

УДК 656.072

DOI: 10.34670/AR.2024.95.50.025

Алгоритмы K-means и модель XGBoost в транспортной отрасли**Мадяр Ольга Николаевна**

Кандидат технических наук, доцент,
Российский университет транспорта,
127994, Российская Федерация, Москва, ул. Образцова, 9;
e-mail: zox2.777@mail.ru

Попова Елена Александровна

Кандидат технических наук, доцент,
Воронежский филиал,
Ростовский государственный университет путей сообщения,
394026, Российская Федерация, Воронеж, ул. Урицкого, 82;
e-mail: Elena-1504@mail.ru

Аннотация

В данном исследовании использовалась аналитика больших данных для изучения влияния территориальных факторов и особенностей транспортной сети на распределение пассажиропотока. Для обеспечения принятия решений дифференцированных стратегий оперативного управления для различных типов железнодорожных станций с целью повышения оптимизации работы и устойчивого развития городских железнодорожных систем в статье сравнивались результаты кластеризации с использованием различных функций измерения сходства в рамках алгоритма K-means и выбиралась оптимальная функция измерения сходства для построения моделей кластеризации. Также в статье рассматривались факторы, влияющие на распределение пассажиропотока, на основании которых был сформирован набор признаков, которые в сочетании с результатами кластеризации модели послужили входом для модели XGBoost для анализа взаимосвязи между различными признаками и распределением пассажиропотока на станции.

Для цитирования в научных исследованиях

Мадяр О.Н., Попова Е.А. Алгоритмы K-means и модель XGBoost в транспортной отрасли // Экономика: вчера, сегодня, завтра. 2024. Том 14. № 11А. С. 226-239. DOI: 10.34670/AR.2024.95.50.025

Ключевые слова

Распределение пассажиропотока, K-средних, K-means, кластеризация, распределение пассажиропотока, МЦД-3, факторы распределения, моделирование, временные ряды, XGBoost.

Введение

Изучение классификации станций городского железнодорожного транспорта на основе характеристик распределения пассажиропотока, а также анализ взаимосвязи между распределением пассажиропотока и характеристиками транспортной сети помогает уточнить стратегии распределения ресурсов для различных типов станций. Это помогает снизить эксплуатационные расходы железнодорожного транспорта и добиться рациональной эксплуатации.

Ученые широко исследовали особенности пассажиропотока на железнодорожных станциях, изучая такие аспекты, как пассажиропотоки при посадке и высадке [Пазойский, Савельев, Середов, 2022], пассажиропотоки в часы пик [Филиппов, Смирнов, Язев, Александров, 2021] и пассажиропотоки в будние дни по сравнению с пассажиропотоками в выходные. С развитием информационных технологий данные исследований перешли от традиционных опросных данных к большим данным, что позволило более точно выразить показатели влияющих факторов [Пазойский, Савельев, Середов, 2022; Кокорев, Корнеева, Сидоренко, Шаш, 2021]. Для анализа взаимосвязи между характеристиками пассажиропотока на городских железнодорожных станциях и влияющими факторами широко использовались модели линейной регрессии [Кокорев, Корнеева, Сидоренко, Шаш, 2021; Ефимов, Шапкин, Шмаль, Минаков, [www...](#)]. Чтобы лучше выявить общие черты между станциями, многие ученые используют алгоритмы кластеризации для классификации станций перед изучением характеристик пассажиропотока и влияющих факторов для различных типов станций.

Несмотря на обширные исследования характеристик и факторов, влияющих на пассажиропоток на городских железнодорожных станциях, исследований, специально посвященных распределению пассажиропотока на городских железнодорожных станциях на этапах эксплуатации, относительно мало. Кроме того, существующие исследования по классификации станций часто опираются на упрощенные алгоритмы кластеризации, упуская из виду значительное влияние качества кластеризации на последующие анализы, а многие анализы влияющих факторов по-прежнему в значительной степени опираются на линейные регрессионные модели, игнорируя потенциальные нелинейные связи между характеристиками пассажиропотока на станции и влияющими факторами, а также ограничения линейных регрессионных моделей при решении задач классификации.

Поэтому в статье рассматриваются факторы, влияющие на распределение пассажиропотоков на станциях городского железнодорожного транспорта с точки зрения классификации станций. Понимание этих факторов имеет решающее значение для оптимизации управления станциями и повышения качества обслуживания, что, в свою очередь, способствует устойчивости городских транспортных систем за счет эффективного использования ресурсов и снижения эксплуатационных расходов. В модуле классификации станций использовался подход кластеризации временных рядов, а также выбор соответствующих функций измерения сходства для повышения эффективности кластеризации. В модуле анализа влияющих на распределение пассажиропотока факторов использовались возможности классификации и способность модели XGBoost к изучению нелинейных связей. Наконец, объединив результаты двух модулей, были проанализированы временные и пространственные характеристики различных типов станций, чтобы создать основу для принятия решений по уточнению оперативного управления железнодорожными станциями.

Основное содержание

Исследование железнодорожных станций МЦД-3 показало, что такие факторы, как протяженность пешеходных дорог в зоне охвата, количество автобусных маршрутов, наличие парковочных мест и возможность пересадки на другие виды транспорта оказывают значительное влияние на пассажиропоток. С развитием информационно-коммуникационных технологий большие данные, особенно пространственно-временные данные, широко используются в различных исследованиях. OpenStreetMap может предоставить более точные данные о дорожно-транспортной инфраструктуре, что позволяет рассчитывать такие характеристики, как длина дорог в рамках РСА [Кулешова, Широченко, 2013].

Для повышения эффективности работы городских железнодорожных станций на основе факторов, влияющих на пассажиропоток, необходимо провести дополнительный анализ факторов для различных типов станций. Это позволит разработать дифференцированные операционные стратегии для разных типов станций. Многие ученые начали использовать алгоритмы кластеризации для классификации железнодорожных станций. Обычно используется алгоритм кластеризации K-means (K-средних) [Митин, Панов, 2023] но чтобы избежать необходимости заранее определять количество кластеров, было предложено применять кластеризацию с максимизацией ожиданий (ЕМ). Для классификации станций часто использовалось евклидово расстояние (ED), которое не позволяет сравнивать сходство формы данных. Однако для кластеризации временных рядов выбор подходящей функции меры сходства может эффективно улучшить результаты кластеризации. Поэтому функции измерения сходства, такие как расстояние по форме (SBD) [Бычкова, 2020], динамическое искривление времени (DTW) и пространственно-временное сходство [Савруков, Савруков, 2020], постепенно стали применяться для анализа кластеризации временных рядов.

В качестве примера для исследования было предложено рассмотреть станции МЦД-3. Исходные данные включают следующее:

- 1. Данные о пассажиропотоке были получены на станциях МЦД-3 с 9 сентября 2024 года (понедельник) по 22 сентября 2024 года (воскресенье). Данные охватывают 39 станций МЦД-3. Выборка данных учитывала суточные пассажиропотоки во временные периоды до 9 утра и в вечерний и ночной периоды до 1 ночи.
- 2. Данные о POI в пределах 900-метрового РСА были получены с платформы Amap Development Platform. Данные POI были разделены на 69 типов, в блоках госучреждения, культура и досуг, образование, туризм, услуги общественного питания и размещения. В общей сложности 222 147 точек интереса. Часть выборки данных представлена в таблице 1.

Таблица 1 – Фрагмент выборки с данными
о месторасположении основных объектов в пределах
800-метрового расстояния от железнодорожных станций МЦД-3

POI NAME	WGS84		Тип
	Долгота	Ширина	
Школа №1528	55.97687355622217	37.19086495714495	Образование
ТЦ Зеленоградский	55.97928983389853	37.17206804574363	Досуг и услуги
Автостанция Крюково А-10	55.979533802302086	37.17303782691413	Общественный транспорт
...			

Обработка больших данных в этом исследовании в основном состояла из двух частей: реструктуризация данных о пассажиропотоке на входе и выходе во временные ряды, их стандартизация и реклассификация данных о POI.

В соответствии с требованиями метода кластеризации временных рядов для набора данных, данные о пассажиропотоке на входе и выходе с железнодорожных станций МЖД-3 были взяты на основе «вход в будний день - выход в будний день - вход в выходной день - выход в выходной день». Эта реструктуризация была проведена для создания одномерного временного ряда для каждой станции как показано в уравнении (1):

$$x_j = \{x_j^1 \dots x_j^n\} \quad (1)$$

где X_j представляет собой временной ряд, соответствующий железнодорожной станции j , а x_j^n - данные о пассажиропотоке в n -й временной точке для железнодорожной станции j , где $n = 39$.

Из-за низкого пассажиропотока на некоторых железнодорожных станциях на начальных этапах строительства, а также для исключения влияния случайных изменений пассажиропотока, из набора данных были исключены городские железнодорожные станции с ежедневным пассажиропотоком менее 500 человек в день. Чтобы уменьшить влияние случайных факторов, были рассчитаны средние значения пассажиропотока на входе и выходе для городских железнодорожных станций в будние дни (с понедельника по пятницу) и в выходные (с субботы по воскресенье). Это позволило получить средний пассажиропоток для будних и выходных дней на каждой станции. В итоге было получено 132 достоверных временных ряда. В связи со значительными различиями в пассажиропотоке на разных городских железнодорожных станциях и для уменьшения влияния величины пассажиропотока на формы данных в данном исследовании использовалась процентная нормализация для стандартизации данных. Метод расчета процентной нормализации был следующим [Бычкова, 2020] (2):

$$p_j^i = \frac{x_j^i}{\sum_{i=1}^m x_j^i} \quad (2)$$

где p_{ij} - доля входного (выходного) пассажиропотока в i -й час к общему входному (выходному) пассажиропотоку в j -й день, x_{ij} - входной (выходной) пассажиропоток в i -й час, n - общее количество рабочих часов в день для станции, а $m = 19$.

Сравнивая рисунок 1, можно заметить, что, устраняя влияние величины данных, процентная нормализация имеет тот же эффект, что и нормализация Z-оценки. Нормализация Z-оценки не придает никакого практического значения числовым значениям, соответствующим каждой временной точке, в то время как процентная нормализация отражает вклад каждого часа в общий ежедневный пассажиропоток на входе (выходе). Использование процентной нормализации сохраняет содержательный смысл данных, тем самым повышая интерпретируемость результатов кластеризации.

Чтобы всесторонне представить особенности территориального месторасположения, была проведена реклассификация данных POI на основе специфических атрибутов территории. Как показано в таблице 2, данные POI, полученные с платформы разработки Amap, были систематически классифицированы по 6 отдельным классам, а именно: коммерческие услуги, общественные услуги, туристические достопримечательности, жилые районы, офисные помещения и транспортные услуги.

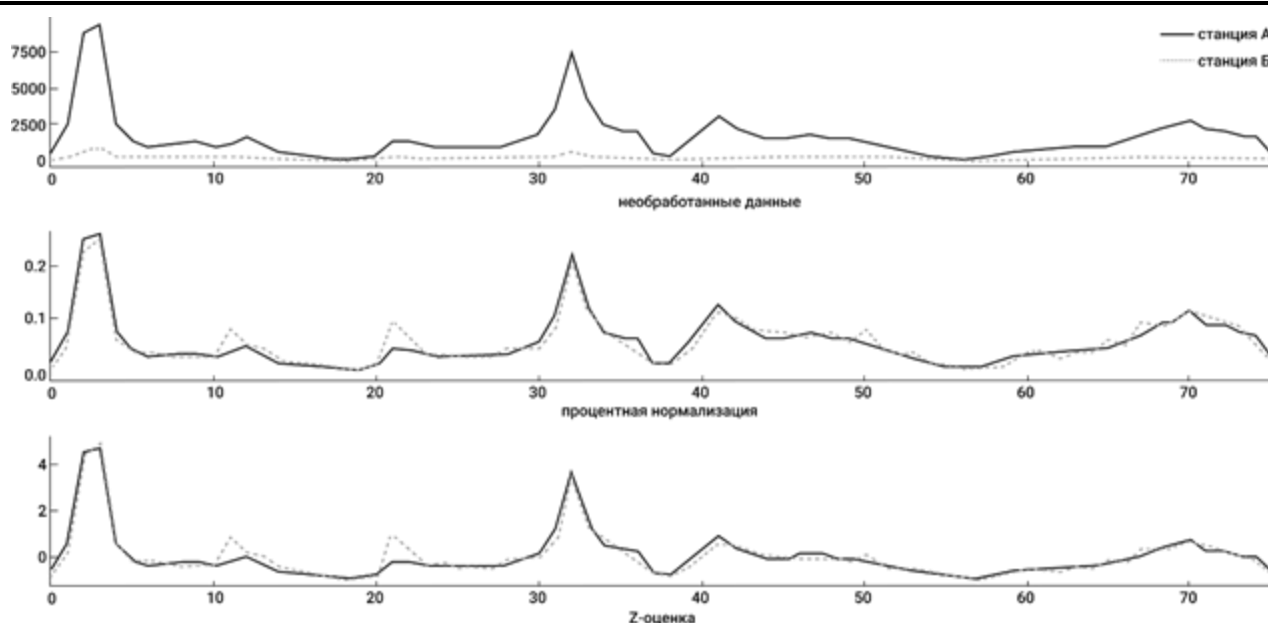


Рисунок 1 - Сравнение процентной нормализации и Z-оценки.

Таблица 2- Реклассификация данных POI

Новые названия категорий	Категории
Коммерческие услуги	Прокат автомобилей
	Продажа автомобилей
	Техническое обслуживание автомобилей
	Услуги для мотоциклистов
	Услуги кейтеринга
	Розничные услуги
	Услуги по организации досуга
	Отдых
	Услуги по проживанию
Государственные услуги	Здравоохранение
	Государственные учреждения и общественные организации
	Научные, образовательные и культурные услуги
	Общественные учреждения
Туристические достопримечательности	Туристические достопримечательности
Жилые районы	Коммерческие жилые районы
Офисные помещения	Финансовые учреждения
	Компании
Транспортные услуги	Транспортные объекты
	Придорожная инфраструктура

Алгоритм кластеризации К-средних — это алгоритм кластеризации на основе разбиения, который опирается на набор образцов [Кокорев, Корнеева, Сидоренко, Шаш, 2021]. Основные шаги:

Шаг 1. Случайным образом выбираются k образцов в качестве начальных центров кластеров.

Шаг 2. Каждый образец назначается в кластер, центр которого ближе всего по расстоянию.

Шаг 3. Производится вычисление центра каждого кластера по среднему значению его образцов.

Шаг 4. Повторяет эти процедуры до тех пор, пока центры кластеров не перестанут меняться или не будет достигнуто заданное количество итераций.

В кластеризации K-средних функция измерения сходства используется для вычисления расстояния между каждой точкой данных и центром кластера, что облегчает отнесение точек данных к ближайшему центру кластера. В данном исследовании в качестве функций измерения сходства в модели были использованы ED, DTW и SBD.

При использовании ED для оценки сходства двух временных рядов необходимо установить соответствие один-к-одному между временными узлами двух последовательностей и вычислить расстояние в каждом соответствующем временном узле, как показано на рисунке 2.

Из-за присущего ED ограничения в отражении сходства форм временных рядов, кластеризация на основе этой метрики может привести к объединению двух последовательностей с существенно различными формами кривых, но с относительно небольшим ED.

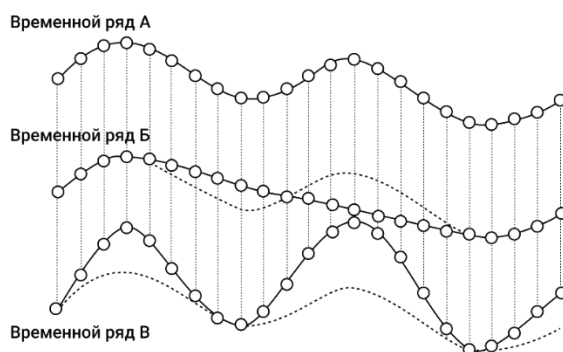


Рисунок 2 - ED временных рядов

DTW обычно используется для измерения сходства между двумя временными рядами разной длины. Как показано на рисунке 3, при использовании DTW для измерения сходства между временным рядом $A_n = (a_1, a_2, \dots, a_n)$ длины n и временным рядом $B_m = (b_1, b_2, \dots, b_m)$ длины m , необходимо найти непрерывное соответствие, которое включает все точки обоих временных рядов. Первоначально строится матрица размера $n \times m$, где элемент в i -й строке и j -м столбце представляет собой расстояние w_{ij} (обычно ED) между точкой a_i во временном ряду A_n и точкой b_j во временном ряду B_m . Затем ставится задача найти монотонно возрастающий, непрерывный диагональный путь в матрице с минимальной суммой расстояний (D_{dtw}), который представляет собой оптимальное выравнивание. Формула расчета D_{dtw} выглядит следующим образом:

$$D_{dtw} = \frac{\sum_{k=1}^p w_{ij}^k}{p}$$

где D_{dtw} - DTW, w_{ij}^k - значение элемента матрицы, соответствующее k -й точке на пути, а p - количество точек на пути.

SBD - это усовершенствование, предложенное на основе ED, с учетом того, что масштабирование и трансляция амплитуды не изменяют тенденцию формы волны временных рядов. Как показано на рисунке 5, при измерении сходства между временным рядом $A_m = (a_1, a_2, \dots, a_m)$ и временным рядом $B_m = (b_1, b_2, \dots, b_m)$ с помощью SBD, B_m является прогрессивно

сдвигается по отношению к A_m в каждой временной точке. Внутреннее произведение между A_m и B_m вычисляется на каждом шаге, в результате чего формируется последовательность значений внутреннего произведения, обозначаемая как $C_\omega = (c_1, \dots, c_\omega)$, $\omega \in (1, 2m - 1)$, длиной $2m - 1$. Наконец, максимальное значение в последовательности C_ω выбирается для вычисления SBD, как выражено в формуле (3):

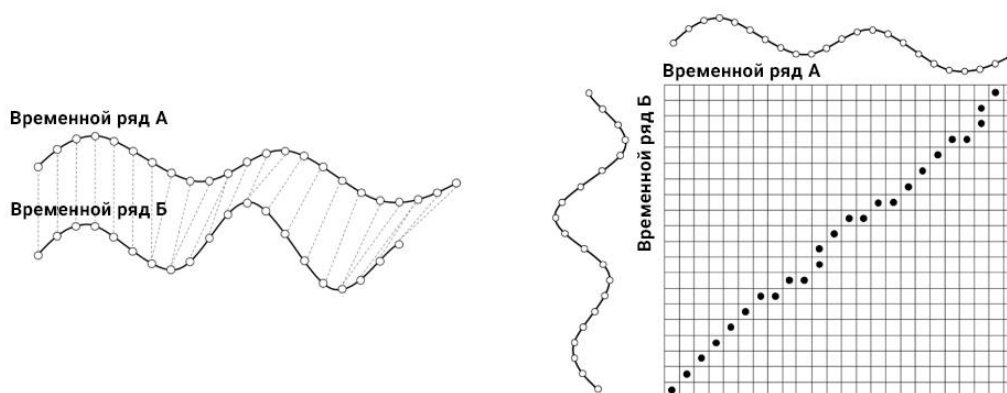


Рисунок 3 - DTW двух временных рядов

$$B_{(s)} = \begin{matrix} |s| \\ \overbrace{(0, \dots, 0, b_1, b_2, \dots, b_{m-s})}^{s \geq 0} \\ |s| \\ \underbrace{(b_{m+s}, \dots, b_{m-1}, b_m, 0, \dots, 0)}_{s \leq 0} \end{matrix} \quad (3)$$

В уравнении $B(s)$ представляет временной ряд при скольжении с размером шага s : s обозначает количество шагов скольжения от начальной точки выравнивания, где положительное значение означает скольжение вправо, отрицательное - влево, и $s \in [1 - m, m - 1]$.

$$SBD = 1 - \frac{\max(C_\omega)}{\sqrt{\sum_{i=1}^m a_i^2 \cdot \sum_{i=1}^m b_i^2}} \quad (4)$$

В этом уравнении SBD — это расстояние, основанное на форме, C_ω - последовательность значений внутреннего произведения, а a_i и b_i представляют собой i -е элементы временного ряда B_m .

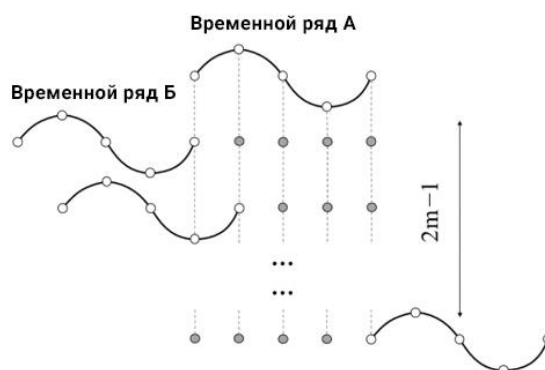


Рисунок 4 - SBD двух временных рядов

В данном исследовании для оценки эффективности кластеризации различных моделей кластеризации временных рядов использовались индекс СН, индекс DB и индекс S_Dbw. Индекс СН измеряет эффективность кластеризации путем вычисления отношения ковариации между кластерами к ковариации внутри кластера. Более высокий индекс СН указывает на лучшую эффективность кластеризации, где меньшая внутрикластерная ковариация и большая межкластерная ковариация способствуют более высокому индексу СН. Индекс DB измеряет эффективность кластеризации путем вычисления отношения среднего расстояния между двумя точками в кластере к расстоянию между центрами кластера. Меньший индекс DB указывает на лучшую эффективность кластеризации, при этом меньшие средние расстояния между точками внутри кластера и большие расстояния между центрами кластеров способствуют меньшему индексу DB. Индекс S_Dbw измеряет эффективность кластеризации путем вычисления суммы компактности внутри кластеров и плотности между кластерами. Меньший индекс S_Dbw указывает на лучшую эффективность кластеризации, и результаты кластеризации не зависят от используемого алгоритма. Уравнения для расчета индекса S_Dbw выглядят следующим образом:

$$S_{S_Ddw} = \text{Scat} + \text{Dens_bw} \quad (5)$$

$$\text{Scat} = \frac{1}{k} \sum_{i=1}^k \frac{\|\sigma(A_i)\|_2}{\|\sigma(E)\|_2} \quad (6)$$

$$\text{Dens_bw} = \frac{1}{K(K-1)} \sum_{i=1}^k \sum_{j=1, i \neq j}^k \frac{\text{density}(c_{ij})}{\max(\text{density}(c_i), \text{density}(c_j))} \quad (7)$$

$$\begin{cases} \text{stdev} = \frac{1}{k} \sqrt{\sum_{i=1}^k \|\sigma(A_i)\|} \\ d(x, c_i) = \|x - c_i\|_2 \\ f(x, c_i) = \begin{cases} 0, & d(x, c_i) > \text{stdev} \\ 1, & d(x, c_i) \leq \text{stdev} \end{cases} \\ \text{density}(c_i) = \sum_{x \in A_i} f(x, c_i) \\ \text{density}(c_{ij}) = \sum_{x \in (A_i \cup A_j)} f(x, c_{ij}) \end{cases} \quad (8)$$

В этих уравнениях SS_Dbw - индекс S_Dbw, Scat - компактность внутри кластеров, Dens_bw - плотность между кластерами, k - количество кластеров, A_i - набор данных кластера i, E - набор данных образцов, $c_{i,j}$ - центроиды кластеров i и j, c_{ij} - конечная точка центроидов кластеров i и j.

Для удовлетворения требований по обучению на наборе данных, железнодорожные станции подверглись маркировке с помощью алгоритмов кластеризации. Что касается характеристик сети, были выбраны три ключевые метрики: центральность степени, центральность между и центральность близости. Центральность степени определяет важность узлов непосредственно в сети, а центральность между узлами отражает частоту использования станции через количество кратчайших путей. Центральность близости показывает близость узлов к другим узлам.

Алгоритм кластеризации K-средних требует предварительного определения количества кластеров. Для определения количества кластеров был использован метод силуэтного коэффициента. Силуэтный коэффициент был относительно идеальным, когда количество кластеров составляло 4.

При анализе данных железнодорожных станций МЖД-3 были вычислены индексы оценки результатов кластеризации с использованием различных функций измерения сходства. Как

показано в таблице 3, результаты кластеризации на основе DTW превзошли другие функции измерения сходства, продемонстрировав лучшие значения индекса CH, индекса DB и индекса S_Dbw по сравнению с ED и SBD. Таким образом, предпочтение было отдано модели кластеризации К-средних с использованием DTW в качестве функции измерения сходства для кластеризации станций в данном случае.

Таблица 3 – Результаты расчета К-средних

Индексы	К-средние алгоритмы		
	DTW	ED	SBD
CH индекс	64.779	63.981	62.899
DB индекс	1.446	1.461	1.476
S_Dbw индекс	0.664	0.681	0.670

Результаты кластеризации, представленные на рисунке 7, свидетельствуют о наличии четких временных закономерностей в пассажиропотоках четырех кластеров. Примечательно, что в будние дни кривые пассажиропотоков подвержены большим колебаниям по сравнению с выходными. Кроме того, наблюдались различия во временном распределении входящих и исходящих пассажиропотоков (рисунок 8).

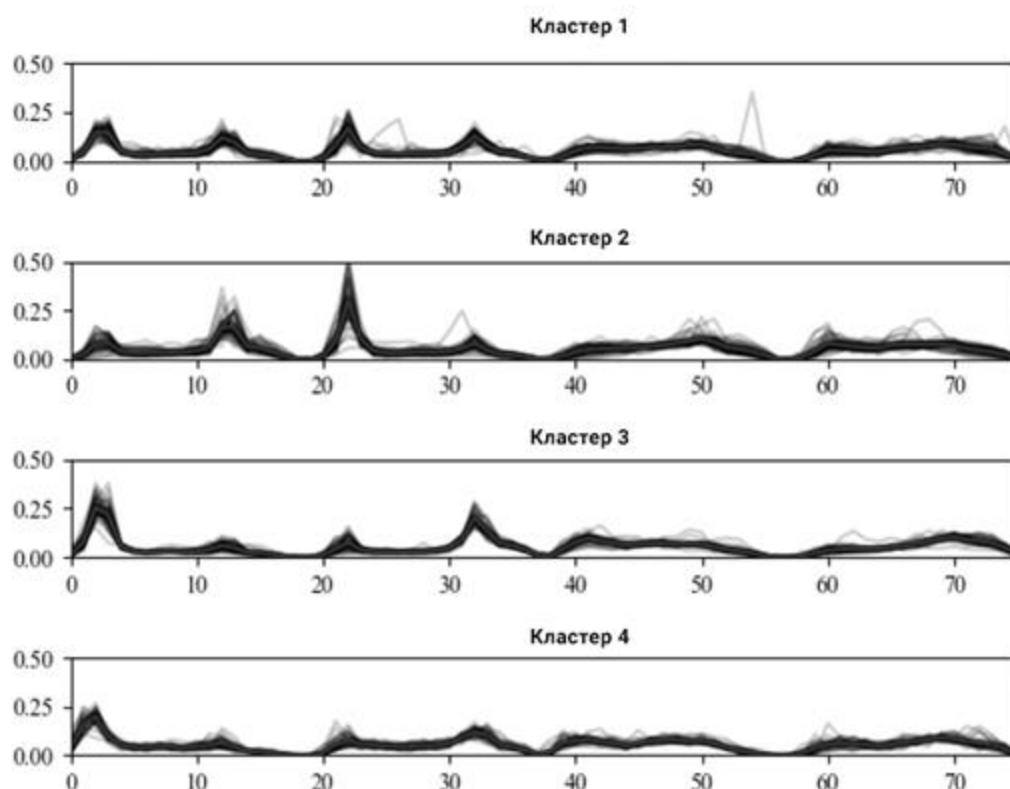


Рисунок 7 – Результаты кластеризации временных рядов

Кластер 1: В будние дни кривые распределения входящих и исходящих пассажиропотоков демонстрировали «сбалансированную» двухпиковую картину. Утренний пик для входящего пассажиропотока приходился на период с 7:00 до 8:00, а вечерний - на период с 17:00 до 18:00. Для исходящего пассажиропотока утренний пик был с 7:00 до 9:00, а вечерний - с 18:00 до 19:00.

В выходные дни на кривых распределения входящих и выходящих пассажиропотоков не было выраженных пиковых значений.

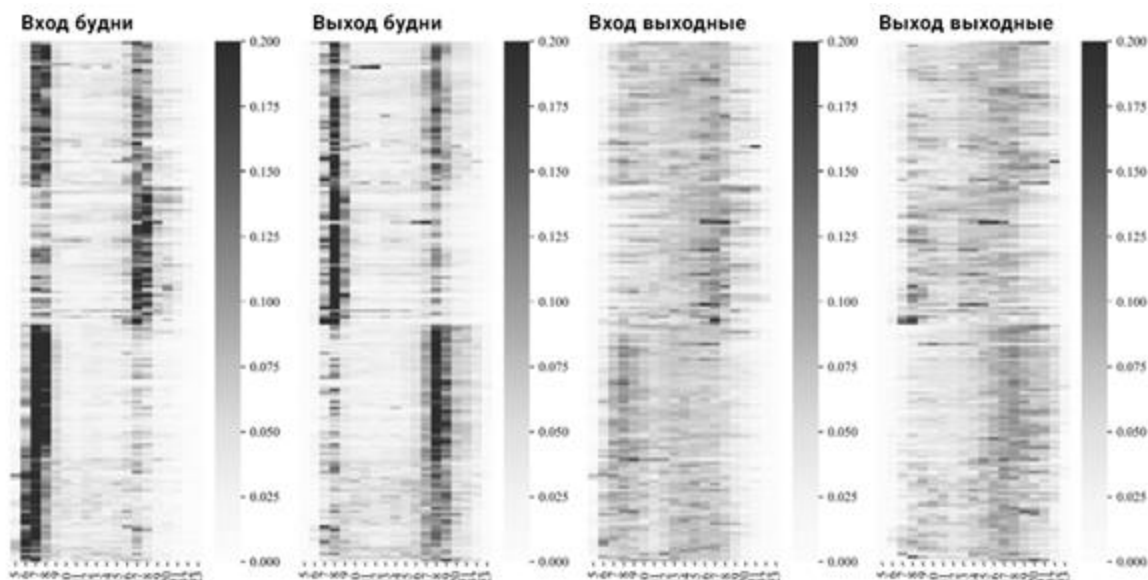


Рисунок 8 – Тепловая карта результатов кластеризации

Кластер 2: В будние дни кривые распределения входящих и выходящих пассажиропотоков демонстрировали «размерный» двухпиковый паттерн. Утренний и вечерний пики для въезжающего и выезжающего пассажиропотоков были такими же, как и в кластере 1, но по сравнению с кластером 1 утренний пик для выезжающего пассажиропотока и вечерний пик для въезжающего пассажиропотока имели более высокие пиковые значения. Вклад утреннего пика в пассажиропоток выходящих пассажиров составлял около 25 %. В выходные дни на кривых распределения входящих и выходящих пассажиропотоков не было выраженных пиковых значений, а по сравнению с кластером 1 наблюдались большие колебания пассажиропотоков.

Кластер 3: В будние дни кривые распределения входящих и выходящих пассажиропотоков имели однопиковый характер. Пик входящего пассажиропотока приходился на период с 7:00 до 8:00, а пик выходящего пассажиропотока - на период с 18:00 до 19:00. Доля пассажиропотока в часы пик составляла около 25 %. В выходные дни кривые распределения входящих и выходящих пассажиропотоков демонстрировали асимметричную пиковую картину, как и в будние дни. Входящий пассажиропоток был сосредоточен в утреннем пике, а выходящий - в вечернем. Доля пассажиропотока в часы пик составляла около 10 %.

Кластер 4: В будние дни кривые распределения входящих и выходящих пассажиропотоков были несколько похожи на кластер 3, но пик для входящего пассажиропотока приходился на период с 6:00 до 7:00, на 1 час раньше, чем в кластере 3, а пиковое значение было ниже, чем в кластере 3. В выходные дни распределение входящих и выходящих пассажиропотоков имело ярко выраженную двухпиковую тенденцию, причем пассажиропотоки в основном концентрировались в утренний и вечерний пиковые периоды. Вклад пассажиропотока в часы пик составлял около 10 %.

Согласно результатам, представленным на рисунке 9, пространственное распределение четырех типов городских железнодорожных станций можно представить следующим образом: станции кластера 2 в основном сосредоточены в центральной части города, т.е. в ядре города.

Станции из кластера 1, напротив, были в основном распределены в периферийных районах кластера 2. Даже когда станции из кластера 2 и кластера 1 оказывались в периферийных районах железнодорожной сети, они все равно находились в относительно центральных частях. Станции кластера 3 находились в основном в пригородных районах города, а станции кластера 4 - в самых отдаленных районах железнодорожной сети, то есть на максимально дальних расстояниях от центра города. В целом, все четыре типа станций демонстрируют характерное пространственное распределение, постепенно переходя от ядра города к его окраинам.

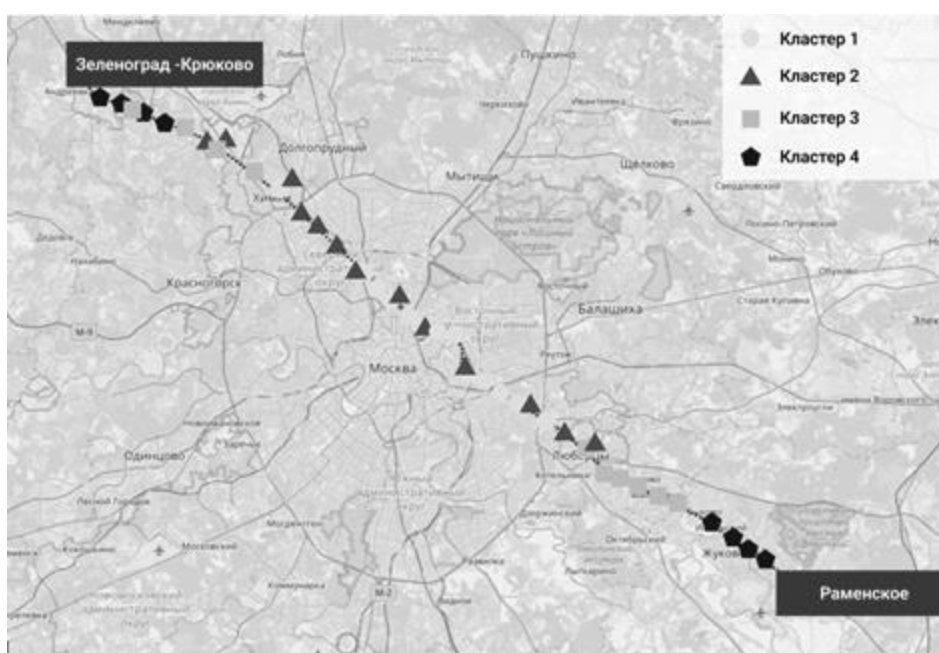


Рисунок 9 - Пространственное распределение железнодорожных станций МЦД-3

Анализ результатов кластеризации показал, что на различных типах железнодорожных станций наблюдаются различия в характеристиках распределения пассажиропотока. Кроме того, на железнодорожных станциях одного типа наблюдаются различия в распределении пассажиропотока в будние и выходные дни. Из-за разнообразия сети и территориальных особенностей на железнодорожных станциях пассажиропоток имеет свои особенности распределения. По сути, эти различия обусловлены различиями в пространственной планировке городов и интенсивности застройки, что приводит к явному разделению пассажиропотоков между работой и домом.

Заключение

Вклад данного исследования заключается в усилении эффекта кластеризации городских железнодорожных станций путем выбора соответствующих функций измерения сходства и анализа факторов, влияющих на распределение пассажиропотока на станциях с точки зрения особенностей сети. Используя большие данные, в данном исследовании были применены алгоритмы кластеризации временных рядов и модель XGBoost, чтобы обеспечить всестороннее понимание того, как эти факторы влияют на распределение пассажиропотока. Данное исследование предлагает целевые стратегии эксплуатации и управления, основанные на отличительных характеристиках распределения пассажиропотока на станциях, что

способствует оптимизации городских железнодорожных систем. Повышая эффективность и сокращая потери ресурсов, эти стратегии способствуют устойчивому развитию городских транспортных сетей.

В данном исследовании в основном использовались данные о точках интересов пассажиров, но при этом не учитывалась площадь территории. В то же время в качестве факторов влияния на характеристики сети были выбраны только характеристики, связанные с железнодорожной сетью, а показатели, связанные с дорожной сетью, такие как количество перекрестков и длина автодорог, не рассматривались, что может быть учтено в последующих исследованиях.

Библиография

1. Пазойский, Ю. О. Закономерность распределения пассажиропотока в пассажирских поездах дальнего следования / Ю. О. Пазойский, М. Ю. Савельев, Е. А. Середов // Экономика железных дорог. – 2022. – № 7. – С. 29-39. – EDN USFVBI.
2. Кулешова, А. В. Моделирование транспортной системы города с учетом распределения пассажиропотоков / А. В. Кулешова, В. А. Широченко // Материалы, оборудование и ресурсосберегающие технологии : материалы международной научно-технической конференции, Могилев, 18–19 апреля 2013 года / Главный редактор: И. С. Сазонов, Редакционная коллегия: В. М. Пашкевич (зам. гл. редактора), В. И. Кошелева (отв. секретарь), доц. М. Е. Лустенков, В. П. Куликов, М. Ф. Пашкевич, В. П. Тарасик, С. Д. Семенюк, И. И. Маковецкий, Рецензенты: И. В. Лесковец, С. А. Рынкевич, С. Д. Семенюк, С. С. Сергеев, И. И. Маковецкий. Том Часть 2. – Могилев: Межгосударственное образовательное учреждение высшего образования "Белорусско-Российский университет", 2013. – С. 235-236. – EDN FNMROC.
3. Определение потребных размеров движения пригородных поездов с учетом суточной и внутрисуточной неравномерностей распределения пассажиропотока / А. Г. Филиппов, С. С. Смирнов, Д. В. Язев, И. П. Александров // Известия Петербургского университета путей сообщения. – 2021. – Т. 18, № 4. – С. 528-536. – DOI 10.20295/1815-588X-2021-4-528-536. – EDN CRFEG1.
4. Александров, В. Ю. Анализ эффективности работы наземного транспорта на основе анализа пространственных характеристик остановок и их пассажиропотоков / В. Ю. Александров, А. Я. Бутыркин // Академик Владимир Николаевич Образцов - основоположник транспортной науки : Труды международной научно-практической конференции, Москва, 17 ноября 2023 года. – Москва: ЗАО "Университетская книга", 2023. – С. 413-419. – EDN BGARXM.
5. Математические методы интеллектуального управления / Д. С. Кокорев, Е. В. Корнеева, В. Г. Сидоренко, А. М. Шаш. – Москва : Российская открытая академия транспорта федерального государственного бюджетного образовательного учреждения высшего образования "Российский университет транспорта" (МИИТ), 2021. – 66 с. – EDN EYNNHU.
6. Савруков, А. Н. Методика расчета индекса и оценка транспортной доступности субъектов РФ / А. Н. Савруков, Н. Т. Савруков // Финансы и кредит. – 2020. – Т. 26, № 11(803). – С. 2410-2426. – DOI 10.24891/fc.26.11.2410. – EDN IYTOFQ.
7. Митин, Г. В. Модификация алгоритма DBSCAN с использованием гибридных подходов к определению границ кластеров для обработки потоковых данных / Г. В. Митин, А. В. Панов // ИТ-Стандарт. – 2023. – № 4(37). – С. 36-57. – EDN ECFWNX.
8. Интеллектуальные технологии в эксплуатационной работе на железнодорожном транспорте / Р. А. Ефимов, И. Н. Шапкин, В. Н. Шмаль, П. А. Минаков. – Москва : ФГБУ ДПО "Учебно-методический центр по образованию на железнодорожном транспорте", 2024. – 272 с. – EDN NJYDYH.
9. Бычкова, А. А. Автокорреляционный анализ пассажиропотока железнодорожных вокзалов в городах Свердловской области / А. А. Бычкова // VIII Информационная школа молодого ученого : Сборник научных трудов, Екатеринбург, 21–24 сентября 2020 года / Ответственный редактор П. П. Трескова; составители А. И. Кирсанова, Ю. Д. Прокофьева, М. А. Пекшева, Т. А. Осипенко, О. Н. Зырянова, Л. А. Оболенская. – Екатеринбург: ООО "Издательство УМЦ УПИ", 2020. – С. 197-202. – DOI 10.32460/ishmu-2020-8-0020. – EDN WQJGGM.
10. Азаренкова, З. В. Транспортная составляющая социально-экономических связей / З. В. Азаренкова // Градостроительство. – 2016. – № 2(42). – С. 56-62. – EDN VZVLLF.
11. Гуриева, Л. К. Особенности использования кластерного подхода при организации государственно-частного партнерства в сфере туризма / Л. К. Гуриева, А. В. Кучумов, П. Ю. Еремичева // Вестник Северо-Осетинского государственного университета имени К. Л. Хетагурова. – 2024. – № 1. – С. 108-118. – DOI 10.29025/1994-7720-2024-1-108-118. – EDN CITOPY.

K-means Algorithms and the XGBoost Model in the Transport Industry

Ol'ga N. Madyar

PhD in Technical Sciences, Associate Professor,
Russian University of Transport,
127994, 9, Obraztsova str., Moscow, Russian Federation;
e-mail: zox2.777@mail.ru

Elena A. Popova

PhD in Technical Sciences, Associate Professor,
Voronezh Branch of Rostov State University of Transport,
394026, 82, Uritskogo str., Voronezh, Russian Federation;
e-mail: Elena-1504@mail.ru

Abstract

This study utilized big data analytics to examine the influence of territorial factors and transport network characteristics on passenger flow distribution. To support decision-making for differentiated operational management strategies for various types of railway stations, aiming to optimize performance and ensure sustainable development of urban railway systems, the article compares clustering results using different similarity measurement functions within the K-means algorithm and selects the optimal similarity measurement function for building clustering models. Additionally, the article explores factors influencing passenger flow distribution, based on which a set of features was formed. These features, combined with the clustering results, served as input for the XGBoost model to analyze the relationship between various factors and passenger flow distribution at stations.

For citation

Madyar O.N., Popova E.A. (2024) Algoritmy K-means i model' XGBoost v transportnoi otrasli [K-means Algorithms and the XGBoost Model in the Transport Industry]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 14 (11A), pp. 226-239. DOI: 10.34670/AR.2024.95.50.025

Keywords

Passenger flow distribution, K-means, clustering, MCD-3, distribution factors, modeling, time series, XGBoost.

References

1. Pazoyskiy, Yu. O., Saveliev, M. Yu., Seredov, E. A. (2022). O zakonomernosti raspredeleniya passazhiropotoka v passazhirs-kikh poezdakh dal'nego sleduvaniya [On the regularity of passenger flow distribution in long-distance passenger trains]. *Ekonomika zheleznikh dorog* [Economics of Railways], 7, pp. 29-39. EDN USFVBI.
2. Kuleshova, A. V., Shirochenko, V. A. (2013). Modelirovanie transportnoi sistemy goroda s uchetom raspredeleniya passazhiropotokov [Modeling the transport system of the city taking into account the distribution of passenger flows]. In I. S. Sazonov (Ed.), *Materialy, oborudovanie i resursosberegayushchie tekhnologii: materialy mezhdunarodnoi nauchno-tekhnicheskoi konferentsii, Mogilev, 18–19 aprelya 2013 goda* (Vol. 2, pp. 235-236). Mogilev:

- Mezhhgosudarstvennoe obrazovatel'noe uchrezhdenie vysshego obrazovaniya "Belorussko-Rossiyskii universitet". EDN FNMROC.
3. Filippov, A. G., Smirnov, S. S., Yasev, D. V., Aleksandrov, I. P. (2021). Opredelenie potrebnnykh razmerov dvizheniya prigorodnykh poezdov s uchetom sutochnoi i vnutrisutochnoy neravnomernosti raspredeleniya passazhiropotoka [Determination of required sizes for suburban train movements considering daily and intraday unevenness in passenger flow distribution]. *Izvestiya Peterburgskogo universiteta putey soobshcheniya* [News of the St. Petersburg University of Transport], 18(4), pp. 528-536. <https://doi.org/10.20295/1815-588X-2021-4-528-536> EDN CRFEGI.
 4. Aleksandrov, V. Yu., Butyrkin, A. Ya. (2023). Analiz effektivnosti raboty nazemnogo transporta na osnove analiza prostranstvennykh kharakteristik ostanovok i ikh passazhiropotokov [Analysis of the efficiency of ground transport based on the analysis of spatial characteristics of stops and their passenger flows]. In *Akademik Vladimir Nikolaevich Obraztsov - osnovopolozhnik transportnoi nauki: Trudy mezhdunarodnoi nauchno-prakticheskoi konferentsii*, Moskva, 17 noyabrya 2023 goda (pp. 413-419). Moscow: ZAO "Universitetskaya kniga". EDN BGARXM.
 5. Kokorev, D. S., Komeeva, E. V., Sidor-enko, V. G., Shash, A. M. (2021). Matematicheskie metody intellektual'nogo upravleniya [Mathematical methods of intellectual management]. Moscow: Rossiyskaya otkrytaya akademiya transporta federal'nogo gosudarstvennogo byudzheta-nogo obrazovatel'nogo uchrezhdeniya vysshego obrazovaniya "Rossiyskii universitet transporta" (MIIT). EDN EYHHHU.
 6. Savrukov, A. N., Savrukov, N. T. (2020). Metodika rascheta indeksa i otsenka transportnoi dostupnosti sub'ektov RF [Methodology for calculating the index and assessing the transport accessibility of subjects of the Russian Federation]. *Finansy i kredit* [Finance and Credit], 26(11), pp. 2410-2426. <https://doi.org/10.24891/fc.26.11.2410> EDN IYTOFQ.
 7. Mitin, G. V., Panov, A. V. (2023). Modifikatsiya algoritma DBSCAN s ispol'zovaniem gibridnykh podkhodov k opredeleniyu granits klasterov dlya obrabotki potokovykh dannykh [Modification of the DBSCAN algorithm using hybrid approaches to defining cluster boundaries for processing streaming data]. *IT-Standart* [IT Standard], 4(37), pp. 36-57. EDN ECFWNX.
 8. Efimov, R. A., Shapkin, I. N., Shmal', V. N., Minakov, P. A. (2024). Intellektual'nye tekhnologii v eksploatatsionnoi rabote na zheleznodorozhnom transporte [Intelligent technologies in operational work on rail transport]. Moscow: FGBU DPO "Uchebno-metodicheskii tsentr po obrazovaniyu na zheleznodorozhnom transporte". EDN NJYDYH.
 9. Bychkova, A. A. (2020). Avtokorrelatsionnyi analiz passazhiropotoka zheleznodorozhnykh vokzalov v gorodakh Sverdlovskoi oblasti [Autocorrelation analysis of passenger flow at railway stations in Sverdlovsk region]. In *VIII Informatsionnaya shkola molodogo uchyonogo: Sbornik nauchnykh trudov*, Ekaterinburg, 21–24 sentyabrya 2020 goda (pp. 197-202). Ekaterinburg: OOO "Izdatel'stvo UMT UPI". <https://doi.org/10.32460/ishmu-2020-8-0020> EDN WQGJGM.
 10. Gurieva, L. K. Osobennosti ispol'zovaniya klasternogo podkhoda pri organizatsii gosudarstvenno-chastnogo partnerstva v sfere turizma / L. K. Gurieva, A. V. Kuchumov, P. Yu. Eremicheva // *Vestnik Severo-Osetinskogo gosudarstvennogo universiteta imeni K. L. Khetagurova*. – 2024. – № 1. – S. 108-118. – DOI 10.29025/1994-7720-2024-1-108-118. – EDN CITOPY.
 11. Azarenkova, Z. V. (2016). Transportnaya sostavlyayushchaya sotsial'no-ekonomicheskikh svyazey [The transport component of socio-economic ties]. *Gradostroitel'stvo* [Urban Planning], 2(42), pp. 56-62. EDN VZVLLF.