



Soil and Water Science Department
University of Florida
2169 McCarty Hall, P.O. Box 110250
Gainesville, FL 32611-0250
Phone: 352-392-1161 ext. 233
Fax: 352-392-3902
E-mail: gmvasesques@yahoo.com.br

Comparison of Methods for the Assessment of Soil Organic Carbon Using Visible/Near-Infrared Reflectance Spectroscopy

Gustavo M. Vasques, Sabine Grunwald, and James O. Sickman

Soil and Water Science Department, University of Florida



Introduction

In the last decades, models to predict soil properties have become more accurate and less costly. Advances in information technology and the development of new sensors and instruments have facilitated the collection and analysis of data, making possible the formulation of more complex models. Carbon is of great importance to soils. It has a strong relationship with soil organic matter, influencing the soil physical, chemical and biological processes. In addition, soil is a potential reservoir to sequester atmospheric CO₂ and mitigate global warming. Hence, the analysis of the distribution and dynamics of soil carbon is an essential requirement for sustainable land management. Visible/near-infrared spectroscopy (VNIRS) is a fast, cheap and accurate alternative for the investigation of soil properties, and is gradually becoming recognized as a powerful analytical tool in soil science.

Objectives

To identify, among thirty pre-processing transformations of soil VNIR reflectance spectra, and five calibration methods, the best combination to estimate soil total organic carbon using VNIRS.

Methods

Lab analysis/spectroscopy: Total organic carbon (TOC) was measured with a Thermo Electron FlashEA Elemental Analyzer; VNIR spectra were derived with an ASD QualitySpec Pro spectroradiometer (350-2500 nm).

Pre-treatment of TOC: Log-normalization using base-10 logarithm.

Validation: 400 observations were randomly separated for calibration, leaving 154 observations for validation.

Pre-processing transformations: 30 techniques were tested for each method, except ANN.

Methods: Stepwise Multiple Linear Regression (SMLR) with a stepping probability of 0.05 (SPSS 11); Principal Components Regression (PCR) (Unscrambler 9.5); Partial Least-Squares Regression (PLSR) (Unscrambler 9.5); Regression Tree (RT) (CART 5.0); Artificial Neural Networks (ANN) using a perceptron, with hyperbolic tangent transfer function, conjugate gradient learning, and 20,000 epochs (NeuroSolutions 4.0).

Comparison of methods and pre-processing transformations: The best methods and pre-processing techniques were selected based on the coefficient of determination of calibration (R_c^2) for the RT and of validation (R_v^2) for all other methods.

Pre-processing technique*	Technique details	Search window
Savitzky-Golay Smoothing ¹ , and Averaging ² (SA)		9, and ² 10
Baseline Correction		
Kubelka-Munk Transformation (K-M)		
Log(1/Reflectance)		
Normalization	By the maximum By the mean By the range	
Norris Gap Derivative (NGD)		3, 5, 7, and 9
Savitzky-Golay First Derivative (SG-1D)	1 st -order polynomial	3, 5, 7, and 9
	2 nd -order polynomial	3, 5, 7, and 9
	3 rd -order polynomial	5, 7, and 9
Savitzky-Golay Second Derivative (SG-2D)	2 nd -order polynomial	3, 5, 7, and 9
	3 rd -order polynomial	5, 7, and 9
Standard Normal Variate (SNV)		

* Savitzky-Golay smoothing, and averaging, were used as a standard preparation of the soil spectral curves to reduce noise and match the resolution of the instrument. This standard curve was used as the input to all other pre-processing transformations.

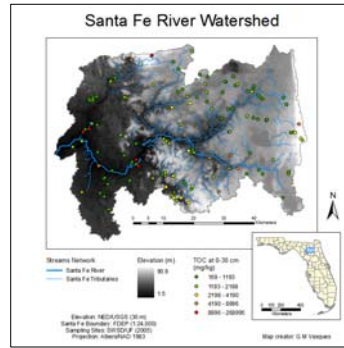
Study Area

Study area: Santa Fe River watershed (3,585 km²) in north-central Florida.

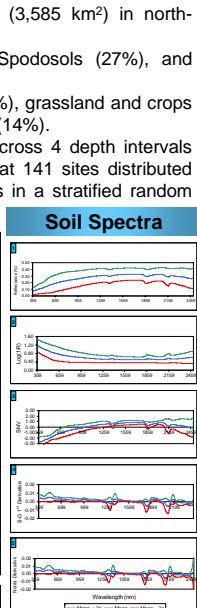
Dominant soil orders: Ultisols (47%), Spodosols (27%), and Entisols (17%).

Land use/land cover: Pine plantation (30%), grassland and crops (29%), upland forest (11%) and wetlands (14%).

Sampling design: Composite sampling across 4 depth intervals (0-30, 30-60, 60-120, and 120-180 cm) at 141 sites distributed across different land uses and soil types in a stratified random design.



Soil Spectra



(1) Savitzky-Golay smoothing, and averaging; (2) Log(1/R); (3) SNV transformation; (4) SG-1D with 1st-order polynomial and window of size 9; (5) NGD with window of size 5.

Results

SMLR: The best pre-processing transformations were Log(1/R) and SNV, both with a R_v^2 of 0.854. The Log(1/R) model selected 14 predictors, while the SNV model selected 23.

PCR: The best transformations were SG-1D with 1st or 2nd-order polynomial and window of size 9 ($R_v^2 = 0.834$), using 13 principal components (PCs), followed by SA (11 PCs, $R_v^2 = 0.830$). For the SG transformations, the degree of the derivative and the size of the search window were more sensitive factors than the order of the polynomial.

PLSR: Like PCR, the best transformations were SG-1D with 1st or 2nd-order polynomial and window of size 9 (7 PCs, $R_v^2 = 0.855$), followed by SA (8 PCs, $R_v^2 = 0.854$).

RT: The best pre-processing transformation was NGD with window of size 5, followed by NGD with window of size 7 ($R_c^2 = 0.739$).

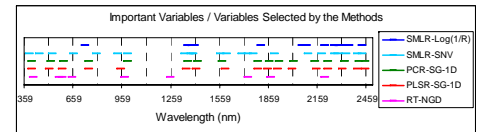
Method	$R_v^{2(1)}/R_c^{2(2)}$			
	Minimum	Maximum	Mean	Std. Deviation
SMLR ¹	0.656	0.854	0.814	0.040
PCR ¹	0.560	0.834	0.770	0.075
PLSR ¹	0.741	0.855	0.830	0.028
RT ²	0.493	0.754	0.659	0.056

Method	Number of Predictors/PCs/Terminal Nodes ³			
	Minimum	Maximum	Mean	Std. Deviation
SMLR ¹	6	35	19	7
PCR ²	7	17	12	2
PLSR ²	6	13	7	2
RT ³	3	22	10	5

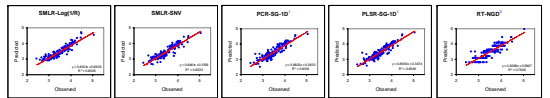
All the methods were sensitive to the regions of absorption features of C-H, O-H and H₂O. Except for RT, all methods included variables in the absorbance region of N-H.

Results

Overall, PLSR gave the most accurate predictions of Log(TOC). Some advantages of PLSR are: rapidness, ease of use, and flexibility to deal with correlated and missing data.



Multicollinearity and missing data are potential problems in SMLR, but they were not observed in this analysis. Linearity is assumed in SMLR, and to some extent in PCR and PLSR. Alternatively, non-parametric methods are more flexible to deal with non-linear relationships. In this study, RT was not well suited for the estimation of Log(TOC) using VNIRS, as it predicted discontinuous values and produced the worst results.



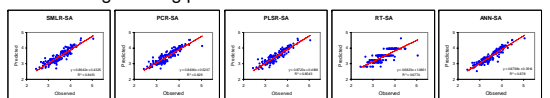
(1) SG-1D with 1st-order polynomial and window of size 9; (2) NGD with window of size 5.

ANN vs. Other Methods

ANN: The ANN method was performed using SA pre-processing. A single-layer perceptron was used because it approximates a linear least-squares estimator. An exhaustive comparison among different transfer functions, learning rules and numbers of epochs was performed to identify the best combination of learning parameters.

Comparative results: ANN outperformed all the other methods when they were calibrated using the same dataset (SA). The main advantage of ANN is its flexibility to adjust to any dataset. The main drawbacks are the non-transparency, the difficulty to find the optimum set of parameters and the long learning period.

Method	R_v^2
SMLR	0.841
PCR	0.830
PLSR	0.854
RT	0.675
ANN	0.878



Conclusions

- The best model to predict Log(TOC) using VNIRS was obtained by ANN.
- Overall performance of the methods (excluding ANN): PLSR > SMLR > PCR > RT
- Performance of the methods with SA pre-processing (including ANN): ANN > PLSR > SMLR > PCR > RT

Future Research

- To compare the 30 pre-processing transformations using ANN, and explore other configurations of ANN, including multi-layer perceptron, radial basis functions, and wavelength selection with genetic algorithm.
- To compare the 30 pre-processing transformations using Boosted Regression Trees (BRT).

Acknowledgements

To Carolyn Olson, Steve Bloom and Sanjay Lamsal. This work was funded by the Cooperative Ecosystem Studies Unit (CESU) – Natural Resources Conservation Service (NRCS).