Приложение на NIR спектроскопия, комбинирана с хемометрични методи на моделиране за екологичен анализ на замърсени почви

С. Касабов, Е. Кирилова

Application of NIR Spectroscopy Combined with Chemometric Modelling Methods for Environmental Analysis of Contamined Soils

S. Kasabov, E. Kirilova

Institute of Chemical Engineering, Bulgarian Academy of Sciences, Acad. G. Bontchev St., Bl. 103, 1113 Sofia, Bulgaria, e.kirilova@iche.bas.bg

Key Words: *NIR spectroscopy; chemometric methods; environmental analysis; contaminated soils.*

Abstract. *Conducting quantitative analysis of soil properties and contaminants in them is a crucial for an achievement of a good understanding of dynamics of ecosystems and a sustainable soil management. This study presents the recent progress on the developed combined approaches including NIR spectroscopy and chemometrics for accurate, low-cost and reliable quantitative analysis of soil content, published in prestigious journals for the last 7 years. A classification of the approaches depending on the types of soils components, the type of applied pre-processing and chemometric methods is done. Their main advantages and disadvantages are shown and some trends for future development are outlined.*

1. Introduction

A major challenge for humanity nowadays is preserving Earth's ecosystems and minimizing the damage that has been done to the environment following the rapid and unrestricted industrialization and urbanization. We have seen examples of oil spillages, toxic waste from various factories contaminating water around the world, which leads to accumulation of these contaminants in soil. Society's awareness on the topic has emerged progressively in recent years because health and environmental risks are at stake. Petroleum hydrocarbons and heavy metals (mercury, lead, cadmium, etc.) are the most common contaminants which contribute to 60% of soil contamination. Thus, soil monitoring is essential in order to assess soil quality and protect biodiversity. Conventional analysis consists of sampling, sample pretreatment, analyte separation and analyte detection – all these processes are time consuming, heavily dependent on qualified analysts and on top of that are associated with usage of organic solvents, which, if not handled properly, can contribute to pollution of the environment. By contrast, near infrared (NIR) reflectance spectroscopy offers rapid, non-destructive and ecofriendly analysis involving diffuse reflectance measurement in the near-infrared region. The spectrum is obtained as a result of vibrational energy transitions of molecular bonds with high dipole moment and the spectral data can be used for quantitative analysis of multicomponent complex samples.

However, the interpretation of the data is not immediate and requires the use of chemometrics, which can be broadly defined as "the use of mathematical and statistical methods to analyze chemical data". The quantitative analysis of the obtained data includes implementation of clustering techniques and regression methods. The aim of the present work is to review the available literature on the combined methods developed in the last 7 years, including NIR spectroscopy and chemometric methods applied for environmental analysis of contaminated soils. Classification of the methods according to the type of the analyzed soil properties and contaminants in them, the used methods for pre-treatment of the samples, the applied chemometric methods and used measures for estimation of the performance of the applied method is done. In addition, the main advantages and disadvantages are shown and some trends for future development of this type of methods are also shown.

2. Soil Properties and Contaminants

The unreasonable utilization of natural resources (including soil) by humans caused numerous ecological and environmental issues. Human and anthropogenic activities, such as excessive use of fertilizers, pesticides, sewage irrigation, and discharge of waste affected soil environment significantly [12].

In order to achieve sustainable soil management, it is necessary to have a deep knowledge of the mechanisms of chemical and biological processes that take place in them, as well as their properties. For example, the content of soil organic carbon (SOC) [1-3], total nitrogen (TN) [2,3], available phosphorus (P) [2], available potassium (K) [2] and moisture content (MC) [3] have a considerable influence on soil quality and plant growth.

Another important parameter which have a crucial role for improvement of the fertilization and softening of the soil is content of organic matter in the soil (SOM) [7]. The latter is composed of living plant, animal and microbial biomass, dead roots and other plant residues in various stages of rot and soil humus. It promotes the processes of decomposition of nutrients and the growth and development of plants. Clay and sand also are important parameters of soil, because they are directly bound with water availability and nutrient adsorption [5]. Clay and sand interact with the other physical, chemical and biological properties of the soil. Other important parameters to define the soil quality are the cation exchange capacity (CEC) and sum of exchange bases (SB) [5]. CEC is a parameter which represents the capability of soil to attract, retain and hold exchangeable cations, such as K^+, Na^+, Ca^{2+}, Mg^{2+}, Al^{3+} and H^+.

Also, various rare earth elements (REEs) such as La, Ce, and Nd [6] can be found in the soils which have become socially important due to their various industrial and technological applications (e.g., superconductors, supermagnets, catalysts, automotive electrification, medicine, fertilizers).

The presence of contaminants in soils also affects the process of soil management. One of the most common pollutants are oil spillages. They can be caused by people's mistakes, natural disasters while transporting, equipment breakage, etc. Petroleum hydrocarbons (PHCs) [7] are main constituent of those leakages and consist of mixture of short and long-chain hydrocarbon compounds such as phenanthrene [8], alkanes [9] and polycyclic aromatic hydrocarbons (PAHs) [8,9], which once discharged into to the environment undergo degradation processes such as photolysis, oxidation and hydrolysis. These processes lead to products with increased hydrophobicity, thus make them more water-soluble and readily available for accumulation in soils. The latter are harmful to the environment because some are potentially carcinogenic or mutagenic [8].

On the other hand, occurrence of heavy metals in environmental samples is result of negligence and elements such as Cr, Mn, Ni, Cu, Zn, As, Cd[10], Pb [11], Co, Fe [12] and Hg [11,12,13], are extremely dangerous due to their high toxicity.

3. Description of NIR Spectroscopy

Laboratory analysis has been the main key to better understand the soil system and to assess its quality and functions. However, the traditional laboratory methods for chemical analysis of soil samples are slow, time consuming, expensive and require significant labor and reagents and generate a lot of residue. Compared to the conventional analytical methods, the attractiveness of the NIR spectroscopy analysis techniques is that measurements are rapid, non-destructive, cost-effective and estimates of soil properties are inexpensive compared to conventional soil analyses. In NIR technique, the diffuse reflectance spectrum of a sample is obtained in the range of the NIR region (780–2500 nm), as a result from vibrational energy transitions of molecular bonds with the high dipole moment. The overtones and combinational bands of the stretching vibration of carbon-hydrogen (C-H), oxygen-hydrogen (O-H) and nitrogen-hydrogen (N-H) groups are the main molecular vibrations absorbing in the NIR region which makes the diffuse reflectance spectrum useful to predict concentrations of chemicals containing these bonds.

4. Spectral Pre-Processing Methods

In order to improve the accuracy of the calibration models with NIR spectral data and to reduce the random noise in them, the raw spectral data are often pre-processed before modeling. Pre-processing is usually regarded as an integral part of chemometrics modeling with spectral data. The prevailing pre-processing methods for NIR spectroscopy are two main groups: scatter-correction methods and spectral derivatives. The scatter-corrective group includes Multiplicative Scatter Correction (MSC), de-trending, Standard Normal Variate (SNV) and normalization. Among methods based on derivatives two techniques are frequently used: Norris-Williams (NW) derivatives and Savitzky-Golay (SG) polynomial derivative filters. Both are normally applied to minimize baseline deviations. Multiplicative scatter correction (MSC) and standard normal variate (SNV) are typical examples of scatter-corrective methods that can remove additive or multiplicative signal effects and reduce the particle size effect and curvilinear trend of the spectrum [11]. Spectral derivatives, including first derivative (FD), second derivative (SD) and Savitzky-Golay (SG) polynomial derivative filters, have the capability to eliminate both additive and multiplicative effects from reflectance spectra. In particular, Savitzky–Golay smoothing filter has the capability to remove noise from the spectra and to decrease the detrimental effect on the signal-to-noise ratio that conventional finite-difference derivatives would have [9].

Sometimes preprocessing methods such as (log1/R) is applied for reduction of the nonlinearities that probably exist in the spectra. The transformed data are then pre-treated so that they became mean centred, with a standard deviation equal to 1 (autoscaling) [3]. Following this, soil spectra are subjected to the Savitzky and Golay filter for smoothing, using a first derivative transformation. The first derivative transformation enhances small spectral absorptions and eliminates the background effect. Scatter removal from the transformed data was followed by implementing the standard normal variate technique (SNV), which centres each spectrum on its mean and then scales it by its standard deviation in order to remove the path length variations [3]. The outliers are omitted from further modelling steps.

5. Chemometric Calibration Methods

However, the interpretation of the data obtained by NIR spectroscopy is not immediate, and thus requires the use of multivariate analysis to acquire prediction models for each measured variable, that is possible because of reference data based on the traditional destructive methods. The use of chemometric methods is essential to obtain relevant information from NIR spectra, and the combination of both methodologies is necessary for the development of calibration models [5]. Chemometric methods most often used in the analysis of multivariate data can be divided into two categories: clustering and regression. Clustering tries to identify clusters of samples or variables with similar characteristics, which enables an assessment of the general structure of the dataset. When the spectral information is directly used, clustering techniques are applied prior to any other chemometric technique, such as regression analysis. Cluster Analysis (CA), Principal Component Analysis (PCA), Multivariate Curve Resolution (MCR) and Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) are the most popular two-dimensional techniques. For example, PCA has been performed before regression to detect spectral outliers or to reduce the amount of spectral data to deal with by decreasing its dimensionality. Clustering techniques are also useful to define correctly the calibration and validation sets before regression analysis or to compare the results obtained by different spectroscopic techniques. Regression techniques are being used in multivariate calibration to establish the relation between two matrices by means of a mathematical model. Each matrix contains different information about the same group of samples. Most often, the first matrix contains information obtained by means of a reference method, and the second matrix contains information obtained by an alternative method. Calibration involves regression, validation and prediction which means that the data should be divided into two subsets – calibration and validation set. The calibration data set is used to construct a mathematical model by regression analysis and this model is then validated through the validation set in order to check its quality. Later on, in prediction the variables corresponding to unknown samples (the prediction set, completely independent from the calibration and validation sets) are predicted using the previously validated mathematical models. The most popular multivariate regression techniques include Multilinear Regression (MLR), Principal Component Regression (PCR) and Partial Least Squares (PLS). There are some other techniques derived from PLS or MLR, such as, Genetic Algorithms Partial Least Squares (GA-PLS), Moving Window Partial Least Squares (MW-PLS), Uninformative Variable Elimination Partial Least Squares (UVEPLS), Random forest regression (RF), Successive Projections Algorithm Multiple Linear Regression (SPA-MLR) and Equidistant combination Multiple Linear Regression (EC-MLR). All these techniques are based on PLS, or MLR, but they all include a previous step which investigates all the possible combinations among spectral wavelenghts to design the best calibration model. Partial least-squares regression (PLSR) is the most common multivariate analysis method, as it is capable to model several response variables simultaneously while effectively addressing strongly collinear and noisy predictor variables [4,8,11,12]. It is particularly useful when we need to predict a set of dependent variables from a very large set of independent variables.

Chemometric methods successfully model the linear relationship between spectral data and chemical components, especially when multi-dimension and multi-collinearity exist in raw spectra data. However, nonlinearity between the spectra data and chemical components often exists due to instrument variations (lamp aging and sensor sensitivity) and heterogeneous soil characteristics. Thus, nonlinear calibration methods, such as support vector machine regression (SVMR), Cubist model, Random forest (RF) can provide a more reasonable solution than linear methods.

Least squares support vector machines (LS-SVM) [3,13] is robust approach for the classification and regression analysis of linear and nonlinear multivariate problems, using linear equations set and not quadratic programming as in the classical SVM [1].

The Cubist model is a data mining technique which is based on the construction of an unconventional type of regression tree, where the prediction is based on the intermediate linear models at each step [3]. It creates subsets of sample of the original data set that have similar attributes and creates multi-linear regression rules by selecting the optimum predictor variables to be used as regression variables among all of the spectral variables. If the tested sample falls into the restrictions of the first subset, it performs the regression rule that was chosen for that subset, or else it moves to the next rule. The main advantages of the Cubist regression method is its ability to handle non-linear relationships between dependent and independent variables and the ability to use both discrete and continuous variables as inputs.

Random forest (RF) is an ensemble method based on decision of classification or regressions trees [5,7,9]. The latter is very suitable in a case of increasing the variability and the number of samples. Decision trees are methods that use trees to solve classification and regression problems based on rules to binary split data. For classification can be used the Gini index value to split the data, while for the regression problems the trees are trained by minimizing the sum of squared deviations about the mean. Random forest regression can incorporate complex, linear or nonlinear relationships and interactions. Moreover, it has the possibility of use of a data set with missing values.

6. Prediction Performance of Calibration Methods

Different types of measures are applied to evaluate the effectiveness of prediction of the applied chemometric models. The most commonly used are: root mean square error of prediction (RMSEP); coefficient of determination (R^2); residual prediction deviation (RPD), which is a ratio of standard deviation (SD) to RMSEP. In general, a good model prediction would

have high values of R^2 and RPD, and small value of RMSEP. In some cases, in addition to these errors, others are used, such as: ratio of performance to interquartile range (RPIQ), which is expressed as the difference between the third and first per root mean square error.

7. Advantages/Disadvantages and Trends for Future Development

The table presents information on the combined approaches, published between 2013 and 2020, that include regression analysis of datasets with spectral data obtained from the analysis of soil samples using NIR spectroscopic techniques.

Research works on combined approaches found in the available literature for period 2013-2020

Pre-processing method	Chemometric method	Soil properties, contaminants	Model performance	Reference
Log(1/R), smoothing Log(1/R), SG, FD, smoothing FD, SD, smoothing SD, SNV, mean center (MC), MSC	SVM, SPA, PLSR	SO	R^2, RMSE, RPD	[1]
MC, FD, MSC, SG, SNV	PCA, PLSR	OC, TN, P, K	R^2, RPD, RMSEC, RMSEP, RMSEE	[2]
Log(1/R), SG	PCA, PLSR, LS-SVM, Cubist	TN, OC, MC	RMSEC, RMSEP, RMSE, R^2, RPD	[3]
R, SG, SNV, FD	PLSR	SO	Pc, Pc^2, RMSEC, RMSEP, SEC, Slope, Offset, RPD,	[4]
MC, auto scale, SNV, MSC, FD, SD,	RF, PLSR	CEC, EB, OM, clay and sand	RMSEC, RMSEP	[5]
MSC, SG, SNV	PLSR, iPLSR, iSPA-PLSR	La, Nd, Ce	$RMSE_{CV}$, R^2_{CV}, $bias_{CV}$, $RMSE_{pred}$, R^2_{pred}, $bias_{pred}$	[6]
Noise cut, maximum normalization, FD and smoothing	PLSR, RF	TPHs	R^2, RPD, RMSEP	[7]
FD, SG	PLSR	PAHs	R^2, RPD, RMSE	[8]
FD and smoothing, second-order polynomial approximation, SG	PLSR, RF	PAHs and alkanes	R^2, RPD, RMSEP	[9]
SG	PLSR, Genetic algorithm	Cd	R^2, RPD	[10]
log (1/R), SG smoothing, FD, SD, second-order polynomial in conjunction with SNV, MSC, NOR, and log (1/R)	PLSR	Cr, Mn, Ni, Cu, Zn, As, Cd, Hg, and Pb	R^2, RMSE, SE, RPD	[11]
Smoothing MC, SNV, MSC, Normalization, Detrending, Derivatives	PLSR	Cr, Mn, Ni, Cu, Zn, As, Cd, Hg, Pb and Co, Fe	R^2	[12]
SD, SNV	LS-SVM	Hg in plant leaves	R^2	[13]

As it can be observed in *the table*, PLS is the regression technique preferably used. The latter successfully model the linear relationship between spectral data and chemical components, especially when multi-dimension and multi-collinearity exist in raw spectra data. Results for quantification of heavy metals show excellent performance of the PLS method for Hg, Pb, Cr, Ni [11,12]. As for soil properties, excellent results were obtained with PLS model for OC, TN, available P [2]. On the other hand, the machine learning methods such as LS-SVMs and the Cubist method are capable of tackling non-linear problems in the dataset. The latter out-performed the linear multivariate methods for the prediction of the soil properties providing more reliable results [3]. The random forest regression models have better performance than PLS regression models for CEC, OM, clay and sand, demonstrating resistance to overfitting, attenuating the effect of outlier samples and indicating the most important variables for the model [5]. There is a strong indication that NIR spectroscopy signal acquisition followed by RF algorithm can be trusted for real application in hydrocarbon analysis in petroleum-contaminated sites where limited data are available [8]. The application of spectral pre-processing methods increases the prediction capacity of the calibration models. However, the number of soil samples as well as their concentrations affected the prediction capacity of the calibration models based on NIR spectra. The smaller the number of samples and the lower their concentrations, the worse the models perform [5,9]. The calibration models work best within the ranges between the minimum and maximum values of concentrations of the soil properties [2,3]. The inclusion of other soil samples with different concentra-

tions may increase the concentration range of the models and significantly reduce its ability to predict well [3].

In general, there is a tendency to combine several pre-processing methods and chemometric methods, the latter usually being linear with nonlinear to obtain better results.

8. Conclusions

This review has shown an analysis of the latest combined NIR spectroscopy and chemometrics approaches found in the available literature for the prediction of soil properties and the presence of contaminants in them. They were classified according to the type of components studied in them, the used pre-processing and chemometric methods. Their main advantages and disadvantages were pointed out and some tendencies for future development were outlined.

References

1. Peng, X., T. Shi, A. Song, Y. Chen, W. Gao. Estimating Soil Organic Carbon Using VIS/NIR Spectroscopy with SVMR and SPA Methods, Remote Sens. 6, 2014, 2699-2717.

2. Carra, J. B., M. Fabris, J. Dieckow, O. R. Brito, P. R. S. Vendrame and L. M. D. Santos Tonial. Near-infrared Spectroscopy Coupled with Chemometrics Tools: a Rapid and Non-destructive Alternative on Soil Evaluation. – *Communications in Soil Science and Plant Analysis*, 50, 2019, 4, 421-434.

3. Morellos, A., X. E. Pantazi, D. Moshou, T. Alexandridis, R. Whetton, G. Tziotzios, J. Wiebensohn, R. Bill, A. M. Mouazen. Machine Learning Based Prediction of Soil Total Nitrogen, Organic Carbon and Moisture Content by Using VIS-NIR Spectroscopy. – *Biosystems Engineering: Special Issue: Proximal Soil Sensing*, 152, 2016, 104-116.

4. Ba, Y., J. Liu, J. Han, X. Zhang. Application of Vis-NIR Spectroscopy for Determination the Content of Organic Matter in Saline-alkali Soils. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 229, 2020, 117863.

5. De Santana, F. B., A. M. de Souza, R. J. Poppi. Visible and Near Infrared Spectroscopy Coupled to Random Forest to Quantify Some Soil Quality Parameters. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 191, 2018, 454–462.

6. Maia, A. J., Y. J. A. B. da Silva, C. W. A. do Nascimento, G. Veras, M. E. O. Escobar, C. S. M. Cunha, Y. J. A. B. da Silva, R. C. Nascimento, L. H. de Souza Pereira. Near-infrared Spectroscopy for the Prediction of Rare Earth Elements in Soils from the Largest Uranium-phosphate Deposit in Brazil Using PLS, iPLS, and iSPA-PLS Models. Environmental Monitoring and Assessment, 192, 2020, Article Number: 675.

7. Douglas, R. K., S. Nawar, M. C. Alamar, A. M. Mouazen, F. Coulon. Rapid Prediction of Total Petroleum Hydrocarbons Concentration in Contaminated Soil Using vis-NIR Spectroscopy and Regression Techniques. Science of the Total Environment, 616-617, 2018, 147–155.

8. Okparanma, R. N., A. M. Mouazen. Visible and Near-Infrared Spectroscopy Analysis of a Polycyclic Aromatic Hydrocarbon in Soils. – *The Scientific World Journal (Hindawi)*, 2013, Article ID 160360, 9, http://dx.doi.org/10.1155/2013/160360.

9. Douglas, R. K., S. Nawar, M. C. Alamar, F. Coulon, A. M. Mouazen. Rapid Detection of Alkanes and Polycyclic Aromatic Hydrocarbons in Oil-contaminated Soil with Visible Near-infrared Spectroscopy. - *European Journal of Soil Science*, 70, 2019, 1, 140-150.

10. Liu, J., J. Han, J. Xie, H. Wang, Y. Ba. Assessing Heavy Metal Concentrations in Earth-cumulic-orthicanthrosols Soils Using Vis-NIR Spectroscopy Transform Coupled with Chemometrics. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 226, 2020, 117639.

11. Omran, E. S. E. Inference Model to Predict Heavy Metals of Bahr El Baqar Soils. Egypt Using Spectroscopy and Chemometrics Technique. – *Modeling Earth Systems and Environment*, 2, 2016, 1–17.

12. Xia, Z., S. Weichao, C. Yi, Z. Lifu, W. Nan. Predicting Cadmium Concentration in Soils Using Laboratory and Field Reflectance Spectroscopy. Science of the Total Environment, 650, 2019, 321-334.

13. Xu, L., Q. Shi, B. C. Tang, S. Xie. A New Plant Indicator (Artemisia lavandulaefolia DC.) of Mercury in Soil Developed by Fourier-Transform Near-Infrared Spectroscopy Coupled with Least Squares Support Vector Machine. – *Hindawi Journal of Analytical Methods in Chemistry*, 2019, Article ID 3240126, 6, https://doi.org/10.1155/2019/3240126.

За контакти:
Доц. д-р **Елисавета Кирилова**
Институт по инженерна химия – БАН
e-mail: e.kirilova@iche.bas.bg