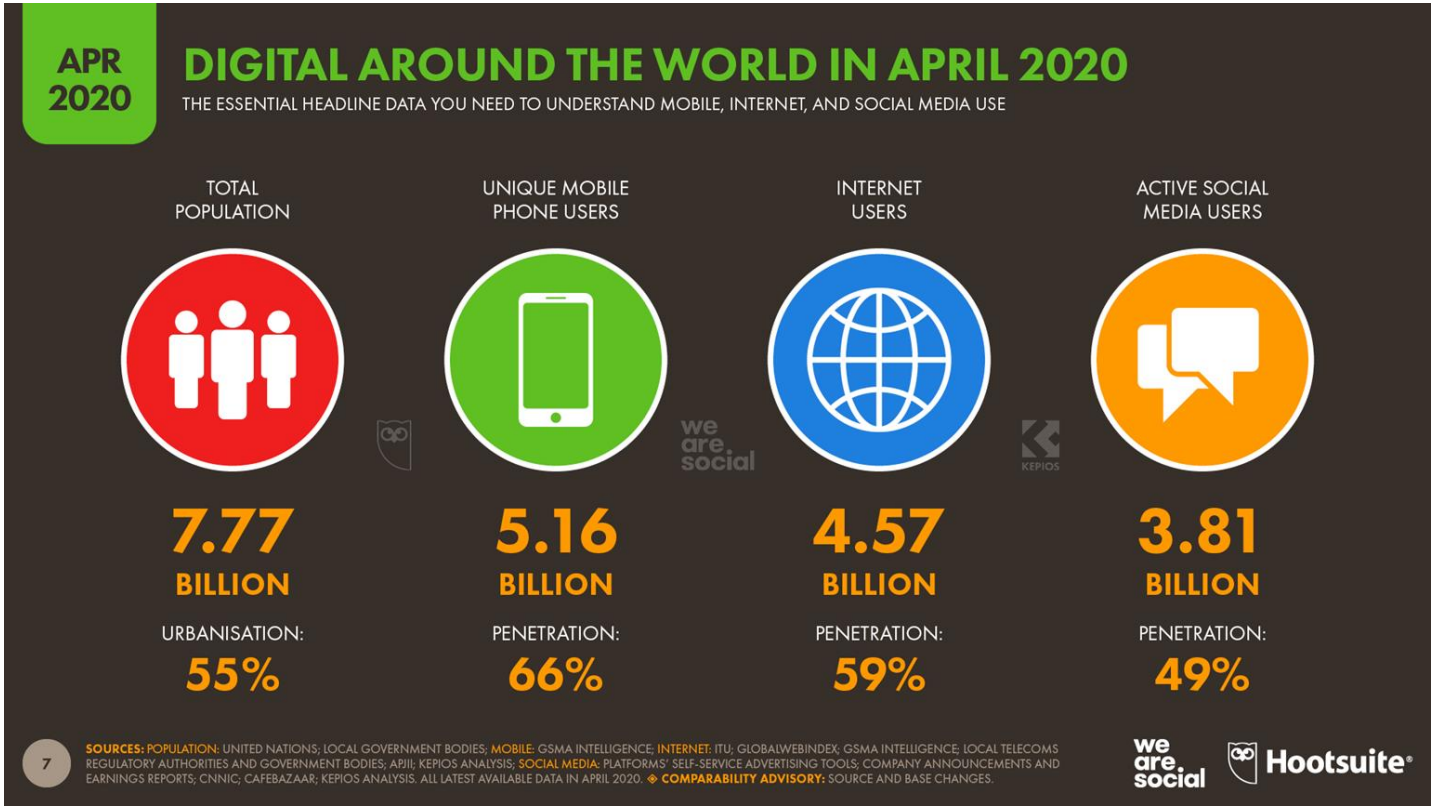
Анализ данных сотовых операторов в цифровой урбанистике

Выполнил:
студент группы М-210
Булыгин Марк
Валерьевич
Научный руководитель:
к.ф.-м.н., с.н.с лаб. ОИТ
Намиот Дмитрий
Евгеньевич

2020

Введение


APR 2020
DIGITAL AROUND THE WORLD IN APRIL 2020
THE ESSENTIAL HEADLINE DATA YOU NEED TO UNDERSTAND MOBILE, INTERNET, AND SOCIAL MEDIA USE
TOTAL POPULATION
UNIQUE MOBILE PHONE USERS
INTERNET USERS
ACTIVE SOCIAL MEDIA USERS
we are social
KEPIOS
7.77 BILLION
5.16 BILLION
4.57 BILLION
3.81 BILLION
URBANISATION: 55%
PENETRATION: 66%
PENETRATION: 59%
PENETRATION: 49%
7
SOURCES: POPULATION: UNITED NATIONS; LOCAL GOVERNMENT BODIES; MOBILE: GSMA INTELLIGENCE; INTERNET: ITU; GLOBALWEBINDEX; GSMA INTELLIGENCE; LOCAL TELECOMS REGULATORY AUTHORITIES AND GOVERNMENT BODIES; APJII; KEPIOS ANALYSIS; SOCIAL MEDIA: PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; COMPANY ANNOUNCEMENTS AND EARNINGS REPORTS; CNNIC; CAFEBAZAAR; KEPIOS ANALYSIS. ALL LATEST AVAILABLE DATA IN APRIL 2020. ◈ COMPARABILITY ADVISORY: SOURCE AND BASE CHANGES.
we are social
Hootsuite

Агрегированные данные сотовых операторов

Для Москвы и Московской области доступны следующие данные, агрегированные по получасовым интервалам:

1. Общее количество человек, начавших поездку в заданный получасовой интервал из района A в район B
2. Количество человек, начавших поездку в заданный получасовой интервал из района A в район B, при этом во время поездки использовалось метро
3. Количество человек, начавших поездку в заданный получасовой интервал из района A (из дома) в район B (на работу)
4. Количество человек, начавших поездку в заданный получасовой интервал из района A (с работы) в район B (домой)
5. Количество человек, находящихся на территории района не менее 60 минут и не совершавших поездок в заданный получасовой интервал

Зафиксировав район отправления и район прибытия, можно получить соответствующие временные ряды для анализа

Цель и задачи

В настоящее время **актуальной** является проблема анализа данных сотовых операторов. **Целью** работы является предложение новых алгоритмов кластеризации районов и связей между ними на основе агрегированных данных сотовых операторов, а также выявления аномалий в них.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1) Исследование существующих методов анализа данных сотовых операторов
2) Исследование существующих методов кластеризации и выявления аномалий
3) Разработка алгоритмов кластеризации агрегированных данных сотовых операторов
4) Проведение вычислительных экспериментов, иллюстрирующих предложенные методы

Существующие подходы

Решаемая задача	Используемые методы	Исследование
Шесть различных задач городского планирования	Разработанный DFL-алгоритм, методы сравнения с пороговыми значениями, методы визуализации и картирования	« Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome» Ф.Калабрезе и др.
Предсказание уровня загруженности по областям города	Классификация на основе случайных лесов	«Data-Driven Prediction System of Dynamic People-Flow in Large Urban Network Using Cellular Probe Data» С.Чен и др.
Поиск социальных событий при помощи анализа мобильных данных	Вейвлет-анализ, TA и SA-алгоритмы	««Anomaly detection mechanisms to find social events using cellular traffic data» Д.Росарио и др.
Анализ индивидуальных и коллективных паттернов поведения	Статистические методы, анализ математических ожиданий и стандартных отклонений.	«Uncovering individual and collective human dynamics from mobile phone records» М.С. Гонсалез и др.
Картирование населения	Диаграммы Вороного, регрессионные модели	«Dynamic population mapping using mobile phone data» П.Девиль и др
Поиск мест для размещения зеленых микрозон	2SFCA-алгоритм	«Evaluation and Planning of Urban Green Space Distribution Based on Mobile Phone Data and Two-Step Floating Catchment Area Method» X.By
Предсказание транспортного потока пассажиров	Сверточные нейронные сети	«Passenger Demand Prediction with Cellular Footprints» Дж.Чу и др.

Математические методы

Выявление аномалий:

1) Методы, не учитывающие связь данных со временем
2) Методы на основе моделирования временных рядов (ARIMA)
3) Методы на основе машинного обучения

Кластеризация:

1) KMeans
2) Иерархическая кластеризация
3) Методы основанные на плотности

Разработка алгоритмов. Выявление аномалий

Группа методов	Преимущества	Недостатки	Примечания
Методы, не учитывающие связь данных со временем	Вычислительная легкость, хорошая интерпретируемость, легкая адаптация к новизне в данных.	Не учитывается информация о типе дня и времени. В случае перцентильного подхода какая-то доля данных будет признаваться аномальной, даже если данные аномальными не являются.	Данные модели не учитывают тип дня (выходной/будний) и время наблюдения. В случае появления новизны в данных пороговые значения могут быть легко пересчитаны. Могут быть применены как к необработанным данным, так и к разностям.
На основе ARIMA	Возможность предугадывания аномалий путем построения прогноза	Высокая сложность получения хорошего алгоритма, в случае применения SARIMAX может потребоваться ручная разметка.	Необходимо приводить ряды к стационарному виду, для каждой связи нужна своя модель. Наблюдаются ложные сигналы в случае, если за неделю (период сезонности) была аномалия (при использовании SARIMA)
На основе алгоритмов машинного обучения	Возможность предугадать аномалии заранее. Возможность учета дополнительных факторов, например, погодных условий, типа дня (выходной/праздничный)	Требуется разметка данных специалистом для обучения. Требуются различные модели для разных типов районов (жилой и рабочий). Трудность адаптации к новизне	Специалисту потребуется отметить аномалии для обучения, это трудоемкий процесс. В случае появления новизны в данных разметку и обучение нужно повторить. Если модель учитывает данные прошлой недели, а там были аномалии, возможны ложные срабатывания

Разработка алгоритмов. Выявление аномалий

Собственный метод:

$$anomality(value, day, time, type, delay, hist, th) = [|hist\_est(day, time, type, delay, hist) - val| > th]$$

- value - значение временного ряда. для которого проверяется аномальность
- day - день наблюдения, для которого проверяется аномальность
- time - время наблюдения (получасовой интервал)
- type - тип дня наблюдения (выходной, будний)
- delay - период рассмотрения исторических данных в днях
- hist[a, b] - исторические данные для данного района в день a, время b.
- th - пороговое значение
- hist_est - функция, возвращающая значения на основе исторических данных

Разработка алгоритмов. Выявление аномалий

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

- *get_type(i)* - функция, возвращающая тип дня *i* (будний, выходной)

Рекомендации по использованию алгоритма:

1) delay должен выбираться больше или равным 7, чтобы на рассматриваемом промежутке гарантированно присутствовали дни всех типов
2) th - должен задаваться специалистом предметной области, он показывает, какое отклонение от исторических данных следует считать аномальным.
3) Может быть использован и для анализа данных по завершению временного периода. В алгоритм нужно передать в таком случае значение day, соответствующее концу периода, а delay - его длине

Разработка алгоритмов. Выявление аномалий

Данный алгоритм обладает следующими преимуществами:

1) Учет типа дня и времени при выявлении аномальности
2) Автоматическая адаптация при возникновении новизны в данных
3) Используя функцию *hist_est,* возможно реализовать подсчет средних для всех временных интервалов для будних и рабочих дней. Эти данные могут быть использованы для последующего анализа.
4) Не требует разметки аномалий для всех районов. Специалисту важно задать лишь порог (например, 30% от медианного значения для данного интервала)

Недостатки:

1) Нет возможности предугадывания аномалий (требуется отдельная прогнозная модель)
2) Нет возможности учета дополнительных данных

Разработка алгоритмов. Кластеризация

В рамках данной работы рассматривается два типа кластеризации:

1) Кластеризация районов
2) Кластеризация связей между районами

Кластеризовать необработанные данные не получится:

1) Трудность объяснения получившихся кластеров
2) Слишком большой объем данных

Основная задача:

Выбор пространства, характеризующего районы/связи, в котором можно провести кластеризацию

Разработка алгоритмов. Кластеризация районов.

1. По количеству людей, пребывающих в районе
    a. Среднее количество людей, пребывающих в районе в рабочее время (с 10:00 до 18:00 по будням) (working_rate)
    b. Среднее количество людей, пребывающих в районе в ночное время (с 23:00 до 7:00) (night_rate)

Для того, чтобы исключить влияние размера района была использована минимаксная нормализация.

$$working\_rate = \frac{\frac{\sum_{i=1}^{n} static(i) * [i \in work\_time]}{\sum_{i=1}^{n} [i \in work\_time]} - min(static)}{max(static) - min(static)}$$

$$night\_rate = \frac{\frac{\sum_{i=1}^{n} static(i) * [i \in night\_time]}{\sum_{i=1}^{n} [i \in night\_time]} - min(static)}{max(static) - min(static)}$$

где static(i) - количество людей в районе в i-ый временной интервал, n - общее количество временных интервалов, work_time - номера временных интервалов рабочего времени, night_time - ночного.

Разработка алгоритмов. Кластеризация районов.

2. По доле людей уезжающих на работу в утренний час пик и по доле поездок на метро

a. Средняя доля людей, уезжающих в утренние часы пик на работу (с 7:00 до 10:00 по будням) (working_percent)
b. Средняя доля людей, использовавших метро в этот же временной интервал (metro_percent)

$$working\_percent = \frac{\sum_{i=1}^{n} \frac{work(i)}{customers(i)} * [i \in morning\_time]}{\sum_{i=1}^{n} [i \in morning\_time]}$$

$$metro\_percent = \frac{\sum_{i=1}^{n} \frac{metro(i)}{customers(i)} * [i \in morning\_time]}{\sum_{i=1}^{n} [i \in morning\_time]}$$

где customers(i) - количество людей. выехавших из района в i-ый временной интервал, metro(i) - количество людей, выехавших из района с использованием метро, work(i) - количество людей, выехавших в i-ый временной интервал на работу, n - общее количество временных интервалов, morning_time - индексы временных интервалов, соответствующим утреннему часу пик

Разработка алгоритмов. Кластеризация связей

- Много различных связей между районами Москвы. Трудно выделить кластеры
- Стоит зафиксировать один район интереса и кластеризовать лишь его связи с другими районами.

Каждая такая связь может быть описана точкой в одном из следующих признаковых пространств:

1) Количество людей, переместившихся из зафиксированного района в связанный район (в долях от максимальной) и доля из них, перемещавшихся на работу (2-мерное)
2) Количество перемещений из зафиксированного района в связанный район (1-мерное)
3) Количество перемещений в связанный район с использованием метро (1-мерное)

Разработка алгоритмов. Кластеризация

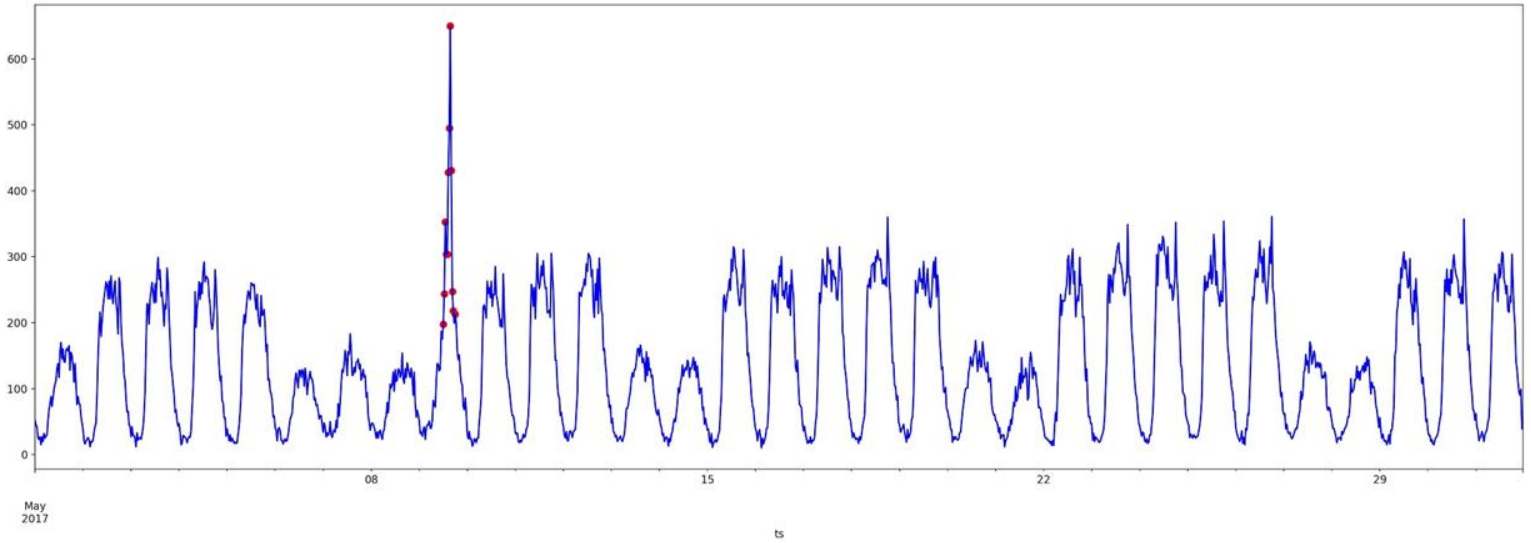
Для кластеризации в двумерных пространствах выбран метод K-Means

Для проверки наличия кластерной структуры используется статистика Хопкинса, а для оценки такой кластеризации используется коэффициент силуэта.

Для кластеризации в одномерных пространствах выбран иерархический метод кластеризации. Он позволяет строить наглядные дендрограммы и удобен в интерпретации.

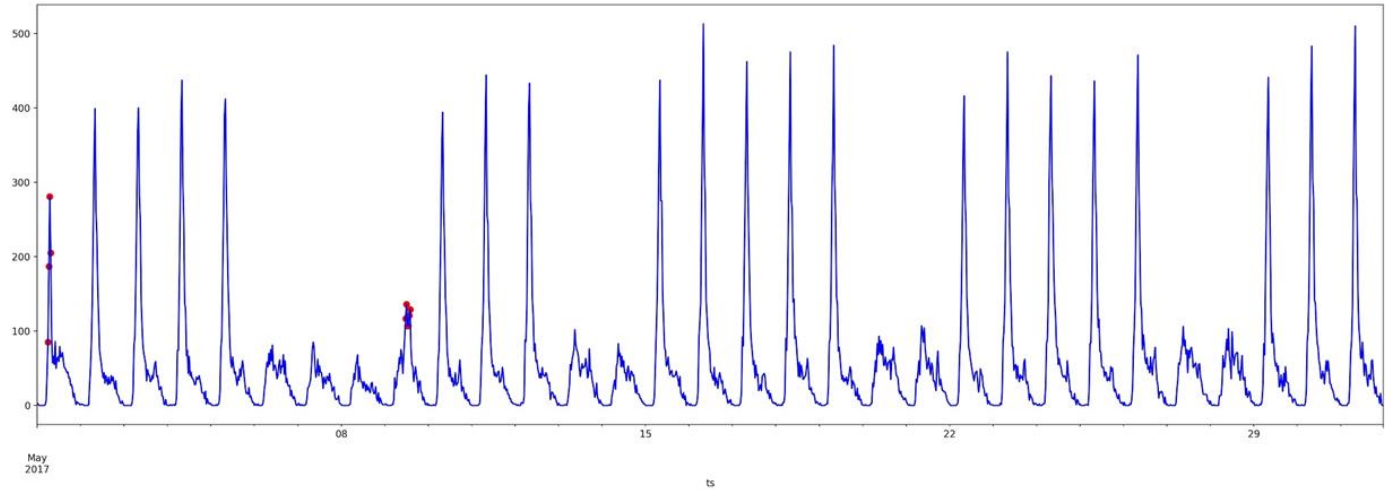
Вычислительный эксперимент. Выявление аномалий.

По оси X данного графика отложены получасовые интервалы с первого по 31 мая 2017-го года включительно. По оси Y отложено количество людей, перемещавшихся из района Беговой в район Тверской. Аномалии, выделенные предложенным алгоритмом, выделены красными точками. Эти аномалии соответствуют акции “Бессмертный полк”, проходившей 9-го мая 2017-го года. При визуальной диагностике временного ряда других аномалий не выявлено.



Вычислительный эксперимент. Выявление аномалий.

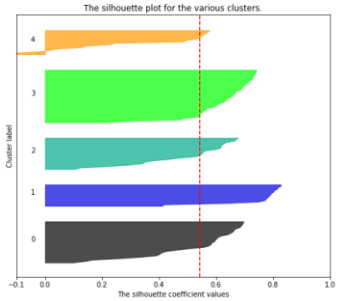
По оси X данного графика отложены получасовые интервалы с первого по 31 мая 2017-го года включительно. По оси Y отложено количество людей, перемещавшихся из района Марьино в район Тверской. Аномалии, выделенные предложенным алгоритмом, выделены красными точками. Эти аномалии соответствуют первомайским гуляниям и акции “Бессмертный полк”. При визуальной диагностике временного ряда других аномалий не выявлено.

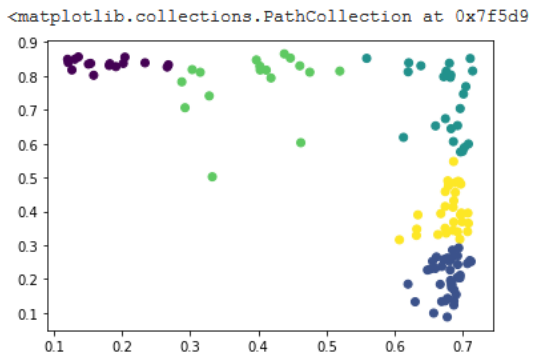


Вычислительный эксперимент. Выявление аномалий.

По результатам вычислительного эксперимента предложенный алгоритм работает, выявляемые аномалии связаны с социальными событиями. Пропусков, в отличие от пороговых алгоритмов, или ложных срабатываний (по визуальному анализу графика) не обнаружено. Для более точной оценки качества и настройки предложенного алгоритма необходимо продолжить работу с экспертами предметной области.

Вычислительный эксперимент. Кластеризация





По оси X отложен night_rate, по оси Y working_rate

Выделяется 5 кластеров:

1) Рабочий кластер, характерен night_rate близкий к 0.2, working_rate близкий к 0.85, типичные представители - Арбат, Тверской
2) Кластер районов, которые являются более рабочими, чем жилыми. Для них характерен working_rate близкий к 0.8, а также night_rate близкий к 0.45. Типичные представители: Капотня, Войковский, Южнопортовый и.т.д.
3) Кластер районов, являющихся в равной степени жилыми и рабочими. Оба показателя близки к 0.7. Типичные представители: Матушкино, Академический, Раменки
4) Более жилые районы, чем рабочие. В них показатель night_rate близок к 0.65, а показатель working_rate - к 0.4. Типичные представители: Ярославский, Рязанский, Кунцево
5) Жилые районы: в них показатель working_rate менее 0.3. К таким районам относятся: Кузьминки, Выхино-Жулебино, Южное Бутово, Марьино и.т.д.

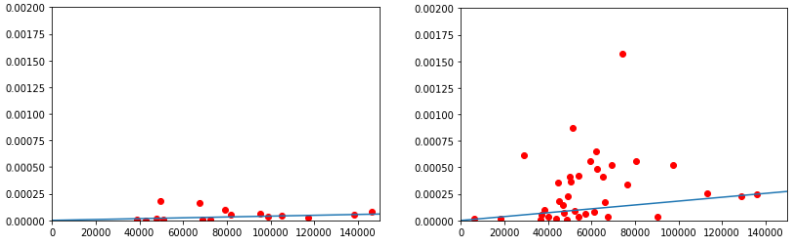
Вычислительный эксперимент. Кластеризация

Кластеры, полученные в результате такой кластеризации, могут использоваться в прикладных задачах.

В качестве признака. Была обучена RandomForest-модель для предсказания стоимости квартир в Москве. При добавлении признака, обозначающего кластер района качество модели увеличивалось (точечная оценка размера эффекта: уменьшение MedAE на 279 тыс. рублей, а также уменьше RMSE на 926 тыс. рублей)

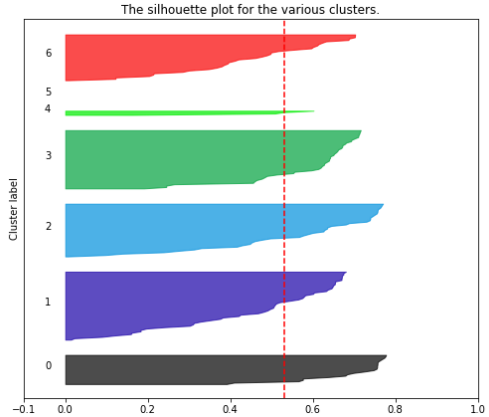
Ошибка/Признаки	default	default+cluster	default+hood	default+cluster+hood
RMSE	2901538.76	2579360.58	2548347.42	2269059.1
MedAE	1298552.52	1169841.73	1105516.41	1012914.33

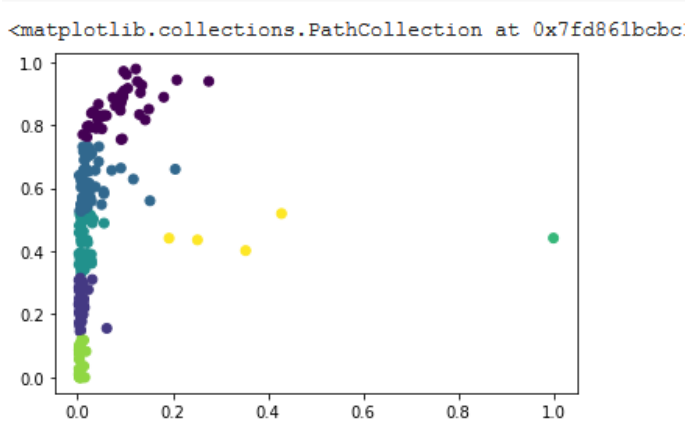
Построение своей модели для каждого кластера. Были взяты данные о площадях парковых зон города Москвы. Для каждого кластера была построена своя RANSAC-модель на основе линейной регрессии, показывающая зависимость площади парковых зон в районе от населения. Интерпретация и анализ такой модели позволит выявить районы с недостатком зеленых зон и обратить на них внимание при благоустройстве



На изображениях по оси X отложено население районов Москвы, а по Y отложена территория парков в этих районах в условных единицах. На левом изображении представлены районы рабочего кластера, а на правом - жилого. Как видно, в жилых районах парковых зон больше (у.е. площади на человека), однако существуют районы с недостатком зелёных зон.

Вычислительный эксперимент. Кластеризация





На слайде представлена кластеризация связей Долгопрудного. Рассматривается кластеризация по двум признакам: каждому району соответствует количество людей, переместившихся в район из Долгопрудного (в долях от максимального), а также доля людей, которые едут на работу в данный район.

В отдельный кластер выделяются Химки. В другой кластер выделяются Мытищи, Федоскинское, Лобня, район Северный - в эти части Москвы и Московской области жители Долгопрудного перемещаются достаточно часто, не только на работу. В остальные районы перемещений довольно мало, они кластеризуются в основном по доле людей, которые ездят в эти районы на работу. В фиолетовом кластере лежат точки, соответствующие связям с районами центра города, такими как Тверской и Арбат. Туда люди ездят в основном на работу.

Заключение

В результате работы было **выполнено исследование** существующих методов анализа агрегированных данных сотовых абонентов, **были разработаны** алгоритмы кластеризации данных сотовых операторов и выявления аномалий в них. Для анализа полученных алгоритмов **был проведен** вычислительный эксперимент, который показал корректность предложенных методов.

Дальнейшее развитие работы может продолжаться в двух направлениях:

1) Совместная работа с экспертами предметной области для более глубокого исследования полученных результатов с точки зрения урбанистики
2) Предложение новых алгоритмов кластеризации и выявления аномалий в данных