Near-infrared-based Identification of Sesame Oil Authenticity

Qingsong Luo, Yang Chen, Qiang Xu, Yaru Yu and Xiao Zheng*

School of Mechanical Engineering, Wuhan Polytechnic University

*Corresponding author e-mail: xnsjd85773@163.com

Abstract. Near-infrared spectroscopy (NIRS) combined with chemometrics were employed to determine the quality of sesame oil. Sesame oil authenticity identification and content determination models with two variables (binary systems) were established. 52 sesame oil samples were mixed with soybean and rapeseed oil for the preparation of fake sesame oil samples. NIRS, in association with the support vector machine classification (SVC), was used to establish full-range-wavelength models. The competitive adaptive reweighted sampling (CARS) and backward interval partial least squares (BIPLS) methods were adopted in the optimization process for characteristic wavelengths for the modeling. The results showed that all the models established could effectively identify the authenticity of sesame oil. The data were preprocessed with the standard normal variable transformation algorithm (SNV), and the preprocessed data were then used to establish an SNV-SVC model with the prediction set recognition accuracy of up to 99.4975%. The correlation coefficient R of the prediction model for content determination was higher than 99%, and the mean square error (MSE) was lower than 0.0605, indicating that the models based on NIRS and support vector machine regression (SVR) can realize the content determination of fake sesame oil.

1. Introduction

Sesame oil is high-quality and nutrient-rich oil. However, with the adulteration of sesame oil in the market, it is necessary to establish a fast, effective, low-cost detection approach for oils.

Recently, the technologies of chemometrics and spectral analysis have been greatly developed in the field of computer. Especially, near-infrared spectroscopy is one of the most attractive analytical techniques, because it has advantages of short scan time, high efficiency, no chemical treatment and damage for samples, low cost, and little impact on the environment.

2. Experimental

2.1. Materials and samples preparation

Sesame oil, rapeseed oil, and soybean oil with different origins and different brands were purchased. Meanwhile, sesame was purchased to derive 52 sesame oil samples after squeezing. The fake sesame oil samples were prepared by mixing sesame oil with soybean oil or rapeseed oil with different contents. As depicted in Table 1, the contents of soybean oil or rapeseed oil were 2%, 4%,…, and 34% (step = 2% at < 35%); and 35%, 40%…, and 95% (step = 5% at > 35%).
| Oil type                        | Sesame oil | Soybean-oil-doped sesame oil | Soybean oil | Rapeseed-oil-doped sesame oil | Rapeseed oil |
|--------------------------------|------------|-----------------------------|------------|-------------------------------|-------------|
| Number of samples              | 52         | 360                         | 12         | 360                           | 12          |
| Entry                          | 1          | 2                           | 3          | 4                             | 5           |

2.2. Instruments
A home-made laser near-infrared spectral instrument was used for the fast detection of the quality of edible oil. The host of the spectral instrument was an AxsunXL410 laser near-infrared spectrometer with a super luminescent diode as the light source. The sesame oil was squeezed with a Delong-type squeezer (Wuhan City). The impurities in the oil separation were separated with a TG16-II-type medical centrifuge. The samples were prepared with a JA1003N-type electronic balance. The samples were stored in 15-mL centrifuge tubes.

2.3. Spectra acquirement
The samples stored in centrifuge tubes were heated to 60°C in an electronic thermostat water bath. Then, the samples were separately transferred into a cuvette for scanning. Each sample was scanned three times, and the average spectrum was used in the following procedures.

2.4. Preprocessing of spectral data
The original NIR spectra were subject to factors such as temperature and noise. Consequently, the spectra would overlap, the signal to noise ratio would be low, and the information specificity would be low. Therefore, the original NIR spectra should be preprocessed. Using MATLAB, we preprocessed the data with the multiplicative scatter correction (MSC), standard normal variate (SNV), and detrending (DT) methods. Besides, we employed the competitive adaptive reweighted sampling (CARS), backward interval partial least squares (BIPLS), and partial least squares (PLS) methods to extract the spectra at characteristic wavelengths. The modeling was implemented following the support vector machine classification (SVC) and support vector regression (SVR) method.

2.5. Extraction methods for spectra at characteristic wavelengths
Original NIR spectra usually contain overlapped data and noise, which affect the prediction ability of models and should be eliminated. In the present work, the spectra at characteristic wavelengths were extracted through partial least squares (PLS), interval partial least squares (IPLS), backward interval partial least squares (BIPLS), and synergy interval partial least squares (SIPLS).

2.6. Modeling methods
The models for qualitative identification were built by using SVC, which can be divided into linear and nonlinear classification. The former establishes classification functions, for the acquirements of optimal classification hyperplanes and reasonable sample classification. The latter uses kernel functions to map data to high-dimension space, for the acquirements of optimal classification hyperplane and nonlinear classification.

The models for quantitative prediction were built by using SVR, which establishes a loss function. Like SVC, SVR is also divided into linear and nonlinear regression. SVR establishes a relationship with a kernel function. In the modeling process with SVR, an RBF kernel function with fewer parameters should be selected, and the RBF kernel should have great learning ability in practical applications.

In the modeling process with support vector machine (SVM), penalty factor C and kernel function radius g are two important parameters, which have a great impact on the accuracy of a model. Therefore, the optimization of (C, g) is necessary. In this study, cross validation (CV) and genetic
algorithm (GA) were utilized to roughly optimize \((C, g)\). And, more measures would be taken to further optimize \((C, g)\) in the following sections for modeling.

3. Experimental section

3.1. Spectra preprocessing

The averaged NIR spectra of sesame oil samples were separately preprocessed with the MSC and SNV-DT methods, and the preprocessed spectra were shown in Figures 1–2.

![Preprocessed spectra with the MSC method.](image1)

**Figure 1.** Preprocessed spectra with the MSC method.

![Preprocessed spectra with the SNV-DT method.](image2)

**Figure 2.** Preprocessed spectra with the SNV-DT method.

3.2. Spectra extraction at characteristic wavelengths

The spectra at characteristic wavelengths were extracted with the CARS and BIPLS methods, and the spectra in the range of the characteristic wavelengths were selected as the input data for the modeling in the following sections.

3.3. Models for the qualitative authenticity identification

The preprocessed entire spectra and CARS/BiPLS-extracted spectral data were selected as the input data of the SVC models. The parameters \((C, g)\) of the SVC models were optimized through CV. The preprocessed entire spectra and extracted spectra, with optimized parameters, functioned as the
modeling parameters of the SVC models for qualitative identification of the authenticity of sesame oil. Table 2 shows the parameters (C, g) of these preprocessed and extracted spectra and accuracy of models. Figure 3a-b gives the prediction results of calibration and prediction sets of the SNV-CARS-SVC model after the SNV preprocessing and CARS extraction for the identification of the authenticity of sesame oil (binary systems).

Table 2. Parameters and prediction results of the NIRS-SVC models.

| Oil type        | Data processing method | Number of variables | Parameters (C, g) | Accuracy/% |
|-----------------|------------------------|---------------------|------------------|------------|
| Sesame oil      |                         |                     |                  |            |
|                 | MSC                    | 451                 | 1024 32          | 96.4824    | 97.9879    |
|                 | SNV                    | 451                 | 128 16           | 99.665     | 99.4975    |
|                 | SNV-DT                 | 451                 | 512 8           | 99.8325    | 98.995     |
|                 | MSC-CARS               | 56                  | 1024 32          | 98.66      | 97.9899    |
|                 | SNV-CARS               | 74                  | 1024 16          | 98.325     | 98.4925    |
|                 | SNV-DT-CARS            | 83                  | 1024 8           | 97.8224    | 95.9799    |
|                 | MSC-BiPLS              | 45                  | 1024 32          | 86.5997    | 90.9548    |
|                 | SNV-BiPLS              | 45                  | 1024 32          | 91.1223    | 95.9799    |
|                 | SNV-DT-BiPLS           | 315                 | 1024 2           | 96.4824    | 95.9799    |

As shown in Table 2, these identification models based on NIR spectra combined with support vector machine classification can realize the fast qualitative detection of sesame oil. In particular, the SNV-SVC model involving the entire wavelengths had the highest accuracy of up to 99.4975%, with the penalty coefficient C of 128 and kernel function parameter g of 16, indicating that the prediction ability of this model met the requirement.

3.4. Quantitative prediction of the content of impurity in sesame oil

The quantitative prediction herein only involved binary soybean-oil- and rapeseed-oil-doped sesame oil systems. 412 soybean-oil-doped sesame oil samples and 412 rapeseed-oil-doped sesame oil samples were used for modeling. The original NIR spectra were illustrated in Figures 4 and 5.
3.5. Soybean-oil-doped sesame oil samples
The preprocessed NIR spectral data and BiPLS-optimized variables of the soybean-oil-doped sesame oil samples were used to build SVR models. Following the CV method, we optimized the parameters (C, g) in the SVR models. Table 3 shows the prediction results of the SVR models based on the entire and BiPLS-extracted spectra for the content of soybean oil in sesame oil. Figure 6a-b exhibit the prediction results of the calibration and prediction sets of the SNV-DT-BiPLS-SVR model for the content of soybean oil in sesame oil, respectively.

| NIRS data processing method | Number of variables | Parameters (C, g) | Calibration set | Prediction set |
|----------------------------|---------------------|--------------------|-----------------|----------------|
|                            |                     | C                 | g               | R (%)          | MS E          | R (%)          | MS E          |
| MSC                        | 451                 | 1024              | 64              | 99.8547        | 0.0133        | 99.5293        | 0.0418        |
| SNV                        | 451                 | 1024              | 4               | 99.8681        | 0.0121        | 99.5407        | 0.0408        |
| SNV-DT                     | 451                 | 256               | 16              | 99.9674        | 0.0030        | 99.5492        | 0.0400        |
| MSC-BiPLS                  | 159                 | 1024              | 256             | 99.8684        | 0.0121        | 99.3919        | 0.0533        |
| SNV-BiPLS                  | 111                 | 256               | 16              | 99.3271        | 0.0623        | 98.8382        | 0.1017        |
| SNV-DT-BiPLS               | 361                 | 256               | 16              | 99.9398        | 0.0056        | 99.5266        | 0.0421        |
Figure 6. Prediction results of (a) the calibration and (b) prediction sets of the SNV-DT-BiPLS-SVR model for the contents of soybean oil in sesame oil.

Table 3 shows that the NIRS combined with support vector regression could effectively predict the composition of the mixture of sesame oil and soybean oil. The prediction correlation coefficient was ≥ 98.8382%. Among these models, the SNV-DT-BiPLS-SVR model adopted 361 samples as the input variables for the modeling. The modeling was fast. The parameter C of the prediction set was 256 and parameter g was 16, demonstrating the great learning ability of this model. The correlation coefficients R of the calibration and prediction sets were 99.9398% and 99.5266%, respectively. The mean square error MSE was 0.0056 and 0.0421, respectively. In short, this model had small error and a good prediction performance. In the modeling for quantitative prediction of the content of soybean oil in sesame oil, the application of BiPLS to the optimization of the characteristic variables could effectively improve the modeling speed. Meanwhile, the number of SVR parameters (C, g) of the prediction set was reduced, and the learning ability and practicability of the model were improved.

3.6. Rapeseed-oil-doped sesame oil samples
Table 4 shows the prediction results of the SVR models based on the entire and BiPLS-extracted spectra for the content of rapeseed oil in sesame oil. Figure 7a-b exhibit the prediction results of the calibration and prediction sets of the SNV-DT-BiPLS-SVR model for the content of rapeseed oil in sesame oil, respectively.

Table 4. Prediction results of the contents of rapeseed oil in sesame oil.

| NIRS data processing method | Number of variables | Parameters (C, g) | Calibration set | Prediction set |
|----------------------------|---------------------|-------------------|-----------------|----------------|
|                            |                     | C     | g     | R (%) | MSE   | R (%) | MSE   |
| MSC                        | 451                 | 1024  | 64    | 99.8135 | 0.0176 | 99.378 | 0.0507 |
| SNV                        | 451                 | 1024  | 4     | 99.8303 | 0.0160 | 99.3721 | 0.0512 |
| SNV-DT                     | 451                 | 1024  | 4     | 99.7818 | 0.0200 | 99.4549 | 0.0477 |
| MSC-BiPLS                  | 178                 | 64    | 1024  | 99.8767 | 0.0115 | 99.3256 | 0.0558 |
| SNV-BiPLS                  | 178                 | 64    | 64    | 99.8947 | 0.0098 | 99.3041 | 0.0576 |
| SNV-DT-BiPLS               | 111                 | 1024  | 16    | 99.5221 | 0.0443 | 99.3112 | 0.0605 |
As shown in Table 4, the NIRS combined with support vector regression could efficiently predict the composition of the mixture of sesame oil and rapeseed oil. The prediction correlation coefficient was $\geq 99.3041\%$. Among these models, the SNV-DT-SVR full-range-wavelength model showed the best prediction results. The R values of the calibration and prediction sets were 99.7818\% and 99.4549\%, respectively. The MSE values were 0.0200403 and 0.0476783, respectively. The SNV-DT-BiPLS-SVR model adopted 111 characteristic variables, selected by the BiPLS approach, as the input variables. The total number of variables was reduced, and the modeling speed was improved. As shown in Figure 7, compared to those SNV-DT-SVR models which did not adopt characteristic variables, the R values of the SNV-DT-BiPLS-SVR model were slightly decreased, the MES values were slightly increased, and the g values were larger, affecting the generalization ability of this model. In summary, among these quantitative prediction models for the content of rapeseed oil in sesame oil shown in Table 4, the application of BiPLS to the optimization of the characteristic variables could improve the modeling speed, but had a negative impact on the generalization ability of the model.

4. Conclusion
In the identification for the authenticity of sesame oil, the modeling speed is always limited by the large number of samples. In the present work, the spectral data were simplified through the extraction of part of spectra at characteristic wavelengths, and hence the modeling speed was evidently improved. Therefore, it is necessary to extract these spectra at characteristic wavelengths for data with a large number of samples. On the other hand, the full-range-wavelength spectra should be selected for the modeling of data with a small number of samples.

In the qualitative identification, the accuracy of the SVC model was as high as 99.4975\%, providing an effective method for the identification of binary mixture of sesame oil and another type of oil. The prediction performance was dependent on the number of sesame oil samples, indicating that enough samples are very important for the prediction models.

In the quantitative prediction, the contents of soybean oil or rapeseed oil in sesame oil was modeled. The R values were higher than 98\%, and the MSE values were lower than 0.06, indicating that NIRS combined with SVR could provide a reliable method for the quantitative determination of binary edible-oil systems. Although the sesame oil was repeatedly sampled to a great extent, with the same content gradient of impurity (2\% or 5\%) for the prepared samples, the prediction ability of the models did not largely change. Therefore, the content gradient of impurity should be as small as possible to expand the data for modeling.
Acknowledgments
Innovation and Transformation of Grain Science and Technology of Hubei Province (20165104); Key Science and Technology Project of Wuhan City (2013010501010147).

References
[1] Yang Fan, Xue Changyong. Research advancement of nutritional characteristics and functions of common edible oils [J]. Food and Nutrition in China, 2013, 19 (03):63-66.
[2] Cheng Cheng, Gong Kai, Luo can, et al. Edible oil and human health [J]. Education in Chemistry, 2014 (11):81-86.
[3] Zhang Yan. Near infrared spectroscopy for the fast determination and detection of adulteration of edible oils [D]. Henan University of Technology, 2015.
[4] Yao Yunping. The application of oil fingerprint for vegetable oils recognition [D]. Tianjin University of Science and Technology, 2012.
[5] Yang Jia, Wu Yanwen, Li Bingning et al. Fourier Transform Near Infrared Spectroscopy in the Authentication and Adulteration of Sesame Oil [J]. Journal of the Chinese Cereals and Oils Association, 2014, 29 (03):114-119.
[6] Zeng Yan Ming, Du Shuxin. Review of on-line detection of food safety based on spectroscopy analysis [J]. Science and Technology of Food Industry, 2010, 31 (11):389-393.
[7] Ma Wei, Wang Xiu, Li Weiguo. Application of near infrared spectroscopy (NIRs) detection technology in the production of fruits and vegetables [J]. Agricultural Engineering Technology, 2015 (31):60-63.
[8] Zhu Yutian, Li Jincai, Gao Sujun, et al. Progress in rapid determination field of edible oils by near infrared spectroscopy technology [J]. China oils and fats, 2017, (7):140-143.
[9] Zeng Lu Lu, Tu Bin, Yin Cheng et al. Qualitative and quantitative analysis of peanut oil adulteration by laser near infrared spectroscopy combined with SVM [J]. Journal of the Chinese Cereals and Oils, 2016, 31 (08):126-130+137.
[10] Vapnik, Vladimir Naumovich. The Nature of Statistical Learning Theory [M]. New York:Springer-Verlag, 1999.
[11] Tu Bin, Song Zhiqiang, Zheng Xiaoy et al. Qualitative-Quantitative Analysis of Rice Bran Oil Adulteration Based on Laser Near Infrared Spectroscopy [J]. Spectroscopy and Spectral Analysis, 2015, 35 (6):1539-1545.
[12] Chen Jinfeng. Study and application of support vector machine regression algorithm [D]. Jiangnan University, 2010.
[13] Zhu Fang, Zhao Qingping, Wang Jiangtao. The Classification and Application of SVM Based on the RBF Kernel [J]. Journal of Tangshan University, 2017, 30 (03):13-17+39.