Identification of adulterated olive oil by fusion of near infrared and Raman spectroscopy

Qiang Xu, Xingxing Yang, Zhui Hu, Yuan Li and Xiao Zheng

School of Mechanical Engineering, Wuhan Polytechnic University

Abstract. Aiming at the identification of adulterated olive oil, based on the fusion of near-infrared and Raman data on the feature level, a model for the identification of adulterated olive oil was established. 24 pure olive oil samples and 248 adulterated olive oil samples were used for model establishment in the experiment. With three parameter optimization methods, support vector machine classification and regression models were established based on fusion of near-infrared and Raman data on the feature level with canonical correlation analysis (CCA). The results showed that the accuracy of correction and prediction sets of olive oil adulteration classification model based on the fusion of near-infrared and Raman data was high, and the correlation coefficient of prediction model based on the fusion of near-infrared and Raman data reached 99.13%. The results showed that the classification and prediction models were feasible and effective in the identification of adulterated olive oils, and this work shed light on the development and applications of fusion of spectral data.

1. Introduction

The contents of monounsaturated fats in olive oils are as high as 60%. The intake of olive oil can promote blood circulation, prevent arteriosclerosis and its complications, prevent hypertension, and prevent cerebrovascular and cardiovascular diseases. Therefore, olive oils are popular in markets. Due to the high price of olive oils, driven by the interests, immoral providers fake olive oils with low-grade oils, jeopardizing the interests of consumers. Currently, for the rapid identification of olive oil adulteration, the establishment of olive oil adulteration identification models and databases is prevalent on the basis of the acquirement of oil samples characteristics by chromatography and spectroscopy and chemometrics such as principal component analysis [1], cluster analysis [2] and BP neural network. Among these studies, near-infrared spectroscopy [3-4] and Raman spectroscopy [5-6] are non-destructive and fast, and have been studied and applied in the identification of adulterated olive oils.

For the fusion of the near-infrared and Raman data, in this paper, the original near-infrared and Raman spectra data of adulterated olive oil samples were preprocessed by multivariate scattering correction (MSC), second derivative, and combination of standard normal variate and de-trend (SNV + DT) [7]. After preprocessing, the characteristic wavelengths of near-infrared and Raman spectra were extracted and variables were screened through principal component analysis (PCA), interval partial least squares (iPLS) and combination of competitive adaptive reweighted sampling algorithm and partial least squares method (CARS-PLS) [8-10]. Canonical correlation analysis was used to fuse the near-infrared and Raman data after preprocessing and characteristic wavelengths extraction on a
The prediction accuracy of the olive oil adulteration identification models established was analyzed.

2. Materials and methods

2.1. Samples and sample sets
The oil samples with popular brands from different manufacturers were purchased from online markets and local supermarkets. The samples were representative, eliminating the influence of external factors, such as production technology and reducing the random error in the experiment. In this experiment, peanut oil, rapeseed oil and sunflower seed oil were selected as the dopants. The types and numbers of the adulterated oil samples are shown in Table 1.

| No. | Oil for adulteration       | Subtotal | Oil sample                      |
|-----|----------------------------|----------|---------------------------------|
| 1   | None                       | 24       | pure olive oil                  |
| 2   | Peanut oil                 | 75       | olive oil adulterated with peanut oil |
| 3   | Sunflower seed oil         | 75       | olive oil adulterated with sunflower seed oil |
| 4   | Rapeseed oil               | 98       | olive oil adulterated with rapeseed oil |
| total|                            | 272      |                                  |

Peanut oil, rapeseed oil and sunflower oil were separately added to three types (24 samples) of pure olive oil, and 248 samples with the dopant contents of 2%–34% with an interval of 2% and 35%–70% with an interval of 5% were prepared.

2.2. Instruments and software
A custom-made laser near-infrared (NIR) vegetable-oil-quality detection instrument was employed in the tests. The host was an Axsun XL410-type laser near-infrared spectrometer (The United States), with the scanning range of roughly 1350–1800 nm. A RamTraceer-200 Raman spectrometer (OptoTrace Technologies, Inc., the United States) was used to collect the Raman spectra with 330mW laser power, and the wavenumber range was 250–2340 cm⁻¹ (40000–4273.5nm). The preprocessed and models of spectral data were built and optimized by MATLAB and Unscrambler X10.4 software.

2.3. Acquisition of spectra
The sampling conditions were maintained constant to reduce the influences of temperature, humidity and other environmental factors. First, the oil samples were heated to 60°C with a thermostatic water bath. Then, a sample was fast moved into a 2-mm cuvette, which was then fast placed into a sample cell at 60°C in the NIR detection instrument. After 1 min, the NIR spectral data of the sample were acquired. For the acquirement of Raman spectra, the samples were transferred to special cuvettes, which were placed in the detection device for Raman measurements. The 272 original NIR and Raman data and spectra of true olive oil and adulterated oil are acquired by above instruments (spectra shown in Figure 1 and 2.)
As we see, spectra of true olive oil and adulterated oil from Figure 3 (choose a representative sample of the four types for easy observation), it proves that different types have different characteristic peaks and it is feasibility to classify olive and adulterated oil, and so did Raman spectra.

2.4. Preprocessing of spectra and characteristic wavelength extraction
Spectral preprocessing can eliminate the influence of random noise, background interference and instrumental fluctuations on the calibrated measurement results. Meanwhile, the representative spectral region can be screened to improve the modeling efficiency. In this work, the multiple scattering correction (MSC), second derivative, and combination of standard normal variate and de-trend (SNV-DT) methods were separately adopted to preprocess the original spectra.

Subsequently, getting feature layer variables of preprocessed spectral by the principal component analysis (PCA), interval partial least squares (iPLS), and combination of competitive adaptive
reweighted sampling algorithm and partial least squares method (CARS-PLS) for the preparation of latter data fusion.

2.5. Modeling and spectral fusion methods

2.5.1. Spectral fusion. Fusion on a feature level is to independently process the spectral data, to extract the characteristic wavelengths, and to fuse the spectral data at characteristic wavelengths. With this method, the possibility of successful extraction of featured information from the data is increased, and useful composite information can be obtained.

The preprocessed and extracted near-infrared and Raman spectra on a feature level were fused with canonical correlation analysis (CCA). (Principle of CCA aiming to find one or a few comprehensive variables to replace the original variables according to the correlation between the variables). Apply CCA to fuse feature layer variables of near-infrared and Raman preprocssed spectral, and establish Model with fused data.

2.5.2. Establishment of qualitative and quantitative models. In the experiment, the support vector classification (SVC) method was used for the modeling with the spectral data fused, to realize the qualitative identification of adulterated olive oils. Then, the support vector regression (SVR) approach was selected to quantitatively predict the contents of dopants in adulterated olive oil samples. The principle of SVM is based on the theoretical basis of structural risk minimization principle and defines the classifier with the largest interval in the feature space (basic formula show as 1.1), it could solve both linear and nonlinear classification problems.

\[
\min \frac{1}{2} || \omega ||^2 \\
\text{s.t. } y_i (\omega^T x_i + b) \geq 1, i = 1,2...n
\]

Genetic algorithm (GA), grid calculus (GC), particle swarm optimization (PSO) approaches were used to optimize the SVC models and important parameters (C, g) in the SVR models.

3. Results and discussion

The 272 samples derived included 248 samples adulterated with peanut oil, rapeseed oil and sunflower seed oil and 24 pure olive oil samples. 204 samples were selected as the correction set and the other 68 samples were selected as the prediction set by SPXY. After the preprocessing, characteristic wavelength extraction and parameters optimization, spectra data and corresponding number of sample were used as the X and y input of the model, correction set were establish SVC model and prediction set were tested prediction performance of SVC model. The outputs prediction results of the olive oil adulteration identification SVC models are shown in Table 2.

| Data fusion | Preprocessing method | Characteristic wavelength extraction method | Modeling method | Calibration set results | Prediction set results |
|-------------|----------------------|---------------------------------------------|-----------------|-------------------------|-----------------------|
|             | MSC                  | iPLS-CCA                                    | GC-SVC          | 100(0/204)              | 97.05(2/68)           |
|             |                      | CARS-PLS-CCA                                | PSO-SVC         | 100(0/204)              | 100(0/68)             |
|             |                      | PCA-CCA                                     | GA-SVC          | 100(0/204)              | 98.53(1/68)           |
| Second order|                      | CARS-PLS-CCA                                | PSO-SVC         | 100(0/204)              | 100(0/68)             |
|             | SNV+DT               | PCA-CCA                                     | GA-SVC          | 100(0/204)              | 100(0/68)             |
According to the data shown in Table 2, the calibration and prediction sets of SVC models established by CCA with fusion of spectral data on a feature level showed high accuracy. The prediction classification results of the SNV+DT-PCA-CCA-SVC model in Figure 4 exhibited high accuracy. Besides, the SVC models preprocessed by the MSC-iPLS and MSC-PCA methods also exhibited high prediction accuracy with only one or two errors in the prediction sets. The results show that with the fusion of near-infrared and Raman spectral data, the identification of adulterated olive oils can be effectively implemented.

![Figure 4. Prediction classification results of the SNV+DT-PCA-CCA-SVC model.](image)

Meanwhile, for the verification of the feasibility and accuracy of the fusion spectra in the detection of contents of dopants, the fusion of near-infrared and Raman spectra were set as the input of SVR models to establish SVR models. Calculating relative error between predicted and true value to assess model performance. The prediction results are shown in Table 3.

| Preprocessing and characteristic wavelength extraction methods | Number of variables | Parameters | Calibration set | Prediction set | RMSEP |
|-------------------------------------------------------------|---------------------|------------|----------------|---------------|-------|
| MSC-iPLS-GC                                                 | 136                 | 256        | 9.76e-04       | 99.98         | 91.97 | 0.51 |
| MSC-PCA-PSO                                                 | 10                  | 56.05      | 0.01           | 94.05         | 99.13 | 0.04 |
| MSC-PCA-GA                                                  | 10                  | 129.56     | 0.01           | 91.61         | 98.57 | 0.11 |
| Second-order CARS-PLS-PSO                                   | 36                  | 72.72      | 0.10           | 99.80         | 78.13 | 0.34 |
| SNV+DT-PCA-GA                                               | 10                  | 63.98      | 0.06           | 93.30         | 98.51 | 0.08 |

Table 3 shows that the R values of the calibration sets of the SVR models based on the fusion spectra after the MSC-iPLS and second-order CARS-PLS preprocessing and characteristic wavelength extraction were as high as 99.98% and 99.80%. Especially, the R value of the prediction set of the SVR model based on the fusion spectra after the MSC-PCA-PSO preprocessing and characteristic wavelength extraction reached 99.13%, and the RMSEP value (root mean square error of the predicted
value) was as low as 0.04, indicating that the SVR prediction model is superior in generalization and accuracy.

**Table 4.** Relative errors of the SVR models based on the fusion of near infrared and Raman spectra preprocessed by MSC-PCA-PSO.

| No. | Real value (proportion of dopants) | Predicted value | Relative error(%) |
|-----|-----------------------------------|-----------------|------------------|
| 1   | 34                                | 37.01           | 8.13             |
| 2   | 20                                | 18.91           | 5.76             |
| 3   | 4                                 | 3.59            | 11.42            |
| 4   | 26                                | 24.92           | 4.33             |
| 5   | 32                                | 31.65           | 1.11             |
| 6   | 14                                | 15.18           | 7.77             |
| 7   | 8                                 | 7.51            | 6.52             |
| 8   | 34                                | 37.01           | 8.13             |
| 9   | 20                                | 18.91           | 5.76             |
| 10  | 22                                | 23.66           | 7.02             |
| 11  | 12                                | 11.35           | 5.73             |
| 12  | 50                                | 48.39           | 3.33             |
| 13  | 20                                | 20.85           | 4.08             |
| 14  | 6                                 | 4.43            | 35.44            |
| 15  | 14                                | 15.18           | 7.77             |
| 16  | 22                                | 22.15           | 0.68             |
| 17  | 55                                | 55.39           | 0.70             |
| 18  | 66                                | 66.38           | 0.57             |
| 19  | 75                                | 73.76           | 1.68             |
| 20  | 32                                | 32.36           | 1.11             |
| 21  | 70                                | 69.21           | 1.14             |
| 22  | 55                                | 55.39           | 0.70             |

Average relative error 5.85

Note: relative error = (|real value - predicted value|)/real value * 100. (The actual prediction set contained 68 prediction results, and this table only shows a part of them.)

For the clear expression of the prediction results, the predicted results and real values (actual proportions of dopants) of the prediction set of the model preprocessed by MSC-PCA-PSO are given in the form of relative error table (Table 4). The overall relative error (average relative error) was relatively small, which shows that the SVR model based on the fusion of near-infrared and Raman data on a feature level is feasible and accurate in the prediction of dopants contents in olive oil.

**Acknowledgments**

Funds: Innovation and Transformation of Grain Science and Technology of Hubei Province in 2016 (20165104); Key Science and Technology Project of Wuhan City (2013010501010147).

**References**

[1] Wang Y 2014 Identification of edible oils by second derivative infrared spectroscopy combined with principal component analysis [A]. Chinese Academy of Optics. *Proceedings of the 18th National Molecular Spectroscopy Academic Conference* [C]. Chinese Academy of Optics: Spectral Committee of Chinese Academy of optics

[2] Deng P J, Geng Y J, Liang Y, Yang D Y, Li H and Yang Y C 2015 Rapid identification of adulterated olive oil by Raman spectroscopy and clustering analysis [J]. *Chinese Oils* **40(02)** 50-55
[3] Zhang Y 2015 *Rapid identification and detection of adulterated edible oils by near infrared spectroscopy* [D]. Henan University of Technology

[4] Mo X X 2018 *Rapid determination of trans fatty acids in edible vegetable oils by near infrared and Raman spectroscopy* [D]. Jiangxi Agricultural University

[5] Li Y P 2016 *Qualitative identification and content detection of adulterated edible oils based on Raman spectroscopy* [D]. Jinan University

[6] Huang S, Wang Q, Ying R F, Wang Y S, Hu F J and Huang M G 2019 Applications of Raman spectroscopy in the identification of adulteration and quality of olive oils [J]. *Food Industry Technology* **40**(11) 334-341 + 350

[7] Zhao L Y 2018 *Research on data processing and analysis methods in spectral measurements* [D]. Tianjin University of Technology

[8] Wu J Z, Shi R J, Chen Y and Liu C L 2015 Screening of near infrared characteristic spectral regions for the detection of oleic acid in edible oils [J]. *Journal of Grain and Oil* **30**(02) 118-121

[9] Wang H L, Yang G G, Zhang Y, Bao Y D and He Y 2017 Detection of tomato leaf fungal diseases through extraction of characteristic wavelengths with competitive adaptive reweighting algorithm and correlation coefficient methods [J]. *Spectroscopy and Spectral Analysis* **37**(07) 2115-2119

[10] Liu X M, Zhang H L, Sun X D 2010 NIR sensitive wavelength selection based on different methods[P]. *Mechanic Automation and Control Engineering (MACE), International Conference on*

[11] Xing L Y, Wang Q Y and Yin X Y 2018 Detection of glucose based on Raman spectroscopy and Multi-space fusion pretreatment [J]. *Spectroscopy and Spectral Analysis* **38**(S1) 175-176

[12] Gharbia R, El Baz A H, Hassanien A E, Schaefer G, Nakashima T and Azar A T 2014 Fusion of multi-spectral and panchromatic satellite images using principal component analysis and fuzzy logic[P]. *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on*