Rapid Detection of Green Sichuan Pepper Geographic Origin Based on Near-Infrared Spectroscopy

Yan Cai[*], Dianxu Ma, Xiaopan Li

School of Physics and Information Engineering, Zhaotong University, Zhaotong, Yunnan, China

*Corresponding author

Keywords: Near-infrared spectroscopy; Green Sichuan pepper; Origin identification; Pretreatment; Rapid detection

Abstract: This study explores a method for rapid detection of the origin of Green Sichuan Pepper based on near-infrared spectroscopy. A total of 260 samples of peppercorns from 5 producing areas including Longtoushan Town in Ludian County, Suoshan Town in Ludian County, Xiaozhai Town in Ludian County, Jiangdi Town in Ludian County and Tianba Town in Zhaoyang District were collected. Spectra are collected. The original spectra were preprocessed by methods such as wavelet decomposition and denoising, and then radical basic function (RBF), support vector machine (SVM), and partial least squares (PLS) were used to establish the origin identification model. The research shows that the identification accuracy of SVM and RBF neural network models are significantly improved, up to 100%; wavelet decomposition denoising and baseline correction can significantly improve the accuracy of the Green Sichuan Pepper identification model. The rapid detection method of Green Sichuan Pepper based on near-infrared spectroscopy is feasible

1. Introduction

Zanthoxylum is a perennial woody deciduous shrub or small tree of Rutaceae. There are about 250 species in the world, which are distributed in tropical and subtropical regions of Asia, America, Africa and Oceania. Among them, there are about 39 species and 14 varieties in my country. Most of the varieties of Green Sichuan Pepper are still in the wild state. The main varieties of artificially cultivated Green Sichuan Pepper are bamboo leaf Green Sichuan Pepper and Green Sichuan Pepper, which are usually divided into Green Sichuan Pepper and red Green Sichuan Pepper according to the color of the fruit. The cultivation areas are mainly concentrated in China, Japan and South Korea in Asia. my country is the country of origin of Green Sichuan Pepper cultivation, and it is also the country with the largest cultivation area. At present, the planting area has exceeded 120,000 hectares, and Green Sichuan Pepper planting bases such as Zhaotong in Yunnan, Fengxian in Shaanxi, Baoji in Shaanxi, Ruicheng in Shanxi, and Hancheng in Shaanxi have gradually formed.

The main chemical components in Zanthoxylum are alkaloids, amides, lignans, volatile oils, fatty acids, coumarin, etc. Modern natural product chemistry and pharmacology studies have shown that these bioactive components in Zanthoxylum bungeanum have antioxidant, anti-tumor, anti-inflammatory and Antibacterial and antiseptic function. Many scholars at home and abroad have done a lot of research and reports on the composition and content of Zanthoxylum bungeanum. The research results show that there are indeed certain differences in the chemical composition of the same species of Zanthoxylum bungeanum in different cultivation areas. The chemical composition is also quite different. The content and composition of the chemical components of Zanthoxylum bungeanum are not only related to the extraction site and extraction method, but also largely depend on the species and production environment of Zanthoxylum bungeanum. However, with the in-depth development and utilization of Green Sichuan Pepper getting more and more attention, the market demand has grown rapidly, followed by the phenomenon of replacing the inferior with the superior, using the old to fake the new, and adulteration. On the one hand, consumers are starting to care more about the origin and authenticity of the agricultural products they consume. On the other hand, companies and the origin of Zanthoxylum are eager to seek

DOI: 10.25236/icited.2022.001

effective ways to protect their own brands, so they have researched and developed a simple, fast, and non-destructive the identification and detection method of the origin of Green Sichuan Pepper has important practical significance.

Near-infrared spectroscopy has the characteristics of non-polluting, non-destructive, low analysis cost and fast speed and is widely used in qualitative and quantitative analysis of agricultural products, food, medicine, etc. At present, near-infrared spectroscopy technology has been successfully applied to trace the origin of maca, the origin of tea oil, and the origin of coffee. In this paper, the complex chemical components of Zanthoxylum bungeanum are taken as a whole, and the near-infrared diffuse reflectance spectrum is used to construct the atlas library of Zanthoxylum bungeanum. At the same time, the pattern recognition method of Zanthoxylum bungeanum samples from different origins is combined with the pattern recognition method to establish a fast, simple and accurate origin of Zanthoxylum bungeanum identification method.

2. Experimental Part

2.1. Material collection

Collect samples from the main producing areas of Green Sichuan Pepper from the local Green Sichuan Pepper market and Green Sichuan Pepper growers in Zhaotong City. There was no obvious difference in appearance between the purchased Green Sichuan Pepper. The variety of Green Sichuan Pepper is determined by the assignment method. The Green Sichuan Pepper in Longtoushan Town, Ludian County is assigned as 1, the Green Sichuan Pepper in Suoshan Town, Ludian County is assigned as 2, the Green Sichuan Pepper in Xiaozhai Town in Ludian County is assigned as 3, and the Green Sichuan Pepper in Jiangdi Town in Ludian County is assigned as 3. The Green Sichuan Pepper is assigned a value of 4, and the Green Sichuan Pepper in Tianba Town, Zhaoyang District is assigned a value of 5.Before the spectrum collection experiment, the collected Green Sichuan Pepper samples were stored in a dry environment at room temperature. Use a hand-held pulverizer (model A-11-B-S25) produced by IKA company to pulverize the samples, pass through an 80-mesh sieve to ensure that the particle size of the samples is consistent, and prepare 260 samples of 20.0g each, which are stored Sealed and numbered in a ziplock bag. Among the 260 samples, 60 samples were from Longtoushan Town, Ludian County, 50 samples were from Suoshan Town, Ludian County, 50 samples were from Xiaozhai Town, Ludian County, 50 samples were from Jiangdi Town, Ludian County, and 50 samples were from Tianba Town, Zhaoyang County.

2.2. Instruments and equipment

Bruker MPA near-infrared spectrometer; OPUS 6.0; MATLAB R2015a; Unscrambler 10.4; Origin 8.0.

2.3. Spectrum acquisition

The spectrometer was turned on and warmed up for 30 min before scanning the sample to ensure the stability of the sample test. At about 25° C, 260 samples of Zanthoxylum bungeanum were scanned in the full spectrum: The range was 12000-4000 cm^{-1}, the number of times was 20, and the resolution was 4 cm^{-1}.

2.4. Spectral preprocessing method

Spectral images obtained by scanning often contain noise, which is caused by the environment where the instrument is placed and the instrument itself; on the other hand, the light source with other spectral interferences or the matrix of the sample will also affect the spectrum. Instrument and background noise can affect the accuracy of the analysis. Preprocessing can reduce high-frequency random noise, strengthen the characteristic information of samples, and make the model more stable. Commonly used methods for smoothing include baseline correction, smoothing, wavelet decomposition and denoising [1]. The wavelet modulus maximum denoising method in wavelet decomposition denoising has a large amount of calculation and low efficiency. When the number of layers is low, the coefficients are greatly affected by noise, resulting in pseudo-extreme points; when the number of layers is high, local characteristics will be lost, and the low-frequency coefficients will be directly reconstructed. It is easy to lose useful components in high frequency coefficients. Threshold-based denoising methods can be approximately optimal under the minimum mean square error. Due to the large number of wavelet basis functions, it is difficult to conduct comprehensive experiments on all parameters of wavelet denoising to find the optimal parameter combination. According to the previous literature, the wavelet basis functions coif2, haar, sym5, etc. with better performance are selected as the candidate wavelet basis functions [2]. After many screenings and comparisons, the optimal parameter combination is obtained, namely haar, db5 and sym5 wavelet functions, the number of decomposition layers is 5, and the threshold scheme is sqtwolog rule [3].

2.5. Data analysis methods

2.5.1. Partial least squares discriminant method

Partial least squares (PLS) refers to decomposing the concentration matrix $Y=(y_{i,j})_{n\times m}$ of m components of n samples and the absorbance matrix $X=(x_{i,j})_{n\times m}$ at p wavelength points of n samples measured by the instrument into the form of eigenvectors [4].

$$Y= UQ+ F \tag{1}$$

$$X=TP+E \tag{2}$$

Where U and T are the concentration eigenfactor matrix and absorbance eigenfactor matrix of n rows and d columns, respectively, Q is the $d\times m$ order concentration loading matrix, P is the $d\times p$ order absorbance loading matrix, and F and E are the $n\times m$, $n\times p$ order concentration residual matrix, respectively and absorbance residual matrix.

Then, build the PLS regression model

$$U=TB+E_d \tag{3}$$

Where E is the random error matrix, and B is the d-dimensional diagonal regression coefficient matrix.

For the unknown sample to be tested, if the absorbance vector is x, its concentration can be solved as

$$y=x(UX)^{'}BQ \tag{4}$$

In the PLS algorithm, if the concentration variable in the matrix is replaced by a binary variable representing the category attribute to calculate the correlation between the spectral vector and the category vector, it is called the discriminant partial least squares method. $Y=(y_{i,j})_{n\times m}$ is changed to the following category matrix form.

$$Y=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{5}$$

Among them, each column of the Y matrix represents a sample category. In the mixture, "1" indicates that it belongs to this category, and "0" indicates that it belongs to other categories, that is, $y_{ij}=1$ indicates that the ith sample belongs to the jth category, and $y_{ij}=0$ indicates that the ith sample does not belong to the jth category. belongs to category j. Establish a PLS regression model and set the category attribute discrimination threshold θ. When predicting, for k unknown samples, the predicted value $X=(x_{i,j})_{n\times p}$ of the model can be obtained according to its spectral matrix $X=(x_{i,j})_{n\times p}$, and the maximum value of the i-th row of $y_{k\times m}$ can be obtained.

$$y_i = \max(y_{i1}, y_{i2}, \cdots y_{in}) \tag{4}$$

If $y_i > \theta$, it is considered that the column number j where y_i is located is the category to which the ith sample belongs; otherwise, the ith sample does not belong to any known category.

The evaluation index of the qualitative analysis model is the correct identification rate (CIR), that is, the percentage of correctly judged samples in the total number of samples.

2.5.2. Support vector machines

Support vector machine is a machine learning method suitable for processing small sample and nonlinear data, and is widely used in data classification, model prediction and various regression analysis. In low-dimensional space, vector sets are often difficult to divide, and SVM maps them to high-dimensional space for analysis. The data is partitioned by finding a hyperplane in a high-dimensional space. The complexity of data calculation in high-dimensional space can be solved by different kernel functions. The diversity of kernel functions greatly increases the diversity and flexibility of the SVM algorithm. The selection of the kernel function is based on the known data. The error in this process is corrected by determining the relaxation coefficient. The fewer the number of support vectors, the smaller the error [5].

2.5.3. RBF neural network

In 1980 Powell introduced the concept of RBF neural network. RBF neural network has input layer, output layer and hidden layer. The input layer is the perception unit, which is the bridge between the inside and outside of the network. The radial basis function constitutes the hidden layer as a hidden unit, completes the nonlinear transformation, and the output layer is responsible for responding. The basis of RBF neural network is function approximation theory, and interpolation is an important part of function approximation. The RBF neural network has the characteristics of simple and easy training method, fast convergence, and good fitting effect to nonlinear functions [6].

2.5.4. Linear discriminant analysis

Linear discriminant analysis was proposed by Fisher in 1936. Project multi-dimensional data in a direction so that all data in this direction satisfies the maximum distance between classes. The same class of sample data has the smallest intra-class distance. The categorical separation of the data works best. Both the dimensionality is compressed and the features are extracted. Linear discriminant methods are often used in the research of face recognition, image classification and forest coverage [7].

3. Results and Discussion

3.1. Near-infrared spectrum of the sample

Figure 1 is the original near-infrared spectrum of the Zanthoxylum bungeanum sample. It can be seen that the original spectrum has obvious absorption peaks near wavenumbers 8420, 6750, 5840, 5170, 4700 and 4300cm^{-1}, and the absorption peaks at 8420, 5840 and 4300cm^{-1} may be it is caused by the combined frequency, double frequency and triple frequency absorption of CH groups. Because the double frequency region of H stretching vibration is near 6 700 cm^{-1}, and a combined frequency absorption region of H_2O is near 5155 cm^{-1}, the absorption at 6750 and 5170 cm^{-1} should be caused by the water in Zanthoxylum bungeanum. In addition, it can be seen from the figure that due to the serious overlap of near-infrared spectral information, it is difficult to visually identify the characteristic information of each sample from the peak position and other aspects. Therefore, the qualitative analysis of Zanthoxylum bungeanum samples by near-infrared combined with chemometric methods must be processed and extracted by appropriate mathematical methods [8].

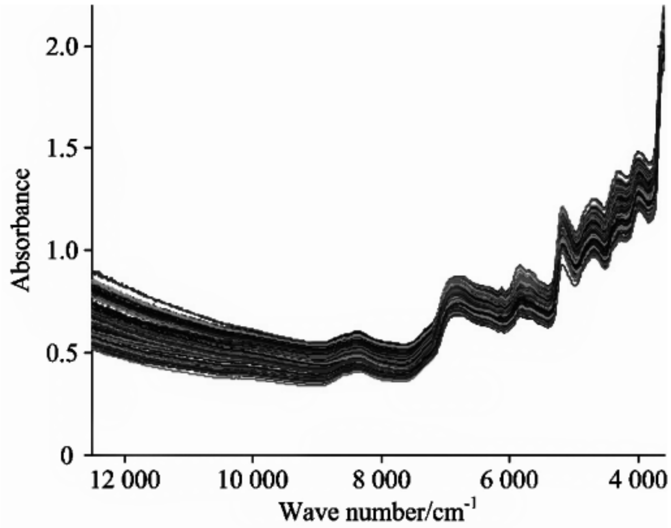

Figure 1 Near-infrared spectra of 260 samples of Zanthoxylum bungeanum.

3.2. Identification of origin

The experimental results are preprocessed, all methods weaken the effect of scattering, the spectral curve after preprocessing is smoother, and the characteristic peak points of the waveform are not changed. The comparison found that the data after baseline correction, SNV, baseline correction combined with MSC, baseline correction combined with SNV and sym5 wavelet decomposition were highly consistent with the original data in the spectral line trend [9].

However, the spectral trend of the data after nonlinear trend removal (DET) processing is significantly different from the original data. Baseline correction is a method of subtracting the minimum value from the original spectral value, so that the spectral line variation and value distribution are closest to the original data. For signals with better continuity, sym5 wavelet has better denoising effect among the four wavelet methods selected. Some useful information is filtered out at the same time as random noise. Among many preprocessing methods, the model accuracy of wavelet decomposition denoising method is generally higher than that of other preprocessing methods. In the calibration set and prediction set, the root mean square error of the model after wavelet denoising preprocessing is lower than 0.01. Among them, the model accuracy after db5 wavelet preprocessing is the best. Comprehensive consideration, this study chooses baseline correction and wavelet decomposition denoising method as the best preprocessing method for classification modeling.

The Kennard-Stone algorithm was used to select 185 samples from the 260 samples of Zanthoxylum bungeanum as the calibration set, and the remaining 75 samples were used as the prediction set [10].

Among the 185 samples, 45 samples were from Longtoushan Town, 40 samples from Suoshan Town, 40 samples from Xiaozhai Town, 30 samples from Jiangdi Town, and 30 samples from Tianba Town. Among the 75 samples, there are 15 samples from Longtoushan Town, 9 samples from Suoshan Town, 8 samples from Xiaozhai Town, 23 samples from Jiangdi Town and 20 samples from Tianba Town. The numbers of Longtoushan Town, Suoshan Town, Xiaozhai Town, Jiangdi Town and Tianba Town are "1", "2", "3", "4" and "5" respectively. PCA dimension reduction is performed on the preprocessed spectral data decomposed by baseline correction, sym5 wavelet, db5 wavelet and haar wavelet [11].

Select the appropriate number of principal components to establish three origin identification models of SVM, PLS and RBF. The classification results of the prediction and output model of origin identification established by RBF neural network are shown in figure 2.

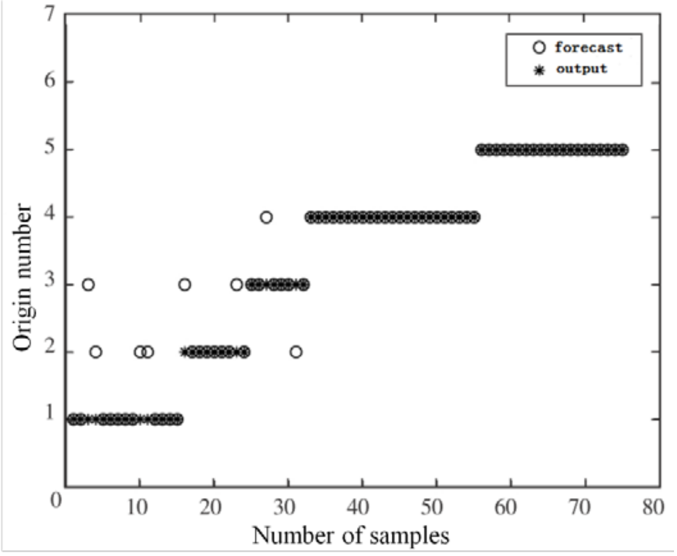

Figure 2(a) Raw spectrum.

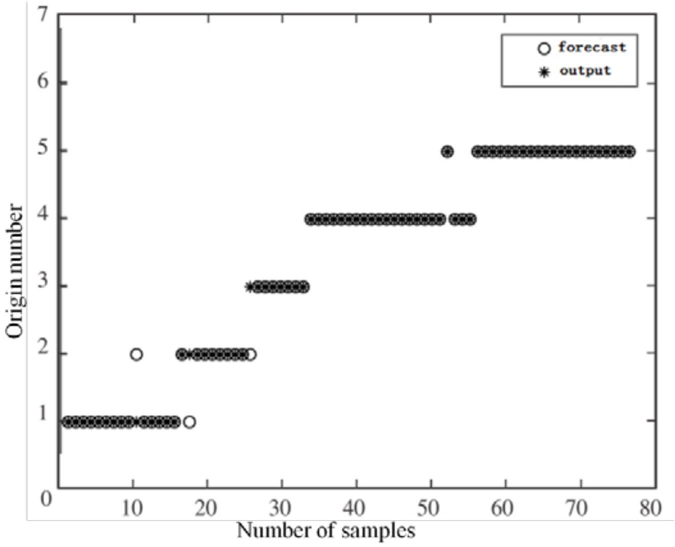

Figure 2(b) Baseline correction.

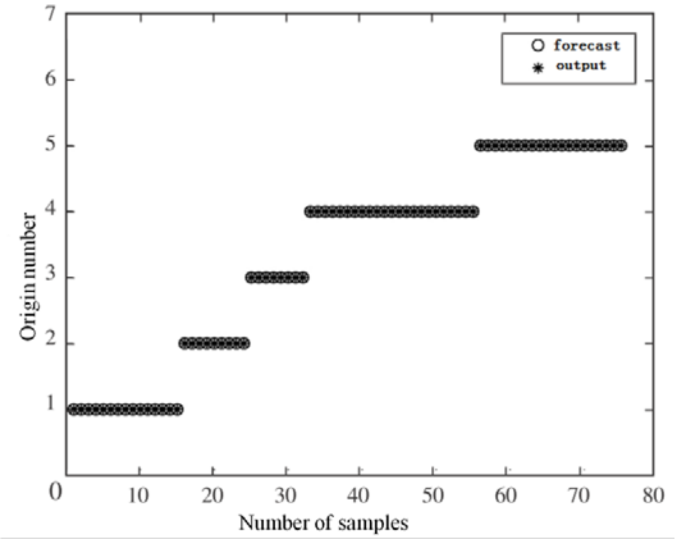

Figure 2(c) Dbs wavelet denoising.

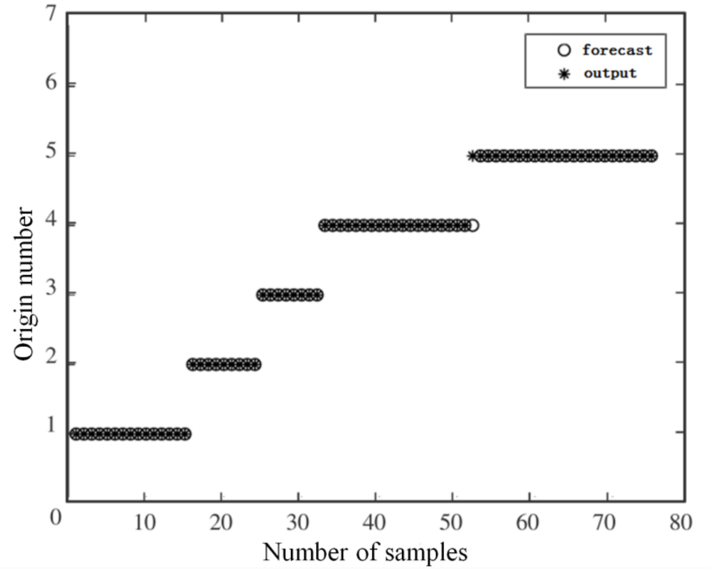

Figure 2(d) Haar wavelet denoising.

Figure 2 RBF neural network origin identification prediction and output results

Figure 2(a) shows the discrimination result of the RBF neural network without preprocessing the original spectrum. In the prediction set, there are 4 deviations in the 15 pepper samples from Longtoushan Town, 2 deviations in the 9 pepper samples from Suoshan Town, and 2 deviations in the 8 samples from Xiaozhai Town. The identification accuracy rate was 89.33%.

Figure 2(b) shows the discrimination results of the RBF neural network after baseline correction. In the prediction set, there is 1 deviation in the 15 pepper samples from Longtoushan Town, 1 deviation in the 9 pepper samples from Suoshan Town, and 1 deviation in the 8 pepper samples from Xiaozhai Town. The identification accuracy rate was 96.00%.

Figure 2(c) shows the identification results of the RBF neural network after db5 wavelet denoising preprocessing. There is no prediction bias in the 75 samples in the prediction set, and the identification accuracy rate is 100.00%.

Figure 2(d) shows the identification results of the RBF neural network after the haar wavelet denoising preprocessing. There is one deviation in the 20 Jiangdi Town pepper samples in the prediction set. The identification accuracy rate was 98.67%. Table 1 shows the results of three modeling methods, namely RBF, SVM, and PLS, for identifying and classifying the origin of Green Sichuan Pepper under different preprocessing methods.

Table 1 Origin identification and classification results under different pretreatment methods

Rreprocessing method	Principal Component Score	Prediction accuracy rate (%)		
		RBF	SVM	PLS
Raw spectrum	16	89.33	95.50	91.80
Baseline correction	16	96.00	96.40	94.24
sym5 wavelet decomposition	6	100.00	100.00	81.00
db5 wavelet decomposition	7	100.00	100.00	80.00
haar wavelet decomposition	12	98.67	95.50	73.00

In the full spectrum range, different preprocessing methods have different modeling effects. In the case of no preprocessing, the PCA dimensionality reduction is performed on the data, the optimal number of principal components is 16, and the best classification model among the three models is SVM, and the accuracy rate reaches 95.50%. The original spectrum is preprocessed by baseline correction and sym5 and other three different wavelet decomposition, and then the PCA dimension reduction is used to select the best principal component fractions for modeling. Among them, the discrimination accuracy of the SVM model and the RBF neural network model is greater than or equal to the accuracy of the original spectral discrimination model, while the accuracy of the PLS discrimination model has decreased, which may be caused by insufficient features extracted while compressing the dimensions. of. The spectral data after baseline correction and db5 wavelet denoising can achieve better modeling and classification effect after PCA dimensionality reduction, up to 100%. Overall, the classification effect of the SVM model is better than that of the PLS and RBF neural network models.

4. Conclusion

A qualitative identification model of the origin of Zanthoxylum bungeanum samples was established by using SVM, PLS and RBF neural network. When the spectrum is not preprocessed, the accuracy of the three models is up to 95.50%. After the spectrum is preprocessed by baseline correction and wavelet analysis, and the data dimension is reduced by PCA, the discrimination accuracy of SVM and RBF neural network models are significantly improved, up to 100%. The analysis shows that wavelet decomposition denoising and baseline correction can significantly improve the accuracy of the Zanthoxylum bungeanum discriminant model, and the rapid detection method of Zanthoxylum bungeanum based on near-infrared spectroscopy is feasible.

Acknowledgements

Yunnan Provincial Department of Education Scientific Research Fund (2022J0978)

References

[1] Xi Yu Wu et al. (2017) Quantitative Identification of Adulterated Green Sichuan Pepper Powder by Near-Infrared Spectroscopy Coupled with Chemometrics. Journal of Food Quality, 2017, 1-7.

[2] Xi Yu Wu et al. (2017) Quantitative Identification of Adulterated Green Sichuan Pepper Powder by Near-Infrared Spectroscopy Coupled with Chemometrics. Journal of Food Quality, 2017, 1-7.

[3] Donghui Luo et al. (2015) The application of stable isotope ratio analysis to determine the geographical origin of wheat. Food Chemistry, 174, 197-201.

[4] Zhao Li et al. (2014) Application of Vis/NIR Spectroscopy for Chinese Liquor Discrimination. Food Analytical Methods, 7, 1337-1344.

[5] Ma Inmaculada González-Martín et al. (2014) Chilean flour and wheat grain: Tracing their origin using near infrared spectroscopy and chemometrics. Food Chemistry, 145, 802-806.

[6] Niedzielski Przemysław and Krueger Michał and Brandherm Dirk (2020) Effects of sample processing on XRF results from archeological pottery. Materials and Manufacturing Processes, 35, 1455-1460.

[7] Syah Rahmad et al. (2021)Implementation of artificial intelligence and support vector machine learning to estimate the drilling fluid density in high-pressure high-temperature wells. Energy Reports, 7, 4106-4113.

[8] Ernest Teye et al. (2014)Feasibility study on the use of Fourier transform near-infrared spectroscopy together with chemometrics to discriminate and quantify adulteration in cocoa beans. Food Research International, 55, 288-293.

[9] Ahmed Rady et al. (2019) The Effect of Light Intensity, Sensor Height, and Spectral Pre-Processing Methods When Using NIR Spectroscopy to Identify Different Allergen-Containing Powdered Foods. Sensors, 20, 230-230.

[10] Congming Zou et al. (2019) Scalable calibration transfer without standards via dynamic time

warping for near-infrared spectroscopy. Analytical Methods, 11, 4481-4493.

[11] Salvatore Genovese et al. (2014) HPLC analysis of 4'-geranyloxyferulic and boropinic acids in grapefruits of different geographical origin. Phytochemistry Letters, 8, 190-192.