Soil properties determinations by vis-near infrared spectroscopy using boosted regression trees

R. Joffre,[a,*] F. Gogé,[a] C. Jolivet[b] and N. Saby[b]

[a]*CEFE-CNRS, UMR5175, 1919 Route de Mende, F-34293 Montpellier cedex 5, France. E-mail: richard.joffre@cefe.cnrs.fr*

[b]*INRA, Unité Infosol, BP 20619, F-45166 Olivet, France*

Introduction

Several chemometric techniques allow us to unravel spectra and to calibrate the VIS-NIRS signal, i.e. to relate the spectra of samples to their laboratory reference values. The most common include Partial Least Squares Regression (PLSR), Multiple Linear Regression (MLR) and Principal Component Regression (PCR). There are also non parametric methods such as Artificial Neural Network (ANN) or Regression Trees (RT).[1] RT partition samples into groups having similar values for the response variable based on a series of binary rules constructed from the predictor variables, and model complex interactions between predictors. Boosted methods improve the performance by creating an ensemble of hundreds of simple trees, each of which is adapted to each sample, and by combining them in a forward procedure.[2,3] One of the main interests of the regression tree family of methods consists in the integration of additional data either quantitative or qualitative in the model. Though direct graphic representation of the complete tree model is impossible with boosted regression trees, the model interpretation is made easy by identifying the variables most relevant for prediction, and then visualising the partial effect of each predictor variable, after accounting for the average effect of the other variables.[2] A few papers have used RT models to model soil properties using Vis-NIR spectroscopy.[4,5] Our objective in this study was to evaluate the accuracy of calibrations for prediction of soil properties, comparing PLS and BRT models, based on Vis-NIR spectra and additional data on a large database encompassing a wide array of soils.

Materials and methods

The soils came from the RMQS network soil library (French soil quality monitoring network) representing soils sampled with a 16 × 16 km systematic grid covering the whole French territory. A NIRSystem Model 6500 spectrophotomer (Foss Analytical) was used to record reflectance spectra of 1986 soils at 2 nm intervals between 400 to 2500 nm. Classical physico-chemical properties including organic carbon (OC), nitrogen (N), cation exchange capacity (CEC), metal concentration, textural fraction, were analysed.

PLS calibrations were built on the first derivative of spectra using the NIPALS algorithm included in the PLS Toolbox (Eigenvector Research Inc., Manson, WA, USA). Boosted regression trees (BRT) were built using the gbm R package.[6] The main parameters for fitting BRT are learning rate (LR), tree complexity (TC), minimum number of observations (Min obs.) per terminal node and bag fraction. The learning rate determines the contribution of each tree to the model. Tree complexity defines the maximal numbers of nodes in the individual trees. A bag fraction of 0.75 was used which means that, at each step of the boosting procedure, 75% of the data in the training set were drawn at random without replacement. A variety of models were next run combining three learning rates (0.05 0.005 and 0.0005), four levels of tree complexity (5, 10, 15, 20) and three thresholds of minimum observations (5, 10, 15).

A first experiment compared PLS and BRT regressions for five chemical variables – organic carbon (OC), total nitrogen (N), cation exchange capacity (CEC), total iron (Tot-Fe) and total magnesium Tot-Mg - built on 210 spectral predictors (each 10 nm between 400 and 2500 nm). A second experiment explored the BRT regressions for the same variables built on spectral plus chemical (pH, CEC) and textural (sand, clay) and geographic (latitude, longitude, elevation, land use type) predictors. CEC was not considered as a predictor when calibrating it. Calibrations were done on 1486 samples and validation on 500 independent samples.

Results

Figure 1 showed the variation of the *RMSEP* of OC with the number of trees for distinct values of learning rate, tree complexity and minimum observation respectively.

Convergence between *RMSEP* was obtained with learning rate values of 0.05 and 0.005 whilst a lower rate of 0.0005 did not allow *RMSEP* to decrease rapidly [Figure 1(a)]. As a consequence a LR value of 0.005 was fixed for all subsequent models. The curves between *RMSEP* and number of trees were quite distinct when modifying the minimum number of observations with significant lower *RMSEP* with a value of 3 [Figure 1(b)]. No clear pattern emerged due to interactions between tree complexity and number of trees in Figure 1(c). Accordingly, three levels of minimum observations (3, 6 and 9) and five values of TC ranging for 10 to 20 were systematically tested in

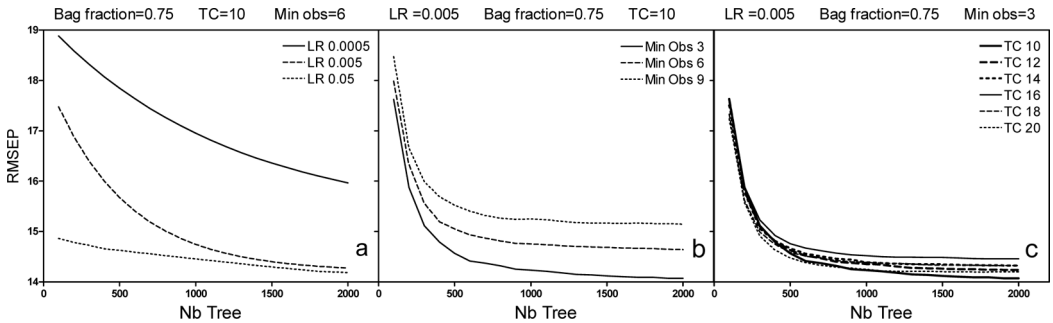

Figure 1. Variation of the *RMSEP* of organic carbon with the number of tree for distinct values of learning rate LR (a), minimum number of observations in the terminal node Min Obs (b) and tree complexity TC (c).

Table 1. Validation statistics for vis-NIR PLS and BRT models.

		PLS model on vis-NIR data		BRT model on vis-NIR data		BRT model on vis-NIR data + chemical and geographic descriptors		*RMSEP* improvement Between complete BRT and PLS calibrations
		RMSEP	r^2	*RMSEP*	r^2	*RMSEP*	r^2	
OC	g kg^{-1}	4.82	0.80	5.01	0.78	3.63	0.88	24.7
N	g kg^{-1}	0.38	0.81	0.39	0.80	0.29	0.89	23.7
CEC	cmol kg^{-1}	2.66	0.87	2.78	0.85	1.81	0.94	32.0
Tot-Fe	%	0.51	0.77	0.52	0.75	0.44	0.82	13.7
Tot-Mg	%	0.14	0.67	0.14	0.65	0.13	0.73	7.1
Clay	g kg^{-1}	43.7	0.83	44.5	0.82	34.3	0.90	21.5
Silt	g kg^{-1}	94.7	0.70	95.8	0.68	86.5	0.74	8.7

the subsequent models for all studied variables. We presented hereafter only the best calibration for each constituent.

Validation statistics of the PLS model and the BRT models for each constituent were presented in Table 1.

When based only over spectral data, BRT models gave very close results to PLS models whatever the variable under study. Adding non-spectral descriptors as chemical, textural and geographic descriptors improved the calibrations for all constituents significantly. The improvment

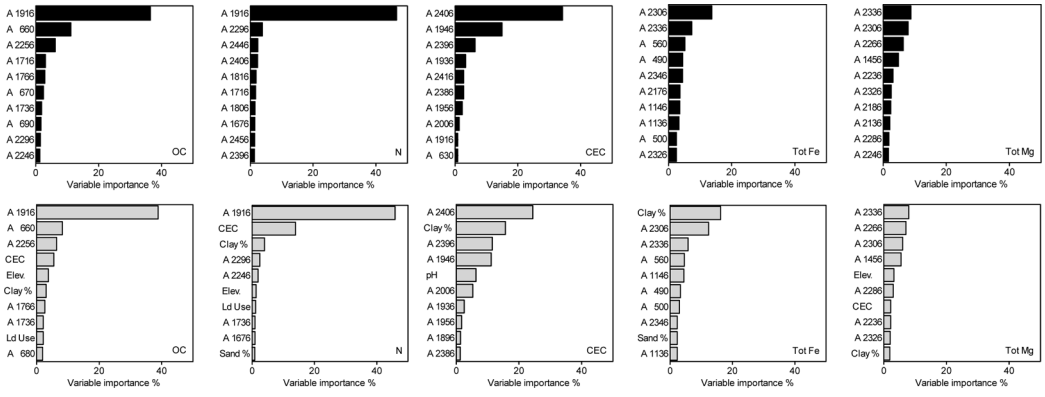

Figure 2. Relative variable importance for the ten most important predictors of boosted regression models based on vis-NIR predictors (upper graph, black bars) and vis-NIR plus physic-chemical and geographic descriptors (lower graph, grey bars). Figures correspond to organic carbon (OC), nitrogen (N), cation exchange capacity (CEC), toal iron (Tot-Fe) and total magnesium (Tot-Mg).

of root mean square error of prediction (*RMSEP*) ranged from 7.1% (Tot-Mg) to 32.0% for the chemical parameters and from 8.7% (Silt) to 21.5% (Clay) for the textural fractions.

Figure 2 showed the relative variable importance of the ten first predictors for each of the chemical variable in the two BRT models.

Not surprisingly, the wavelengths selected by the model based on spectral absorbance were also selected and generally equally ranked by the more complex model, including other non-spectral predictors. Clay content was included in all complex BRT models with a very high rank in Tot-Fe, CEC and N models. CEC improved models for OC, N and Tot-Mg significantly. pH was selected only to predict CEC values. Elevation improved OC, N and Tot-Mg and the categorical descriptor of land-use was included in the OC and N models.

Conclusion

The BRT calculation procedures allow mixing of qualitative and quantitative predictors to build predictive models. The near infrared spectra of soil are dependent on the mineral matrix and in this case, calibration of organic constituents could be difficult due to the "noise" of the physical parameters. This study suggests that the BRT approach gave the same results as PLS models when based only on spectral descriptors, but could improve the precision of the obtained calibrations significantly when other predictors relative to land use, parent material or other geographic information could be included in the dataset.

Acknowledgements

This work was funded through the ECOMIC-RMQS program by the French National research agency (ANR).

References

1. G. De'ath and K.E. Fabricius, *Ecology* **81,** 3178 (2000).
2. J.H. Friedman and J.J. Meulman, *Statistics in Medicine* **22,** 1365 (2003).
3. G. De'ath, *Ecology* **88,** 243 (2007).
4. D.J. Brown, K.D. Shepherd, M.G. Walsh, M.D. Mays and T.G. Reinsch, *Geoderma* **132,** 273 (2006).
5. G. Vasques, S. Grunwald and J. O. Sickman. *Geoderma* **146,** 14 (2008).
6. G. Ridgeway, *Generalized boosted regression models, R package*, Version 1.5 (2006).