FRI-10.326-1-EEEA-11

SELECTION OF INFORMATIVE WAVELENGTHS IN HYPERSPECTRAL IMAGE ANALYSIS TO DETERMINE THE CONTENT OF NITROGEN COMPOUNDS IN SOIL SAMPLES

Georgi Manchev
Department of Automatic and Electronics,
"Angel Kanchev" Univesity of Ruse
E-mail: gmanchev@uni-ruse.bg

Assoc. Prof. Stanislav Penchev, PhD
Department of Automatic and Electronics,
"Angel Kanchev" Univesity of Ruse
E-mail: msp@uni-ruse.bg

Principal Asst. Prof. Eleonora Nedelcheva, PhD
Department of Automatic and Electronics,
"Angel Kanchev" Univesity of Ruse
E-mail: ekirilova@uni-ruse.bg

Assoc. Prof. Tsvetelina Georgieva, PhD
Department of Automatic and Electronics,
"Angel Kanchev" Univesity of Ruse
E-mail: cgeorgieva@uni-ruse.bg

Prof. Plamen Daskalov, PhD
Department of Automatic and Electronics,
"Angel Kanchev" Univesity of Ruse
E-mail: daskalov@uni-ruse.bg

Abstract: *The article reviews existing popular solutions for wavelength selection and shows the possibilities of using Machine Learning models as a quick feedback and assessment in the analysis of spectral data for determining the nitrogen content in synthetic soil samples. The goal was to verify the possibilities for quantitative assessment of nitrogen content in samples captured with a hyperspectral camera in the near infrared region, using modern open-source solutions and tools. Data processing and analysis were performed using Orange software, with special attention paid to the built-in tools and methods for determining informative wavelengths. As a result of the experiments, the effectiveness of the applied solutions for determining nitrogen in soil samples and the potential for future development by upgrading the functionality of the used software package were confirmed.*

Keywords: *Hyperspectral Imaging, Near-Infrared, Machine Learning, Wavelength selection, Orange.*

INTRODUCTION

Quantitative assessment of macroelements in soil using optical methods is a process that combines scientific achievements in a number of fields, such as chemistry, physics, mathematics, optoelectronics, etc. Regarding the near infrared region (NIR), the analysis and interpretation of spectral characteristics is a complex and laborious process, especially when it comes to complex mixtures such as soil, in which different types of minerals, clay, plant residues, bacteria, etc. are present (Ma, Y.et al., 2023). Nitrogen in the soil is bound in different types of chemical compounds. Some of them, under the influence of temperature, moisture and microbiological processes, are transformed into other types of nitrogen-containing compounds, creating a rich palette of spectral responses for each of them. The absorption spectra in the NIR region of some

of the compounds overlap, which further creates serious difficulties in directly linking the observed results with spectral responses of chemical compounds at fixed wavelengths determined by analytical chemistry. The presence of moisture in some cases leads to a spectral shift (Chang, C. et al., 2005), and temperature to a sharpening of the spectral peaks (Wullfert, F. et al., 1998). The observed absorption in the NIR is not at the fundamental vibrational frequencies of the molecular bonds, but is a manifestation of frequency overtones associated with proportionally reduced energy relative to their order, which is characterized by weaker spectral responses. The high spectral resolution of HSI further complicates the analysis, since each band is considered as a separate factor. All these features require the use of specialized software products.

This paper presents an approach for selecting informative wavelengths for quantitative assessment of nitrogen content in samples captured with a hyperspectral camera in the near infrared region. The possibility of using ML models as a fast feedback and assessment in spectral data analysis is verified. Experimental studies have been conducted to confirm the effectiveness of the proposed solutions for quantitative assessment of nitrogen content in soil samples.

EXPOSITION

Software instruments

There are a number of commercial and also free solutions, and for the purposes of this study, Orange3 was chosen – an open-source software developed in Python by the University of Ljubljana. It is equipped with built-in modules for spectral data processing, statistical analysis, a set of ML models, as well as evaluation tools.

Spectral characteristics of soil samples

The study used averaged spectral characteristics from hyperspectral images (HSI) of synthetic soil samples, captured with a hyperspectral camera Specim N17E with 256 spectral bands in the range 850nm … 1700nm. The soil samples were created by introducing an aqueous solution of urea (CH_4N_2O), for which the exact mass ratio of the introduced nitrogen to the mass of the soil sample was calculated. The study included samples with concentrations: 0 – 5000mg N/kg soil and 0 – 25000 mg N/kg soil in the presence and absence of moisture.

Spectral data preprocessing

The data from all soil samples are normalized in the interval 0÷1 in order to be effectively compared and analysed (1). They are then merged into a common data stream using the Concatenate widget, and the general appearance of their spectral characteristics is shown in Fig. 1. It shows the presence of differences in the baseline of the individual samples, pronounced absorption in the water band around 1400nm, as well as the absence of clearly pronounced areas of variation.

$$I_{norm} = \frac{I_r - I_d}{I_w - I_d} \quad (1)$$

I_{norm} – normalized intensity; I_w – intensity of white target with >99% reflectance; I_d – intensity of dark (fully covered lens).

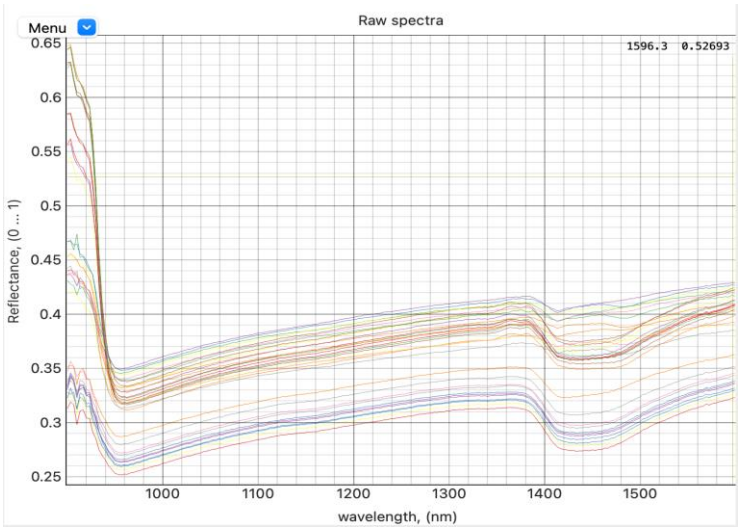

Fig. 1. Normalized spectral characteristics of soil samples

It is known that data transformation and the use of derivatives emphasizes variations in spectral regions and removes the baseline effect. The calculation of first derivatives was done by applying Savitzky-Golay filtering with parameters (5,2,1), the result of which is shown in Fig. 2.

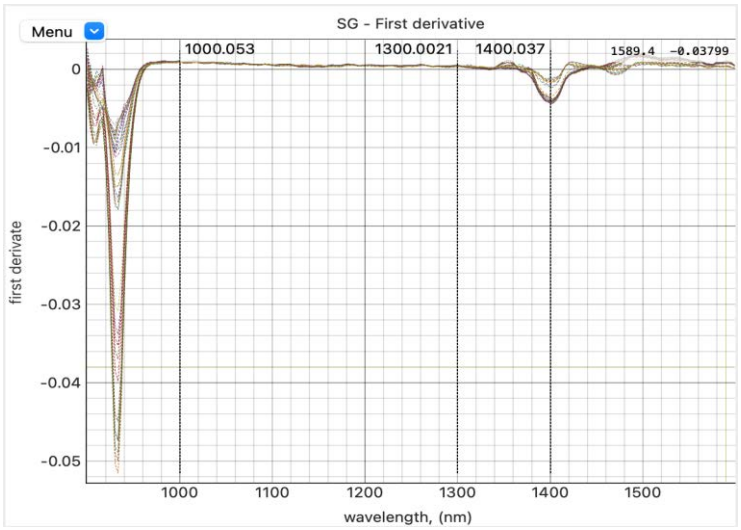

Fig. 2. First derivative of spectral characteristics

Multivariate analysis

Since each band of the spectral data is a factor, selecting and limiting them to a list of those with the most informative character is a key stage of the analysis and model building. There are many algorithms for selecting informative wavelengths, some of which are included as tools in Orange. The "traditional" selection methods based on PCA and PLSR have been widely used over the years (Cordella, C. 2012, Dardenne, P. et al., 2000). There are also others, with proven effectiveness, such as genetic algorithm (GA), competitive adaptive reweighted sampling (CARS), forward recursive feature selection (FRFS), recursive feature elimination (RFE), etc. (Zhang, X. et al., 2023).

The software package used includes tools based mainly on PCA and PLSR. Due to the relatively high nitrogen concentration in the samples, working with them shows satisfactory results. In this study, a combination of PCA and Rank modules was built, shown in Fig 3, in which the final selection was made based on two evaluation methods: univariate regression and RReliefF. As a result, the 20 most significant bands were selected, shown in Table 1.

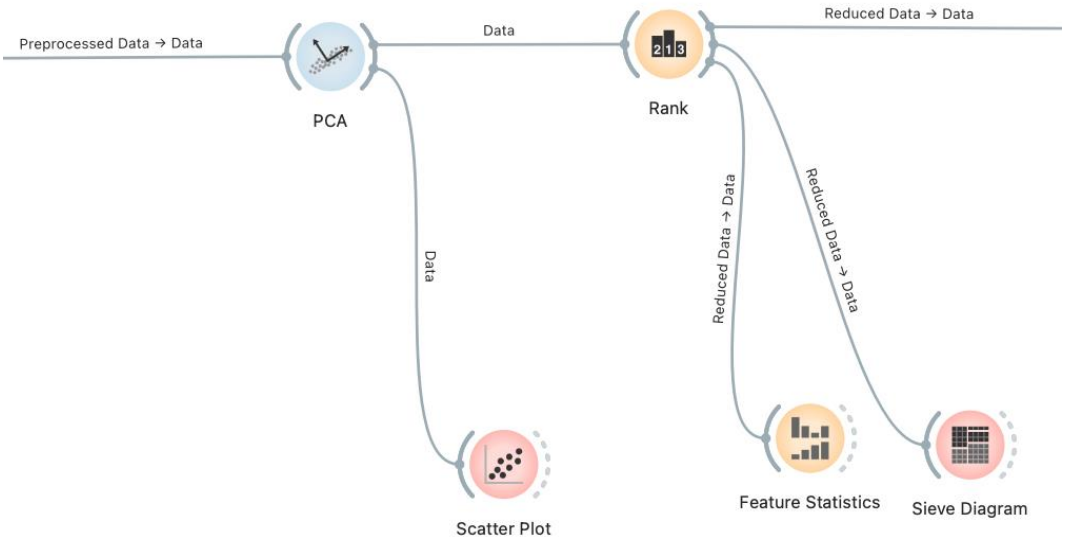

Fig. 3. A combination of PCA and Rank widgets for wavelength selection.

Table 1. Most informative wavelengths selected by Rank widget

Wavelength (nm)	Univariate Regression	RReliefeF
913.51	33.982	0.396
916.82	46.669	0.420
966.42	29.100	0.275
969.73	29.933	0.255
1198.01	30.070	0.227
1307.24	35.619	0.291
1310.55	28.468	0.270
1313.86	34.906	0.272
1317.17	28.47	0.306
1320.48	23.020	0.388
1360.21	25.200	0.384
1363.52	45.708	0.412
1452.93	89.018	0.349
1456.25	183.748	0.388
1459.56	117.476	0.419
1462.87	55.179	0.395
1466.18	52.789	0.423
1469.49	39.011	0.417
1472.81	31.912	0.414
1476.12	25.033	0.404

ML models as a tool for quick feedback of wavelength selection

For feedback on the selection result, a comparison of results from a set of ML models was applied: PLS, Stochastic Gradient Descent, Tree, Neural Network (Multi-layer perceptron), kNN, AdaBoost, Random Forest, Gradient Boosting, Linear Regression. The constructed structure and the results are shown in Fig. 4 and Table 2.

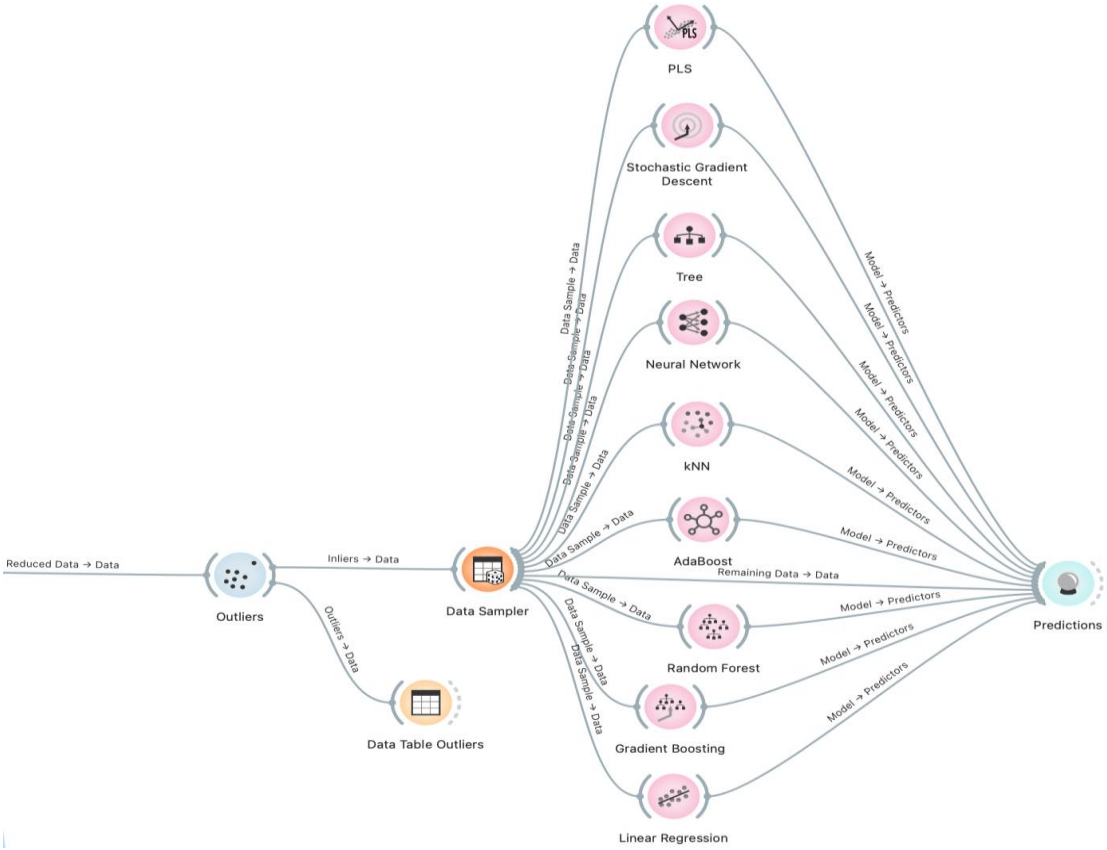

Fig. 4. A structure for wavelength selection evaluation based on ML models.

Machine-learning models strive to achieve optimal prediction metrics but may struggle to identify the correct fundamental wavebands relevant to soil studies (Ma, Y.et al., 2023) but in combinations with algorithms for wavelength selection such in this approach allows quick assessment of the effectiveness, as well as indicative values for the potential predictive ability of the models. Due to the relatively small number of samples in the study, the assessment was performed only on the basis of the data from the coefficient of determination (R^2) for each of the ML models. Some authors (Sarathjith, C. et al., 2016, Reda, R. et al., 2019) point out that despite R^2, it is more correct to base the estimate on root mean squared error (RMSE) and residual prediction deviation (RPD), which would be applicable when increasing the representative sample studied. To ensure the adequacy of the models, outliers were preliminary excluded from the data stream, using the Outliers widget.

It can be seen from the results that almost all models have potential to predict the nitrogen content in soil samples (except Linear Regression). The best results are achieved when using Neural Network, PLS, Stochastic Gradient Descent and Tree, with values of the coefficient of determination above 0.9.

Table 2. Results of ML models for quantitative evaluation of nitrogen content

ML model	R^2
AdaBoost	0.851
Gradient Boosting	0.844
kNN	0.751
Linear Regression	0.479
Neural Network	0.952
PLS	0.950
Random Forest	0.782
Stochastic Gradient Descent	0.963
Tree	0.930

CONCLUSION

The presented approach is able to detect the presence of nitrogen in synthetic soil samples, narrows the spectral range to areas correlating with the amount of nitrogen, and shows promising results for achieving qualitative quantitative assessment. The precise manual adjustment of the selected regions leads to increased accuracy. The combination of wavelength selection methods and ML models allows for express monitoring of intermediate results and quick manual corrections for more precise analysis, significantly reducing data processing time due to the reduction of the number of input factors. The built-in software modules in Orange3 allow for efficient wavelength selection, but mainly cover traditional methods based on PCA and PLS. On the other hand, it has an open interface, allowing for expansion of functionality by integrating modern algorithms, which opens new opportunities for future development.

ACKNOWLEDGMENTS

The report reflects the results of the work on project No. 24-FEEA-05, financed by the "Scientific Research" fund of the University of Ruse, Bulgaria.

REFERENCES

Chang, C., Laird, D. & Hurburgh, C. (2005). Influence of soil moisture on near-infrared reflectance spectroscopic measurement of soil properties. *Soil Science*, 170(4), 244-255

Cordella, C. (2012) PCA: The Basic Building Block of Chemometrics. *Analytical Chemistry*. InTech. Available at: http://dx.doi.org/10.5772/51429.

Dardenne, P., Sinnaeve, G. & Baeten, V. (2000). Multivariate calibration and chemometrics for near infrared spectroscopy: which method?, *Journal of Near Infrared Spectroscopy,* 8, 229-237.

Ma, Y., Minasny, B., Demattê, J. & McBratney, A. (2023). Incorporating soil knowledge into machine-learning prediction of soil properties from soil spectra. *European Journal of Soil Science*, 74(6), e13438.

Reda, R., Saffaj, T., Ilham, B., Saidi, O., Issam, K., Brahim, L. & El Hadrami, M. (2019). A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 195, 103873.

Sarathjith, C., Das, S., Wani, P. & Sahrawat, L., (2016). Variable indicators for optimum wavelength selection in diffuse reflectance spectroscopy of soils. *Geoderma*, 267, 1–9.

Wullfert, F., Kok, W. & Smilde, A. (1998). Influence of Temperature on Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models. *Analytical Chemistry*, 70(9), 1761-1767.

Zhang, X., Xue, J., Xiao, Y., Shi, Z. & Chen, S. (2023). Towards Optimal Variable Selection Methods for Soil Property Prediction Using a Regional Soil Vis-NIR Spectral Library. *Remote Sensing*, 15(2), 465.

