Determination of the Real Contents of Olive Oil in Blend Oils by near Infrared Spectroscopy

Zihao Cai, Qiang Xu, Xingxing Yang, Zhui Hu, Xiao Zheng* and Jiafan Zhang

(School of Mechanical Engineering, Wuhan Polytechnic University)

Corresponding Author Email: zhengxiao@whpu.edu.cn

Abstract. Aiming at the determination of real contents of olive oil in blend oils, near infrared spectroscopy and support vector machine were combined to establish regression models for the determination of olive oil contents. The data were fused on the feature level with canonical correlation analysis to improve the prediction performance of the models. The results showed that the SVM regression models could effectively predict the real contents of olive oil in 135 samples containing 2%-20% olive oil. Among these models, the highest correlation coefficient R2 was 99.74%, and the root mean square error of the prediction set was 0.08. In addition, the prediction performances of some regression models were deteriorated after the curves were smoothed with the Savitzky-Golay method. But, the prediction performances were improved in the models established with data fused with the canonical correlation analysis approach.

1. Introduction

Olive oil is beneficial for blood circulation and can prevent arteriosclerosis and complications of arteriosclerosis, hypertension, and cardiovascular and cerebrovascular diseases. Market survey shows that organic, green, natural, additive-free, nutritious, healthy, and high-quality edible oils are popular with the consumers, and thereby olive-oil-containing blend oils have drawn more and more attention [1-2]. The models and databanks constructed by the combination of near infrared spectroscopy, chemometrics, and machine learning technologies have been widely studied and applied in the determination of edible oils due to their advantages of nondestructive and rapid detection [3-4].

In this paper, adaptive iteratively reweighted penalized least squares (airPLS), direct orthogonal signal correction (DOSC), and multiplication scatter correction (MSC) preprocessing methods were adopted to eliminate the noises and interference from unrelated variables in the original spectral data. Interval partial least squares (iPLS) and combination of competitive adaptive reweighted sampling algorithm and partial least squares (CARS-PLS) methods were used to extract the characteristic wavelengths of the preprocessed NIR spectral data for the establishment of the support vector machine regression models for the prediction of the qualities of olive-oil-containing blend oils. Meanwhile, canonical correlation analysis (CCA) was carried out for the fusion of data in the models with poor prediction performances on a feature level. The effect of data fusion on the stabilities of models was analyzed [5-6].

2. Materials and Methods

2.1. Samples and Sample Sets

The oil samples for testing were purchased from online markets and local supermarkets. According to the market survey, the contents of olive oil in blend oils were lower than 20% under most circumstances, so the contents of olive oil in the blend oil samples were no more than 20%. For
eliminating the influence of external factors such as production technology and production technology and reducing the random error in the experiments, representative samples with popular brands from different manufacturers were purchased from online markets and local supermarkets. The types and codes of the oil samples are shown in Table 1.

| Code | Type of blend oil                        | Subtotal (share) | Remarks |
|------|-----------------------------------------|------------------|---------|
| 1    | rapeseed oil - corn oil - olive oil      | 45               |         |
| 2    | rapeseed oil - olive oil                | 45               |         |
| 3    | sunflower seed oil - olive oil          | 45               |         |
|      | Total                                   | 135              |         |

2.2. Instruments and Software

A custom-made laser near-infrared (NIR) vegetable-oil-quality detector was used in the experiments. The host was an Axsun XL410-type laser near-infrared spectrometer (The United States), with the scanning range of 1350–1800 nm and spectral resolution of 3.5 cm⁻¹. The scanning was performed 32 times. The temperature could be adjusted within the range of 20–100 °C. An RamiTraceer-200 Raman spectrometer (OptoTrace Technologies, Inc., the United States) was adopted to acquire the Raman data. The highest laser power was 330 mW, and the wavenumber range was 250–2340 cm⁻¹ (4273.5–40000 nm).

The spectral data were preprocessed, and the models were constructed and optimized with the Unscrambler X10.4 and MATLAB software.

2.3. Acquisition of Spectra

The sampling conditions were maintained constant to reduce the impact of temperature, humidity and other environmental factors. The oil samples were heated to 60°C in a thermostatic water bath. A sample was placed into a 2-mm cuvette, which was then placed into the sample cell at 60 °C. After 1 min, the NIR spectra of samples with different olive oil contents are shown in Figure 1.

Fig. 1 shows that in the wavelength range of 1350–1800 nm, blend oils with different olive oil contents showed distinctive characteristic bands, indicating that it was feasible to determine the real content of olive oil by NIR.

![Figure 1. Near infrared spectra of blend oils with different olive oil contents](image_url)

2.4. Preprocessing of Spectral Data and Characteristic Wavelength Extraction

Spectral preprocessing can exclude the influences of random noises, background interference and instrumental fluctuations on the calibrated measurement results to improve the modeling efficiency. In this work, the airPLS, DOSC, and MSC preprocessing methods were separately adopted to eliminate
the noises and interference from unrelated variables in the original spectral data.

In the experiments, the characteristic wavelengths of the original spectra were extracted through dimension reduction with the interval partial least squares (iPLS) and combination of competitive adaptive reweighted sampling algorithm and partial least squares method (CARS-PLS) approaches. The canonical correlation analysis (CCA) method was performed for the fusion of data in the models with poor prediction performances on a feature level.

2.5. Data Fusion and Modeling Methods

2.5.1. Data fusion. Data fusion on a feature level means to independently process the data, to extract the characteristic wavelengths, which are used to fuse the spectral data. In this way, the probability of extraction of featured information from the data is increased, and useful complex information can be acquired[7-8].

CCA was adopted to fuse two groups of near infrared spectral data on a feature level after preprocessing and wavelength extraction in different ways. The principle of CCA is to study the relationship between one group of variables and another group, and then one or several comprehensive variables are found to replace the original variables according to the correlations between variables. The modeling with fused data can effectively improve the poor prediction performances of the models caused by the loss of characteristic data during data processing [9].

2.5.2. Regression of models. The support vector regression (SVR) method was employed to quantitatively predict the contents of olive oil in blend oil samples [10-11]. Genetic algorithm (GA), grid calculus (GC), particle swarm optimization (PSO) algorithms were adopted to optimize the important parameters (C, g) in the SVR models. According to MSEPs (root mean square error) of the prediction sets and relative prediction errors, the optimal model among three schemes was selected as the final prediction model [12].

3. Results and Discussion

3.1. Regression Models for the Prediction of Real Content of Olive Oil

The quantitative models for the detection of olive oil contents in blend oils were established with the as-processed data. Three optimization algorithms PSO, GS and GA were separately adopted to find the optimal parameters for the SVR models, and multiple SVR models for the quantitation of olive oil content were obtained. The prediction results are shown in Table 2.

| Spectros copy | Preprocessing and characteristic wavelength extraction methods | Optimization algorithm | Number of variables | Parameters | Calibration set | Prediction set | Mean square error |
|---------------|---------------------------------------------------------------|------------------------|--------------------|------------|----------------|--------------|------------------|
| NIR airPLS-CARS-PLS | PSO | 63 | 1000 | 344.17 | 98.84 | 99.98 | 7e-3 |
| DO ASC-CARS-PLS | GS | 157 | 64 | 9.76e-4 | 99.09 | 99.74 | 0.08 |
| SG-CARS-PLS | PSO | 70 | 1000 | 17.03 | 76.05 | 97.59 | 0.83 |
| MSC-CARS-PLS | PSO | 48 | 1000 | 1000 | 99.22 | 99.78 | 0.09 |
| MSC-biPLS | GS | 157 | 1024 | 262.53 | 97.48 | 99.65 | 0.14 |
| MN-biPLS | GA | 157 | 363.46 | 522.85 | 99.99 | 99.94 | 0.02 |
As shown in Table 2, the correlation coefficient $R^2$ of the correction set of the airPLS-CARS-PLS-PSO-SVR model reached 98.84%, and that of the prediction set was as high as 99.98%, higher than those of other models, indicating that the prediction accuracy was high and regression fitting performance was good. However, the model parameters (C, g) were greater after the optimization, and the generalization of the model was poor. In contrast, the generalization performance of the DOSC-CARS-PLS-GC-SVR model was greater, and the $R^2$ value of the prediction set was 99.74, and MSEP value was 0.08, indicating that this model had small prediction deviations. And, the prediction stability and accuracy were high. The relative errors of the prediction results of the DOSC-CARS-PLS-GS-SVR model are shown in Table 3.

### Table 3. Relative errors of the prediction results of the DOSC-CARS-PLS-GS-SVR model for the quantitative detection of olive oil contents in blend oils

| No. | Calibration set (real value) | Calibration set (predicted value) | Relative error (%) | Prediction set (real value) | Prediction set (predicted value) | Absolute error (%) | Relative error (%) |
|-----|-----------------------------|----------------------------------|--------------------|-----------------------------|----------------------------------|--------------------|--------------------|
| 1   | 16.00                       | 15.30                            | 4.38               | 8.00                        | 8.03                             | 0.03               | 0.43               |
| 2   | 10.00                       | 10.11                            | 1.12               | 12.00                       | 12.26                            | 0.26               | 2.21               |
| 3   | 14.00                       | 14.45                            | 3.18               | 16.00                       | 16.36                            | 0.36               | 2.22               |
| 4   | 10.00                       | 9.64                             | 3.57               | 18.00                       | 17.96                            | 0.04               | 0.23               |
| 5   | 18.00                       | 17.61                            | 2.17               | 18.00                       | 17.72                            | 0.28               | 1.54               |
| 6   | 18.00                       | 17.92                            | 0.46               | 8.00                        | 8.64                             | 0.64               | 8.06               |
| 7   | 8.00                        | 7.81                             | 2.39               | 2.00                        | 2.23                             | 0.23               | 11.69              |
| 8   | 4.00                        | 4.28                             | 6.99               | 2.00                        | 1.63                             | 0.37               | 18.72              |
| 9   | 12.00                       | 12.31                            | 2.60               | 6.00                        | 6.12                             | 0.12               | 1.98               |
| 10  | 8.00                        | 8.54                             | 6.69               | 6.00                        | 6.21                             | 0.21               | 3.43               |
| 11  | 10.00                       | 9.49                             | 5.12               | 14.00                       | 14.15                            | 0.15               | 1.07               |
| 12  | 6.00                        | 6.15                             | 2.50               | 18.00                       | 18.16                            | 0.16               | 0.91               |
| 13  | 2.00                        | 2.13                             | 6.56               | 20.00                       | 19.95                            | 0.05               | 0.25               |
| 14  | 14.00                       | 14.93                            | 6.65               | 2.00                        | 2.15                             | 0.15               | 7.65               |
| 15  | 16.00                       | 16.31                            | 1.95               | 10.00                       | 9.74                             | 0.26               | 2.57               |

Average relative error: 3.78
Mean error: 0.21
Overall average relative error: 5.68

3.2. Svr Fusion Models for the Detection of Real Content of Olive Oil

Table 2 shows that the prediction performance of the SG-CARS-PLS-SVR model was not ideal compared with those of other models using different preprocessing and feature wavelength extraction methods. In the absence of spectral information acquirement using other instruments, the NIR-based featured data after different processing steps were fused, and the fused data were used to establish the SVR fusion models for the determination of olive oil content. And, the modeling performance was assessed. The prediction results of these fusion models are shown in.
Table 4. Prediction results of the fusion models for the determination of olive oil contents in blend oils

| Spectroscopy | Preprocessing and characteristic wavelength extraction methods | Optimization algorithm | Number of variables | Parameters | Calibration set | Prediction set | Mean square error |
|--------------|---------------------------------------------------------------|------------------------|---------------------|------------|----------------|---------------|------------------|
| NIR SG-CARS-PLS | PSO | 60 | | 1000 | 17.03 | 76.05 | 97.59 | 0.83 |
| MSC-biPLS | GS | 157 | | 1024 | 262.53 | 97.48 | 99.65 | 0.14 |
| MN-biPLS | GA | 157 | | 363.46 | 522.85 | 99.99 | 99.94 | 0.02 |
| SG-CARS-PLS | GS | | | 1024 | 0.01 | 99.99 | 99.91 | 0.04 |
| MSC-biPLS | GA | 60 | | 732.31 | 0.01 | 99.99 | 99.91 | 0.04 |
| CCA SG-CARS-PLS | PSO | | | 850.12 | 0.01 | 99.98 | 99.90 | 0.04 |
| MN-biPLS | GA | | | 1024 | 0.01 | 99.99 | 99.86 | 0.07 |
| | PSO | | | 1024 | 0.01 | 99.99 | 99.91 | 0.04 |
| | GA | | | 727.15 | 0.01 | 99.99 | 99.86 | 0.06 |

Table 4 shows that the R² values of the correction and prediction sets of the SG-CARS-PLS-SVR model are relatively low. After the fusion of the feature variables of SG-CARS-PLS-SVR, MSC-biPLS, and MN-biPLS models with CCA, the modeling was conducted. The models showed improved prediction accuracy and stability to a certain extent, and the MSEP value was reduced from 0.83 in the beginning to 0.04 in the end. Although the modeling time was increased to some extent in this approach, the flexible applications of the spectral fusion data and improvement of the modeling performance can be fulfilled in this way.

4. Conclusions
The results show that the fast determination of the real content of olive oil in blend oil can be realized by using near-infrared spectral data and support vector machine regression modeling. The generalization performance of the DOSC-CARS-PLS-GC-SVR model is great. Moreover, after the fusion of near-infrared data on a feature layer after preprocessing and feature wavelength extraction, the prediction results show that the prediction accuracy of the data fusion SVR model can be improved to a certain extent, indicating that the CCA technology can not only find the correlations between data, but also improve the performances of regression models with the data.

5. References
[1] Wu Meijun, Pu Hongzheng, Gou Yao, Diao Yue, Li Yuanyuan. Research progress on the effective components and applications of olive oil [J]. Anhui Agronomy Bulletin, 2020,26 (10): 34-35
[2] E.Pardo, T. Lescot, P. Massanet, N. Lapidus, V. Fraipont, F. Tamoin, L. Petit, C. Ichai, G. Taverny, P. Boizeau, C. Alberti, J.-C. Preiser. SUN-LB642: Beneficial Effect of Olive Oil Based Parenteral Nutrition Compared to Fish Oil in Critically Ill Patients[J]. Elsevier Ltd,2019,38.
[3] Zhu Yutian, Li Jincai, Gao Sujun, Zheng Yanyan. Research progress of near infrared spectroscopy in the field of edible oil rapid detection [J]. Chinese Oil, 2017,42 (07): 140-143
[4] Zhang Yan. Rapid identification and detection of adulteration of edible oil by near infrared spectroscopy [D]. Henan University of Technology, 2015
[5] Zheng Kaiyi, Fan Wei, Wu Ting, Hu Huilian, Du Yiping, Liang Yizeng. Calibration of near infrared spectra measured by different instruments and at different times by canonical
correlation analysis [J]. Computer and Application, 2013,30 (03): 246-250
[6] Correlation Analysis; Reports from University of Zurich Add New Data to Findings in Correlation Analysis (Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis)[J]. Science Letter, 2015.
[7] Li Zhiyong, Kuang Gangyao, Zou Huanxin, Wu Hao. Research on anomaly detection algorithm of hyperspectral image based on feature layer fusion [J]. Journal of Remote Sensing, 2003 (04): 304-308
[8] Wu Shuang, Wang Jie, Yu Yaru, Tu bin, Zheng Xiao, He Dongping. Detection of MUFA and PUFA contents in edible oils based on the fusion of Raman and near infrared spectra on their feature layers [J]. Journal of China Grain and Oil Journal, 2017,32 (11): 158-164
[9] Han weiqi, Sun Zi Ru, Zhao Bing. Study on the coordination of higher education and regional economy in Beijing, Tianjin and Hebei Province based on canonical correlation analysis [J]. Tianjin Science and Technology, 2020,47 (02): 5-8 + 11
[10] Xiong Shen, Xiao Zheng, Zhiqiang Song, Dongping He, Peishi Qi. Rapid identification of waste cooking oil with near infrared spectrum based on support vector machine [C]. China Agricultural University, China Agricultural Engineering Society, Beijing Agricultural Information Society. Computer and Computing Technologies in Agriculture VI—Proceedings of 6th IFIP WG 5.14 International Conference(CCTA2012) Part I. China Agricultural University, China Agricultural Engineering Society, Beijing Agricultural Information Society: Beijing Agricultural Internet of Things Engineering Technology Research Center, China Agricultural University, 2012:28-35
[11] Chen Jinfeng. Research and applications of support vector machine regression algorithms [D]. Jiangnan University, 2008
[12] Yu Yaru. Detection of the adulteration of sesame oil and camellia oil by near infrared spectroscopy [D]. Wuhan Light Industry University, 2018