Talanta 115 (2013) 548–555

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Preliminary study on classification of rice and detection of paraffin in the adulterated samples by Raman spectroscopy combined with multivariate analysis

Xinwei Feng, Qinghua Zhang, Peisheng Cong, Zhongliang Zhu [*],[1]

Department of Chemistry, Tongji University, Siping Road 1239#, Shanghai 200092, PR China

ARTICLE INFO

Article history:
Received 8 April 2013
Received in revised form
24 May 2013
Accepted 30 May 2013
Available online 19 June 2013

Keywords:
Raman spectroscopy
Rice
Multivariate data analysis
Geographical origin
Adulteration

ABSTRACTS

Rice has played an important role in staple food supply of over approximately one-half of the world population. In this study, Raman spectroscopy and several multivariate data analysis methods were applied for discrimination of rice samples from different districts of China. A total of 42 samples were examined. It is shown that the representative Raman spectra in each group are different according to geographical origin after baseline correction to enhance spectral features. Moreover, adulteration of rice is a serious problem for consumers. In addition to the obvious effect on producer profits, adulteration can also cause severe health and safety problems. Paraffin was added to give the rice a desirable translucent appearance and increase its marketability. Detection of paraffin in the adulterated rice samples was preliminarily investigated as well. The results showed that Raman spectroscopy data with chemometric techniques can be applied to rapid detecting rice adulteration with paraffin.

1. Introduction

Rice (*Oryza sativa* L.) is one of the leading food crops of the world and is a staple food of over approximately one-half of the world population [1]. The rice produced in different areas varies significantly in composition. One of the major problems in the agricultural-food industry is to set down objective tools to determine the origin of raw materials as well as finished products to ensure their traceability [2]. Previous research to determine the geographical origins of diverse agricultural products has focused especially on the utilization of near-infrared (NIR) spectroscopy [3–11]. Although the NIR spectral features are not highly distinct because of the nature of over tone and combination bands, its speed and simplicity make it the most frequently selected spectroscopic method for the determination of the origins of agricultural products.

Raman spectroscopy is also a non-invasive technique, which allows obtaining detailed chemical information on a sample without a need for labelling. It is a non-destructive technique that yields reliable results for solid and liquid multicomponent samples [12]. As same as NIR spectroscopy, this method usually does not require the dissolution or extraction of the samples being analyzed, substantially simplifying and shortening the analysis. In addition, after recording Raman spectra, one can then perform additional analyses of the same sample using other analytical tools. Another benefit of the method, especially in comparison with IR spectroscopy, is the fact that the presence of water does not hinder the analysis of liquid samples. As early as 1928 Sir C.V. Raman reported the effect that was later named after him [13], but only in the last two decades has Raman spectroscopy developed into a mature technique for chemical analysis in a broad range of application areas. In industrial applications, its high capacity for automation is also very important. As such, Raman spectroscopy is a convenient tool for the quantitative analysis and is applied in biochemistry, forensic science [14], pharmaceuticals [15,16] and food products analysis [17–19].

Raman spectroscopy is capable of collecting data quickly and providing more selective spectral features than NIR spectroscopy; however, it has not been as commonly utilized. Only a few researches describing the use of Raman spectroscopy to determine the origins of agricultural products have been reported [8,20–22].

One of the reasons for this is the fact that during the application of Raman spectroscopy, fluorescence of organic compounds in the samples, which sometimes several orders of magnitude more intense than the weak Raman scatter [23], interfere with the Raman signals. A phenomenon of baseline drift shows up, and the resolution and analysis of Raman spectra become impractical consequently.

Both instrumental [24] and mathematical methods are developed to reduce the fluorescence interference. The use of an

[*] Corresponding author. Tel.: +86 216 598 2653 8211; fax: +86 216 598 2670.
E-mail address: zhuzl@tongji.edu.cn (Z. Zhu).

[1] Present address: Department of Chemistry, Tongji University, Siping Road 1239#, Shanghai 200092, PR China.

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.05.072

excitation wavelength that does not stimulate fluorescence, such as 785 nm lasers, is the most traditional instrumental methods. Utilization of anti-Stokes Raman spectroscopy is another way [25]. Mathematical methods include first and second order derivatives [26], wavelet transform [27], and manual polynomial fitting. They are useful in certain situation, but still with some limitations. For example, Manual polynomial fitting needs the user identify the "non-Raman" locations manually [23], after which the baseline curve is formed by fitting these locations. Consequently, the result involves the subjective factors inevitably. Herein, a modified polynomial fitting method proposed in previous work was adopted [28]. The baseline was automatically estimated and subtracted, leaving only the Raman peaks.

Adulteration of food products involves the replacement of high-cost ingredients with lower grade and cheaper substitutes [29,30]. Adulteration of rice is a serious problem for regulatory agencies, suppliers and consumers [31–33]. Adulteration of rice with additives has become a problem for Chinese rice producers and consumers. Especially paraffin has been used to adulterate rice, reported as "Toxic rice" [34]. It is the rice that has been treated with chemicals and coated with paraffin to increase its marketability. This gives the rice a desirable translucent appearance. Unfortunately, the treated rice which contains carcinogens causes negative health effects. In some cases, paraffin is regarded as mineral oil saturated hydrocarbons (MOSH) contamination in foodstuffs. Such contamination is mainly related to migration of printing inks applied directly to the packaging surface or the inks present in newspapers or other types of paper used to generate recycled fiber, which is then used for production of cardboard. Multidimensional liquid–gas chromatography (LC–GC) systems were used by Purcaro et al. [35] to determine the mineral oil contamination of a series of food products (rice, pasta, icing sugar, olive oil). Nevertheless, no previous work and basic information specifically concerned the rice adulteration with paraffin has been reported.

Previous works on determinations of the geographical origins of rice grains normally rely on the differences of the trace elements. For example, Gonzalvez et al. [36] determine the concentrations of 32 elements (Al, As, Ba, Bi, etc.) with ICP-OES to achieve geographical origin classification of samples from different countries including Spain, Japan, Brazil and India. Li et al. [37] has established a methodology to determine the geographical origin of rice from Fujian, China with element profiles (Ca, K, Mg, P, etc.). These works have proved that the concentrations of elements are different according to their origins, which could be the basis of geographical classification.

In this research, we focused on the main chemical compositions on the surface of rice samples and proposed a Raman scheme with the goal of rapid classification rice of several different geographical origins. The rice samples used were de-husked grains that were not further milled. Raman spectra of all these grains were directly collected by confocal microscope Raman spectrometer. Before further spectral processing, all the spectra were baseline corrected and normalized. The Raman spectral features of several rice types originating from different locations were very similar. Only a minor spectral difference could be recognized for differentiation. Therefore, PCA [38] was used to convert the resulting Raman spectral features into fewer dimensions of scores, and visually display the differentiation. And other multivariate data analysis techniques were further used to solve the problem, including Soft independent modeling of class analogy (SIMCA), Partial least squares-discriminant analysis (PLS-DA), K-nearest neighbors (KNN) technique, and Support vector machines (SVM) [39]. The geographical origins of rice were preliminarily identified by the method described above. Additionally, another aim of this study is to demonstrate the applicability of Raman spectroscopy as a rapid analysis method to detect paraffin in the adulterated rice samples.

2. Materials and methods

2.1. Rice materials

A total of 42 samples, including 20 indica and 22 japonica rice were collected. The samples composed of different species and were cultivated in diverse geographical regions of China (Table 1). The raw materials were washed thoroughly with deionized water re-purified by Millipore Synergy Ultrapure Water Systems (Millipore Co., Ltd., USA).

2.2. Adulterated rice samples

Samples of BY were selected as unadulterated rice samples. And parts of them were adulterated with different quantities of paraffin. Paraffin was obtained from Merck Company Inc. (Darmstadt, Germany). A set of seven adulterated samples in the range between 0.003% and 0.166% (w/w) (0.003%, 0.008%, 0.017%, 0.033%, 0.040%, 0.066% and 0.166%) were prepared for each adulterant. These ranges were chosen to demonstrate the adulteration detection limit in rice adulteration studies. Samples were mixed well and kept at room temperature to equilibrate before Raman measurements.

2.3. Spectrum collection

Raman spectra of samples were recorded at room temperature under ambient conditions using a Renishaw Raman InVia spectrometer (Renishaw plc., Wotton–under–Edge, UK) equipped with an air-cooled charge-coupled device (CCD) detector. And the monochromatic light source was a 514 nm Argon ion laser. The laser was focused on the solid samples which were placed on microscope slides. And the power was kept at about 20 mW to prevent irreversible thermal damage to the samples. Raman spectra with a resolution of 1.5 cm^{-1} were collected with an exposure time of 20 s and 2 scans. The Raman spectra of each sample were plotted as relative intensity (arbitrary units) against Raman shift in wavenumber (cm^{-1}). All analyses were performed in triplicates.

2.4. Baseline removal based on modified polynomial fitted method

As mentioned in Ref. [28], the background elimination algorithm could be described as: The original spectrum S ($m \times 1$) was taken as initial values of y_0 which to be polynomial fitted. The fitted curve, marked as y_n, was compared to y_0 one by one. If anyone of y_n greater, they were reassigned to the corresponding values of y_0. The changed y_n would be polynomial fitted again. This

Table 1
The origins of 42 rice grains samples.

Species	Geographical origins	Sample number	Label
Indica	Sichuan	5	SC
	Guizhou	5	GZ
	Fujian	6	FJ
	Hainan	4	HN
Japonica	Baoyin, Jiangsu	4	BY
	Xinghua, Jiangsu	4	XH
	Haifeng, Jiangsu	4	HF
	Sheyang, Jiangsu	3	SY
	Wuchang, Heilongjiang	3	WC
	Anqing, Heilongjiang	4	AQ

iteration was repeated until y_0 and y_n converge, which considered as baseline. The estimated baseline was then subtracted from the original spectrum to yield corrected spectrum. The details of this algorithm can be referenced in Ref. [28].

2.5. Raman spectra pre-processing

Besides the baseline correction pre-processing step introduced in Section 2.4, other mathematical pre-treatment should be performed to smooth out the inherent noise associated with instrumental or sampling variability. Before the further model building, different kinds of pre-processing methods such as mean centering (MC), mean scattering correction (MSC), etc. have been applied to the data. For the purposes of this analysis, a Savitsky–Golay smoothing filter [40] using a 20 points smoothing window and a second order polynomial deconvolution (SGD2) was applied to the data, which resulted in no distortion of the spectral profiles. The spectra were then normalized using the method of Standard Normal Variate (SNV) proposed by Barnes et al. [41]. SNV is a scatter correction treatment which centers each spectrum at zero and scales them by the standard deviation of the spectral data. Thus, pre-treatment procedure (SGD2-SNV) which produced the best results was selected throughout the study.

2.6. Multivariate data analysis

2.6.1. Principal component analysis

PCA is a way of identifying patterns in data, and expressing the data in such a way as to emphasize their similarities and differences. It can compress the data, that is, by reducing the number of dimensions without much loss of information based on their similarities and differences, and define a limited number of “principal components” which describe independent variation structures in the data. When more than three variables have been measured, visualization of the data by various plotting systems is then possible [42]. Therefore, PCA can indicate relationships among groups of variables in a data set and show relationships that might exist between objects.

Other four different multivariate data analysis methods including SIMCA, PLS-DA, KNN and SVM are adopted in this study for further analysis. Some basic remarks about these methods are given as follows. Detailed information can be referenced in Ref. [39,43].

2.6.2. Soft independent modeling of class analogy

SIMCA requires a data set consisting of samples with a set of attributes and their class membership. Herein “soft” refers to that the classifier can identify samples as belonging to multiple classes [44]. To build the classification models, the samples in each class need to be analyzed using PCA. A hyperplane is described in a given class. The mean orthogonal distance of data samples from the hyperplane is used to determine a critical distance for classification. A new observation is assigned to the model class when its residual distance from the model is below the statistical limit for the class. The observation may be found to belong to multiple classes and a measure of goodness of the model can be found from the number of cases where the observations are classified into multiple classes.

2.6.3. Partial least squares-discriminant analysis

PLS-DA is an extension of the multiple linear regression model and PCA method. PLS is normally used for calibration model building, but it can also be applied for classification [45]. For PLS-DA classification, the class (one of M) of each sample is coded with a binary vector of length M with zeros and 1 one. The M predicted scores by PLS-DA method are used to predict the class of each sample. If the normalized $j_{\text{-score}}$ is greater than or equal to parameter δ, the predicted class of the sample was j. More information about PLS-DA classification can be found in Ref. [46].

2.6.4. K-nearest neighbors

KNN proposed by Fix and Hodges [47] include the distances calculated between all data points. Then, K-closest neighbors are found by sorting the distance matrix. The K-closest data points are analyzed to determine which class label is the most common among the set. KNN has good performance in dealing with multi-class problem.

2.6.5. Support vector machines

SVM have been developed by Vapnik [48]. Based on “beautifully simple ideas” [49], SVM correspond to a linear method in a very high dimensional feature space that is nonlinearly related to the input space. It does not involve any calculations in that high dimensional space. By the use of kernels (radial basis function kernel used here), all necessary computations are performed directly in the input space [50]. Support vector machines map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data and maximizes the distance between the two parallel hyperplanes.

2.6.6. Model efficiency estimation

Model efficiency was defined with error of cross validation:

$$E = \frac{N_{\text{wrong}}}{N} = \frac{N - N_{\text{right}}}{N} \quad (1)$$

where N_{wrong} and N_{right} are the numbers of wrongly and rightly classified samples; N is the total number of samples. The value E is a measure of model sensitivity which stands for the ability to classify in a correct manner the objects belonging to the class [43]. In this study, leave-one-out cross validation was used to evaluate model's efficiency.

2.6.7. Model optimization

To compare different classification methods, the efficiency of the best possible model should be found. The result of model usage depends on the parameters, and the following parameters were optimized to achieve the best results.

SIMCA: number of principal components
PLS-DA: number of latent variables and threshold parameter
KNN: number of neighbors
SVM: different types of kernel functions.

All calculations and multivariate analysis, including baseline correction, data pretreatment were conducted using MATLAB version 7.6 software (The Math-Works Inc., MA, USA).

3. Result and discussion

3.1. Baseline correction

Before examining spectral features or performing further spectral processing, all the spectra were baseline corrected and pre-processed according to Sections 2.4 and 2.5. Fig. 1A shows the 6 raw Raman spectra of the samples (selected from the Japonica rice samples). The most noticeable feature is the baseline variation among spectra. As there was nonlinear background superimposed in the sample spectra caused by fluorescence, the estimated

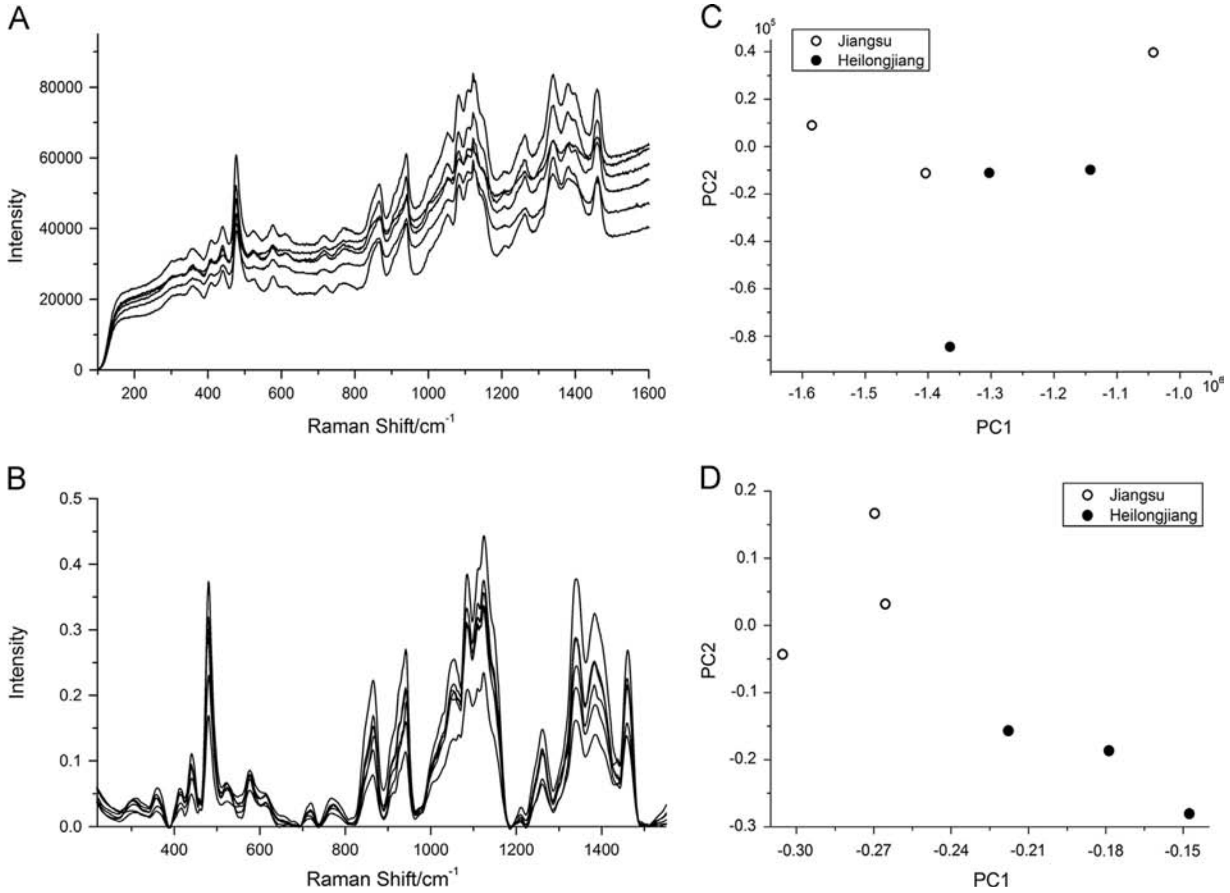

Fig. 1. Original (A) and corrected (B) Raman spectra of 6 rice samples from Jiangsu and Heilongjiang, and score scatter plots generated using original (C) and corrected (D) Raman spectra of 6 samples.

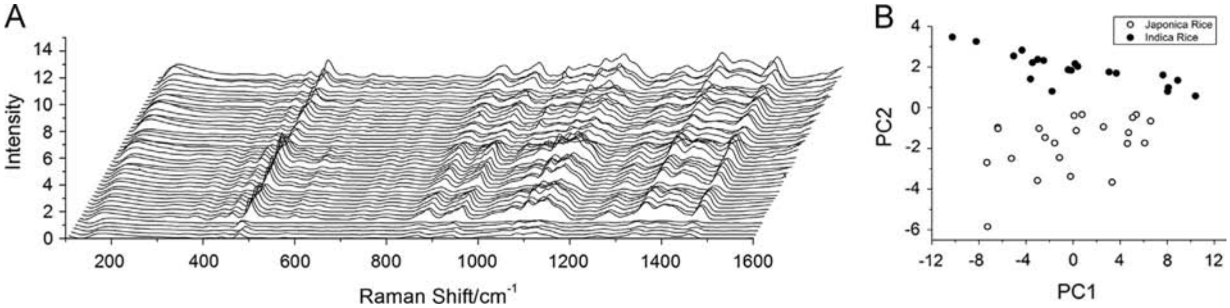

Fig. 2. Corrected Raman spectra (A) and their score scatter plot (B) of all 42 rice samples.

baselines of spectra were fitted by modified least-squared method, and then be subtracted from the original spectrum to yield corrected spectrum. Since the surface of the rice grain is round and irregular, the location of a grain with respect to the focal point will not be consistent in the measurements. This also leads to considerable baseline variation. The peak area under the 1600–100 cm^{-1} range was calculated. Then, each baseline-corrected spectrum was divided by the corresponding peak area under the 1600–100 cm^{-1} range. Through normalization, the possible Raman peak intensity variation resulting from a change in sample position with respect to the focal point could be greatly reduced. The resulting normalized spectra are shown in Fig. 1B.

Because the existence of fluorescence, some Raman peaks which might stand for characteristic vibrations were covered. Samples cannot be distinguished visually from the spectra. PCA was applied to both the original and corrected spectra of rice grains for classifications. The rice grains are not classified before baselines corrected, samples are close along the second principal component (PC2) axis (Fig. 1C). The first two principal components might just represent the mutual characteristics of fluorescence. After baseline corrected, samples were clearly classified to two groups, which refer to Jiangsu and Heilongjiang province (Fig. 1D). The influence of baseline drift has been eliminated effectively. Corrected Raman spectra of all rice samples are shown in Fig. 2A.

3.2. The main characteristic bands and their assignments

Rice is composed of starch, protein, and lipids as well as other types of nutrition. Primary rice Raman bands are observed at 476, 856, 941, 1004, 1032, 1082, 1155, 1204, 1253, 1342, 1391, 1458, and 1551 cm^{-1}. Each rice type was classified and its various spectral components were associated with vibrations, rotations, etc., of the chemical bonds in the rice's constituent components. Raman bands have been attributed to biological compounds such as starch, protein, and lipids as well as to chemical bonds in these compounds such as, C–H bending (1458 cm^{-1}), C–O–H bending, CH_2 twisting (1440–1320 cm^{-1}), C–O–H stretching, and –OH twisting (1200–1000 cm^{-1}). Raman bands observed below 600 cm^{-1} were associated with the backbone of the ring composition and torsional vibration modes [51].

Strong and clearly resolved Raman signals appear at 1155 cm^{-1} and correspond to carbon–carbon single-bond stretch vibrations of the conjugated backbone [52]. Raman bands of rice result primarily from starch vibrations, such as amylose at ~850 and 1253 cm^{-1}, amylopectin at ~905 and 1391 cm^{-1}, CH_2 twisting and wagging vibration at 1342 and 1314 cm^{-1}, CH_2 (or CH_3) deformation vibration at 1452 cm^{-1}. A series of bands are also found at 941, 1037, 1082, and 1132 cm^{-1}. According to the literature [51], those bands may be from the characteristic vibration of cyclohexaamylose. Furthermore, the other bands, such as 1360, 1032, 1004 cm^{-1}, were attributed to certain vibration of side chain of proteins. For example, the peaks at 1360 and 1032 cm^{-1} are attributed to vibration of tryptophan and praline, respectively [53]. The peak at 1004 cm^{-1} is attributed to ring-breathing vibration of phenylalanine molecules. Changes in the relative intensities, position, and width of these bands are related to the composition of the sample. The main Raman bands and their preliminary assignments are listed in Table 2.

3.3. Indica and japonica rice

Oryza sativa contains two major subspecies [54]: the sticky, short grained japonica or sinica variety, and the non-sticky, long-grained indica variety. Japonica are usually cultivated in dry fields, in temperate East Asia, upland areas of Southeast Asia and high elevations in South Asia, while indica are mainly lowland rices, grown mostly submerged, throughout tropical Asia [55].

To better interpret the data obtained, the baseline corrected Raman spectra of 42 samples were analyzed by PCA, to find a possible separation between the indica and japonica rice.

Applying PCA to the data matrix, two components were extracted, representing jointly 97.1% of the whole system variance. The PCA results of the all 42 Raman spectra acquired from rice samples is shown in Table 3. And the factor indicator (IND) and Fisher variance ratio (*F*-ratio) proposed by Malinowski [56] were also calculated besides the Eigen values and residual standard deviation (RSD). It is confirmed that the first two PCs could represent most information of the system. In Fig. 2B, the scatter plot of the scores of PC1 versus PC2 is reported. The score plot shows that there is not a clear separation between indica rice and japonica rice, but it is possible to observe a tendency of solid stamp samples to cluster in the area upper of the score plot, while hollow stamp samples seem to cluster in the lower side of the score plot. So PC2 (14.8% of variance) is more important than PC1 (82.3% of variance) for the visually discrimination according to species.

Table 2
The approximate peak positions and their tentative assignments.

Approximate peak positions (cm^{-1})	Assignments
1458	C–H bending
1452	CH_2(or CH_3) deformation
1440–1320	C–O–H bending
1391, 905	Amylopectin
1360	Tryptophan
1342	CH_2 twisting
1314	CH_2 wagging
1253, 850	Amylose
1200–1000	C–O–H stretching
1155	C–C stretching
1132, 1082, 1037, 941	Cyclohexaamylose
1032	Praline
1004	Phenylalanine

Table 3
The PCA results of the Raman spectra of 42 rice samples.

No.	Eigen value	Ratio	RSD	IND/10^{-6}	*F*-ratio
1	93.15	6.40	0.08306	47.09	1368.483
2	14.55	3.23	0.03807	22.65	158.914
3	4.504	1.41	0.03056	19.10	23.640
4	3.194	1.15	0.02594	17.06	16.493
5	2.787	1.28	0.02163	14.99	18.01

PCA loading data indicated that the Raman bands that determine the scores on PC2 correspond to the wave number of 941, 1037, 1082, 1132, 1004, 1244, 1360 and 1452 cm^{-1}. The 941, 1037, 1082 and 1132 cm^{-1} bands contributing to the positive value of PC2, are characteristics of indica rice. These bands stand for cyclohexaamylose according to the assignments in 3.2. It is indicated that the quantity of amylose are different between two varieties [57]. And the other bands contributing to the negative value of PC2 are characteristics of japonica rice. Most were attributed to side chain of proteins. Thus, different quantity of amylose and side chain proteins might be the basis of discrimination of indica and japonica rice.

3.4. Geographical classification of indica and japonica rice respectively with PCA

To evaluate the possibility of differentiating the samples taking into account different geographical regions, we applied PCA for the Raman spectra of indica and japonica rice, respectively.

As for indica rice, the application of the PCA showed four distinct groups (Fig. 3A). The first group was composed by indica rice from SC province. The second one was collected from FJ and located on the top center of the scores-plot. The third group was composed by four samples harvested from HN areas. Finally, the fourth group was found in a different quadrant of the plot which included five samples collected from GZ. These conclusions are in good agreement with the results presented in Table 1.

Fig. 3B shows PCA plots of japonica rice data. Based on elaborated methods it was possible to differentiate between samples. All rice samples of Heilongjiang formed one cluster distinct from other samples. Those samples exhibited negative scores according to both PC1 and PC2. All the other samples of Jiangsu province centered more towards the negative PC1 axis. However, there is less distinction between groups of same province. It was unable to identify the rice's geographical regions in the further level.

3.5. Geographical classification with other multivariate statistical analysis methods

In Section 3.4, PCA is operated on the Raman spectra of indica and japonica rice, respectively. The availability of classification and group tendency according geographical origins is visually observed on the PCA plot in Fig.3. To establish a robust model, more sophisticated classification methods are utilized to handle the Raman spectral data set. Four advanced multivariate data analysis methods including SIMCA, PLS-DA, KNN and SVM are adopted here for further analysis.

Table 4 and Table 5 summarize the classification results of indica and japonica rice with four methods, respectively. They clearly show the non-equivalency of the classes in terms of classification accuracy. For indica rice samples from SC and GZ, 100% effectiveness in classification was reached using any one of

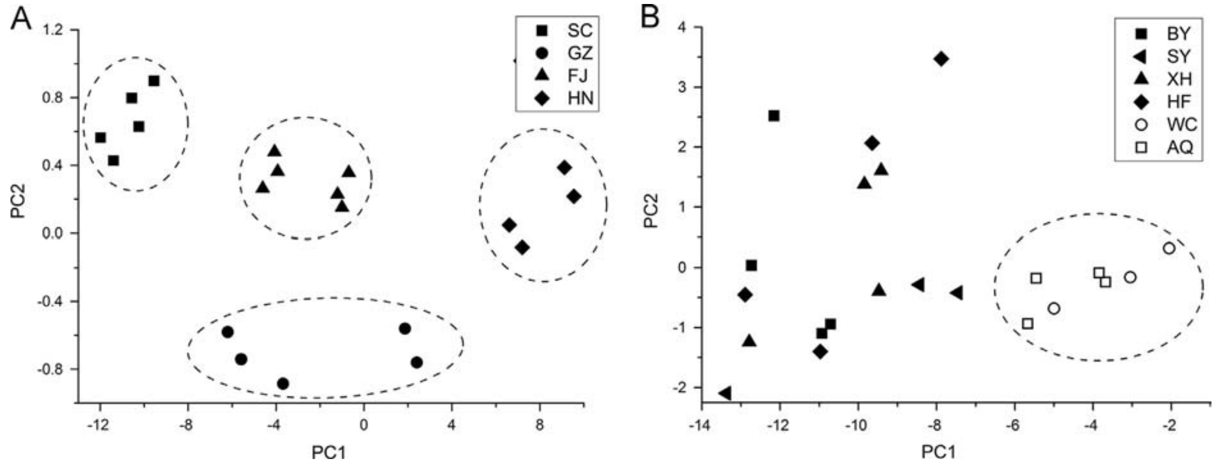

Fig. 3. Score scatter plots generated using Raman spectra of indica rice samples (A) and japonica rice samples (B).

Table 4
Classification of 20 indica rice grains samples with different multivariate methods.

	SIMCA		PLS-DA		KNN		SVM	
	N_{right}/N	E (%)	N_{right}/N	E (%)	N_{right}/N	E (%)	N_{right}/N	E (%)
SC	5/5	0	5/5	0	5/5	0	5/5	0
GZ	5/5	0	5/5	0	5/5	0	5/5	0
FJ	4/6	33	4/6	33	6/6	0	6/6	0
HN	4/4	0	4/4	0	3/4	25	4/4	0
Total	18/20	10	18/20	10	19/20	5	20/20	0

Table 5
Classification of 22 japonica rice grains samples with different multivariate methods.

	SIMCA		PLS-DA		KNN		SVM	
	N_{right}/N	E (%)	N_{right}/N	E (%)	N_{right}/N	E (%)	N_{right}/N	E (%)
Jiangsu	13/15	13	14/15	7	15/15	0	15/15	0
Heilongjiang	7/7	0	7/7	0	7/7	0	7/7	0
Total	20/22	9	21/22	5	22/22	0	22/22	0

the methods. But SIMCA and PLS-DA have some difficulties on discriminating FJ samples, thus inaccuracy of 33% are observed with both methods, which is also the largest inaccuracy value in investigations of indica rice samples. Rice samples from these two groups are difficult to label correctly only using SIMCA and PLS-DA. Same phenomenon is showed in dealing with japonica rice, and samples from Jiangsu are misclassified according to Table 5. KNN classification model is a simple but highly effective classification algorithm. Only one classification error is observed from HN samples in Table 4. The general accuracy of a classification model seems to be greatly dependent on its ability to correctly distinguish most of the rice samples. The KNN and SVM methods are more successful in two cases. Especially, support vector machines can be regarded as the most effective classification model for rice samples. Accurate and robust classification models can be built using SVM approach.

Comparison of the effectiveness of SIMCA, PLS-DA, KNN, and SVM for different indica and japonica rice samples is presented in Fig. 4. The general trend that relates classification model order with respect to accuracy is the same in both cases: SIMCA < PLS-DA < KNN < SVM. In these two cases, the classification accuracy is all above 90%. SIMCA and PLS-DA methods yield approximately the same error of classification, which is relatively larger compared to KNN and SVM. SVM-based approach shows significant superiority over others. According to the conclusions reported by Balabin et al. [43], the choice between KNN and SVM methods is also dependent on the computational resources available. While the KNN model is easy to build, SVM is rather computationally demanding in terms of CPU time, and the computer resources needed.

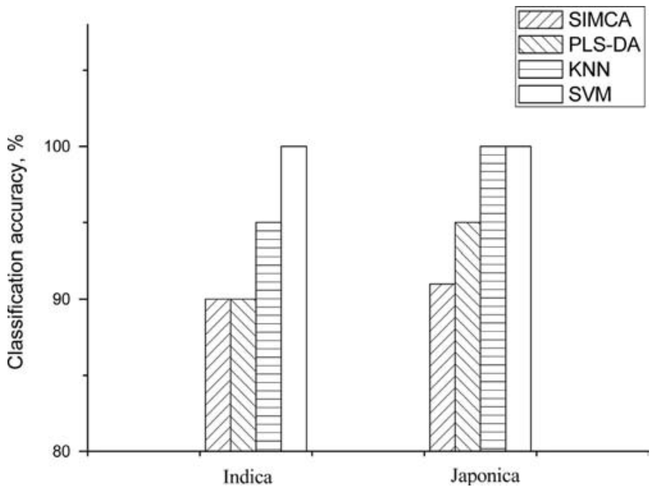

Fig. 4. Comparison of the efficiency of classification models for indica and japonica rice sample sets.

3.6. *Detection of paraffin in the adulterated rice*

Liquid paraffin wax or products derived from petroleum distillation, is a mineral oil. High purity products can be used in medicine and cosmetics, products that contain low-level impurities are harmful if mixed with food. Adding liquid paraffin or other mineral oil mixture make the surface of stale rice smooth and beautiful, but it masks mycotoxins that may exist in stale rice. To the knowledge of the authors, no published work has focused on the detection of paraffin in adulterated rice with rapid and non-destructive method.

Strong Raman peaks were observed at 1062, 1132, 1295, 1417, 1440 and 1462 cm^{-1} in the spectrum of paraffin sample (Fig. 5A). Although some of the bands were also appeared in the Raman spectrum of rice sample, due to the sharp shape and great intensities, these could be considered as characteristics to distinct

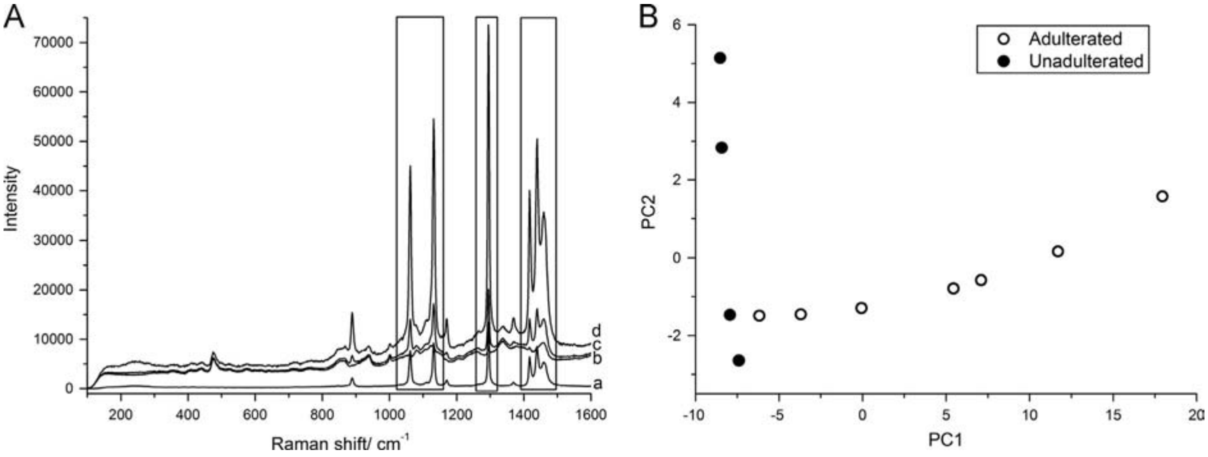

Fig. 5. Original Raman spectra (A) and their score scatter plot (B) of paraffin and adulterated rice samples. (a) Pure paraffin (b) 0.008% adulterated samples (c) 0.040% adulterated samples (d) 0.166% adulterated samples.

the adulterated and unadulterated samples. In fact, the determination of adulteration could be done directly on the spectra by observation.

PCA was performed adulterated group with pure rice samples. Pure rice samples include four samples from BY region. Scores plot (Fig. 5B) revealed that 0.003% adulterated samples of rice is plotted too close to pure rice samples which hinder the discrimination. It is consistent with the direct comparison of two Raman spectra. Besides, each sample point represents different adulteration percentages are in ascending order from left to right. Score plot reveals the discrimination of adulterated samples from non-adulterated rice samples. Consequently, the method constructed here can significantly detect the paraffin in rice samples and discriminate adulterated rice samples.

4. Conclusions

The results described in this research open the possibility of discriminating rice according to their geographic origin using Raman spectroscopy combined with pattern recognition methods, such as PCA, SIMCA, PLS-DA, KNN and SVM. Practically, direct Raman scanning of rice samples without grinding, which makes this measurement much more convenient, is easy and possible. Without baseline correction of the spectra, the rapid identification of origins of rice was impossible or only limited to human eye examination that is lack of objectiveness. This study shows that, the combination of experimental data along with chemometric approaches can be successfully employed by researchers to determine the geographic origin of rice. It might be an application for quality control, process control and quick classification of rice in the industry. In addition, Raman spectroscopy combined with PCA was preliminarily investigated in identification of unadulterated and adulterated rice with paraffin. It indicates that adulterated rice sample could be satisfactorily discriminated from unadulterated rice sample. Therefore, Raman technique is suitable for determining geographical origins and authenticity of rice with non-destructive and cost-efficient characteristics, especially as a fast screening tool for rice producer and regulatory authorities.

References

[1] N. Singh, L. Kaur, N. Singh Sodhi, K.Singh Sekhon, Food Chem. 89 (2005) 253–259.
[2] O. Galtier, N. Dupuy, Y. Le Dréau, D. Ollivier, C. Pinatel, J. Kister, J. Artaud, Anal. Chim. Acta 595 (2007) 136–144.
[3] H.Y. Yu, Y. Zhou, X.P. Fu, L.J. Xie, Y.B. Ying, Eur. Food Res. Technol. 225 (2007) 313–320.
[4] T. Woodcock, G. Downey, J.D. Kelly, C. O'Donnell, J. Agric. Food Chem. 55 (2007) 9128–9134.
[5] X.P. Fu, Y.B. Ying, Y. Zhou, H.R. Xu, Anal. Chim. Acta 598 (2007) 27–33.
[6] L. Liu, D. Cozzolino, W.U. Cynkar, M. Gishen, C.B. Colby, J. Agric. Food Chem. 54 (2006) 6754–6759.
[7] R. Karoui, A.M. Mouazen, E. Dufour, L. Pillonel, E. Schaller, J. De Baerdemaeker, J.O. Bosset, Int. Dairy J. 16 (2006) 1211–1217.
[8] Y.A. Woo, H.J. Kim, H. Chung, Analyst 124 (1999) 1223–1226.
[9] M. Casale, N. Sinelli, P. Oliveri, V. Di Egidio, S. Lanteri, Talanta 80 (2010) 1832–1837.
[10] M.A. Al-Ghouti, Y.S. Al-Degs, M. Amer, Talanta 76 (2008) 1105–1112.
[11] R. Balabin, R. Safieva, J. Near Infrared Spectrosc. 15 (2007) 343–349.
[12] S. Mazurek, R. Szostak, Food Chem. 125 (2011) 1051–1057.
[13] E.V. Efremov, F. Ariese, C. Gooijer, Anal. Chim. Acta 606 (2008) 119–134.
[14] A.G. Ryder, G.M. O'Connor, T.J. Glynn, J. Raman Spectrosc. 31 (2000) 221–227.
[15] T.R.M. De Beer, W.R.G. Baeyens, A. Vermeire, D. Broes, J.P. Remon, C. Vervaet, Anal. Chim. Acta 589 (2007) 192–199.
[16] S. Mazurek, R. Szostak, J. Pharm. Biomed. Anal. 49 (2009) 168–172.
[17] S. Armenta, S. Garrigues, M. de la Guardia, Anal. Chim. Acta 521 (2004) 149–155.
[18] N. Peica, J. Raman Spectrosc. 40 (2009) 2144–2154.
[19] E. Guzmán, V. Baeten, J.A.F. Pierna, J.A. García-Mesa, Talanta 93 (2012) 94–98.
[20] A.B. Rubayiza, M. Meurens, J. Agric. Food Chem. 53 (2005) 4654–4659.
[21] B. Muik, B. Lendl, A. Molina-Diaz, D. Ortega-Calderon, M.J. Ayora-Canada, J. Agric. Food Chem. 52 (2004) 6055–6060.
[22] E.C. Lopez-Diez, G. Bianchi, R. Goodacre, J. Agric. Food Chem. 51 (2003) 6145–6150.
[23] C.A. Lieber, A. Mahadevan-Jansen, Appl. Spectrosc. 57 (2003) 1363–1367.
[24] J. Funfschilling, D.F. Williams, Appl. Spectrosc. 30 (1976) 443–446.
[25] P.A. Mosier-Boss, S.H. Lieberman, R. Newbery, Appl. Spectrosc. 49 (1995) 630–638.
[26] A. O'Grady, A.C. Dennis, D. Denvir, J.J. McGarvey, S.E.J. Bell, Anal. Chem. 73 (2001) 2058–2065.
[27] V.J. Ashish, et al., Meas. Sci. Technol. 19 (2008) 065102.
[28] X.W. Feng, Z.L. Zhu, M.J. Shen, P.S. Cong, Comput. Appl. Chem. 26 (2009) 759–762.
[29] A. Tay, R.K. Singh, S.S. Krishnan, J.P. Gore, Lebensmittel-Wissenschaft und-Technologie 35 (2002) 99–103.
[30] F.C.C. Oliveira, C.R.R. Brandão, H.F. Ramalho, L.A.F. da Costa, P.A.Z. Suarez, J.C. Rubim, Anal. Chim. Acta 587 (2007) 194–199.
[31] E. Christopoulou, M. Lazaraki, M. Komaitis, K. Kaselimis, Food Chem. 84 (2004) 463–474.
[32] L. Chen, X. Xue, Z. Ye, J. Zhou, F. Chen, J. Zhao, Food Chem. 128 (2011) 1110–1114. (URL:http://dx.doi.org/10.1016/j.foodchem.2010.10.027).
[33] V. Morales, N. Corzo, M.L. Sanz, Food Chem. 107 (2008) 922–928.
[34] ⟨http://chinadigitaltimes.net/space/Poison_rice⟩.
[35] G. Purcaro, M. Zoccali, P.Q. Tranchida, L. Barp, S. Moret, L. Conte, P. Dugo, L. Mondello, Anal. Bioanal. Chem. 405 (2013) 1077–1084.
[36] A. Gonzalvez, S. Armenta, M. de la Guardia, Food Chem. 126 (2011) 1254–1260.
[37] G. Li, L. Nunes, Y.J. Wang, P.N. Williams, M.Z. Zheng, Q.F. Zhang, Y.G. Zhu, J. Environ. Sci. 25 (2013) 144–154.
[38] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical GuideWiley-Interscience, New York, 1998.
[39] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Anal. Chim. Acta 671 (2010) 27–35.
[40] A. Savitzky, M.J. Golay, Anal. Chem. 36 (1964) 1627–1639.
[41] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.
[42] A. Kamal-Eldin, R. Andersson, J. Am. Oil Chem. Soc. 74 (1997) 375–380.
[43] R.M. Balabin, R.Z. Safieva, Anal. Chim. Acta 689 (2011) 190–197.

[44] S. Wold, M. SjÖStrÖM, SIMCA: A method for analyzing chemical data in terms of similarity and analogy, in: Chemometrics: Theory and Application, American Chemical Society, 1977, pp. 243–282.

[45] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Chemom. Intell. Lab. Syst. 88 (2007) 183–188.

[46] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1–17.

[47] E. Fix, J.L. Hodges Jr., Int. Stat. Rev./Revue Internationale de Statistique 57 (1989) 238–247.

[48] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1999.

[49] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, Chemom. Intell. Lab. Syst. 69 (2003) 13–20.

[50] R.M. Balabin, S.V. Smirnov, Talanta 85 (2011) 562–568.

[51] F.S. Parker, Applications of Infrared, Raman, and Resonance Raman Spectroscopy in Biochemistry, Plenum Press, New York, 1983.

[52] M.G. Shim, B.C. Wilson, Photochem. Photobiol. 63 (1996) 662–671.

[53] C.J. Frank, R.L. McCreery, D.C.B. Redd, Anal. Chem. 67 (1995) 777–783.

[54] H. Ikehashi, Rice Sci. 16 (2009) 1–13.

[55] H.J. Kang, I.K. Hwang, K.S. Kim, H.C. Choi, J. Agric. Food Chem. 54 (2006) 4833–4838.

[56] E.R. Malinowski, J. Chemom. 3 (1989) 49–60.

[57] Y. Takemoto-Kuno, K. Suzuki, S. Nakamura, H. Satoh, K. Ohtsubo, J. Agric. Food Chem. 54 (2006) 9234–9240.