RAPID, NONDESTRUCTIVE AND SIMULTANEOUS PREDICTIONS OF SOIL CONTENT IN WULING MOUNTAIN AREA USING NEAR INFRARED SPECTROSCOPY

LUO, J.[1,2§*] – WANG, Y.[1§] – LOU, W. L.[3] – ZHOU, X. L.[1,2*] – TIAN, Y. X.[1,2]

[1]Hunan Academy of Forestry, Changsha, Hunan Province, China

[2]Hunan Cili Forest Ecosystem State Research Station, Cili, Hunan, China

[3]Torch High Technology Industry Development Center, Ministry of Science & Technology, Beijing, China

[§]Contributed equally to this work.

[*]Corresponding authors
e-mail: luojia@hnlky.cn (Luo, J.); ZXL@hnlky.cn (Zhou, X. L.)

Abstract. Estimating the soil content using near infrared (NIR) spectroscopy with an appropriate method of multivariate regression analysis, is a rapid and nondestructive testing technique with the virtue of high analysis speed and easy operation. Partial least squares (PLS) and least square support vector machine (LS-SVM) as the existing models widely used in other studies were improved and employed to develop an optimal regression model for the prediction of typical soil nutrient in Wuling mountain, Hunan Province, including total nitrogen, available phosphorus, and organic carbon. The performance of models established in this research was assessed by the coefficient of determination (R) and the root mean square error of calibration (RMSEC) and prediction (RMSEP). The result showed that the pre-processing method MSC+SG displayed the highest R values in PLS and LS-SVM models, which were 0.89 and 0.91, respectively. However, compared to the PLS model, LS-SVM displayed more desirable performance on the predictions. The RSMEC and RMSEP values of LS-SVM (4.84 and 4.75, respectively) were much better compared to PLS (6.15 and 6.58, respectively).
Keywords: *near infrared spectroscopy, chemometrics, soil nutrient analysis, prediction models, partial least squares*

Introduction

Soil nutrients are crucial for plant growth. The traditional analysis methods in laboratory restrict sample numbers due to the exhausting test time. Investigating the nutrient content in laboratory is time-consuming and demands considerable expenses. In the past decades, researches have been implemented to estimate the soil nutrient content from reflectance in the near infrared ranges (vis-NIR, 700-2500 nm) since spectrum data in the vis-NIR range has been proved to be reasonably accurate and fast in practical applications (Chung et al., 2015; Chen et al., 2017; Sampaio et al., 2018). With the development of chemometric methods and combination with NIR spectroscopy practices, these techniques have been

extensively applied to estimate soil nutrient content, such as total nitrogen, available phosphorus, and organic carbon (Pandit et al., 2010; Kwon et al., 2014; Sudarno et al., 2017; Yu and He, 2017). Compared with traditional test methods, the NIR spectroscopy techniques are not only cost- and time-effective, but also non-destructive, and demand no chemical reagents during the measurement (Al-Harrasi et al., 2017; Genisheva et al., 2018). Ecological environment monitoring requires rapid investigation and response to the drastic change of soil properties. Therefore, NIR spectroscopy techniques as promising test methods used in ecological environment monitoring will be effectively monitoring the change of soil properties in certain areas.

Partial least squares (PLS) regression is a statistical method, associated with principal component regression (PCR) (Yang and Ying, 2011). However, it is not to find the hyperplane of the minimum variance between dependent and independent variables, but to find a linear regression model by projecting predictive variables and observational variables into a new space. Under one algorithm, PLS can combine multiple linear regression (MLR), principal component analysis (PCA) and correlation analysis between two groups of variables together (Suykens and Vandewalle, 1999; Stella et al., 2015). PLS is a regression modeling method of multiple dependent variables for multiple independent variables, which can solve many problems that cannot be solved by ordinary multiple regression. Support vector machine (SVM) is a new general learning method proposed by Vapnik et al. (1997) according to statistical theory. It is based on the VC dimension (Vapnik Chervonenks Dimension) theory of statistical theory and the structural risk minimization principle. It can solve small sample, nonlinear, high dimension and local minima better. Practical problems such as classification, function approximation and time series prediction have become one of the hotspots in machine learning. Least square support vector machine (LS-SVM) is an improvement of SVM. Contrasted to the standard SVM model, this method has obvious advantages: (1) using equality constraints rather than inequality constraints in standard SVM algorithm; (2) transforming quadratic programming problem into solving linear equations directly (Suykens and Vandewalle, 1999; Li et al., 2017).

In China, a large sum of farmland has been converted to forests under the policy of returning farmland to forests in the past two decades. A lot of studies focused on the effects of this policy on ecosystem processes. However, little concerns were cast on the influence on the change of soil content, especially in hilly erosion areas. Therefore, the object of this paper is to evaluate the performance of PLS and LS-SVM modeling techniques to predict the typical nutrient content in soil, including total nitrogen, available phosphorus, and organic carbon content of soil sampled in the hilly erosion area of Cili where is located in Wuling mountain area of Hunan Province.

Materials and Methods

Soil sampling and laboratory analysis

The sample sites where the soil samples collected were in Liangxi Village, Wuling mountains of West Hunan Province in China (E111°12′42.836″, N29°25′27.582″), about 7 km northwest of Cili County. The annual average temperature and precipitation in this location are 18.1°C and 1436 mm, respectively. The landscape mainly includes mountains, posts and plains, in which is dominated by red soil, and the layer of soil is more than 80 cm. The forest in the sample sites is a concentrated area of returning farmland to forest, due to its coverage rate in this area over 80%, and the vegetation types oriented at returning farmland to forests and secondary forests. In this area the soil parent rock is mainly sand shale, and the soil is primarily yellow-red soil and acid soil. *Pinus massoniana* forest, *Eucommia ulmoides* forest, *Citrus reticulata* and miscellaneous shrub forest are the dominant vegetation types in and around the sampled sites. All samples were taken randomly from 0-50 cm below the ground. In May 2018, the soil of this location was sampled in 29 different sites spanning mainly alongside the Li River, which was illustrated in *Fig. 1*. All soil samples of laboratory analysis were performed in accordance with Chinese Standards, including GB 9834, NY/T 1121.7, and HJ 615.
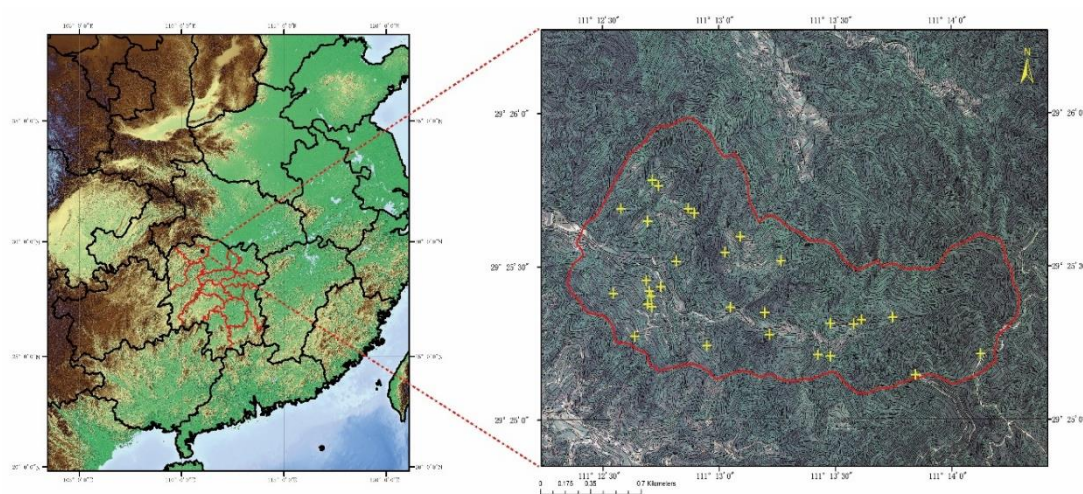

Figure 1. The location of the sites sampled: Cili, Hunan Province

Optical measurement

Vis-NIR spectra from 400 nm to 2500 nm, 2 nm/step with 1050 points was obtained by using Antaris II FT-NIR spectrometer (Thermo Fisher Scientific, the United States) charged with a sample cup. Every spectrum of all soil samples was collected with an average of 32 scans with 16 references scans before each sample.

Spectral data pre-treatment

It is prone to scatter effects for NIR spectroscopy when the samples are solid. The scattering effects are primarily caused by the distribution of sample size. Therefore, data pre-processing was crucial to establish the best models. The pre-processing techniques applied in this paper were standard normal variate (SNV), multiplicative scatter correction (MSC), 1st derivative using the Savitsky-Golay (SG) routine with 7 points and a 3rd order polynomial fitted to the data. All pre-processing steps were performed in Unscrambler X 10.4 (CAMO Software Inc., USA).

Comparison of calibration algorithms

In order to predict the of the soil nutrients content from the NIR spectrum data demands the calibration of the NIR spectra and the laboratory analysis values of soil nutrients. In our research, PLS and LS-SVM were used and compared to construct models to predict soil nutrient content. The data of the matrices are comprised of NIR spectra that was dominated by X and the values of soil content including total nitrogen, available phosphorus, and organic carbon that dominated by the vector Y.

Partial least squares regression

PLS regression is applied widely and considered as one of the most common multivariate methods for the establishment of predictive models. Usually, multiple linear regression (MLR) is utilized to build prediction models with a lot of factors. However, once the factor number is getting too large, the model will be defined as over-fitting which means that the model fits the sampled data perfectly but fails to predict the new sample data well. Under such circumstances, there are many dominant factors, but there may be few latent factors that account for most of the variation in the response. The primary concept of PLS is to attempt to extract these latent factors, accounting for as much of the dominant factor variation as possible while constructing the perfect models to fit the responses well. Therefore, PLS is also called ''projection to latent structure''(Geladi and Kowalski, 1986; Tobias, 2017).

Least square support vector machine

LS-SVM is least squares version of SVM, which is a set of related supervised learning methods that analyze data and recognize patterns to be utilized for classification and prediction analysis. In these versions, the solution is obtained by solving a set of linear equations rather than a convex quadratic programming (QP) problem for classical SVMs (Suykens and Vandewalle, 1999). By this method, the improved LS-SVM avoids the complicated calculations of the classical SVMs. Also, it can also solve the problem of the multivariate calibration relatively fast and avoid over-fitting.

Evaluation of model performance

The quantitative models between NIR spectral data and soil content were built by PLS and LS-SVM, respectively. The modeling parameters were optimized and evaluated by the root mean square error of calibration (RMSEC) and prediction (RMSEP), and the coefficient of determination (R). RMSE and R are defined as follow (*Eq. 1* and *Eq. 2*).

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

The coefficient of determination is illustrated as below.

$$R=\sqrt{1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\hat{y})^2}} \tag{Eq.2}$$

where n is the soil sample number in the validation test set, y_i and $\hat{y}_i$ are the experimental data measured by laboratory analysis and the predictive results of the model for the corresponding test sample i, respectively.

Results

Soil content

Table 1 summarizes the laboratory analysis values of total nitrogen, available phosphorus, and organic carbon of all 306 soil samples, which displays the calibration set and the validation set, respectively. The total nitrogen content for all the sets ranging from 50.88 mg·kg⁻¹ to 648.93 mg·kg⁻¹, which has an average of 205.42 mg·kg⁻¹. The contents of available phosphorus and organic carbon for all sets are varying from 0.81 mg·kg⁻¹ to 90.45 mg·kg⁻¹ and 1.86 g·kg⁻¹ to 25.44 g·kg⁻¹, respectively. The average values of these two contents are 0.15 mg·kg⁻¹ and 13.72 g·kg⁻¹ separately. The mean values in these sets of total nitrogen, available phosphorus, and organic carbon content are similar to each other, while the standard deviations are also close to each other. Therefore, it can be confident in the validation.

NIR spectrum of soil samples

The original NIR spectrum (10,000-4000 cm⁻¹) of soil samples are represented in *Fig. 2*. NIR spectra displayed different degrees of migration or drift that made pretreatment

crucial for the establishment of prediction models with desirable performance. Moreover, the bands were broad and overlap, which made the spectrum difficult to interpret. Although the variables of the soil samples had quite a wide range of values, some useful information was represented in the vis-NIR region related to the total nitrogen, available phosphorus, and organic carbon of soil samples. There were mainly three absorption bands observed at 7083 cm^{-1}, 5205 cm^{-1} and 4532 cm^{-1}, among which 5205 cm^{-1} displayed the strongest absorption relevant with the combination of bending and stretching of hydroxyl group (O-H) of soil content (Cho et al., 1998; Cécillon et al., 2009; An et al., 2015). The absorption band observed at 7083 cm^{-1} was owing to the combination of C-H second overtone and the first overtone of O-H anti-symmetric stretching of soil (Cécillon et al., 2009; Viscarra Rossel and Hicks, 2015). The NIR spectrum of soil also displayed a major absorption band at 4532 cm^{-1}, due to the stretching and bending vibrations of C-H groups linked to organic carbon (Douglas et al., 2018).

Table 1. *Statistic description of soil nutrition content*

	Data set	Samples	Min.	Max.	Mean	Std dev.	CV (%)
Total nitrogen (mg·kg^{-1})	Calibration set	204	50.22	614.51	201.39	15.25	114
	Validation set	102	51.45	652.26	213.51	16.11	112
	All data	306	50.88	648.93	205.42	15.87	113
Available phosphorus (mg·kg^{-1})	Calibration set	204	0.84	89.54	52.13	9.84	108
	Validation set	102	0.79	91.21	53.25	9.68	106
	All data	306	0.81	90.45	52.43	9.79	106
Organic carbon (g·kg^{-1})	Calibration set	204	2.35	20.01	13.87	1.25	115
	Validation set	102	1.58	24.69	13.77	1.04	112
	All data	306	1.86	25.44	13.72	1.12	114

Min. is minimum, Max. is maximum, Std dev. is standard deviation, CV is coefficient variation
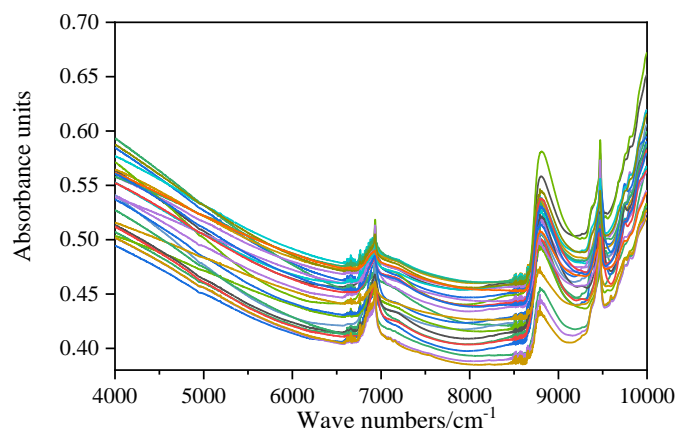

Figure 2. *Weak absorption peaks in NIR spectra of soil samples*

Performance of PLS and LS-SVM models

Pre-processing methods are quite crucial for the accuracy of models. For PLS, the most optimum calibration was obtained by applying the pre-processing method with MSC combined with SG pretreatment (RMSEC=6.15 and R_c=0.89). The validation results of MSC+SG pretreatment was also the best of all (RMSEP=6.58 and R_p=0.90).

For the spectra analyzed by LS-SVM, the most optimum calibration was achieved by using the pre-processing method also with MSC combined with SG pretreatment for prediction of total nitrogen, available phosphorus, and organic carbon content (RMSEC=4.84 and R_c=0.91), which maintained the best validation results (RMSEP=4.75 and R_p=0.93). The optimum calibrated PLS and LS-SVM models were employed to conduct predictions for total nitrogen, available phosphorus, and organic carbon content that the results confirmed the desirable accuracy of these selected models in bold in *Table 2*.

Table 2. Analysis of PLSR and LS-SVM models with selected pre-processing methods including SG, ANV+SG, MSC+SG and 1stD+SD

Model types	Data pre-processing	Calibration set		Validation set	
		R_c	RMSEC	R_p	RMSEP
PLSR	None	0.78	8.55	0.76	8.01
	SG	0.84	7.21	0.86	7.45
	SNV+SG	0.86	7.02	0.84	6.89
	MSC+SG	**0.89**	**6.15**	**0.90**	**6.58**
	1stD+SG	0.88	6.54	0.89	6.67
LS-SVM	None	0.86	6.21	0.87	6.34
	SG	0.88	5.98	0.86	5.74
	SNV+SG	0.90	5.25	0.91	5.12
	MSC+SG	**0.91**	**4.84**	**0.93**	**4.75**
	1stD+SG	0.86	4.97	0.87	4.87

The results emphasized in bold are selected models

Discussion

NIR spectra of solid samples is usually prone to scatter noise due to the uneven distribution of particle size of samples, for example soil samples (Casale and Simonetti, 2014; Acquah et al., 2015). To take full advantage of the spectra data re sample corded by NIR and to diminish the noise to the most extent, the data pre-processing methods are usually crucial to the establishment of the calibration model. The results of prediction are related to some uncertainties which mainly comprise environment conditions, measurement error, model selection and system settings. A simultaneous applications of data pre-processing methods and PLS models were applied to validate which model could

be more accurate in the predictions of soil nutrients including total nitrogen, available phosphorus, and organic carbon content. A large number of computer runs were required in order to establish different models derived from analysis of NIR spectrum data corresponding to different pre-processing methods. It is crucial to enhance the predictive ability of the analysis of the NIR spectroscopy data, particularly for batch samples of more sophisticate systems (Suykens and Vandewalle, 1999).

In addition, it is obvious that the optimal smoothing mode should be chosen in order to implement the pretreatment optimization for the following establishment of the best prediction models. The performance of PLS and LS-SVM was different to some extent when SG smoothing and the pre-processing method (MSC, 1stD or SNV) was applied separately or combined, respectively. Spectra could be normalized by SNV when there were variations of the effective path length among samples. Moreover, this kind of path length could emerge especially when investigating powdered samples. In this study, the particle size, structure and color were varying among the soil samples.

SG filter is widely used in data flow smoothing and de-noising. It is known as a smoothing method based upon the local polynomial least squares fitting in time domain (Tobias, 2017). The most important merit of this method is that it can keep the shape and width of the signal unchanged while filtering out the noise. When smoothing filter is used to filter the signal, the low-frequency components of the signal are fitted and the high-frequency components are smoothed out. If the noise is at high frequency, then the result of filtering is to remove the noise. Otherwise, if the noise is at low frequency, then the result of filtering is to leave the noise (Geladi et al., 1985; Norgaard et al., 2000). MSC is a multivariate scattering correction technique, which can effectively eliminate the baseline shift and offset caused by scattering between samples and improve the signal-to-noise ratio of the original spectra. This pre-processing method can be regarded as a suitable way to smooth the spectroscopy data of samples with different structure and size. It could be concluded from *Table 2* that the data pre-processing method (MSC+SG) displayed the highest R values in PLS and LS-SVM models, which were 0.89 and 0.91, respectively. This step of pre-processing methods could eliminate the spectrum differences in the same series of samples caused by the inhomogeneity of particle size. However, model effect could be different when the pre-processing methods were applied separately or combined. The R_c values of PLS and LS-SVM models used SG pretreatment method separately were both lower than SG combined with SNV, MSC or 1stD. Therefore, the appropriate combination should be screened for the pre-processing optimization.

Comparing the performances of PLS and LS-SVM, LS-SVM displayed better prediction ability. For PLS and LS-SVM models, all variables from the soil sample spectrum were applied to the calibration of the models. The values of RSMEC and RMSEP of LS-SVM (4.84 and 4.75, respectively) were better compared to PLS (6.15 and 6.58, respectively). LS-SVM uses the least squares linear system as the loss function to replace it, so the inequality constraint is replaced by the equality constraint method, and

the LS-SVM evolves into the solution of linear equations, which improves the speed of solution and the convergence precision of solution. Therefore, compared with PLS model, LS-SVM demonstrated more desirable ability of predictions on the content of soil nutrients.

Conclusions

Based on NIR spectroscopy and chemometrics, two quantitative prediction models of soil nutrients including total nitrogen, available phosphorus, and organic carbon were established, which were PLS and LS-SVM models. As the existing prediction models, these models were widely applied in other studies, which were improved and adopted in this study for the prediction of typical soil nutrient in Wuling mountain, Hunan Province, including total nitrogen, available phosphorus, and organic carbon. The results illustrated that these two models both had good predictive ability. The pre-processing methods could enhance the ability of prediction models. MSC+SG method displayed the highest R values in PLS and LS-SVM models, which were 0.89 and 0.91, respectively. However, compared with PLS model, LS-SVM displayed more desirable performance on the predictions. The values of RSMEC and RMSEP of LS-SVM (4.84 and 4.75, respectively) were much better compared to PLS (6.15 and 6.58, respectively). The proposed methods in this study demonstrated obvious superiority over traditional laboratory test. The results demonstrated that NIR spectroscopy combined with chemometric technique is a rapid, low cost and reliable method for the analysis of soil content in the soil monitoring and improvement along the area of Wuling mountain. Further studies will be performed to validate the model in another place located in Hunan Province. In addition, more predictive models will be developed for practical applications in future.

REFERENCES

[1]   Acquah, G. E., Via, B. K., Fasina, O. O., Eckhardt, L. G. (2015): Non-destructive prediction of the properties of forest biomass for chemical and bioenergy applications using near infrared spectroscopy. – Journal of Near Infrared Spectroscopy 23: 93-102.

[2]   Al-Harrasi, A., Rehman, N. U., Mabood, F., Albroumi, M., Ali, L., Hussain, J., Hussain, H., Csuk, R., Khan, A. L., Alam, T., Alameri, S. (2017): Application of NIRS coupled with PLS regression as a rapid, non-destructive alternative method for quantification of KBA in

Boswellia sacra. – Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 184: 277-285.

[3] An, X., Li, M., Zheng, L., Sun, H. (2015): Eliminating the interference of soil moisture and particle size on predicting soil total nitrogen content using a NIRS-based portable detector. – Computers and Electronics in Agriculture 112: 47-53.

[4] Casale, M., Simonetti, R. (2014): Review: Near infrared spectroscopy for analysing olive oils. – Journal of Near Infrared Spectroscopy 22: 59-80.

[5] Cécillon, L., Barthès, B. G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., Brun, J. J. (2009): Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). – European Journal of Soil Science 60: 770-784.

[6] Chen, Y., Delaney, L., Johnson, S., Wendland, P., Prata, R. (2017): Using near infrared spectroscopy to determine moisture and starch content of corn processing products. – Journal of Near Infrared Spectroscopy 25: 348-359.

[7] Cho, R. K., Lin, G., Kwon, Y. K. (1998): Nondestructive analysis for nitrogens of soils by near infrared reflectance spectroscopy. – Journal of Near Infrared Spectroscopy 6: A87-A91.

[8] Chung, I. M., Kim, J. K., Lee, J. K., Kim, S. H. (2015): Discrimination of geographical origin of rice (Oryza sativa L.) by multielement analysis using inductively coupled plasma atomic emission spectroscopy and multivariate analysis. – Journal of Cereal Science 65: 252-259.

[9] Douglas, R. K., Nawar, S., Alamar, M. C., Mouazen, A. M., Coulon, F. (2018): Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. – Science of The Total Environment 616-617: 147-155.

[10] Geladi, P., Macdougall, D., Martens, H. (1985): Linearization and scatter-correction for near-infrared reflectance spectra of meat. – Applied Spectroscopy 39: 491-500.

[11] Geladi, P., Kowalski, B. R. (1986): Partial least-squares regression: a tutorial. – Analytica Chimica Acta 185: 1-17.

[12] Genisheva, Z., Quintelas, C., Mesquita, D. P., Ferreira, E. C., Oliveira, J. M., Amaral, A. L. (2018): New PLS analysis approach to wine volatile compounds characterization by near infrared spectroscopy (NIR). – Food Chemistry 246: 172-178.

[13] Kwon, Y. K., Bong, Y. S., Lee, K. S., Hwang, G. S. (2014): An integrated analysis for determining the geographical origin of medicinal herbs using ICP-AES/ICP-MS and 1H NMR analysis. – Food Chemistry 161: 168-175.

[14] Li, Y., Zhang, J., Li, T., Liu, H., Li, J., Wang, Y. (2017): Geographical traceability of wild Boletus edulis based on data fusion of FT-MIR and ICP-AES coupled with data mining methods (SVM). – Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 177: 20-27.

[15] Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., Engelsen, S. B. (2000): Interval partial least-squares regression (ipls): a comparative chemometric study with an example from near-infrared spectroscopy. – Applied Spectroscopy 54: 413-419.

[16] Pandit, C. M., Filippelli, G. M., Li, L. (2010): Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression. – International Journal of Remote Sensing 31: 4111-4123.

[17] Sampaio, P. S., Soares, A., Castanho, A., Almeida, A. S., Oliveira, J., Brites, C. (2018): Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms. – Food Chemistry 242: 196-204.

[18] Stella, E., Moscetti, R., Haff, R. P., Monarca, D., Cecchini, M., Contini, M., Massantini, R. (2015): Review: Recent advances in the use of non-destructive near infrared spectroscopy for intact olive fruits. – Journal of Near Infrared Spectroscopy 23: 197-208.

[19] Sudarno, Silalahi, D. D., Risman, T., Widyastuti, B. L., Davrieux, F., Yuan, Y. Y., Caliman, J. P. (2017): Rapid determination of oil content in dried-ground oil palm mesocarp and kernel using near infraredspectroscopy. – Journal of Near Infrared Spectroscopy 25: 338-347.

[20] Suykens, J. A. K., Vandewalle, J. (1999): Least squares support vector machine classifiers. – Neural Processing Letters 9: 293-300.

[21] Tobias, R. D. (2017): An introduction to partial least squares regression. – Sas Institute Inc., Cary, NC.

[22] Vapnik, V. N. (1997): The nature of statistical learning theory. – Technometrics 38: 409-409.

[23] Viscarra Rossel, R. A., Hicks, W. S. (2015): Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. – European Journal of Soil Science 66: 438-450.

[24] Yang, D., Ying, Y. (2011): Applications of Raman Spectroscopy in Agricultural Products and Food Analysis: A Review. – Applied Spectroscopy Reviews 46: 539-560.

[25] Yu, X. L., He, Y. (2017): Challenges and opportunities in quantitative analyses of lead, cadmium, and hexavalent chromium in plant materials by laser-induced breakdown spectroscopy: A review. – Applied Spectroscopy Reviews 52: 605-622.