Структурно-временной анализ пассажиропотока метрополитена

Д.Т. Оспанов, Д.Е. Намиот, О.Н. Покусаев

***Аннотация*— Статья посвящена одному подходу к анализу транспортных потоков. Исходными данными для анализа являются так называемые матрицы корреспонденции, которые описывают количество поездок в единицу времени между двумя точками. Конкретный датасет, который анализировался в работе представляет собой матрицу корреспонденций московского метро (поездки пассажиров между станциями метро) за февраль 2018 года. Целью анализа является структурно-временной анализ пассажиропотока (как и когда перемещаются пассажиры). В работе предложена методика анализа транспортного трафика, основанная на связке сингулярного разложения и методов кластеризации машинного обучения. Сингулярное разложение используется здесь для понижения размерности. Концепция использовать в связке упомянутые инструменты не нова, применялась в других областях, но в данной работе она была успешно адаптирована именно для транспортной сферы. В статье представлен библиотечный программный модуль для реализации каждого этапа разработки предложенной модели. Модуль способен обрабатывать большие объемы данных, имеет потенциал для легкого масштабирования и расширения. В работе представлен пример реализации предложенной методики применительно к историческим данным о пассажиропотоке московского метрополитена.**

***Ключевые слова*—пассажиропоток, метро, сингулярное разложение, методы кластеризации, машинное обучение, матрица корреспонденции.**

I. ВВЕДЕНИЕ

В настоящее время во время ежедневных поездок генерируются сразу несколько цифровых следов, многочисленные наборы данных, представляющие объекты, перемещающиеся между географическими точками. Эти источники данных могут быть использованы для реализации новых подходов к моделированию для анализа городской мобильности. Могут быть решены различные вопросы, а именно определение целей поездки, выявление моделей мобильности или достижение лучшего понимания и прогнозирования пассажиропотоков.

Чтобы удовлетворить спрос на перевозку, этот спрос должен быть известен. Потребность в перевозках может быть представлена с помощью матрицы отправления-назначения (origin-destination - OD), называемой также матрицей корреспонденции. Матрица отправления-назначения – это матрица, в которой каждая ячейка представляет номер поездки от отправителя (строки) до места назначения (столбца). Потоки OD являются фундаментальными предпосылками для анализа транспорта и могут обеспечить модели поездок между географическими зонами, которые могут отражать трафик и экономическую деятельность. Надежное прогнозирование потоков OD может улучшить планирование и операции в управлении трафиком в режиме реального времени. Кроме того, с развитием подключенных и автономных транспортных средств динамическая информация матриц корреспонденции может облегчить процесс назначения транспортного средства и выбора маршрута, что может повысить эффективность интеллектуальных транспортных систем. Модель OD является весьма полезным средством для исследований в области транспорта, поскольку в ней обобщается динамика городов и мобильность людей.

II. ВОЗМОЖНОСТИ АНАЛИЗА ПАССАЖИРОПОТОКА

Методы анализа данных можно разделить на две большие группы: статистические модели и машинные методы обучения. Эти два подхода различаются по философии, целям, процессу разработки моделей и приобретению знаний. Статистические методы касаются умозаключений и оценки, направлены на создание модели, которая предлагает понимание данных, априори принимает функциональную форму, выдвигает ряд гипотез и накладывает ограничения, а также являются жесткими дисциплинами. Методы машинного обучения в свою очередь касаются реализации, направлены на обеспечение эффективного прогнозирования, приближения функциональной формы посредством обучения, возможны без предварительных моделей или спецификаций распределения ошибок и просты в применении с использованием программных пакетов. Из 130 извлеченных статей в 48 исследованиях использовались статистические методы, в 59 исследованиях использовались методы машинного обучения, а в 23 исследованиях использовались как статистические, так и машинные методы обучения.

С экономической точки зрения процесс дорожного движения можно рассматривать поездку как спрос на поездки, а производительность сети представляет собой предложение доступности поездок. Эти два элемента имеют дезагрегированную и агрегированную форму, которые взаимодействуют в соответствии с

поведением человеческого выбора и физикой динамики движения. Спрос, как и предложение, является периодическими и стохастическими, краткосрочное прогнозирование связано с отслеживанием динамики этого взаимодействия [1]. Технология прогнозирования может быть разделена на две парадигмы: эмпирическую, включающую довольно стандартную статистическую методологию, и основанную на теории транспортных процессов, понимании их структуры, спроса и предложения [2].

Эмпирические методы включают регрессионную нелинейную регрессию и модели Box-Jenkins типа ARIMA (авторегрессионная скользящая средняя) с сезонными моделями функций корректировки и переноса и без них. Используются различные обобщения с использованием нелинейных компонентов и адаптивной оценки изменяющихся во времени параметров. Другие статистические методы дискриминационного и кластерного анализа были использованы в методах распознавания образов, основанных на таксономии исторических баз данных [1]. В рамках теоретических подходов усилия были затрачены на физическое моделирование переменных прохода транспортных средств со стороны предложения; и поведенческое моделирование потоков поездок и ОД для стороны спроса. Физические модели основаны на динамике движения и используют технологию моделей пространства состояний и фильтрации Калмана для связи оценок скрытых состояний с наблюдениями на дорогах. Поведенческие модели представляют потоки OD как скрытые состояния и требуют сложных алгоритмов назначения трафика для динамического назначения потоков OD путям и каналам в сети. Эти методы требуют оценок и прогнозов фракций поворота на узлах принятия решений в сети автомагистралей для моделирования выбора маршрута.

Набирает интерес применение нейронных сетей. Нейронная сеть является естественным методологическим кандидатом для прогнозирования с несколькими входами и несколькими выходами. Кроме того, поскольку известно, что процесс трафика является нелинейной системой, нейронные сети привносят в задачу когерентную структуру нелинейных регрессионных моделей. Методы, основанные на нейронной сети, использовались в [3] и [4] с различными моделями оптимизации и смешивания с целью прогнозирования пассажиропотоков.

В работе [5] предлагается структура, которая объединяет нейронные сети графа и фильтр Калмана для прогнозирования потоков OD. Показано, что графовые нейронные сети эффективны при обработке данных с конкретной сетевой топологией, которая обозначается матрицей смежности. Фильтр Калмана — это классическая модель, включающая этапы прогнозирования и обновления для минимизации неопределенности прогнозирования [6]. Поскольку графовые нейронные сети и фильтр Калмана используют различные матрицы топологии и механизмы оптимизации, параметр смешивания используется для балансировки гетерогенного прогнозирования двумя методами. Кроме того, новые сверточные сети графов FL-GNN, включают свертку графа ссылок и свертку графа узлов, предназначены для прогнозирования потоков O-D. Предлагаемые сети обеспечивают общую структуру глубокого обучения для решения проблем, связанных с пространственно-временным отображением от ссылок до узлов.

Для краткосрочного прогнозирования спроса на высокоскоростные железные дороги использовались эмпирическая декомпозиция режимов и серые опорные вектора [7]. Позднее этот метод был улучшен с помощью кластеризации [8].

В работе [9] авторы не обнаружили существенной разницы между байесовской сетью (BN) и моделью авторегрессивной интегрированной скользящей средней ARIMA в прогнозировании транспортного потока. Однако предложенная линейная условная гауссовская модель BN превзошла модель ARIMA на уровнях 5, 10 и 15 минут. В исследовании [10] сравнили модели прогнозирования нейронных сетей исторического среднего значения (HA), векторной авторегрессии (VAR) и общей регрессии, прогнозируя транспортный поток шести выбранных каналов, извлеченных из шоссе 290 США в Хьюстоне, штат Техас. Авторы пришли к выводу, что модель VAR имеет самую высокую точность прогнозирования, когда нет или мало недостающих данных. Они также отметили, что модель HA, которая уже была применена к городским системам управления дорожным движением и другим информационным системам путешественников, поскольку она проста в реализации и понимании, имеет худшую точность прогнозирования.

В работе [11] сравнили пять методов машинного обучения: (1) нейронная сеть обратного распространения (BPNN), (2) нелинейная авторегрессионная модель с экзогенными входами нейронной сети (NARXNN), (3) машина опорных векторов с радиальной базисной функцией как функция ядра (SVM-RBF), (4) машина опорного вектора с линейной функцией (SVM-LIN) и (5) многолинейная регрессия с тремя статистическими методами: авторегрессионная интегрированная модель скользящей средней (ARIMA), VAR и пространственно-временной (ST). Используя скорость движения трех каналов, собранных с 4-й кольцевой дороги в Пекине, они пришли к выводу: BPNN, NARXNN и SVM-RBF явно превосходят две традиционные статистические модели ARIMA и VAR, а также то, что по мере увеличения временного шага

модель ST и модель VAR в некоторых случаях обеспечивают наименьшие ошибки.

В исследовании [12] сравнили три хорошо зарекомендовавшие себя модели прогнозирования транспортных потоков, а именно авторегрессионную скользящую среднюю ARMA, пространственно-временную ARMA и искусственную нейронную сеть (ANN) в разных условиях движения. Результаты были неоднозначными. Авторы не обнаружили какого-либо превосходства метода машинного обучения над статистическими методами на коротких промежутках. Однако на более длинных временных промежутках метод машинного обучения превзошел статистические методы.

Для прогнозирования транспортного потока в городской дорожной сети Шанхая [13], протестированы авторегрессия (AR), многомерные адаптивные регрессионные сплайны (MARS), регрессия опорного вектора (SVR), сезонная авторегрессионная интегрированная скользящая средняя (SARIMA), пространственно-временный байесовский MARS и SVR на основе выбора переменных. Результаты показали превосходство методов машинного обучения над статистическими методами.

Произведено сравнение также простых одномерных ARIMA и SARIMA с моделью графовой нейронной сети (GNN), прогнозируя поток трафика [14]. Авторы пришли к выводу, что ARIMA или SARIMA немного превосходят модель GNN в задачах прогнозирования на один или на два шага вперед. Тем не менее, производительность GNN больше, чем статистические методы, когда речь идет о более длинном горизонте прогнозирования. Затем было замечено, что GNN более эффективен в извлечении долгосрочных зависимостей и изучении динамики сети, в отличие от простых моделей временных рядов. Более ранние исследования также показали превосходство методов машинного обучения, включая модель нейронных сетей [15], модель Элмана [16] и адаптивную гибридную нечеткую систему на основе правил [17] над ARIMA.

Согласно данным исследованиям, можно остановиться на выводе, что методы машинного обучения, которые набирают популярность в последние годы, превосходят наивные статистические методы, такие как историческое среднее и экспоненциальное сглаживание. Однако нет определенного превосходства, когда методы машинного обучения сравниваются с передовыми статистическими методами, такими как пространственно-временная авторегрессионная интегрированная скользящая средняя и VAR. Чтобы можно было прийти к однозначным выводам, нужны более всесторонние исследования, сравнивающие точность распространенного машинного обучения и статистических методов.

III. ЗАДАЧА ПОНИЖЕНИЯ РАЗМЕРНОСТИ

Стремительное развитие технологий неминуемо приводит к взрывному росту объема данных. Данное обстоятельство обрекает традиционные способы распознавания данных на серьезные проблемы, поскольку они могут не справиться с огромными массивами данных. В нашем случае их затруднительно применять к матрицам OD большой размерности. Когда векторы данных являются многомерными, с вычислительной точки зрения невозможно использовать алгоритмы анализа или кластеризацией данных из-за повторяющихся вычислений отношений к определенному классу или метрик в исходном пространстве данных. В таких условиях задача уменьшения размерности [19] становится крайне актуальной, или даже необходимой. Для ее решения можно выделить два основных метода: SVD разложение и PCA.

Матричное представление сингулярного разложения записывают обычно в виде

$$\mathrm{M} = USV^T. (1)$$

Где U — матрица с векторами-столбцами $\{q^i\}_{i=1}^m$; V — матрица с векторами-столбцами $\{e^i\}_{i=1}^n$; S — матрица, у которой главный минор порядка r диагональный, на диагонали расположены сингулярные числа матрицы M, все остальные элементы S заполняются нулями [18].

Основные цели как PCA, так и SVD заключаются в том, чтобы найти основные компоненты, которые имеют максимально возможную дисперсию, чтобы зафиксировать основные характеристики матрицы. Эти компоненты будут сохранены, а другие удалены. После матричного преобразования матрица высокой размерности уменьшается до приличной степени. Низкоразмерные данные более пригодны для анализа данных и распознавания образов. Процедуры линейного преобразования PCA и SVD имеют то преимущество, что сохраняют основные особенности предыдущей высокомерной матрицы [20]. Затем на этой основе проводится кластерный или классификационный анализ.

SVD и PCA для обработки матриц корреспонденции

Трудно получить паттерны спроса непосредственно из их пространственного или временного распределения, а огромное количество столбцов затрудняет обнаружение потенциальных особенностей. Объемы трафика между станциями метро очень сильно связаны друг с другом. Поэтому использование исходной матрицы OD для проведения кластеризации может не охватить основные факторы, описывающие конкретную структуру спроса за один день, поскольку учитываются все столбцы и перекрывающаяся информация, поэтому уменьшение размерности является необходимостью уменьшения избыточности и повышения эффективности перед процедурами кластеризации [21].

Анализ главных компонент (Principal component analysis - PCA) и разложение по сингулярным числам (singular value decomposition - SVD) также уже были применены для уменьшения размерности матрицы OD метро Шэньчжэня [22]. После матричного преобразования для кластеризации низкоразмерной матрицы и определения моделей спроса в ежедневной матрице OD в городских условиях было выбрано аффинитивное распространение (affinity propagation - AP). Исходная матрица с 13 924 столбца сжата до 100 столбцов с применением преобразования оси. 100 столбцов восстанавливают более 96% исходной матрицы, а сокращение от высокого измерения к низкому также облегчает выявление основных свойств каждого дня с точки зрения понимания спроса на поездки.

Affinity propagation. В статистике, а также в интеллектуальном анализе данных алгоритм AP известен также как алгоритм кластеризации, основанный на концепции «передачи сообщений» между точками данных [23]. Количество кластеров в традиционных методах кластеризации, таких как k-means и k-medoids, обычно необходимо задавать заранее. Однако в нашем случае для матрицы корреспонденции это не всегда возможно, прежде необходимо изучить модели спроса. В отличие от этих алгоритмов, AP не требует предварительного определения или оценки количества кластеров. Этот фактор делает его более эффективным, чем другие методы.

AP находит экземпляры, которые являются членами входного набора, который может представлять кластеры. Кластеризация данных путем определения подмножества репрезентативных примеров играет важную роль в обнаружении закономерностей в данных. Такие образцы можно найти, выбирая начальное подмножество точек данных случайным образом, а затем итеративно уточняя его, но это хорошо работает, только если случайная инициализация близка к удовлетворительному решению. Фрей и Дьюек [23] разработали метод, называемый «распространение по сходству», принимая входные данные как меры подобия между парами точек данных, где сходство $s(i,k)$ предполагает, насколько хорошо точка данных с индексом *к* подходит для использования в качестве образца для точки данных *i*. Чтобы свести к минимуму квадратичную ошибку, каждому сходству затем присваивается отрицательный квадрат ошибки (евклидово расстояние): для точек x_i и x_k, $s(i,k) = -||x_i - x_k||^2$.

AP ищет допустимые конфигурации меток $c = \{c_1, c_2, \dots, c_n\}$, чтобы минимизировать энергию (n равно общему количеству столбцов) [10/12]:

$$E(c) = -\sum_{i=1}^{n} s(i,c).$$

Соответствие (responsibility - r) отражает то, насколько хорошо k подходит для того, чтобы служить примером для пункта *i*, обновляется: $r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\}$. Доступность (availability – a) обновляется путем $a(i,k) \leftarrow \min(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max(0, r(i',k)))$ для $i \neq k$, или $a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k))$. Идентификация экземпляров, соответствие и доступность могут быть собраны вместе в поиске экземпляров. Для точки *i* значение k, которое максимизирует $a(i,k) + r(i,k)$, либо определяет точку *i* в качестве примера, если *k* = *i*, либо идентифицирует точку данных, которая является примером для точки *i*. Процедура передачи сообщений может завершиться через фиксированное количество итераций после этого, или когда изменения в сообщениях ниже порогового значения. Коэффициент демпфирования вводится, чтобы избежать числовых колебаний, возникающих в некоторых случаях. l = 0,5 выбран в качестве коэффициента демпфирования по умолчанию в следующей кластеризации AP. В каждой итерации Соответствие r и доступность a обновляются с использованием взвешенной суммы r и a предыдущей итерации.

$r_i = (1-l) * r_i + l * r_{i+1}$,
$a_i = (1-l) * a_i + l * a_{i+1}$.

Доступность и соответствие обновляются по очереди в каждой итерации, а также объединяются для мониторинга примерных решений. Когда $r(k,k) + a(k,k) > 0$, k рассматривается как пример. Алгоритм окончательно завершается, когда эти решения не меняются в течение 10 итераций.

Матрицы OD метро Шэньчжэня 290*100 и 290*200 [45] группируются с использованием AF. Но прежде размерность первой матрицы уменьшается PCA, вторая уменьшается SVD. Матрица уменьшенной размерности OD, соответствующая методу PCA, сгруппирована в 11 категорий. Матрица уменьшенной размерности OD, соответствующая методу SVD, сгруппирована в 17 категорий. Таким образом, 290 дней распределены по 11 и 17 категориям соответственно. По мнению авторов, число категорий, сгруппированных по SVD, превышает соответствующий уровень, который не отражает основные характеристики среди моделей спроса в метрополитене, метод PCA показал себя эффективнее в данном исследовании. Хотя SVD достигает удовлетворительных результатов, он не в состоянии эффективно уменьшить многомерную матрицу до приличного размера. Другими словами, SVD хорошо работает за счет высокой временной сложности и намного медленнее, чем матрица PCA при выполнении кластеризации AP [24].

SVD оказался полезным в качестве вспомогательного инструмента в рамках задачи классификации станций московского метрополитена на 3 группы-зоны: рабочую, жилищную и смешанного

типа [25]. Было выдвинуто предположение о том, что данные на соседних станциях могут быть информативным признаком - 8 соседей и 1 целевой поток. Если ещё установить размер временного окна *t=5,* получится n=45 признаков. Получившаяся матрица n*m (где m - количество наблюдений) содержала много избыточной информации, поэтому перед применением классификации была проведена нормализация данных.

Также SVD показал себя инструментом для понимания внутренней структуры ежедневного спроса на поездки в метро [26]. Уравнение (1) можно переписать в виде:

$$M = USV^T = \sum_{i=1}^{r} \delta_i u_i v_i^T,$$

Где M – $m*n$ матрица с рангом r ; U – $m*r$ матрица и u_i i-й столбец U; V является матрицей n*r, и v_i является i -м столбцом V ; u_i и v_i являются единичными ортонормированными векторами; $\boldsymbol{S}$ — диагональная матрица, i -й диагональный элемент равен δ_i, который называется сингулярным значением. Была использована формула Фробениуса [27] для оценки подобия восстановленной по сингулярным значениям матрицы спроса $\widehat{M}$ и исходной матрицы M.

$$E = \frac{||\widehat{M} - M||_F}{||M||_F} * 100\%,$$

Суточная матрица ОП может быть разложена на различные матрицы OD, которые могут быть преобразованы из v_i^T. В этом исследовании v_i^T определяется как i -й образец спроса, а u_i определяется как i -й временной поток. Примечательно, что существует однозначное соответствие между моделью спроса, временным потоком и единичной ценностью. Модель спроса обозначает конкретное распределение спроса на поездки. Временной поток представляет собой временной ряд, представляющий изменение конкретной модели спроса в период исследования. Установлено, что ежедневный спрос на проезд в метро можно разложить на три составляющие: периодическую часть, пакетную часть и прочую часть [49]. Периодическая часть меняется еженедельно и составляет большую часть матрицы корреспонденции. Пакетная часть демонстрирует кратковременные всплески, вызванные особыми событиями или праздниками. Прочая часть варьируется случайным образом, её можно сопоставить с шумами, и она составляет лишь небольшую часть спроса на поездки. Периодическая часть, соответствующая двум самым большим сингулярным значениям, очень стабильна в течение половины исследуемого промежутка, пакетную часть можно использовать для анализа воздействия аварии на трафик, прочая часть мала и её воздействие на общее понимание потоков можно нивелировать.

ВРЕМЕННАЯ КЛАСТЕРИЗАЦИЯ ТРАФИКА

Первоначальный датасет является журналом записей поездок пассажиров в московском метро за февраль 2018 года (28 дней, 203 станции) с получасовыми интервалами и представлен в виде таблицы со следующими полями: дата и время; идентификатор станции отправления; идентификатор станции назначения; количество пассажиров, начавших обозначенный идентификаторами маршрут в заданное время; количество пассажиров, закончивших маршрут. Если сгруппировать записи по временным интервалам, для каждого получится своя матрица OD.

Кластеризация дней

Нашей задачей является кластеризация дней. Для дальнейшего исследования составим следующую таблицу: в качестве объектов возьмем дни месяца (от 1 до 28), а в качестве признаков все возможные маршруты между станциями с суммарным количеством поездок на нем в заданный день. Результатом будет двумерная матрица размерностью 28*41006.

Очевидно, применять стандартные методы кластеризации к объектам, имеющим 41006 признаков не представляется возможным. В качестве задачи уменьшения размерности применим сингулярное разложение к нашей матрице. Из получившихся матриц U, sigma, V нас интересует вторая, а точнее ее диагональные элементы – сингулярные числа, отсортированные по невозрастанию.

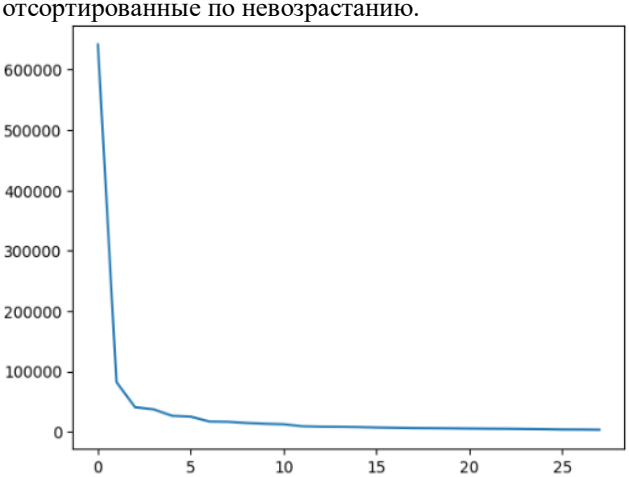
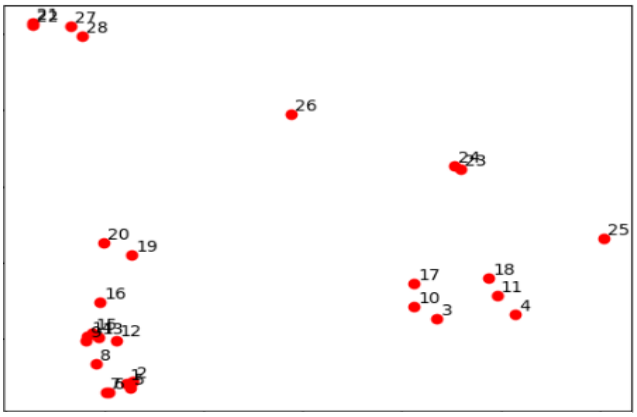


Рис. 1. Распределение сингулярных чисел

На графике сингулярных чисел видно, что заметнее остальных выделяется самое первое число - оно несет в себе 63% всей информации, со второго по шестое – 20%, тогда как на остальные 22 чисел приходится только 17%. По первым шести сингулярным значениям с помощью матрицы U восстановим уже гораздо более сжатую исходную матрицу размерностью 28*6. Для наглядности по первым двум признакам можно изобразить наши объекты на плоскости.

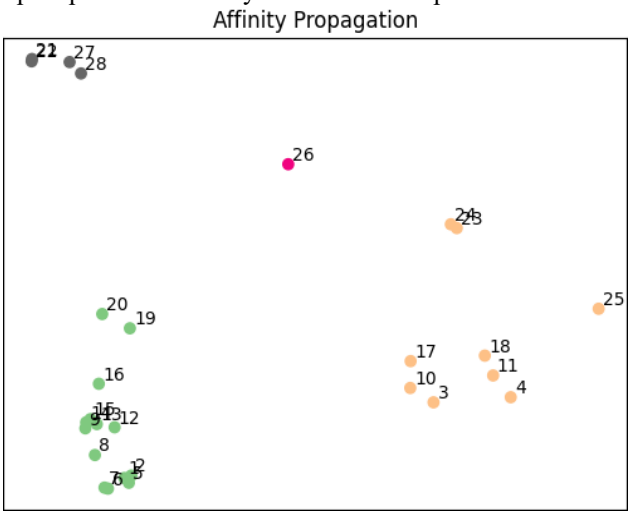


ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28				

Рис. 2. Объекты-дни на плоскости

Какое количество чисел-признаков (N) выбрать - вопрос дискуссионный, тем не менее в любом случае цель операции достигнута: количество признаков с 41006 существенно уменьшено, что открывает возможности применения методов кластеризации.

Так как количество кластеров заранее неизвестно, разумным решением будет использовать аффинное преобразование. Получилось 4 кластера:



ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28				

Рис. 3. Кластеризация дней аффинным преобразованием

1) Будние дни в начале и середине месяца

2) Будние дни в конце месяца (21, 22, 27, 28 февраля)

3) Выходные дни (в том числе праздничный день 23 февраля)

4) 26 февраля

Интерпретация результатов не является целью данного исследования, но можно попробовать предположить, что отделение будней в конце февраля в отдельный кластер (2) связано с сезонным движением, приближением весны. Либо же наоборот, с необычным похолоданием в этот период в Москве. Причем изменения касаются именно направлений движения пассажиров – общее количество поездок за день осталось на том же уровне, что и у кластера (1).

Рис. 4. Среднесуточная температура в феврале 2018

Выделение 26 февраля от остальных можно попробовать объяснить, во-первых, минимальной среднесуточной температурой за месяц, а также тем, что этот день следует за сокращенным днем 22 февраля и тремя выходными днями, что подталкивало людей отправляться в дальние поездки, возвращаясь уже 26 февраля на работу или учебу в другое время и другим маршрутом.

Также хороший результат показал метод кластеризации DBSCAN. Он в дополнение к уже указанным смог выделить 3 кластера: разделил дни субботы и воскресенья; 25 февраля; 23 и 24 февраля.

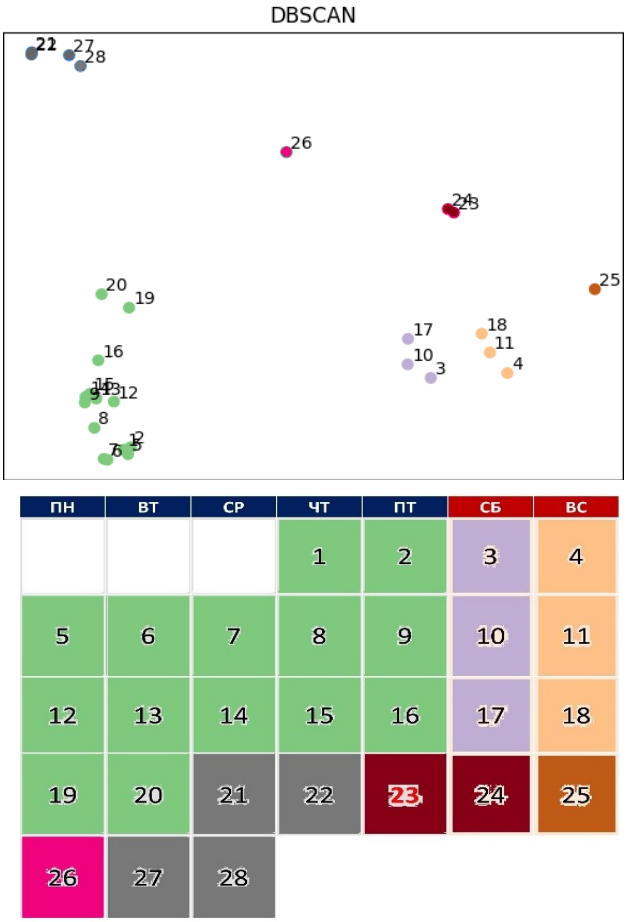


Рис. 5. Кластеризация дней методом DBSCAN

При увеличении количества признаков в восстановленной матрице изменения в результатах кластеризации обоих приведенных методов незначительны.

Кластеризация трафика по часам

Далее за основу возьмем кластеризацию дней, полученную аффинным преобразованием. Для каждого кластера построим разбиение по часам. В качестве объектов соответственно возьмем номер часа (для удобства пронумеруем от 0 до 23), а в качестве признаков суммарное количество поездок для каждой станции в роли начальной и конечной точки маршрута. Составленная таблица имеет размерность 24*406. Произведя схожую последовательность действий, что и при кластеризации дней, получим результаты для каждого кластера дней и отобразим на графике с средним количеством поездок для каждого часа.

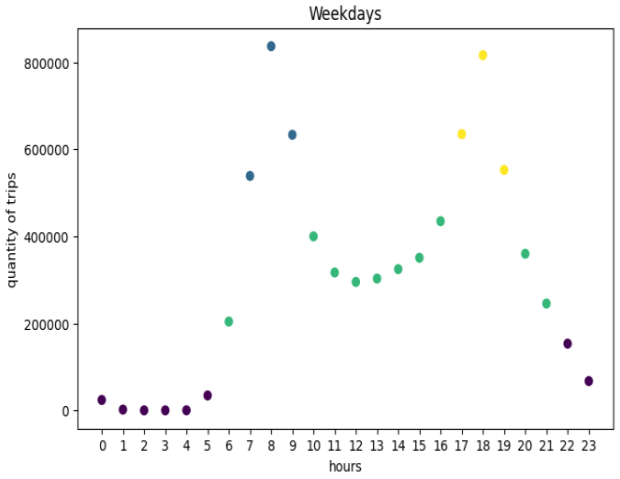


Рис. 6. Разбиение часов в будние дни

Для будней отчетливо видны кластеры час пик. Причем разделены они на 2 кластера: для утреннего (7, 8, 9) и вечернего (17, 18, 19), несмотря на то что объем трафика в эти часы почти одинаковый. Также в кластер объедены ночные часы или же раннего утра (до 6 утра) и два заключительных часа в сутках. Можно назвать этот кластер наименьшей активности, забегая вперед, он присутствует для всех кластеров дней.

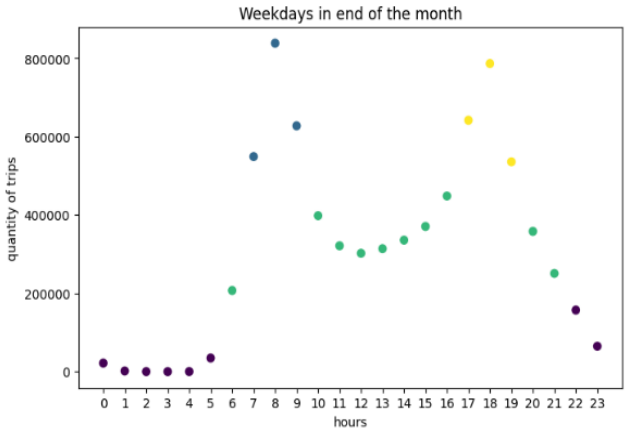


Рис. 7. Разбиение часов в будние дни в конце месяца

Будни в конце месяца оказались неотличимы по разбиению часов от остальных будней месяца, как, впрочем, и ожидалось.

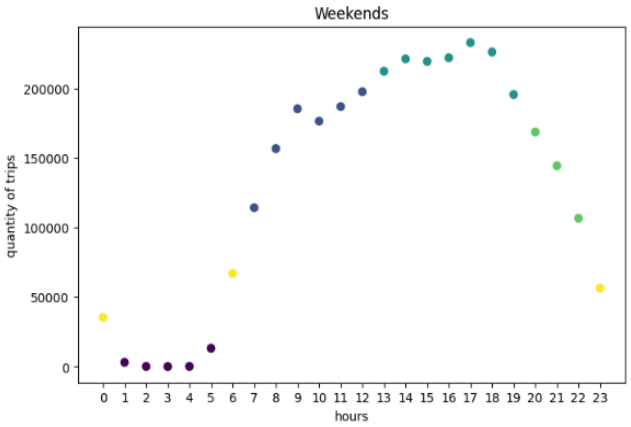


Рис. 8. Разбиение часов в выходные дни

Структура выходных дней разительно отличается от будней. О часах пик говорить не приходится, объем трафика планомерно возрастает к полудню, держится примерно на одном уровне в течении дня и угасает к позднему вечеру. Можно предположительно обозначить кластеры так: 1) Люди постепенно выезжают из своих домов в первой половине дня - кластер с 7 до 12 часов; 2) Совершают поездки между центрами притяжения - с 13 до19 часов; 3) Возвращаются домой – с 20 до 22 часов.

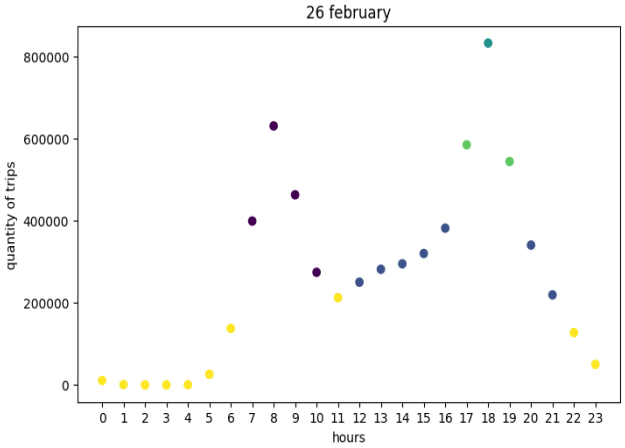


Рис. 9. Разбиение часов 26 февраля

26 февраля является стандартным будним днем, не сокращенным. Можно отметить, что в сравнении с остальными буднями, ощутимо уменьшен объем трафика утренних часов пик, тогда как объем вечерних остался примерно на том же уровне. И к кластеру наименьшей активности добавились шестой и одиннадцатый часы.

IV. АНАЛИЗ СТАНЦИЙ МЕТРО

В этом разделе будем использовать только кластер будних дней в начале и середине месяца, для остальных кластеров дней выполняется та же последовательность действий.

В качестве объектов возьмем 203 станции метро, в качестве признаков суммарное количество поездок по каждому кластеру часов, где каждая станция, выбранная, как объект, выступает в качестве начальной или конечной точки маршрута. К полученной матрице 203*8 применим метод кластеризации DBSCAN с параметром n_samples=2, чтобы найти все близкие по евклидовой метрике станции по заданным признакам. Получилось 12 групп станций, из которых 11 целевых (в 12-ую включены станции, ни попавшие ни в одну целевую). Для всех групп сохраняется признак равноудаленности включенных станций от центра города. То есть группы содержат только центральные станции, только станции на окраине города и так далее.

Не выполняется это условие только для одной станции в одной из групп (рис 10). Слева приписано название ветки, справа название станции, также каждой ветке сопоставлен свой цвет.

Рис. 10. Один из кластеров похожих друг на друга станций

Станция Третьяковская находится в центре города, но попала в один кластер со станциями на периферии.

V. ЗАКЛЮЧЕНИЕ

В процессе анализа предметной области была выработана методика анализа транспортного трафика, основанная на связке сингулярного разложения и методов кластеризации машинного обучения. Концепция использовать в связке упомянутые инструменты не нова, применялась в других областях, нам же удалось успешно адаптировать ее для транспортной сферы, дополнив многочисленными модификациями. Написан библиотечный программный модуль для реализации каждого этапа разработки, упомянутого в исследовании. Модуль способен обрабатывать большие объемы данных, имеет потенциал для легкого масштабирования и расширения.

БЛАГОДАРНОСТИ

Авторы благодарны сотрудникам Центра Высокоскоростных транспортных систем РУТ (МИИТ) за предоставленные данные. Также хотелось поблагодарить В.П. Куприяновского [28, 29, 30], чьи

многочисленные работы с соавторами способствовали как открытию направления транспортных исследований, так и множеству соответствующих публикаций в журнале INJOIT. В работе, среди прочих материалов, использовались и ранние исследования матриц корреспонденции [31, 32, 33].

Библиография

[1] Van Arem B. et al. Recent advances and applications in the field of short-term traffic forecasting //International journal of forecasting. – 1997. – Т. 13. – №. 1. – С. 1-12.

[2] Papageorgiou M. (ed.). Concise encyclopedia of traffic and transportation systems. – Pergamon, 1991. – Т. 6.

[3] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, “Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach,” Transportation Research Part C: Emerging Technologies, vol. 13, no. 3, pp. 211–234, 2005.

[4] Y. Wei and M.-C. Chen, “Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks,” Transportation Research Part C: Emerging Technologies, vol. 21, no. 1, pp. 148–162, 2012.

[5] Xiong X. et al. Dynamic origin–destination matrix prediction with line graph neural networks and kalman filter //Transportation Research Record. – 2020. – Т. 2674. – №. 8. – С. 491-503.

[6] Ashok, K., and M. E. Ben-Akiva. Alternative Approaches for Real-Time Estimation and Prediction of TimeDependent Origin–Destination Flows. Transportation Science, Vol. 34, No. 1, 2000, pp. 21–36.

[7] X. Jiang, L. Zhang, and X. M. Chen, “Short-term forecasting of highspeed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in china,” Transportation Research Part C: Emerging Technologies, vol. 44, pp. 110–127, 2014.

[8] Y. Chen, H. S. Mahmassani, and Z. Hong, “Data mining and pattern matching for dynamic origin–destination demand estimation: Improving online network traffic prediction,” Transportation Research Record: Journal of the Transportation Research Board, no. 2497, pp. 23–34, 2015.

[9] Zhu, Z., Peng, B., Xiong, C., & Zhang, L. (2016). Short-term traffic flow prediction with linear conditional gaussian Bayesian network. Journal of Advanced Transportation, 50(6), 1111–1123.

[10] Zhang, Y., & Zhang, Y. (2016). A comparative study of three multivariate short-term freeway traffic flow forecasting methods with missing data. Journal of Intelligent Transportation Systems, 20(3), 205–218.

[11] Jiang, H., Zou, Y., Zhang, S., Tang, J., & Wang, Y. (2016). Short-term speed prediction using remote microwave sensor data: Machine learning versus statistical model. Mathematical Problems in Engineering, 2016, 1–13.

[12] Wu, Y.-J., Chen, F., Lu, C.-T., & Yang, S. (2016). Urban traffic flow prediction using a spatio-temporal random effects model. Journal of Intelligent Transportation Systems, 20(3), 282–293.

[13] Zhu, Z., Peng, B., Xiong, C., & Zhang, L. (2016). Short-term traffic flow prediction with linear conditional gaussian Bayesian network. Journal of Advanced Transportation, 50(6), 1111–1123.

[14] Shahsavari, B., & Abbeel, P. (2015). Short-term traffic forecasting: Modeling and learning spatio-temporal relations in transportation networks using graph neural networks. University of California at Berkeley, Technical Report No. UCB/EECS-2015-243.

[15] Qing, L., Yongqin, T., Yongguo, H., & Qingming, Z. (2014). The forecast and the optimization control of the complex traffic flow based on the hybrid immune intelligent algorithm. Open Electrical & Electronic Engineering Journal, 8, 245–251.

[16] Dong, C., Shao, C., Richards, S. H., & Han, L. D. (2014). Flow rate and time mean speed predictions for the urban freeway network using state space models. Transportation Research Part C: Emerging Technologies, 43, 20–32.

[17] Dimitriou, L., Tsekeris, T., & Stathopoulos, A. (2008). Adaptive hybrid fuzzy rule-based system approach for modeling and predicting urban traffic flow. Transportation Research Part C: Emerging Technologies, 16(5), 554–573.

[18] Карчевский Е. М., Карчевский М. М. Лекции по линейной алгебре и аналитической геометрии: учебное пособие. – 2012.

[19] Roweis, Sam T., and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding., Science 290. 5500 (2000): 2323-2326.

[20] S. Sun, C. Zhang, and G. Yu, “A Bayesian network approach to traffic flow forecasting,” IEEE Trans. Intell. Transp. Syst., vol. 7, no. 1, pp. 124-132, Mar. 2006.

[21] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", Science, vol. 290, no. 5500, pp. 2323-2326, Dec. 2000.

[22] Yang C, Yan FF, Xu XD. Clustering Daily Metro Origin-Destination Matrix in Shenzhen China. AMM 2015;743: 422–32.

[23] Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." *science* 315.5814 (2007): 972-976.

[24] C. Yang, F. Yan and X. Xu, "Daily metro origin-destination pattern recognition using dimensionality reduction and clustering methods," 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 2017, pp. 548-553.

[25] Некраплённая М.Н., and Намиот Д.Е.. "Анализ матриц корреспонденции метро" International Journal of Open Information Technologies, vol. 7, no. 7, 2019, pp. 68-80.

[26] Duan Z. Lei Z. Zhang M. et al.: ‘Understanding multiple days metro travel demand at aggregate level’, IET Intell. Transp. Syst., 2018, 13, (5), pp. 756– 763.

[27] Zhi Y. Li H. Wang D. et al.: ‘Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data’, Geo Spat. Inf. Sci., 2016, 19, (2), pp. 94–105.

[28] Куприяновская, Ю. В., et al. "Умный контейнер, умный порт, BIM, Интернет Вещей и блокчейн в цифровой системе мировой торговли." International Journal of Open Information Technologies 6.3 (2018): 49-94.

[29] Николаев, Д. Е., et al. "Цифровая железная дорога-инновационные стандарты и их роль на примере Великобритании." International Journal of Open Information Technologies 4.10 (2016): 55-61.

[30] Куприяновский В. П., Намиот Д. Е., Синягов С. А. Демистификация цифровой экономики //International Journal of Open Information Technologies. – 2016. – Т. 4. – №. 11. – С. 59-63.

[31] Misharin, A., D. Namiot, and O. Pokusaev. "On Processing of Correspondence Matrices in Transport Systems." 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon). IEEE, 2019.

[32] Namiot, Dmitry, and Oleg Pokusaev. "On mobility patterns in Smart City." CEUR Workshop Proceedings. 2019.

[33] Pokusaev, O., and D. Namiot. "Anomalies in Transport Data." IOP Conference Series: Earth and Environmental Science. Vol. 666. No. 5. IOP Publishing, 2021.

Статья получена 1 мая 2023 года.

Д.Т. Оспанов – МГУ имени М.В. Ломоносова, email: daulet705@gmail.com

Д.Е. Намиот – МГУ имени М.В. Ломоносова email:dnamiot@gmail.com

О.Н. Покусаев – РУТ (МИИТ), email: o.pokusaev@rut.digital

Structural and temporal analysis of metro passenger traffic

Daulet Ospanov, Dmitry Namiot, Oleg Pokusaev

Abstract — **The article is devoted to one approach to the analysis of traffic flows. The initial data for analysis are the so-called correspondence matrices, which describe the number of trips per unit of time between two points. The specific dataset that was analyzed in the work is a matrix of Moscow metro correspondence (passenger trips between metro stations) for February 2018. The purpose of the analysis is a structural-temporal analysis of passenger traffic (how and when passengers move). The paper proposes a method for analyzing transport traffic based on a combination of singular value decomposition and machine learning clustering methods. Singular value decomposition is used here for dimensionality reduction. The concept of using these tools in conjunction is not new, it has been used in other areas, but in this work, it has been successfully adapted specifically for the transport sector. The article presents a library software module for the implementation of each stage of the development of the proposed model. The module is capable of processing large amounts of data and has the potential for easy scaling and expansion. The paper presents an example of the implementation of the proposed methodology in relation to historical data on the passenger flow of the Moscow metro.**

Keywords— **passenger traffic, metro, singular value decomposition, clustering methods, machine learning, correspondence matrix.**

REFERENCES

[1] Van Arem B. et al. Recent advances and applications in the field of short-term traffic forecasting //International journal of forecasting. – 1997. – T. 13. – #. 1. – S. 1-12.

[2] Papageorgiou M. (ed.). Concise encyclopedia of traffic and transportation systems. – Pergamon, 1991. – T. 6.

[3] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach," Transportation Research Part C: Emerging Technologies, vol. 13, no. 3, pp. 211–234, 2005.

[4] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," Transportation Research Part C: Emerging Technologies, vol. 21, no. 1, pp. 148–162, 2012.

[5] Xiong X. et al. Dynamic origin–destination matrix prediction with line graph neural networks and kalman filter //Transportation Research Record. – 2020. – T. 2674. – #. 8. – S. 491-503.

[6] Ashok, K., and M. E. Ben-Akiva. Alternative Approaches for Real-Time Estimation and Prediction of TimeDependent Origin–Destination Flows. Transportation Science, Vol. 34, No. 1, 2000, pp. 21–36.

[7] X. Jiang, L. Zhang, and X. M. Chen, "Short-term forecasting of highspeed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in china," Transportation Research Part C: Emerging Technologies, vol. 44, pp. 110–127, 2014.

[8] Y. Chen, H. S. Mahmassani, and Z. Hong, "Data mining and pattern matching for dynamic origin–destination demand estimation: Improving online network traffic prediction," Transportation Research Record: Journal of the Transportation Research Board, no. 2497, pp. 23–34, 2015.

[9] Zhu, Z., Peng, B., Xiong, C., & Zhang, L. (2016). Short-term traffic flow prediction with linear conditional gaussian Bayesian network. Journal of Advanced Transportation, 50(6), 1111–1123.

[10] Zhang, Y., & Zhang, Y. (2016). A comparative study of three multivariate short-term freeway traffic flow forecasting methods with missing data. Journal of Intelligent Transportation Systems, 20(3), 205–218.

[11] Jiang, H., Zou, Y., Zhang, S., Tang, J., & Wang, Y. (2016). Short-term speed prediction using remote microwave sensor data: Machine learning versus statistical model. Mathematical Problems in Engineering, 2016, 1–13.

[12] Wu, Y.-J., Chen, F., Lu, C.-T., & Yang, S. (2016). Urban traffic flow prediction using a spatio-temporal random effects model. Journal of Intelligent Transportation Systems, 20(3), 282–293.

[13] Zhu, Z., Peng, B., Xiong, C., & Zhang, L. (2016). Short-term traffic flow prediction with linear conditional gaussian Bayesian network. Journal of Advanced Transportation, 50(6), 1111–1123.

[14] Shahsavari, B., & Abbeel, P. (2015). Short-term traffic forecasting: Modeling and learning spatio-temporal relations in transportation networks using graph neural networks. University of California at Berkeley, Technical Report No. UCB/EECS-2015-243.

[15] Qing, L., Yongqin, T., Yongguo, H., & Qingming, Z. (2014). The forecast and the optimization control of the complex traffic flow based on the hybrid immune intelligent algorithm. Open Electrical & Electronic Engineering Journal, 8, 245–251.

[16] Dong, C., Shao, C., Richards, S. H., & Han, L. D. (2014). Flow rate and time mean speed predictions for the urban freeway network using state space models. Transportation Research Part C: Emerging Technologies, 43, 20–32.

[17] Dimitriou, L., Tsekeris, T., & Stathopoulos, A. (2008). Adaptive hybrid fuzzy rule-based system approach for modeling and predicting urban traffic flow. Transportation Research Part C: Emerging Technologies, 16(5), 554–573.

[18] Karchevskij E. M., Karchevskij M. M. Lekcii po linejnoj algebre i analiticheskoj geometrii: uchebnoe posobie. – 2012.

[19] Roweis, Sam T., and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding., Science 290. 5500 (2000): 2323-2326.

[20] S. Sun, C. Zhang, and G. Yu, "A Bayesian network approach to traffic flow forecasting," IEEE Trans. Intell. Transp. Syst., vol. 7, no. 1, pp. 124-132, Mar. 2006.

[21] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", Science, vol. 290, no. 5500, pp. 2323-2326, Dec. 2000.
[22] Yang C, Yan FF, Xu XD. Clustering Daily Metro Origin-Destination Matrix in Shenzhen China. AMM 2015;743: 422–32.
[23] Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." science 315.5814 (2007): 972-976.
[24] C. Yang, F. Yan and X. Xu, "Daily metro origin-destination pattern recognition using dimensionality reduction and clustering methods," 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 2017, pp. 548-553.
[25] Nekrapljonnaja M.N., and Namiot D.E.. "Analiz matric korrespondencii metro" International Journal of Open Information Technologies, vol. 7, no. 7, 2019, pp. 68-80.
[26] Duan Z. Lei Z. Zhang M. et al.: 'Understanding multiple days metro travel demand at aggregate level', IET Intell. Transp. Syst., 2018, 13, (5), pp. 756– 763.
[27] Zhi Y. Li H. Wang D. et al.: 'Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data', Geo Spat. Inf. Sci., 2016, 19, (2), pp. 94–105.
[28] Kuprijanovskaja, Ju. V., et al. "Umnyj kontejner, umnyj port, BIM, Internet Veshhej i blokchejn v cifrovoj sisteme mirovoj torgovli." International Journal of Open Information Technologies 6.3 (2018): 49-94.
[29] Nikolaev, D. E., et al. "Cifrovaja zheleznaja doroga-innovacionnye standarty i ih rol' na primere Velikobritanii." International Journal of Open Information Technologies 4.10 (2016): 55-61.
[30] Kuprijanovskij V. P., Namiot D. E., Sinjagov S. A. Demistifikacija cifrovoj jekonomiki //International Journal of Open Information Technologies. – 2016. – T. 4. – #. 11. – S. 59-63.
[31] Misharin, A., D. Namiot, and O. Pokusaev. "On Processing of Correspondence Matrices in Transport Systems." 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon). IEEE, 2019.
[32] Namiot, Dmitry, and Oleg Pokusaev. "On mobility patterns in Smart City." CEUR Workshop Proceedings. 2019.
[33] Pokusaev, O., and D. Namiot. "Anomalies in Transport Data." IOP Conference Series: Earth and Environmental Science. Vol. 666. No. 5. IOP Publishing, 2021.