IMPROVEMENTS OF THE VIS-NIRS MODEL IN THE PREDICTION OF SOIL ORGANIC MATTER CONTENT USING SPECTRAL PRETREATMENTS, SAMPLE SELECTION, AND WAVELENGTH OPTIMIZATION**

Z. D. Lin [1,2,3], Y. B. Wang [1*], R. J. Wang [1*], L. S. Wang [1], C. P. Lu [1], Z. Y. Zhang [1], L. T. Song [1], Y. Liu [1]

[1] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China; e-mail: ybwang@iim.ac.cn
[2] Department of Automation, University of Science and Technology of China, Hefei 230026, China
[3] Electronic Engineering Institute, Hefei 230037, China

A total of 130 topsoil samples collected from Guoyang County, Anhui Province, China, were used to establish a Vis-NIR model for the prediction of organic matter content (OMC) in lime concretion black soils. Different spectral pretreatments were applied for minimizing the irrelevant and useless information of the spectra and increasing the spectra correlation with the measured values. Subsequently, the Kennard–Stone (KS) method and sample set partitioning based on joint x-y distances (SPXY) were used to select the training set. Successive projection algorithm (SPA) and genetic algorithm (GA) were then applied for wavelength optimization. Finally, the principal component regression (PCR) model was constructed, in which the optimal number of principal components was determined using the leave-one-out cross validation technique. The results show that the combination of the Savitzky–Golay (SG) filter for smoothing and multiplicative scatter correction (MSC) can eliminate the effect of noise and baseline drift; the SPXY method is preferable to KS in the sample selection; both the SPA and the GA can significantly reduce the number of wavelength variables and favorably increase the accuracy, especially GA, which greatly improved the prediction accuracy of soil OMC with R_{cc}, RMSEP, and RPD up to 0.9316, 0.2142, and 2.3195, respectively.

Keywords: *Vis-NIR spectroscopy, organic matter content, spectral pretreatment, sample selection, wavelength optimization.*

УСОВЕРШЕНСТВОВАНИЕ МОДЕЛИ ДЛЯ АНАЛИЗА СОДЕРЖАНИЯ ОРГАНИКИ В ПОЧВЕ МЕТОДОМ СПЕКТРОМЕТРИИ В ВИДИМОМ И БЛИЖНЕМ ИК ДИАПАЗОНАХ ЗА СЧЕТ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ СПЕКТРОВ, ВЫБОРА ОБРАЗЦОВ И ОПТИМИЗАЦИИ ДЛИН ВОЛН

Z. D. Lin [1,2,3], Y. B. Wang [1*], R. J. Wang [1*], L. S. Wang [1], C. P. Lu [1], Z. Y. Zhang [1], L. T. Song [1], Y. Liu [1]

УДК 543.42.062

[1] Институт интеллектуальных машин, Китайская академия наук, Хэфэй, 230031, Китай; e-mail: ybwang@iim.ac.cn
[2] Отдел автоматизации, Китайский университет науки и технологии, Хэфэй, 230026, Китай
[3] Институт электронной техники, Хэфэй, 230037, Китай

(Поступила 21 марта 2016)

При создании модели для анализа методом спектрометрии в видимом и ближнем ИК диапазонах содержания органических веществ (ОМС) в черноземе с известковыми включениями использованы 130 образцов поверхностного слоя почвы, взятых в графстве Гуоян провинции Аньхой, Китай. Для минимизации несоответствующей и бесполезной информации в полученных из спектров данных

** Full text is published in JAS V. 84, No. 3 (http://springer.com/10812) and in electronic version of ZhPS V. 84, No. 3 (http://www.elibrary.ru/title_about.asp?id=7318; sales@elibrary.ru).

и повышения степени их корреляции с измеряемыми величинами проведена предварительная обработка спектров. Для выбора обучающего набора использованы метод Кеннарда–Стоуна (KS) и разбиение набора образцов с совместным учетом расстояний по x-y (SPXY). С помощью алгоритма последовательных проекций (SPA) и генетического алгоритма (GA) осуществлена оптимизация длин волн. В итоге построена модель регрессии на главные компоненты (PCR), в которой оптимальное число главных компонент определено с помощью методики перекрестной проверки с исключением по одному. Показано, что комбинация фильтра Савицкого–Голея (SG) для сглаживания и мультипликативной поправки на рассеяние (MSC) может исключить влияние шума и дрейфа базовой линии; метод SPXY имеет преимущество перед KS при выборе образцов; SPA и GA могут значительно уменьшить число переменных по длинам волн и эффективно повысить точность GA и анализа OMC в почве до коэффициентов корреляции (R_{cc}), среднеквадратичной ошибки предсказания (RMSEP) и остаточной погрешности прогноза (RPD) 0.9316, 0.2142 и 2.3195 соответственно.

Ключевые слова: *спектроскопия видимого и ближнего ИК диапазонов, содержание органических веществ, предварительная обработка спектра, выбор образца, оптимизация длины волны.*

Introduction. The application of precision agriculture needs vast amounts of accurate, real-time and low-cost soil data over a large area. It sets high requirements for soil data collection. An ideal measurement method should be more efficient in both detection time and cost. Visible and near infrared reflectance spectroscopy (Vis-NIRS) is a non-destructive, rapid, and repeatable method that can provide inexpensive and real-time prediction of soil physical, chemical, and biological properties [1–6]. The NIR band ranges from 700 to 2500 nm, and the spectra in this range are mainly composed of the absorption peaks corresponding to the overtones and combinations of the fundamental vibrations due to the stretching and bending of N–H, O–H, and C–H groups [7]. Owing to the significant development in chemometrics, NIRS now is widely applied in many fields for qualitative and quantitative analyses.

Soil organic matter content (OMC) plays a major role in many chemical and physical properties of soil and significantly affects its reflectance spectra. As to the prediction of OMC using the VIS-NIRS technique, many researchers have conducted a great number of studies and achieved many favorable results. Based on the collected VIS-NIR data, the international scholars used different multivariate regression methods and systematically compared their capabilities in OMC prediction [8–12]. Principal component regression (PCR) is a well-accepted method in constructing a soil prediction model. Chang et al. [13] evaluated the ability of near infrared spectroscopy to predict soil OMC with PCR, with a determination coefficient up to 0.87 and a residual prediction deviation (RPD) of 2.79. Vasques et al. [8] identified the best combination to predict soil OMC with five multivariate techniques including PCR. Wang et al. [14] analyzed the potential of VIS-NIRS to predict soil OMC using two spectrometers, and the results showed that both spectrometers could achieve favorable results. All of those proved the feasibility of the PCR-based NIRS model in the prediction of soil OMC. However, the noise and some irrelevant or collinear information included in VIS-NIRS can affect the accuracy of the PCR-based method, and the influence should be eliminated with some measures before using the model.

In this paper, PCR was used to relate the VIS/NIR spectra with soil OMC, while spectral pretreatments, sample selection, and wavelength optimization were conducted for improving the prediction accuracy of the constructed model. By comparing the prediction results obtained with different methods, the role of spectral pretreatment, sample selection and wavelength optimization in the improvement of the soil OMC prediction capacities was evaluated.

Materials and methods. An experimental field in Tongfeng Seed Industry located in Guoyang County, Bozhou City, Anhui Province, China (33°27′~33°47′N; 115°53′~116°33′E) was selected in the present study. Mean annual temperature is about 14.6°C, and mean annual precipitation is about 830 mm. The overall flat fields in Guoyang County can be classified as lime concretion black soil. As one type of ancient cultivated soils, lime concretion black soil presents a highly localized distribution in Huaibei Plain, China. According to the China Soil Scientific Database (http://www.soil.csdb.cn/), it is composed of two layers from top to bottom, namely, a black-soil layer and a lime-concretion layer. The black-soil layer exhibits neutral or weakly alkaline properties with the measured pH value ranging from 6.0 to 8.6. It should be noted that OM, N, and P contents of the lime concretion black soil are not very high while K content is relatively high [15, 16].

A total of 130 topsoil samples was collected using an S-shaped sampling scheme (i.e., the sampling points were arranged in an S-shaped pattern). As specified in the *Technical Specification of Balanced Fertilization by Soil Testing* (NY/T 1118-2006) and Soil Testing-Part 1: *Soil Sampling, Processing and Reposi-*

tion (NY/T1121.1-2006) released by the Ministry of Agriculture of the People's Republic of China, the S-shaped soil sampling method was generally adopted for soil testing in rectangle fields, which could overcome errors caused by tillage and fertilization. The soil samples were collected from the surface layer at the bottom of a 20 cm deep trench using a special soil sampler. Each sample was about 2000 g and was placed into a tightly sealed plastic bag to avoid external contamination. After the rejection of weeds and small pieces of rocks, the soils were then naturally air-dried and sieved to less than 2 mm [17]. The soil samples after preprocessing were divided by a four-way division method into two portions A and B. The samples in Group A were placed in properly closed bags and taken to a chemical laboratory for the analysis of OMC. The soil OMC was determined by the potassium dichromate volumetric method coupled with a watering heating technique [18]. Statistical data of the measurement results are $[OMC]_{min}$ = 1.29%, $[OMC]_{max}$ = = 3.85%, mean value 2.34%, and standard deviation value 0.525%. The samples in group B were subjected to spectral measurements, which are described in the next section.

The experimental instrument employed in the present work was a VIS/NIR soil sensor by Veris Technology Incorporation. As a tractor-mounted sensor that can collect real-time soil information, the spectrophotometers and the optical system of Veris VIS/NIR soil sensor were built into a shank mounted on a toolbar and then pulled by a tractor during field investigations. Many factors in the field measurements such as temperature, humidity and light intensity can seriously affect the collected spectra and the subsequent measurement results. Thus, this study was firstly focused on the laboratory measurements, in which many environmental and soil factors were assumed to be identical. In the present work, the spectrophotometers (Ocean Optics USB4000 and Hamamatsu C9914GB) and the optical system (a tungsten halogen bulb and fibers) in the Veris Soil Sensor were dismounted from the tractor-mounted mobile platform to perform the laboratory measurements. All the data processing procedures in the present work were compiled with Matlab. A total of 130 samples was adopted for the measurements. They were put in a Petri dish, and the surface was smoothed beforehand. During the measurements, the sample surface was pressed against a sapphire window of the Veris soil sensor. As a tradeoff between the minimization of the measured errors and time consumptions, each sample was detected three times while rotated by 120° for the next scan. After each measurement, the collected three spectra were averaged. It can be found that the first two large absorption peaks are located at around 1420 and 1930 nm, both of which are coincident with the characteristic absorption peaks of H_2O [14].

The measured spectra are easily influenced by individual differences (sample particle size, intensity of light, measurement conditions, etc.), baseline variations, and substantial noises. Therefore, the pretreatment should be applied to minimize irrelevant and useless information of the spectra and increase the correlation between the spectra and the measured values. The frequently-adopted pretreatment methods include normalization, the first- and second-order derivatives, multiplicative scatter correlation (MSC), standard normal variate (SNV), and detrending or any combination thereof [19, 20].

The pretreatment methods employed in this work were MSC, SNV, Savitzky–Golay (S–G) filter for smoothing, the first-order and second-order derivative, and a combination thereof. The SNV and MSC transformations could remove the baseline drift from spectra caused by scattering and variations of particle sizes. To remove high frequency noise, the Savitzky–Golay polynomial smoothing filter can digitally smooth a given spectrum by approximating it within a specified data window using a specified order polynomial. Accordingly, the data can be best matched in the window on a least-square basis. In the present work, a filter with a polynomial of order 3 and a window with width 7 data points were used [14].

The selection of a representative training set plays a determinative role in the construction of prediction models, since the models established with the representative-characteristics samples can lead to accelerated regressions, an improvement of the prediction accuracy, and a reduction of storage space and costs. Moreover, the application range of the established models can be expanded by adding a small amount of representative samples beneficial to the model update and improvement.

In the present study, the random sample (RS), The Kennard–Stone (K-S) methods, and the method of sample set partitioning based on joint *x-y* distance (SPXY) were used for the sample selection. The RS method refers to the random selection of a certain number of samples as the training set. The K-S algorithm aims at covering the multidimensional space in a uniform manner by maximizing the Euclidean distances between the instrumental response vectors (x) of the selected samples [21]. The SPXY method extends the K-S algorithm by encompassing both x and y differences in the calculation of inter-sample distances [22].

Wavelength optimization on the full spectrum with the aim of enhancing accuracy is still a challenging task, especially when the collected spectra display strong overlapping and imperceptible distinctive features. The VIS-NIR range spectra are mainly composed of the overtones and combination bands of hydrogen

groups, and the absorption peaks are of weak intensity and relatively low sensitivity and have wide absorption band width, serious overlaps, and multiple correlations in spectral information. If the full spectrum were involved in the model, it would not only increase the complexity of the model and calculation load, but also reduce the prediction accuracy of the model owing to the irrelevant variables and collinearity between the variables.

The successive projection algorithm (SPA) selects the wavelengths according to the contribution value sequence of the test samples and looks for the original spectral data with minimum redundant information. Accordingly, the overlapping information can be avoided and the redundant information can be eliminated in the selected wavelength data. This method can greatly reduce the calculation amount and improve the model stability and accuracy [23].

The genetic algorithm (GA) is a kind of random search optimization algorithm based on the rule of biological evolution. Owing to the invisible parallelism and adaptive and global optimization capacity, GA has become a common method for the optimization of wavelength in the construction of NIR prediction models. In combination with GA algorithm, the constructed NIR prediction models exhibit relatively high predictive abilities [24].

Calibration and validation. Among the training sets after selection with 130 samples, the calibration and validation sets were selected with the aim of predicting an unknown sample scientifically and exactly. The selected 100 samples were divided as the calibration data set, and the rest of the 30 samples were used as the validation set. In the calibration stage, the spectra were compressed using principal component analysis (PCA), and an optimum number of the principal components (PCs) was determined using the leave-one-out cross validation (LOOCV) technique, in which each sample was omitted and predicted using the calibration model established by the remaining samples [25]. With the pre-processed spectra, values of the prediction residual error sum of squares (PRESS) in the leave-one-out cross validation for different numbers of PCs and soil contents were calculated. Thus, the number of PCs corresponding to the least PRESS values was determined as the optimal number of PCs. A suitable number of PCs is quite crucial for taking full advantage of the spectral information and noise-filtering, while some useless information, such as the measured errors, can be over-included for a greater number of PCs, also known as "over-fitting". Subsequently, the models were constructed using the PCR method, in which multiple linear regressions were performed using the obtained optimum numbers of PCs. In the validation stage, the above-described calibration model developed from the training set (100 samples) was used to predict contents of the soil samples in the validation set (30 samples), and the predicted values were compared with the measured ones. The statistic parameters for evaluating the predictive capability of the models include the correlation coefficient (R_{cc}), RMSEP, and RPD [26].

Results and discussion. The selection of an appropriate pretreatment method is very important for the establishment of the NIR prediction model. As stated above, the NIR spectra are affected by many factors, such as collinearity, physical properties, light scattering, machine noise, and so on. Generally, different methods were applied and compared for selecting the most appropriate one according to the requirements. The pretreatment could mine the weak signals and imperceptible information through some transformations on the original spectra. The combination and the sequence of different pretreatment methods need to be optimized in the practical applications. In this study, the PCR-based prediction models for OMC after different pretreatments were developed using the training set, and the prediction results are listed in Table 1. It can be seen that different methods have different influences, and the combination of the S-G filter for smoothing and MSC exhibit the most favorable results. Thus, it was chosen for later elaborate discussions.

TABLE 1. Prediction Results of PCR Models with Different Pretreatments

Pretreatment	PCs	R_{cc}	RMSEP	RPD
Smooth	15	0.5307	0.4123	1.189
1st derivative	7	0.6296	0.3547	1.3013
2nd derivative	2	0.6808	0.3292	1.3116
MSC	9	0.6324	0.4053	1.3005
SNV	9	0.6206	0.4093	1.2879
MSC+2nd derivative	7	0.6261	0.3837	1.2704
MSC+1st derivative	6	0.5857	0.3927	1.2559
MSC+ smooth	20	0.5779	0.4076	1.2079
SNV+ smooth	18	0.7086	0.3434	1.4337
SNV+1st derivative	6	0.5854	0.3928	1.2556

Continue Table 1

Pretreatment	PCs	R_{cc}	RMSEP	RPD
SNV+2nd derivative	7	0.6415	0.3738	1.3058
Smooth +MSC	17	0.7198	0.3432	1.5155
1st derivative +MSC	8	0.5956	0.3438	1.2317
2nd derivative +MSC	3	0.5801	0.3766	1.2215
BS+SNV	19	0.5818	0.3823	1.1593
1st derivative +SNV	8	0.5981	0.3432	1.2337
2nd derivative +SNV	3	0.5805	0.3761	1.2233

Table 2 shows the results of the PCR models with different sample selection methods. It can be seen that, using the random selection (RS), the prediction results are relatively poor; the prediction results using the K-S sample selection are better than the results using RS as R_{cc} increases from 0.329 to 0.7198, RMSEP decreases from 0.4537 to 0.3432, and RPD increases from 1.2826 to 1.5155; the prediction results using the SPXY results are best, with R_{cc}, RMSEP, and RPD up to 0.8298, 0.2738, and 1.7216, respectively. Consequently, the SPXY method was chosen for further in-depth discussions in calibration and validation.

TABLE 2. Prediction Results of PCR Models with Different Sample Selection Methods

Method	PCs	R_{cc}	RMSEP	RPD
RS	17	0.329	0.4537	1.2826
KS	17	0.7198	0.3432	1.5155
SPXY	17	0.8298	0.2738	1.7216

It can thus be concluded that, using the RS method, the samples are randomly selected by the computer, with strong randomness and poor repeatability, the selected training set cannot adequately represent the whole samples, and the constructed NIRS models are generally not accurate and robust enough; using the KS method, the number of similar samples in the training set can be reduced, and thus the constructed prediction model is superior to the model using the RS method; using the SPXY method, both the spectral and target variables were taken into account, so the selected training set is more representative than that selected using the KS method and the established model is best in prediction accuracy.

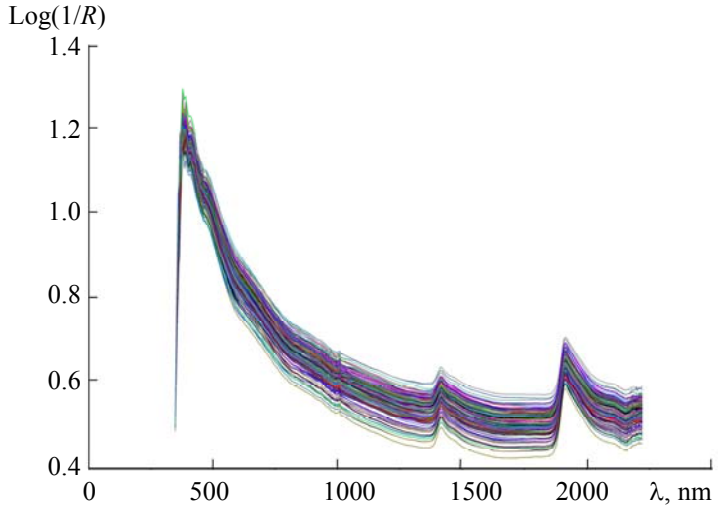

Fig. 1. Correlation between the predicted and the measured values of OMC using GA-based PCR model.

Table 3 lists the results of the PCR models with the use of different wavelength optimization methods, and Fig. 1 shows a comparison between the predicted results and the measured values using the GA-based PCR model. As shown in Table 3, SPA and GA can both increase the PCR model accuracy and greatly decrease the predicting errors. Both two methods can contribute to the optimization of wavelengths so as to

remove the effects of noise and enhance the predictive capability. This can be reflected in the increased correlation coefficient R_{cc} and RPD as well as the decreased RMSEP.

TABLE 3. Prediction Results of PCR Models with Different Wavelength Optimization Methods

Method	PCs	R_{cc}	RMSEP	RPD
NULL	17	0.8298	0.2738	1.7216
SPA	7	0.8712	0.2362	1.9957
GA	18	0.9316	0.2142	2.3195

SPA and GA employ simple operations in a vector space to obtain the subsets of variables with small collinearity, which can effectively eliminate the redundant information of the wavelength variables and thus improve the prediction precision. Specifically, SPA requires less computational work load than GA; however, the GA-based prediction model has a better prediction precision. Since GA takes the serial correlation characteristics of the spectral data into account in the optimization of wavelength, more spectral information is included, and the GA-PCA model is superior to the SPA-PCA model in the prediction of soil OMC.

Conclusion. Visible-near infrared spectroscopy (Vis-NIRS) was proved to be an effective tool in the prediction of soil properties. Spectral pretreatments, sample selection and wavelength optimization play important roles in the construction of the Vis-NIRS prediction model. In order to establish an accurate and robust prediction model for soil OMC, 17 pretreatment methods, three sample selection methods and two wavelength optimization methods were applied in the Vis-NIRS model establishment. Results show that the combination of the S–G filter for smooth and MSC can effectively eliminate the effects of noise and baseline drift. Both KS and SPXY can select the representative samples, but the SPXY method overall considers both x and y differences in the calculation of inter-sample distances and can select the more representative samples and give more accurate results. SPA and GA can decrease the number of wavelengths enormously, simplify the model and increase the accuracy remarkably. SPA requires less computational work, but GA is preferable in prediction accuracy. In conclusion, after spectral processing and selection of the training set, the GA-PCR model can accurately predict soil OMC and fully meet actual needs.

This work was supported by the Science and Technology Service Network Initiative of Chinese Academy of Sciences (No. KFJ-EW-STS-069).

REFERENCES

1. E. Ben-Dor, A. Banin, *Soil Sci. Soc. Am. J.*, **59**, 364–372 (1995).
2. J. B. Reeves, G. W. McCarty, J. J. Meisinger, *JNIRS*, **8**, 161–170 (2000).
3. J. B. Reeves, G. W. McCarty, V. B. Reeves, R. F. Follet, J. M. Kimble, *Abstr. Am. Chem. Soc.*, **223**, U141–U142 (2002).
4. B. W. Dunn, H. G. Beecher, G. D. Batten, S. Ciavarella, *Aust. J. Exp. Agric.*, **42**, 607–614 (2002).
5. K. D. Shepherd, M. G. Walsh, *SSSA*, **66**, 988–998 (2002).
6. K. Islam, B. Singh, A. B. McBratney, *Aust. J. Soil Res.*, **41**, 1101–1114 (2003).
7. E. Ben-Dor, J. Irons, G. F. Epema, In: Soil Reflectance: Remote Sensing for the Earth Science, Ed. N. Andrew Rencz, 3rd edn., Manual of Remote Sensing, **3** (1999).
8. G. M. Vasques, S. Grunwald, J. O. Sickman, *Geoderma*, **146**, 14–25 (2008).
9. G. M. Vasques, S. Grunwald, J. O. Sickman, *Soil Sci. Soc. Am. J.*, **73**, 176–184 (2009).
10. A. M. Mouazen, B. Kuang, J. De Baerdemaeker, H. Ramon, *Geoderma*, **158**, 23–31 (2010).
11. A. Stevens, T. Udelhoven, A. Denis, B. Tychon, R. Lioy, L. Hoffmann, B. Van Wesemael, *Geoderma*, **158**, 32–45 (2010).
12. R. A. Viscarra Rossel, T. Behrens, *Geoderma*, **158**, 46–54 (2010).
13. C. W. Chang, A. L. David, J. M. Maurice, R. H. Charles, *Soil Sci. Soc. Am. J.*, **65**, 480–490 (2001).
14. Y. B. Wang, T. Y. Huang, J. Liu, Z. D. Lin, S. H. Li, R. J. Wang, Y. J. Ge, *Comput. Electron. Agric.*, **111**, 69–77 (2015).
15. L. W. Liu, *Pedosphere*, **1**, 3–15 (1991).
16. L. J. Li, X. S. Guo, D. Z. Wang, Y. X. Sun, C. L. He, P. P. Wu, J. *Anhui Agric. Sci.*, **34**, 722–723 (2006).
17. M. Nathan, Ron. Gelderman, *Recommended Chemical Soil Test Procedures for the North Central Region*, North Central Regional Research Publication No. 221 (Revised, 2012).

18. R. Lu, *Chemical Analysis Method of Agricultural Soil*, China Agricultural Science Press, Beijing, 106–107 (in Chinese) (2000).
19. A. Rinan, F. Vanden Berg, S. B. Engelsen, *Trends Anal. Chem.*, **28**, 1201–1222 (2009).
20. A. Savitzky, M. J. E. Golay, *Anal. Chem.*, **36**, 1627–1639 (1964).
21. W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I. R. Last, K. A. Prebble, *Chemometr. Intell. Lab. Syst*., **33**, 35–46 (1996).
22. R. Kawakami Harrop Galvão, M. César Ugulino Araujo, G. Emıdio Jose, M. Jose Coelho Pontes, E. Cirino Silva, T. Cristina Bezerra Saldanha, *Talanta*, **67**, 736–740 (2005).
23. M. C. U. Argújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama. H. C. Chame, V. Visani, *Chemometr. Intell. Lab. Syst*., **57 (2)**, 65–73 (2001).
24. W. Z. Lu, H. F. Yuan, G. T. Xu, *Model NIR Spectroscopy*, Beijing, China Petro-chemical Press, 56–67 (2001).
25. H. Abdi, L. J. Williams, *Wiley Interdiscipl. Rev.:Comput. Stat.*, **2**, 433–459 (2010).
26. A. M. Mouazen, J. D. Baerdemaeker, H. Ramon, *JNIRS*, **14**, 189–199 (2006).