УДК 535.343.32, 543.421/.424

КЛАССИФИКАЦИЯ ГЕОГРАФИЧЕСКОГО ПРОИСХОЖДЕНИЯ ЛЕКАРСТВЕННЫХ ТРАВ С ПОМОЩЬЮ МНОГОПАРАМЕТРИЧЕСКОГО СПЕКТРАЛЬНОГО АНАЛИЗА

П. С. КОЛОДОЧКА[1)]***, М. А. ХОДАСЕВИЧ***[1)]

[1)]*Институт физики им. Б. И. Степанова НАН Беларуси, пр. Независимости, 68, 220072, г. Минск, Беларусь*

Аннотация. На примере ромашки аптечной, изготовленной в России и Беларуси, проведена классификация географического происхождения и производителя лекарственных трав методами многопараметрического анализа спектров оптической плотности 70 % спиртовых настоек в диапазоне длин волн 230–2600 нм. Для построения классификационных моделей применялись метод главных компонент, метод построения деревьев классификации и метод выбора спектральных переменных. Метод главных компонент позволяет существенно уменьшить размерность пространства признаков, в котором осуществляется построение деревьев классификации. Максимальное количество рассматриваемых главных компонент ограничено величиной 10, что дает возможность описать более 0,999 общей дисперсии измеренных спектров. Деревья классификации при проведении десятикратной кроссвалидации идентифицируют страну происхождения образцов в четырехмерном пространстве и производителя в трехмерном пространстве главных компонент широкополосных спектров оптической плотности с точностью более 0,93. Ранжирование спектральных переменных в порядке уменьшения модуля среднего отклонения оптической плотности от усредненной величины позволяет повысить точность классификационных моделей. Достоверная классификация географического происхождения ромашки аптечной достигается в пространстве главных компонент 20 из 2623 переменных, имеющихся в широкополосных спектрах. Точность классификации производителя была повышена до 0,94 за счет выбора 14 спектральных переменных.

Ключевые слова: спектральный анализ; метод главных компонент; дерево классификации; выбор спектральных переменных; лекарственное растительное сырье.

Образец цитирования:
Колодочка ПС, Ходасевич МА. Классификация географического происхождения лекарственных трав с помощью многопараметрического спектрального анализа. *Журнал Белорусского государственного университета. Физика.* 2024;3:10–16.
EDN: WPFCRK

For citation:
Kolodochka PS, Khodasevich MA. Classification of the geographical origin of medicinal herbs using multivariate spectral analysis. *Journal of the Belarusian State University. Physics.* 2024;3:10–16. Russian.
EDN: WPFCRK

Авторы:
Полина Сергеевна Колодочка – младший научный сотрудник центра «Диагностические системы».
Михаил Александрович Ходасевич – доктор физико-математических наук, доцент; главный научный сотрудник центра «Диагностические системы».

Authors:
Polina S. Kolodochka, junior researcher at the centre «Diagnostic systems».
kolodochka.p@gmail.com
Mikhail A. Khodasevich, doctor of science (physics and mathematics), docent; chief researcher at the centre «Diagnostic systems».
m.khodasevich@tut.by

CLASSIFICATION OF THE GEOGRAPHICAL ORIGIN OF MEDICINAL HERBS USING MULTIVARIATE SPECTRAL ANALYSIS

***P. S. KOLODOCHKA*[a], *M. A. KHODASEVICH*[a]**

[a]*B. I. Stepanov Institute of Physics, National Academy of Sciences of Belarus, 68 Niezaliezhnasci Avenue, Minsk 220072, Belarus*

Corresponding author: P. S. Kolodochka (kolodochka.p@gmail.com)

Abstract. Classification of the geographical origin and manufacturer of medicinal herbs was carried out by multivariate analysis of the optical density spectra of 70 % alcohol tinctures in the wavelength range 230–2600 nm using the example of chamomile from Russia and Belarus. Principal component analysis, classification and regression tree method, and spectral variable selection were used to build the models. The principal component analysis allows one to significantly reduce the dimension of the feature space. Classification and regression trees are being constructed in it. The maximum number of principal components considered is limited to 10, which made it possible to describe more than 0.999 of the total dispersion of the measured spectra. Classification and regression trees with tenfold cross-validation classify the country of origin of samples in a four-dimensional space and the manufacturer in a three-dimensional space of the principal components of broadband optical density spectra with an accuracy of more than 0.93. Ranking the spectral variables in decreasing order of the absolute value of the average deviation of optical density from the average value made it possible to improve the accuracy of classification models. A reliable classification of the geographical origin of chamomile is achieved in the space of principal components of 20 variables out of 2623 available in the broadband spectra. The manufacturer's classification accuracy was improved to 0.94 by selecting 14 spectral variables.

Keywords: spectral analysis; principal component analysis; classification and regression tree; spectral variable selection; medicinal herbs.

Введение

Широкое использование лекарственного растительного сырья (ЛРС) на протяжении многих столетий является основой народной медицины. Научные исследования ЛРС ограничены отсутствием общепринятой исследовательской методологии для оценки нетрадиционной медицины [1]. Для определения качества и подлинности ЛРС при качественном и количественном анализе отдельных трав или многокомпонентных препаратов, как правило, используются один или два фармакологически активных компонента. Такая оценка не дает полного представления о ЛРС, поскольку за терапевтический эффект может отвечать множество активных компонентов, которые трудно или невозможно разделить. Кроме того, набор и содержание химических компонентов, входящих в состав ЛРС, могут варьироваться в зависимости от сезона сбора, географического происхождения, процессов заготовки и многих других факторов. В связи с этим для идентификации ЛРС можно использовать «отпечаток пальца» [2] – характерный профиль, который отражает сложный химический состав анализируемого образца и может быть получен с помощью хроматографических, спектроскопических или иных методов. Этот профиль должен характеризоваться фундаментальными признаками сходства и различия. С помощью «отпечатков пальцев» можно с некоторой вероятностью определять подлинность ЛРС и идентифицировать его [3–5], даже если набор и (или) концентрации характерных компонентов отличаются для разных образцов.

Целью проводимого исследования являются разработка модели классификации, основанной на применении «отпечатков пальцев» в ультрафиолетовой, видимой и ближней инфракрасной абсорбционной спектроскопии с использованием метода главных компонент (*principal component analysis*, PCA) и метода построения деревьев классификации (*classification and regression tree*, CART), для определения географического происхождения и производителя ЛРС на примере ромашки аптечной и повышение точности классификации с помощью выбора спектральных переменных.

Материалы и методы исследования

Спектры оптической плотности 70 % спиртовых настоек образцов ромашки аптечной были зарегистрированы на спектрофотометре Shimadzu UV-3101PC (Япония) со спектральной шириной щели 1 нм в диапазоне длин волн 230–480 нм с шагом 0,5 нм и в диапазоне длин волн 480–2600 нм с шагом 1 нм. Выбранная ширина щели на порядок меньше характерных ширин особенностей спектров рассматриваемых объектов. Для проведения исследований использовалась ромашка аптечная двух производителей из Беларуси (Могилёвская и Витебская области; суммарно 38 образцов) и двух производителей из России (Тверь и Алтайский край; суммарно 35 образцов).

В качестве метода исследований применялась оптическая абсорбционная спектроскопия, объектами исследования являлись спиртоводные настойки лекарственного сырья. Для анализа спектров использовались методы PCA [6] и CART [7].

В данном исследовании метод PCA применяется для анализа информации, выявления выбросов и уменьшения размерности пространства признаков, в котором будет проводиться классификация образцов. Вместо исходного множества спектральных переменных набор данных может быть описан с использованием небольшого количества главных компонент без значительной потери данных. Подробно основы метода PCA изложены в статье [8].

В настоящей работе дерево классификации строится в пространстве найденных главных компонент [9–12]. Используется следующий наиболее часто встречающийся алгоритм реализации метода CART. В пространстве главных компонент все образцы разбиваются на две группы различным образом. Далее выбирается вариант разбиения, при котором максимальное количество образцов одного класса попадают в одну группу (это первый узел дерева). Дальнейшее разбиение продолжается подобным образом до тех пор, пока не будет достигнуто ограничивающее условие. В данном случае таким условием является ухудшение точности классификационной модели при проведении десятикратной кросс-валидации.

Результаты и их обсуждение

Перед применением метода PCA выполнялась предварительная обработка спектров посредством центрирования. После этого проводилось уменьшение размерности матрицы исходных данных (2623 спектральных переменных) с помощью PCA. На этапе применения PCA учитывается избыточное количество главных компонент, ограниченное величиной 10. На рис. 1 показано, что 10 главных компонент описывают 0,999 2 общей дисперсии данных. На этапе построения классификационных моделей выбираются компоненты, которые являются наиболее значимыми для определения страны происхождения и производителя лекарственного средства. Ниже показано, что количество таких компонент не превышает 5.

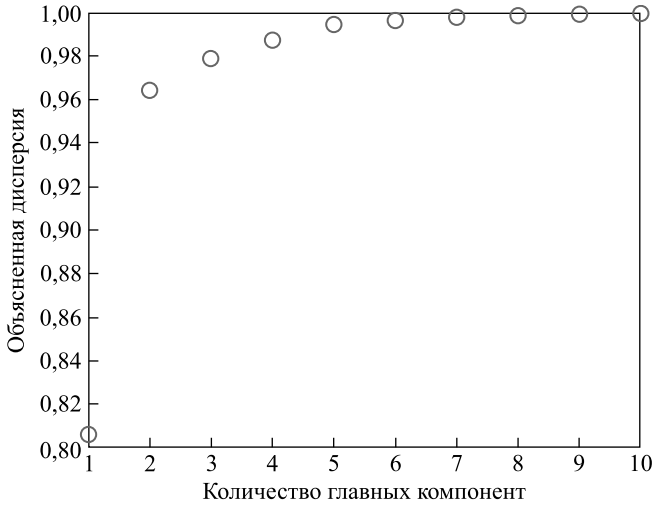

Рис. 1. Объясненная дисперсия спектров оптической плотности 73 настоек ромашки аптечной в зависимости от количества главных компонент

Fig. 1. Explained dispersion of the optical density spectra of 73 chamomile tinctures depending on the number of principal components

На рис. 2 изображены счета в пространстве наиболее информативных первой (PC1) и второй (PC2) главных компонент, которые суммарно описывают более 0,96 дисперсии данных. Видно, что на двумерном графике два образца ЛРС одного из российских производителей попадают в кластеры, состоящие из продукции других производителей. При этом с точки зрения определения страны происхождения значимым является попадание в кластеры, состоящие из продукции белорусских производителей. Попадание обозначенного кружком образца в кластер продукции, обозначенной звездочкой, не сказывается на точности определения страны происхождения, а существенно только при определении региона производства в пределах одной страны.

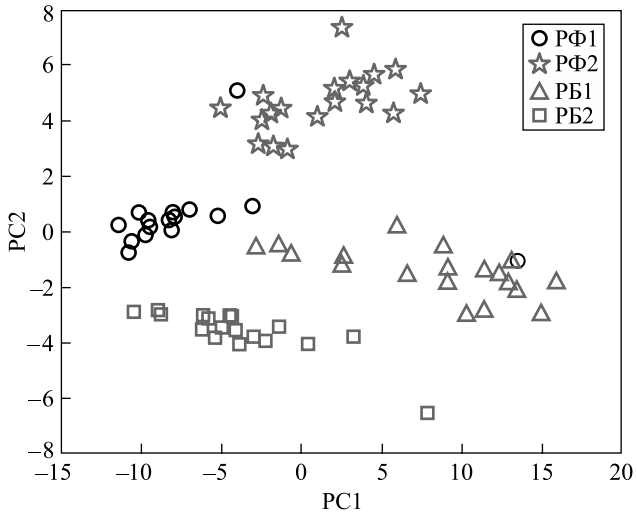

Рис. 2. Счета спектров оптической плотности настоек ромашки аптечной в пространстве первой и второй главных компонент (РФ1, РФ2 – российские производители; РБ1, РБ2 – белорусские производители)

Fig. 2. First and second principal components scores of the optical density spectra of chamomile tinctures (РФ1, РФ2 – Russian manufacturers; РБ1, РБ2 – Belarusian manufacturers)

Классификационное дерево было построено на основе полученных 10 главных компонент с применением десятикратной кросс-валидации. Оно представлено на рис. 3. Точность определения географического происхождения ромашки аптечной составила более 0,93.

Из рис. 3 видно, что для построения модели использовались только первая (PC1), вторая (PC2), четвертая (PC4) и пятая (PC5) главные компоненты. Можно сделать вывод о том, что именно они несут в себе информацию, наиболее полезную для классификации географического происхождения рассматриваемого ЛРС методом CART в пространстве главных компонент широкополосных спектров оптической плотности 70 % спиртовых настоек. В пространстве тех же главных компонент была построена модель CART для классификации ромашки аптечной по производителю (рис. 4), точность которой при десятикратной кросс-валидации превышает 0,93.

Классификационное дерево для определения производителя ромашки аптечной может также выполнять функцию определения географического происхождения образцов. Модели на рис. 3 и 4 показывают вариативность классификационных деревьев: различные модели могут характеризоваться одинаковой точностью. Однако классификация географического происхождения ромашки аптечной является более универсальной, поскольку не ограничена четырьмя производителями, но требует дополнительной валидации.

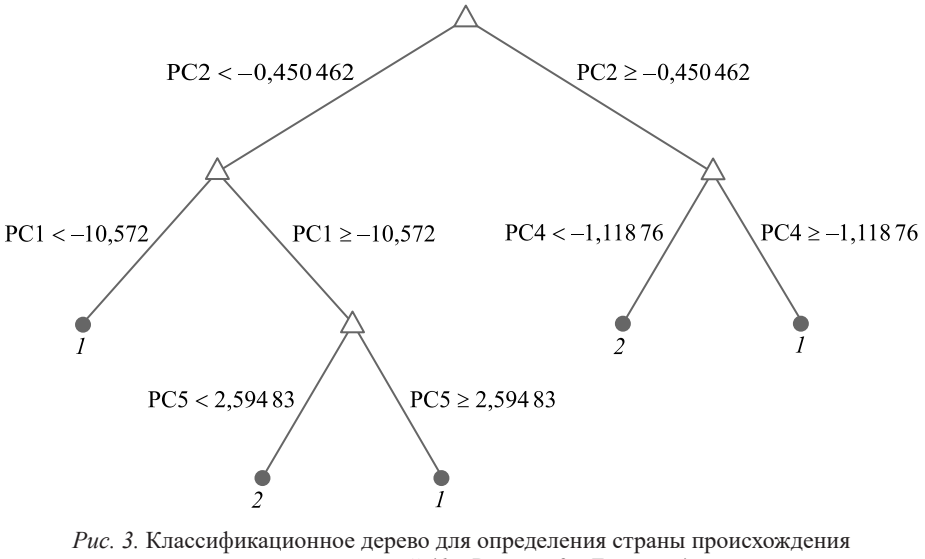

Рис. 3. Классификационное дерево для определения страны происхождения ромашки аптечной (*1* – Россия; *2* – Беларусь)

Fig. 3. Classification tree for determining the country origin of chamomile (*1* – Russia; *2* – Belarus)

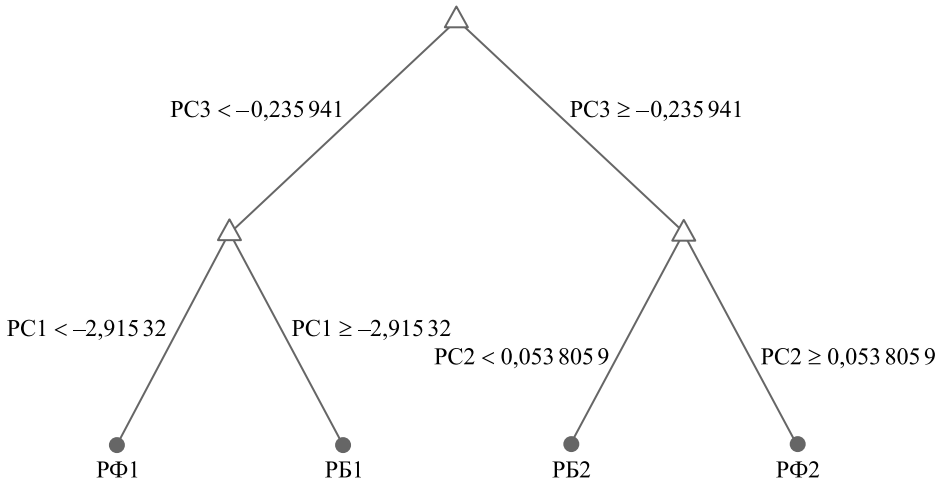

Рис. 4. Классификационное дерево для определения производителя ромашки аптечной

Fig. 4. Classification tree for determining the manufacturer of chamomile

Для повышения точности классификации ко всему спектральному диапазону был применен метод выбора спектральных переменных по значению модуля среднего по всем образцам отклонения оптической плотности от усредненной величины (рис. 5). Количество выбранных спектральных переменных должно превышать количество рассматриваемых главных компонент. По этой причине метод PCA был реализован для пространств спектральных переменных размерности от 11 до 2623, т. е. от 11 спектральных переменных, характеризующихся наибольшими величинами сортирующего параметра, до всего измеренного спектра в диапазоне длин волн 230–2600 нм, который был упорядочен по убыванию модуля среднего отклонения оптической плотности от усредненной величины.

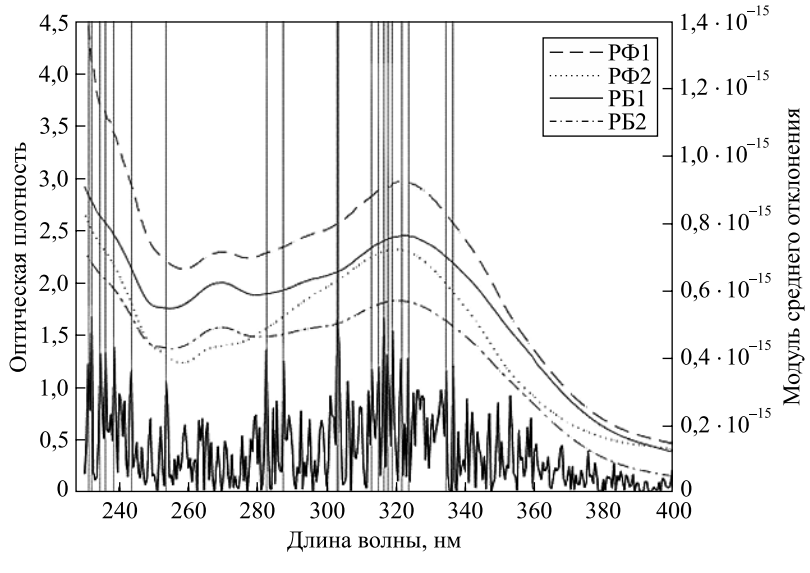

Рис. 5. Модуль среднего отклонения оптической плотности
от усредненной величины (график черного цвета),
спектры оптической плотности 70 % спиртовых настоек
образцов ромашки аптечной и выбранные для проведения классификации страны
происхождения 20 спектральных переменных (вертикальные линии)

Fig. 5. Absolute value of the average deviation
of optical density from the average value (black graph),
optical density spectra of 70 % alcohol tinctures
of chamomile samples and 20 spectral variables (vertical lines)
selected for country origin classification

Существенно возросшие потребности вычислительных ресурсов, по сравнению с классификацией по широкополосным спектрам, привели к ограничению размерности пространства главных компонент для построения классификационного дерева величиной не более 3. Классификационные модели были

разработаны для всех комбинаций 3 из 10 главных компонент, построенных в пространствах выбираемых спектральных переменных. Таким образом, одновременно решались две задачи. Первая задача заключалась в выборе спектральных переменных из их упорядоченного множества и построении десятимерного пространства главных компонент. Вторая задача состояла в построении калибровочной модели максимальной точности в пространстве главных компонент размерности не более 3 из 10 рассматриваемых главных компонент. На рис. 5 представлены выбранные для построения оптимальной модели 20 спектральных переменных вместе со спектрами оптической плотности случайных образцов ромашки аптечной из каждой группы. Дерево, наиболее точно классифицирующее географическое происхождение данного ЛРС, представлено на рис. 6. Классификационное дерево в двумерном пространстве третьей (PC3) и четвертой (PC4) главных компонент, построенном по 20 выбранным спектральным переменным, характеризуется достоверностью классификации рассмотренных образцов ромашки аптечной при проведении десятикратной кросс-валидации.

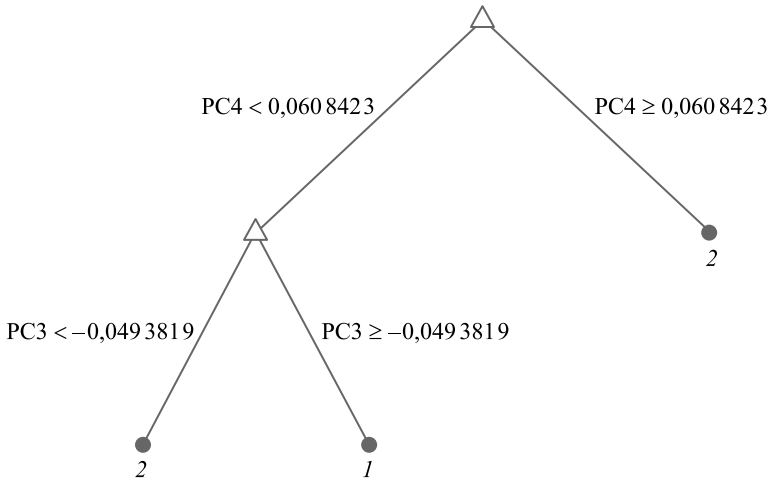

Рис. 6. Классификационное дерево для определения страны происхождения ромашки аптечной (*1* – Россия; *2* – Беларусь) с выбором спектральных переменных

Fig. 6. Classification tree for determining the country origin of chamomile (*1* – Russia; *2* – Belarus) with spectral variable selection

Аналогичный метод выбора спектральных переменных был применен и для улучшения модели классификации ромашки аптечной по производителю. Наилучшая точность (более 0,94) была достигнута в пространстве второй (PC2), третьей (PC3) и пятой (PC5) главных компонент (рис. 7) при выборе 14 спектральных переменных (рис. 8). В этом случае удалось добиться незначительного повышения точности классификации. Такая особенность согласуется с более высокой сложностью задачи классификации производителя по сравнению с задачей классификации географического происхождения.

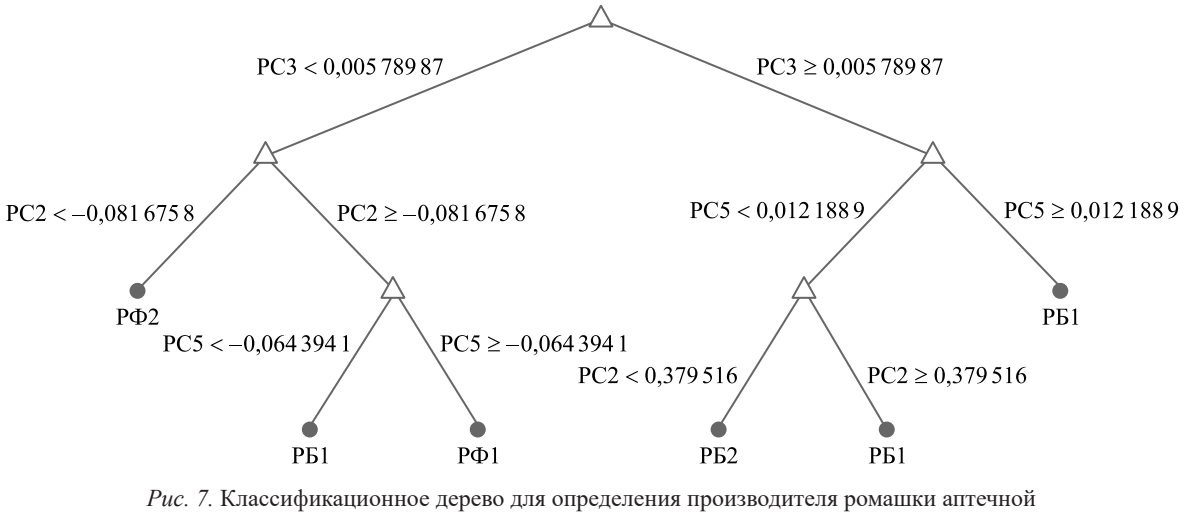

Рис. 7. Классификационное дерево для определения производителя ромашки аптечной

Fig. 7. Classification tree for determining the manufacturer of chamomile

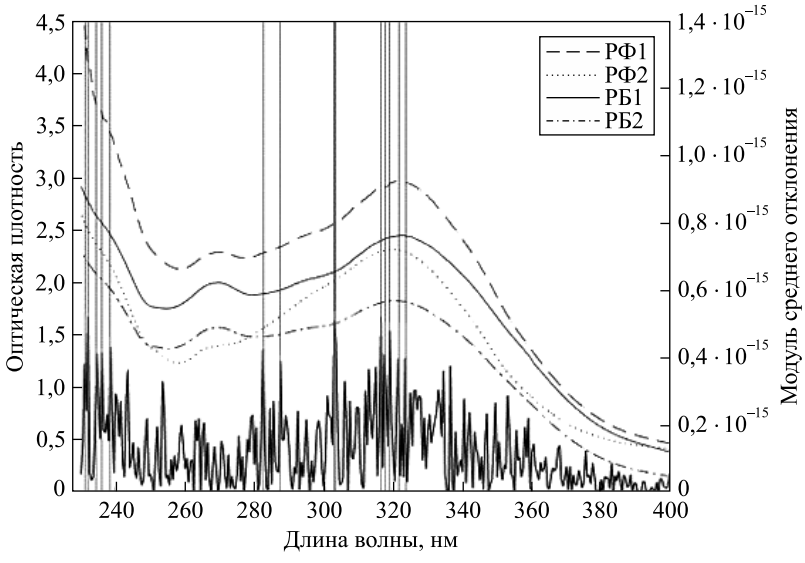

Рис. 8. Модуль среднего отклонения оптической плотности от усредненной величины (график черного цвета), спектры оптической плотности 70 % спиртовых настоек образцов ромашки аптечной и выбранные для проведения классификации производителя 14 спектральных переменных (вертикальные линии)

Fig. 8. Absolute value of the average deviation of optical density from the average value (black graph), optical density spectra of 70 % alcohol tinctures of chamomile samples and 14 spectral variables (vertical lines) selected for manufacturer classification

Заключение

В работе продемонстрированы возможности применения методов многопараметрического анализа для построения моделей классификации географического происхождения и производителя ЛРС по спектрам оптической плотности спиртовых настоек.

Использование метода выбора спектральных переменных по уменьшению модуля средней по всем образцам величины разброса оптической плотности для построения маломерного пространства главных компонент и применения в нем методов кластерного анализа позволило достичь достоверной классификации образцов ромашки аптечной российского и белорусского производства. Достигнутая точность определения производителя данного лекарственного сырья составила 0,94, что достаточно для практического применения.

Библиографические ссылки

1. Liang Y-Z, Xie P, Chan K. Quality control of herbal medicines. *Journal of Chromatography B.* 2004;812(1–2):53–70. DOI: 10.1016/j.jchromb.2004.08.041.

2. Noviana E, Indrayanto G, Rohman A. Advances in fingerprint analysis for standardization and quality control of herbal medicines. *Frontiers in Pharmacology.* 2022;13:853023. DOI: 10.3389/fphar.2022.853023.

3. Wang P, Yu Z. Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: a review. *Journal of Pharmaceutical Analysis.* 2015;5(5):277–284. DOI: 10.1016/j.jpha.2015.04.001.

4. Klein LC Jr, de Souza MR, Viaene J, Bresolin TMB, de Gasper AL, Henriques AT, et al. Quality control of herbal medicines: from traditional techniques to state-of-the-art approaches. *Planta Medica.* 2021;87(12–13):964–988. DOI: 10.1055/a-1529-8339.

5. Chen R, Liu F, Zhang C, Wang W, Yang R, Zhao Y, et al. Trends in digital detection for the quality and safety of herbs using infrared and Raman spectroscopy. *Frontiers in Plant Science.* 2023;14:1128300. DOI: 10.3389/fpls.2023.1128300.

6. Drivelos SA, Georgiou CA. Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. *Trends in Analytical Chemistry.* 2012;40:38–51. DOI: 10.1016/j.trac.2012.08.003.

7. Resce G, Vaquero-Piñeiro C. Predicting agri-food quality across space: a machine learning model for the acknowledgment of geographical indications. *Food Policy.* 2022;112:102345. DOI: 10.1016/j.foodpol.2022.102345.

8. Li S, Yu X, Zhen Z, Huang M, Lu J, Pang Y, et al. Geographical origin traceability and identification of refined sugar using UPLC-QTof-MS analysis. *Food Chemistry.* 2021;348:128701. DOI: 10.1016/j.foodchem.2020.128701.

9. Bro R, Smilde AK. Principal component analysis. *Analytical Methods.* 2014;6(9):2812–2831. DOI: 10.1039/C3AY41907J.

10. Loh W-Y. Fifty years of classification and regression trees. *International Statistical Review.* 2014;82(3):329–348. DOI: 10.1111/insr.12016.

11. Mishra S, Datta-Gupta A. *Applied statistical modeling and data analytics: a practical guide for the petroleum geosciences.* [S. l.]: Elsevier; 2018. Chapter 5, Multivariate data analysis; p. 97–118. DOI: 10.1016/B978-0-12-803279-4.00005-5.

12. Kolodochka PS, Khodasevich MA. Classification of sugar types by UV-VIS-NIR spectroscopy and multivariate analysis. In: *The 12th International conference on photonics and applications (ICPA-12); 2022 September 28 – October 1; Con Dao, Ba Ria – Vung Tau, Vietnam.* [S. l.]: [s. n.]; 2023. p. 244–247.

Получена 11.06.2024 / исправлена 17.07.2024 / принята 28.07.2024.
Received 11.06.2024 / revised 17.07.2024 / accepted 28.07.2024.