Improving the predictive performance of a national vis-NIR spectroscopic library by comparing clustering data transformation, and data-mining calibration techniques

Suzana Romeiro Araújo [1]
José Alexandre Melo Demattê [1]
Marston Héracles Domingues Franceschini[1]
Rodnei Rizzo [1]
Bo Stenberg [2]
Johanna Wetterlind[2]

[1] Universidade de São Paulo - USP/ESALQ
Caixa Postal 9 - 13416-000 - Piracicaba - SP, Brasil
suzanaromeiro@yahoo.com.br, jamdemat@usp.br, marston.franceschini@usp.br, rodnei.rizzo@gmail.com

[2] Swedish University of Agricultural Sciences - SLU
Box 234, SE-532 23 Skara, Sweden
bo.stenberg@mark.slu.se, Johanna.wetterlind@mark.slu.se

Abstract. Effective agricultural planning requires basic soil information. In recent decades near-infrared diffuse reflectance spectroscopy (NIRS) has been shown to be a viable alternative for rapidly analyzing soil properties. We studied 7171 samples of the soil Brazilian spectral library. The aim was to explore the possibility of enhancing the performance of NIRS data in predicting organic matter and clay content in this library by dividing it into smaller sub-libraries based on their vis-NIR spectra and to compare these results to two nonlinear calibration techniques (BT and SVM) applied to the whole library. The general predictive models for clay performed well ($R^2 > 0.79$), reflecting the influence of the direct spectral responses of this property in the NIRS range. Predictions of OM were reasonably good, especially with clustering, and in view of the very low variation in this parameter. Results showed that the division of the large library into smaller subsets based on the variation in the mean-normalized spectra was the best alternative for using vis-NIR spectra to quantify soil attributes in tropical soils by Partial Least Square regressions. This divided the global data set into clusters that were more uniform in mineralogy, regardless of geographical origin, and improved predictive performance. Another alternative would be to use boosted regression trees for the whole library. It was also possible to identify regions of the vis-NIR spectrum that showed absorption features due to water, iron oxides and clay minerals that their variation might be responsible for the cluster divisions.

Palavras-chave: espectroscopia de reflectância difusa, matéria orgânica do solo, argila, regressão PLS, regressão de árvores, máquina de vetor.

Key words: diffuse reflectance spectroscopy, soil organic matter, clay, PLS regression, support vector machine learning, boosted regression trees

1. 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. Methods used to determine texture (Gee et al., 1986) and organic matter content (Raij et al., 2001) in conventional soil laboratories in Brazil and elsewhere are expensive, time-consuming and can be environmentally hazardous. Thus there is 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 reasonable costs. 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 (e.g. Nanni and Demattê, 2006; Bellinaso et al., 2010; Wetterlind et al., 2008; Stenberg et al., 2010; Demattê et al., 2010).

The absorption of vis-NIR light occurs due to overtones and combinations of fundamental molecular absorptions in the mid-infrared region and is associated with soil moisture, organic materials, and mineralogy. As clay particles consist mainly of clay minerals, vis–NIR spectra can be assumed to be of value for predicting clay content (Stenberg et al., 2010). OM can be related directly to the absorption of vis-NIR spectra through a number of functional groups such as the carboxyl, hydroxyl and amine groups (Clark et al., 1990).

It is often suggested that libraries containing smaller soil variation at the field scale would result in better OM predictions than more general ones collected over larger geographic areas (e.g. Kuang and Mouazen, 2011). However, Stenberg et al. (2010), reviewing published predictions, found that variation in the texture or SOC variables themselves accounted for a majority of the variation in model accuracy for texture and SOC, respectively, and that the size of the geographic area had a small influence. Thus, attempts to improve the prediction accuracy of a large global/national spectral library may benefit from dividing the library into smaller sub-libraries with more similar soils, regardless of the geographical origin of the samples. Because clay minerals and SOC tend to have the largest influence on soil vis-NIR spectra (Stenberg et al., 2010), dividing a global library into smaller models based on the variation in the spectra is one potential strategy for improving vis-NIR calibrations.

Partial least square regression (PLSR) is one of the most commonly used techniques to analyze this type of data. When dealing with a highly heterogeneous sample set in which measured parameters may vary considerably, the precision of linear regression techniques tends to decrease due to the non-linear nature of the relationship between spectral data and the dependent variable.

This study aims to (i) 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 pre-treatments of the spectra; continuum removal, first derivative, and mean normalization before dividing the library; (ii) compare the predictive performance of the sub-models with global models using PLSR and two multivariate data-mining techniques: boosted regression trees (BT) and support vector machines (SVM). Given the non-linear and contingent relationships between VNIR reflectance and soil composition (Clark, 1999), it was expected that BT and SVM would perform better than PLSR, since they can incorporate complex, non-linear relationships and interactions whereas PLSR is built upon linear, continuous relationships between predictors and the target variable of interest.

2. Material and Methods

For this study we used 7172 samples in the soil spectral library of the Remote Sensing Laboratory at the Soils Department, University of São Paulo. The soils in this spectral library are diverse and represent several orders of the World Reference for Soil Resources (WRB, 2006). The samples were air-dried and ground to a particle size of <2 mm before being submitted to chemical and spectral analyses. Sand (2-0.05mm), 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 et al., 1986). Organic matter (OM) content was determined by a colorimetric method (Raij et al., 2001).

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; Hatchell, 1999) (Henderson, 1992).

Prior to any model development the spectral library was randomly divided into a calibration set (CS) with 5169 samples and a validation set (VS) with 2003 samples, keeping the layers of the same soil profile together to ensure independence between CS and VS. The general approach in model development was that two major lines of calibration procedures were performed and compared. One involved straight forward on the calibration set as a whole (global models), and one involved calibrations that were performed cluster by cluster after the calibration set had beendivided into spectrally similar clusters(clustered models). The first derivative using a 2^{nd} order polynomial Savitzky-Golay smoothing over 11 points was applied as spectral pre-processing for all calibrations. This led to improved results for both clay and OM as compared to a range of other pre-treatments tested in our preliminary evaluation.

Global models were calibrated on the full, undivided calibration set (CS; n=5161). Three different calibration techniques were tested: PLSR, SVM, and BT. The PLSR technique is widely used, showing a good capacity for estimating attributes based on the spectral behavior of the soil (Vasques et al., 2008). It was performed in Unscrambler v.10.1 software on the calibration set using the orthogonalized PLSR algorithm for one Y-variable (PLS-1) and full cross-validation. The number of partial least-square (PLS) factors was chosen to minimize the root mean square error (RMSE) in the cross validation.

For the clustered models three different transformations prior to clustering were evaluated. The raw reflectance data were transformed to 1) the 1^{st} derivative Savitzky-Golay (2^{nd} order with 11 smoothing points; The Unscrambler v 10.1), 2) mean normalized (dividing each spectrum by its mean; The Unscrambler v 10.1), and 3) continuum removal (CR) calculated by a convex hull (Envi 4.5, 2008; Clark and Roush, 1984). The main purpose of the transformations was to see if they would divide the data differently and to assess what influence this would have on predictive performance for OM and clay (Figure 1).

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

For the clustered models, the validation sample first had to be assigned to one of the clusters. Thus, the success of this assignment step was included in the validation of calibrations. Discriminant analysis (Wold, 1982) models, one for each transformation, were developed to define the spectral features that separate the clusters. Score vectors from the 10 first principle 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 1).

The coefficients of determination (R^2), the root mean square error (RMSE), and the ratio of performance to deviation (RPD) were used to compare the results, calculated using the following equations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - yi^{\wedge})^2} \quad \text{and} \quad RPD = SD/RMSE$$

where *n* is the number of samples and *SD* is the standard deviation of laboratory-measured values for the property in question.

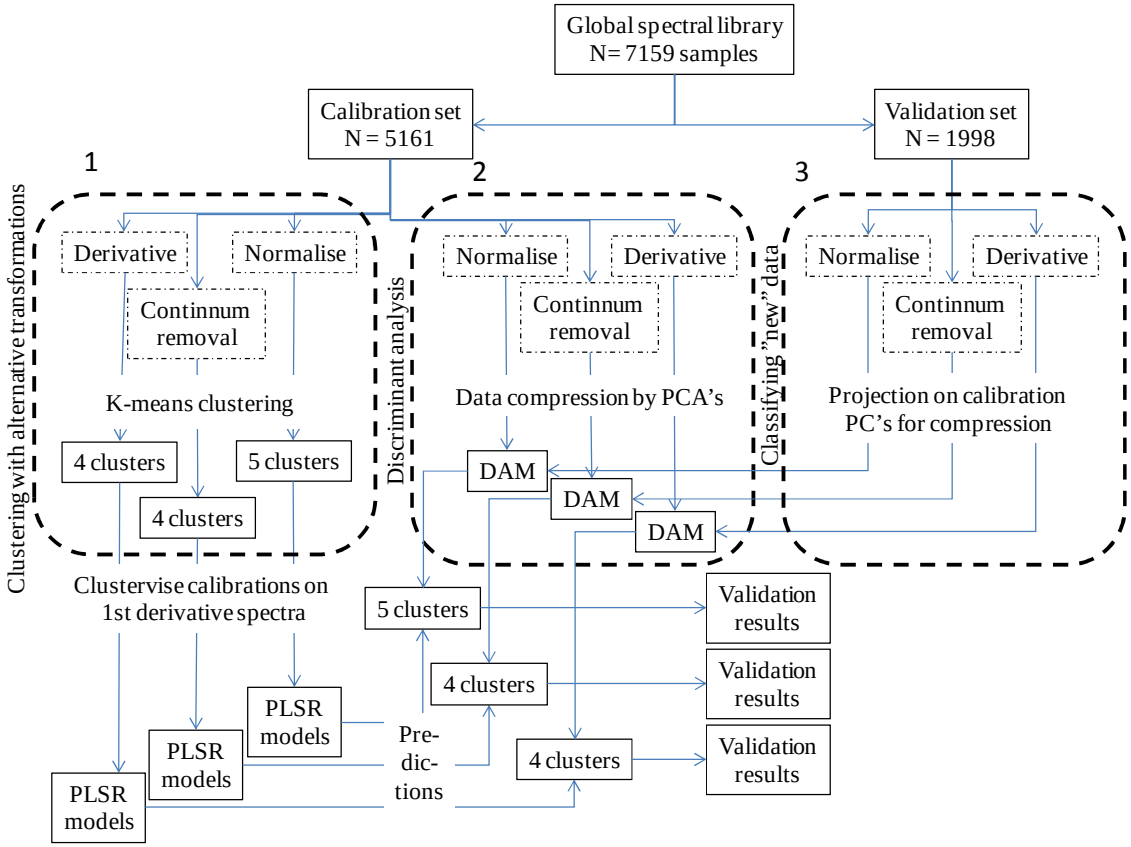


Figure 1.Overview of steps taken during the pre-processing and analyses. To be able to allocate unknown samples (the validation set) to one of the spectrally defined clusters, the spectral features defining the clusters were identified by discriminant analysis models (DAM).

3. Results and discussion

The validation results of the global predictions produced using the PLS, BT and SVM methods are summarized in Figure 2. Different regression methods provided different levels of predictive accuracy for OM and clay content. In general, we observed a tendency towards better results when using the boosted regression trees technique than SVM and PLSR, but the differences were small. These results corroborate Brown et al. (2006), who compared BT and PLS techniques for analyzing soil properties with vis-NIR and found BT to be a superior modeling approach. Those authors used 4184 compositionally diverse, well-characterized, and largely independent soil samples. In our study we also used a large number of samples (more than 7000) and a heterogeneous data set with soil properties measured in the topsoil

and subsoil in different soil orders. According to Friedman and Meulman (2003), the BT technique tends to be resistant to the effects of outliers, handling missing values and correlated variables. It also allows the inclusion of a potentially large number of irrelevant predictors (Jalabert et al., 2010). However, Viscarra Rossel and Behrens (2010) and Vasques et al. (2008) observed that BT and regression trees models produced the worst results among many multivariate techniques, including PLSR and SVM, tested to analyze total carbon, organic carbon, and clay. Those studies used 1104 samples from four regions in Australia (50% of them surface soils) and 554 samples collected to a depth of 180 cm in north-central Florida, respectively. The contrasting results reported by these authors may be due to the very diverse origin of the data sets. SVM also provided slightly better RMSE and RPD statistics than PLSR.

The spectral library data were divided into spectrally defined clusters with different numbers of samples according to the transformation employed (CR, 1st derivative, and mean normalized). This division can be attributed to the variation of clay content in the dataset, given that soil mineralogy is one of the principal factors influencing soil reflectance (Hartmann and Appel, 2006) and that the type and concentration of soil minerals are strongly correlated with soil texture through the amount of clay minerals. The key requirement for empirical modeling, that validation samples are similar to the calibration samples (Dardenne et al., 2000), was fulfilled.

The validation statistics calculated based on 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) (Table 1) showed that transforming the data using continuum removal and especially mean normalization prior to cluster analysis provided more accurate models than transforming the data by applying the 1st derivative.

Predictive accuracy differed among and within the clustering methods (Table 2). Comparing the validation results obtained from the global library and from combined cluster predictions, the data that was transformed by normalization before the clustering analysis resulted in improved validation results (Table 2).We observed a larger improvement in accuracy of models for OM than for clay with clustered models, with a reduction of the RMSE of 30 % and 17% for OM and clay, respectively.

When normalization was used as a pre-processing treatment, the global spectral library was divided into 5 clusters. The independent validation results for clusters 2 and 5 presented the highest values of RPD and R^2, followed by clusters 1, 4, and 3, respectively for clay and 3, 1, and 4 for OM.

The results of the RMSE (Table 1) reveal higher values of model errors when the 1st derivative was applied before the cluster analysis, with mean values of RMSE of 12.84 for clay and 0.59 for OM. According to Brown et al. (2005), the 1st derivative analysis can introduce instability and noise to soil reflectance data because of changing spectral contributions of soil minerals (Clark and Roush, 1984; Kokaly and Clark, 1999). This may have reduced the accuracy of the discriminant analysis of our data.

The success of assigning the validation samples to the right cluster by discriminant analyses on normalized data can be seen in Figure 2, which shows the linear discriminate analyses projection of the 5 clusters. The relevance of using normalization transformation is in accordance with other authors who found this pre-processing to improve soil property calibrations. For example, Kuśnierek (2011), in a study of Polish soils, observed that this transformation was the best of several for SOC modeling.

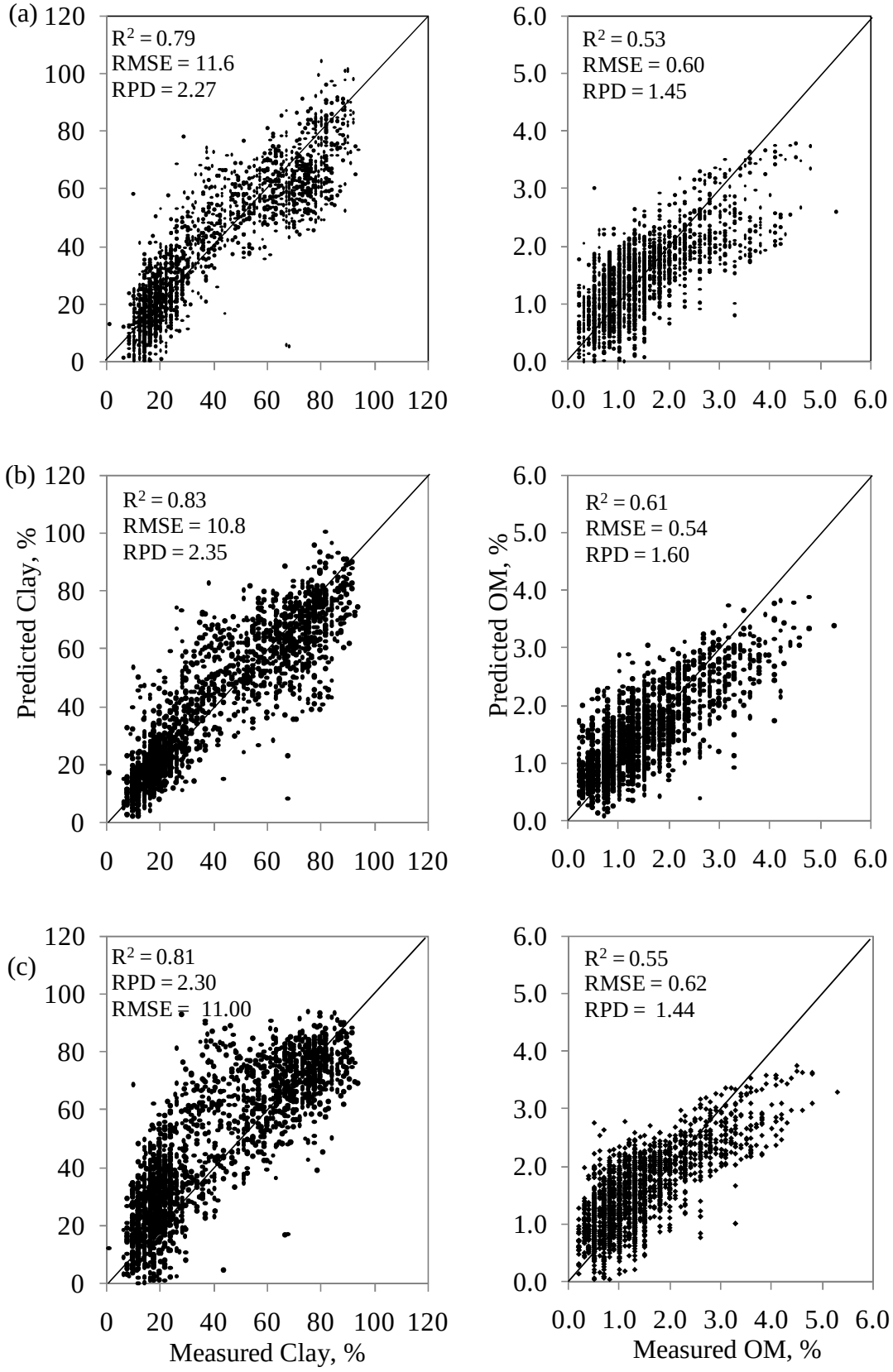
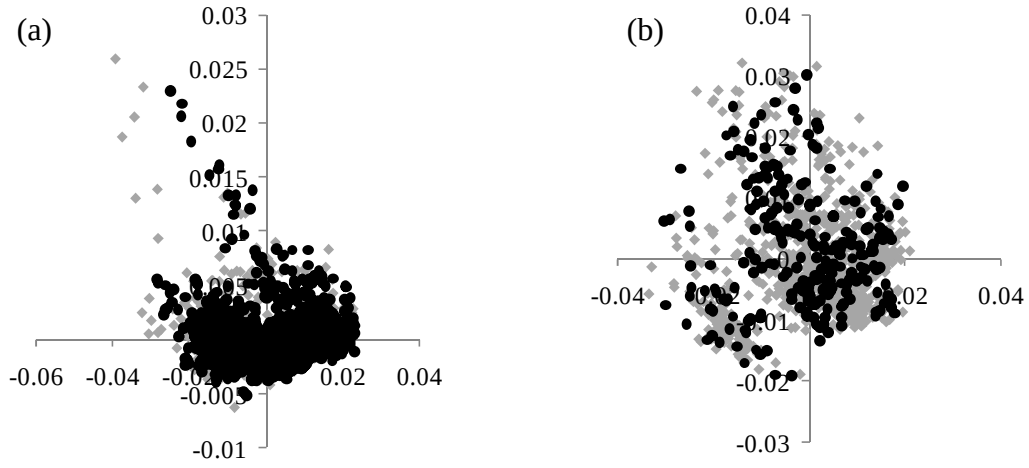


Figure 2.Validation scatter plot of laboratory-measured data versus vis-NIR predictions obtained from (a)partial least square regression, (b) boosted tree regression and (c) support vector machine for organic matter (OM) and clay content.

Table 1 - Summary statistics of validation results of calibrations for clay (%) and OM (%) using an independent validation dataset

Preprocessing	Cluster	Number of samples	Clay			OM		
			R^2	RMSEv	RPD	R^2	RMSEv	RPD
	1	512	0.75	9.63	1.96	0.54	0.51	1.47
Normalized	2	238	0.83	10.43	2.37	0.76	0.53	2.02
	3	347	0.40	11.13	1.46	0.53	0.47	1.81
	4	175	0.68	9.15	1.73	0.40	0.55	1.29
	5	726	0.77	7.80	2.07	0.58	0.27	2.85
CPR - N			**0.87**	**9.28**	**2.74**	**0.60**	**0.42**	**2.07**
1st derivative	1	802	0.61	13.34	1.66	0.40	0.56	1.35
	2	687	0.79	12.65	2.18	0.62	0.53	1.78
	3	195	0.57	11.62	1.19	0.30	0.61	1.24
	4	314	0.52	12.75	1.45	0.27	0.76	1.14
CPR - D			**0.76**	**12.84**	**1.98**	**0.30**	**0.59**	**1.47**
CR	1	289	0.53	12.22	1.43	0.60	0.53	1.54
	2	312	0.61	13.89	1.54	0.60	0.65	1.55
	3	794	0.85	9.90	4.46	0.60	0.30	2.61
	4	603	0.76	10.30	2.02	0.59	0.38	2.25
CPR - CR			**0.81**	**10.98**	**2.31**	**0.56**	**0.41**	**2.12**
PLS	**all**	**1988**	**0.79**	**11.16**	**2.27**	**0.52**	**0.60**	**1.45**
BT	**all**	**1988**	**0.83**	**10.80**	**2.35**	**0.61**	**0.54**	**1.60**
SVM	**all**	**1988**	**0.81**	**11.00**	**2.30**	**0.55**	**0.62**	**1.40**

The number of samples was based on discriminant analyses. *PLS, 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, andCPR - CR refer to combined prediction results(CPR) with mean normalization (N), 1st derivative (D),and continuum removal (CR) as pre-transformation treatments, respectively.



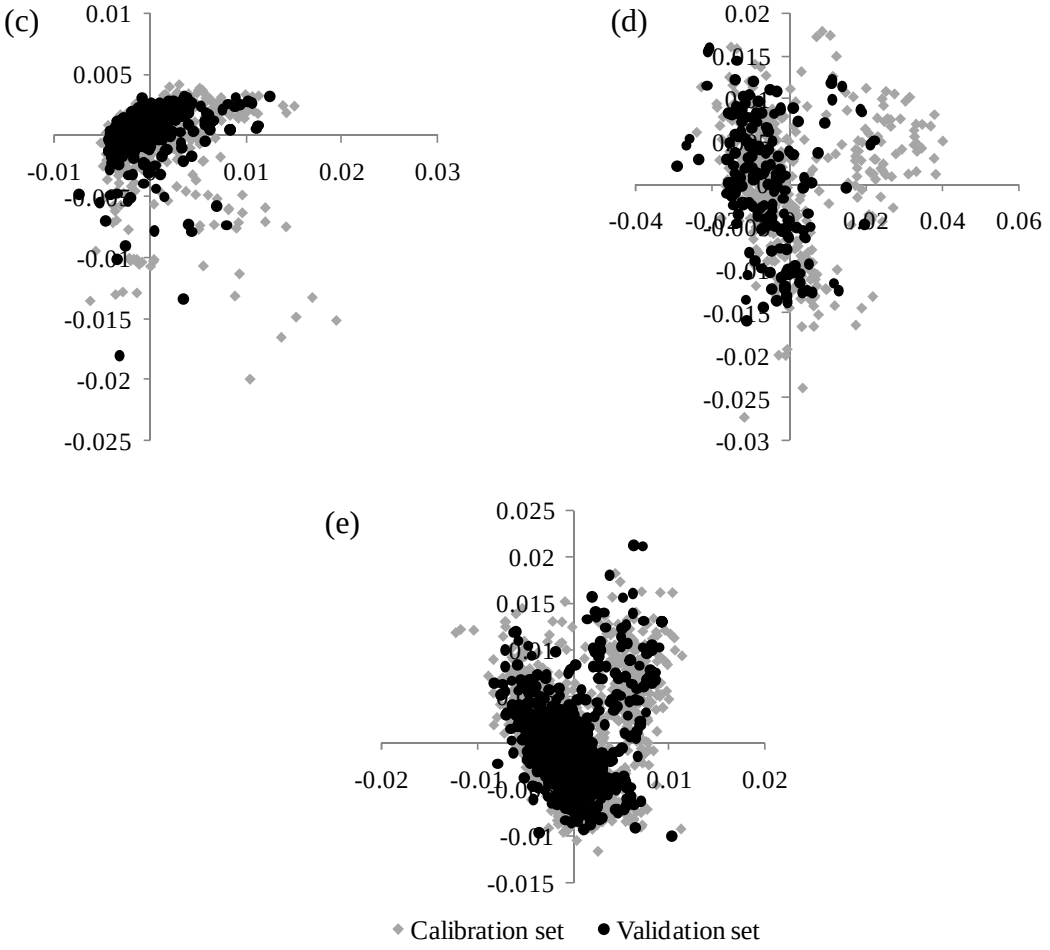


Figure 2. Projection of linear discriminant analysis clusters, obtained with normalized transformation, on the corresponding K-means cluster. The letters a, b, c, d, and e refer to cluster numbers 1, 2, 3, 4, and 5, respectively.

In our study, the additional step of assigning validation samples to the right class in the prediction process did not add substantially to the overall prediction error. We observed that cross validation (which does not involve sample-to-cluster assignment) and independent validation (which does) results did not differ substantially more for the clustered models as compared to the global models. If the assignment of validation samples to clusters added substantially to the prediction error, a larger difference for the clustered models would be expected.

4. 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. OM predictions were reasonably good, especially with clustering and in view of the very low variation in OM levels in the data set. The division of the large library into smaller subsets based on variation in the mean-normalized spectra was the best alternative for using vis-NIR spectra to quantify soil attributes in tropical soils by Partial Least Square regressions. It divided the global data set into more mineralogically uniform clusters, regardless of geographical origin, and improved predictive performance. The additional step of assigning the validation samples to the right class in the prediction process (clustered models) did not add substantially to the overall

prediction error. Another alternative would be to use Boosted regression trees for the whole library. Comparing the results of this study and previously published ones indicates that the selection of the best performing pre-processing method is dataset-dependent.

References

Bellinaso, H.; Demattê, J.A.M.; Araújo, S.R. Soil spectral library and its use in soil classification. 2010. *RevistaBrasileira de Ciência do Solo*, Viçosa, 34, p. 861-870.

Brown, D.; Shepherd, K. D.; Walsh, M. G.; Mays, M. D.; Reinsch, G. 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, Amsterdan, 132, p. 273-290.

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

Clark, R.N. 1999. Spectroscopy of rocks and minerals and principles of spectroscopy. In: RENCZ, A.N. (Ed.). *Remote sensing for the earth sciences*. Toronto: John Wiley, chap. 1, p. 3-58.

Clark, R. N.; King, T. V. V.; Klejwa, M.; Swayze, G.; Vergo, N. 1990. High resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research*, Washington, v. 95, p. 12653-12680.

Clark, R.N.; Roush, T.L. 1984.Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research*, Washington,89, p. 6329–
6340.

Dardenne, P.; Sinnaeve, G.; Baeten, V. 2000. Multivariate calibration and chemometrics for near infrared spectroscopy: which method? J. *Near Infrared Spectroscopy*,8 (4), pp. 229–237.

Demattê, J. A. M.; Nanni, M. R.; Silva, A. P.; Melo Filho, J. F., Santos, W. C.; Campos, R. C. 2010. Soil density evaluated by spectral reflectance as an evidence of compaction effects. *International Journal of Remote Sensing* (Print), 31, p. 403-422.

ENVI. Environment for Visualizing Images.Guia do ENVI 3.5 em Português. Disponível em: <http://www.envi.com.br>. Acesso em: 2012.

Friedman, J. H.; Meulman, J. J. 2003. Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22,1365–1381.

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

Hartmann, H. P.; Appel, T. 2006. Calibration of near infrared spectra for measuring decomposing cellulose and green manure in soils. Soil Biology & Biochemistry, 38 pp. 887–897.

Jalabert, S. S. M.; Martin, M. P.; Renaud, J. P.; Boulonne, L.; Jolivet, C.; Montanarella, L.; Arrouays, D. 2010. Estimating forest soilbulk density using boosted regression modelling. *Soil Use Management.* 26: 516−528.

Kokaly, R. F.; Clark, R. N. 1999. Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. *Remote Sensing of Environment*, 67 pp. 267–287.

Kuang, B.; Mouazen, A. M. 2011. Calibration of a visible and near infrared spectroscopy for soil analysis at field scales across three European farms. *European Journal of Soil Sciences*, 62, 629–636.

Nanni, M. R. ;Demattê, J. A. M. 2006. Spectral reflectance methodology in comparison to traditional soil analysis. Soil Science Society of America Journal, Madison, 2, n. 70, p. 393-407.

Raij, B. van.; Andrade, J. C.; Cantarela, H.; Quaggio, J. A. Análise química para avaliação de solos tropicais. Campinas: IAC, 2001. 285p.

Stenberg, B.; Viscarra Rossel, R. A.; Mouazen, A. M.; Wetterlind, J. 2010.Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*, Amsterdam, 107, n. 107, p. 163-215.

Vasques, G.M.; Grunwald, S.; Sickman, J.O. 2009.Modeling of soil organic carbon fractions using visible-near-infrared spectroscopy.*Soil Science Society of America Journal*, Madison, 73, p. 176-184.

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

Wetterlind, J.; Stenberg, B.; Söderström, M.2008.The use of near infrared (NIR) spectroscopy to improve soil mapping at the farm scale.*Precision Agriculture*, Berlin, 9, p. 57-69.

Wold, H. 1982. Soft modeling. The basic design and some extensions. In: Joreskog, K.-G., Wold, H. (Eds.), Systems Under Indirect Observation, Vols. I and II. North-Holland, Amsterdam (Chapter 1 of Vol. II).

IUSS Working Group WRB. 2006: World reference base for soil resources 2006 (2nd ed). World Soil Resources Report 103. FAO, Rome.