GLOBAL SYMPOSIUM ON SOIL ORGANIC CARBON, Rome, Italy, 21-23 March 2017

Soil Spectral Libraries for monitoring and reporting on Sustainable Development Goal indicators in Northern Greece

Nikolaos L. Tsakiridis[1], Nikolaos Tziolas[2], George Galanis[3], Eyal Ben-Dor[4], George C. Zalidis[2, 3,]*

[1] Automation and Robotics Laboratory, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki 54214, Greece; Tel.: +302310996377; E-mail: tsakirin@ece.auth.gr
[2] Laboratory of Remote Sensing, Spectroscopy, and GIS, Faculty of Agriculture, Aristotle University of Thessaloniki, Thessaloniki 54214, Greece; Tel.: +302310991779; E-mail: ntziolas@agro.auth.gr, zalidis@agro.auth.gr
[3] Interbalkan Environment Center, 18 Loutron Str., Lagadas, Greece; Tel.:+302394023485; E-mail: info@i-bec.org
[4] The Remote Sensing Laboratory, Department of Geography and Human Environment, Tel-Aviv University, P.O. Box 39040, Ramat Aviv, 69978 Tel-Aviv, Israel; Tel.: +972036407049; E-mail: bendor@post.tau.ac.il

Abstract

In this paper, we present a case-study of utilizing vis-NIR spectroscopy to estimate the content of Soil Organic Carbon (SOC) remotely in Northern Greece. In this agricultural area, SOC plays a pivotal role in the physical, chemical, and biological function of the soils and hence requires rapid and in situ analysis. 474 Entisol soil samples were collected from the two top soil horizons. The wet analytical evaluation of SOC using the Walkley-Black method yielded an average of 0.6599%, with a standard deviation of 0.3908%. The reflectance spectra of these soils were acquired across the vis-NIR region (350-2500 nm) in the laboratory using a standardization protocol. For the chemometric analysis, three pre-processing methods were considered, namely the absorbance transformation, the continuum removal, and the first-derivative. We used two state-of-the-art machine learnings algorithms (Partial Least Squares Regression and Cubist), to estimate the SOC from the spectra. The best results were achieved using the first derivative, by the Cubist algorithm, where an RMSE of 0.1174% was achieved. These results indicate that precise mapping of SOC can be achieved with vis-NIR spectroscopy, facilitating the regular updating of SOC maps for sustainable agriculture, in line with the Sustainable Development Goals 2.4 and 15.3.

Keywords: soil spectral library, soil spectroscopy, soil organic carbon, precision agriculture, vis-NIR spectroscopy, reduced input agriculture

Introduction, scope and main objectives

Soil Organic Carbon (SOC) plays a crucial role in agro-ecosystem function, influencing the productive potential of the soil and water holding capacity. The United Nations Statistical Commission's Interagency and the Expert Group on Sustainable Development Goals (SDGs) Indicators agreed on "Proportion of agricultural area under productive and sustainable agriculture" and "Proportion of land that is degraded over total land area" as indicators to monitor the SDG 2.4 and 15.3 targets, respectively. Reliable, accurate and timely information of SOC stocks at a variety of scales, would help for monitoring and reporting progress towards the implementation of the SDG targets.

However, many countries have significant gaps and great level of fragmentation in infrastructure to monitor and report these indicators. Building capacity in this regard, provides essential knowledge to farmers, environmental policy makers and regional stakeholders, enabling enhanced decision making and effective management of natural resources.

The purpose of this paper is to highlight the use of vis-NIR spectroscopy to accurately estimate the SOC of soil samples in a regional soil spectral library and simultaneously overcome the shortage of soil data in Greek

territory. In particular, only few and sporadic spectra have been recorded in the latest attempt for the development of a global soil spectral library (SSL) [1], as well as in the European LUCAS spectral library [2].

Recently, soil spectroscopy has emerged as a fast, simple, and relatively inexpensive way to estimate the physical and chemical properties of soil samples [3]. Among the most commonly employed applications of soil spectroscopy, is the estimation of SOC (or soil organic matter), which has proven to be accurately estimated.

Several papers have already assessed SOC from spectra, but the models are not robust and it is essential to generate different models for each area. This is mainly because SOC is a complex material composed of varying molecules, strongly dependent on the environmental conditions within the field in question.

The objective of this work is thus to assess the ability of vis-NIR spectroscopy to accurately estimate SOC using a vast soil spectral library generated recently in Greece and demonstrate its potential usage. The derived models can be used to map the SOC of the entire region without the necessity to measure SOC directly in the laboratory and save time and money. To this end, two state of the art machine learning algorithms were used to correlate the input vis-NIR spectra with the observable values of SOC. The predictive accuracy of the derived models was investigated, to identify the best model.

Methodology

Initially, a soil spectral library was developed comprised of 474 Entisol soil samples (~250g) from soil horizons A (0-30 cm) and B (30-60 cm). These soil samples were collected from the agricultural lands surrounding the Nestos river delta, in the Eastern Macedonia and Thrace region, located in northern Greece. The Nestos river delta spans a region of roughly 300 square kilometers. From 235 different sampling points both layers A and B were sampled, while from 4 sampling points only the top layer was sampled.

The collected samples were subsequently divided into two equal parts. The first half was sent to a chemical laboratory, which measured SOC using the Walkley-Black method, and yielded an average of 0.6599%, while the standard deviation was 0.3908%. The distribution is positively skewed (1.03). The second half of the soil sample was air dried, and gently crushed to pass through a <2 mm sieve. It was subsequently placed into a dark chamber, and its reflectance spectrum in the vis-NIR region (350-2500 nm) was collected. The PSR+ spectrometer from Spectral Evolution was used, which covers the 350-2500 nm range using a spectral resolution of 3 nm at 700 nm, 8 nm at 1500 nm, and 6 nm at 2100 nm. It further provides a data output with a 1nm sampling resolution. A standardization procedure was applied to correct from potential nonsystematic and systematic spectral variations [4].

Initially, the 5 first principal components (explaining 99.53% of the variance) of the reflectance spectra were used, in order to calculate the Mahalanobis distance of each spectrum. Using the cumulative chi-squared distribution, and by applying a threshold of 97.5%, 27 outliers were identified and removed from the dataset. Thus, the soil spectral library considered in this study was comprised of a total of 447 soil samples.

The recorded reflectance spectra were then pre-processed using the following independent methods: 1) the (pseudo) absorbance transformation (log(1/reflectance)), 2) the continuum removal of the reflectance spectra, and 3) the first-derivative of the reflectance spectra using a Savitzky-Golay filter of width 7. To these 4 datasets (including the initial reflectance spectra), two algorithms were applied, namely Partial Least Squares Regression (PLSR), and the Cubist algorithm [5] to correlate the input spectra with the output SOC. The R package caret [6] was used to apply these algorithms.

Two sets of experiments were considered. Initially, a 5-fold cross-validation experiment was conducted. Thus, one fold was kept for testing the performance of the model, and the rest 4 folds comprised the training set.

This was repeated 5 times, with each of the fold used once as a testing test. For each training dataset, an internal repeated 10-fold experiment (with 5 repetitions) was used to determine the based parameters of the algorithms (the latent variables for PLSR, and the number of committees and neighbors for Cubist).

In the second set of experiments, the dataset was split into two parts using the Kennard-Stone algorithm [7]; 2/3 of the dataset were used to build the model, and 1/3 to validate it. The distance metric used was the Mahalanobis distance computed over the principal components' space. Once again, to select the optimal set of parameters for both the algorithms, a repeated 10-fold experiment was conducted.

To compare the generated models, the following measures were calculated in the independent test set (y_i is the SOC of the i-th sample, $\hat{y}_i$ is the predicted SOC for the i-th sample, and $\bar{y}$ is the mean SOC of all samples):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

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

Results

The performance of the algorithms using as predictors the different sources is presented in Table 1. We note that for the cross-validation experiment, the average results in the testing (prediction) dataset across the folds are presented. For PLSR, the average number of latent variables (LV) is also given. Furthermore, for the Cubist algorithm the average numbers of committees and neighbours are depicted. The results of the best model for both experiments and both algorithms were derived using as a source the first derivative of the reflectance spectra. The models developed by Cubist tend to be more accurate, with an average RMSE of 0.1718 compared to an average RMSE of 0.2377 for the first experiment, and an average RMSE of 0.2410 compared to an average RMSE of 0.2467 when Kennard-Stone was used. With an RMSE of 0.1174 the model created by the Cubist algorithm using the first derivative spectra as the predictors, exhibit the best predictive accuracy.

5-fold cross-validation

	PLSR			**Cubist**			
	R^2	RMSE	LV	R^2	RMSE	Committees	Neighbours
Reflectance	0.6053	0.2460	10.2	0.7977	0.1761	18.0	0.0
Absorbance	0.6359	0.2363	13.0	0.7714	0.1872	18.0	5.4
Continuum Removed	0.6192	0.2416	13.2	0.7220	0.2064	16.0	0.0
First derivative	0.6638	0.2270	11.4	0.9102	0.1174	16.0	9.0

Splitting with Kennard-Stone

	PLSR			**Cubist**			
	R^2	RMSE	LV	R^2	RMSE	Committees	Neighbours
Reflectance	0.6081	0.2450	10	0.6147	0.2430	20	0
Absorbance	0.5770	0.2537	10	0.6024	0.2469	20	0
Continuum Removed	0.5914	0.2471	12	0.6043	0.2463	20	0
First derivative	0.6216	0.2409	9	0.6613	0.2279	20	9

Table 1: Results of the derived models for both set of experiments

Discussion

The above results underscore the fact that vis-NIR spectroscopy can effectively estimate the SOC content. The significant difference between the models generated by PLSR and Cubist is attributed to the fact that Cubist employs boosted regression trees, i.e. an ensemble of models, each creating local models, while in contrast PLSR is a single global model. Moreover, the 5-fold cross-validation experiment generated better results over the use of the Kennard-Stone algorithm. This is attributed to the following reasons: a) when a 5-fold cross-validation experiment is considered, more percentage of the dataset is used in each fold to build the model (80% compared to 66.6%), thus more variance is covered, and b) the Kennard-Stone algorithm has been shown to not always optimally represent the initial vis-NIR distribution [8].

Conclusions

By utilizing vis-NIR SSLs and deploying state-of-the-art machine learning methods, essential information can be extracted in order to promote the development of an integrated Nexus framework, supporting the strengthening of capacities in the areas of food security monitoring and adaptation to climate change.

Forthcoming relevant activities and outcomes envisaged will be in the direction of extending, improving and strengthening the vision for a global SSL. Especially, in less developed countries, where monitoring systems for SOC, spanning from the absolute absence of monitoring capacities to the execution of timely and high cost field campaigns.

A further step that could be conducted, to ascertain that the model is working correctly, is to compare the spectral assignments of the model with the spectral regions that SOC influences the most, as reported in the literature. In this context, the encoded information could be the basis for new developments based on Micro Electro Mechanical Systems technology, in order to revolutionize the agricultural sector by providing more cost-efficient and targeted tools.

As future work, using the best model created, it would be possible precisely map the SOC of the region using only the vis-NIR spectra of the soil samples. A common spectral library enabling data interoperability and comparability, might help to regularly update digital soil mapping products with better resolution.

Acknowledgments

This research work was funded by the project "AGRO_LESS: Joint reference strategies for rural activities of reduced inputs", of the European Territorial Cooperation Programme Greece-Bulgaria 2007-2013.

References

[1] R. A. Viscarra Rossel, T. Behrens, E. Ben-Dor, D. J. Brown, J. A. M. Demattê, K. D. Shepherd, Z. Shi, B. Stenberg, A. Stevens, V. Adamchuk, H. Aïchi, B. G. Barthès, H. M. Bartholomeus, A. D. Bayer, M. Bernoux, K. Böttcher, L. Brodský, C. W. Du, A. Chappell, Y. Fouad, V. Genot, C. Gomez, S. Grunwald, A. Gubler, C. Guerrero, C. B. Hedley, M. Knadel, H. J. M. Morrás, M. Nocita, L. Ramirez-Lopez, P. Roudier, E. M. R. Campos, P. Sanborn, V. M. Sellitto, K. A. Sudduth, B. G. Rawlins, C. Walter, L. A. Winowiecki, S. Y. Hong, and W. Ji, "A global spectral library to characterize the world's soil," *Earth-Science Rev.*, vol. 155, no. February, pp. 198–230, Apr. 2016.

[2] G. Tóth, A. Jones, and L. Montanarella, "The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union," *Environ. Monit. Assess.*, vol. 185, no. 9, pp. 7409–7425, Sep. 2013.

[3] M. Nocita, A. Stevens, B. van Wesemael, M. Aitkenhead, M. Bachmann, B. Barthès, E. Ben Dor, D. J. Brown, M. Clairotte, A. Csorba, P. Dardenne, J. A. M. Demattê, V. Genot, C. Guerrero, M.

Knadel, L. Montanarella, C. Noon, L. Ramirez-Lopez, J. Robertson, H. Sakai, J. M. Soriano-Disla, K. D. Shepherd, B. Stenberg, E. K. Towett, R. Vargas, and J. Wetterlind, "Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring," no. September, 2015, pp. 139–159.

[4] V. Kopačková and E. Ben-Dor, "Normalizing reflectance from different spectrometers and protocols with an internal soil standard," *Int. J. Remote Sens.*, vol. 37, no. 6, pp. 1276–1290, 2016.

[5] J. R. Quinlan, "Combining Instance-Based and Model-Based Learning," *Mach. Learn.*, vol. 76, pp. 236–243, 1993.

[6] M. Kuhn, "Building Predictive Models in R Using the caret Package," *J. Stat. Softw.*, vol. 28, no. 5, pp. 1–26, 2008.

[7] R. W. Kennard and L. A. Stone, "Computer Aided Design of Experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[8] L. Ramirez-Lopez, K. Schmidt, T. Behrens, B. van Wesemael, J. A. M. Demattê, and T. Scholten, "Sampling optimal calibration sets in soil infrared spectroscopy," *Geoderma*, vol. 226–227, no. 1, pp. 140–150, Aug. 2014.