2023-09

DRIFT spectroscopy and permutation importance algorithm in quantitative analysis of organic matter in soil model systems

Jović Branislav, Panić Marko

DRIFT spectroscopy and permutation importance algorithm in quantitative analysis of organic matter in soil model systems

Branislav D. Jović[1*], Marko N. Panić[2]

[1] University of Novi Sad, Faculty of Sciences, Department of Chemistry Biochemistry and Environmental protection, Novi Sad, Serbia; e-mail: branislav.jovic@dh.uns.ac.rs
[2] University of Novi Sad, Biosense Institute, Novi Sad, Serbia; e-mail: panic@biosense.com

* *Corresponding author*

DOI: 10.46793/ICCBI23.104J

Abstract In order to obtain useful MIR spectrochemical data of soil organic matter for the development of remote sensing methods, synthetically prepared soils with artificial precisely defined organic matter fractions, DRIFT spectroscopy and permutation importance algorithm were used in this paper. In terms of imitation of soil organic matter, sample model systems were prepared with precisely defined added values of added organic components. After MSC and SNV spectral treatments using PCA and LDA techniques and DRIFT spectra, the soil was classified according to the percentage of organic matter. Using the KDE+permutation importance algorithm, three significant MIR spectral regions were obtained for percentage grouping: 600-1000cm^{-1}(skeletal vibrations of organic matter); 1750-2250cm^{-1} (Total reflectance+quartz overtones) and 3250-3950cm-^{1} (Hydroxyl groups). In terms of the potential for quantitative analysis, the calculated wavelength ranges match well with the classical spectrochemical theoretical basis of analytical methodologies. Also, extracted useful spectrochemical data can be potentially used in the development of new remote-satellite detection methods (ASTER satellite in SWIR and MIR range).

Keywords: FTIR spectroscopy, Soil organic matter, PCA, Feature permutation algorithm

1. Introduction

In recent times, the need for analysis of soil organic matter for many samples at a field scale has emerged, especially for the purpose of precision agriculture implementation. Soil organic matter has great implications on all physical and plenty of chemical soil properties and thus characterizes fertility and other potentials of this valuable natural resource. Standard laboratory analysis of soil organic matter is complex, expensive and time-consuming compared to possible indirect measurement by reflectance spectroscopy. Diffuse Reflectance Infrared Fourier Transform *Spectroscopy* (*DRIFTS*) is a method widely used in different fields of soil studies. Examination of structural differences of humic acids [1], and vertical distribution of coal [2] are just some of the many beneficial uses of the DRIFT method. In this research, the aim was to explore the possibilities of using DRIFT spectroscopy in the determination of organic

matter. In this sense, model systems were created by adding a precisely defined amount of organic matter to the soil samples. About six hundred DRIFT spectra were recorded and subjected to multivariate analysis methods (LDA, Feature permutation algorithm) and useful information was extracted for the development of further remote sensing methods. This paper is a part of our continuing and systematic investigation of the application of spectroscopic-chemometric methods in soil analysis [3,4].

2. Experimental

To imitate soil organic matter, a mixture of humic acid, starch and nicotinamide was used. These model systems were prepared due to the presence of characteristic functional groups of organic molecules (aliphatic, aromatic, Hydroxyl, nitrogenous...). To examine and influence the soil type and obtain a robust model mixture of the most common soil types from northern Serbia was used. A total of 5 grams of soil with a percentage of organic matter of 0.5 to 4% was prepared for each sample of the system model. Infrared spectra were obtained using the Thermo-Nicolet Nexus iS20 instrument on the diffuse reflectance module. The spectral range was 4000-400 cm^{-1}, a total of 32 scans per spectrum were recorded at a resolution of 4 cm^{-1}. Savitzky-Golay digital filter, multiplicative scatter correction (MSC) and standard normal variate (SNV) techniques are utilized for the removal of (physical) variability among the data samples due to scatter caused by variations in sample positioning, irregularities in its surface and differences in particle size.

3. Results and discussion

3.1. Classification of samples based on the concentration of organic matter

Linear dimensionality reduction technique LDA achieve moderately good separation in latent space among four groups with different concentration of organic matter denoted as G05, G1, G2 and G4. MSC preprocessing yields significantly better classification model performance than SNV treatment. ROC stands for receiver operating characteristic presented with a curve, defined by true positive rate (TPR) against the false positive rate (FPR), and serves as a visual diagnostic tool for the evaluation of the binary classifier.

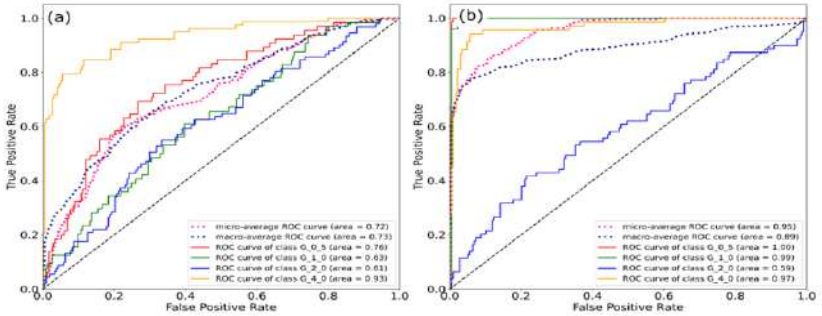

Figure 1. ROC curves of binary models for classification of samples into four groups preprocessed with (a) SNV and (b) MSC correction techniques.

One-vs-all strategy for fitting the model for the classification of fractions is used during the training procedure and the LDA with one component is selected as the model to be fitted using 50% of data within the dataset. These results in trained for binary classifiers (models) whose ROC curves are estimated from predictions on the test set, are presented in Figure 1.

3.2. Analysis of feature importance for classification

The permutation importance algorithm is applied to the created dataset using trained binary models to evaluate which wavelengths (features) or part of the spectrum has influenced the most model performances.

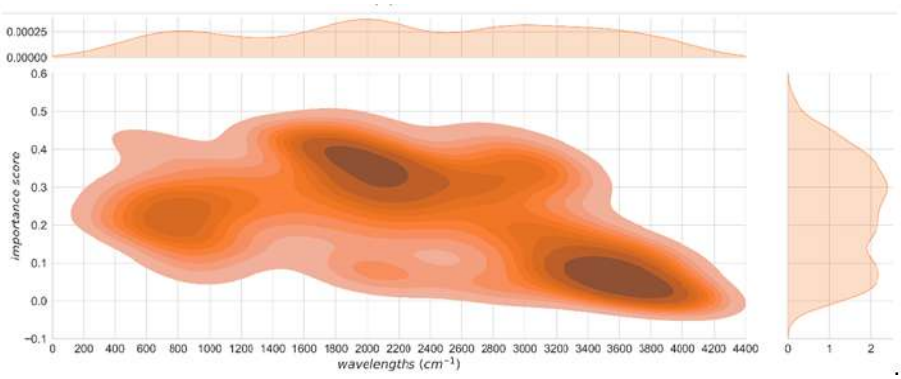

Figure 2 Estimated bivariate probability density functions (pdfs) of an importance score and wavelength from a list of scores and selected wavelengths obtained from a permutation importance algorithm runs.

Starting with the condensed distance matrix, obtained from Spearman rank-order correlations a Ward's linkage matrix is calculated, and a hierarchical clustering is performed on the predefined train dataset for creation of models for the classification of percent of organic matter. Visualizing the dendrogram of obtained clusters the following range of values are selected for the threshold of [0.1, 5.0] for the percent dedicated train dataset used for building corresponding models for classification. Importance scores show how much trained binary model performance decreased in accuracy with a random shuffling of features. Presenting the importance scores for the representative wavelengths together with their frequency of occurrences in the form of a pdf, give more insight into important part of the spectrum for classification than that only importance scores are used. Models for classification of samples according to percent of organic matter divide the wavelength-score region into three regions: (600 - 1000,0.18 - 0.26), (1750 - 2250, 0.30 - 0.41) and (3250 - 3950, 0.01 - 0.11). In the research of Demate et al, [5] after high correlations of extracted spectral MIR regions and ASTER simulated spectral bands were determined. Namely, in this research the following potential spectral regions for future satellite sensors were suggested: 2760 – 2500 cm^{-1}, 2150–1875cm^{-1} and 840 – 740 cm^{-1}. In our research, using a completely different approach in laboratory conditions, very similar data were obtained. We believe that the obtained wavelength

ranges lead to the convergence of useful data necessary for the calibration of existing and the development of new sensors and remote sensing methods.

4. Conclusions

Obtained wavelength ranges indicate the importance of the aliphatic (CHx) and aromatic (C=C) vibration MIR region as well as the soil color (total reflectance related to the mineral content, quartz) for the potential quantification of soil organic matter. In terms of the potential for quantitative analysis, the calculated wavelength ranges match well with the classical spectrochemical theoretical basis of analytical methodologies. Obtained MIR wavelength ranges can be potentially used in the development of satellite detection methods (ASTER satellite).

Acknowledgment

This work was supported by the Provincial Secretariat for Science and Technological Development, Autonomous Province of Vojvodina, project no. 142–451-2198/2022–01.

References

[1] Francioso O., Montecchio D., Gioacchini P., Cavani L., Ciavatta C., Trubetskoj O., Trubetskaya O., 2009. *Structural differences of Chernozem soil humic acids SEC–PAGE fractions revealed by thermal (TG–DTA) and spectroscopic (DRIFT) analyses*, Geoderma, 152; 264–268

[2] Hobley E., Willgoose G.R., Frisia S., Jacobsen G., 2014. *Vertical distribution of charcoal in a sandy soil: evidence from DRIFT spectra and field emission scanning electron microscopy*, European Journal of Soil Science, 65; 751–762

[3] Jović, B., Ćirić, V., Kovačević, M., Šeremešić, S., et al. *Empirical equation for preliminary assessment of soil texture*. Spectrochim. Acta A Mol. Biomol. Spectrosc. 2019. 206: 506–511.

[4] Jović, B., Maletić, S., Kordić, B., Beljin, J. *DRIFT spectroscopic determination of clay and organic matter in sediment by mixed soil-sediment calibration approach*, Environmental Monitoring and Assessment. Volume 195, 3 2023

[5] N. E. Q. Silvero, L. A. Di Loreto Di Raimo, G. Silva Pereira, L. P. de Magalhães, F. da Silva Terra, M. A. Ananias Dassan, D. F. Urbina Salazar, José A.M. Demattê, *Effects of water, organic matter, and iron forms in mid-IR spectra of soils: Assessments from laboratory to satellite-simulated data* Geoderma. 2020. 375:114480