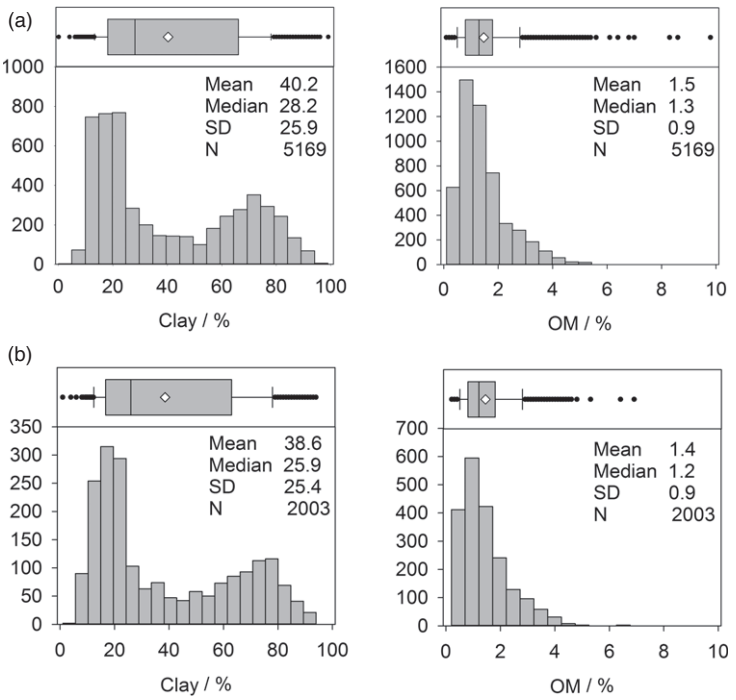

Figure 1 Descriptive statistics, histograms and outlier box-plots for clay and organic matter content in the calibration (a) and validation (b) datasets. The mean (diamond), the median (line), the 25th and 75th percentiles (box), the 10th and 90th percentiles (whiskers) and outliers (dots) are indicated.

Global models were calibrated on the full calibration set (CS; $n = 5161$). Three different calibration techniques were tested: PLSR, SVM and BT. The PLSR technique is widely used and has a good capacity for estimating attributes resulting from the spectral behaviour of the soil (Vasques *et al.*, 2008). It was performed in the Unscrambler v.10.3 software using the orthogonalized PLSR algorithm for one Y-variable (PLSR-1) and cross-validation in 50 random segments. The number of partial least-square factors was chosen to minimize the root mean square error (RMSE) in the cross-validation.

The SVM approach is a relatively new non-linear technique and is used in classification and multivariate calibration problems (Kovačević *et al.*, 2009). In this technique, model complexity is limited by the learning algorithm itself, which prevents over-fitting. In the present study the Kernel radial basis function was used, which allows learning of non-linear decision functions (Jain *et al.*, 2012). Optimal model parameters that minimize the root mean square error (RMSE) were determined by 10-fold random cross-validation of CS with Statistica 10 software (StatSoft Inc, Tulsa, OK, USA).

The BT technique makes multiple predictions that are based on resampling and weighting and belongs to the group of ensemble techniques (Friedman, 2001). It has the ability to include a large number of weak relationships in a predictive model and it is insensitive to outliers in the calibration dataset. Moreover, BT has a relative immunity to over-fitting (Brown, 2007). A maximum number of seven nodes and committees of 900 trees for organic matter and 600 for clay were used for calibrations. The number of nodes and trees was optimized to minimize the root mean square error (RMSE) by reserving a 30% test set from CS: these analyses were performed with Statistica 10.

Prior to clustering, three different spectral transformations were applied and evaluated. In these cases the absorption spectra were (i) transformed to the first derivative Savitzky-Golay (D; second order with 11 smoothing points), (ii) mean normalized (N; dividing each spectrum by its mean) and the reflectance spectra were (iii) transformed to continuum removal (CR; Clark, 1999) by determining the convex hull (ENVI 4.5; www.envi.com.br). The main purpose of the transformations was to see if they would divide the data differently in the clustering process and to assess what influence this would have on the predictive performance of calibrations for OM and clay. The procedures for calibrating and validating clustered datasets are summarized in Figure 2 as follows: (i) The calibration spectra were clustered with k-means cluster analysis by their spectral features as they appear with the three different transformations. Subsequently, PLSR calibrations were used to produce predictive models for OM and clay in each cluster. The models were in all cases calibrated on first derivative spectra only, based on absorbance. (ii) To be able to allocate unknown samples (the validation set) to one of the spectrally defined clusters, the spectral features defining the clusters were identified by discriminant analysis models (DA). (iii) The validation set was clustered by the DAs in step (ii). For validation of the PLSR models in step (i), the validation samples clustered through step (iii) were predicted and validation statistics compared.

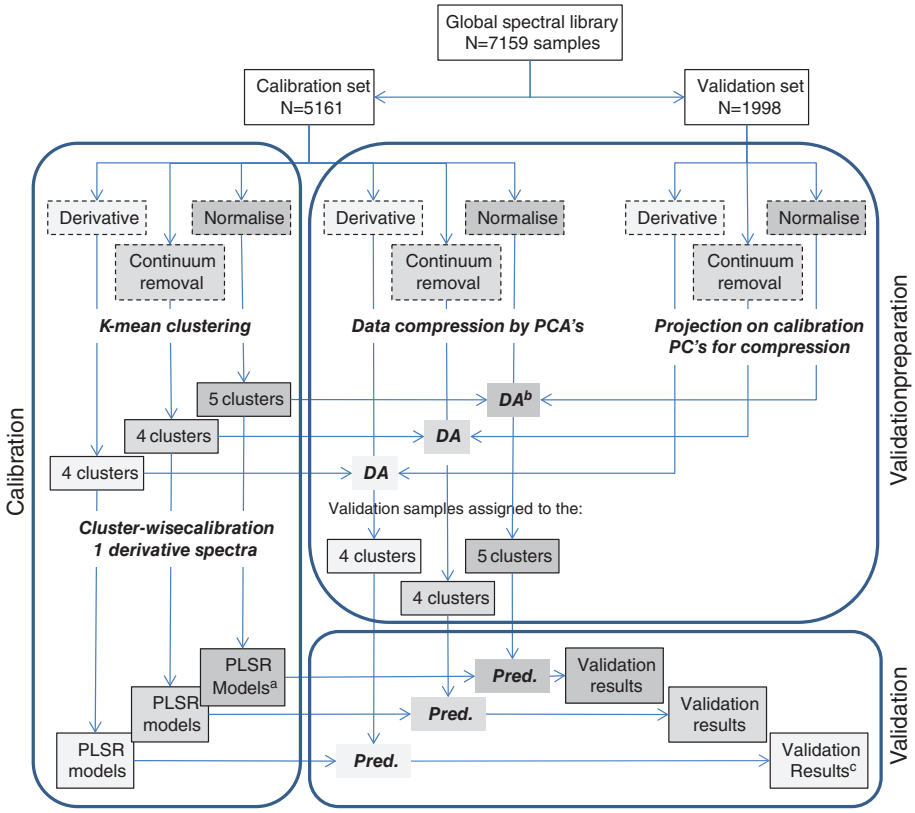

Figure 2 Overview of steps taken during the preprocessing and analyses. The dataset is restricted to OM of < 6%. [a]One PLSR model per cluster. [b]Discriminant analysis to assign the validation samples to the different clusters for prediction of unknown samples using the cluster-vise PLSR models. [c]The validation results were compared (i) between clusters within the same transformation, (ii) between the three transformations using the combined prediction results (CPR) from all the clusters within a method and (iii) the CPRs were compared with the global model (also using 1 derivative spectra).

The three differently transformed datasets were submitted to a k-means clustering algorithm with the statistical software Statistica 10. This analysis starts with *k* random clusters, and then moves objects between those clusters in order to minimize the intra-group variability and to maximize the distances between groups. The software iteratively moves objects in and out of clusters, minimizing the square of the within-cluster sum of distances to get the most significant ANOVA results between clusters. For CR and first derivative transformations four clusters were optimal and for normalized data the optimum number of classes was five. Optimizing the number of clusters was performed by a cross-validation procedure to minimize the misclassification. PLSR models were calibrated cluster-wise to test the influence of different transformations for clustering on total prediction performance. For these calibrations the first derivative Savitzky-Golay transformations of absorption spectra were used independently of transformation used for clustering.

Validation

All predictions of OM and clay content by global (PLSR, BT and SVM) and clustered models were validated using the predefined validation set (VS; $n = 1998$).

For the clustered models, the validation sample had first to be assigned to one of the clusters. Thus, the success of this assignment step was included in the validation of calibrations. Discriminant analysis models, one for each transformation, were developed to define the spectral features that separate the clusters. The Euclidian metric distance method in Statistica 10 was used to separate the predefined classes. For computational reasons, the analyses were performed on dimensionally compressed data. Thus, score vectors from the 10 first principal components of a PCA based on the calibration set were used. Scores for the validation samples were calculated by projecting the transformed spectral data on the PCA based on the calibration set. Each validation sample was then assigned to one of the clusters for each transformation by the corresponding discriminant analysis model (Figure 2). Principal components and scores were calculated by the NIPALS algorithm in Unscrambler 10.3.

Finally, the modelling efficiency (*ME*), the root mean squared error (RMSE) and the ratio of performance to deviation (RPD) were used to compare the results, calculated by using the Equations (1)–(3). The ME indicates the proportion of the total variation explained by the model (the 1:1 line) and includes the relationship between measured and predicted values as well as systematic errors. In Equations (1)–(3) where y denotes the measured value and $\hat{y}$ the predicted value, n is the number of samples and SD is the standard deviation of laboratory-measured values for the property in question.

$$ME = 1 - \frac{\sum_{i=1}^{n} \left(yi - \hat{y}i\right)^2}{\sum_{i=1}^{n} \left(yi - \overline{yi}\right)^2}, \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(yi - \hat{y}i\right)^2}, \quad (2)$$

and

$$RPD = SD/RMSE. \quad (3)$$

To achieve combined prediction results (CPR) of the clustered models ME, RMSE and subsequently RPD were calculated after merging predictions of all clusters to the full VS. The RPD values are included for comparative evaluations. We did not allocate predictions according to any published performance thresholds based on RPDs as the required accuracy needs to be evaluated for each application and should not be dependent on the size of the standard deviation.

As mentioned earlier, organic matter and especially clay minerals and iron-oxides (which are also included in the clay fraction) have a large influence on the vis-NIR spectra and are therefore expected to also influence the clustering. To indicate the difference in OM and clay content between clusters, Kruskal-Wallis ANOVAs by ranks and median test, including pair-wise *post-hoc* comparisons of mean ranks, were run on the calibration set in Statistica 10.

Results and discussion

Global calibrations: prediction accuracy obtained with PLSR, BT and SVM techniques

The validation results of the global predictions produced using the PLSR, BT and SVM methods are summarized in Figure 3. In general, we observed better results with SVM and the boosted regression trees technique than with PLSR. The SVM and BT methods performed equally well for OM, reducing RMSEv by 13% when compared with PLSR. For clay, SVM reduced RMSEv by almost twice as much as BT and by 18% when compared with PLSR. These results agree with Brown (2007), who compared BT and PLSR techniques for analysing soil properties with vis-NIR and found BT to be the superior approach. These authors used 4184 diverse, well-characterized and largely independent soil samples. The BT technique tends to be resistant to the effects of outliers and can handle missing values and correlated variables. It also allows the inclusion of a potentially large number of irrelevant predictors (Jalabert *et al.*, 2010). On the other hand, Viscarra Rossel & Behrens (2010), using 1104 samples from four regions in Australia, and Vasques *et al.* (2008), using 554 samples collected in profiles to a depth of 180 cm in north-central Florida, observed that BT and regression trees models produced the worst results among many multivariate techniques, including PLSR, when tested for total carbon, organic carbon and clay. Viscarra Rossel & Behrens (2010) also found SVM to be about equal to PLSR, while Pierna & Dardenne (2008) found that SVM was superior to PLSR for predicting total N, total C and CEC in a validation set of 207 soils from a calibration set of 618 cultivated Belgian soils. Stevens *et al.* (2013) also found that SVM was better for organic C predictions when they compared several data mining calibration techniques on a diverse sample set of 20 000 samples covering most soil types (but divided into subsets according to land cover) in the EU.

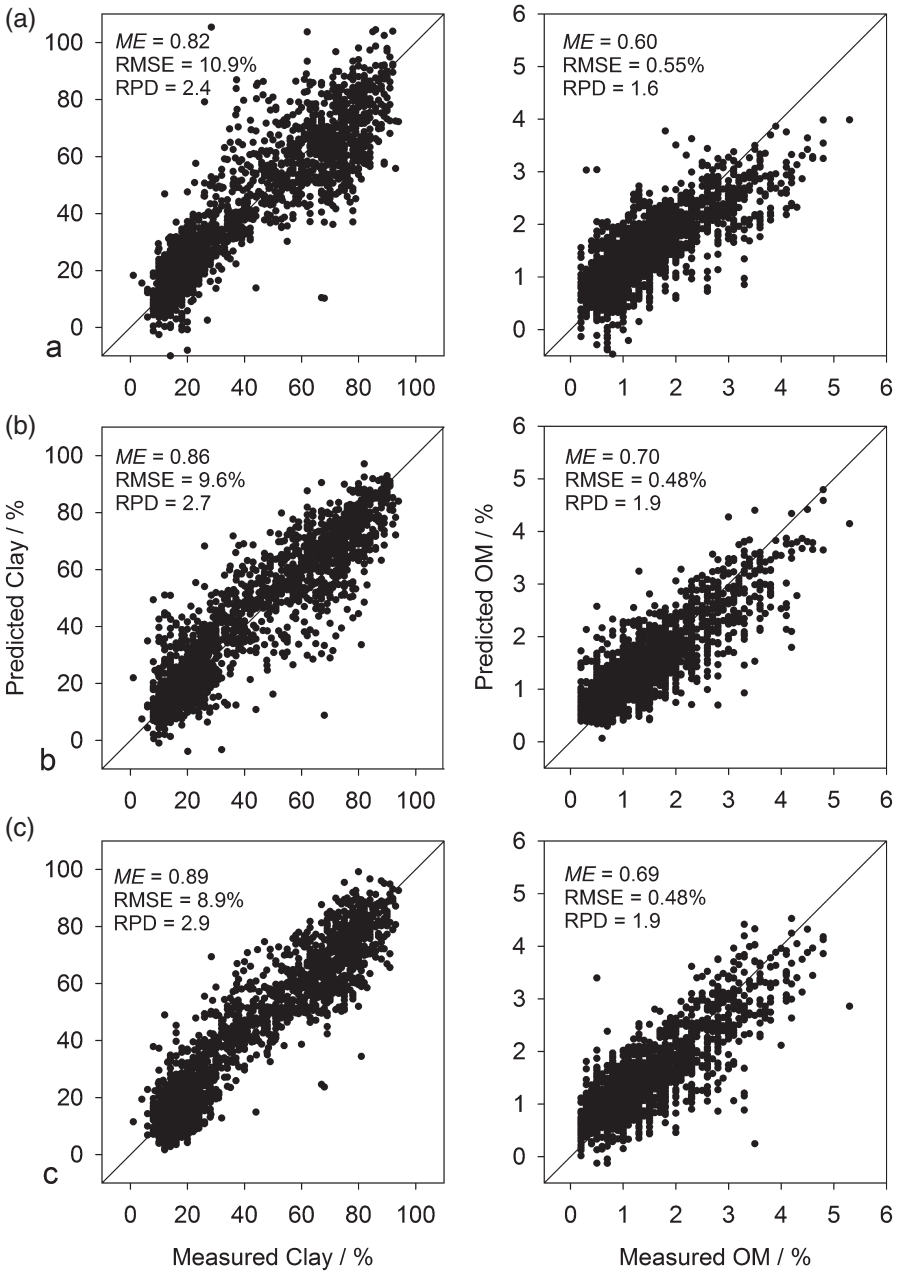

Figure 3 Validation scatter plot of laboratory-measured data against vis-NIR predictions obtained from (a) partial least square regression, (b) boosted tree regression and (c) a support vector machine for organic matter (OM) and clay content.

Clustering

The spectral library data were divided into spectrally defined clusters of different sizes and samples depending on the transformation employed (CR, first derivative or mean normalized; Table 1). All cluster medians within a transformation differed significantly from each other for either OM or clay and in most cases for both. Typically the differences between clusters were larger for clay than for OM. The standard deviation was less than in the total dataset in many clusters (Figure 1), but not in all. The reductions of SD were generally larger in clay than in OM. These results were expected as OM and especially clay minerals have a large influence on the spectra.

Cluster predictions

The validation statistics calculated from the combined prediction results (CPR) of all validation samples in all clusters by the respective pre-transformations (CPR-N, CPR-D and CPR-CR, respectively) showed that transforming the data by using the first derivative and mean normalization prior to cluster analysis provided slightly more accurate models than transforming the data by continuum removal (Table 2).

We observed a larger improvement in accuracy of predictions for clay than for OM with clustered models, with a largest reduction of the RMSEv of 15 and 21% for OM and clay, respectively. The combined prediction results with mean normalization (CPR-N)

Table 1 Calibration and validation set statistics for the clusters acquired by cluster *k*-means and discriminant analysis, respectively

		N	25%	75%	Median	SD	N	25%	75%	Median	SD
Transformations	Classes	*Calibration set*					*Validation set*				
		OM / %									
Normalized	1	807	0.8	2.0	1.3	1.0	308	0.8	1.8	1.1	0.9
	2	788	0.9	1.8	1.3	1.0	336	0.8	1.8	1.2	0.9
	3	657	1.0	2.0	1.4	0.8	297	1.0	2.2	1.5	0.9
	4	1035	1.1	2.3	1.7	0.9	395	1.2	2.3	1.7	0.9
	5	1874	0.7	1.5	1.0	0.8	662	0.6	1.3	1.0	0.7
Derivative	1	933	0.8	2.5	1.5	1.1	377	0.8	2.5	1.3	1.1
	2	456	0.8	1.8	1.3	0.7	189	0.8	1.8	1.3	0.7
	3	1101	1.1	2.3	1.6	0.9	404	1.1	2.3	1.6	0.8
	4	2671	0.8	1.5	1.0	0.7	1028	0.7	1.5	1.0	0.7
CR	1	745	1.1	2.3	1.7	0.9	289	1.2	2.3	1.6	0.8
	2	740	1.0	2.3	1.5	1.0	312	1.0	2.3	1.5	1.0
	3	2086	0.7	1.5	1.0	0.8	794	0.7	1.5	1.0	0.8
	4	1590	0.9	1.8	1.3	0.9	603	0.8	1.8	1.2	0.9
		Clay / %									
Normalized	1	807	22.0	64.0	30.0	23.6	308	20.0	64.8	28.4	24.6
	2	788	12.0	33.7	16.0	24.4	336	12.0	20.0	15.8	20.6
	3	657	44.0	78.0	61.0	20.7	297	41.5	78.0	60.0	20.6
	4	1035	46.0	75.0	67.0	18.7	395	44.6	75.0	64.0	17.9
	5	1874	15.6	25.1	20.0	20.3	662	16.0	24.0	19.3	17.1
Derivative	1	933	25.5	82.0	64.0	27.8	377	24.0	81.0	60.0	28.0
	2	456	36.0	59.0	47.5	15.8	189	32.5	59.0	44.0	16.6
	3	1101	44.0	75.0	66.0	19.3	404	41.0	75.0	63.0	19.5
	4	2671	14.0	24.0	18.0	18.9	1028	14.0	24.0	18.0	17.5
CR	1	745	44.0	75.0	66.0	19.0	289	44.4	74.0	63.0	17.4
	2	740	38.0	69.0	55.0	20.4	312	32.2	67.0	51.0	21.4
	3	2086	18.0	68.0	24.0	26.1	794	18.0	63.0	23.4	25.5
	4	1590	12.7	26.9	17.4	21.0	603	12.3	25.7	16.0	20.8

N, D and CR mean normalization, first derivative and continuum removal transformations, respectively.

are shown in Figure 4. McDowell *et al.* (2012) also used mean normalization for a Hawaiian dataset of 307 samples, but in contrast to our results they did not find any significant improvement by k-means clustering on total C predictions. This may be because of their small number of samples that were divided in both validation and calibration sets as well as in clusters. Potential advantages with clustering may have been dominated by unstable calibrations. For clustering, McDowell *et al.* (2012) also focused on spectral bands known to be absorbed by organic compounds. This may have resulted in clusters with reduced diversity in carbon-related properties, but with retained diversity in other characteristics such as mineralogy.

When normalization was used as a preprocessing treatment, the global spectral library was divided into five clusters. For clay the independent validation results for clusters 2 and 5 provided the largest values of *RPD* and *ME*, followed by clusters 1, 3 and 4; for OM the values were largest in clusters 1 and 2, followed by 3, 5 and 4.

The success of assigning the validation samples to the right cluster by discriminant analyses on normalized data is indicated in Figure 5(a,b), which shows the clusters' positions and distribution in PCA scores of calibration and validation samples. In addition, the proportional size distributions between clusters were very similar in CS and VS (Table 1).

In our study, the additional step of assigning validation samples to the right cluster in the prediction process, which is required in a real situation with unknown samples, did not increase the overall prediction error substantially. We observed that calibration and cross-validation results (which do not involve sample-to-cluster assignment) and independent validation results (which do) did not differ more for the clustered models than for the global PLSR models (Table 3). If the assignment of validation samples to clusters increased the prediction error substantially, a larger difference for the clustered models would have been expected. In fact, SVM and BT had the largest differences between calibration and independent validation, indicating some degree of over-fitting,

Table 2 Summary statistics of validation results of calibrations for clay (%) and OM (%) using an independent validation dataset

			Clay			OM		
Preprocessing	Cluster	Number of samples	*ME*	RMSEv	RPD	*ME*	RMSEv	RPD
Normalized	1	308	0.86	9.2	2.7	0.77	0.45	2.0
	2	336	0.89	6.8	3.0	0.75	0.44	2.0
	3	297	0.80	9.2	2.2	0.68	0.50	1.8
	4	395	0.55	12.0	1.5	0.57	0.57	1.6
	5	662	0.87	6.1	2.8	0.64	0.40	1.8
CPR - N		**1998**	**0.88**	**8.6**	**3.0**	**0.71**	**0.47**	**1.9**
First derivative	1	377	0.88	9.6	2.9	0.79	0.50	2.2
	2	189	0.88	8.9	1.9	0.50	0.52	1.3
	3	404	0.64	11.7	1.7	0.55	0.56	1.4
	4	1028	0.83	7.3	2.4	0.70	0.40	1.8
CPR - D		**1998**	**0.88**	**8.8**	**2.9**	**0.71**	**0.47**	**1.9**
CR	1	289	0.52	12.1	1.4	0.59	0.53	1.5
	2	312	0.73	11.1	1.9	0.63	0.61	1.6
	3	794	0.88	8.8	2.9	0.65	0.46	1.7
	4	603	0.86	7.7	2.7	0.64	0.51	1.8
CPR - CR		**1998**	**0.86**	**9.5**	**2.7**	**0.66**	**0.51**	**1.8**
PLS	**All**	**1998**	**0.82**	**10.9**	**2.4**	**0.60**	**0.55**	**1.6**
BT	**All**	**1998**	**0.86**	**9.6**	**2.7**	**0.70**	**0.48**	**1.9**
SVM	**All**	**1998**	**0.89**	**8.9**	**2.9**	**0.69**	**0.48**	**1.9**

Bold values represent the full validation set and are comparable across all methods.
The validation samples assigned to each cluster were based on discriminant analyses. *PLSR*, *BT and SVM* refer to non-clustered models obtained by partial least squares regression, boosted regression trees and support vector machines, respectively; CPR - N, CPR - D and CPR - CR refer to combined prediction results (CPR) with mean normalization (N), first derivative (D) and continuum removal (CR) as transformation applied prior to clustering, respectively.

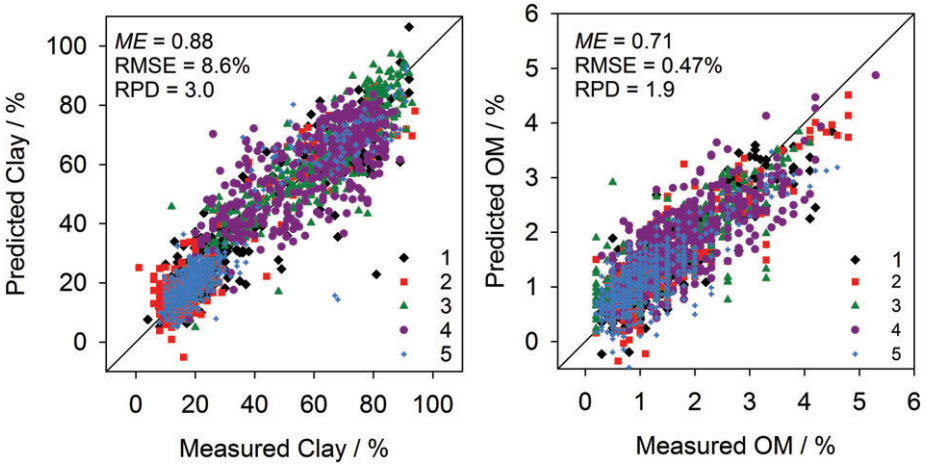

Figure 4 (a,b) Validation scatter plot of laboratory measured data plotted against vis-NIR predictions obtained from partial least square regression (CPR - N).

but as the validation statistics were still substantially better than PLS the effect was not large.

We compared our results with the standard deviations and RMSE or R^2 of most large-scale published datasets with all three values available (Stenberg *et al.* (2010) (Figure 6). In most studies the R^2 values obviously correspond to our *ME* values. Because the relationship between SD and RMSE is strong, it is more relevant to compare results with other studies by this relationship rather than simply comparing RMSE, R^2 or RPD values. We observe that our *ME* value for global PLSR is less than expected from previously published data, but not so for global SVM and BT. Values of RMSE were, on the other hand, more or less as expected, but while global PLSR and CR clustering were slightly above the regression line, the other methods were slightly below the line, which indicates a comparatively more successful result than those previously published. Similarly, predictions of clay were equal to, or slightly better than, what could be expected from the standard deviation of 25% clay in the global library. This is based on the strong correlation between standard deviations and RMSE values in published texture data by Stenberg *et al.* (2010), corresponding to that for OM in Figure 6.

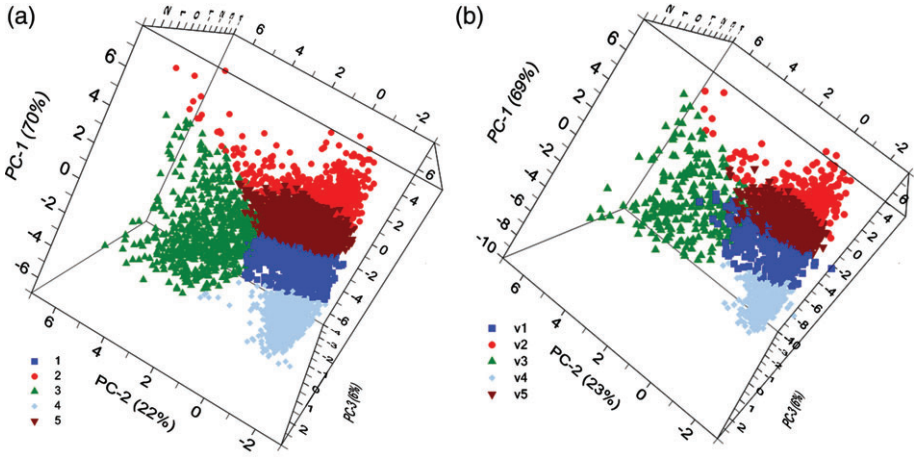

Figure 5 Scores of the three first principal components of PCAs of the calibration (a) and validation (b) sets. Numbers 1–5 correspond to clusters 1–5 from k-means clustering in the calibration set and v1–v5 as validation samples were allocated to clusters.

Table 3 Prediction results of cross-validation (RMSEcv) and independent validation (RMSEv) of clay (%) and OM (%)

	Clay			OM		
Models	RMSEc	RMSEcv	RMSEv	RMSEc	RMSEcv	RMSEv
CPR - N	8.3	8.6	8.6	0.48	0.50	0.47
CPR - D	8.0	8.5	8.8	0.47	0.49	0.47
CPR - CR	8.4	8.8	9.5	0.48	0.50	0.51
PLS	10.4	10.4	10.9	0.57	0.57	0.55
BT	7.9	–	9.6	0.40	–	0.48
SVM	7.5	–	8.9	0.43	–	0.48

PLSR, *BT and SVM* refer to non-clustered models obtained by partial least squares regression, boosted regression trees and support vector machines, respectively; CPR - N, CPR - D and CPR - CR refer to combined prediction results (CPR) with mean normalization (N), first derivative (D) and continuum removal (CR) as pre-transformation treatments, respectively.

Soil spectra

Bands around 1100, 1600, 1700, 2000 and 2300 nm have been identified as being particularly important for SOC and total N calibration (Malley *et al.*, 2000; Stenberg, 2010). Although we observed spectral features in these regions by removing the continuum from the average reflectance spectra of the classes, these features were not enhanced with increasing OM content, as observed by Stenberg (2010).This may be explained by the small concentrations, and the narrow range, of OM values in the current library (Table 1).

The mean spectral curves of clusters based on mean normalization spectra were analysed in detail as these classes provided the best regression model results (Table 3). The spectral features were, however, studied by continuum-removed spectra and not by the mean normalized ones (Figure 7). From Figure 7 it can be observed that soil mineralogy had a substantial influence on spectral clustering, as discussed in the next paragraph. This is not surprising as clay minerals together with organic matter have large influences on soil spectra (Stenberg *et al.*, 2010). As the OM content is very small in this dataset, clay mineralogy should dominate. Apparently, the more homogenous mineralogical clusters allowed improved overall prediction accuracy. Clay predictions gained the most from clustering and this may be because this fraction in our soil library was a mixture of both iron oxides and clay minerals, with fundamentally different features in the vis-NIR region. It is probably also this non-linearity that is analysed better by global data mining techniques than by global PLS. On the other hand, our results do not support the explanation that a reduction in SD caused by clustering should improve prediction accuracy, as indicated by Figure 6. There is no obvious relationship between cluster SDs in Table 1 and RMSEs in Table 2.

The 1400 and 1900 nm bands are associated with water vibrations connected to bonds of lattice layers as hydrated cations (structural), combined with water adsorbed on the particle surfaces. In all clusters we observed absorptions caused by charge transfers near 400–780 nm, which are indicative of the presence of iron oxides (Sherman & Waite, 1985). In turn, bands near 489 and 530 nm are attributed to absorptions edges of intense charge transfer absorptions that occur in the UV (Sherman & Waite, 1985). The reduced reflectance observed around 510–560 nm (Figure 7b) suggests that soils belonging to clusters 1, 2, 4 and 5 have larger haematite contents than cluster 3 (Demattê & Garcia, 1999). Scheinost *et al.* (1998) reported that the most intense absorption band for haematite occurred at 521–565 and 870 nm, clearly separated from the more yellowish Fe oxides (479–499 and 930 nm). In fact, the mean spectral reflectance of cluster 3 had a shift to the right near 900 nm and to the left near 500 nm, which indicates a greater presence of

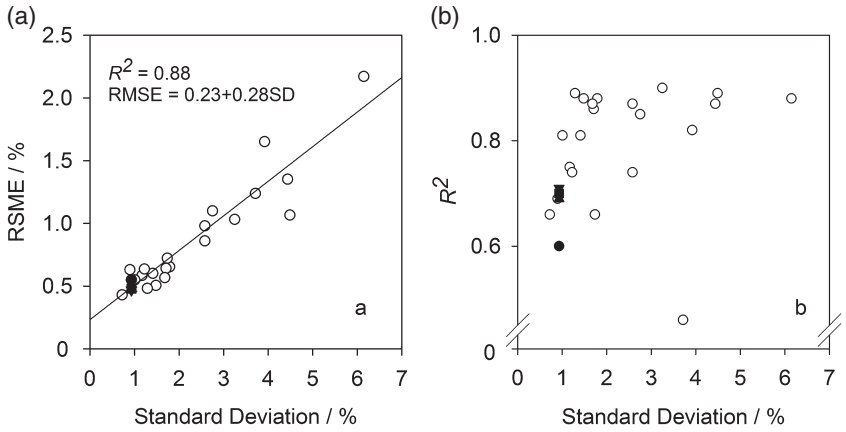

Figure 6 Relationships between the standard deviation and RMSE (a) and R^2 (b). Published data on organic matter predictions with vis-NIR spectroscopy extracted from Malley *et al.*, 2000; Chang *et al.*, 2001, 2005; Chang & Laird, 2002; Dunn *et al.*, 2002; Fystro, 2002; Martin *et al.*, 2002; Moron & Cozzolino, 2002; Stenberg *et al.*, 2002; Islam *et al.*, 2003; Udelhoven *et al.*, 2003; Sørensen & Dalsgaard, 2005; Todorova *et al.*, 2009; Stenberg, 2010; Viscarra Rossel & Behrens, 2010; Nocita *et al.*, 2011; Vohland *et al.*, 2011; Cambule *et al.*, 2012; Goge *et al.*, 2012; and Tekin *et al.*, 2012 (○). PLSR results data, global model (●); SVM results, global model (▲); BT results, global model (■); PLSR results data, normalized cluster models (▼). Adapted from Stenberg *et al.* (2010) . R^2 in (b) refer to published data denoted as r^2 or R^2, which obviously in most cases correspond to model efficiency (*ME*).

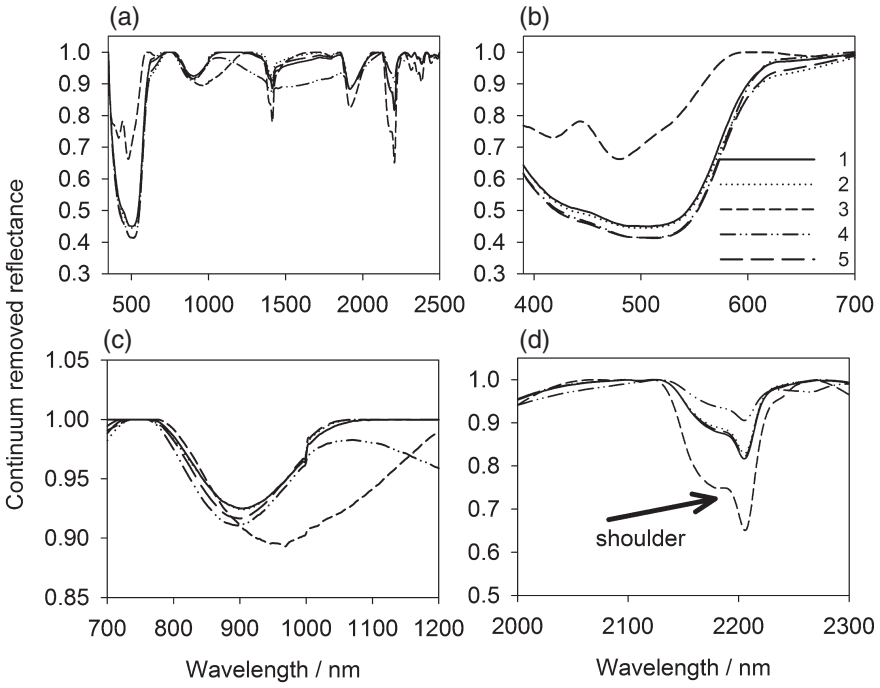

Figure 7 The continuum-removed spectra average of classes obtained by k-means clustering analyses when the global data were submitted to mean normalization. Numbers 1–5 refer to clusters 1–5. Full-range spectra (a) and enlargements (b–d).

goethite than haematite (Figure 7c). For soil clays, it is known that kaolinite (1:1) and 2:1 minerals have characteristic patterns near 1400 and 2200 nm because of the vibrations of molecules of OH of their structures (Hunt & Salsbury, 1970). Different minerals have, however, different signatures. Kaolinite minerals have a shoulder near 2200 nm that does not occur when there is a predominance of a 2:1 mineral in the soil (Demattê *et al.*, 2006). This shoulder is more pronounced for cluster number 3, indicating greater proportions of 1:1 minerals (Figure 7c). Hunt & Salisbury (1970) argued that the intensity of the kaolinite trait at 2200 nm is associated with the dioctahedral layers of the mineral structure. Although the absorption near 2345 nm may represent illite or mixtures of smectite and illite (Post & Noble, 1993), we observed only a slight difference in reflectance between clusters 3 and 4 and the other clusters in this region (Figure 7a).

Conclusions

The general predictive models for clay were good, which reflects the influence of the direct spectral responses of this property in the NIR range. Organic matter predictions were reasonably good,

especially with clustering and in view of the very small variation in OM contents in the dataset.

The division of the large library into smaller subsets using variation in the mean-normalized spectra or first derivative spectra were the best alternatives for using vis-NIR spectra to quantify soil attributes in tropical soils by partial least square regressions. An alternative would be to use boosted regression trees or support vector machines for the whole library, which were almost as good and are more straightforward methods as the clustering and cluster assignment steps are avoided. While clustering reduced RMSEv by 21 and 15% for clay and OM, respectively, the corresponding values for SVM were reduced by 18 and 13% when compared with PLSR. This suggests that clustering and data mining calibration techniques are preferable over global PLS to handle non-linearity in large and complex datasets.

Clustering divided the global dataset into more mineralogically uniform clusters, regardless of geographical origin, and improved prediction. The additional step of assigning the validation samples to the correct cluster in the prediction process (clustered models) did not increase the overall prediction error. It was possible to identify regions of the vis-NIR spectrum that showed absorption features from water, iron oxides and clay minerals that seemed to be largely responsible for the cluster divisions.

Acknowledgement

This work has been funded by FAPESP (São Paulo research Foundation), (CNPq) National Council for Scientific and Technological Development, the Swedish Farmers' Foundation for Agricultural Research and the Swedish Research Council Formas.

References

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**, 444–453.

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J. & Smaling, E.M.A. 2012. Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. *Geoderma*, **183**, 41–48.

Chang, C.W. & Laird, D.A. 2002. Near-infrared reflectance spectroscopic analysis of soil C and N. *Soil Science*, **167**, 110–116.

Chang, C.W., Laird, D.A., Mausbach, M.J. & Hurburgh, C.R. 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. *Soil Science Society of America Journal*, **65**, 480–490.

Chang, G.W., Laird, D.A. & Hurburgh, G.R. 2005. Influence of soil moisture on near-infrared reflectance spectroscopic measurement of soil properties. *Soil Science*, **170**, 244–255.

Clark, R.N. 1999. Spectroscopy of rocks and minerals and principles of spectroscopy. In: Remote Sensing for the Earth Sciences (ed. A.N. Rencz), pp. 3–58. John Wiley & Sons, Chichester.

Demattê, J.A.M. & Garcia, G.J. 1999. Alteration of soil properties through a weathering sequence as evaluated by spectral reflectance. *Soil Science Society of America Journal*, **63**, 327–342.

Demattê, J.A.M., Sousa, A.A., Alves, M.C., Nanni, M.R., Fiorio, P.R. & Campos, R.C. 2006. Determining soil water status and other soil characteristics by spectral proximal sensing. *Geoderma*, **135**, 179–195.

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. *Australian Journal of Experimental Agriculture*, **42**, 607–614.

Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.

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 & Soil*, **246**, 139–149.

Gee, G.W. & Bauder, J.W. 1986. Particle-size analysis. In: Methods of Soil Analysis. Part 1: Physical and Mineralogical Methods, 2nd edn (ed. A. Klute), pp. 383–411. Soil Science Society of America, Madison, WI.

Goge, F., Joffre, R., Jolivet, C., Ross, I. & Ranjard, L. 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemometrics & Intelligent Laboratory Systems*, **110**, 168–176.

Hunt, G.R. & Salisbury, J.W. 1970. Visible and near-infrared specta of minerals and rocks: I. Silicate minerals. *Modern Geology*, **1**, 283–300.

Islam, K., Singh, B. & McBratney, A. 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Australian Journal of Soil Research*, **41**, 1101–1114.

IUSS Working Group WRB. 2006. World Reference Base for Soil Resources 2006. 2nd edition. World Soil Resources Reports No. 103. FAO, Rome. ISBN 92-5-105511-4, 128 pp.

Jain, P., Kulis, B., Davis, J.V. & Dhillon, I.S. 2012. Metric and kernel learning using a linear transformation. *Journal of Machine Learning Research*, **13**, 519–547.

Jalabert, S.S.M., Martin, M.P., Renaud, J.P., Boulonne, L., Jolivet, C., Montanarella, L. *et al.* 2010. Estimating forest soil bulk density using boosted regression modelling. *Soil Use & Management*, **26**, 516–528.

Kovačević, M., Bajat, B., Trivic, B. & Pavlovic, R. 2009. Geological units classification of multispectral images by using support vector machines. In: 2009 International Conference on Intelligent Networking and Collaborative Systems (eds Y.K. Badr, S. Caballe, F. Xhafa, A. Abraham & B. Gros), pp. 267–272. Ieee, New York.

Kuang, B. & Mouazen, A.M. 2011. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. *European Journal of Soil Science*, **62**, 629–636.

Malley, D.F., Martin, P.D., McClintock, L.M., Yesmin, L., Eilers, R.G. & Haluschak, P. 2000. Feasibility of analysing archived Canadian prairie agricultural soils by near infrared reflectance spectroscopy. In: Near Infrared Spectroscopy: Proceedings of the 9th International Conference (eds A.M.C. Davies & R. Giangiacomo), pp. 579–585. NIR Publications, Chichester.

McDowell, M.L., Bruland, G.L., Deenik, J.L. & Grunwald, S. 2012. Effects of subsetting by carbon content, soil order, and spectral classification on prediction of soil total carbon with diffuse reflectance spectroscopy. *Applied & Environmental Soil Science*, **2012**, Article ID 294121.

Moron, A. & Cozzolino, D. 2002. Application of near infrared reflectance spectroscopy for the analysis of organic C, total N and pH in soils of Uruguay. *Journal of Near Infrared Spectroscopy*, **10**, 215–221.

Nocita, M., Kooistra, L., Bachmann, M., Mueller, A., Powell, M. & Weel, S. 2011. Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. *Geoderma*, **167–68**, 295–302.

Pierna, J.A.F. & Dardenne, P. 2008. Soil parameter quantification by NIRS as a chemometric challenge at 'Chimiometrie 2006'. *Chemometrics & Intelligent Laboratory Systems*, **91**, 94–98.

Post, J.L. & Noble, P.N. 1993. The near-infrared combination band frequencies of dioctahedral smectites, micas, and illites. *Clays & Clay Minerals*, **41**, 639–644.

Sankey, J.B., Brown, D.J., Bernard, M.L. & Lawrence, R.L. 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. *Geoderma*, **148**, 149–158.

Savitzky, A. & Golay, M. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, **36**, 1627–1639.

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. *Clays & Clay Minerals*, **46**, 528–536.

Sherman, D.M. & Waite, T.D. 1985. Electronic spectra of Fe3+ oxides and oxiyhydroxides in the near infrared to ultraviolet. *American Mineralogist*, **70**, 1262–1269.

Sørensen, L.K. & Dalsgaard, S. 2005. Determination of clay and other soil properties by near infrared spectroscopy. *Soil Science Society of America Journal*, **69**, 159–167.

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**, 15–22.

Stenberg, B., Jonsson, A. & Börjesson, T. 2002. Near infrared technology for soil analysis with implications for precision agriculture. In: Near Infrared Spectroscopy: Proceedings of the 10th International Conference (eds A. Davies & R. Cho), pp. 279–284. NIR Publications, Chichester.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M. & Wetterlind, J. 2010. Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*, **107**, 163–215.

Stevens, A., Nocita, M., Toth, G., Montanarella, L. & van Wesemael, B. 2013. Prediction of soil organic carbon at the European scale by visible and near-infrared reflectance spectroscopy. *PLoS One*, **8**, e66409. doi: 10.1371/journal.pone.0066409.

Tekin, Y., Tumsavas, Z. & Mouazen, A.M. 2012. Effect of moisture content on prediction of organic carbon and pH using visible and near-infrared spectroscopy. *Soil Science Society of America Journal*, **76**, 188–198.

Todorova, M., Atanassova, S. & Ilieva, R. 2009. Determination of soil organic carbon using near-infrared spectroscopy. *Agricultural Science and Technology*, **1**, 45–50.

Udelhoven, T., Emmerling, C. & Jarmer, T. 2003. Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: a feasibility study. *Plant & Soil*, **251**, 319–329.

van Raij, B., Andrade, J.C., Cantarela, H. & Quaggio, J.A. 2001. Análise química para avaliação de solos tropicais. IAC, Campinas.

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

Viscarra Rossel, R.A. & Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, **158**, 46–54.

Vohland, M., Besold, J., Hill, J. & Fruend, H.-C. 2011. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma*, **166**, 198–205.

Wetterlind, J. & Stenberg, B. 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science*, **61**, 823–843.