Application of LS-SVM-NIR spectroscopy for carbon and nitrogen prediction in soils under sugarcane

Sandra Oliveira Sá[A],, Marco Flores Ferrão[B], Marcelo Valadares Galdos[C], Carla Maris Machado Bittar[D] and Ronei Jesus Poppi[E]

[A]Universidade Estadual do Maranhão, São Luis, MA, 65055-310. Brazil, E-mail: sa.oliveiras@gmail.com
[B]Universidade de Santa Cruz do Sul, 96815-900, Santa Cruz do Sul, RS, Brazil.
[C]Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, 13400-970, Piracicaba, SP, Brazil.
[D]Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, 13418-900, Piracicaba, SP, Brazil.
[E]Instituto de Química, Universidade Estadual de Campinas, 13083-970, Campinas, SP, Brazil.

Abstract
In this paper, Least-Square Support Vector Machine (LS-SVM) regression is used to for a rapid and accurate quantification of total carbon (total-C) and total nitrogen (total-N) in soil samples collected in forest and sugarcane areas in Sao Paulo State, Brazil. NIRS spectra were recorded on a NIRS 5000 scanning monochromator. The concentration ranges of 0.401-3.101 %, for total-C, and 0.030-0.252 %, for total-N were obtained for the references. The performance and robustness of LS-SVM regression are compared to Partial Least Square Regression (PLSR). For total-C, correlation coefficients (R^2cal) of 0.99 and 0.92, RMSECV of 0.132 and 0.176 %, and RMSEP of 0.110 and 0.141 % were obtained for LS-SVM and PLSR, respectively. For total-N, correlation coefficients (R^2cal) of 0.98 and 0.92, RMSECV of 0.015 and 0.013 %, and RMSEP of 0.008 and 0.009 % were obtained for LS-SVM and PLSR, respectively. At the same time, results indicate that LS-SVM-NIRS can be used with advantage as an analytical method for rapid, accurate, reliable and cost-effective routine analysis of total carbon and nitrogen in tropical soil.

Key Words
NIR, chemometrics, sugarcane, tropical soil, carbon, nitrogen.

Introduction
The Support Vector Machine (SVM) is a relatively new nonlinear technique in the field of chemometrics and is employed basically in classification and multivariate calibration problems (Thissen *et al*. 2003). Recently an extension of SVM, called Least Square Support Vector Machines (LS-SVM) was introduced (Codgill and Dardenne, 2004). LS-SVM is capable of dealing with linear and nonlinear multivariate calibration and resolves multivariate calibration problems in a relatively fast way. In the LS-SVM a linear function ($y = w x + b$) is fitted between the dependent (y) and independent (x) variables. As in SVM, it is necessary to minimize a cost function (L) containing a penalized regression error. The aim of this work was to propose the use of least-squares support vector machine (LS-SVM) and NIR spectroscopy with diffuse reflectance as a methodology for quantification of total carbon (total-C) and total nitrogen (total-N) in Brazilian soils from sugarcane cultivation. Brazil is today the largest sugarcane produce in the world attaining nearly 5 million hectares of planted area to produce 643,7 million tones of cane stalks in the 2007/08 season. About 386 million tones are produced in the State of Sao Paulo alone.

Methods
Study area and sampling
A total of 250 soil samples (0-100 cm depth) were collected in a sugarcane plantation located in Pradópolis (21^0 22' of S. 48^0 03' W.) in São Paulo State, Brazil. The soil is a Typic Haplodux, with clayey texture. According to the Köppen classification, the climate is an Aw type, tropical wet with a dry winter, with average annual precipitation close to 1,560 mm/year. The average annual air temperature is 22.9 ^{o}C, and the average monthly temperatures are above 18.0 ^{o}C. Sugarcane had been grown for the factory for at least fifty years. Four fields with mechanical pre-harvesting were selected, using the method of chronossequence, where sugarcane had been harvested, without replanting or soil disturbance, for 8, 6, 4, and 2 years. Soil samples were also collected in an area of native forest, as a reference. The sampling was done in a grid system with nine replications in each field, to depths of 0-10, 10-20, 20-30, 40-50, 70-80 and 90-100 cm.

Reference analyses and Spectral measurements
Samples were air-dried, sieved and grounded to 60 mesh before analysis. Reference analyses for total C and total N were performed by dry combustion on a LECO CN 2000 elemental analyzer (furnace at 1200 ^{0}C in

pure oxygen). The principle is to convert all the different forms of carbon into CO_2 to be measured quantitatively by infrared. In addition, the combustion process converts any nitrogen forms into N_2 and NO_x and an aliquot of the sample gas is purified by catalyst heater (NO_x gas are reduced to N2), Lecosorb (to remove CO_2) and Anhydrone (to remove H_2O). Then the N_2 can be measured by thermoelectric detector.

NIRS spectra were recorded on a NIRS 5000 scanning monochromator (Foss NIRSystems, MD). Sample ware scanned in a spinning micro sample cup and the spectra were recorded at 2-nm intervals in the range of 1100 - 2498 nm by using WINISI II version 1.05 software (Infrasoft International, Silver Spring, MD) for data acquisition. A ceramic standard was used for the background spectra and the spectra was collected as log(1/R), where R is reflectance. For NIR calibration, two multivariate regression methods were used: (i) NIR-PLS using the PLS program from PLS-Toolbox version 3.5 with Matlab from Eigenvector Research Inc. (Wise *et al*. 2005); and (ii) NIR-LS-SVM using the LS-SVMlab (Matlab/C Toolbox for Least Squares Support Vector Machines) (Suykens *et al*. 2002). All programs were run on an IBM-compatible Intel Pentium 4 CPU 3.00 GHz and 1 Gbyte RAM microcomputer. Data were treated using a multiplicative scattered correction (MSC) technique before further multivariate analysis. To evaluate the error of each calibration model, the root mean square error was used, calculated by eq. 1.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (1)$$

Results

All spectra were pre-processed by multiplicative scatter correction (MSC), aiming at correcting the baseline deviation between the spectra. Figure 1 shows the results of the PCA for all soil samples studied. We have mapped the data on three most important principal components PC_1, PC_2 and PC_3 presenting the distribution on data in three-dimensional plot which explain 99.90% of the original information. Analysing the distribution of data mapped on the principal components analysis it is evident that the data points belonging to three distinct classes, which represent three different stages of the dynamic soils for C and N throughout the time. The first cluster (A) contains soil samples for two years and forest, the second (B) contains soil samples for four years, and the third cluster (C) contains soil samples with six and eight years of sugarcane cultivation (Chronosequence).

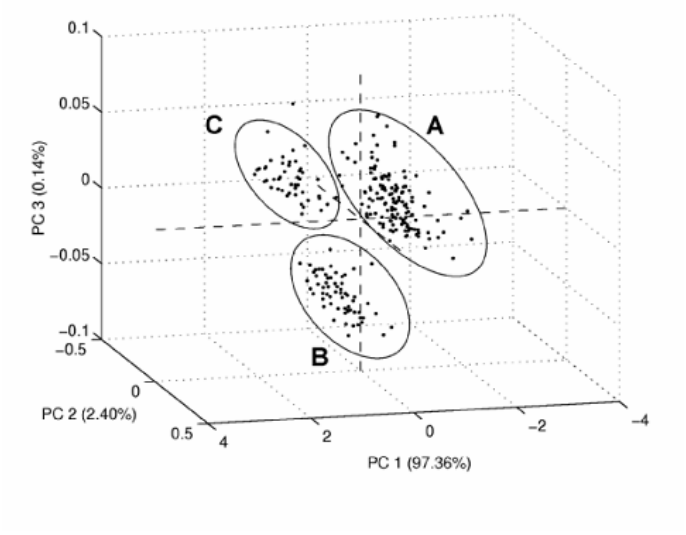

Figure 1. The three-dimensional scores (PC_1xPC_2xPC_3) of PCA plots using spectra soil sugarcane data.

In Figure 2 the optimized surfaces result for the LS-SVM model, using the calibration set is shown. The γ and σ^2 parameters were a manageable task, similar to the process employed to select the number of factors for PLS models, but in this case in two-dimensional problem. The cross-validation procedure has been used for to the determination of the RMSECV.

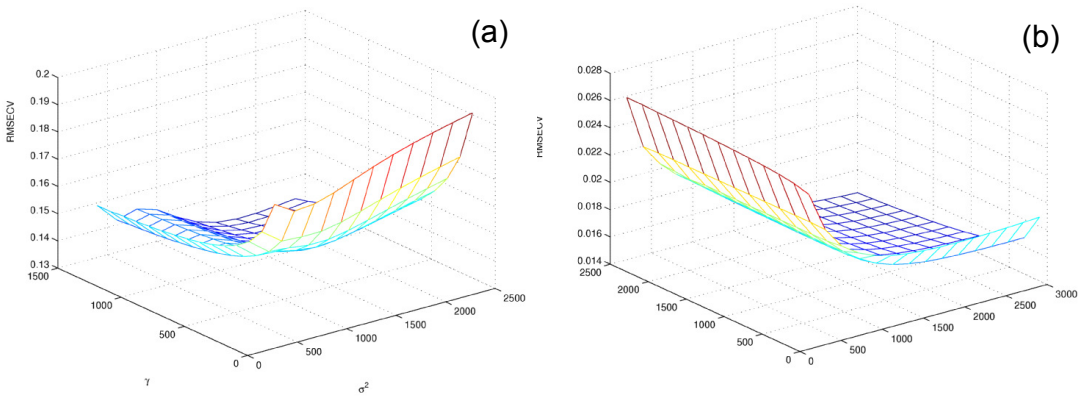

Figure 2. Parameter optimization response surfaces using LS-SVM for total-C (left) and total-N (right) in soil samples.

We compare the coefficient of correlation (R^2cal) for the calibration model, and RMSECV and RMSEP for the different γ and σ^2 parameter combinations. If both parameters are increased the RMSECV and RMSEP values decrease (Figure 2). For total-C quantification model, σ^2 = 1400 was used and increased values of γ, the RMSE values were optimized even for γ = 1600, and for values above of 1600 the RMSECV and RMSEP values did not vary. When γ = 1600 was used and σ^2 decreased the RMSECV and RMSEP values increased. It was possible to observe that when the γ values increased towards infinity, the RMSECV value tends to a minimum. However, this would likely lead to an over-fitted calibration. For total-N when γ = 1500 was used and it increased σ^2 the RMSE values were optimized even for σ^2 = 2200, and for values above 2200 the RMSECV value did not vary and RMSEP increased with the risk of over-fit. When σ^2=2200 was used and γ was decreased the RMSECV and RMSEP values increased.

Table 1 presents the results for LS-SVM models for total-C determination using γ =1600 and σ^2=1400 and for total-N using γ =1500 and σ^2= 2200. When the results for PLS and for LS-SVM models were compared, both presented good correlation coefficients (R^2cal), but when procedures were compared RMSEP values for total-C and total-N quantification LS-SVM was better.

Table 1. Performance comparison results between PLS and LS-SVM for total-C and total-N quantification.

	total-C		total-N	
	PLS	**LS-SVM**	**PLS**	**LS-SVM**
$\mathbf{R^2_{cal}}$	0.920	0.995	0.918	0.985
RMSECV (%)	0.1762	0.1318	0.0133	0.0151
RMSEP (%)	0.1410	0.1101	0.00948	0.00837
LV	9	-	12	-
γ	-	1600	-	1500
σ^2	-	1400	-	2200

The graphs between reference and NIR predicted total-C values are presented in Figure 3 for the 166 calibration and 78 prediction spectra samples. The correlation coefficients (R^2cal) between PLS and LS-SVM models for total-C content were found to be the best. However it is possible the PLS model presents greater error for the samples with high and minors values of total-C, which indicates that the LS-SVM calibration model could be used for extreme values.

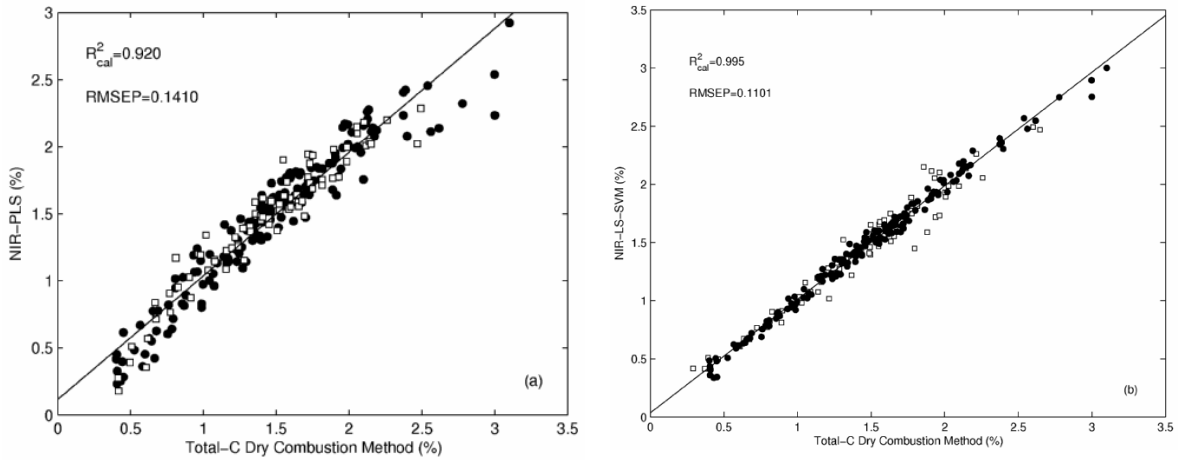

Figure 3. Calibration and prediction plot of % total-C by NIR-PLS model using 9 latent variables (a) or LS-SVM model (b).

For total-N 167 calibration and 75 prediction spectra samples were used. The correlation coefficients of calibration models (R2cal) between PLS and LS-SVM models were found to be the best. It is possible see that the LS-SVM presents excellent prediction abilities when compared with PLS regression.

Conclusion

The use of the LS-SVM for to quantify total-C and total–N in soil samples under sugarcane cultivation results in robustness calibration method with respect to a heterogeneous set soil samples. PCA analyses indicated that there are three distinct groups in the set of samples according to land use and time of cultivation of sugarcane. Furthermore, we demonstrate the capacity for generalization and flexible application of the LS-SVM procedure when combining different data sets. Finally, LS-SVM are promising techniques to use for estimation of soil quality from indirect but fast and reliable measurements such as near infrared spectra.

Acknowledgements

This work was supported by the FAPESP, FAPEMA, CAPES and CNPq.

References

Codgill RP, Dardenne P (2004) *Journal of Near Infrared Spectroscopy* **12**, 93-100.

Suykens JAK, van Gestel T, Brabanter J, Moor B and Vandewalle J (2002) Least-Squares Support Vector Machines. World Scientific, Singapore.

Thissen U, van Brakel R, Weijer AP, Melssen WJ, Buydens LMC (2003) *Chemometrics and Intelligent Laboratory Systems* **69**, 35-49.

Wise BM, Gallagher NB, Bro R, Shaver JM, Windig W, Koch RS (2005) PLS Toolbox 3.5 for use with MATLAB, Eigenvector Research Inc., Manson, WA.