# КЛАССИФИКАЦИЯ ШОКОЛАДА НА ОСНОВЕ АНАЛИЗА МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ ПРЕДОБРАБОТАННЫХ СПЕКТРОВ ПРОПУСКАНИЯ ТЕРАГЕРЦОВОГО ДИАПАЗОНА

**М. А. Ходасевич [1*], А. В. Ляхнович [1], H. Eriklioğlu [2]**

*УДК 535.343.2;543.421/422*
*https://doi.org/10.47612/0514-7506-2022-89-2-198-203*

*[1] Институт физики НАН Беларуси,*
*Минск, Беларусь; e-mail: m.khodasevich@ifanbel.bas-net.by*
*[2] Ближневосточный технический университет, Анкара, Турция*

*(Поступила 14 февраля 2022)*

*Продемонстрирована эффективность классификации образцов шоколада по типу и производителю методом "спектрального отпечатка" с использованием спектров пропускания в терагерцовом частотном диапазоне. С целью устранения негативного влияния шума и эффекта Фабри—Перо методом адаптивных итеративно взвешенных наименьших квадратов со штрафом по спектрам определены их базовые линии. Классификация проведена путем построения маломерного пространства главных компонент базовых линий и применения методов кластерного анализа в этом пространстве. Точность и полнота классификации образцов шоколада методами k-среднего, построения классификационного дерева и иерархического кластерного анализа составили 0.85 и 0.83, 0.91 и 0.90, 0.94 и 0.93 соответственно. Для случаев, когда проведение попарной классификации наиболее проблематично, успешно применен метод опорных векторов.*

***Ключевые слова:*** *терагерцовая спектроскопия во временной области, метод главных компонент, базовая линия, кластерный анализ, метод опорных векторов.*

*We demonstrate the efficiency of the chocolate sample classification by type and manufacturer using the "spectral print" method using THz transmission spectra. To suppress the noise and the Fabry–Perot effect, spectra baselines are determined using the adaptive iteratively reweighted penalized least squares (airPLS) method. The classification was carried out by constructing a low-dimensional space of the principal components of the baselines and applying the methods of cluster analysis in this space. The precision and recall values of the classification of chocolate samples by the k-means, classification and regression tree and hierarchical cluster analysis are 0.85 and 0.83, 0.91 and 0.90, 0.94 and 0.93, respectively. The support vector machine is successfully applied to consider two cases where pairwise classification is most problematic.*

***Keywords:*** *terahertz time-domain spectroscopy, principal component analysis, baseline, cluster analysis, support vector machine.*

**Введение.** В настоящее время в мире интенсивно развиваются неразрушающие и бесконтактные методы диагностики состава, качества и аутентичности продукции пищевой промышленности. Интерес к этой области обусловлен необходимостью пищевой безопасности, защиты не только прав потребителя, но и интересов производителя. Наличие доступных и релевантных диагностических приборов и методик может как благотворно отразиться на качестве жизни потребителей, так и снизить расходы на обслуживание производственного оборудования. Бесконтактные, неинвазивные методы должны быть востребованы в условиях неблагоприятной эпидемиологической обстановки.

---

# CHOCOLATE SAMPLE CLASSIFICATION BY PRINCIPAL COMPONENT ANALYSIS OF PREPROCESSED TERAHERTZ TRANSMISSION SPECTRA

**M. A. Khodasevich [1*], A. V. Lyakhnovich [1], H. Eriklioğlu [2]** (*[1] B. I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus, Minsk, Belarus; e-mail: m.khodasevich@ifanbel.bas-net.by; [2] Middle East Technical University, Ankara, Turkey*)

На сегодняшний день известны многочисленные разработки средств диагностики, основанные на спектроскопической технике измерений (см., например, [1, 2]). К традиционным видам УФ, видимой и ИК-спектроскопии [3] сравнительно недавно добавился терагерцовый (ТГц) диапазон частот [4], занимающий определенную нишу в соответствии с присущими ему особенностями и преимуществами. Разработка средств диагностики состава и качества продукции во многом основана на современных способах обработки данных. Так, эффективность спектроскопических методов диагностики значительно повышается в сочетании с многопараметрическими методами анализа данных [3, 5]. Современные подходы позволяют выявлять скрытые зависимости и делать обоснованные выводы об аутентичности продукции, не проводя при этом детального анализа каждого компонента в сложных по составу образцах.

**Эксперимент.** В качестве объекта для анализа выбран шоколад различных производителей (“Коммунарка”, “Спартак” (Беларусь), “Ulker” (Турция), “Волшебница” (Россия) и “AVK” (Украина)) и типов (горький, десертный, молочный). Спектры образцов толщиной 3.0±0.1 мм зарегистрированы на разработанном в Институте физики НАН Беларуси импульсном когерентном спектрометре [6], схема которого представлена на рис. 1. В спектрометре используются фотопроводящие антенны (Teravil, Литва) на основе низкотемпературного GaAs, допированного висмутом, в качестве эмиттера ТГц-импульсов (ТЭ) и детектора (ТД). Они возбуждаются и синхронизируются излучением KYW:Yb-лазера FLINT (Light Conversion, Литва) с длиной волны 1030 нм и длительностью импульса ~80 фс. Пути оптического излучения показаны тонкой сплошной линией, пути ТГц-излучения — жирным пунктиром, штрихпунктир соответствует электрическим соединениям антенн с источником питания $U_{см}$ и регистрирующей аппаратурой. Средняя мощность ТГц-излучения не превышает ~2 мкВт, а соответствующий ток в детекторе составляет единицы наноампер. Применение синхронного детектирования для регистрации слабых сигналов в условиях насыщенного электромагнитными полями окружения позволяет существенно (до 70 дБ) повысить соотношение сигнал/шум. Для этого в системе регистрации объединены модулятор SR540 и синхронный усилитель SR830 (Stanford Research Systems, США). Линия задержки (ЛЗ) обеспечивает сканирование временного профиля ТГц-импульса с требуемой точностью. Зарегистрированный профиль длительностью 100 пс с шагом ~0.067 пс соответствует разрешению 0.01 ТГц при ширине спектра 7.5 ТГц. Сравнительно малый шаг приводит к избыточной ширине регистрируемого диапазона, но позволяет точнее определить фазовые изменения импульса. При этом за счет выбора минимально необходимого разрешения обеспечивается приемлемое время регистрации профиля. Регистрирующее оборудование и исполнительные элементы спектрометра работают под управлением компьютера (ПК), оснащенного специально разработанным программным обеспечением.

Эффективное фокусное расстояние параболических зеркал 156 мм обеспечивает фокусировку ТГц-излучения на образце в пятно диаметром ~5 мм. Это позволяет на площади образца 20×30 мм выбирать для измерения пропускания не менее шести точек. Кроме того, по площади пятна излучения на поверхности шоколада характеристики образцов интегрировались, что нивелировало погрешности изготовления и мелкомасштабную неоднородность некоторых образцов. Отметим, что факторами,

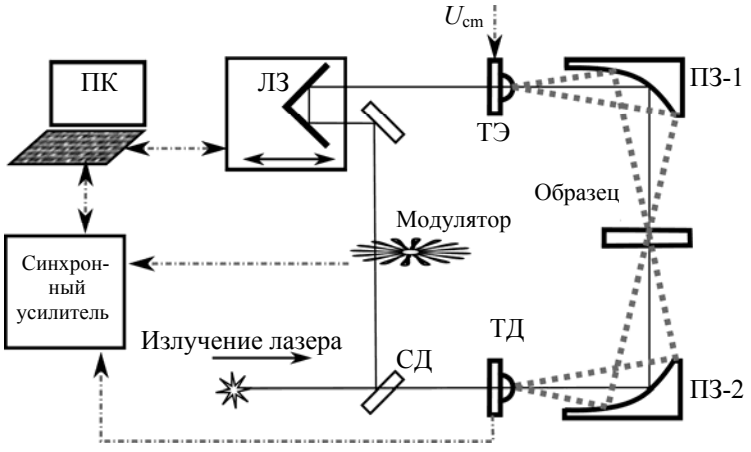


Рис. 1. Схема ТГц-спектрометра: СД — делитель пучка,
ПЗ-1, ПЗ-2 — внеосевые параболические зеркала

затрудняющими извлечение полезной информации из измеренных спектров, являются интерференционные эффекты и наличие паров воды в измерительном тракте [7]. Мы сознательно не принимали мер к удалению паров, в отличие от [8], моделируя условия измерения в обычных производственных либо складских помещениях. В диапазоне 0.2—2.0 ТГц шоколад демонстрирует достаточную прозрачность. Сложность состава образцов существенно затрудняет детальный компонентный анализ. В ряде работ предложены методы диагностики качества продукции на основе ТГц-изображений (безотносительно состава) либо отдельных составляющих продукта, таких как сахар и какао-масло [9].

Цель настоящей работы — демонстрация возможности классификации продукции по типу и производителю методом “спектрального отпечатка” [10].

**Результаты и их обсуждение.** Применение метода главных компонент [11] к спектрам образцов шоколада, обязательная предобработка которых заключается только в центрировании на каждой длине волны, показывает, что первая и вторая главные компоненты описывают 93.5 % суммарной дисперсии данных, а на представленном графике счетов спектров в первую и вторую главные компоненты образцы кластеризуются неудовлетворительно с точки зрения их принадлежности к определенным типам шоколада (рис. 2, *а*). Для улучшения кластеризации в качестве дополнительной предобработки спектров для коррекции базовой линии использован метод адаптивных итеративно взвешенных наименьших квадратов со штрафом (airPLS) [12]. Простейшим и эффективным способом проведения такой коррекции считается полиномиальная аппроксимация, недостатками которой являются требуемое вмешательство пользователя и возможность больших отклонений при низком отношении сигнал/шум. Алгоритм airPLS не требует предварительного экспертного обнаружения пиков или других спектральных особенностей. Метод работает путем итеративного изменения весовых коэффициентов ошибок суммы квадратов разности между подобранной базовой линией и исходным спектром. Весовые коэффициенты определяются адаптивно по разности, найденной на предыдущем шаге итераций базовой линии и исходного спектра. Единственный подгоночный коэффициент метода airPLS, определяемый пользователем, обеспечивает достижение компромисса между точностью аппроксимации спектра и степенью гладкости базовой линии. Если этот коэффициент слишком мал, найденная с помощью алгоритма базовая линия искажает пики путем излишнего учета особенностей спектра, если слишком велик — базовая линия избыточно гладкая. Для рассматриваемых ТГц-спектров подгоночный коэффициент равен 1000.

Описанный вид предобработки ТГц-спектров позволяет избавиться от существенно затрудняющих извлечение полезной информации шумов, вызванных интерференционными эффектами и наличием паров воды в измерительном тракте. В отличие от классического применения метода airPLS для дальнейшего многопараметрического анализа использовались найденные базовые линии, двумерный график счетов которых представлен на рис. 2, *б*. В этом случае первая и вторая главные компоненты базовых линий спектров описывают 95.6 % суммарной дисперсии данных. Видно, что качество кластеризации исследуемых образцов шоколада существенно улучшилось.

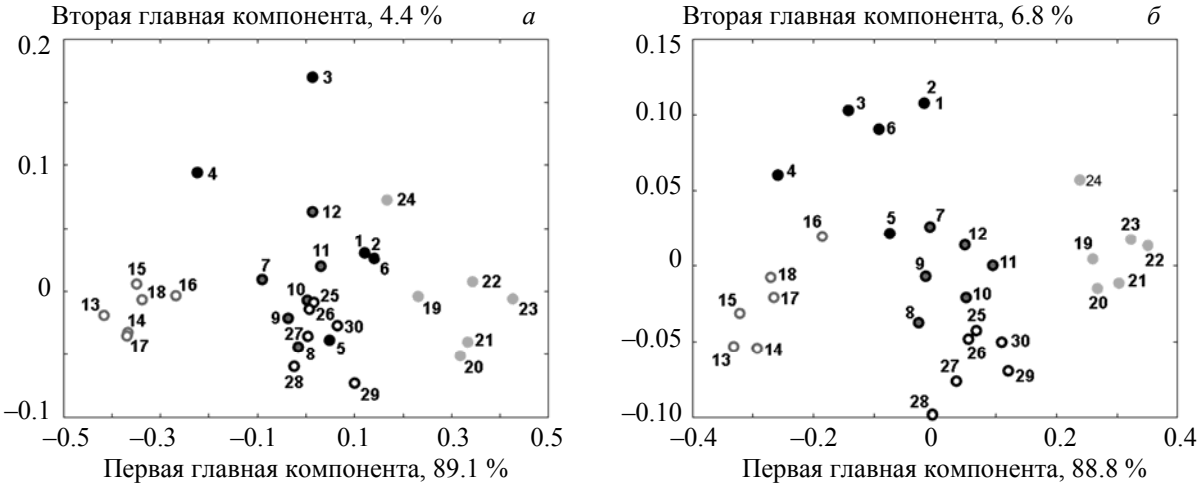


Рис. 2. Графики счетов в первую и вторую главные компоненты спектров пропускания образцов шоколада (*а*) и их базовых линий (*б*)

Далее к счетам в первые две главные компоненты базовых линий применялись методы кластерного анализа: *k*-средних, построение классификационного дерева (CART), иерархический кластерный анализ (HCA) и метод опорных векторов (SVM). Цель этих методов — получение знания о структуре исследуемой выборки путем разбиения ее на группы схожих объектов.

Метод *k*-средних [13] заключается в разделении выборки исследуемых объектов на *k* кластеров и отнесении объекта к кластеру с минимальным расстоянием до центроида. В результате кластеризации методом *k*-средних пять образцов шоколада (4, 5, 10, 11 и 12) на рис. 2, *б* ошибочно отнесены к другим классам. Усредненная по классам точность *P* представляет собой отношение истинно положительных (*TP*) решений классификатора к сумме истинно положительных и ложноположительных (*FP*) решений и составляет 0.85. Усредненная полнота классификации *R* является отношением истинно положительных решений к сумме истинно положительных и ложноотрицательных (*FN*) и составляет 0.83. Для характеризации качества классификации применим также гармонический средний между *P* и *R* параметр $F_1$ [14]:

$$F_1 = 2\frac{P\,R}{P+R},\ P = \frac{TP}{TP+FP},\ R = \frac{TP}{TP+FN}.$$

Для метода *k*-средних $F_1 = 0.84$.

Алгоритм кластеризации CART [15] находит структурированный набор решений, определяющих правила классификации. CART имеет недостаток возможной сходимости к локальному и не обязательно оптимальному решению. Интерпретируемость такой модели часто является определяющим для ее применения фактором. На рис. 3, *а* представлено дерево решений в двумерном пространстве главных компонент ТГц-спектров, которое приводит к ошибочной классификации трех образцов (1, 2 и 16) и следующим параметрам качества: $P = 0.91$, $R = 0.90$ и $F_1 = 0.90$.

Метод HCA, предполагающий наличие вложенных кластеров различного порядка, рекурсивно находит их агломеративным или дивизимным способом. На первом этапе применяемого агломеративного HCA каждый спектр рассматривается как собственный кластер. Затем происходит объединение похожих пар кластеров на основе расчета выбранной меры расстояния между ними, что приводит к построению иерархии. Недостатки HCA — невозможность уточнения параметров иерархической кластеризации с помощью дополнительной выборки [16] и необходимость знания целевого количества кластеров. На рис. 3, *б* представлена дендрограмма иерархии кластеров, в результате построения которой ошибочно классифицированы образцы 5 и 16. Точность классификации $P = 0.94$, полнота $R = 0.93$, $F_1 = 0.93$.

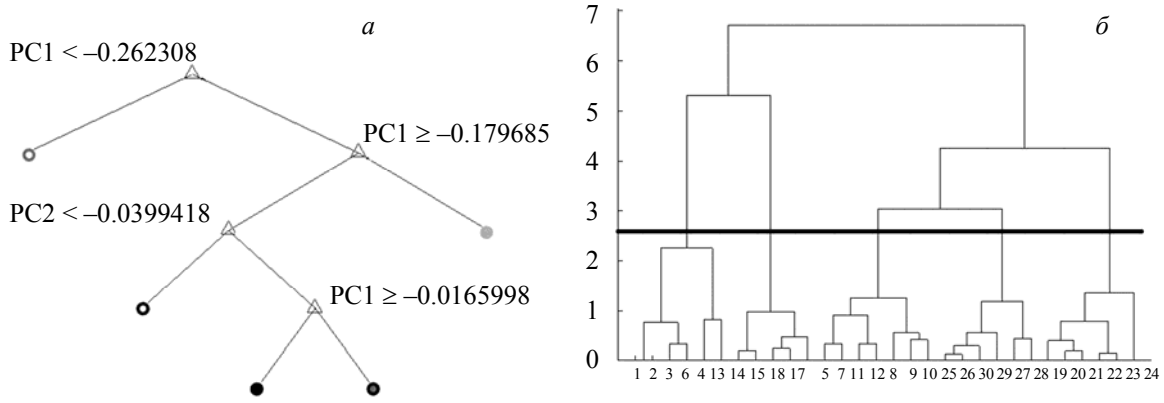


Рис. 3. Построенное методом CART дерево решений (*а*) (символы как на рис. 2) и найденная методом HCA дендрограмма иерархии кластеров (*б*) с выделенным разбиением исследуемой выборки на пять классов

Для случаев, когда проведение попарной классификации наиболее проблематично (первые два класса обозначены на рис. 2 черными кругами под номерами 1—6 и серыми под номерами 13—18, вторые — образцы 1—6 и обозначенные серыми кругами с черной границей образцы 7—12), применен метод SVM — метод получения оптимальной границы двух кластеров в векторном пространстве независимо от вероятностных распределений векторов обучающей выборки. Для линейно не разде-

лимых данных метод SVM позволяет ввести дополнительные переменные, характеризующие величину ошибки в классах и минимизируемый штраф за суммарную ошибку [17], а также применить преобразование пространства переменных. В рассматриваемом случае в пространстве главных компонент применено полиномиальное преобразование третьей степени, позволившее достоверно определить указанные пары классов образцов. На рис. 4 представлены найденные методом SVM границы классов. Окружностями большего радиуса обозначены образцы, являющиеся опорными векторами. Особенность метода состоит в том, что построение границы зависит только от опорных векторов и не зависит от остальных образцов, что делает моделирование более устойчивым при добавлении дополнительных образцов по сравнению, например, с методом *k*-средних.

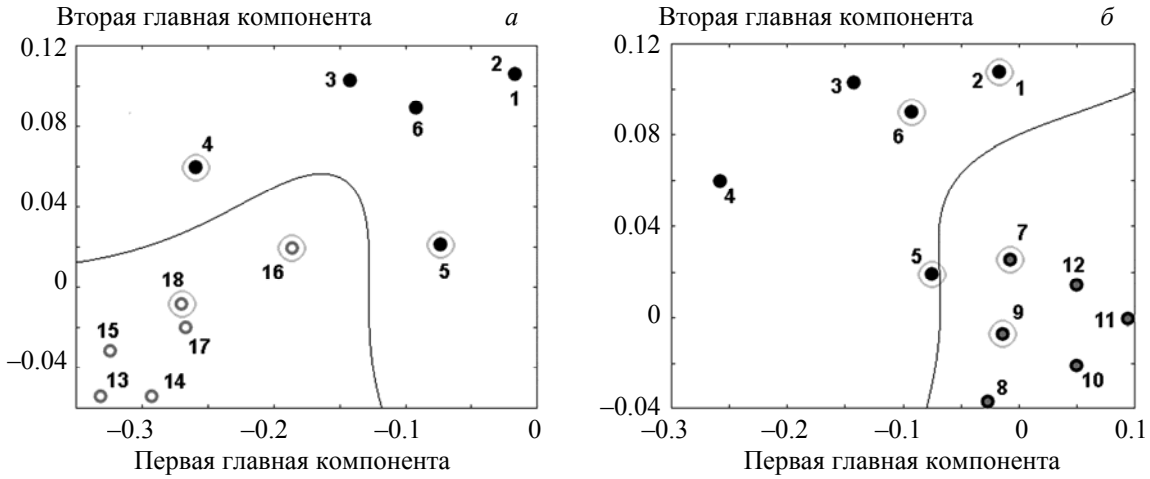


Рис. 4. Найденные методом SVM границы классов образцов 1—6 и 13—18 (*а*) и 1—6 и 7—12 (*б*) в пространстве первой и второй главных компонент

**Заключение.** Несмотря на значительные различия, рассмотренные методы кластеризации опираются на априорную гипотезу компактности расположения объектов в пространстве главных компонент базовых линий терагерцовых спектров пропускания исследуемых образцов — близкие объекты должны относиться к одному кластеру, а различные должны находиться в различных кластерах. Продемонстрирована возможность классификации образцов шоколада по типу и производителю с высокими точностью и полнотой методом “спектрального отпечатка” с помощью измерения спектров пропускания образцов в терагерцовом частотном диапазоне, определения их базовых линий методом адаптивных итеративно взвешенных наименьших квадратов со штрафом, построения маломерного пространства главных компонент базовых линий и применения методов кластерного анализа в этом пространстве.

Работа выполнена при частичной финансовой поддержке в рамках исследовательской и инновационной программы Европейского Союза Horizon 2020 (грант № 101008228).

[1] **C. McVey, C. T. Elliott, A. Cannavan, S. D. Kelly, A. Petchkongkaew, S. A. Haughey.** Trends Food Sci. Technol. B, **118** (2021) 777—790
[2] **A. Arroyo-Cerezo, A. M. Jimenez-Carvelo, A. González-Casado, A. Koidis, L. Cuadros-Rodríguez.** LWT – Food Sci. Technol., **149** (2021) 111822(1—8)
[3] **H. N. Moghaddam, Z. Tamiji, M. A. Lakeh, M. R. Khoshayand, M. H. Mahmoodi.** J. Food Comp. Anal., **107** (2022) 104343
[4] **X. Sun, D. Cui, Y. Shen, W. Li, J. Wang.** Infrared Phys. Technol., **121** (2022) 104018
[5] **I. Magnus, M. Virte, H. Thienpont, L. Smeesters.** Food Control, **130** (2021) 108342(1—10)
[6] **А. М. Гончаренко, Г. В. Синицын, А. В. Ляхнович, В. Л. Малевич.** Сб. науч. тр., IV Конгресс физиков Беларуси, 24—26 апреля, Минск (2013) 82—83
[7] **R. Fastampa, L. Pilozzi, M. Missori.** Phys. Rev. A, **95** (2017) 063831(1—6)
[8] **J. Oblitas, J. Ruiz.** Proceedings, **70**, N 1 (2021) 109(1—6), doi: 10.3390/foods_2020-08029

[9] **S. Weiller, T. Tanabe, Y. Oyama.** World J. Eng. Technol., **6** (2018) 268—274
[10] **R. Ríos-Reina, J. M. Camiña, R. M. Callejón, S. M. Azcarate.** Trends Anal. Chem., **134** (2021) 116121(1—21)
[11] **K. H. Esbensen, P. Geladi.** In: Comprehensive Chemometrics, Eds. S. Brown, R. Tauler, B. Walczak, Elsevier (2009) 211—226
[12] **Z. M. Zhang, S. Chen, Y. Z. Liang.** Analyst, **135**, N 5 (2010) 1138—1146
[13] **P. Govender, V. Sivakumar.** Atm. Poll. Res., **11** (2020) 40—56
[14] **J. Huang, J. Liu, K. Wang, Z. Yang, X. Liu.** Spectrochim. Acta A: Mol. Biomol. Spectrosc., **198** (2018) 198—203
[15] **K. Herberger.** In: Medical Applications of Mass Spectrometry, Eds. K. Vékey, A. Telekes, A. Vertes, Elsevier (2008) 141—169
[16] **T. W. Liao.** Pattern Recognition, **38** (2005) 1857—1874
[17] **Y. Xu, S. Zomer, R. G. Brereton.** Critical Rev. Anal. Chem., **36** (2006) 177—188