Predicting soil properties for sustainable agriculture using vis-NIR spectroscopy - a case study in northern Greece

Nikolaos L. Tsakiridis[a,*], Nikolaos Tziolas[b], Agathoklis Dimitrakos[c], Georgios Galanis[c], Eleftheria Ntonou[b], Anastasia Tsirika[b], Evangelia Terzopoulou[c], Eleni Kalopesa[c] and George C. Zalidis[b, c]

[a] 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

[b] 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, ntonou@agro.auth.gr, atsirika@agro.auth.gr, zalidis@agro.auth.gr

[c] Interbalkan Environment Center, 18 Loutron Str., Lagadas, Greece; Tel.:+30 23940 23485; E-mail: info@i-bec.org

ABSTRACT

Soil Spectral Libraries facilitate agricultural production taking into account the principles of a low-input sustainable agriculture and provide more valuable knowledge to environmental policy makers, enabling improved decision making and effective management of natural resources in the region. In this paper, a comparison in the predictive performance of two state of the art algorithms, one linear (Partial Least Squares Regression) and one non-linear Cubist), employed in soil spectroscopy is conducted. The comparison was carried out in a regional Soil Spectral Library developed in the Eastern Macedonia and Thrace region of Northern Greece, comprised of roughly 400 Entisol soil samples from soil horizons A (0-30 cm) and B (30-60 cm). The soil spectra were acquired in the visible – Near Infrared Red region (vis-NIR, 350nm-2500nm) using a standard protocol in the laboratory. Three soil properties, which are essential for agriculture, were analyzed and taken into account for the comparison, namely Organic Matter, Clay content and the concentration of nitrate-N. Additionally, three different spectral pre-processing techniques were utilized, namely the continuum removal, the absorbance transformation, and the first derivative. Following the removal of outliers using the Mahalanobis distance in the first 5 principal components of the spectra (accounting for ~99.8% of the variance), a five-fold cross-validation experiment was considered for all 12 datasets. Statistical comparisons were conducted on the results, which indicate that the Cubist algorithm outperforms PLSR, while the most informative transformation is the first derivative.
Keywords: Soil Spectroscopy, Soil Spectral Library, Soil Modeling, Nestos River

1. INTRODUCTION

The 2030 Agenda for Sustainable Development has highlighted the importance of food security and the promotion of sustainable agriculture, through the specific indicators 2.4.1 (“Proportion of agricultural area under productive and sustainable agriculture”) and 15.3.1 (“Proportion of land that is degraded over total land area”) of the Sustainable Development Goals (SDGs) . Soil is the foundation of agriculture and maintaining healthy and sustainable soils is a pre-requisite to maintain and achieve sustainable agriculture. Due to the unprecedented pressures on soils in recent years from degradation and over-exploitation which threaten the agro-ecosystem, food security is endangered. If no action is taken to maintain and enhance the agricultural lands, which have their nutrients depleted due to the repetitive harvesting of crops, and proper soil management practices are not followed, it is impossible to achieve sustainable agriculture. In order to mitigate the aforementioned predicaments, understanding and monitoring of the problem is of key essence. Therefore,

detailed soil maps should be produced using a dedicated and repeated process in order to accurately depict the pressures in agricultural soils. These maps will enable farmers, industry and governmental agencies to explicitly identify the key areas endangered, and allow them to take targeted actions towards soil restoration and conservation.

Soil Spectral Libraries (SSLs) contain meticulously recorded data and metadata of soil samples, attempting to capture the variability of the soils in an area. The data recorded concerns the physical and chemical soil properties of each sample measured in a chemical laboratory, the reflectance spectrum in the visible to near-infrared region (vis-NIR, 350-2500 nm), and additional metadata such as the location of each sample, the soil class etc. The importance of highly precise SSLs has been highlighted in recent years[1–3] due to their ability to produce accurate, detailed, within short time and cost-effective thematic soil maps of the area. This is achieved by developing models using partial least squares regression[4] and machine learning algorithms[5], which correlate each soil's spectrum with its properties. It is thus possible to apply these models to a sample's spectral signature and derive its soil properties using only the information contained in the signature, and with no a priori knowledge of any soil property.

Although several other papers have estimated essential for agriculture soil properties[6,7], the models are not robust and it is important to generate different models for each area. The reason is that soils are vastly complex, and most of the properties investigated are complex materials composed of varying molecules. These properties are additionally strongly dependent on the underlying conditions within each respective field.

The objective of this paper is hence to assess the ability of vis-NIR to accurately estimate the following soil properties: Soil Organic Matter (SOM), Clay content, and concentration of nitrate-N (N-NO3) using a recently generated regional SSL in Greece. The aforementioned properties are among the most important soil properties that define the ability of soil to support productivity, and are thus considered vital for agricultural activity.

To this end, a number of machine learning models were developed and the most accurate ones were identified. The models developed in this work can be utilized in the future to map these properties and assist in the provision of knowledge based recommendations for sustainable agricultural management strategies as well as an informed and transparent framework to meet policy regulations.

The rest of the paper is organized as follows. Section 2 presents the methodology applied to create the regional SSL, as well as the machine learning methods used to establish the models. Section 3 describes in details the results, and presents the models' accuracies. The conclusions of this work are drawn in Section 4, where furthermore suggestions regarding the application of this work are made.

2. MATERIALS AND METHODS

2.1 Populating the regional spectral library

The regional soil spectral library was developed from the agricultural lands surrounding the Nestos river delta, in the Eastern Macedonia and Thrace region, in Northern Greece, which is a part of the largest and most diverse currently SSL available in Greece. This area is one of the more important agricultural areas in Greece. Actually 96% of the whole Eastern Macedonia and Thrace region is covered by agricultural land. The importance of agricultural production is highlighted by the fact that it produces 8.7% of the National Gross Domestic Product, which is 3 times higher than the Greek average. This is due to the adequate availability of water resources and soils' fertility. But many of those agricultural areas are within the Natura 2000 network, so protection and conservation of resources is of high importance. Intense agricultural practices though have resulted in soil degradation at some parts, so it is necessary to develop systems of recording soil indicators and monitoring the soil quality.

The sampling area spans at a region of roughly 400 square kilometers, and is composed mainly of Entisol soils. A random stratified sampling procedure was employed, to select 474 Entisol soil samples (~250g) from soil horizons A (0-30 cm) and B (30-60 cm). From 235 different sampling points both layers A and B were sampled, while from 4 sampling points only the top layer was sampled. The sampling campaign took place during the summer of 2015, within the frame of the AgroLess Project. The geographical location of the sampling points are given in Figure 1.

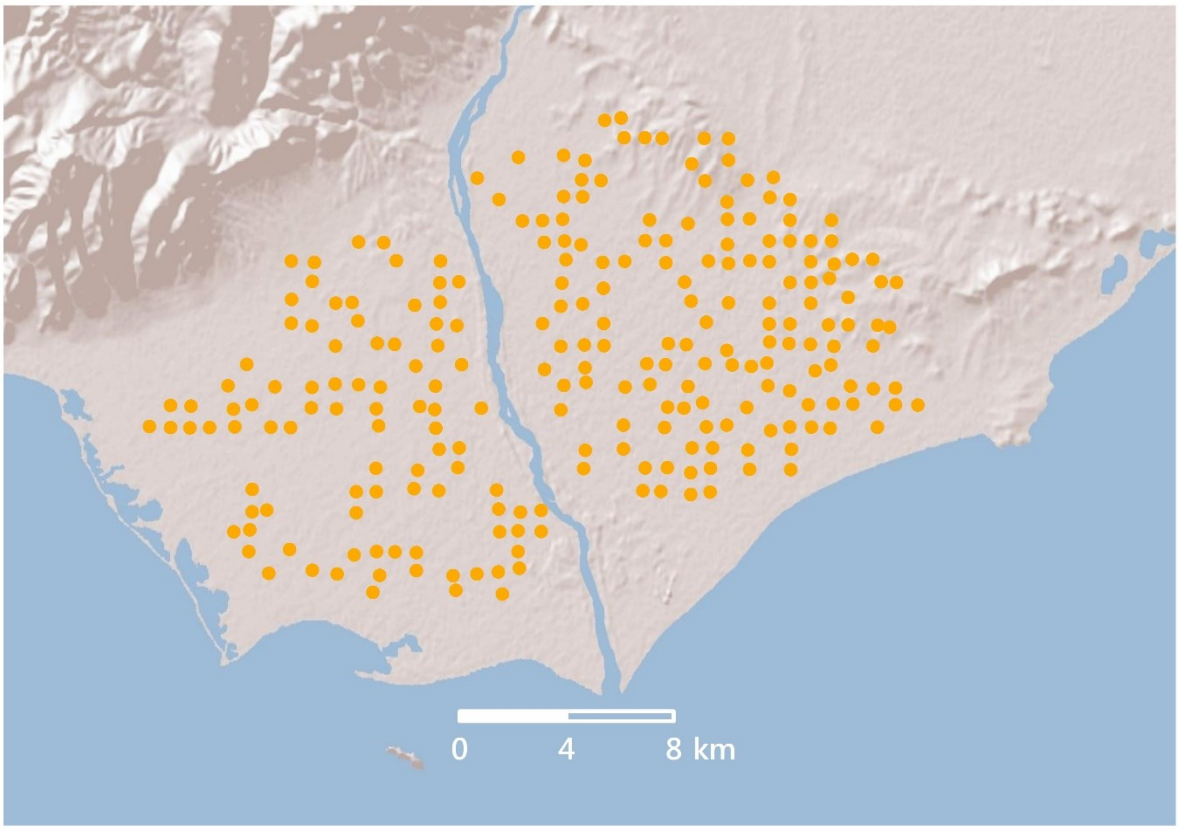

Figure 1: The sampling points in the Nestos river delta

The collected soil samples were subsequently divided into two equal parts. The first half was sent to a chemical laboratory, which measured SOM using the Walkley-Black method, the Clay content using the Bouyoucos hydrometer method, while the Kjeldah method was employed for the measurement of N-NO3. The most important statistical moments of the chemical results are presented in Table 1, while the respective box plots are illustrated in Figure 2. All soil properties are positively skewed, with the concentration of the nitrate-N exhibiting the largest positive skewness. Their Pearson correlation coefficients are given in Table 2; SOM and Clay are the most correlated properties.

Table 1: Results of the chemical analysis of the soil samples. Q25, Q50 and Q75 refer to the 25th, 50th and 75th quartile respectively, while SD denotes the standard deviation.

Property	Min	Q25	Q50	Mean	Q75	Max	SD	Skewness
SOM [%]	0.00	0.60	1.10	1.14	1.50	4.18	0.67	1.03
Clay [%]	0.00	9.00	13.00	15.03	19.00	75.00	9.77	2.01
N-NO3 [ppm]	0.00	3.60	8.50	25.96	26.48	661.20	50.49	6.07

Table 2: Cross-correlations between the measured soil properties

	SOM	Clay
Clay	0.4101	-
NO3-N	0.2573	0.0668

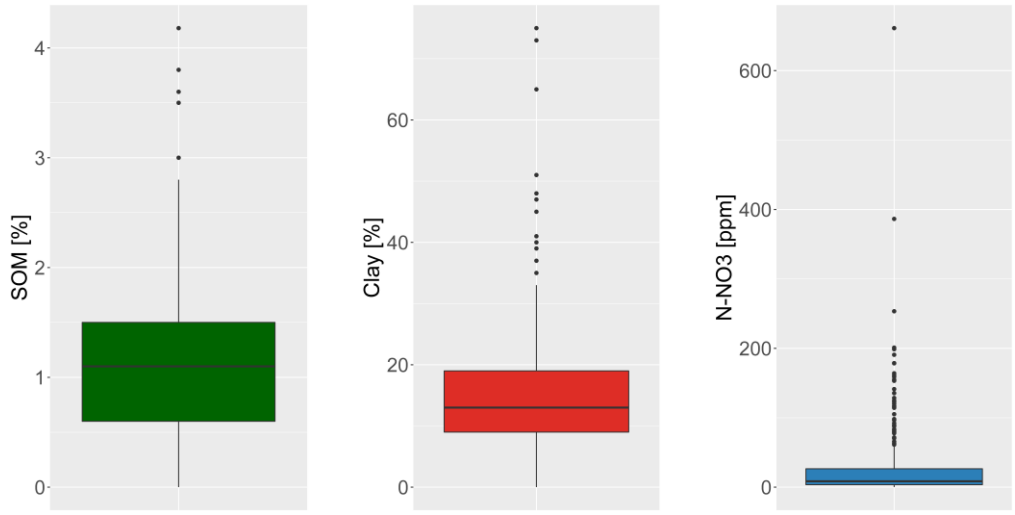

Figure 2: Box-plots of the concentrations of the measured soil properties

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 to collect the spectral signatures, covering 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[8]. Additionally, considering that PSR+ uses internally 3 arrays (a 512 element Si array, and two 256 element extended InGaAs arrays), the step-like artefacts inserted at the two splices of the spectrometer's sensors were corrected using the visible portion of each spectrum as the base spectrum, in order to create continuous spectra.

2.2 Pre-processing of the spectral data

Initially, the first and last 50nms corresponding to the fringes of the spectra were removed, due to the noise they exhibited. Thus, the spectral range considered henceforth is 400-2450nm. After this step, the following process was used to identify and remove potential outliers: First, the 5 first principal components (explaining 99.76% of the variance) of the recorded reflectance spectra were used, in order to calculate the Mahalanobis distance d_i of each spectrum. Assuming the data are normally distributed, d_i^2 is approximately chi-square distributed with 5 degrees of freedom. By selecting a critical value[9], 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 (log10(1/reflectance))
2) the continuum removal (CR) of the reflectance spectra, and
3) the first-derivative of the reflectance spectra using a Savitzky-Golay filter of width 7.

The generated datasets are presented in Figure 3.

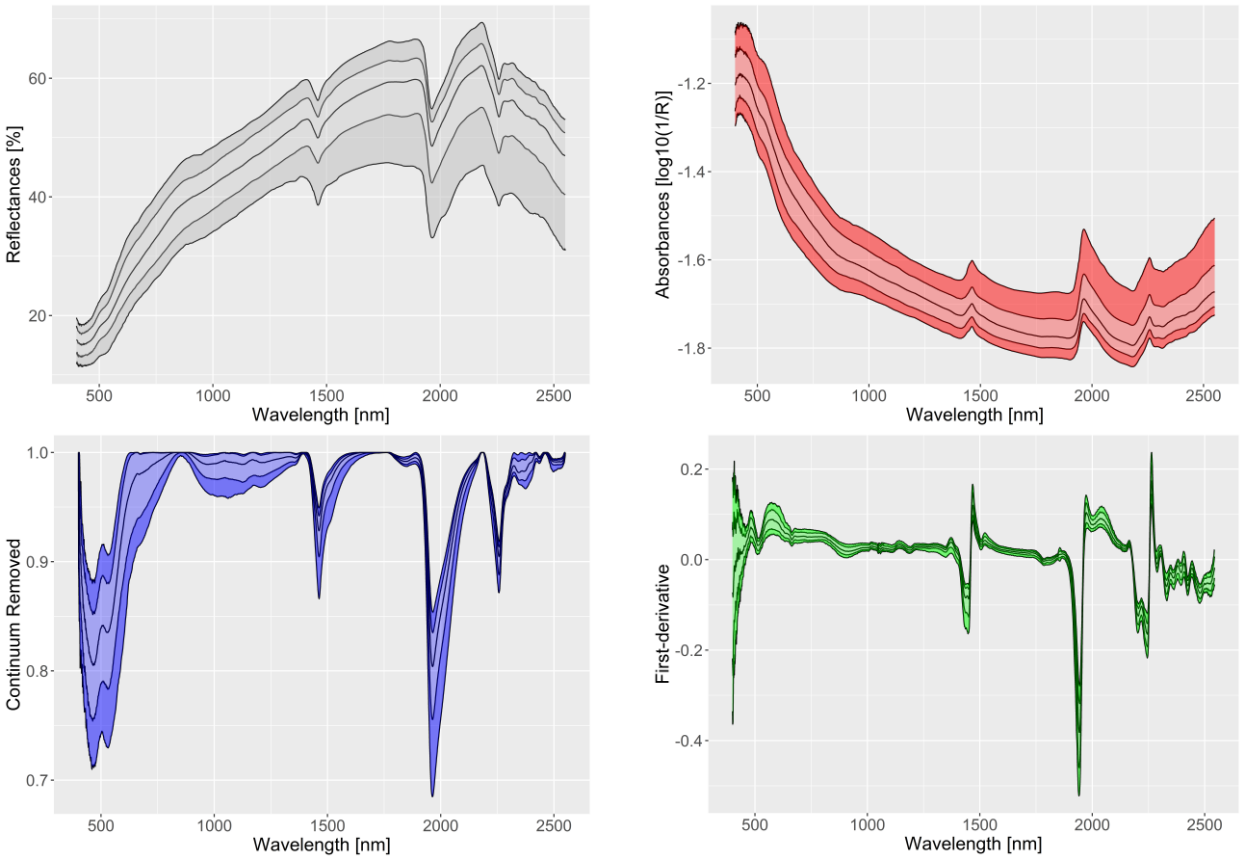

Figure 3: The initial reflectance spectra, and the 3 different pre-processing techniques used. Depicted are the 5th, 16th, 50th, 84th and 95th percentiles of the spectra.

2.3 Building of the models and estimating their performance

In total, 12 datasets were considered (3 chemical attributes times 4 different spectral sources). For each dataset, we applied two state of-the-art algorithms, namely the Partial Least Squares Regression (PLSR) algorithm and the Cubist algorithm using the caret package in R[10]. The PLSR algorithm[11] constructs a few number of orthogonal factors, termed latent variables, which are linear combinations of the initial predictor variables. They are created in a way to explain as much as possible the covariance between the input and the output. The Cubist algorithm[12] creates a rule-based model, by constructing tress with its terminal leaves containing linear regression models. It further can use a boosting-like scheme where iterative models (i.e. rule bases) named committees are constructed to enhance the accuracy and robustness of the derived model[13].

To create the models, a double 5-fold cross-validation method was used. One fold was left out as an independent set, and the four rest were used as calibration; this was repeated for all folds. Internally within the calibration set, a 5-fold cross-validation experiment was used to estimate the parameters of each algorithm. For PLSR, the parameter estimated was the number of latent variables, while for the Cubist algorithm the parameters estimated were the number of committees and neighbors.

To assess and evaluate the performance of the models, the following measures were used in the independent test set:

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

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

where y_i is the soil property of the i-th sample, $\hat{y}_i$ is the predicted property for the i-th sample, and $\bar{y}$ is the mean property of all samples.

3. EXPERIMENTAL RESULTS

The experimental results for the prediction set are presented in Tables 3, 4, and 5 for the SOM, the Clay Content and N-NO3 respectively. It should be noted that each number presented in the aforementioned tables refers to the average of the 5 folds. Both accuracy as well as structural parameters are given for both algorithms, in order to assess both the predictive performance as well as the complexity of the underlying models. The structural parameters given are the latent variables (LV) for PLSR, and the number of committees and neighbors for the Cubist algorithm.

Table 3: Results for Soil Organic Matter

	PLSR			Cubist			
	R^2	RMSE	LV	R^2	RMSE	Committees	Neighbors
Reflectance	0.6675	0.3893	14.8	0.7270	0.3527	20.0	0.0
Absorbance	0.6215	0.4156	12.8	0.7338	0.3485	20.0	0.0
Continuum Removed	0.6179	0.4152	13.0	0.8724	0.2399	18.0	1.8
First derivative	0.6874	0.3775	7.4	0.9142	0.1978	16.0	9.0

Table 4: Results for Clay Content

	PLSR			Cubist			
	R^2	RMSE	LV	R^2	RMSE	Committees	Neighbors
Reflectance	0.6643	4.6200	12.4	0.7535	3.9591	20.0	0.0
Absorbance	0.6671	4.4366	10.6	0.7315	3.9844	18.0	0.0
Continuum Removed	0.6620	4.7866	10.8	0.8725	2.9398	18.0	0.0
First derivative	0.6739	4.6890	5.0	0.9070	2.5035	14.2	7.4

Table 5: Results for N-NO3

	PLSR			Cubist			
	R^2	RMSE	LV	R^2	RMSE	Committees	Neighbors
Reflectance	0.2471	45.0361	9.0	0.4173	39.6216	16.0	5.4
Absorbance	0.2394	45.2658	9.0	0.7446	26.2442	16.0	4.6
Continuum Removed	0.2528	44.8653	6.0	0.8087	22.6367	16.0	0.0
First derivative	0.2960	43.5490	3.4	0.7718	24.7546	14.0	9.0

The results indicate that Cubist outperforms PLSR in terms of accuracy, with an average R^2 of 0.7712, compared to an average R^2 of 0.5247. PLSR fails to accurately estimate the concentration of N-NO3, whereas the Cubist algorithm attains a significant R^2, with a maximum value of 0.8087 when the continuum removed spectra are considered. By applying the Wilcoxon signed rank test[14] between the achieved R^2 values of both algorithms, we can test whether the difference of accuracy is statistically significant. The p-value of the test is 4.8828e-04, which rejects the null hypothesis, i.e. that the algorithms attain similar accuracies. Therefore we can conclude that Cubist statistically outperforms PLSR.

Additionally, the first derivative transformation scored the highest average R^2 (0.7084) and was ranked first among the different pre-processing methods (Table 6). Moreover, the CR transformation was ranked second, closely following the

first derivative transformation. These results underscore the fact that spectral pre-treatment can have a large impact on the derived chemometric models, by enhancing the spectral information.

Table 6: Ranking and average performance of each pre-processing method when both algorithms are considered

	Reflectance	Absorbance	Continuum Removal	First derivative
Ranking	4	3	2	1
Average R^2	0.5794	0.6230	0.6810	0.7084

To identify the most important wavelengths for each soil property, we used the best model as identified from Tables 3, 4, and 5. The variable importance for the Cubist models is a linear combination of the usage of each feature both in the rule conditions, as well as in the model. The identified features are presented in Figure 4. These features are similar to the ones reported as important wavelengths of spectral absorptions[15,16]. More concretely, important regions for SOM are around the following wavelengths (in nm): 1100, 1600, 2000 and 2200-2400, while the visible range has been shown to improve the accuracy results. For Clay minerals, the 2200-2400 region is the most important.

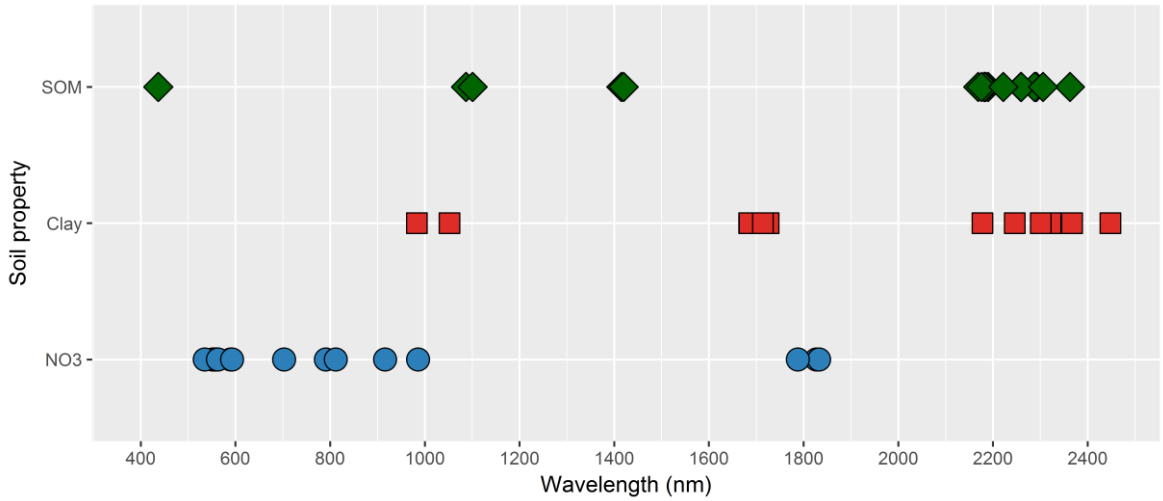

Figure 4: The 20 most important features (wavelengths) of each soil property, as identified by the best model

4. CONCLUSION

Evaluating soil indicators by using soil spectroscopy could provide essential tools for land management according to the low input agriculture principles. Low input agriculture requires detailed and dense spatial temporal data of soil indicators in order to develop sustainable management plans to help minimize soil fertility loss and eliminate irrigation overdose. Measuring those soil indicators with laboratory methods can be quite costly and time consuming, enabling the risk that the results would be available later than the management plan should have been defined. By using soil spectroscopy results can be obtained on time while minimizing the cost of the operation, which could prove to be vital in order to increase the competitiveness of agricultural production. In this context harmonized data are provided with sufficient accuracy for several applications in the field of smart agriculture. Utilizing spectra pretreatment techniques prior to the development of the algorithms facilitated the removal of unwanted background effects and noise. In order to derive the best chemometric models from soil spectra, different pre-processing techniques, as well as different machine learning algorithms, should be tested.

This research has focused on the implementation of state of the art methodologies and algorithms to develop an easily updatable soil spectral library in order to fully contribute to the implementation of SDG indicators, and assist the provision of novel in-situ observation methods which enables sustainable farm management.

Future work should focus on extending this Soil Spectral Library with samples from the Balkans and the MENA region.

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] Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., Shi, Z., Stenberg, B., Stevens, A., et al., "A global spectral library to characterize the world's soil," Earth-Science Reviews 155(February), 198–230 (2016).

[2] Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D. J., Clairotte, M., et al., "Chapter Four – Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring," Advances in Agronomy 132, 139–159 (2015).

[3] Tóth, G., Jones, A., Montanarella, L., "The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union," Environmental Monitoring and Assessment 185(9), 7409–7425 (2013).

[4] Mouazen, A. M., Kuang, B., De Baerdemaeker, J., Ramon, H., "Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy," Geoderma 158(1-2), 23–31 (2010).

[5] Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., "Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy," PLoS One 8, H. Y. Chen, Ed., e66409 (2013).

[6] Ben-Dor, E., Chabrillat, S., Demattê, J. A. M., Taylor, G. R., Hill, J., Whiting, M. L., Sommer, S., "Using Imaging Spectroscopy to study soil properties," Remote Sensing of Environment 113, 38–55 (2009).

[7] Rossel, R. A. V., Webster, R., "Predicting soil properties from the Australian soil visible-near infrared spectroscopic database," European Journal of Soil Science 63(6), 848–860 (2012).

[8] Kopacková, V., Ben-Dor, E., "Normalizing reflectance from different spectrometers and protocols with an internal soil standard," International Journal of Remote Sensing 37(6), 1276–1290 (2016).

[9] Filzmoser, P., Garrett, R. G., Reimann, C., "Multivariate outlier detection in exploration geochemistry," Computers & Geosciences 31(5), 579–587 (2005).

[10] Kuhn, M., "Building Predictive Models in R Using the caret Package," Journal of Statistical Software 28(5), 1–26 (2008).

[11] Wold, S., Sjöström, M., Eriksson, L., "PLS-regression: A basic tool of chemometrics," Chemometrics and Intelligent Laboratory Systems 58(2), 109–130 (2001).

[12] Quinlan, J. R., "Combining Instance-Based and Model-Based Learning," Machine Learning 76, 236–243 (1993).

[13] Kuhn, M., Johnson, K., [Applied Predictive Modeling], Springer, New York, (2013).

[14] Wilcoxon, F., "Individual Comparisons by Ranking Methods," Biometrics Bulletin 1(6), 80 (1945).

[15] Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. a., Macdonald, L. M., McLaughlin, M. J., "The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties," Applied Spectroscopy Reviews 49(2), 139–186 (2014).

[16] Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., Wetterlind, J., "Visible and Near Infrared Spectroscopy in Soil Science," Advances in Agronomy 107, 163-215 (2010).