МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ МЕТРИК И УЛЬТРАМЕТРИК

(автореферат бакалаврской работы)

студента 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Тарабрина Данилы Валерьевича

Научный руководитель
Доцент, кандидат ф.-м. наук _____ Л.Б. Тяпаев
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2025

ВВЕДЕНИЕ.

Актуальность темы исследования “Кластеризация на основе метрик и ультраметрик” состоит в том, что стремительный рост объёмов данных в различных областях, включая социальные и экономические процессы, повышает актуальность их анализа. Развитие цифровых технологий и методов обработки больших данных предоставляет новые возможности для углублённого изучения этих процессов, особенно в контексте динамики и взаимосвязей сложных систем, где традиционные методы анализа могут быть недостаточными [1].

Кластерный анализ, который применяется для группировки объектов на основе их характеристик, является важным инструментом в таких областях, как экономика, социология и экология, для работы с большими многомерными данными [2].

Страны ОЭСР, обладая различиями в экономическом развитии и социальной структуре, являются подходящими объектами для анализа. Временные ряды социально-экономических показателей этих стран позволяют выявить закономерности и взаимосвязи, способствующие пониманию глобальных процессов [3].

Степень разработанности проблемы исследования: Кластерный анализ широко применяется для обработки больших данных в таких областях, как экономика, социология и демография. Методы кластеризации помогают выявить группы объектов с похожими характеристиками, что важно для анализа структуры и динамики сложных систем. Иерархическая кластеризация, с её визуализацией в виде дендрограмм, стала популярной благодаря наглядности [4].

Использование бинарных префиксных кодов для представления дендрограмм упрощает их сравнение и позволяет более точно оценивать сходства с помощью ультраметрических расстояний, расширяя возможности анализа.

Применение кластеризации в социально-экономических исследованиях помогает выявить группы стран с схожими динамическими моделями развития.

Интеграция метода префиксных кодов и ультраметрических метрик для анализа временных рядов остаётся перспективной областью для дальнейшего изучения.

Настоящее исследование направлено на разработку и практическую реализации комплексного подхода к кластерному анализу социально-экономических временных рядов с использованием метрик и ультраметрик.

В рамках настоящего исследования были поставлены следующие задачи:

1. Обзор существующих методов кластерного анализа - анализ классических и современных методов кластеризации.

2. **Описание методологии анализа дендрограмм, интерпретируемых как максимальные префиксные коды** - описание методологии для представления дендрограмм в виде бинарных префиксных кодов и использования метрик и ультраметрик для их сравнения.

3. Практическая реализация программы по анализу социально-экономических признаков стран ОЭСР - разработка программного инструмента для реализации предложенной методологии анализа временных рядов с использованием ультраметрик и метрик.

Объектом исследования являются социально-экономические данные стран ОЭСР, представленные в виде временных рядов. Эти данные включают различные социальные и экономические показатели, такие как валовой внутренний продукт (ВВП), население, продолжительность жизни, выбросы CO2 и другие параметры, характеризующие развитие стран.

Предметом исследования является метод анализа временных рядов социально-экономических показателей стран ОЭСР, с использованием метрик и ультраметрик для анализа и сравнения иерархических структур, представленных в виде дендрограмм.

Структура ВКР представлена в виде: введения, трех глав, заключения, списка использованных источников и приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ.

В первой главе «Теоретические аспекты исследования социально-экономических характеристик стран ОЭСР» раскрываются основные понятия анализа социально-экономических показателей стран ОЭСР.

Анализ социально-экономических показателей стран ОЭСР направлен на изучение механизмов развития и выявление взаимосвязей в социально-экономической динамике. ОЭСР, включающая 38 развитых экономик, обеспечивает доступ к стандартизированным данным высокого качества [5], а их экономическая и институциональная однородность минимизирует структурные различия, позволяя сосредоточиться на вариациях в социальной политике, экологической ответственности и экономической диверсификации [6].

Обширные временные ряды данных ОЭСР охватывают десятилетия, что позволяет анализировать долгосрочные тенденции. Иерархическая кластеризация учитывает временную динамику, группируя страны по траекториям экономического роста, социальных и экологических показателей. Стандартизированные методы сбора данных упрощают интерпретацию результатов, способствуя разработке рекомендаций в области международных отношений и устойчивого развития.

Для анализа использовались показатели, отражающие экономическое, социальное и экологическое развитие: ВВП, ВВП на душу населения, численность населения, выбросы CO2, энергоемкость экономики, ожидаемая продолжительность жизни, детская смертность, рождаемость, смертность, уровень занятости, занятость в сельском хозяйстве и расходы на образование. Эти показатели характеризуют экономическую мощь, благосостояние, экологическую устойчивость, состояние здравоохранения, рынок труда и инвестиции в человеческий капитал.

В заключение можно сказать, что выбранные показатели легли в основу кластерного анализа и иерархической кластеризации временных рядов, позволяя

выявить закономерности и взаимосвязи в социально-экономических процессах с высокой точностью и сопоставимостью результатов.

Во второй главе «Теоретические основы кластерного анализа» раскрываются основные понятия кластерного анализа социально-экономических показателей стран ОЭСР. Кластерный анализ — статистический метод, направленный на группировку объектов в кластеры, где объекты внутри кластера схожи, а между кластерами различны. Он применяется в экономике, социологии, экологии и биологии для выявления скрытых закономерностей в данных без предварительных меток, что делает его универсальным в условиях неопределенности [7].

Кластерный анализ включает иерархическую и неиерархическую кластеризацию. Иерархическая строит дендрограмму, отображая последовательность объединения или разделения объектов. Неиерархические методы, такие как K-средних [8] и DBSCAN [9], требуют задания количества кластеров и обладают высокой вычислительной эффективностью.

Выбор метрики расстояния критически важен, так как она определяет различия между объектами (например, временными рядами или векторами) и влияет на распределение по кластерам.

Евклидова метрика является одной из самых распространённых и интуитивно понятных метрик. Она рассчитывает расстояние между двумя точками $x = (x_1,x_2,\ldots,x_n)$ и $y = (y_1,y_2,\ldots,y_n)$ в пространстве n-мерных векторов как корень из суммы квадратов различий между соответствующими компонентами:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Эта метрика измеряет кратчайшее расстояние между двумя точками в пространстве. По сути, это длина отрезка, соединяющего две точки. Используется в большинстве стандартных алгоритмов кластеризации, таких как K-средних, так как она легко интерпретируется и вычисляется. Евклидова

метрика эффективно работает, когда данные являются непрерывными и имеют схожие масштабы, однако она может быть чувствительна к выбросам в данных и плохо справляется с неевклидовыми структурами.

Манхэттенская метрика измеряет расстояние между двумя точками как сумму абсолютных разностей между их компонентами:

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Она более устойчива к выбросам по сравнению с евклидовой метрикой, что делает её подходящей для данных с редкими аномалиями. Метрика эффективна для сеточных данных, таких как координаты или сеточные модели. Однако она менее точно отражает глобальную схожесть объектов, оценивая различия только по отдельным признакам, что может привести к ошибкам при сложной структуре данных.

Расстояние Чебышёва измеряет максимальное различие между компонентами двух объектов:

$$d(x,y) = \max_i |x_i - y_i|$$

Метод эффективен для выделения наибольшего различия по любому признаку и отличается высокой скоростью вычислений, что подходит для больших данных. Однако он учитывает только максимальное отклонение, игнорируя остальные различия, что может привести к потере информации о других значимых признаках, особенно при анализе мелких различий между объектами.

Расстояние Хеллингера используется для измерения различий между двумя вероятностными распределениями. В случае дискретных данных оно вычисляется по следующей формуле:

$$d(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} (\sqrt{P_i} - \sqrt{Q_i})^2}$$

где P_i и Q_i — это вероятностные плотности для двух распределений P и Q в точке i.

Метод эффективен для сравнения распределений, особенно при анализе временных рядов или данных с редкими событиями, благодаря низкой чувствительности к таким событиям. Однако он неприменим, если данные нельзя представить как вероятностные распределения или они имеют дискретные значения, не являющиеся вероятностными оценками.

Выбор метрики расстояния определяется структурой данных, их масштабом, наличием выбросов и целями кластеризации. Евклидова метрика подходит для данных с нормальным распределением и сферической структурой кластеров, манхэттенская — для сеточных или категориальных данных, расстояние Чебышёва — для задач, где важны максимальные отклонения, а расстояние Хеллингера — для вероятностных распределений.

Преобразование дендрограмм в бинарные префиксные коды упрощает анализ и сравнение иерархических структур. Этот метод кодирует путь от корня дендрограммы к каждому листу, обеспечивая компактное представление и облегчая вычисление расстояний между дендрограммами. Бинарные коды делают анализ более структурированным, позволяя применять метрики для оценки сходства иерархий, особенно при работе с многомерными данными, такими как социально-экономические показатели. Это обеспечивает точное определение сходства и различий на разных уровнях иерархий, что важно для анализа сложных данных.

Архимедова метрика используется в кластерном анализе для измерения различий между дендрограммами, представленными бинарными префиксными кодами. Она основана на сравнении длин ветвей дендрограмм, учитывая их вложенность и топологическое сходство.

Формула для вычисления архимедовой метрики выглядит следующим образом:

$$\delta(a,b) = \sqrt{\sum_{n=1}^{m} (2^{-\Lambda(\omega_a\ (n))} - 2^{-\Lambda(\omega_b\ (n))})^2}$$

где $\Lambda(\omega_a\ (n))$ и $\Lambda(\omega_b\ (n))$ — это длины ветвей $\omega_a\ (n)$ и $\omega_b\ (n)$, которые ведут к конечной вершине с номером n в дендрограммах a и b, соответственно. Эта метрика основана на вычислении разницы между длинами соответствующих ветвей в дендрограммах на каждом уровне иерархии. Учитывая, что длина ветви отражает степень схожести между объектами, архимедова метрика позволяет более детально анализировать структуры дендрограмм на разных уровнях иерархий [10].

Архимедова метрика обладает несколькими важными преимуществами. Она учитывает вложенность иерархий и позволяет измерять расстояние между дендрограммами с учётом их структуры. Это даёт возможность более точно моделировать топологические различия, которые могут быть не видны при применении иных метрик.

Ультраметрическое расстояние используется для сравнения дендрограмм и является основой для более точного и глубокого анализа иерархий, чем традиционные метрики расстояния. Основным подходом в ультраметрическом расстоянии является представление дендрограмм как максимальных префиксных кодов, которые интерпретируются через 2-адическую структуру.

Ультраметрическое расстояние между двумя дендрограммами определяется через соответствие их бинарных кодов на разных уровнях вложенности. Сравниваются префиксные коды, соответствующие ветвям, и расстояние между двумя дендрограммами a и b вычисляется как степень, до которой их коды совпадают по модулю 2^k. Математически это представляется следующим образом:

$$d(a,b) = 2^{-k}$$

где k — максимальное натуральное число, при котором коды ветвей дендрограмм a и b совпадают по модулю 2^k. [10]

В заключение можно сказать, что архимедова метрика, основанная на сравнении длин ветвей дендрограмм, обеспечивает точный анализ топологического сходства и вложенности иерархий, позволяя выявлять различия в структурах данных. Ультраметрическое расстояние, рассматривающее дендрограммы как 2-адические структуры, повышает точность сравнения иерархий за счёт учёта максимальных совпадений префиксных кодов на разных уровнях, что особенно эффективно для анализа сложных многомерных данных.

В третьей главе «Практическая реализация программы по анализу социально-экономических признаков стран ОЭСР» раскрываются основные шаги работы программы по анализу социально-экономических показателей стран ОЭСР. Кластерный анализ позволяет группировать страны ОЭСР по сходству социально-экономических характеристик, выявляя закономерности и иерархии в данных, что особенно важно для анализа временных рядов.

Предобработка временных рядов, таких как ВВП, выбросы CO2 и продолжительность жизни, необходима для повышения точности кластеризации. Пропуски в данных устраняются методами интерполяции или заполнения значениями соседних периодов, сохраняя динамику и минимизируя потери информации. Нормализация методом Min-Max приводит показатели к единому диапазону, обеспечивая их сопоставимость и корректное применение метрик расстояния, таких как евклидова и манхэттенская [11].

Качественная предобработка данных формирует основу для иерархической кластеризации, улучшая анализ долгосрочных тенденций и взаимосвязей между социально-экономическими показателями, что способствует точным выводам о развитии стран ОЭСР.

Для измерения сходства между временными рядами признаков применяется расстояние Хеллингера - метрика, эффективная для сравнения распределений и неотрицательных векторов. Расчёт расстояния производится по формуле:

$$H(x,y) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} (\sqrt{x_i} - \sqrt{y_i})^2}$$

где x и y - сравниваемые векторы.

Для агломеративной иерархической кластеризации выбран метод средней связи (average linkage). Этот метод основан на вычислении расстояния между кластерами как среднего арифметического расстояний между всеми парами объектов, принадлежащих разным кластерам. Формально, расстояние между двумя кластерами A и B определяется как:

$$d(A,B) = \frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{x \in B} d(x,y)$$

где |A| и |B| - количество элементов в кластерах A и B соответственно, а d(x,y) - расстояние между объектами x и y.

Преимущество метода среднего объединения заключается в том, что он учитывает структуру всего кластера и менее подвержен влиянию выбросов по сравнению с методами минимального (single linkage) или максимального (complete linkage) расстояния. Это делает его более стабильным и пригодным для анализа сложных многомерных данных, таких как временные ряды социально-экономических показателей.

В результате применения метода кластерного анализа формируется дендрограмма, которая отображает последовательное объединение объектов и кластеров по возрастанию расстояния между ними. Эта структура помогает выявить естественные иерархии и группировки в данных, что важно для глубокого анализа и интерпретации.

Далее строятся структуры отношений, определяющие иерархию и пути в дереве. Для каждого узла дендрограммы рекурсивно вычисляется двоичный префиксный код, где переход влево кодируется как «0», а вправо — как «1». Такое кодирование обеспечивает компактное представление иерархической структуры данных.

Полученные дендрограммы сохраняются для каждой страны, и формируются словари, которые сопоставляют каждому признаку его бинарный код. Для визуализации создаются графики дендрограмм с подписями признаков и их кодов, что облегчает анализ и интерпретацию данных.

Результаты кодирования дендрограмм выводятся в виде префиксных бинарных кодов для каждого признака по всем анализируемым странам. Для удобства интерпретации введен заранее определённый порядок признаков, что отражает значимость социально-экономических индикаторов, таких как ВВП, население, выбросы CO2, продолжительность жизни и другие ключевые параметры.

Эти коды позволяют эффективно сравнивать иерархии и оценивать сходства между странами с помощью таких метрик, как архимедова метрика.

Архимедова метрика позволяет точнее сравнивать иерархические отношения между странами, особенно когда данные включают долгосрочные временные ряды, отражающие изменения в социально-экономическом контексте. Программа реализует вычисление расстояния между двумя дендрограммами на основе длин их префиксных кодов. Архимедова метрика рассчитывается по формуле:

$$d(A,B) = \sqrt{\sum_{n=1}^{m} (2^{-\lambda_a(n)} - 2^{-\lambda_b(n)})^2}$$

где m — общее количество признаков, λa(n) и λb(n) — длины бинарных кодов (длины ветвей) признаков n-го элемента дендрограмм A и B соответственно, отражающие путь от корня до соответствующей ветви.

Основная идея метода заключается в том, что различия в длинах кодов бинарных префиксов между двумя дендрограммами количественно характеризуют различия в их иерархической структуре.

Далее, для всех стран из набора данных формируется матрица попарных расстояний, где каждая ячейка содержит значение архимедовой метрики между парой стран. Коды признаков для каждой страны извлекаются в порядке,

заданном заранее определённым списком признаков, что обеспечивает сопоставимость вычислений.

Далее реализована процедура вычисления ультраметрического расстояния между дендрограммами, представленными в виде наборов бинарных префиксных кодов признаков, что позволяет количественно оценить степень сходства их иерархической структуры.

Процесс вычисления ультраметрического расстояния основан на поэтапном анализе 2-адических редукций бинарных кодов. На каждом уровне k, из каждой бинарной строки выделяются k младших бит, которые затем интерпретируются как десятичные числа.

Затем множества редукций двух дендрограмм сравниваются на совпадение. Максимальное значение k, при котором множества совпадают, фиксируется как показатель уровня структурного совпадения между иерархиями. Если на уровне k редукции не совпадают, процесс сравнения прерывается.

Ультраметрическое расстояние между дендрограммами вычисляется как:

$$p = \frac{1}{2^k}$$

где ρ отражает степень различия: чем выше k, тем меньше расстояние и тем ближе структуры иерархий.

Далее для всех пар стран с помощью ультраметрики вычисляется матрица расстояний, которая выносится в двумерный массив и преобразуется в двумерную табличную структуру данных для удобства анализа и визуализации. Матрица наглядно демонстрирует степени сходства и различий между социально-экономическими иерархиями стран, выявленными посредством кластеризации.

Для полного отображения результатов в pandas отключается усечение строк и столбцов, а данные выводятся с округлением до четырёх десятичных знаков. Итоговые данные сохраняются в форматах Excel и CSV для дальнейшего анализа.

Визуализация дендрограмм с использованием бинарных префиксных кодов отобразила структуру кластеров и взаимосвязи между странами ОЭСР на основе социально-экономических показателей: ВВП, население, ВВП на душу населения, выбросы CO2, использование энергии, продолжительность жизни, детская смертность, рождаемость, смертность, занятость, занятость в сельском хозяйстве и расходы на образование. Эти показатели отражают текущие значения и динамику, позволяя анализировать долгосрочные тренды.

Анализ матрицы расстояний с архимедовой метрикой выявил степень схожести социально-экономических структур. Например, расстояние 0.2451 между Австралией и Канадой указывает на высокое сходство их экономических моделей, уровня жизни и экологической политики, тогда как расстояние 0.4927 между Грецией и Швецией отражает значительные различия. Аналогично, Австралия и Бельгия (0.2619) близки по характеристикам, но различаются в политических системах или подходах к устойчивому развитию, а Австрия и Чили (0.4292) демонстрируют различия, обусловленные географическими и историческими факторами.

Ультраметрическая метрика, учитывающая иерархическую структуру данных, подтвердила близость Австралии и Канады (0.125) по социально-экономическим показателям и подходам, в то время как Австрия и Турция (0.25) показали значительные различия в социальной и экономической структуре.

В заключение можно сказать, что использование архимедовой и ультраметрической метрик позволяет глубже понять взаимосвязи между странами ОЭСР. Обе метрики дают уникальные представления о сходствах и различиях между странами, что помогает выделить ключевые закономерности и оценить влияние долгосрочных изменений на их развитие. Методология, основанная на этих метриках, показала свою гибкость и эффективность для анализа социально-экономических характеристик стран и формирования более точных и обоснованных рекомендаций для международных политических стратегий и устойчивого развития.

ЗАКЛЮЧЕНИЕ

В ходе исследования была разработана методология кластерного анализа временных рядов социально-экономических показателей стран ОЭСР с использованием архимедовой и ультраметрической метрик. Созданный программный продукт автоматизировал процессы предобработки данных, нормализации, построения иерархических дендрограмм и их сравнения, что обеспечило эффективную работу с большими многомерными данными и повысило точность анализа.

Основным результатом работы является метод количественного сравнения сходства и различий между иерархическими структурами данных, представленными в виде бинарных префиксных кодов. Этот подход позволил выявить скрытые закономерности и взаимосвязи в социально-экономических показателях стран ОЭСР. Анализ с использованием ультраметрических расстояний позволил выделить группы стран с схожими моделями развития и указать на различия, обусловленные уникальными характеристиками.

Разработанный инструмент обеспечивает воспроизводимость и масштабируемость анализа, что позволяет применять методологию к данным других регионов и временных периодов. Это открывает перспективы для более глубокого изучения динамики социально-экономических процессов и разработки политических рекомендаций.

Таким образом, выполненная работа предоставляет эффективный инструмент для кластерного анализа стран ОЭСР и способствует развитию методов анализа временных рядов, что может быть полезным для дальнейших экономических исследований и разработки стратегий для устойчивого развития.