https://doi.org/10.35627/2219-5238/2022-30-4-44-53
Оригинальная исследовательская статья

УДК 613.2, 664.66.014

Подходы к алгоритму анализа результатов исследований микро- и макронутриентного состава хлебобулочных изделий. Сообщение первое

Г.Д. Щербаков[1,2], В.В. Бессонов[1]

[1]ФГБУН «Федеральный исследовательский центр питания, биотехнологии и безопасности пищи», Устьинский пр-д, д. 2/14, г. Москва, 109240, Российская Федерация

[2]ФБУЗ «Федеральный центр гигиены и эпидемиологии» Роспотребнадзора, Варшавское ш., д. 19А, г. Москва, 117105, Российская Федерация

Резюме

Введение. Данные химического состава пищевых продуктов являются востребованными для решения многих задач как в медицинской, так и в социальной сфере. Востребованной является разработка механизмов актуализации действующих баз данных химического состава пищевых продуктов, в том числе требуется изменение подходов к получению первичных данных и разработка алгоритмов их обработки.

Цель: разработка алгоритма получения статистически корректных значений средних концентраций и вариабельности основных микро- и макронутриентов в хлебобулочных изделиях.

Материалы и методы. Для разработки и апробации алгоритма использовались данные лабораторных исследований хлебобулочных изделий, выполненные в рамках федерального проекта «Укрепление общественного здоровья» в 2020 году лабораториями Роспотребнадзора.

Результаты. Хорошую разделяющую способность продемонстрировала кластеризация методом k-средних на две группы по показателю содержания жира. Предложен алгоритм генерализации данных, полученных от разных лабораторий, в связи с тем что не представляется возможным провести оценку совокупности ошибок (аналитической, лабораторного персонала, ввода и других). Для оценки результативности каждого этапа и алгоритма в целом использовалась величина отклонения получаемой вариабельности от исходной. В результате обработки этот показатель составил от 5 % для содержания углеводов и до 72 % для содержания жира. Для содержания углеводов, золы, пищевых волокон, витамина B_1, натрия и влажности в обоих кластерах получены статистически значимые различия между обработанными значениями и исходными данными. Данный результат и сопоставимость полученных значений среднего и вариабельности со справочными могут свидетельствовать о корректности работы алгоритма. Для полученных значений содержания жира и белка статистически значимые отличия отсутствуют, но также фиксируется совпадение порядков значений со справочными.

Заключение. Разработанный алгоритм позволил получить актуальные сведения о химическом составе хлебобулочных изделий. Дальнейшие исследования должны быть направлены на апробацию и, в случае необходимости, корректировку алгоритма для всех основных групп пищевых продуктов.

Ключевые слова: качество пищевых продуктов, база данных химического состава пищевых продуктов, цифровая нутрициология, стандартизация данных, обработка результатов лабораторных исследований, классификация пищевых продуктов.

Для цитирования: Щербаков Г.Д., Бессонов В.В. Подходы к алгоритму анализа результатов исследований микро- и макронутриентного состава хлебобулочных изделий. Сообщение первое // Здоровье населения и среда обитания. 2022. Т. 30. № 4. С. 44–53. doi: https://doi.org/10.35627/2219-5238/2022-30-4-44-53

Сведения об авторах:

✉ **Щербаков** Григорий Дмитриевич – начальник отдела социально-гигиенического мониторинга анализа и прогнозирования ФБУЗ «Федеральный центр гигиены и эпидемиологии» Роспотребнадзора, аспирант ФГБУН «Федеральный исследовательский центр питания, биотехнологии и безопасности пищи»; e-mail: sherbakovgrigory@gmail.com; ORCID: https://orcid.org/0000-0002-9046-6837.

Бессонов Владимир Владимирович – д.б.н., заведующий лабораторией химии пищевых продуктов ФГБУН «Федеральный исследовательский центр питания, биотехнологии и безопасности пищи»; e-mail: bessonov@ion.ru; ORCID: https://orcid.org/0000-0002-3587-5347

Информация о вкладе авторов: концепция и дизайн исследования: *Бессонов В.В., Щербаков Г.Д.*; сбор данных: *Щербаков Г.Д.*; анализ и интерпретация результатов: *Щербаков Г.Д.*; литературный обзор: *Щербаков Г.Д.*; подготовка рукописи: *Бессонов В.В., Щербаков Г.Д.* Все авторы ознакомились с результатами работы и одобрили окончательный вариант рукописи.

Соблюдение этических стандартов: данное исследование не требует представления заключения комитета по биомедицинской этике или иных документов.

Финансирование: исследование не имело спонсорской поддержки.

Конфликт интересов: авторы декларируют отсутствие явных и потенциальных конфликтов интересов в связи с публикацией данной статьи.

Статья получена: 04.02.22 / Принята к публикации: 04.04.22 / Опубликована: 29.04.22

Approaches to the Algorithm of Analyzing the Results of Laboratory Testing of Micro- and Macronutrient Content of Bakery Products: Part 1

Grigory D. Shcherbakov,[1,2] Vladimir V. Bessonov[1]

[1]Federal Research Center for Nutrition, Biotechnology and Food Safety, 2/14 Ustyinsky Driveway, Moscow, 109240, Russian Federation

[2]Federal Center for Hygiene and Epidemiology, 19A Varshavskoe Highway, Moscow, 117105, Russian Federation

Summary

Introduction: Data on the chemical composition of food products are important for solving many problems in medical and social spheres. The development of mechanisms for updating existing databases of the chemical composition of foodstuffs, including the need to change approaches to obtaining primary data and develop algorithms of their processing, is in demand.

Objective: To develop an algorithm of obtaining statistically correct values of average concentrations and variability of the main micro– and macronutrients in bakery products.

Materials and methods: To develop and test the algorithm, we used the results of testing bakery products obtained in 2020 within the Federal Project on Public Health Strengthening by the laboratories of the Russian Federal Service for Surveillance on Consumer Rights Protection and Human Wellbeing (Rospotrebnadzor).

Results: A good separating power was demonstrated by *k-means* clustering into two groups by the fat content. An algorithm for generalization of data obtained from different laboratories is proposed due to impossibility to assess the whole aggregate of potential errors related to testing, laboratory personnel, data entry, etc. To assess the effectiveness of each stage and the algorithm as a whole, we used the value of the deviation of the resulting variability from the initial one. As a result of processing, this indicator ranged from 5 % for the carbohydrate content to 72 % for the fat content. For the contents of carbohydrates, ash, dietary fiber, vitamin B1, sodium and moisture in both clusters, statistically significant differences were obtained between the processed and original data. This result and the comparability of the obtained values of the mean and variability with the reference ones may indicate the correctness of the algorithm. There were no statistically significant differences between the obtained values of fat and protein content, but the consistency of the order of values with the reference ones was also recorded.
Conclusion: The developed algorithm made it possible to obtain up-to-date information about the chemical composition of bakery products. Further research should be aimed at testing and, if necessary, adjusting the algorithm for all major food groups.
Keywords: food quality, databases of the chemical composition of food products, digital nutrition, data standardization, laboratory data processing, food classification.

For citation: Shcherbakov GD, Bessonov VV. Approaches to the algorithm of analyzing the results of laboratory testing of micro- and macronutrient content of bakery products: Part 1. *Zdorov'e Naseleniya i Sreda Obitaniya.* 2022;30(4):44–53. (In Russ.) doi: https://doi.org/10.35627/2219-5238/2022-30-4-44-53

Author information:
✉ Grigory D. **Shcherbakov**, Head of the Department of Public Health Monitoring, Analysis and Forecasting, Federal Center for Hygiene and Epidemiology; postgraduate student, Federal Research Center of Nutrition, Biotechnology and Food Safety; e-mail: sherbakovgrigory@gmail.com; ORCID: https://orcid.org/0000-0002-9046-6837.
Vladimir V. **Bessonov**, Dr. Sci. (Biol.), Head of the Laboratory of Food Chemistry, Federal Research Center of Nutrition, Biotechnology and Food Safety; e-mail: bessonov@ion.ru; ORCID: https://orcid.org/0000-0002-3587-5347.
Author contributions: study conception and design: *Bessonov V.V., Shcherbakov G.D.*; data collection: *Shcherbakov G.D.*; analysis and interpretation of results: *Shcherbakov G.D.*; literature review: *Shcherbakov G.D.*; draft manuscript preparation: *Bessonov V.V., Shcherbakov G.D.* Both authors reviewed the results and approved the final version of the manuscript.
Funding: The authors received no financial support for the research, authorship, and/or publication of this article.
Conflict of interest: The authors declare that there is no conflict of interest.
Received: February 4, 2021 / Accepted: April 4, 2022 / Published: April 29, 2022

Введение. Обеспечение здорового питания является одним из ключевых приоритетов для каждой страны, в том числе и для Российской Федерации. Нездоровое питание и низкая физическая активность выделены Всемирной организацией здравоохранения как одни из наибольших вкладов в риски для здоровья во всем мире [1]. Таким образом, и исследование состояния питания населения становится крайне важной задачей.

На самом верхнеуровневом представлении исследования состояния питания можно разложить на два этапа. Первый этап – получение данных о потреблении человеком пищевых продуктов и соответствующих микро- и макронутриентов, второй – сравнение полученных данных с рекомендуемыми значениями. Рекомендации в Российской Федерации установлены двумя основными документами[1,2]. Данные документы проходят процесс актуализации и приведения в соответствие с современными представлениями о здоровом питании [2]. Важным является определение подходов к переходу от фактического к химическому составу рациона. Так, классическим подходом является использование данных справочника химического состава российских пищевых продуктов[3]. Исследования, которые легли в основу данного документа, проводились более чем 40 лабораториями СССР по одинаковым методикам и на схожих продуктах.

Воспроизведение подобного эксперимента в настоящий момент потребует подключение значительных ресурсов различных лабораторий, а также существенных финансовых затрат на приобретение всей постоянно растущей номенклатуры пищевой продукции, что, особенно в период распространения COVID-19, является практически нереализуемой задачей. В том числе даже при работе множества научных учреждений, имеющих целью создание базы данных, требуется серьезная поддержка государства для, например, признания ее в качестве национальной, а также выделение соответствующих информационных ресурсов для поддержания доступности и информационной безопасности [3]. В связи с этим особо значимой становится разработка методологии актуализации данных о химическом составе пищевых продуктов на основе данных регулярных исследований, проводимых Роспотребнадзором[4]. Учитывая, что сопряжение баз данных является непростой задачей, при этом каждая база данных носит национальный характер, применение данных, полученных именно путем химического анализа, считается наиболее корректным и используется для сравнения результатов, полученных математически и непосредственно на реальном объекте [4]. Этот же подход к выбору источников данных позволит оценить изменения в практически реальном времени, что является важным аспектом с точки зрения оценки изменения состояния питания населения и принятия соответствующих мер [5]. Для решения данной задачи становится некорректной и нецелесообразной оценка одного конкретного результата лабораторного исследования. Оценка же базы данных – множества измерений одного и того же показателя в разных продуктах одной группы различными испытательными лабораторными центрами – должна проводиться по некоторым мерам центральной тенденции,

[1] Приказ Министерства здравоохранения Российской Федерации от 19 августа 2016 г. № 614 «Об утверждении Рекомендаций по рациональным нормам потребления пищевых продуктов, отвечающих современным требованиям здорового питания».
[2] МР 2.3.1.0253–21 «Нормы физиологических потребностей в энергии и пищевых веществах для различных групп населения Российской Федерации». Утверждены Главным государственным санитарным врачом РФ 22.07.2021.
[3] Химический состав российских пищевых продуктов: Справочник / под ред. член-корр. МАИ, проф. И.М. Скурихина и академика РАМН, проф. В.А. Тутельяна. М.: ДеЛи принт, 2002. 236 с.
[4] Паспорт федерального проекта «Формирование системы мотивации граждан к здоровому образу жизни, включая здоровое питание и отказ от вредных привычек», реализующийся в рамках исполнения Указа Президента Российской Федерации от 07.05.2018 № 204 «О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года».

https://doi.org/10.35627/2219-5238/2022-30-4-44-53
Оригинальная исследовательская статья

таким как среднее, мода или медиана. Однако получение данных мер и соответствующих им ошибок и диапазонов доверительных интервалов является комплексной задачей, включающей в себя как работу с первичными данными (удаление выбросов, обработка пропущенных значений, поиск недостоверных (подобранных) групп данных), так и оценку точности и правдоподобности полученной меры.

Актуальность данной темы также отражена в одном из направлений Стратегии повышения качества пищевой продукции[5], а именно «Совершенствование и развитие методологической базы для оценки соответствия показателей качества пищевой продукции», содержащем в себе задачу создания базы данных показателей качества пищевой продукции, учитывающих естественную вариабельность энергетической и пищевой ценности.

Несмотря на актуальность вопроса в Российской Федерации, анализ статей за последнее время показал крайне малое внимание к вышеуказанной задаче. Так, в статье, посвященной именно базам данных химического состава [6], рекомендуется рассмотрение процесса создания двух существующих международных программ сбора и компиляции данных (RAOLIROOBOB и EigoRNS), которые реализованы в качестве ресурсов для оказания помощи в создании и поддержании собственных национальных баз данных по пищевой химии. Основа базы данных включает в себя изучаемый продукт, способ его изготовления или приготовления, методы исследования химического состава и прочие сведения. Они отличаются от существующих печатных таблиц химического состава тем, что содержат более подробную информацию о всех соответствующих показателях. Кроме того, так как основой базы являются литературные данные о химическом составе (научные статьи, отчеты и др.), которые часто реализуются в виде электронных ресурсов, то это позволяет постоянно проводить обновление и актуализацию. Такая база продуктов питания может обеспечить обмен данными между странами и легкий доступ к соответствующим изменениям в ней [7].

Разработка базы данных по химическому составу пищевых продуктов по опыту RAOLIROOBOB и EigoRNS представляет собой трехэтапную систему. Вначале осуществляется поиск литературных источников и определение их надежности. Выбор осуществляется на основании рекомендаций, таких как те, которые были разработаны в рамках программ ФАО/INFOODS [8, 9], EuroFIR [10]. На втором этапе необходимо, помимо собственно базы данных о всех интересующих свойствах, производстве, производителях и так далее, составить базу данных источников, откуда вся эта информация была получена и как она в дальнейшем будет обновляться. И на финальном этапе требуется создание справочной базы данных. Это означает переоценку достоверности данных о химическом составе архивной базы данных, расчет статистических параметров образцов и объединение информации для одних и тех же продуктов, обновление недостающих данных. Результатом этого этапа является создание справочной базы данных, которая может быть частично или полностью доступна для публичного использования.

Однако и у такого подхода есть определенная проблема, связанная в первую очередь с объемом входящих данных и, следовательно, ограничениями по применению базы данных. Таблицы и базы данных о составе пищевых продуктов доступны в большинстве стран, однако содержащиеся в них данные неизменно подвергаются критике за то, что они слишком неточны для многих целей. Например, использование таблиц состава пищевых продуктов для расчета потребления питательных веществ отдельными лицами считается слишком ненадежным для клинических исследований и исследований, связанных со здоровьем, поскольку содержание питательных веществ в пищевых продуктах сильно различается.

Точно так же производители пищевых продуктов не могут полагаться на существующие данные о составе для обеспечения точности, необходимой для регулятора, или собственной работы. Пользователям данных о составе пищевых продуктов требуется информация, выходящая за рамки простых значений питательных веществ или компонентов. Им требуются более точные описания пищевых продуктов, включающие происхождение данных. Запрашиваются данные о питательных микроэлементах и различных биологически активных формах питательных веществ, а также дополнительная информация о репрезентативности и качестве существующих данных. Поскольку такие данные и описания недоступны для пользователей в настоящее время, таблицы и базы данных нельзя с этой точки зрения считать полными и адекватными.

Когда в 1950–1960-х годах в Соединенных Штатах и Европе были созданы таблицы и базы данных о составе пищевых продуктов, произошел значительный обмен данными, чтобы можно было составить списки для интерпретации национальных обследований питания. В развивающихся странах масштабы заимствования данных для составления таблиц были еще выше. Большая часть информации о составе пищевых продуктов основана на устаревших технологиях и аналитических методах, которые были усовершенствованы с тех пор, когда впервые были собраны данные. Когда эти базы данных были созданы, разработчики предоставили однозначные результаты для состава питательных веществ продукта питания. Пользователи не были достаточно осведомлены о естественных вариациях в составе пищевых продуктов или различиях в составе продуктов из разных регионов. Описания и вариации не регистрировались большинством учреждений. Таким образом, невозможно определить, повлияли ли на сообщаемые значения факторы, которые, как теперь известно, являются важными. Кроме того, пользователи не могут быть уверены в ассортименте продуктов, представленных этими средними значениями. Таким образом, данные о составе пищевых продуктов недостаточны для нескольких важных целей, включая торговлю пищевыми продуктами, клинические исследования и международную эпидемиологию.

Большинство баз данных не содержат информации о качестве данных. Включенные значения являются теми, которые аналитики, предоставившие данные, считают надежными; редко

[5] «Стратегия повышения качества пищевой продукции в Российской Федерации до 2030 года», утвержденная распоряжением Правительства Российской Федерации от 29 июня 2016 г. № 1364-р.

бывает какая-либо гарантия того, что данные сопоставимы, или для пользователей объясняется сопоставимость данных.

Пользователи должны полагаться на объяснения аналитиков, чтобы знать влияние аналитических методов на данные о составе и судить, достаточно ли точны данные для их целей. Для оценки качества данных был разработан код с тремя категориями, от «ненадежный» до «настолько точный, насколько позволяют современные технологии и методы» [11], хотя до настоящего времени он применялся только для малопитательных веществ.

Значения, выбранные в качестве репрезентативных, будут использоваться для определения того, соответствует ли потребление отдельными лицами рекомендуемым нормам. В результате эти значения влияют на решения, принимаемые дистрибьюторами и производителями продуктов питания. Тем не менее профессионалы, которые рассчитывают индивидуальное потребление питательных веществ, часто не могут решить, какие записи в базе данных представляют фактически съеденную пищу, потому что информация не включает подробностей о переменных факторах, влияющих на состав пищи, таких как условия выращивания, стадия зрелости или рецептура продукта. Такую неопределенность можно свести к минимуму за счет более точного описания источника данных и условий выращивания, хранения и обработки. Разработчики баз данных могут способствовать лучшему принятию решений, предоставляя детали, объясняющие значения пищевых компонентов.

Изменения в аналитических методах привели к новым значениям питательных веществ. Например, данные о содержании витаминов во многих пищевых продуктах были пересмотрены в национальных таблицах после более широкого использования оборудования для жидкостной хроматографии высокого давления (ВЭЖХ). Так, например, повторный анализ значений бета-каротина в Восточной Африке [12] показал, что ранее примененные методы завышали количество этого питательного вещества в пищевых продуктах в два раза. Эти результаты могут повлиять на продовольственную политику в странах Восточной Африки. В этом случае предположения об содержании витамина А в рационе детей резко изменятся в результате изменения данных о составе пищи. Мало того, что проблемы с питанием определены более точно, последующие вмешательства будут более эффективными для спасения детей от необратимого ущерба. Наконец, затраты на вмешательства могут быть рассчитаны более точно.

Поскольку для повторного анализа широко употребляемых пищевых продуктов в развивающихся странах используются более новые методы, разумно ожидать, что аналогичные преимущества будут документально подтверждены и в других случаях. В результате вмешательства обещают быть более эффективными в плане улучшения состояния питания населения.

Также достаточно широко рассматривается проблема вариабельности отдельных показателей в пищевых продуктах, таких как витамин D в молоке [13], антиоксиданты в томатах в зависимости от сезона [14], фенольная кислота в различных бобовых культурах [15].

В качестве продукта для апробации подходов была выбрана группа хлебобулочных изделий, как лежащая в основе пищевой пирамиды [16], так и являющаяся традиционной основой рациона во многих странах [17, 18].

Цель исследования – разработать методы получения статистически достоверных сведений о микро- и макронутриентном составе хлебобулочных изделий на основе данных неспециализированных исследований.

Методы исследования. В качестве базы данных для оценки результативности и корректности разрабатываемых методов была выбрана база результатов исследований качества и безопасности пищевых продуктов, выполненных в рамках федерального проекта «Укрепление общественного здоровья» национального проекта «Демография»[6]. Для хлебобулочных изделий размер выборки составил 615 исследований (N = 615), количество лабораторий, проводивших исследования, составило 20 учреждений. В группу вошли такие виды хлеба, как ржано-пшеничный, «Дарницкий», белый пшеничный, бородинский и прочие доступные для приобретения в магазинах на территории Российской Федерации.

Для получения корректных значений среднего, а также коэффициента вариации необходимо гомогенезировать исходные данные. С этой целью первым этапом необходимо провести оценку наличия выбросов данных. В данной работе в связи с тем, что распределение исследуемых величин является нормальным (тест Шапиро – Уилка, $p > 0,05$), целесообразно применение правила трех сигм – по всем показателям по исходному массиву определялись среднее и стандартное отклонение, затем все значения, отклоняющиеся от среднего более чем на 3 стандартных отклонения, обозначались как выбросы и соответствующие исследования (строки в базе данных) удалялись из дальнейшего анализа. Стоит сделать уточнение, что при работе с большим количеством показателей и, как следствие, потенциально большим количеством выпадающих значений, стоит использовать более современные методы и оценивать их эффективность для каждого конкретного случая [19].

Для реализации последующих этапов необходимо исключение пропущенных значений в базе данных. В связи с небольшим, по сравнению с общим количеством, пропущенных значений в данной работе они исключались из дальнейшего анализа. Однако при воспроизведении алгоритма на более неполных базах данных целесообразно пропущенные значения заменять средним, модой или медианой, в зависимости от характера распределения исходных значений [20]. Несмотря на то что современные статистические методы, предложенные различными авторами, такие как применение аппарата нечеткой логики [21] или логистической регрессии [22], позволяют повысить качество данных, они требуют дополнительного анализа для адаптации их применения именно к пищевых продуктам, так как необходимо учитывать возможные внутренние корреляции параметров. Так, например, по данным [23], для кукурузы характерно наличие статистически значимой корреляции, обнаруженной между питательными

[6] Паспорт национального проекта Демография (утв. президиумом Совета при Президенте Российской Федерации по стратегическому развитию и национальным проектам, протокол от 24.12.2018 № 16).

https://doi.org/10.35627/2219-5238/2022-30-4-44-53
Оригинальная исследовательская статья

веществами, минералами и тяжелыми металлами, которые могут быть использованы для прогнозирования содержания растворимых волокон, цинка, свинца и фосфора. Одновременно с этим не стоит оставлять без внимания вопрос компьютерной сложности организации подобных вычислений для нестатистических методов [24]. Учитывая растущие объемы лабораторных исследований, необходимо соблюдать баланс между сложностью и соответственно временем вычислений [25].

В связи с тем что анализ проводится по обобщенной группе продуктов, необходимо либо проводить синтаксический анализ названий продуктов для выделения подгрупп продукции и дальнейшего анализа уже внутри них [26, 27], что является сложной задачей, которая при низком качестве ввода исходных данных будет давать некорректные результаты, либо проводить кластеризацию исходного массива данных. Дополнительной причиной применения кластеризации является необходимость усреднения группы продуктов в связи с тем, что нельзя говорить с точки зрения торговых наименований о постоянстве рациона. Кластеризацию результатов исследований предполагалось проводить по показателям, характеризующим потребительские свойства (например, входящие в сортность соответствующего продукта), которые позволяют получить качественное разделение на группы. Начальное число кластеров для анализа определялось исходя из структуры данных или имеющихся сведений о группах продукции по потребительским свойствам, например по имеющимся сортам в соответствии со стандартами. В качестве метода кластеризации использовался метод k-средних. Несмотря на известные проблемы и ограничения данного метода [28], для полученных однородных данных его можно считать достаточно результативным.

Так как результаты получены на неоднотипных продуктах множеством лабораторий на различном оборудовании, с разной точностью, необходимо последним этапом произвести нормализацию данных. Для нормализации данных предлагается использовать методологию, используемую для анализа результатов исследований крови [29]. Несмотря на то что указанный метод имел при разработке другую сферу применения, с точки зрения стандартизации результатов исследований не имеет существенного значения объект исследования. Так, биологическая вариабельность, которая является частью общей вариабельности у пищевых продуктов, характерна и для исследований, проведенных в рамках клинической лабораторной диагностики. Аналогично можно оценить вариабельность, вклад которой обусловлен оборудованием и квалификацией персонала, – для любых неавтоматизированных исследований на прецизионном оборудовании невозможно исключение случайной ошибки, вносимой указанными факторами. В случае рассмотрения серии исследований на неоднородном материале ошибка становится случайной еще и во времени.

В соответствии с упомянутой методикой нормирование данных осуществляется с использованием следующей формулы:

$$y = (x - L_i) \times \frac{U_{CS} - L_{CS}}{U_i - L_i} + L_{CS}, \quad (1)$$

где y – нормализованное отдельное значение показателя;

x – исходное значение показателя;

L_i и U_i – нижний и верхний пределы для результатов отдельной лаборатории;

L_{cs} и U_{cs} – нижний и верхний предел для выбранной общей стандартной лаборатории или некоторой фантомной лаборатории.

В качестве стандартной лаборатории подразумевается лаборатория, обладающая наиболее точными результатами исследований за счет минимального влияния всех ранее указанных ошибок. Теоретически такой лабораторией можно признать одну или даже несколько после проведения внутрилабораторных сличительных испытаний на множестве однородного материала, который в дальнейшем будет исследоваться уже в реальных условиях. Однако это требует не только существенных временных и материальных затрат для организаторов подобных раундов сличительных испытаний, но и предполагает допущение о гомогенности химического состава пищевых продуктов, отобранных случайным образом в торговых точках.

В связи с этим становится целесообразным рассмотреть применение именно фантомной лаборатории. В качестве фантомной лаборатории будем считать некоторую абстракцию, выраженную нижним и верхним пределом обнаружения, которая для данной группы реальных данных о проведенных исследованиях будет характеризоваться наименьшей разницей между данными пределами. В связи с тем, что, как ранее было принято, выборка объектов для исследования не является однородной, поиск наиболее близких пределов из всей выборки для отдельных лабораторий будет являться ошибкой. Появление таких квази-однородностей означает либо подбор результатов под определенный диапазон, либо отбор и последующее исследование однотипных образцов, что является нарушением дизайна исследования.

В соответствии с работами [30] и [31] предложено использовать в качестве верхнего и нижнего пределов фантомной лаборатории соответствующие перцентили. Учитывая достаточно большое количество измерений в каждой группе (> 300), с целью получения более точной оценки статистических показателей были выбраны 20- и 80-процентные перцентили. Решение задачи оптимизации с целью определения минимального допустимого перцентиля, который обеспечит получение достоверных значений для каждой лаборатории по результатам предложенного преобразования при сохранении достоверной структуры данных, в работе не рассматривается. Однако по итогам апробации алгоритма на различных соотношениях экспертным путем именно использование указанного соотношения позволило соблюсти баланс между выходной точностью и потерей данных.

Таким образом, алгоритм получения точного значения приобретает следующий вид:

1) проверка на наличие выбросов данных и их удаление;
2) обработка пропущенных значений;
3) кластеризация полученных значений;
4) нормализация данных.

Дополнительным этапом, который может быть проведен перед реализацией указанного алгоритма, является оценка достоверности данных – исключение результатов отдельных лабораторий,

чьи результаты либо слишком сильно неточны, либо, наоборот, вероятно, являются искусственно подобранными.

Однако важно отметить, что причины и мотивация для корректировки или исключения результатов тестирования не указывают автоматически на мошенничество. Лабораторные процедуры должны позволять исправлять ошибки или исследовать неверные результаты. Неиспользованные, незарегистрированные или бесхозные данные могут быть вызваны чрезмерно упрощенными или слабыми методами документирования, неопытностью персонала или особенно сложными аналитическими методами.

Алгоритм и соответствующие вычисления реализованы на языке R версии 4.1.2 в среде разработки RStudio.

Результаты. Был проведен анализ исходных данных по исследуемой группе продуктов. Были рассчитаны основные меры описательной статистики для исходных данных (табл. 1). Чем ближе коэффициент вариации к 100 %, тем менее воспроизводимой и точной считалась величина.

Был проведен анализ на наличие выбросов значений по каждому из показателей. Для наглядности приведен пример с выпадающими значениями по содержанию белка и жира (рис. 1). Измерения, которые по правилу трех «сигм» должны быть исключены из дальнейшего анализа, обозначены на рисунке квадратами. В качестве некоторой меры эффективности алгоритма было принято решение рассматривать разницу в коэффициенте вариации между значениями после каждого этапа и исходными данными (табл. 2).

Далее в соответствии с алгоритмом было произведено разделение данных на кластеры. В качестве кластерообразующих переменных были выбраны показатели содержания белка и жира, что, во-первых, позволяет отделить так называемый сэндвичный хлеб, характеризующийся высоким содержанием жира, и подобные виды, а во-вторых, дает достаточно хорошую разделяющую картину. Число кластеров 2 было определено с помощью пакета NbClust, который использует 30 различных индексов для определения оптимального количества кластеров, таких как, например, индекс Данна и индекс Дэвиса – Болдуина [32].

Результаты кластеризации представлены на рис. 2. В первый кластер вошло 457 исследований, которые расположены по левую сторону от разделяющей прямой, во второй – 158, они расположены по правую сторону от прямой.

Внутри полученных кластеров уже был произведен процесс нормализации данных в соответствии

Таблица 1. **Меры описательной статистики исходных данных**

Table 1. **Descriptive statistics of initial data**

Показатель / Parameter	Среднее / Mean	Стандартное отклонение / Standard deviation	Коэффициент вариации (CV) / Coefficient of variation (CV)
Содержание жира, г/100 г / Fat content, g/100 g	1,83	1,53	83,19
Содержание белка, г/100 г / Protein content, g/100 g	7,58	1,06	13,99
Содержание золы, % / Ash content, %	1,58	0,46	29,13
Влажность, % / Moisture content, %	40,33	5,43	13,46
Углеводы (расчетные), г/100 г / Carbohydrates (estimated), g/100 g	48,06	4,93	10,25
Пищевые волокна, % / Dietary fiber, %	4,20	2,10	49,92
Витамин B_1, мг/100 г / Vitamin B_1, mg/100 g	0,15	0,14	92,41
Na, мг/кг / Na, mg/kg	4264,09	1440,80	33,79

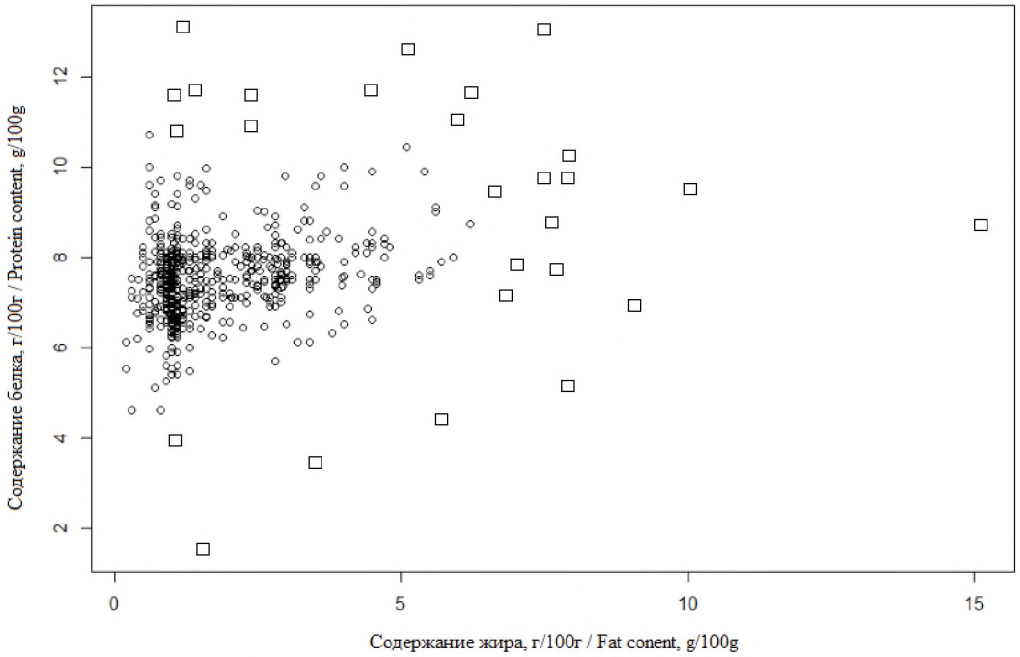

Рис. 1. Выбросы по содержанию белка и жира

Fig. 1. Fat and protein content outliers

https://doi.org/10.35627/2219-5238/2022-30-4-44-53
Оригинальная исследовательская статья

Таблица 2. **Разница в коэффициенте вариации после удаления выбросов**
Table 2. **The difference in the coefficient of variation after exclusion of outliers**

Показатель / Parameter	Среднее / Mean	Стандартное отклонение / Standard deviation	Коэффициент вариации (CV) / Coefficient of variation (CV)	Разница в коэффициенте вариации с исходными данными / Difference in CV compared to initial data
Содержание жира, г/100 г / Fat content, g/100 g	1,66	1,13	67,86	15,33
Содержание белка, г/100 г / Protein content, g/100 g	7,51	0,79	10,46	3,53
Содержание золы, % / Ash content, %	1,58	0,40	25,43	3,70
Влажность, % / Moisture content, %	40,43	5,21	12,87	0,59
Углеводы (расчетные), г/100 г / Carbohydrates (estimated), g/100 g	48,24	4,67	9,67	0,58
Пищевые волокна, % / Dietary fiber, %	4,16	1,97	47,31	2,61
Витамин B_1, мг/100 г / Vitamin B_1, mg/100 g	0,14	0,08	59,38	33,03
Na, мг/кг / Na, mg/kg	4167,47	1256,09	30,14	3,65

Таблица 3. **Результаты кластеризации**
Table 3. **Results of clustering**

Показатель / Parameter	Первый кластер / First cluster		Второй кластер / Second cluster	
	минимум / minimum	максимум / maximum	минимум / minimum	максимум / maximum
Содержание жира, г/100 г / Fat content, g/100 g	0,2	2,3	2,1	6,2
Содержание белка, г/100 г / Protein content, g/100 g	5,11	10	5,7	10,44
Содержание золы, % / Ash content, %	0	2,8	0,567	2,68
Влажность, % / Moisture content, %	23,9	52	22	50
Углеводы (расчетные), г/100 г / Carbohydrates (estimated), g/100 g	36,11	66,4	36	62,2
Пищевые волокна, % / Dietary fiber, %	0,6	9,4	0	9,15
Витамин B_1, мг/100 г / Vitamin B_1, mg/100 g	0	0,48	0	0,38
Na, мг/кг / Na, mg/kg	1369,4	7900	1036,2	8564

с ранее указанными формулами. Результаты нормализации представлены в табл. 4 и 5.

Для того чтобы оценить корректность всего алгоритма, полученные результаты сравнивались с таблицами химического состава (табл. 6).

Обсуждение. Значения получились сопоставимыми, а местами даже более близкими в характере распределения, по сравнению со значениями из справочника[3]. Исключениями являются содержание жира во втором кластере, что обусловлено широким разбросом исходных значений показателя (от 2,1 до 6,2 г/100 г), и содержание витамина B_1 в том же кластере, что, вероятно, можно связать с небольшим количеством исследований в группе.

Для оценки статистически значимых отличий между полученными средними и исходными данными было решение использовать *t*-тест для парного сравнения (табл. 7). Значимыми считали

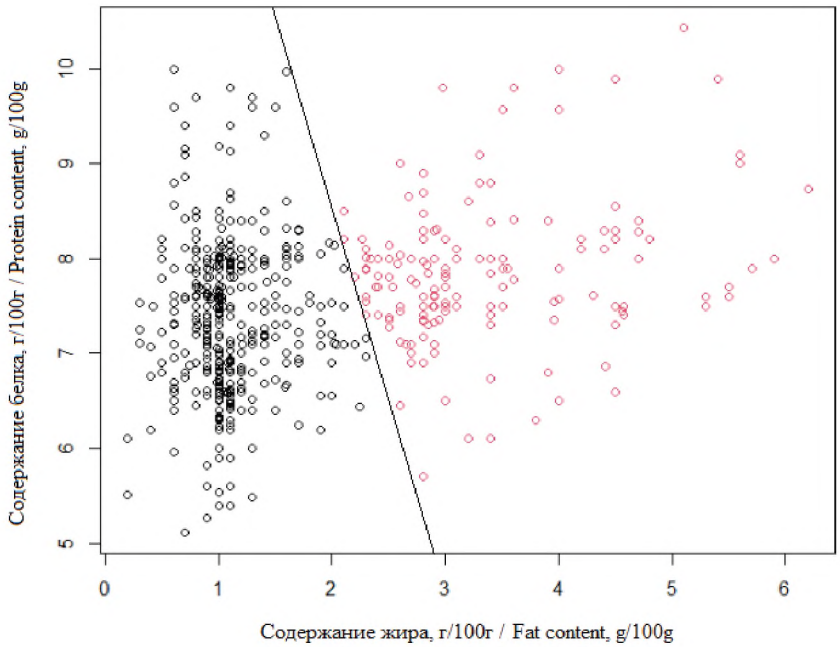

Рис. 2. Результаты кластеризации
Fig. 2. Clustering results

Таблица 4. **Первый кластер. Хлеб с меньшим содержанием жира**

Table 4. **First cluster: Bread with a lower fat content**

Показатель / Parameter	Среднее / Mean	Стандартное отклонение / Standard deviation	Коэффициент вариации (CV) / Coefficient of variation (CV)	Разница в коэффициенте вариации с исходными данными / Difference in CV compared to initial data
Содержание жира, г/100 г / Fat content, g/100 g	1,07	0,11	10,40	72,79
Содержание белка, г/100 г / Protein content, g/100 g	7,39	0,32	4,32	9,67
Содержание золы, % / Ash content, %	1,66	0,14	8,38	20,75
Влажность, % / Moisture content, %	42,67	2,07	4,86	8,60
Углеводы (расчетные), г/100 г / Carbohydrates (estimated), g/100 g	46,89	1,92	4,09	6,16
Пищевые волокна, % / Dietary fiber, %	4,46	1,16	25,91	24,01
Витамин B_1, мг/100 г / Vitamin B_1, mg/100 g	0,14	0,03	20,70	71,71
Na, мг/кг / Na, mg/kg	4016,50	565,91	14,09	19,70

Таблица 5. **Второй кластер. Хлеб с большим содержанием жира**

Table 5. **Bread with a higher fat content**

Показатель / Parameter	Среднее / Mean	Стандартное отклонение / Standard deviation	Коэффициент вариации (CV) / Coefficient of variation (CV)	Разница в коэффициенте вариации с исходными данными / Difference in CV compared to initial data
Содержание жира, г/100 г / Fat content, g/100 g	3,23	0,51	15,65	67,54
Содержание белка, г/100 г / Protein content, g/100 g	7,80	0,30	3,80	10,19
Содержание золы, % / Ash content, %	1,51	0,20	13,14	15,99
Влажность, % / Moisture content, %	36,12	2,05	5,67	7,79
Углеводы (расчетные), г/100 г / Carbohydrates (estimated), g/100 g	50,87	2,20	4,32	5,93
Пищевые волокна, % / Dietary fiber, %	3,30	0,85	25,81	24,11
Витамин B_1, мг/100 г / Vitamin B_1, mg/100 g	0,09	0,05	62,74	29,67
Na, мг/кг / Na, mg/kg	4078,05	506,28	12,41	21,38

Таблица 6. **Сравнение результатов со справочными**

Table 6. **Comparison of the results with reference values**

Показатель / Parameter	Первый кластер (CV), % / First cluster (CV), %	Табличное значение / Reference value	Второй кластер (CV), % / Second cluster (CV), %	Табличное значение / Reference value
Содержание жира, г/100 г / Fat content, g/100 g	10,40	13	**15,65**	**10**
Содержание белка, г/100 г / Protein content, g/100 g	4,32	7	3,80	7
Содержание золы, % / Ash content, %	8,38	–	13,14	–
Влажность, % / Moisture content, %	4,86	–	5,67	–
Углеводы (расчетные), г/100 г / Carbohydrates (estimated), g/100 g	4,09	10	4,32	10
Пищевые волокна, % / Dietary fiber, %	25,91	29	25,81	29
Витамин B_1, мг/100 г / Vitamin B_1, mg/100 g	20,70	44	**62,74**	44
Na, мг/кг / Na, mg/kg	14,09	31	12,41	31

Примечание: величины, превышающие табличные, выделены жирным шрифтом.
Note: The values exceeding reference ones are in bold.

отличия при $p < 0?05$, выделены жирным шрифтом в таблице. Несмотря на то что применение статистических критериев определения достоверности различий после процедуры кластеризации и является обычно нецелесообразным, в данном случае имела место многоэтапная обработка и проверка значимости итогового результата должна быть произведена.

Заключение. Получение корректных значений для включения в базы данных химического состава пищевых продуктов является комплексной задачей, требующей разработки соответствующих алгоритмов обработки исходных массивов и создания методологической основы по самой организации процесса. Разработанный алгоритм позволяет получить актуальные сведения о химическом составе хлебобулочных изделий, а именно о средних значениях и общей вариабельности. Полученные значения могут быть использованы как для актуализации базы данных химического

https://doi.org/10.35627/2219-5238/2022-30-4-44-53
Оригинальная исследовательская статья

Таблица 7. **Меры описательной статистики исходных данных**
Table 7. **Descriptive statistics of initial data**

Показатель / Parameter	*p*	
	первый кластер / first cluster	второй кластер / second cluster
Содержание жира, г/100 г / Fat content, g/100 g	0,39860665	**0,01300045**
Содержание белка, г/100 г / Protein content, g/100 g	0,79341591	0,30806774
Содержание золы, % / Ash content, %	**0,00067599**	**0,04613272**
Влажность, % / Moisture content, %	**0,00000001**	0,18202649
Углеводы (расчетные), г/100 г / Carbohydrates (estimated), g/100 g	**0,00001824**	**0,00816488**
Пищевые волокна, % / Dietary fiber, %	**0,04553885**	**0,01187709**
Витамин B_1, мг/100 г / Vitamin B_1, mg/100 g	**0,00418619**	**$5{,}0824\times10^{-8}$**
Na, мг/кг / Na, mg/kg	**0,00004577**	0,33601625

Примечание: величины, не превышающие заданный уровень значимости, выделены жирным шрифтом.
Note: The values not exceeding a given significance level are in bold.

состава пищевых продуктов в части хлебобулочных изделий, так и для оценки фактического питания населения. Так, для включения в справочник может быть предложена классификация не только по названиям, но и по полученным группам, в том числе возможно указание конкретных полученных показателей вариабельности для каждого показателя. Подобное обобщение может показать свою значимость при проведении эпидемиологических оценок состояния питания в случае невозможности определения потребления конкретного вида хлеба, а более точное определение вариабельности позволит установить реальный диапазон получаемых нутриентов. Предлагаемые подходы могут быть опробованы и внедрены в практику на базе единой информационно-аналитической системы Роспотребнадзора, собирающей результаты контрольно-надзорной деятельности и мониторинга, включая данные лабораторных испытаний. Алгоритм имеет перспективы развития и совершенствования. Дальнейшие исследования должны быть направлены на апробацию и, в случае необходимости, корректировку алгоритма для всех основных групп пищевых продуктов.

Отсутствие значимых различий свидетельствовало либо о качестве исходных результатов, либо, в случае второго кластера, о меньшем по сравнению с первым кластером количестве исследований для каждой лаборатории, что повлияло на эффективность алгоритма.

Список литературы

1. Healthy Diet. WHO Fact Sheets. Accessed April 23, 2022. https://www.who.int/news-room/fact-sheets/detail/healthy-diet
2. Попова А.Ю. Тутельян В.А., Никитюк Д.Б. О новых (2021) Нормах физиологических потребностей в энергии и пищевых веществах для различных групп населения Российской Федерации // Вопросы питания. 2021. Т. 90. № 4 (536). С. 6–19. doi: 10.33029/0042-8833-2021-90-4-6-19.
3. Samman N, Rossi MC. 12th IFDC 2017 Special Issue – Challenges facing the establishment and management of a national food composition database in Argentina. *J Food Compos Anal.* 2019;84(132):103292. doi: 10.1016/j.jfca.2019.103292
4. Silva M, Ribeiro M, Viegas O, *et al.* Exploring two food composition databases to estimate nutritional components of whole meals. *J Food Compos Anal.* 2021;102:104070. doi: 10.1016/j.jfca.2021.104070
5. Jeddi MZ, Boon PE, Cubadda F, *et al.* A vision on the 'foodture' role of dietary exposure sciences in the interplay between food safety and nutrition. T*rends Food Sci Technol.* 2022;120:288-300. doi: 10.1016/j.tifs.2022.01.024
6. Бессонов В.В., Богачук М.Н., Боков Д.О. и др. Базы данных химического состава пищевых продуктов в эпоху цифровой нутрициологии // Вопросы питания. 2020. Т. 89. № 4. С. 211–219. doi: 10.24411/0042-8833-2020-10058
7. Scrimshaw NS. INFOODS: the international network of food data systems. *Am J Clin Nutr.* 1997;65(4 Suppl):1190S-1193S. doi: 10.1093/ajcn/65.4.1190S
8. FAO/INFOODS Guidelines for Converting Units, Denominators, and Expressions – Version 1.0. FAO, Rome; 2012.
9. FAO/INFOODS Guidelines for Checking Food Composition Data prior to the Publication of a User Table/Database – Version 1.0. FAO, Rome; 2012.
10. Machackova M, Giertlova A, Porubska J, Roe M, Ramos C, Finglas P. EuroFIR Guideline on calculation of nutrient content of foods for food business operators. *Food Chem.* 2017;238:35-41. doi: 10.1016/j.foodchem.2017.03.103
11. Schubert A, Holden JM, Wolf WR. Selenium content of a core group of foods based on a critical evaluation of published analytical data. *J Am Diet Assoc.* 1987;87(3):285-99.
12. West CE, Poortvliet EJ. The carotenoid content of foods with special reference to developing countries (Report). 1993. Accessed April 23, 2022. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.1142&rep=rep1&type=pdf
13. Holden JM, Lemar LE, Exler J. Vitamin D in foods: development of the US Department of Agriculture database. *Am J Clin Nutr.* 2008;87(4):1092S-1096S. doi: 10.1093/ajcn/87.4.1092S
14. Raffo A, La Malfa G, Fogliano V, Maiani G, Quaglia G. Seasonal variations in antioxidant components of cherry tomatoes (*Lycopersicon esculentum* cv. Naomi F1). *J Food Compos Anal.* 2006;19(1):11-19. doi: 10.1016/j.jfca.2005.02.003
15. Luthria DL, Pastor-Corrales MA. Phenolic acids content of fifteen dry edible bean (*Phaseolus vulgaris* L.) varieties. *J Food Compos Anal.* 2006;19(2–3):205-211. doi: 10.1016/j.jfca.2005.09.003
16. Sarac I, Butnariu M. Food pyramid – The principles of a balanced diet. *Int J Nutr Pharmacol Neurol Dis.* 2020;5(2):24-31. doi: 10.14302/issn.2379-7835.ijn-20-3199
17. Lockyer S, Spiro A. The role of bread in the UK diet: An update. *Nutr Bull.* 2020;45(2):133-164. doi: 10.1111/nbu.12435
18. Bati A. The role of bread in Hungarian diet today. *Acta Ethnographica Hungarica.* 2012;57(2):253-261. doi: 10.1556/AEthn.57.2012.2.3
19. Xu X, Liu H, Li L, Yao M. A comparison of outlier detection techniques for high-dimensional data. *Int J Comput Intell Syst.* 2018;11(1):652. doi: 10.2991/ijcis.11.1.50
20. Das D, Nayak M, Pani SK. Missing value imputation – A review. *Int J Comput Sci Eng.* 2019;7(4):548-558. doi: 10.26438/ijcse/v7i4.548558
21. El-Bakry M, Ali F, El-Kilany A, Mazen S. Fuzzy based techniques for handling missing values. *Int J Adv Comput Sci Appl.* 2021;12(3). doi: 10.14569/IJACSA.2021.0120306
22. Nadraga V, Smirnov V, Boiko O, Dereko V. Comparison of missing values handling techniques using MICE package tools of R software and logistic regression model. In: Babichev S, Lytvynenko V, Wójcik W, Vyshemyrskaya S, eds. *Lecture Notes in Computational Intelligence and Decision Making.* Springer, Cham; 2021;1246:39-50. doi: 10.1007/978-3-030-54215-3_3
23. Pekel AY, Çalık A, Alataş MS, *et al.* Evaluation of correlations between nutrients, fatty acids, heavy metals, and

deoxynivalenol in corn (*Zea mays* L.). *J Appl Poult Res.* 2019;28(1):94-107. doi: 10.3382/japr/pfy023
24. Pollard S, Namazi H, Khaksar R. Big data applications in food safety and quality. In: *Encyclopedia of Food Chemistry*. Academic Press; 2019:356-363. doi: 10.1016/b978-0-08-100596-5.21839-8
25. Rashid W, Gupta MK. A perspective of missing value imputation approaches. In: Gao XZ, Tiwari S, Trivedi M, Mishra K, eds. *Advances in Computational Intelligence and Communication Technology*. Springer, Singapore; 2021;1086:307-315. doi: 10.1007/978-981-15-1275-9_25
26. Amano S, Aizawa K, Ogawa M. Food category representatives: Extracting categories from meal names in food recordings and recipe data. *2015 IEEE International Conference on Multimedia Big Data*; 2015:48-55. doi: 10.1109/BigMM.2015.54
27. Anzawa M, Amano S, Yamakata Y, Yamasaki T, Aizawa K, Ogawa M. Generation of representative meal names for food recording data by using web search results. *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*; 2016:1-6. doi: 10.1109/ICMEW.2016.7574745
28. Ahmed M, Seraj R, Syed Mohammed Shamsul Islam. The *k-means* algorithm: A comprehensive survey and performance evaluation. *Electronics*. 2020;9(8):1295. doi: 10.3390/electronics9081295
29. Ruvuna F, Flores D, Mikrut B, De La Gana K, Fong S. Generalized lab norms for standardizing data from multiple laboratories. *Drug Inf J*. 2003;37(1):61–79. doi: 10.1177/009286150303700109
30. Brunden MN, Clark JJ, Sutter ML. A general method of determining normal ranges applied to blood values for dogs. *Am J Clin Pathol*. 1970;53(3):332-339. doi: 10.1093/ajcp/53.3.332
31. Herrera L. The precision of percentiles in establishing normal limits in medicine. *J Lab Clin Med*. 1958;52(1):34-42.
32. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R package for determining the relevant number of clusters in a data set. *J Stat Softw*. 2014;61(6):1-36. doi: 10.18637/jss.v061.i06

References

1. Healthy Diet. WHO Fact Sheets. Accessed April 23, 2022. https://www.who.int/news-room/fact-sheets/detail/healthy-diet
2. Popova AYu, Tutelyan VA, Nikityuk DB. On the new (2021) Norms of physiological requirements in energy and nutrients of various groups of the population of the Russian Federation. *Voprosy Pitaniya*. 2021;90(4(536)):6-19. (In Russ.) doi: 10.33029/0042-8833-2021-90-4-6-19
3. Samman N, Rossi MC. 12th IFDC 2017 Special Issue – Challenges facing the establishment and management of a national food composition database in Argentina. *J Food Compos Anal*. 2019;84(132):103292. doi: 10.1016/j.jfca.2019.103292
4. Silva M, Ribeiro M, Viegas O, *et al.* Exploring two food composition databases to estimate nutritional components of whole meals. *J Food Compos Anal*. 2021;102:104070. doi: 10.1016/j.jfca.2021.104070
5. Jeddi MZ, Boon PE, Cubadda F, *et al.* A vision on the 'foodture' role of dietary exposure sciences in the interplay between food safety and nutrition. *Trends Food Sci Technol*. 2022;120:288-300. doi: 10.1016/j.tifs.2022.01.024
6. Bessonov VV, Bogachuk MN, Bokov DO, *et al.* Databases of the chemical composition of foods in the era of digital nutrition science. *Voprosy Pitaniya*. 2020;89(4):211-219. (In Russ.) doi: 10.24411/0042-8833-2020-10058
7. Scrimshaw NS. INFOODS: the international network of food data systems. *Am J Clin Nutr*. 1997;65(4 Suppl):1190S-1193S. doi: 10.1093/ajcn/65.4.1190S
8. FAO/INFOODS Guidelines for Converting Units, Denominators, and Expressions – Version 1.0. FAO, Rome; 2012.
9. FAO/INFOODS Guidelines for Checking Food Composition Data prior to the Publication of a User Table/Database – Version 1.0. FAO, Rome; 2012.
10. Machackova M, Giertlova A, Porubska J, Roe M, Ramos C, Finglas P. EuroFIR Guideline on calculation of nutrient content of foods for food business operators. *Food Chem*. 2017;238:35-41. doi: 10.1016/j.foodchem.2017.03.103
11. Schubert A, Holden JM, Wolf WR. Selenium content of a core group of foods based on a critical evaluation of published analytical data. *J Am Diet Assoc*. 1987;87(3):285-99.
12. West CE, Poortvliet EJ. The carotenoid content of foods with special reference to developing countries (Report). 1993. Accessed April 23, 2022. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.1142&rep=rep1&type=pdf
13. Holden JM, Lemar LE, Exler J. Vitamin D in foods: development of the US Department of Agriculture database. *Am J Clin Nutr*. 2008;87(4):1092S-1096S. doi: 10.1093/ajcn/87.4.1092S
14. Raffo A, La Malfa G, Fogliano V, Maiani G, Quaglia G. Seasonal variations in antioxidant components of cherry tomatoes (*Lycopersicon esculentum* cv. Naomi F1). *J Food Compos Anal*. 2006;19(1):11-19. doi: 10.1016/j.jfca.2005.02.003
15. Luthria DL, Pastor-Corrales MA. Phenolic acids content of fifteen dry edible bean (*Phaseolus vulgaris* L.) varieties. *J Food Compos Anal*. 2006;19(2–3):205-211. doi: 10.1016/j.jfca.2005.09.003
16. Sarac I, Butnariu M. Food pyramid – The principles of a balanced diet. *Int J Nutr Pharmacol Neurol Dis*. 2020;5(2):24-31. doi: 10.14302/issn.2379-7835.ijn-20-3199
17. Lockyer S, Spiro A. The role of bread in the UK diet: An update. *Nutr Bull*. 2020;45(2):133-164. doi: 10.1111/nbu.12435
18. Bati A. The role of bread in Hungarian diet today. *Acta Ethnographica Hungarica*. 2012;57(2):253-261. doi: 10.1556/AEthn.57.2012.2.3
19. Xu X, Liu H, Li L, Yao M. A comparison of outlier detection techniques for high-dimensional data. *Int J Comput Intell Syst*. 2018;11(1):652. doi: 10.2991/ijcis.11.1.50
20. Das D, Nayak M, Pani SK. Missing value imputation – A review. *Int J Comput Sci Eng*. 2019;7(4):548-558. doi: 10.26438/ijcse/v7i4.548558
21. El-Bakry M, Ali F, El-Kilany A, Mazen S. Fuzzy based techniques for handling missing values. *Int J Adv Comput Sci Appl*. 2021;12(3). doi: 10.14569/IJACSA.2021.0120306
22. Nadraga V, Smirnov V, Boiko O, Dereko V. Comparison of missing values handling techniques using MICE package tools of R software and logistic regression model. In: Babichev S, Lytvynenko V, Wójcik W, Vyshemyrskaya S, eds. *Lecture Notes in Computational Intelligence and Decision Making*. Springer, Cham; 2021;1246:39-50. doi: 10.1007/978-3-030-54215-3_3
23. Pekel AY, Çalık A, Alataş MS, *et al.* Evaluation of correlations between nutrients, fatty acids, heavy metals, and deoxynivalenol in corn (*Zea mays* L.). *J Appl Poult Res*. 2019;28(1):94-107. doi: 10.3382/japr/pfy023
24. Pollard S, Namazi H, Khaksar R. Big data applications in food safety and quality. In: *Encyclopedia of Food Chemistry*. Academic Press; 2019:356-363. doi: 10.1016/b978-0-08-100596-5.21839-8
25. Rashid W, Gupta MK. A perspective of missing value imputation approaches. In: Gao XZ, Tiwari S, Trivedi M, Mishra K, eds. *Advances in Computational Intelligence and Communication Technology*. Springer, Singapore; 2021;1086:307-315. doi: 10.1007/978-981-15-1275-9_25
26. Amano S, Aizawa K, Ogawa M. Food category representatives: Extracting categories from meal names in food recordings and recipe data. *2015 IEEE International Conference on Multimedia Big Data*; 2015:48-55. doi: 10.1109/BigMM.2015.54
27. Anzawa M, Amano S, Yamakata Y, Yamasaki T, Aizawa K, Ogawa M. Generation of representative meal names for food recording data by using web search results. *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*; 2016:1-6. doi: 10.1109/ICMEW.2016.7574745
28. Ahmed M, Seraj R, Syed Mohammed Shamsul Islam. The *k-means* algorithm: A comprehensive survey and performance evaluation. *Electronics*. 2020;9(8):1295. doi: 10.3390/electronics9081295
29. Ruvuna F, Flores D, Mikrut B, De La Gana K, Fong S. Generalized lab norms for standardizing data from multiple laboratories. *Drug Inf J*. 2003;37(1):61–79. doi: 10.1177/009286150303700109
30. Brunden MN, Clark JJ, Sutter ML. A general method of determining normal ranges applied to blood values for dogs. *Am J Clin Pathol*. 1970;53(3):332-339. doi: 10.1093/ajcp/53.3.332
31. Herrera L. The precision of percentiles in establishing normal limits in medicine. *J Lab Clin Med*. 1958;52(1):34-42.
32. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R package for determining the relevant number of clusters in a data set. *J Stat Softw*. 2014;61(6):1-36. doi: 10.18637/jss.v061.i06