Variables Selection in the Ultraviolet, Visible and Near Infrared Range for Calibration of a Mixture of Vegetable Oils by Absorbance Spectra

M.A. Khodasevich, D.A. Borisevich

*B.I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus,
Nezavisimosty Ave., 68, Minsk 220072, Belarus*

*Received 29.01.2021
Accepted for publication 01.03.2021*

Abstract

The aim of the work was a multivariate calibration of the concentration of unrefined sunflower oil, considered as adulteration, in a mixture with flaxseed oil. The relevance of the study is due to the need to develop a simple and effective method for detecting the falsification of flaxseed oil which is superior in the content of essential polyunsaturated fatty acids to olive oil. A few works only are devoted to identifying adulteration of flaxseed oil, unlike olive oil.

Multivariate calibration carried out using a model based on the principal component analysis, cluster analysis and projection to latent structures of absorbance spectra in UV, visible and near IR ranges. Calibration uses three methods for spectral variables selection: the successive projections algorithm, the method of searching combination moving window, and method for ranking variables by correlation coefficient.

The application of the successive projections algorithm, ranking variables by correlation coefficient and searching combination moving window makes it possible to reduce the value of the root mean square error of prediction from 0.63 % for wideband projection to latent structures to 0.46 %, 0.50 %, and 0.03 %, respectively.

The developed method of multivariate calibration by projection to latent structures of absorbance spectra in UV, visible and near IR ranges using the spectral variables selection by searching combination moving window is a simple and effective method of detecting adulteration of flaxseed oil.

Keywords: spectral analysis, principal component analysis, projection to latent structures, spectral variables selection.

DOI: 10.21122/2220-9506-2021-12-1-75-81

Адрес для переписки:
*Ходасевич М.А.
Институт физики имени Б.И. Степанова НАН Беларуси,
пр-т Независимости, 68, г. Минск 220072, Беларусь
e-mail: m.khodasevich@ifanbel.bas-net.by*

Address for correspondence:
*Khodasevich M.A.
B.I. Stepanov Institute of Physics of the National Academy
of Sciences of Belarus,
Nezavisimosty Ave., 68, Minsk 220072, Belarus
e-mail: m.khodasevich@ifanbel.bas-net.by*

Для цитирования:
M.A. Khodasevich, D.A. Borisevich.
Variables Selection in the Ultraviolet, Visible and Near Infrared Range for Calibration of a Mixture of Vegetable Oils by Absorbance Spectra.
Приборы и методы измерений.
2021. – Т. 12, № 1. – С. 75–81.
DOI: 10.21122/2220-9506-2021-12-1-75-81

For citation:
M.A. Khodasevich, D.A. Borisevich.
Variables Selection in the Ultraviolet, Visible and Near Infrared Range for Calibration of a Mixture of Vegetable Oils by Absorbance Spectra.
Devices and Methods of Measurements.
2021, vol. 12, no. 1, pp. 75–81.
DOI: 10.21122/2220-9506-2021-12-1-75-81

Выбор переменных в ультрафиолетовом, видимом и ближнем инфракрасном диапазонах для калибровки смеси растительных масел по спектрам оптической плотности

М.А. Ходасевич, Д.А. Борисевич

Институт физики имени Б.И. Степанова Национальной академии наук Беларуси,
пр-т Независимости, 68, г. Минск 220072, Беларусь

Поступила 29.01.2020
Принята к печати 01.03.2021

Целью работы являлась многопараметрическая калибровка концентрации нерафинированного подсолнечного масла, рассматриваемого в качестве фальсификата льняного масла. Актуальность исследования обусловлена необходимостью разработки простого и эффективного метода обнаружения фальсификации льняного масла, превосходящего по содержанию незаменимых полиненасыщенных жирных кислот оливковое масло, выявлению подделок которого в отличие от льняного посвящено большое количество работ.

Многопараметрическая калибровка проводилась с помощью модели, основанной на методе главных компонент, кластерном анализе и проекции на латентные структуры спектров оптической плотности в УФ-, видимом и ближнем ИК диапазонах с применением трех методов выбора спектральных переменных: метода последовательного проецирования, метода поиска комбинации сдвигающихся окон и метода ранжирования переменных по коэффициенту корреляции.

Показано, что применение методов последовательного проецирования, ранжирования переменных по коэффициенту корреляции и поиска комбинации сдвигающихся спектральных окон позволяет уменьшить величину среднеквадратичного отклонения калибровки с 0,63 % для широкополосной проекции на латентные структуры до 0,46 %, 0,50 % и 0,03 %, соответственно.

Разработанный метод многопараметрической калибровки с помощью проекции на латентные структуры спектров оптической плотности в УФ-, видимом и ближнем ИК диапазонах с применением выбора спектральных переменных путём поиска комбинации сдвигающихся окон является простым и эффективным средством обнаружения фальсификации льняного масла.

Ключевые слова: спектральный анализ, метод главных компонент, проекция на латентные структуры, выбор спектральных переменных.

DOI: 10.21122/2220-9506-2021-12-1-75-81

Адрес для переписки:
Ходасевич М.А.
Институт физики имени Б.И. Степанова НАН Беларуси,
пр-т Независимости, 68, г. Минск 220072, Беларусь
e-mail: m.khodasevich@ifanbel.bas-net.by

Address for correspondence:
Khodasevich M.A.
B.I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus,
Nezavisimosty Ave., 68, Minsk 220072, Belarus
e-mail: m.khodasevich@ifanbel.bas-net.by

Для цитирования:
M.A. Khodasevich, D.A. Borisevich.
Variables Selection in the Ultraviolet, Visible and Near Infrared Range for Calibration of a Mixture of Vegetable Oils by Absorbance Spectra.
Приборы и методы измерений.
2021. – Т. 12, № 1. – С. 75–81.
DOI: 10.21122/2220-9506-2021-12-1-75-81

For citation:
M.A. Khodasevich, D.A. Borisevich.
Variables Selection in the Ultraviolet, Visible and Near Infrared Range for Calibration of a Mixture of Vegetable Oils by Absorbance Spectra.
Devices and Methods of Measurements.
2021, vol. 12, no. 1, pp. 75–81.
DOI: 10.21122/2220-9506-2021-12-1-75-81

Introduction

Food adulteration is a serious problem around the world. Products of animal and vegetable origin with a high content of fat are most subject to falsification. Meat, fish, oils, dairy products, etc. account for almost 68 % of adulterated food products [1]. Vegetable oil is one of the most widely demanded foods. Olive oil, which is wrongly considered the most beneficial for human health, is the most often adulterated vegetable oil. A large number of studies are devoted to the detection of falsification of olive oil using optical spectroscopy methods such as fluorescence and UV and visible spectroscopy [2], Raman spectroscopy [3], combination of near and mid IR spectroscopy [4], etc. But olive oil is inferior to flaxseed oil in the content of essential polyunsaturated fatty acids, among which the content of alpha-linolenic (omega-3) acid can reach 64 %. Only a small amount of studies has been focused on detecting flaxseed oil adulteration. For example, Fourier spectroscopy was used to detect the falsification of flaxseed oil with olive oil [5] and mid-IR spectroscopy was used to detect the adulteration of flaxseed oil by soybean and sunflower oils [6].

Earlier [7], we carried out a multivariate calibration of the concentration of unrefined sunflower oil, considered as adulteration, in a mixture with flaxseed oil using a model based on the principal component analysis (PCA) [8], cluster analysis and projection to latent structures (PLS) [9] of absorbance spectra in UV, visible and near IR ranges. To further reduce the root mean square error of prediction ($\mathrm{RMSE_P}$), in this work we compared three methods for spectral variables selection: the Successive Projections Algorithm (SPA) [10], the searching combination moving window interval PLS (scmwiPLS) and the method using correlation coefficients ranging [11].

The objects of the study were specially prepared samples of binary mixtures of unrefined sunflower and flaxseed oils with a percentage from 0 to 100 %. Absorbance spectra were measured on a Shimadzu UV-3101PC spectrophotometer with a step of 1 nm in two ranges: from 335 to 690 nm and from 1130 to 2200 nm with a slit width of 1 nm and 3 nm, respectively. The interval 1698–1766 nm, corresponding to the first overtone of the C – H vibrations of the $–CH_2–$ group [12, 13], is very noisy, therefore, it was not taken into account in further consideration.

Spectra processing and multivariate calibration

Before applying the PCA method, it is necessary to form a rectangular matrix of spectra of the studied samples. In this matrix rows are samples, columns are spectral variables. According to the dependence of the total explained variance of the spectral data on the number of principal components, it was determined that 99.7 % of the total variance is described by the first principal component. Using the linear approximation of the scores to the first principal component, samples that deviate significantly from the general dependence are identified as outliers. These samples correspond to 10 %, 25 %, 30 %, 60 %, 65 %, 70 % and 72.5 % concentrations of sunflower oil and were removed from further consideration.

To create the PLS model, the remaining samples were divided into training sampling and test one by the hierarchical cluster analysis in the Euclidean space of the first principal component of absorbance spectra. For a planned experiment, this method gives smaller values of $\mathrm{RMSE_P}$ [14] compared to uniform partitioning by a calibrated parameter or the frequently used Kennard–Stone algorithm [15]. The values of scores to the first principal component were aggregated to 6 clusters. 6 spectra with scores that were closest to the centers of the clusters were selected to the test sampling. The remaining 18 samples constituted the training sampling. Thus, 75 % of the samples are used to build the model and 25 % to validate it.

After the stage of dividing the samples into training and test samplings, one can proceed to calibrating the content of sunflower oil in a mixture with flaxseed oil using a wideband multivariate PLS with all 1345 spectral variables. Figure 1 shows that the optimal number of latent structures is 6, since $\mathrm{RMSE_P}$ in this case is minimal and equal to 0.63 %.

Due to the collinearity of spectral data and possibility of low signal-to-noise ratio for individual spectral variables and even in rather wide spectral intervals, the use of the entire measured spectral range may not be optimal for calibration accuracy. To improve the quality of the multivariate model, it is advisable to reduce the number of variables taken into account in the simulation. The spectral variables selection is an important step in improving the quality of calibration and stability of the model with possible verification using additional samples.

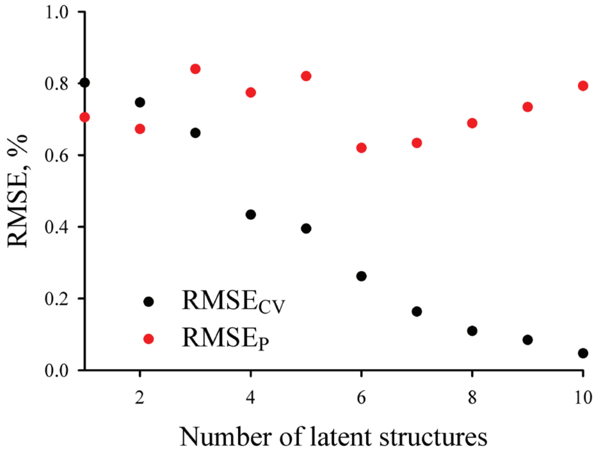

Figure 1 – Dependences of the RMSE_{CV} (root mean square error of cross-validation) for the training sampling and RMSE_{P} for the test sampling for the wideband PLS

We consider three following methods for spectral variables selection. The first method is based on ranking the variables using the correlation coefficients between spectral counts and the calibrated parameter found for wideband PLS with six latent structures. In this case, the spectral variables are excluded from the multivariate model one by one in accordance with the decreasing correlation coefficient. RMSE_{P} is determined at each step. The minimum value of RMSE_{P} specifies an optimum set of spectral variables that corresponds to the best model for the applied method. Figure 2 shows that this minimum value of $\text{RMSE}_{\text{P}} = 0.50$ % is achieved when removing 1106 spectral variables for 239, taken into account in the multivariate model.

The second considered method is SPA. At the first stage of algorithm fulfillment, for the 1345 variables available in our case, a set of 1345 ordered sequences of spectral variables is constructed, the first elements of which are different. In the multidimensional space of spectral variables the remaining 1344 variables are projected onto the space orthogonal to the selected first variable. The largest projection value determines the second in order variable. Similarly, all the following spectral variables in considered sequence are ranked by projections on the subspace orthogonal to the subspace of the variables already selected. For each element of the generated set of ordered sequences of spectral variables, PLS is constructed starting with the first ten spectral variables for certainty, and ending with a set of all 1345 variables. For every number of spectral variables taken into account in the multivariate models for variables sequence considered, the optimal number of latent structures was selected based on the minimum value of RMSE_{P}. The global minimum of RMSE_{P} was found from $1795575 = 1345 \times 1335$ values. Here 1345 is the number of elements in the set of ordered sequences of spectral variables and 1335 is the number of PLS models with an increase in the number of spectral variables from 10 to 1345. Based on the global minimum of RMSE_{P} of the sunflower oil concentration in a binary mixture of vegetable oils, the required sequence of spectral variables was determined, which ensures maximum calibration accuracy for variables selection method applied. In our case, the required sequence of spectral variables began with wavelength of 1781 nm and consisted of only 14 variables. It is rather small number of selected variables and its further reduction is impractical. Often the final stage of SPA execution aims to reduce number of selected variables, taking into account the correlation coefficient of the spectral variables and the calibrated parameter.

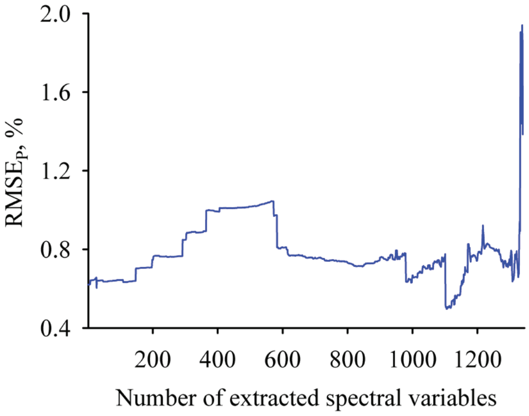

Figure 2 – Dependence of the RMSE_{P} on the number of extracted spectral variables in ranking method using correlation coefficients between spectral counts and calibrated parameter

The third used method is searching combination moving window interval PLS (scmwiPLS) [14]. In contrast to the two previous methods, the described method operates not with individual spectral variables, but with a continuous interval or, as it is often called in multivariate analysis, a window [16]. The algorithm for applying this method is as follows. First, you need to select the width of the windows that shift along the spectrum. In the scmwiPLS modification we use, the number of spectral variables in window exceeds the number of latent structures by one in order for PLS to be able to reduce the dimension of the variable space by at least one. Note that, unlike SPA, the number of latent structures (6 in our case) does not change during the whole algorithm. Second, the spectral position of the first window should be determined. It shifts across the entire spectral range and is fixed in the place where $RMSE_P$ of PLS model based on selected spectral variables is minimal. Third, it is necessary to determine the position of the added windows until they fill the entire measurement range. Subsequent windows are similarly shifted within the entire spectral range of measurements and are alternately combined with the selected ones, provided that the minimum value of $RMSE_P$ is reached for the combined set of windows. And finally the search for the minimum value of $RMSE_P$, depending on the number of windows, determines the desired set of spectral variables for scmwiPLS. Figure 3 shows the dependence of the $RMSE_P$ on the number of combined windows in scmwiPLS. The minimum root mean square error of prediction equals 0.03 % and corresponds to the combination of 38 windows with 7 variables or 266 spectral variables.

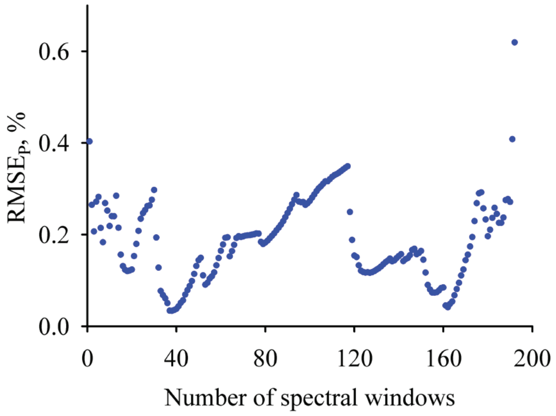

Figure 3 – Dependence of the $RMSE_P$ on the number of combined windows in scmwiPLS

Figure 4 shows dependence of concentration of sunflower oil predicted by the scmwiPLS on its measured concentration in a binary mixture of sunflower and flaxseed oils for training and test samplings. It indicates the high quality of the multivariate model with spectral variables selection, which can be characterized by the value of the residual predictive deviation RPD. RPD is equal to the ratio of the standard deviation of the calibrated parameter and $RMSE_P$. RPD exceeds 1000 for the described scmwiPLS model.

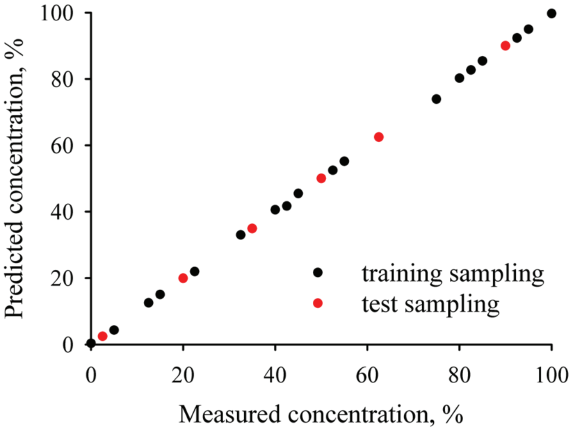

Figure 4 – Concentration of sunflower oil predicted by the scmwiPLS vs measured concentration in a binary mixture of sunflower and flaxseed oils

Figure 5 shows the spectral variables selected using the three investigated methods and the example of the absorbance spectrum of sunflower and flaxseed oils mixture.

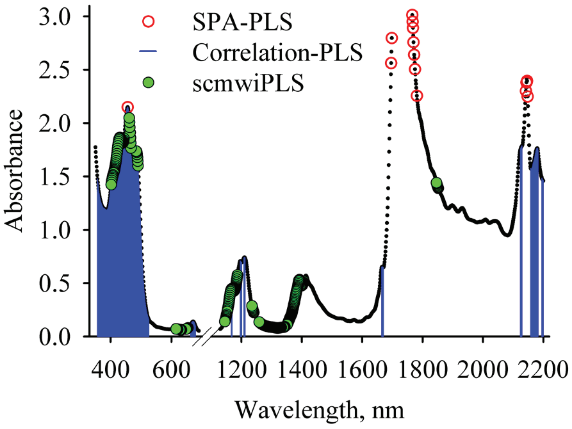

Figure 5 – Absorbance spectra of the mixture of sunflower (12.5 %) and flaxseed (87.5 %) oils and spectral variables selected using the three investigated methods

Spectral variables selection using the ranking of correlation coefficients (239 variables) and the SPA method (14 variables) allows reducing the value of the root mean square error of prediction of sunflower oil concentration from 0.63 % for wideband PLS to 0.50 % and 0.46 %, respectively. These selections are advisable for classical spectroscopy, since the variables selected by both methods are close to the spectral features of the studied objects. The spectral variables selection by scmwiPLS method (266 variables) is less consistent with classical spectroscopy, since a significant part of the selected variables does not describe the characteristic features of the studied spectra, but allows RMSE_P to be reduced by more than an order of magnitude to 0.03 %. Thus, it can be noted that an increase in the calibration accuracy is achieved by using a formal method of variables selection, a feature of which is the use of narrow spectral intervals instead of separate wavelengths.

Conclusion

On the example of the calibration of the concentration of unrefined sunflower oil, considered as a falsified flaxseed oil, it was confirmed that the spectral variables selection is a necessary and important part of multivariate models to improve the accuracy.

It was found that from the considered methods applied to the projection to latent structures of the absorbance spectra for calibrating the concentration of sunflower oil in a mixture with flaxseed oil, a smaller root mean square error of prediction (0.03 %) is achieved for searching combination moving window method in comparison with the successive projection algorithm (0.46 %) and the ranking of spectral variables by the correlation coefficient (0.50 %).

Acknowledgments

This work was supported by the Belarusian Republican Foundation for Fundamental Research grant (project No. F19-060).

References

1. Valdés A., Beltrán A., Mellinas C., Jiménez A., Garrigós M.C. Analytical methods combined with multivariate analysis for authentication of animal and vegetable food products with high fat content. *Trends in Food Science & Technology*, 2018, vol. 77, pp. 120–130.
DOI: 10.1016/j.tifs.2018.05.014

2. Milanez K., Nóbrega T., Nascimento D., Insausti M., Band B., Pontes M. Multivariate modeling for detecting adulteration of extra virgin olive oil with soybean oil using fluorescence and UV–Vis spectroscopies: A preliminary approach. *Food Science and Technology*, 2017, vol. 85, pp. 9–15.
DOI: 10.1016/j.lwt.2017.06.060

3. Lima T., Musso M., Menezes D. Using Raman spectroscopy and an exponential equation approach to detect adulteration of olive oil with rapeseed and corn oil. *Food Chemistry*, 2020, vol. 333, pp. 127454.
DOI: 10.1016/j.foodchem.2020.127454

4. Li Y., Xiong Y., Min S. Data fusion strategy in quantitative analysis of spectroscopy relevant to olive oil adulteration. *Vibrational Spectroscopy*, 2019, vol. 101, pp. 20–27.
DOI: 10.1016/j.vibspec.2018.12.009

5. Elzey B., Pollard D., Fakayode S. Determination of adulterated neem and flaxseed oil compositions by FTIR spectroscopy and multivariate regression analysis. *Food Control*, 2016, vol. 68, pp. 303–309.
DOI: 10.1016/j.foodcont.2016.04.008

6. De Souza L., De Santana F., Gontijo L., Mazivila S., Neto W. Quantification of adulterations in extra virgin flaxseed oil using MIR and PLS. *Food Chemistry*, 2015, vol. 182, pp. 35–40.
DOI: 10.1016/j.foodchem.2015.02.081

7. Khodasevich M.A., Borisevich D.A. Identification of Flax Oil by Linear Multivariate Spectral Analys. *Journal of Applied Spectroscopy*, 2019, vol. 86, no. 6, pp. 880–884.
DOI: 10.1007/s10812-020-00929-z

8. Bro R., Smilde A.K. Principal component analysis. *Analytical Methods*, 2016, vol. 6, pp. 2812–2831.
DOI: 10.1039/C3AY41907J

9. Geladi P., Kowalski B.R. Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta*, 1986, vol. 185, pp. 1–17.
DOI: 10.1016/0003-2670(86)80028-9

10. Soares S.F.C., Gomes A.A., Araujo M.C.U., Filho A.R.G., Galvão R.K.H. The successive projections algorithm. *TrAC Trends in Analytical Chemistry*, 2013, vol. 42, pp. 84–98.
DOI: 10.1016/j.trac.2012.09.006

11. Xiaobo Z., Jiewen Z., Malcolm J.W. Povey, Holmes M., Hanpin M. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, 2010, vol. 667, pp. 14–32.
DOI: 10.1016/j.aca.2010.03.048

12. Li Z., Wang J. , Xiong Y., Li Z., Feng S. The determination of the fatty acid content of sea buckthorn seed oil using near infrared spectroscopy and variable selection methods for multivariate calibration. *Vibrational Spectroscopy*, 2016, vol. 84, pp. 24–29.
DOI: 10.1016/J.VIBSPEC.2016.02.008

13. Wang L., Lee F., Wang X., He Y. Feasibility study of quantifying and discriminating soybean oil adulteration in camellia oils by attenuated total reflectance MIR and fiber optic diffuse reflectance NIR. *Food Chemistry*, 2006, vol. 95, pp. 529–536.
DOI: 10.1016/j.foodchem.2005.04.015

14. Khodasevich M.A., Aseev V.A. Selection of Spectral Variables and Improvement of the Accuracy

of Calibration of Temperature by Projection onto Latent Structures Using the Fluorescence Spectra of Yb^{3+}:CaF_2. *Optics and Spectroscopy*, 2018, no. 124, pp. 713–717.
DOI: 10.1134/S0030400X18050089

15. Nawar S., Mouazen A.M. Optimal sample selection for measurement of soil organic carbon using online vis-NIR spectroscopy. *Computers and Electronics in Agriculture*, 2018, vol. 151, pp. 469–477.
DOI: 10.1016/j.compag.2018.06.042

16. Li Y., Fang T., Zhu S., Huang F., Chen Zh., Wang Y. Detection of olive oil adulteration with waste cooking oil via Raman spectroscopy combined with iPLS and SiPLS. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2018, vol. 189, pp. 37–43.
DOI: 10.1016/j.saa.2017.06.049