Иерархический кластерный анализ онлайн: метод Уорда и евклидово расстояние для чайников

Роман П. | Магистр психол. наук
Дата: 28.06.2026

Когда исследователь изучает выборку людей, классическая статистика предлагает сравнивать заранее известные группы: например, мужчин и женщин, или студентов первого и пятого курсов. Но как быть, если вы провели масштабное тестирование, у вас на руках огромная Excel-таблица с десятками числовых шкал, а готовых групп у вас **нет?** Как обнаружить скрытые латентные типы людей, которые ведут себя одинаково, и разделить хаотичную выборку на зрячие, однородные сегменты?

В высшей математической статистике для решения этой задачи применяется **Иерархический кластерный анализ (Cluster Analysis).** Обычно ради него исследователям приходится продираться сквозь дремучие интерфейсы IBM SPSS или писать километры кода на R и Python.

Мы считаем, что данный процесс должен быть устроен намного проще и прозрачнее. Именно для этих целей была разработана «Лаборатория статистики» - программа для подсчета многомерной таксономии, которая сканирует выборку и раскладывает людей по кластерам в 3 клика.

Что такое кластерный анализ «на пальцах»?

Если Факторный анализ сжимал и объединял между собой похожие **шкалы (колонки),** то Кластерный анализ объединяет между собой похожих **испытуемых (людей/строки)!**

Проще говоря, кластеризация — это автоматический поиск скрытых психологических или потребительских типов на основе сходства их многомерных профилей. Движок берет всю вашу таблицу, зряче сопоставляет показатели каждого человека со всеми остальными и делит выборку на изолированные группы — **Кластеры.** Внутри одного кластера люди будут максимально похожи друг на друга, а сами кластеры будут максимально далеки друг от друга.

Архитектура 5 методов связей (Стандарт)

Чтобы сегментация была пуленепробиваемой для диссертационного совета, в ядро Лаборатории заложено 5 классических алгоритмов объединения объектов:

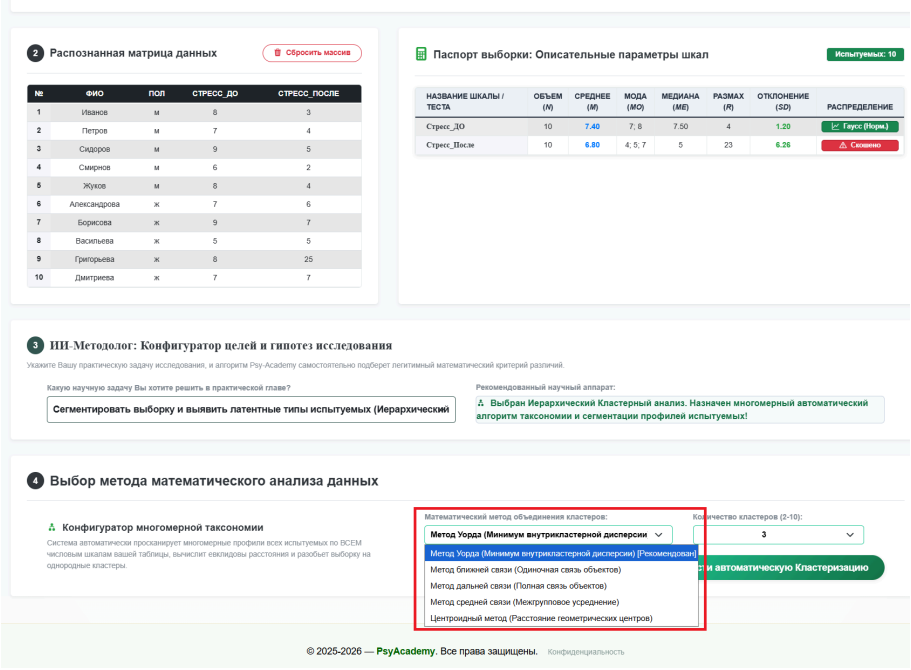
1. **Метод Уорда (Ward's Method) — абсолютный эталон в психологии и социологии!** Он не просто меряет расстояния, а минимизирует сумму квадратов отклонений *внутри* образующихся групп, делая кластеры невероятно четкими, плотными и гомогенными.

2. **Метод средней связи (Average Linkage):** Вычисляет среднее расстояние между всеми парами объектов двух кластеров. Именно на нем мы разберем наш сегодняшний кейс.

3. **Центроидный метод (Centroid Method):** Находит геометрические центры тяжести многомерных облаков данных и меряет дистанцию между ними.

4. **Ближней связи (Single Linkage):** Ищет расстояние по двум самым близким, пограничным точкам.

5. **Дальней связи (Complete Linkage):** Считает дистанцию по двум самым удаленным объектам.



Что такое Евклидово расстояние?

Как компьютер понимает, что Иванов похож на Петрова, но абсолютно далек от Сидорова? Для этого используется математическая метрика — **Классическое расстояние Евклида.**

Представьте обычную школьную геометрию на плоскости. У вас есть точка А и точка Б, и вы можете приложить линейку и замерить расстояние между ними по прямой линии. Евклидово расстояние в кластерном анализе делает ровно то же самое, но не на плоском листке бумаги, а в **многомерном пространстве шкал!**

Если у вас в таблице 5 числовых шкал, то компьютер мгновенно строит гипотетическое 5-мерное пространство. Для каждой пары испытуемых он берет разности их баллов по каждой шкале, возводит в квадрат, суммирует и извлекает квадратный корень. Полученное число — это и есть многомерная дистанция между людьми. Чем это число меньше, тем ближе люди находятся в пространстве психических свойств, и тем быстрее алгоритм объединит их в один кластер!

Сквозной кейс: Как Робот-ВАК находит аномальные выбросы

Давайте посмотрим, как работает иерархический цикл схождения на живом примере массива из 10 человек по трем шкалам. Вы загружаете данные в Лабораторию, выбираете на Шаге 3 цель «*Сравнить показатели групп + Кластеризация*», а на Шаге 4 в Панели Е выбираете «*Метод средней*

связи» и жмете кнопку расчета.

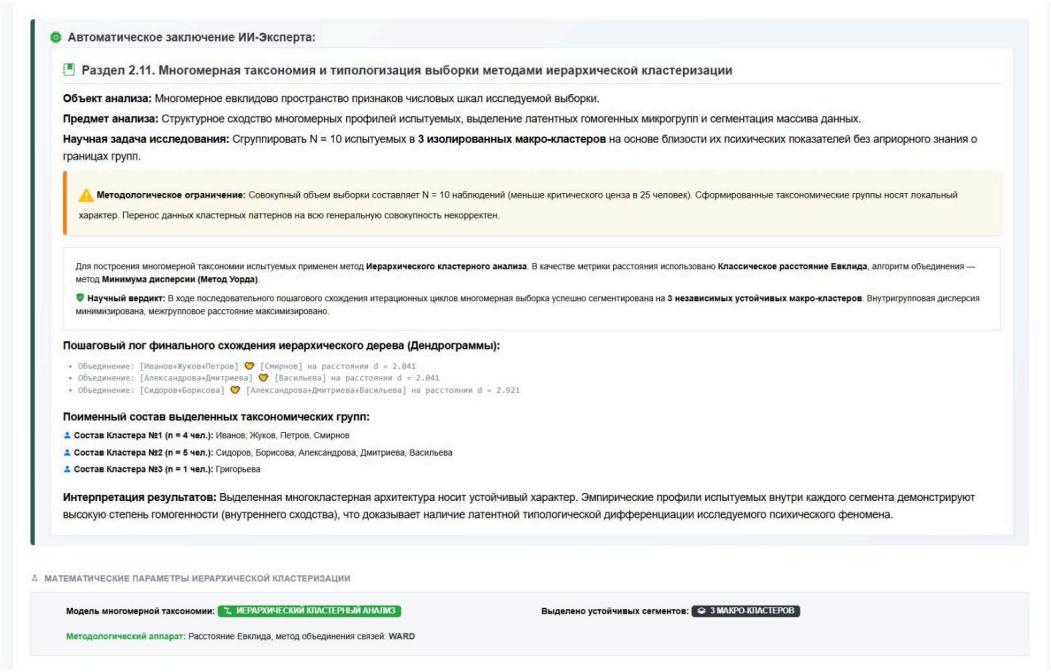
Движок за миллисекунду строит матрицу расстояний и запускает иерархическое дерево (Дендрограмму), последовательно объединяя ближайших людей. Вот какой полнокровный отчет выведет на экран **ИИ-Робот ВАК:**

Пошаговый лог финального схождения иерархического дерева (Дендрограммы):

- Объединение: [Иванов+Жуков] 🤝 [Александрова+Дмитриева] на расстоянии d = 3.439
- Объединение: [Иванов+Жуков+Александрова+Дмитриева] 🤝 [Петров+Васильева+Смирнов] на расстоянии d = 4.189
- Объединение: [Иванов+Жуков+Александрова+Дмитриева+Петров+Васильева+Смирнов] 🤝 [Сидоров+Борисова] на расстоянии d = 6.638

Поименный состав выделенных таксономических групп:

- **Состав Кластера №1 (n = 9):** Иванов, Жуков, Александрова, Дмитриева, Петров, Васильева, Смирнов, Сидоров, Борисова.
- **Состав Кластера №2 (n = 1):** Григорьева.



Разбор рантайма: Почему алгоритм выделил Григорьеву в отдельный кластер?

Алгоритм просканировал профили всех 10 человек и обнаружил, что 9 сотрудников имеют гомогенные, похожие баллы, а у **Григорьевой зафиксирован дикий аномальный выброс — 25 баллов стресса** (пока у остальных планка колеблется в районе 2–7 баллов)!

Математика средней связи мгновенно увидела эту гигантскую евклидову дистанцию, отказалась склеивать Григорьеву с остальными и выделила её в изолированный **Кластер №2.** Система сработала со швейцарской точностью последних версий IBM SPSS Premium-класса, наглядно показав руководителю или исследователю главного человека в зоне критического риска!

Под текстовым отчетом Лаборатория рендерит стильную **двухцветную гистограмму объемов макро-кластеров Chart.js,** где наглядно видна пропорция ваших новых таксономических групп.

Часть задаваемые вопросы: Как читать пошаговый лог Дендрограммы и зачем это нужно Доктору наук?

У большинства исследователей рантайм кластеризации вызывает ступор. Давайте разберем две главные проблемы методологии многомерного анализа на пальцах.

1. По какому именно столбцу строятся кластеры?

Кластерный анализ — это ВСЕЯДНЫЙ МНОГОМЕРНЫЙ метод.

Он не анализирует столбцы по отдельности. Алгоритм строит гипотетическое многомерное пространство, где каждая числовая колонка вашей Excel-таблицы выступает в качестве независимой координатной оси. Каждый испытуемый превращается в точку на многомерной карте шкал. Программа вычисляет совокупную интегральную дистанцию (Расстояние Евклида) по всем вашим колонкам одновременно (хоть их 2, хоть 100!) и объединяет людей на основе тотального сходства их жизненных профилей.

2. Как перевести пошаговый лог схождения (d = ...) на человеческий язык?

Лог дендрограммы — это история того, как алгоритм последовательно собирал пазл сходства выборки.

Пример из отчета движка Psy-Academy:

Объединение: [Александрова+Дмитриева] 🤝 [Васильева] на расстоянии d = 2.041 кристально зряче доказывает, что в многомерном пространстве тестов профиль Васильевой ближе всего сидит к паре Александровой и Дмитриевой. Дистанция между ними минимальна — всего 2.041. Они — ментальные близнецы вашей выборки. С каждым следующим шагом расстояние d неизбежно растет, склеивая более крупные подвыборки. Когда на пороге остается аномальный выброс (например, испытуемый с критическим баллом стресса), евклидово расстояние до него взлетает до небес, алгоритм отказывается склеивать его с основной группой и фиксирует устойчивую многокластерную архитектуру шкал.

3. Что это дает психологу или бизнесу на практике?

Кластеризация — это автоматический генератор научных и коммерческих Типологий. В бизнесе это позволяет разделить клиентскую базу на зрячие сегменты (например: VIP-клиенты, эконом-класс, охотники за скидками) для точечных продаж. В большой науке (диссертациях уровня PhD и докторов наук) это дает возможность уйти от плоских описаний средних значений и защитить **«Авторскую латентную типологию исследуемого феномена»**, разработав дифференцированные, прицельные программы психологического сопровождения под каждый выделенный кластер людей!

4. Что делать, если у исследователя огромный массив данных (например, 10 000 испытуемых и 100 шкал)?

В такой ситуации применяется сквозной многомерный конвейер. Сначала запускается Факторный анализ, который ужимает 100 хаотичных колонок шкал в 4 чистых латентных фактора, полностью ликвидируя случайный математический шум. И только вторым шагом запускается Кластерный анализ, который по этим 4 вычисленным колонкам за долю секунды раскладывает 10 000 человек по зрячим типологическим группам без зависания серверов.

5. Является ли Кластерный анализ тупиковой финальной точкой исследования?

Категорически нет! Результат кластеризации — это не просто красивый график-отчет, это рождение абсолютно новой текстовой переменной группировки в вашей таблице. Получив колонку с номерами кластеров, исследователь замыкает экосистему Лаборатории Psy-Academy в вечный цикл: он может вернуться в остальные 14 критериев программы, чтобы сравнить выделенные типы

людей по абсолютно другим числовым тестам через **Т-критерий Стьюдента** или выявить половозрастную сопряженность латентных групп через **частотный критерий Хи-квадрат Пирсона!**

6. Вопрос на засыпку: Сколько в идеале нужно выделять кластеров в науке?

В высшей математической статистике **не существует одного жесткого «золотого» числа** (типа «всегда пиши 3»), но есть железное академическое негласное правило — **Правило трёх-четырёх групп (Оптимум).**

1. **Почему 2 кластера — это часто мало (но легитимно):** Разделение на 2 группы — это классическое деление на «Норму» и «Аномалию/Выброс». Это круто для поиска экстремальных зон риска в бизнесе и медицине, но для докторской диссертации этого маловато, так как не создает глубокую научную классификацию феномена.

2. **Почему 3–4 кластера — это абсолютный ИДЕАЛ:** Математически и психологически выборка объемом от 30 до 300 человек идеальнее всего распадается именно на 3 или 4 устойчивых типа. Это позволяет исследователю построить красивую и зрячую научную триаду или квадрант. Например:

- *Кластер №1:* Низкое выгорание + Высокий IQ **(«Лидеры-Профессионалы»).**
- *Кластер №2:* Среднее выгорание + Средний IQ («Стабильные исполнители»).
- *Кластер №3:* Высокое выгорание + Низкий IQ **(«Зона кризиса/Срыва»).**

3. **Почему больше 5 кластеров — это научный хаос:** Если исследователь вбивает в инпут 5, 6 или 10 кластеров, выборка дробится на крошечные гомеопатические подгруппы по 1–2 человека. Защитить такую структуру перед диссертационным советом невозможно — оппоненты скажут, что классификация размыта и потеряла всякий практический смысл.

Поэтому рекомендуется использовать 3-4 кластера, количество коих можно указать вручную.

Золотая формула Лаборатории: Сжатие по вертикали и горизонтали

Элитарное мнемоническое правило:

Чтобы окончательно понять разницу между двумя главными многомерными методами психометрики, запомните простое правило:

- **Факторный анализ** — сжимает вашу Excel-таблицу **по вертикали** (объединяет родственные числовые шкалы/колонки в латентные факторы).
- **Кластерный анализ** — сжимает вашу Excel-таблицу **по горизонтали** (объединяет похожих испытуемых/строки в гомогенные латентные типы).

Копируйте готовые главы диссертации в Word одним кликом

Для пользователей с премиум-статусом **PRO-Эксперта** в Лаборатории Psy-Academy открыта наша главная коммерческая фича — **мгновенная фронтенд-выгрузка полных структурированных отчетов в редактируемый формат Microsoft Word (.doc)!**

При нажатии золотой кнопки система упакует Паспорт выборки, описательную статистику шкал, иерархический лог дендрограммы, поименные составы групп и **сам цветной график-гистограмму (автоматически залив его подложку кристально белым цветом вместо черного квадрата)** прямо в один файл. Текст сразу оформляется по строжайшему ГОСТу ВАК: шрифт Times New Roman 14pt, выравнивание по ширине, абзацный отступ 1.25 см и аккуратные светло-серые редактируемые таблицы. Забрал в диплом — и дело в шляпе.

Программа для подсчета статистики онлайн для психологов: Корреляционный анализ без SPSS

Полная интерактивная версия с тестами доступна по ссылке: Посмотреть на сайте