О возможностях применения данных сотовых операторов для решения задач цифровой урбанистики

М.В. Булыгин, Д.Е. Намиот

***Аннотация* — В настоящий момент, широкое проникновение мобильных устройств позволяет использовать данные, собираемые в процессе их эксплуатации, для анализа транспортных потоков. Ассоциируя мобильные телефоны с их владельцами, можно по перемещениям мобильных устройств судить, например, о структуре транспортных потоков в городе. Эти задачи не имеют ничего общего с каким-то слежением за мобильными абонентами, здесь используются только анонимизированные данные. Более того, в силу специфики задач анализа транспортных потоков, здесь принципиально не интересны отдельные перемещения, а необходимы именно общие (агрегированные) данные по перемещениям между выбранными объектами (участками). В данной статье представлены основные направления использования агрегированных данных сотовых операторов такие, как поиск изменений транспортных потоков, соответствующих важным социальным событиям, измерение этих изменений, поиск новизны в транспортных потоках, кластеризация районов и связей между ними, а также оценка распределения людей по районам города. В статье описан подход к выявлению аномалий в данных транспортных потоков (исследуются аномалии во временном домене), который основывается на анализе именно агрегированных данных сотовых операторов. В работе представлен вычислительный эксперимент, демонстрирующий корректность предложенного метода.**

***Ключевые слова*—цифровая урбанистика, анализ данных сотовых операторов, анализ транспортных потоков, поиск аномалий в данных.**

Введение

В статье представлен расширенный и дополненный материал доклада на конференции 28th Fruct [1].

По исследованиям международного агентства “We are social” в апреле 2020 года в мире насчитывалось 5.16 миллиарда пользователей мобильных телефонов, что составляло 66% от общей численности населения [2]. Прирост пользователей мобильных телефонов по сравнению с апрелем 2019 года составил 128 миллионов пользователей.

Статья получена 30 декабря 2020 М.В. Булыгин, МГУ имени М.В. Ломоносова (e-mail: messimm@yandex.ru).
Д. Е. Намиот, МГУ имени М.В. Ломоносова (e-mail: messimm@yandex.ru).

В процессе использования мобильных устройств порождаются данные, которые могут быть использованы для решения прикладных задач. В частности, при совершении вызовов, подключении к Интернету, а также при отправке SMS-сообщений мобильное устройство ведет обмен данными с базовыми станциями сотового оператора. На основе информации с разных базовых станций об уровне сигнала от устройства и задержке передачи сигнала местоположение устройства может быть установлено (рис. 1).

Подробному описанию этих методов, стандартов и технологий, используемых при определении геопозиции абонента, посвящена, например, работа [3].

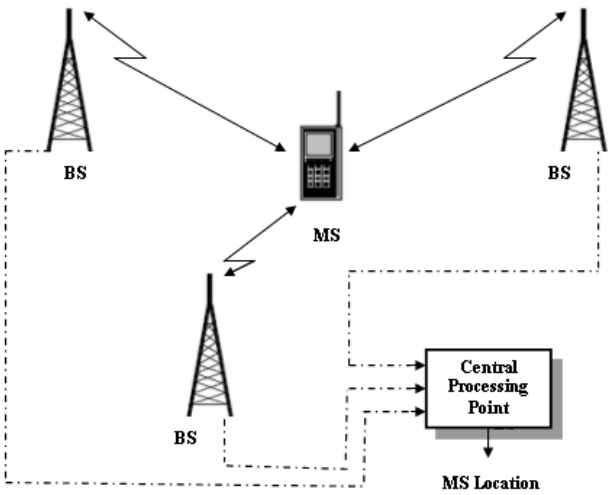

Рис. 1. Обмен информацией между устройством и базовыми станциями.

Такие данные могут сохраняться, агрегироваться, а также обрабатываться сотовыми операторами. Не агрегированные данные являются более подробными и позволяют исследовать индивидуальные паттерны перемещения, но обладают большим объемом, а также имеют задержку получения. Агрегированные данные содержат информацию, собранную по отдельным районам и временным интервалам. Такие данные доступны с меньшей временной задержкой и позволяют исследовать паттерны транспортного поведения в целых районах. В

процессе агрегации также теряются данные об индивидуальных траекториях, что делает использование таких данных конфиденциальным и не нарушает приватность пользователей.

Примеры использования данных сотовых операторов для решения практических прикладных задач можно найти в статье Ф.Калабрезе «Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome». В ней авторами описаны программные решения, которые помогают найти ответы на вопросы о перемещении людей и транспорта в Риме [4].

Одной из наиболее актуальных проблем анализа агрегированных данных сотовых операторов является выявление в них аномалий, соответствующих социальным событиям. Решению подобных проблем при помощи вейвлет-анализа посвящена статья Росарио Д. и др. «Anomaly detection mechanisms to find social events using cellular traffic data»[5].

Для Москвы и Московской области характерно наличие агрегированных по районам данных о перемещениях абонентов сотовых операторов между районами по получасовым интервалам. Благодаря наличию базовых станций сотовых операторов в Московском метрополитене, а также развитию методов анализа данных для исследования доступны данные, представленные в таблице 1.

Таблица 1 ДОСТУПНЫЕ ДАННЫЕ

Временная метка	**Район отправления**	**Район прибытия**	**Количество человек, совершивших поездку**	**Количество человек, совершивших поездку с использованием метро**	**Количество человек, совершивших поездку с работы домой**	**Количество человек, совершивших поездку из дома на работу**	**Количество человек, вернувшихся в район отправления**	**Количество человек, завершивших поездку**	**Количество человек, остававшихся в районе не менее получаса**

Задачам, которые могут быть решены с использованием подобных данных, посвящены работы российских авторов [6-9].

ВОЗМОЖНЫЕ НАПРАВЛЕНИЯ ИСПОЛЬЗОВАНИЯ АГРЕГИРОВАННЫХ ДАННЫХ СОТОВЫХ ОПЕРАТОРОВ

Работы ранних авторов во многом были сосредоточены на нахождении транспортных потоков и их предсказании. С развитием технологий сотовой связи и совершенствованием алгоритмов нахождения местоположения абонентов реальные значения транспортных потоков между районами стали доступны с низкой временной задержкой, а также достаточно высокой точностью благодаря высокой степени проникновения мобильной связи в городах.

Наличие таких данных позволяет по-новому решать широкий круг задач городского планирования с достаточно высокой точностью. Важные социальные события такие, как проведение праздничных мероприятий и больших концертов вызывают изменения в транспортных потоках города. Для того, чтобы сделать жизнь горожан комфортнее, городским властям необходимо реагировать на эти изменения. При этом возникают задачи пространственной и временной оценки этих изменений, а также их детектирования. Значимым социальным событиям соответствуют значения транспортных потоков, отличающихся от характерных для данного направления в данный временной промежуток. Транспортные потоки в районах являются достаточно устойчивыми и постоянными (рис. 2), поэтому типичными, характерными значениями будем называть значения, близкие к среднему для данного временного промежутка, типа дня (будний, выходной, праздничный и др.) и направления. Это важный момент. Мы определяем аномалии не по трафику, предшествующему данному, а по трафику в те же самые промежутки в такие же дни. Трафик с 10:00 до 11:00 в конкретный понедельник может сильно варьироваться, но трафик с 10:00 до 11:00 по понедельникам будет стабильным, если, конечно, не случиться каких-либо отклонений. Значения, сильно отличающиеся от типичных по абсолютному значению, будем называть аномальными. Аномалии определенные именно таким образом и представляют наибольший интерес для управления инфраструктурой города. Отметим, что при таком подходе происходит автоматический учет сезонных изменений. Трафик во вторник в июле будет сравниваться с такими же вторниками в летние месяцы, автоматически принимая во внимание сезонные изменения.

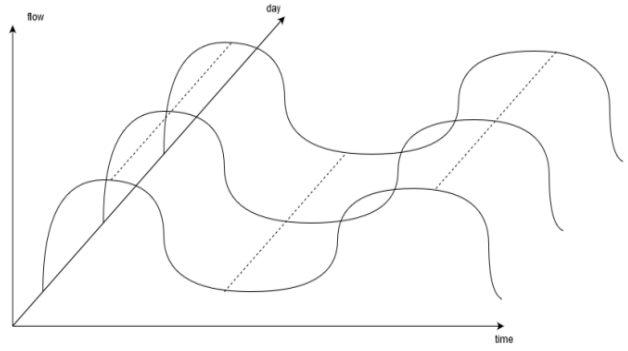

Рис. 2. Транспортные потоки

После нахождения аномалий в данных транспортных потоков возникают задачи их анализа. Интерес может представлять задача кластеризации аномалий. Так как аномалии в транспортных потоках соответствуют важным социальным событиям, то такую кластеризацию также можно считать кластеризацией этих событий.

Отметим, что может производиться как кластеризация отдельных аномальных наблюдений, так и кластеризация наблюдений в некотором контексте предыдущих и последующих наблюдений. Учет такого контекста позволит кластеризовать аномалии, а, главное, соответствующие им социальные события, по специфике их возникновения, развития и завершения. При этом возникают задачи конструирования признаковых пространств, описывающих аномалии и введения на них метрик.

Технически, можно отметить, по крайней мере, несколько типов аномалий. Например, аномалия, связанная с тем, что временно сместилось время поездки. Перемещаются те же самые люди, но позже (раньше) по времени - «аномалия смещения». Или же «аномалия добавления», когда люди поехали после завершения концерта, футбольного матча и т.п. – то есть появились новые пассажиры, которые обычно здесь в такое время не ездили.

Также возможен поиск аномалий в пространственном домене. Для этого необходимо провести анализ распределения транспортных потоков. Поиск аномалий в распределениях по районам входящего транспортного потока позволит понять, жителей каких районов привлекают важные события рассматриваемого района. При рассмотрении распределения исходящего потока возможно построение выводов о том, события каких районов влияют на транспортный поток в рассматриваемом районе. Данные о таких изменениях позволят урбанистам изменять работу городской инфраструктуры (светофоров, общественного транспорта и др.) в соответствии с этими изменениями в направлениях передвижений горожан.

Со временем аномальные наблюдения могут становиться «новой нормальностью». Это может быть вызвано сезонными изменениями, например, горожане стали ездить по вечерам на дачи в весенний период вместо возвращения домой. Также появление такой новизны может быть вызвано строительством и открытием новых жилых комплексов, что вызывает увеличение исходящего транспортного потока по утрам (жители комплекса стали выезжать на работу из дома) и входящего по вечерам (жители комплекса возвращаются на место проживания).

Изменения могут наблюдаться не только во временной области, но и в пространственной. Так характерные изменения могут появиться с открытием и важных объектов: бизнес-центров, парков, торговых центров, которые привлекают население, а, значит, . Подобные изменения могут наблюдаться и при открытии новых маршрутов транспорта, особенно из пригородных территорий (люди из пригорода делали пересадку на метро в районе А, так как туда ходил автобус из пригорода, после открытия же маршрута в район Б).

Анализ данных сотовых операторов может быть одним из методов нахождения этой новизны в транспортных потоках и её оценки. Такая оценка может помочь в планировании долгосрочных изменений транспортной сети, например, введении новых постоянных маршрутов общественного транспорта.

Благодаря установке базовых станций сотовых операторов в метро стало возможно решение задач оценки изменений связанных с проведением изменений в метрополитене. Например, может быть оценена доля поездок, которые были совершены с использованием метро. Увеличение этой доли и величина её прироста может служить показателем успешности открытия новой станции.

Также в данных фиксируется временной интервал окончания поездок между районами. Изменения распределения количества горожан, окончивших поездку, которые не являются постоянными, могут говорить о нарушении функционирования транспортных каналов между районами (крупные ДТП, остановки метро). Постоянные изменения могут говорить об использовании новых транспортных каналов (горожане стали совершать поездки между районами быстрее, что положительно сказывается на качестве их жизни).

Имеющиеся данные могут быть использованы не только для поиска аномалий и новизны в данных и её измерения. Кластеризация исходных данных (характеристик транспортного потока из района *А* в район *Б* в определенный получасовой интервал на заданную дату) не представляет практического интереса. С использованием имеющихся данных возможно проведение кластеризации районов и связей между ними. При этом возникает задача формирования признаков, описывающих районы или связи между ними. В зависимости от выбора признакового пространства выделенные кластеры могут иметь различную интерпретацию. Интересным с точки зрения урбанистики может быть выделение кластеров, соответствующих спальным и «рабочим» районам (а также районам, занимающим промежуточное положение). Результаты такой кластеризации могут быть использованы для предоставления наиболее важных для района сервисов в соответствии с его типом. Также результаты такой кластеризации могут быть включены в качестве признака при решении различных прикладных задач (задача оценки стоимости жилых площадей в районе, прогнозирование числа пассажиров такси или посещаемости нового магазина и др.)

При кластеризации связей между районами можно выбрать признаковое пространство таким образом, чтобы полученные кластеры соответствовали «районам-донорам» (районы, из которых люди прибывают в данный район на работу) и «районам-потребителям» (районы, в которые люди прибывают из данного района на работу). В более широком смысле

«районы-потребители» могут привлекать жителей рассматриваемого района не только как место работы, но и как районы, обладающие какими-то объектами социальной инфраструктуры: парками для прогулок, супермаркетами, ресторанами. В некоторых случаях возможно выявление этих объектов и строительство равноценных в рассматриваемом районе. Это уменьшит необходимость в перемещении между районами и может поднять качество жизни жителей района.

Также отметим, что данные сотовых операторов могут быть использованы не только для оценки транспортных потоков, но и для установления распределения количества людей по районам города (в каких именно районах находятся люди в течение дня и в каком количестве). Этот подход является более точным и дешевым, чем проведение переписи населения. Данные о распределении людей фактически фиксируются каждые полчаса, что позволяет проводить изучение изменения распределения населения города с более высокой точностью. Такие данные, например, дают возможность понять, как изменилось население города (а также его пригородов) и его распределение по районам во время проведения мероприятий по борьбе с COVID-19. Также возможно проанализировать, как влияют новогодние праздники на население города (в каких районах оно увеличивается за счет туристов, а в каких убывает за счет горожан, отправившихся на отдых и в другие регионы). Существовавшие прежде методы подсчета населения, такие как переписи (слишком редкие и дорогостоящие), анализ агрегированных данных приложений, работающих с GPS (низкая массовость) не давали такой возможности.

Обобщая всё выше сказанное, отметим, что описанный круг задач цифровой урбанистики, решение которых может быть проведено с использованием агрегированных данных сотовых операторов, не является полным. Целью данной статьи является не определение всех задач, которые могут быть решены с использованием агрегированных данных сотовых операторов, а демонстрация новых возможностей, открывающихся при проведении анализа подобных данных для специалистов в области цифровой урбанистики.

Далее в статье описан метод выявления аномалий в транспортных потоках (во временном домене), который основан на анализе агрегированных данных сотовых операторов и использует особенности этих данных.

Выявление аномалий в транспортных потоках с использованием агрегированных данных сотовых операторов

Применение существующих методов поиска аномалий в данных

Зафиксировав район отправления и район назначения, можно получить временные ряды перемещений для каждой пары районов. Обзору методов выявления аномальных наблюдений во временных рядах посвящена работа [10].

В данной работе автор выделяет три основных группы методов поиска аномалий в данных. Первая группа – это методы, основанные на основе предсказательных моделей. Наиболее популярными предсказательными моделями для временных рядов являются ARIMA-модели, модели на основе классического машинного обучения (случайных лесов, бустинга и др.), а также модели, основанные на нейронных сетях.

Данные транспортных потоков имеют значимую автокорреляцию между наблюдениями в прошлый момент времени, а также с наблюдениями недельной давности. Эта автокорреляция нарушается в праздничные дни, поэтому ARIMA-модели дают ложные сигналы об аномалиях. По этой причине возможны ложные сигналы и через неделю после праздничных дней, а также через неделю после аномалий. Для решения данной проблемы могут использоваться SARIMAX-модели, где в качестве дополнительных признаков учитываются данные об аномалиях, праздниках и других событиях. Процесс построения SARIMAX-модели также является сложным, так как необходимо приведение временных рядов к стационарному виду, разметка аномальных наблюдений, а также перенастройка модели при появлении новизны в данных. Этими же недостатками, исключая необходимость приведения ряда к стационарному виду, обладают предиктивные модели на основе случайных лесов и градиентного бустинга. Главным достоинством этой группы методов является построение предиктивной модели, которая может использоваться для решения других задач. В случае работы с агрегированными данными сотовых операторов это преимущество не является важным, так как реальные значения доступны с небольшой временной задержкой.

Второй группой моделей выявления аномалий во временных рядах, выделенной в работе [10], являются методы на основе кластеризации. Самым популярным методом этой группы является использование кластеризации DBSCAN с изменяемым пороговым значением. Главным недостатком этого метода, как и других методов этой группы, является то, что значение, которое является аномальным для одного интервала, является обычным для другого интервала. Например, значение, обычное для часа-пик в будний день, может являться аномальным для утреннего интервала выходного дня. Таким образом, для использования алгоритмов этой группы необходимо конструировать признаки для учета даты и времени наблюдения. Преимуществами данной группы методов является отсутствие необходимости построения качественной предиктивной модели, а также возможность находить аномальные значения, которые не являются экстремальными.

Третьей группой методов является статистический профайлинг. Методы данной группы дают самые контролируемые и легко трактуемые результаты. Также к преимуществам можно методов данной группы можно отнести низкую вычислительную сложность, так как вычисляемые статистики, как

правило, имеют низкую вычислительную сложность. Также эти статистики могут быть легко пересчитаны при появлении новизны в данных, например, при появлении нового транспортного канала. К недостаткам методов этой группы относится необходимость правильного выбора статистик. При анализе данных о транспортных потоках и использовании в качестве статистики (100-*a*)-перцентиля *a*% наблюдений будут признаваться аномальными, хотя они таковыми не являются. Аномальные значения, не являющиеся экстремальными, будут пропущены при использовании такой статистики. В случае использования среднего или медианного значения с некоторым порогом также могут быть пропущены некоторые аномалии. Транспортный поток в выходные для спальных районов меньше, чем по будням. При взятии среднего получается значение, которое не является характерным ни для буднего, ни для выходного дня. Такое значение является аномальным, но будет пропущено алгоритмом.

Обобщим существующие преимущества и недостатки использования данных групп методов для выявления аномалий в данных о транспортных потоках в виде таблицы.

Предложенный метод

Данные о транспортных потоках между районами обладают свойствами, которые могут быть использованы при поиске аномалий. Такие временные ряды имеют сильную автокорреляцию: наблюдения дней одного типа (будни, выходные) имеют сильную связь. Данные об интервалах с одинаковыми метками времени и типа дня при отсутствии аномалий не являются сильно осцилирующими и сосредотачиваются в области средних значений. Для поиска аномалий во временных рядах, соответствующих транспортным потокам, предлагается алгоритм, который относится к алгоритмам статистического профайлинга.

На вход алгоритм принимает значение временного ряда, для которого проверяется аномальность, день наблюдения, время наблюдения, тип дня наблюдения, параметр, регулирующий чувствительность к новизне данных, исторические данные о значениях временного ряда, а также порог, регулирующий чувствительность распознавания аномалий. В качестве выходного значения выдается 0, если наблюдение является типичным и 1, если является аномальным.

Суть алгоритма состоит в том, что модуль разности некоторого значения, рассчитанного по историческим данным, с проверяемым значением сравнивается с порогом, передаваемым в качестве параметра. Если этот модуль больше порога, то значение признается аномальным, иначе типичным.

Значение для взятия разности рассчитывается по историческим данным как среднее для получасовых интервалов с такими же метками времени и типа дня, как наблюдение, для которого проверяется аномальность, за некоторое количество предыдущих дней. Это количество задается входным параметром и позволяет регулировать чувствительность алгоритма к новизне. Чем больше исторических дней рассматривается, тем меньше вклад в среднее отдельных дней, тем менее гибким, но более устойчивым становится алгоритм. В случае выбора малого значения данного параметра рассматривается малый промежуток исторических данных, новизна быстро становится типичной.

Схема работы алгоритма представлена на рисунке ниже.

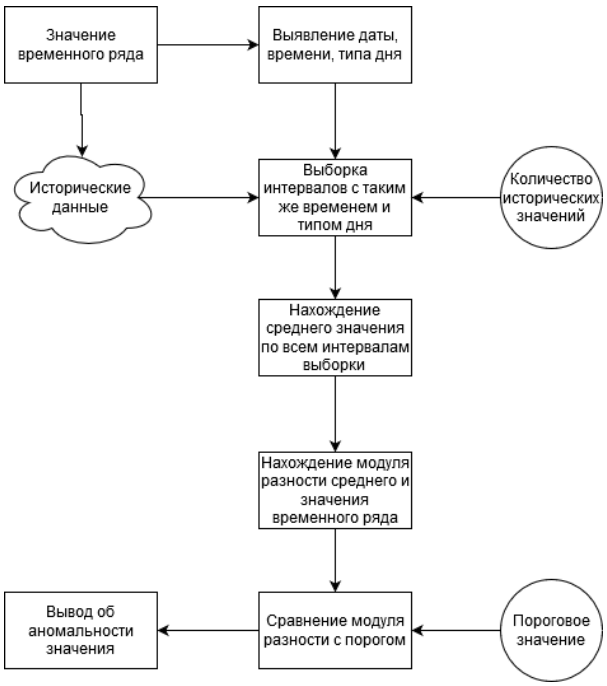

Рис. 3. Схема работы алгоритма

Более формально алгоритм может быть представлен в виде двух шагов.

На первом шаге для проверяемого интервала по дню, времени, типу дня, историческим данным, а также по количеству учитываемых дней рассчитывается среднее по интервалам с тем же временем и типом дня наблюдения. Расчет происходит по формуле (1).

$$hist\_est(day, time, type, delay, hist) = \frac{\sum_{i=delay}^{i=1} [get\_type(i) = type] * hist(i, time)}{\sum_{i=delay}^{i=1} [get\_type(i) = type]} \quad (1)$$

где day - день наблюдения, для которого проверяется аномальность, $time$ – получасовой интервал наблюдения, для которого проверяется аномальность, $type$ – тип дня наблюдения, для которого проверяется аномальность, $delay$ – период рассмотрения исторических данных в днях, а $hist$ - исторические данные, а *get_type* – функция, возвращающая тип дня (будний или выходной).

На втором шаге происходит проверка аномальности наблюдения с использованием значения, найденного на первом шаге. Находится модуль разности

этого значения и проверяемого значения ***value***, а затем он сравнивается с пороговым значением *th*. На основании результата сравнения делается вывод об аномальности проверяемого наблюдения. Данные действия описываются формулой (2)

$$abnormality(value, hist\_est, th) = [|\, hist\_est - val \,| > th] \quad (2)$$

При использовании данного алгоритма выбираются два параметра: *th* – порог, а также *delay* – количество дней для просмотра исторических данных.

Параметр *delay* должен выбираться большим или равным 7, так как иначе в рассматриваемом промежутке исторических данных может не оказаться выходных дней. В процессе нахождения исторического значения в таком случае может появиться деление на 0. Рекомендуемым значением этого параметра является значение 30, которое соответствует длине календарных месяцев. Чем больше данное значение, тем меньше вклад одного наблюдения в среднее значение. При выборе данного параметра, равным 7, среднее значение будет находиться по данным последней недели, новизна в данных, которая появилась неделю назад, будет считаться типичной. Показания алгоритма при таком значении параметра будут быстро перестраиваться под новые условия. Это может быть значимым преимуществом алгоритма, если оцениваются аномалии в районах, где активно вводятся новые транспортные каналы. В районах с уже сформированной транспортной системой следует выбирать данный параметр равным 30 и более. В таком случае отдельные наблюдения, связанные с краткосрочными аномалиями, например, крупными авариями на дорогах или ремонтом дорог, будут оказывать меньшее влияние на среднее, и алгоритм не будет подстраиваться под них, выявляя ложные аномалии.

Также алгоритм можно дополнительно настраивать, изменяя рассматриваемые типы дней. В базовой версии алгоритма используется два типа дней: будни и выходные. Так как в крупных учебных заведениях города суббота также является учебным днём, то выделение суббот в качестве дней особых типов позволяет улучшить распознавание аномалий в районах с крупными учебными заведениями. Во время значимых социальных событий, таких, как проведение салютов и крупных концертов, могут происходить запланированные отклонения в транспортных потоках. В случае, если специалистам необходимо получать данные об аномалиях с учетом таких событий, то следует выделять при рассмотрении таких районов и временных промежутков такие типы дней, как «День концерта», «День салюта» и другие. Данные о днях особых типов следует сохранять отдельно. Это позволит не увеличивать параметр delay до больших значений.

Значение *th* может быть использовано для настройки чувствительности алгоритма. В качестве значения *th* может быть использована константа, заданная специалистом предметной области. В таком случае аномальными будут признаваться интервалы, значения в которых будут отличаться от среднего по подобным интервалам больше, чем на эту константу. Такое понятие аномалии может быть легко воспринято специалистом предметной области. Районы бывают разными по размеру. Для некоторых районов отклонение в 100 человек от среднего в исходящем транспортном потоке не является значительным, в то время как для других районов – это значимая аномалия. В таких случаях можно выражать пороговое значение через *hist_est*. Например, если порог будет выбран равным *0.3*hist_est*, то это будет означать, что аномальными признаются получасовые интервалы, в которых наблюдаемые значения на 30% выше средних для подобных интервалов. Отметим, что при анализе данных метро в ночные часы встречаются ненулевые наблюдения, в то время как среднее оказывается близким к нулю. Такие наблюдения являются редкими и чаще всего связаны с проведением внутренних работ сотрудниками метрополитена. Если урбанистам не нужно распознавание этих работ, как аномалий, то к пороговому значению следует прибавить некоторую константу.

Значения исторических данных *hist_est,* рассчитанные при помощи формулы (1), могут быть использованы для решения других прикладных задач урбанистики. Рассчитав значения *hist_est*_для всех 48 получасовых интервалов буднего и выходного дня по данным месяца, мы можем сохранить эти данные и использовать их для количественной оценки изменения трафика между районами в различные месяцы. При помощи таких усредненных данных о буднях и выходных, например, можно оценить уровень самоизоляции населения по районам города во время пандемии COVID-19.

Алгоритм, описанный выше, обладает преимуществами перед применением существующих алгоритмов поиска аномалий во временных рядах. По сравнению с методами на основе предиктивных моделей, данный метод является более легким в реализации, так как не требует построения предиктивной модели, а также подстраивается под новизну в данных. Для работы алгоритма не требуется разметка аномальных наблюдений для всех районов. Результаты работы алгоритма могут быть легко интерпретированы и объяснены специалисту предметной области. Важным преимуществом по сравнению с применением методов статистического профайлинга и методов, основанных на кластеризации, является возможность нахождения аномальных значений, которые являются аномальными только в контексте даты и времени. Предложенный метод профайлинга верно сигнализирует о таких аномалиях, так как он применяется на выборке данных с такими же метками времени и типа дня. Стоит отметить, что применение методов на основе кластеризации на таких же выборках также может дать успешный результат. Решение на основе профайлинга было выбрано в виду

более широких возможностей в интерпретации и объяснении полученных результатов. Преимущество методов на основе кластеризации по нахождению аномалий со значениями, близкими к среднему, не является значимым в виду специфики данных.

Применение алгоритма и анализ полученных результатов

Для проверки корректности предложенного алгоритма выявления аномалий был проведен вычислительный эксперимент. В ходе данного эксперимента предложенному алгоритму выявления аномалий подавались на вход данные об общих транспортных потоках из всех районов Москвы, а также транспортных потоках метро за май 2017-го года. Для оценки качества выявления аномалий проверялась реакция алгоритма на известные аномалии, такие как праздничные события 1-го мая 2017-го года, акция «Бессмертный полк», которая проходила на территории районов Беговой и Тверской 9-го мая 2017-го года, а также праздничные салюты. Так как в большинстве районов транспортные потоки довольно стабильны, то качество выявления аномалий также проверялось визуально по графикам. Для более подробной оценки качества работы алгоритма, а также его настройки необходимо привлечение специалистов предметной области.

После проведения расчетов были проанализированы результаты. Ниже представлен результат нахождения аномалий в общем транспортном потоке из района Беговой в район Тверской двумя методами статистического профайлинга: предложенным алгоритмом и алгоритмом на основе расчета двух статистик: 95%-перцентиля и 5%-перцентиля.

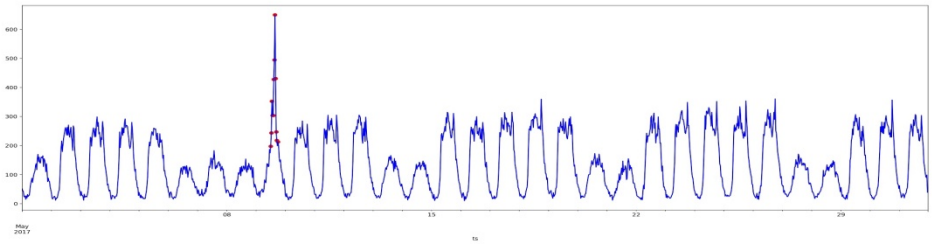

Рис. 4. Выявление аномалий предложенным алгоритмом

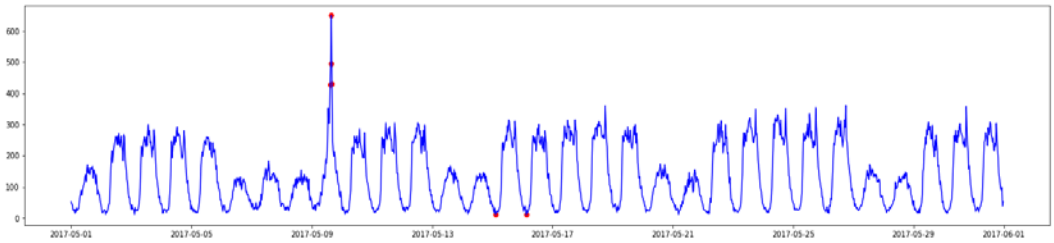

Рис. 5. Выявление аномалий алгоритмом, использующим перцентильные значения

На Рис. 4 и Рис. 5 по горизонтальной оси отмечены получасовые интервалы мая 2017-го года. По вертикальной оси отмечено количество человек, передвигавшихся из района Беговой в район Тверской. Красными точками на графиках отмечены аномалии, найденные при помощи предложенного алгоритма (Рис. 4), а также найденные при помощи метода на основе сравнения с 5%-перцентилем и 95%-перцентилем (Рис. 5).

Предложенный алгоритм лучше справился с поставленной задачей. Все пиковые значения, соответствующие известному аномальному событию выделены, начало сигналов об аномальности наблюдений соответствует дате начала мероприятия, в то время, как при применении профайлинга на основе перцентильных значений алгоритма сигналы подаются на пике аномалии. Также предложенный алгоритм не выявляет ложных аномалий в низких значениях ночного трафика.

Рассмотрим результаты функционирования предложенного алгоритма для других районов. Представлен результат выявления аномалий в потоке из района Марьино в Тверской район (Рис. 6)

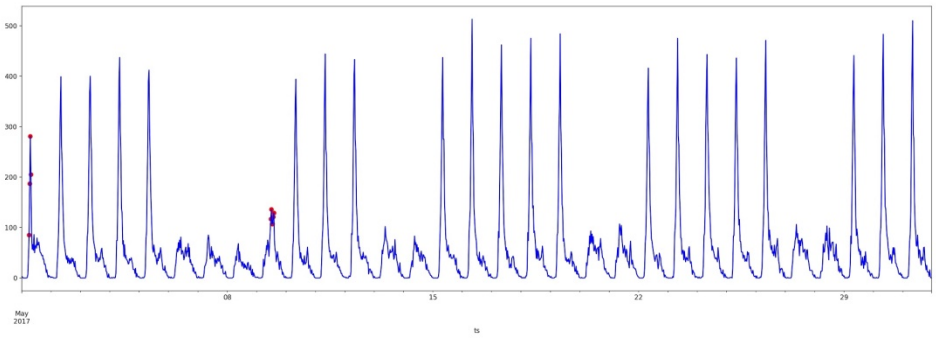

Рис. 6. Выявление аномалий в транспортном поток из района Марьино в район Тверской

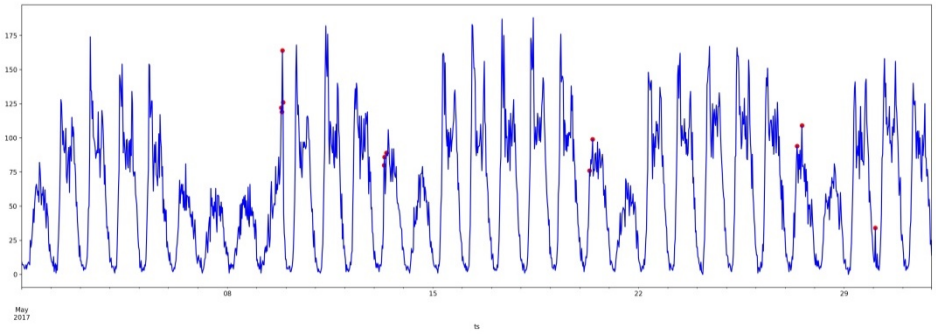

Рис. 7. Выявление аномалий в транспортном потоке из района Академический в район Раменки

На Рис.6 видно, что выделены аномалии, соответствующие празднованиям на 1-ое и 9-ое мая. Ложных срабатываний не наблюдается. Важно отметить, что значения, признанные аномальными, являются таковыми только в контексте даты и времени. Например, первое аномальное значение не является аномальным для выходного дня, но является аномальным конкретно для такого временного промежутка. С выявлением аномалий такого типа не справляются алгоритмы, которые не учитывают связь данных со временем, например методы на основе кластеризации.

Одним из событий, притягивающих множество людей, является салют 9-го мая. Одной из лучших смотровых площадок для просмотра салюта является смотровая площадка, расположенная на Воробьевых горах в районе Раменки. Во временных рядах, соответствующих перемещениям из соседних районов в район Раменки также наблюдаются аномалии. В качестве примера приведено выявление аномалий в потоке из района Академический в район Раменки (Рис. 7).

На данном графике по горизонтальной оси отмечены получасовые интервалы мая 2017-го года. По вертикальной оси отмечено количество человек, передвигавшихся из района Академический в район Раменки. Красными точками на графиках отмечены аномалии, найденные при помощи предложенного алгоритма. Аномалии, соответствующие салюту, выявлены верно. Также выявлено несколько аномалий, соответствующих высокому уровню перемещений между районами по субботам. Возможно, таковые уровни связаны с тем, что в районе Раменки расположен Московский Государственный Университет, в котором суббота является учебным днём. Для того, чтобы избежать реакции на такие явления, которые не являются аномальными, можно либо изменить настройки порогового значения, либо ввести новый тип дня «суббота». Такой тип дня может оказаться актуальным для анализа районов, где расположены крупные учебные заведения, так как в них суббота является учебным днём. Модификация алгоритма в таком случае не понадобится, возможно, придется изменить настройки чувствительности алгоритма.

В целом алгоритм продемонстрировал свою корректность. Все выявленные аномалии могут быть объяснены, но для полного понимания необходимо привлечение специалистов предметной области. Для выявления аномалий во всех районах использовались одни и те же настройки, длина рассматриваемого периода составляла 30 дней, порог задавался как превышение значения, посчитанного по историческим данным на 30% и увеличенного на 5. Последнее увеличение значения, посчитанного по историческим данным, было необходимо, чтобы не выявлялись аномалии по передвижению на метро ночью (средние для таких интервалов близки к 0).

Для оценки качества алгоритмов выявления аномалий необходимо получить разметку аномалий от специалиста в области урбанистики, в которой 0 соответствует типичным значениям, а 1 аномальным, для нескольких временных рядов. Далее с использованием этой разметки в качестве истинных данных могут быть рассчитаны точность, полнота, а

также F1-мера. По этим метрикам может быть оценено качество работы алгоритма с заданными параметрами, проведен выбор лучших параметров, а также проведено сравнение с другими алгоритмами выявления аномалий.

ЗАКЛЮЧЕНИЕ

В статье выявлены возможные способы использования агрегированных данных сотовых операторов для решения задач цифровой урбанистики. Был предложен метод обнаружения аномалий в данных транспортных потоков, использующий подобные данные. Предложенный метод был реализован, а его корректность была проверена путем проведения вычислительного эксперимента.

Библиография

[1] Mark Bulygin and Dmitry Namiot “Anomaly Detection Method For Aggregated Cellular Operator Data” in press
[2] Report of the international agency “We are social” Web: https://digitalreport.wearesocial.com/ Retrieved: Nov, 2020
[3] Huang, Haosheng, et al. "Location based services: ongoing evolution and research agenda." Journal of Location Based Services 12.2 (2018): 63-93.
[4] Calabrese, Francesco, et al. "Real-time urban monitoring using cell phones: A case study in Rome." IEEE Transactions on Intelligent Transportation Systems 12.1 (2010): 141-151.
[5] Garroppo, Rosario & Niccolini, Saverio. (2017). Anomaly detection mechanisms to find social events using cellular traffic data. Computer Communications. 116.10.1016 / j.comcom.2017.12.009.
[6] Namiot, Dmitry, and Manfred Sneps-Sneppe. "A Survey of Smart Cards Data Mining." Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017) Moscow, Russia. 2017
[7] Namiot, D. E., O. N. Pokusaev, and V. S. Lazutkina. "On models of passenger flow for urban railways." International Journal of Open Information Technologies 6.3 (2018)
[8] Namiot, Dmitry, Oleg Pokusaev, and Vasily Kupriyanovsky. "On railway stations statistics in Smart Cities." International Journal of Open Information Technologies 7.4 (2019): 19-24.
[9] Nekraplennaya, M.N., and D.E. Namiot. "Analysis of metro correspondence matrices." International Journal of Open Information Technologies 7.7 (2019).
[10] Shipmon, Dominique T., et al. "Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data." arXiv preprint arXiv:1708.03665 (2017).

On the possibilities of using the data of cellular operators to solve the problems of digital urbanism

Mark Bulygin, Dmitry Namiot

Abstract — **Currently, the widespread penetration of mobile devices makes it possible to use the data collected during their operation to analyze traffic flows. Associating mobile phones with their owners, it is possible to judge by the movements of mobile devices, for example, the structure of traffic flows in the city. These tasks have nothing to do with some kind of tracking of mobile subscribers, only anonymized data is used here. Moreover, due to the specifics of the tasks of analyzing traffic flows, individual movements are not fundamentally interesting here, but it is the general (aggregated) data on movements between the selected objects (sections) that are needed. This article presents the main directions of using aggregated data of cellular operators, such as searching for changes in traffic flows corresponding to important social events, measuring these changes, searching for novelty in traffic flows, clustering areas and connections between them, as well as assessing the distribution of people across city districts. The article describes an approach to identifying anomalies in traffic data (anomalies in the time domain are investigated), which is based on the analysis of aggregated data of cellular operators. The paper presents a computational experiment that demonstrates the correctness of the proposed method.**

Keywords - digital urban studies, data analysis of cellular operators, analysis of traffic flows, search for anomalies in data

REFERENCES

[1] Mark Bulygin and Dmitry Namiot "Anomaly Detection Method For Aggregated Cellular Operator Data" in press
[2] Report of the international agency "We are social" Web: https://digitalreport.wearesocial.com/ Retrieved: Nov, 2020
[3] Huang, Haosheng, et al. "Location based services: ongoing evolution and research agenda." Journal of Location Based Services 12.2 (2018): 63-93.
[4] Calabrese, Francesco, et al. "Real-time urban monitoring using cell phones: A case study in Rome." IEEE Transactions on Intelligent Transportation Systems 12.1 (2010): 141-151.
[5] Garroppo, Rosario & Niccolini, Saverio. (2017). Anomaly detection mechanisms to find social events using cellular traffic data. Computer Communications. 116.10.1016 / j.comcom.2017.12.009.
[6] Namiot, Dmitry, and Manfred Sneps-Sneppe. "A Survey of Smart Cards Data Mining." Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017) Moscow, Russia. 2017
[7] Namiot, D. E., O. N. Pokusaev, and V. S. Lazutkina. "On models of passenger flow for urban railways." International Journal of Open Information Technologies 6.3 (2018)
[8] Namiot, Dmitry, Oleg Pokusaev, and Vasily Kupriyanovsky. "On railway stations statistics in Smart Cities." International Journal of Open Information Technologies 7.4 (2019): 19-24.
[9] Nekraplennaya, M.N., and D.E. Namiot. "Analysis of metro correspondence matrices." International Journal of Open Information Technologies 7.7 (2019).
[10] Shipmon, Dominique T., et al. "Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data." arXiv preprint arXiv:1708.03665 (2017).