ТЕОРЕТИЧЕСКИЕ И МЕТОДИЧЕСКИЕ ПРОБЛЕМЫ АРХЕОЛОГИИ

THEORETICAL AND METHODOLOGICAL PROBLEMS OF ARCHEOLOGY

Научная статья / Research Article
УДК 311.2:004.9
https://doi.org/10.14258/tpai(2024)36(4).-01
EDN: GDGSYM

НОВАЯ ПРОГРАММА КЛАСТЕРИЗАЦИИ ДАННЫХ И ПОСТРОЕНИЯ ДЕНДРОГРАММ

Александр Альбертович Казаков[1]*, Иван Александрович Казаков[2]

[1]Барнаульский юридический институт МВД России, Барнаул, Россия; kaa-2862@mail.ru, https://orcid.org/0000-0003-2652-2002
[2]Сколковский институт науки и технологий, Москва, Россия; ivan.kazakov@phystech.ru, https://orcid.org/0000-0001-7509-1064
**Автор, ответственный за переписку*

Резюме. Статья посвящена решению актуальной проблемы, стоящей перед учеными-гуманитариями, работающими с большими массивами данных, а именно созданию и описанию программного продукта, позволяющего даже начинающим (уверенным) пользователям компьютера полнее использовать его возможности в исследовательской деятельности. На основе анализа отечественной и зарубежной литературы сделан вывод об отсутствии простой программы анализа большого массива данных методом кластеризации, позволяющей значительно упростить процессы анализа больших массивов данных и на основании полученных результатов делать определенные выводы либо проверять гипотезы, полученные методом качественного анализа. Сложность существующих программных продуктов затрудняет этот процесс, так как их применение возможно лишь в сотрудничестве с IT-специалистами. Приводится описание программного продукта DendrogramGenerator, разработанного авторами и позволяющего без особых проблем использовать все возможности кластер-анализа.

Ключевые слова: кластер-анализ, компьютерная программа, алгоритм работы, исходная матрица, DendrogramGenerator, интернет, свободный доступ

Для цитирования: Казаков А.А., Казаков И.А. Новая программа кластеризации данных и построения дендрограмм // Теория и практика археологических исследований. 2024. Т. 36, №4. С. 9–24. https://doi.org/: 10.14258/tpai(2024)36(4).-01

NEW PROGRAM FOR DATA CLUSTERING AND DENDROGRAM CONSTRUCTION

Alexander A. Kazakov[1]*, Ivan A. Kazakov[2]

[1]Law Institute of the Ministry of Internal Affairs of Russia, Barnaul, Russia;
kaa-2862@mail.ru, https://orcid.org/0000-0003-2652-2002
[2]Skolkovo Institute of Science and Technology, Moscow, Russia;
kazakov-ivan95@mail.ru, https://orcid.org/0000-0001-7509-1064
**Corresponding Author*

Abstract. The article is devoted to solving a pressing problem facing scholars of humanities worldwide, namely, creating and describing a software product that allows even novice (confident) computer users to use its capabilities more fully in their research activities. Based on the analysis of domestic and foreign literature, a conclusion is made about the absence of a simple program for analyzing an extensive array of data using the clustering method, which allows for significantly simplifying the processes of analyzing large arrays of data and based on the results obtained, making certain conclusions or testing hypotheses obtained by the qualitative analysis method. The complexity of existing software products complicates this process since their use is possible only in cooperation with IT specialists. A description of the **DendrogramGenerator** software product, developed by the authors, is provided, which allows using all the capabilities of cluster analysis without any problems.

Keywords: cluster analysis, computer program, operating algorithm, initial matrix, DendrogramGenerator, internet, open source

For citation: Kazakov A.A., Kazakov I.A. New Program for Data Clustering and Dendrogram Construction. *Teoriya i praktika arheologicheskih issledovanij = Theory and Practice of Archaeological Research.* 2024;36(4):9–24. (In Russ.). https://doi.org/10.14258/tpai(2024)36(4).-01

Введение

С появлением и широким распространением в последние десятилетия XX в. вычислительной техники появились и новые возможности, оформившиеся в методы исследования. Компьютеризация прочно вошла в нашу жизнь. Несмотря на это до настоящего времени подавляющее большинство исследователей-гуманитариев используют возможности компьютера далеко не в полной мере. Как правило, он выступает в качестве пишущей машинки, хранилища данных или мини-типографии (в комбинации с принтером). Наиболее продвинутые пользователи используют различные графические редакторы и программы для подготовки иллюстративного материала. И лишь немногие используют методы математического анализа и статистики для обработки большого массива данных. А ведь именно эта область является наиболее перспективной в плане применения ЭВМ, однако для исследователя-гуманитария, не вооруженного специальными знаниями и навыками, без сотрудничества со специалистами-естественниками (программисты, математики и т.д.) использование возможностей методов математического анализа является задачей труднодостижимой.

Хорошо понимая все преимущества машинной обработки массового материала, дающей возможность получения сравнительных данных за относительно небольшой промежуток времени, на которое при традиционных методах расчетов (столбиком или с помощью калькулятора) тратится огромное количество времени и рутинного тех-

нического труда, специалисты-гуманитарии начали активно осваивать эту новую площадку для проведения исследований именно в гуманитарных сферах. Однако, как уже отмечалось, недостаток специальных знаний этот процесс заметно тормозит.

Материалы и методы

При массовом внедрении компьютеров в повседневную деятельность начались различные теоретические изыскания по созданию методик использования электронно-вычислительной техники в гуманитарных исследованиях. Создавались специальные лаборатории, творческие коллективы, проводились научные форумы, на которых обсуждались проблемы компьютеризации гуманитарных исследований, издавалось большое количество работ в этой области (Компьютер и историческое знание, 1994; Гарскова, 2018; и др.).

Все разнообразие предлагаемых методик, которые апологеты компьютеризации предлагали чуть ли не как панацею, способную совершить качественный прорыв в гуманитарных науках, после апробации их на практике свелось к двум основным направлениям: это создание электронных баз данных и различные методы статистического анализа, способные из огромного количества признаков формировать определенные похожести, позволяющие их трактовать в ходе дальнейшего качественного анализа.

Данная методика называется кластеризацией. Основной задачей кластерного анализа является формирование из большого количества объектов определенных групп похожих объектов, при этом определяется и число этих групп. Группы, на которые разбивается выборка, называются кластерами. Кластерный анализ является разновидностью многомерного анализа в классификации. Свое название он получил от английского cluster — гроздь, скопление. Основное его достоинство в том, что он формирует кластеры не по одному, а по набору признаков. Он не накладывает ограничений на вид рассматриваемых объектов и позволяет рассматривать множество данных (Торопчина, Двоерядкина, Вохминцева, 2006, с. 6–7).

Это могут быть различные признаки погребальной обрядности, керамического комплекса, различных групп инвентаря, антропологического материала и тому подобные обособленные образования. Весь анализ сводится, по сути, к систематизации больших объемов материала по принципам схожести/различия.

Причем если подобная работа по систематизации уже была проведена ранее без применения машинных методик обработки материала, то результаты этого анализа либо подтверждают сделанные ранее выводы, либо, напротив, ставят их под сомнение и позволяют по-новому взглянуть, по-другому интерпретировать подвергающийся анализу материал.

Не вдаваясь в плюсы и минусы кластерного анализа, возможности различного рода фальсификаций путем искусственного подбора исходных данных, необходимо сказать, что его применение при обработке больших массивов археологического материала крайне перспективно. Более того, накопление материала, сопровождаемое расширением источниковой базы, обобщение материала различных регионов, требующее осознания и обработки огромных массивов информации, без компьютера просто невозможно.

Существующий на сегодняшний день инструментарий не позволяет массово это делать, прежде всего по причине слабой доступности и сложности программного обеспечения и малой компьютерной грамотности гуманитариев. Овладение даже основа-

ми пользования сложными программами требует больших временных затрат, что отвлекает исследователя от основной деятельности и предполагает его сотрудничество со специалистами в этой области.

Необходимость более широкого использования возможностей машинной обработки материала стимулировала разработку программного продукта, простого в использовании и доступного практически всем археологам и другим исследователям-гуманитариям. С учетом достаточно высокой стоимости и сложности других программ кластеризации было принято решение о размещении созданной программы в сети Интернет в свободном доступе, чтобы ее возможностями могли воспользоваться все желающие. Первоначальная идея о патентовании созданной программы после проработки вопроса была отвергнута, так как патентование предполагает определенные ограничения доступа к программному продукту.

Перед реализацией данной задачи был проработан и опыт зарубежных коллег в этой области. В результате изучения специальной литературы пришли к мнению, что задача создания доступных инструментов статистического (кластерного) анализа не решена даже в мировой практике. В зарубежных публикациях широко освещены различные методы машинной обработки археологических данных, не ограниченные кластерным анализом (Troiano et all., 2024; Feuerverger et all., 2008; Kintigh, Keith, 1990; и др.). Статистические методы используются как часть методик по классификации (Ruck, Lana, Brown, 2015). Существуют и отдельные обширные методические работы, ознакомившись с которыми, можно начать самостоятельно обрабатывать данные с помощью открытого программного обеспечения (Baxter, 2015). Наиболее интересующиеся читатели могут ознакомиться с обзором современных методов машинного обучения для применения в археологии по ссылке (Troiano et all., 2024), а также с дискуссионными работами, в которых обсуждается применимость данных методов (Feuerverger et all., 2008; Kintigh, Keith, 1990). Однако все эти работы проводятся исследователями, специализирующимися на данной тематике, и в специализированном ПО, и не встречаются повсеместно в археологических работах, где они могли бы стать весьма ценными.

Созданная программа позволяет любому желающему произвести кластеризацию большого объема данных. Полученные результаты (таблицы расстояний между анализируемыми объектами и дендрограммы) могут быть использованы при проведении различных исследований.

Результаты

Целью настоящей работы является описание алгоритма кластерного анализа материала, чтобы эти методы стали доступны подавляющему большинству исследователей-археологов и превратились из экзотики в повседневную обыденность, как использование линейки, карандаша или миллиметровки.

Разработанная программа кластеризации **DendrogramGenerator** позволяет даже неискушенному человеку, обычному пользователю компьютера, без особого труда использовать все возможности кластерного анализа при обработке большого массива данных. Она разработана Иваном Александровичем Казаковым по техническому заданию, подготовленному Александром Альбертовичем Казаковым, и совместно протестирована при подготовке монографического исследования (Григоров, Казаков, 2018)

и диссертационного исследования на соискание ученой степени доктора исторических наук по специальности «Археология».

За основу взяты наиболее разработанные, апробированные методы кластер-анализа материала по принципу «дальнего соседа», когда расстояние между кластерами определяется как расстояние между наиболее удаленными объектами кластеров. Это дает более выразительные возможности визуализации полученных результатов.

Основой, которая подвергается анализу (**dataset**), является банк данных, который создается исследователем и в дальнейшем подвергается машинной обработке. В дальнейшем мы ее будем называть исходной матрицей. Если анализу подвергается погребальный обряд, то в исходную матрицу вносят основные признаки погребального обряда, если керамический комплекс, то его основные признаки (технологические, морфологические, орнаментальные), если антропологический материал — то его метрические характеристики и т.п. Причем при составлении матрицы-основы желательно как можно более подробно делить различные признаки, так как степень изученности археологического материала различна, и не учтенные при первичном описании признаки могут быть впоследствии потеряны, что может привести к неверным выводам.

При подготовке исходной матрицы, особенно если она включает в себя большой объем информации, наиболее предпочтительной программой будет являться специально созданная корпорацией **Microsoft** для работы с электронными таблицами программа **Microsoft Excel.** Эта программа достаточно широко распространена, однако в большинстве случаев она является пиратской, оригинальная же версия стоит достаточно дорого. Есть практически полные аналоги этого программного продукта, которые по некоторым характеристикам даже превосходят Microsoft Excel и находятся в свободном доступе. Это программа **LibreOffice Calc**, которая достаточно просто скачивается из интернета и устанавливается практически на любой компьютер, без особых требований к системе. Подобная задача вполне под силу начинающему пользователю.

При составлении исходной матрицы желательно по горизонтали (в горизонтальных строках) размещать наименования исследуемых объектов, а по вертикали (в вертикальных строках) — выделенные признаки и их варианты.

Эта исходная матрица ляжет в основу дальнейшей работы с материалом, несмотря на то что в ней не соблюдается важный момент процедуры классифицирования, а именно — выбранные объекты должны быть однородны, однако она отражает индивидуальную характеристику каждого объекта. В исходной матрице допустимо и наличие условных признаков, таких как «сохранность объекта», допустимо и дробное деление метрических характеристик объекта. Как уже говорилось выше, при составлении исходной матрицы основной задачей является как можно более полная характеристика исследуемого объекта или явления.

С учетом принципа работы ЭВМ данные в исходную матрицу необходимо закладывать по простейшему двоичному принципу «да — нет». Требование дробного деления метрических характеристик обусловлено тем, что разработанный алгоритм работает только с бинарными признаками (либо он есть у объекта, либо нет). При наличии

признака в таблицу вводится «1», при отсутствии — «0». В строках матрицы расположены объекты анализа (археологические памятники), а в столбцах — их признаки. Таблица может иметь заголовки, в которых они написаны, занимающие произвольное количество столбцов и строк. Должен быть минимум один столбец, в котором обозначены названия памятников. Есть также возможность группировать памятники с одним основным названием и постфиксом (к примеру, название памятника и номер кургана или другого объекта: могилы, жилища, хозяйственной постройки, планиграфических данных, краниологических показателей и т.п.). Примеры заполненных таблиц приведены в рисунках. В базе данных программы это файлы («Пример 1 простое название. xls» (рис. 1) и «Пример 2 сложное название.xls» (рис. 2)).

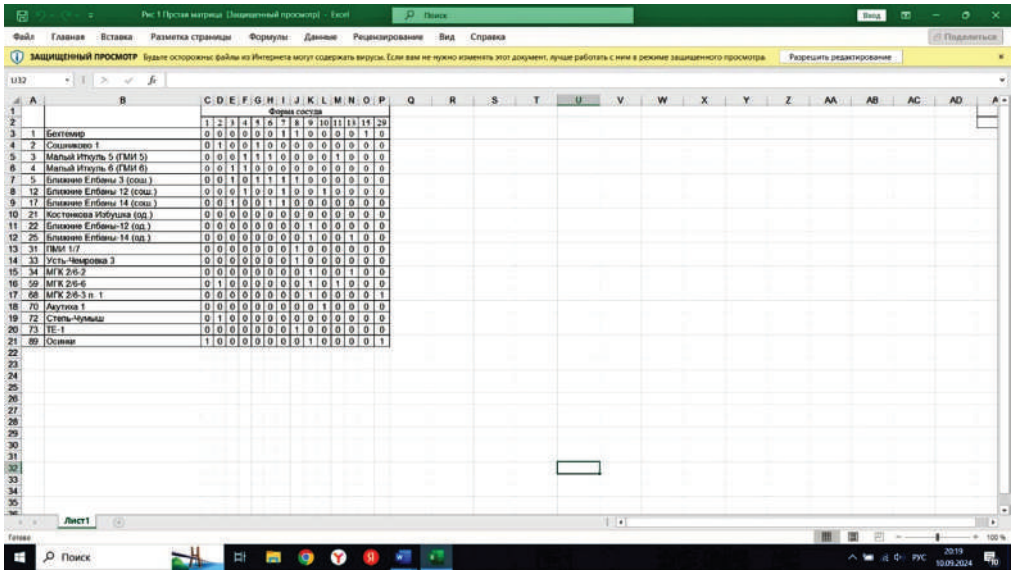

Рис. 1. Пример исходной матрицы с простым названием
Fig. 1. An example of an initial matrix with a simple name

Дальнейшие действия исследователя сводятся к необходимости соблюдения принципа однородности классифицируемых объектов. С этой целью на основе исходной матрицы составляются отдельные матрицы. Например, при анализе погребального обряда это могут быть матрицы для ингумации, для кремации и для характеристики насыпей курганов.

При составлении этих матриц уже соблюден принцип равнозначности признаков. Вполне допустимо, даже желательно с целью исключения фоновых признаков при машинной обработке, когда большое количество малозначимых, но схожих деталей может повлиять на формирование группы объектов со схожими признаками, применить принцип свертывания, т.е. сведение разнообразия вариантов до сути их содержания (Никитина, 1995, с. 10, 116). Еще раз повторяем: излишнее количество признаков может иметь отрицательные последствия для выявления определенных закономерностей в структуре исследуемого объекта или явления.

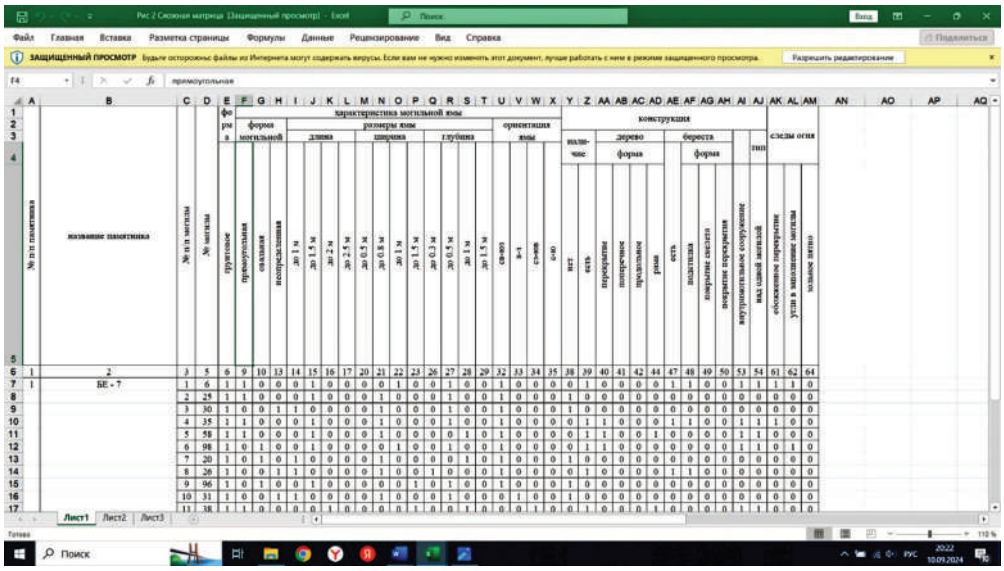

Рис. 2. Пример исходной матрицы со сложным названием
Fig. 2. An example of an initial matrix with a complex name

Составленные подобным образом исходные матрицы, в которых соблюдены принципы равнозначности признаков и принцип свертывания, уже могут служить базой для кластеризации, методы которой активно развиваются в последнее время и дают корректные результаты (Владимиров, Степанова, 1994, с. 3–7; Абдулганеев, Владимиров, 1997, с. 31; Фролов, 2008; Матренин, Тишкин, 2007; Серегин, Матренин, 2016; Кишкурно, Зубова, 2015; Методы экологических исследований…, 2019; Петров, 2013; и др.).

Большинство исследователей для дальнейшей обработки материала используют программу **Statistica** — специально созданную для статистического анализа данных, включающую широкий набор аналитических процедур и методов: более 100 различных типов графиков, описательные и внутригрупповые статистики, разведочный анализ данных, корреляции, быстрые основные статистики и блоковые статистики (Программа STATISTICA).

Эта программа является лицензионной, достаточно дорогостоящей и обладающей огромным спектром возможностей, которые не только не помогают в проведении анализа, а определенным образом путают исследователя. Кроме того, она достаточно сложна в освоении. Для проведения самого распространенного в археологии кластер-анализа с наилучшей визуализацией его результатов — построения дендрограмм по принципу «дальнего соседа» вполне достаточно описываемой программы **DendrogramGenerator**, размещенной авторами в открытом доступе. Ссылка: https://disk.yandex.ru/d/3wCEpnYCcOIMHg

Данная программа реализует алгоритм кластеризации по методу полных связей (complete linkage method), основанный на алгоритме «дальнего соседа». В этом методе

используется метрика, основанная на коэффициенте общего сходства Гауэра. Его основным достоинством является возможность измерения коэффициентов сходства (расстояний) как по качественным, так и по количественным показателям. Он допускает одновременное использование переменных, измеренных по различным шкалам, т.е. он предназначен для смешанных данных (Глазкова, 2017, с. 88; Годяев, Гиголаев, Цуканова, 2017, с. 80; Гайдышев, 2015, с. 276).

Преобразование двоичного кода в коэффициенты Гауэра вызывает затруднения. Некоторые начинают считать их вручную, с калькулятором, тратя на это огромное количество времени и сил. С целью облегчить этот рутинный труд исследователи первоначально проводят классификацию объекта или явления качественными, традиционными методами, после этого выделяют лишь наиболее значимые признаки для уже выделенных групп объектов, составляют двоичную матрицу с малым количеством признаков, после этого просчитывают вручную коэффициенты Гауэра, полученные результаты вводят в машину и на выходе получают либо кластерные дендрограммы, либо иные виды визуализации кластеров.

Формула для просчета этого коэффициента и методика просчета хорошо разъяснены в работе С.С. Матренина и А.А. Тишкина (2007), которая не является библиографической редкостью и размещена в интернете, так что на этом останавливаться не будем.

Предлагаемая к использованию программа предполагает автоматическую процедуру вычисления коэффициента Гауэра и в последующей операции на его основе построение дендрограммы — наиболее наглядного вида визуализации полученных результатов, на котором хорошо читаются расстояния между выделенными кластерами, что дает возможность без особого труда интерпретировать полученные результаты. Поскольку коэффициенты Гауэра имеют дробную природу, возможности программы **Microsoft Excel** или **LibreOffice Calc,** в которой выгружаются готовые результаты машинной обработки, позволяют сократить их до сотых долей десятичной дроби, чего вполне достаточно для целей исследований в гуманитарных сферах.

О формировании исходной матрицы уже говорилось. При работе с программой следует учитывать ее большой объем и сложность выполняемых машиной процедур, что требует в некоторых случаях больших временных промежутков (до 20 минут) в зависимости от конфигурации и мощности компьютера. Алгоритм работы следующий.

1. Находим программу в общем доступе в интернете по ссылке https://disk.yandex.ru/d/3wCEpnYCcOIMHg

Важно!!! Ссылку вставляем не в поисковое окно программы, а в адресную строку (рис. 3).

2. Скачиваем все файлы. Для этого нужно нажать кнопку «Скачать все».

Скачивание данных занимает достаточно большой промежуток времени (до 20 минут), поэтому торопиться не стоит. После скачивания всех данных в папке «Загрузки» появится заархивированная папка с данными, которая называется **DendrogramGenerator**.

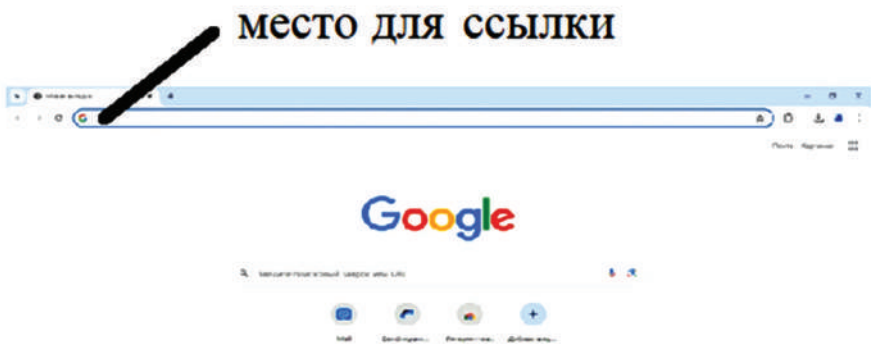

Рис. 3. Адресная строка для поиска программы ***DendrogramGenerator***
Fig. 3. Address bar for searching the DendrogramGenerator program

После активации ссылки попадаете в хранилище данных (рис. 4).

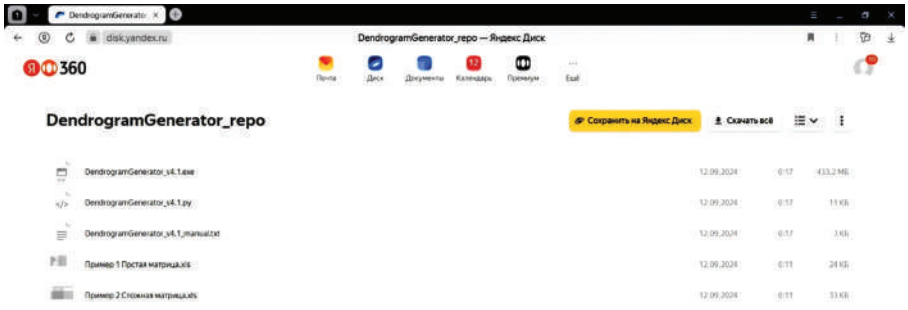

Рис. 4. Вид хранилища данных
Fig. 4. Appearance of the data repository

Эту папку необходимо разархивировать (развернуть). После этого будут доступны следующие файлы:

а) файл **DendrogramGenerator_v4.1.exe** — сама программа кластеризации для построения дендрограмм и таблиц расстояний;

б) файл **DendrogramGenerator_v4.1.py** — программа на языке программирования **python.** Может быть использована пользователями, знакомыми с этим языком и желающими ее модифицировать либо ознакомиться с алгоритмом работы детально, исходные файлы доступны также по ссылке на репозитории **github** по ссылке https://github.com/ivanya-k/archeological_cluster_analysys;

в) файл **DendrogramGenerator_v0.4.1_manual.txt** содержит инструкцию по использованию программы. Дублирует инструкцию, описанную в этой статье;

г) файл **Пример 1 Простая матрица.xls** — пример заполнения исходной таблицы для случая простых названий объектов;

д) файл **Пример 2 Сложная матрица.xls** — пример заполнения исходной таблицы для случая сложных названий объектов.

Примечание 1. Программа не требует повторной установки и может запускаться из скачанного файла.

Примечание 2. В случае появления обновленных версий программы они будут доступны по той же ссылке, только в хранилище файлов дополнительно появятся более новые версии, соответственно для использования новой версии достаточно будет скачать только набор файлов с самой новой версией (наибольшей цифрой после буквы v, например, v5 будет более новой версией по сравнению с текущей v4.1).

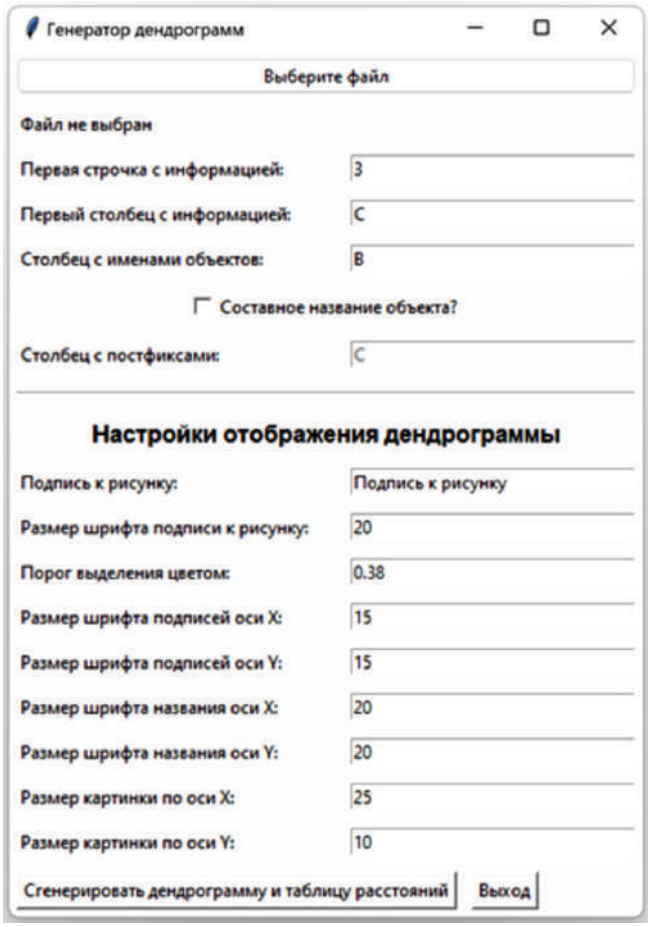

Рис. 5. Командное окно программы ***DendrogramGenerator_v4.1***
Fig. 5. Command window of the ***DendrogramGenerator_v4.1*** *program*

3. При открытии программы появляется командное окно, содержащие три командные клавиши и 14 полей для заполнения (рис. 5).

4. Нажав клавишу, расположенную в самом верху командного окна «Выберите файл», выбираем файл с исходной матрицей.

После того как файл с исходной матрицей, данные которой предназначены для анализа, загружен, в строке ниже, называемой «Выбранный файл», появляется его название.

5. После этого необходимо заполнить поля в командном окне.

Важно! Поля заполняются на латинице!

5.1. В окно «Первая строчка с информацией» вводится порядковый номер строки в исходной матрице, которая содержит качественные данные, подлежащие анализу. На рисунке 1 это будет третья сверху строка, на рисунке 2 это будет седьмая строка. Порядковый номер строки, который следует вводить в это окно, можно посмотреть в крайнем левом столбце программы, в которой заполнена исходная матрица (**Microsoft Excel** или **LibreOffice Calc**).

5.2. В окно «Первый столбец с информацией» вводится буквенное обозначение столбца исходной матрицы, содержащего качественные данные. На рисунке 1 это будет столбец «C», на рисунке 2 — столбец «E». Буквенное обозначение столбца, которое следует вводить в это окно, можно посмотреть в верхней строке программы, в которой заполнена исходная матрица.

5.3. В окно «Столбец с именами объектов» вводится буквенное обозначение столбца исходной матрицы, содержащего название памятника или серии объекта, который подвергается кластеризации. На рисунке 1 это будет столбец «B», на рисунке 2 также столбец «B». Буквенное обозначение столбца, которое следует вводить в это окно, можно посмотреть в верхней строке программы, в которой заполнена исходная матрица. Название объекта будет сгенерировано на дендрограмме по оси «X». Кроме того, эти названия будут сгенерированы в матрице, получаемой на выходе, в которой будут отображены коэффициенты Гауэра.

5.4. Поле для галочки «Составное название объекта» предназначено для просчета матриц, в которых анализируются различные объекты с наименованием, содержащим постфиксы. Например, МГК-2/6-2 (городище Малый Гоньбинский Кордон-2/6-2 — название памятника) и конкретный объект этого памятника (ж. 7 —жилище 7), который является постфиксом. **При отсутствии постфикса ставить галочку не нужно.** Это могут быть номера конкретных погребений могильника, номера курганов и тому подобные объекты.

Если постфиксы есть, то в этом окне нужно поставить галочку, а в строке «Столбец с постфиксами» поставить буквенное обозначение столбца исходной матрицы, содержащего постфикс анализируемого объекта. Это может быть номер могилы, номер жилища или другая конкретизирующая объект анализа информация. На рисунке 1 постфиксов нет, на рисунке 2 это столбец «D».

В итоговой генерации постфиксы присоединяются к названию памятника, позволяя идентифицировать кластеризуемые объекты. Если постфиксов нет и галочка над этой строчкой не поставлена, то данное поле будет заблокировано.

5.5. В окно «Подпись к рисунку» вводится название рисунка, которое затем будет сгенерировано в формате вывода данных **PNG** на дендрограмме. Здесь регистр не важен, так как подпись может быть произвольной. Подписи могут быть набраны кириллицей.

5.6. В окно «Размер шрифта подписи к рисунку» вводится размер кегля, которым будет сгенерирована подпись к рисунку.

5.7. Окно «Порог выделения цветом» позволяет при визуализации более наглядно показать степень сходства или различий анализируемых объектов. Этим порогом является расстояние по Гауэру, которое исследователь желает видеть между объектами при их кластеризации. Определение этого порога субъективно, выбирается исследователем в зависимости от задач, которые он ставит при кластеризации объектов.

Важно! Выбранное значение коэффициента Гауэра пишется латиницей через точку, а не через запятую!

5.8. В окно «Размер шрифта подписи оси X» вводится размер кегля, которым будут сгенерированы подписи оси «X». На ней будут отображены названия кластеризируемых объектов с постфиксами.

5.9. В окно «Размер шрифта подписи оси Y» вводится размер кегля, которым будут сгенерированы подписи оси «Y». На ней будут отображены расстояния между кластеризируемыми объектами по Гауэру.

5.10. В окно «Размер шрифта названия оси X» вводится размер кегля, которым будет сгенерировано название оси «X».

5.11. В окно «Размер шрифта названия оси Y» вводится размер кегля, которым будет сгенерировано название оси «Y».

5.12. В окно «Размер картинки по оси X» вводится размер картинки в дюймах (1 дюйм = 2,54 см) по горизонтали, который влияет на визуальные расстояния между анализируемыми объектами.

Выбирается произвольно в зависимости от количества кластеризируемых объектов и желаемого размера дендрограммы по горизонтали.

5.13. В окно «Размер картинки по оси Y» вводится размер картинки в дюймах по вертикали, который влияет на визуальные расстояния между коэффициентами Гауэра анализируемых объектов по вертикали.

Выбирается произвольно в зависимости от количества кластеризируемых объектов и желаемого размера дендрограммы.

6. После заполнения полей в командном окне нажать на клавишу «Сгенерировать дендрограмму и таблицу расстояний».

Процесс обработки данных может занять несколько минут (иногда до 15–20), в зависимости от объема матрицы исходных данных и быстродействия компьютера, поэтому не следует компьютер торопить или перезагружать.

Когда данные будут обработаны, программа отдельно предложит сохранить сначала дендрограмму, а потом таблицу расстояний. Путь к файлу и его название можно либо редактировать, либо оставить предложенные программой. По умолчанию сгенерированные файлы появятся в папке, в которой расположена исходная матрица, а названия сгенерированных файлов с результатами кластер-анализа будут соответствовать названию файла с исходной матрицей. На файле с дендрограммой к названию файла будет добавлено слово «дендрограмма», на файле с коэффициентами Гауэра к названию файла будут добавлены слова «таблица расстояний».

7. После окончания работы с программой следует ее корректно закрыть, нажав клавишу «Выход» либо нажав крестик в правом верхнем углу окна.

Алгоритм работы и ввода данных крайне прост, что позволяет использовать все возможности кластер-анализа практически любому пользователю, без обращения к специалистам, имеющим соответствующие знания и доступ к более сложным программам с расширенным функционалом.

Заключение

Апробирование этой программы авторами показывает, что ее возможности вполне достаточны для археолога или другого специалиста, работающего с большими объемами данных.

При возникновении сложностей в работе с программой, предложений по ее улучшению, предложений коллаборации для совместной работы — просьба писать на контактный e-mail: ivan.kazakov@phystech.edu или kaa-2862@mail.ru. Авторов особенно интересуют пожелания по ее совершенствованию.

Таким образом, поставленная цель — сделать возможности кластер-анализа доступными для любого специалиста, работающего с большими объемами данных, по мнению авторов, достигнута.

СПИСОК ИСТОЧНИКОВ

Абдулганеев М.Т., Владимиров В.Н. Типология поселений Алтая 6–2 вв. до н.э. Барнаул : Изд-во Алт. ун-та, 1997. 148 с.

Владимиров В.Н., Степанова Н.Ф. Исследование афанасьевского погребального обряда методом автоматической классификации // Археология Горного Алтая. Барнаул : Изд-во Алт. ун-та, 1994. С. 3–8.

Гайдышев И.П. Моделирование стохастических и детерминированных систем: руководство пользователя программы AtteStat. Курган : Б.и., 2015. 484 с.

Гарскова И.М. Историческая информатика: эволюция междисциплинарного направления. СПб. : Алетейя, 2018. 408 с.

Глазкова А.В. Оценка результативности применения расстояний Евклида и Махаланобиса для решения одной из задач классификации текстов // Вестник Дагестанского государственного технического университета. Технические науки. 2017. №44 (1). С. 86–93. DOI:10.21822/2073-6185-2017-44-1-86-93

Годяев А.А., Гиголаев А.В., Цуканова Н.И. Программа построения самоорганизующихся карт Кохонена при категориальных и смешанных данных // Вестник Рязанского государственного радиотехнического университета. 2017. №61. С. 78–87.

Григоров Е.В., Казаков А.А. Барнаульско-Бийское Приобье в I–XII вв. (по данным погребального обряда). Барнаул : Изд-во Алт. ун-та, 2018. 230 с.

Кишкурно М.С., Зубова А.В. Краниология носителей верхнеобского варианта каменской культуры по материалам могильника Верх-Сузун-5 // Вестник археологии, антропологии и этнографии. 2015. №3 (30). С. 92–103.

Компьютер и историческое знание. Барнаул : Изд-во Алт. ун-та, 1994. 205 с.

Матренин С.С., Тишкин А.А. Опыт выделения локально-территориальных групп населения Алтая хуннуского времени (по материалам погребальных памятников) // Теория и практика археологических исследований. Вып. 3. Барнаул : Изд-во Алт. ун-та, 2007. С. 102–115.

Методы экологических исследований. Основы статистической обработки данных: учебно-методическое пособие / Р.М. Городничев и др. Якутск : Издательский дом СВФУ, 2019. 94 с.

Никитина Г.Ф. Анализ археологических источников могильника Черняховской культуры у села Оселивка. М. : Наука, 1995. 230 с.

Петров П.К. Математико-статистическая обработка и графическое представление результатов педагогических исследований с использованием информационных технологий: учеб. пособие. Ижевск : Удмуртский университет, 2013. 179 с.

Программа STATISTICA. URL: http://portable4pro.ru/development/engineering-programs/statistica.html#ixzz4VEV2ePDO

Серегин Н.Н., Матренин С.С. Погребальный обряд кочевников Алтая во II в. до н.э. — XI в. н.э. Барнаул : Изд-во Алт. ун-та, 2016. 272 с.

Торопчина Г.Н., Двоерядкина Н.Н., Вохминцева Г.П. Элементы кластерного анализа: учеб. пособие. Благовещенск : Амурский гос. ун-т, 2006. 40 с.

Фролов Я.В. Погребальный обряд населения Барнаульского Приобья в VI в. до н.э. — II в. н.э. (по данным грунтовых могильников). Барнаул : Азбука, 2008. 479 с.

Baxter M.J. Notes on Quantitative Archaeology and R. Nottingham. Nottingham, 2015.

Feuerverger Andrey, Stephen M. Stigler, Camil Fuchs, Donald L. Bentley, Sheila M. Bird, Holger Höfling, Larry Wasserman et al. Statistical Analysis of an Archeological Find [with Discussion]. The Annals of Applied Statistics. 2008. No. 3. P. 112.

Kintigh, Keith W. Intrasite Spatial Analysis: A Commentary on Major Methods. Mathematics and Information Science in Archaeology: A flexible Framework. 1990. No. 3. Pp. 165–200.

Ruck, Lana, and Clifford T. Brown. Quantitative Analysis of Munsell Color Data from Archeological Ceramics. Journal of Archaeological Science: Reports. 2015. No. 3. Pp. 549–557.

Troiano, Maurizio, Eugenio Nobile, Flavia Grignaffini, Fabio Mangini, Marco Mastrogiuseppe, Cecilia Conati Barbaro, and Fabrizio Frezza. A Comparative Analysis of Machine Learning Algorithms for Identifying Cultural and Technological Groups in Archaeological Datasets through Clustering Analysis of Homogeneous Data. Electronics. 2024. Vol. 13 No. 14. P. 2752.

REFERENCES

Abdulganeev M.T., Vladimirov V.N. Typology of Altai Settlements of the 6th–2nd Centuries BC. Barnaul : Izd-vo Alt. un-ta, 1997. 148 p. (*In Russ.*)

Vladimirov V.N., Stepanova N.F. Study of Afanasievo Funerary Rites by the Method of Automatic Classification. In: Archaeology of the Altai Mountains. Barnaul : Izd-vo Alt. un-ta, 1994. Pp. 3–8. (*In Russ.*)

Gajdyshev I.P. Modeling Stochastic and Deterministic Systems: AtteStat User's Guide. Kurgan : B.i., 2015. 484 p. (*In Russ.*)

Garskova I.M. Historical Informatics: Evolution of an Interdisciplinary Direction. SPb. : Aletejya, 2018. 408 p. (*In Russ.*)

Glazkova A.V. Performance Evaluation of Applying Euclidean and Mahalanobis Distances to a Text Classification Task. *Vestnik Dagestanskogo gosudarstvennogo tehnicheskogo universiteta. Tehnicheskie nauki = Bulletin of Dagestan State Technical University. Technical Sciences.* 2017;44(1):86–93. (*In Russ.*). DOI:10.21822/2073-6185-2017-44-1-86-93

Godyaev A.A., Gigolaev A.V., Cukanova N.I. A Program for Constructing Self-Organizing Kohonen Maps with Categorical and Mixed Data. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta = Bulletin Ryazan State Radio Engineering University*. 2017;61:78–87. (*In Russ.*)

Grigorov E.V., Kazakov A.A. Barnaul-Biysk Priobye in the I–XII Centuries (according to burial rites). Barnaul : Izd-vo Alt. un-ta, 2018. 230 p. (*In Russ.*)

Kishkurno M.S., Zubova A.V. Craniology of the Carriers of the Upper Ob Variant of the Kamensky Culture on the Materials of the Verkh-Suzun-5 Burial Ground. *Vestnik arheologii, antropologii i etnografii = Bulletin of Archaeology, Anthropology and Ethnography*. 2015;30(3). Pp. 92–103. (*In Russ.*)

The Computer and Historical Knowledge. Barnaul : Izd-vo Alt. un-ta, 1994. 205 p. (*In Russ.*)

Matrenin S.S., Tishkin A.A. Experience in Distinguishing Local-Territorial Groups of Altai Population of the Xiongnu Time (based on the materials of funerary monuments). In: Theory and Practice of Archaeological Research. Issue 3. Barnaul : Izd-vo Alt. un-ta, 2007. Pp. 102–115. (*In Russ.*)

Gorodnichev R.M. et al. Methods of Ecological Research. Fundamentals of Statistical Data Processing: Teaching Aid. Yakutsk : Izdatel`skij dom SVFU, 2019. 94 p. (*In Russ.*)

Nikitina G.F. Analysis of Archaeological Sources of the Burial Ground of the Chernyakhovsky Culture near the Village of Oselivka. Moscow : Nauka, 1995. 230 p. (*In Russ.*)

Petrov P.K. Mathematical and Statistical Processing and Graphical Presentation of Pedagogical Research Results Using Information Technologies: Textbook. Izhevsk : Udmurtskij universitet, 2013. 179 p. (*In Russ.*)

STATISTICA program. URL: http://portable4pro.ru/development/engineering-programs/statistica.html#ixzz4VEV2ePDO (*In Russ.*)

Seregin N.N., Matrenin S.S. Burial Rites of Altai Nomads in the 2nd Century BC — 11th Century A.D. Barnaul : Izd-vo Alt. un-ta, 2016. 272 p. (*In Russ.*)

Toropchina G.N., Dvoeryadkina N.N., Vohminceva G.P. Elements of Cluster Analysis: Training Manual. Blagoveshchensk : Amurskij gos. un-t, 2006. 40 p. (*In Russ.*)

Frolov Ya.V. Burial Rites of the Population of the Barnaul Priobie in the 6th century BC — 2nd Century A.D. (according to the data of ground burial grounds). Barnaul : Azbuka, 2008. 479 p. (*In Russ.*)

Baxter M.J. Notes on Quantitative Archaeology and R. Nottingham. Nottingham, 2015.

Feuerverger Andrey, Stephen M. Stigler, Camil Fuchs, Donald L. Bentley, Sheila M. Bird, Holger Höfling, Larry Wasserman et al. Statistical Analysis of an Archeological Find [with Discussion]. *The Annals of Applied Statistics*. 2008;3:112.

Kintigh, Keith W. Intrasite Spatial Analysis: A Commentary on Major Methods. *Mathematics and Information Science in Archaeology: A flexible Framework*. 1990;3:165–200.

Ruck, Lana, and Clifford T. Brown. Quantitative Analysis of Munsell Color Data from Archeological Ceramics. *Journal of Archaeological Science: Reports*. 2015;3:549–557.

Troiano, Maurizio, Eugenio Nobile, Flavia Grignaffini, Fabio Mangini, Marco Mastrogiuseppe, Cecilia Conati Barbaro, and Fabrizio Frezza. A Comparative Analysis of Machine Learning Algorithms for Identifying Cultural and Technological Groups in Archaeological Datasets through Clustering Analysis of Homogeneous Data. *Electronics.* 2024;13(14):2752.

ВКЛАД АВТОРОВ / CONTRIBUTION OF THE AUTHORS

Казаков А.А.: подготовка технического задания, апробация программного обеспечения, написание соответствующего раздела статьи, редактура текста статьи.

A.A. Kazakov: Preparation of technical specifications, software testing, writing the relevant part of the article, editing the text of the article.

Казаков И.А.: разработка программного обеспечения, его апробация, написание соответствующего раздела статьи, перевод на английский.

I.A. Kazakov: Software development, its testing, writing the relevant part of the article, translation into English.

There is no conflict of interest / Конфликт интересов отсутствует.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Казаков Александр Альбертович, доктор исторических наук, доцент, профессор кафедры истории и философии Барнаульского юридического института МВД России, Барнаул, Россия.

Aleksandr A. Kazakov, Doctor of Historical Sciences, Associate Professor, Professor of the Department of History and Phylosophy of the Barnaul Law Institute of the Ministry of Internal Affairs of the Russian Federation, Barnaul, Russia

Казаков Иван Александрович, аспирант Сколковского института науки и технологий, Москва, Россия.

Ivan A. Kazakov, Doctoral Student, Skolkovo Institute of Science and Technology, Moscow, Russia

Статья поступила в редакцию 23.10.2024;
одобрена после рецензирования 15.11.2024;
принята к публикации 25.11.2024.
The article was submitted 23.10.2024;
approved after reviewing 15.11.2024;
accepted for publication 25.11.2024.

