Применение неевклидовых методов кластеризации для определения популярных маршрутов абонентов

Лобанов Артём, 371 группа
Научный руководитель: Невоструев К.Н.

Введение

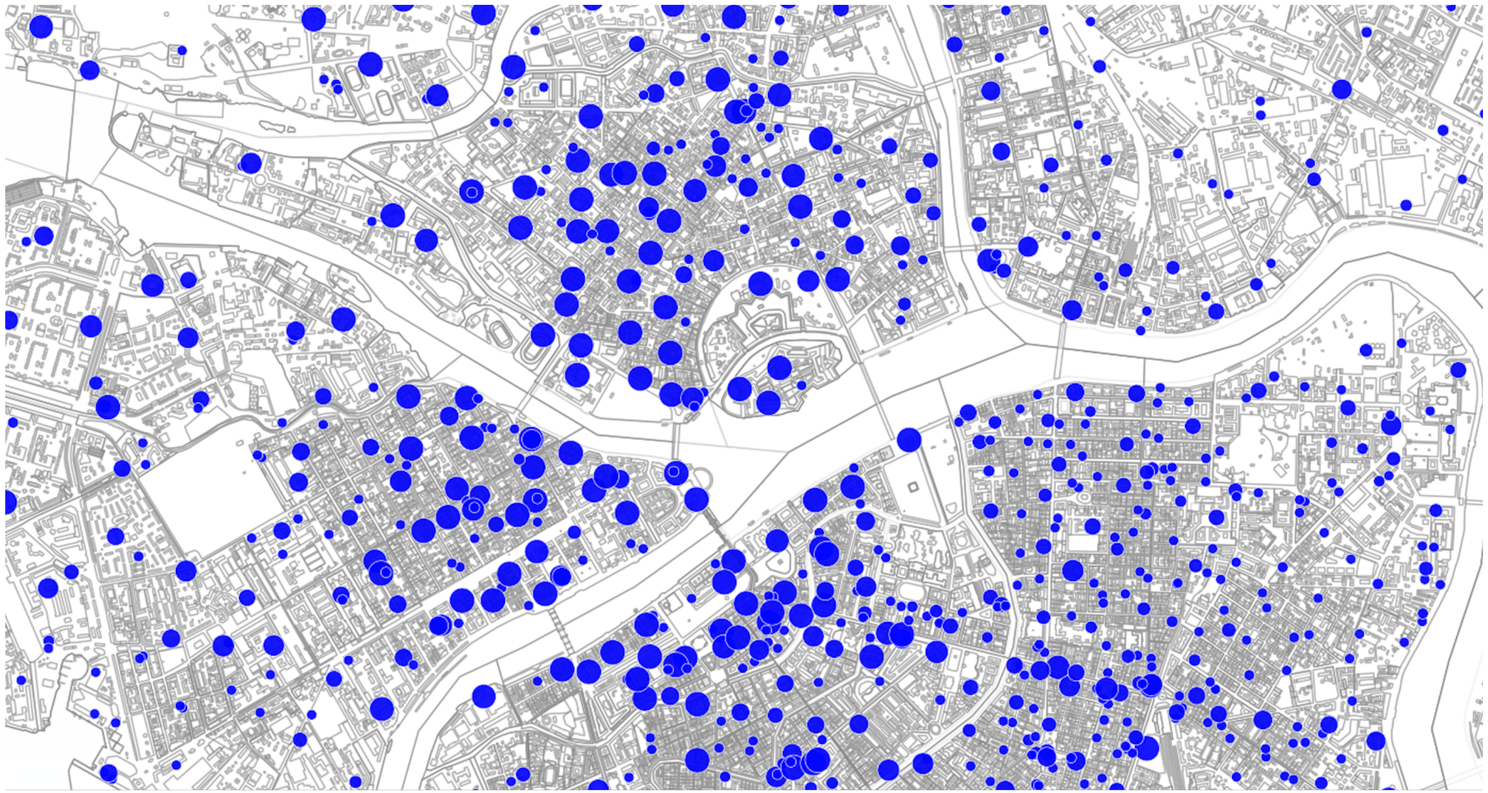
- Мобильным операторам полезно знать популярные маршруты абонентов
- Например, станет ясно, где выгоднее размещать рекламу
- Структура множества путей неочевидна

Постановка задачи

- Есть данные о пользовательской активности в различные промежутки времени
- Последовательность активностей абонента — его “маршрут”
- Цель — найти кластеры популярных маршрутов, построить популярные маршруты

Данные

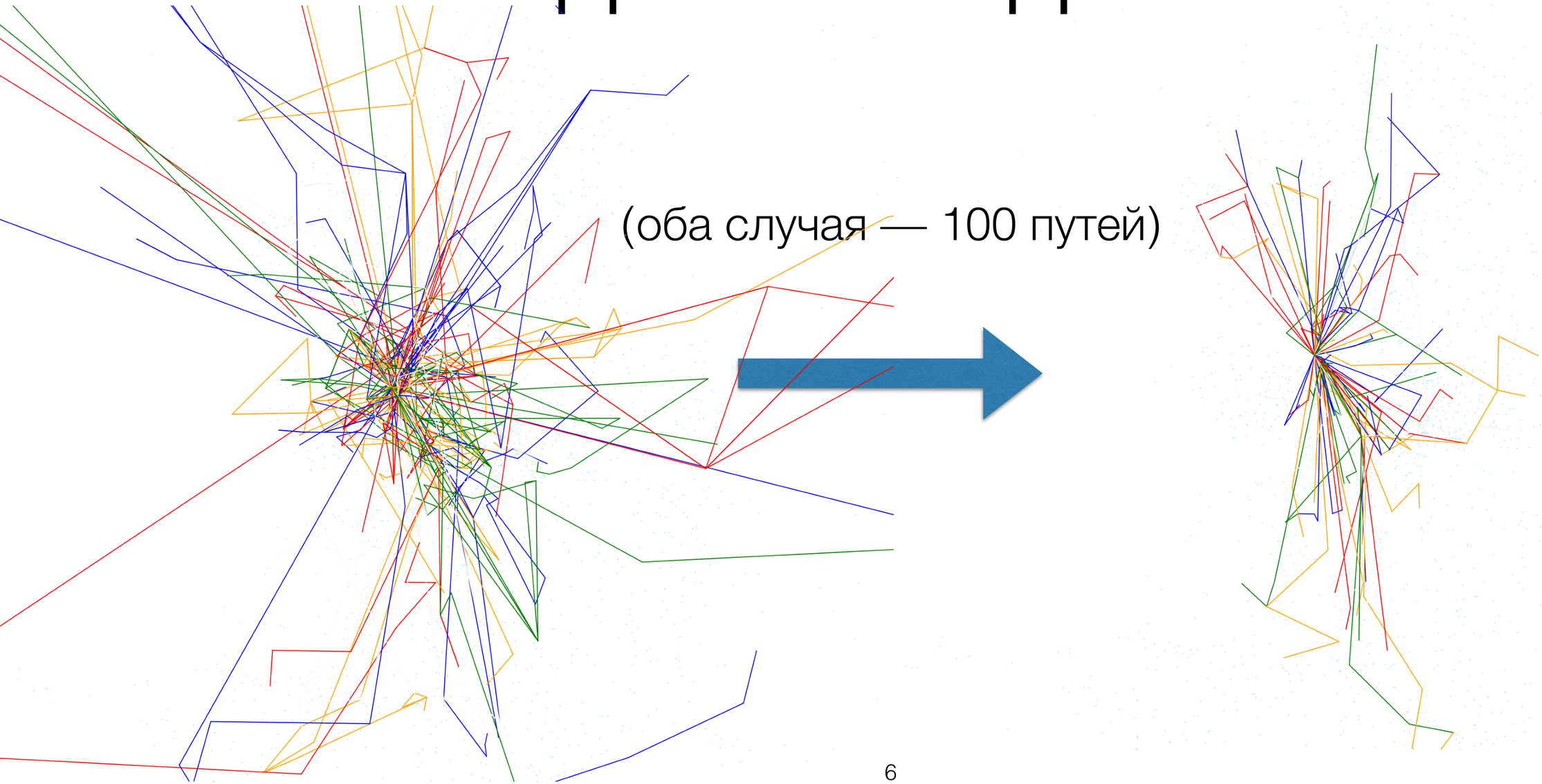
- 2 стадиона
- Координаты базовых станций (≈3500)
- Данные пользовательской активности (≈10 млн. на оба стадиона)



Подготовка данных

- Формирование маршрутов из биллинговых данных
- Представление всех путей “от стадиона”
- Фильтрация/очищение данных
  - По длине максимального отрезка пути
  - Без циклов близко к стадиону
- Единичные “отклонения” от пути игнорируются
- Отбрасываются пути с большим разбросом направлений

Подготовка данных



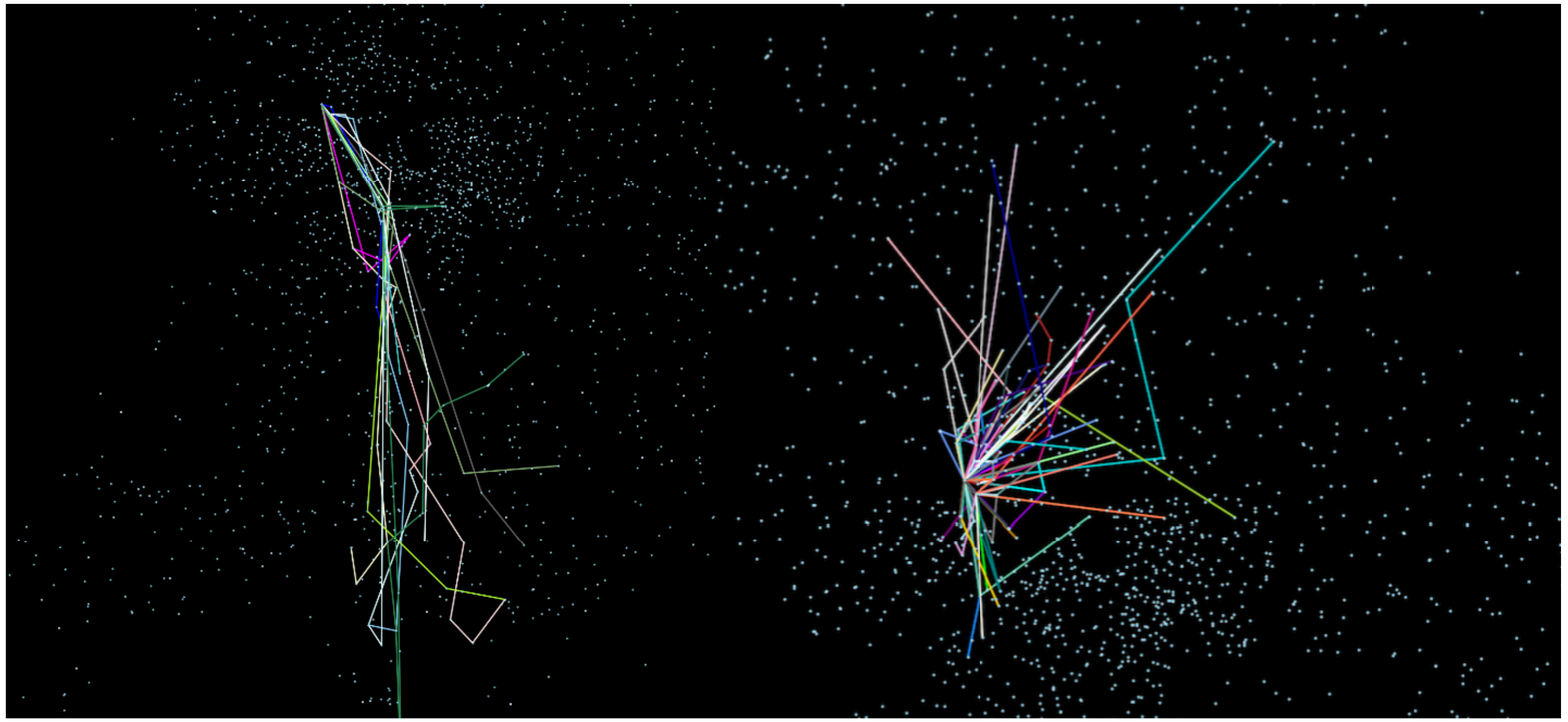
(оба случая — 100 путей)

Агломеративная кластеризация

- Итеративно сливаются 2 самых “похожих” кластера
- Выявлена эффективная мера близости:
  - Количество примерно совпадающих точек/отрезков путей
  - Веса больше на точках/отрезках дальше от стадионов

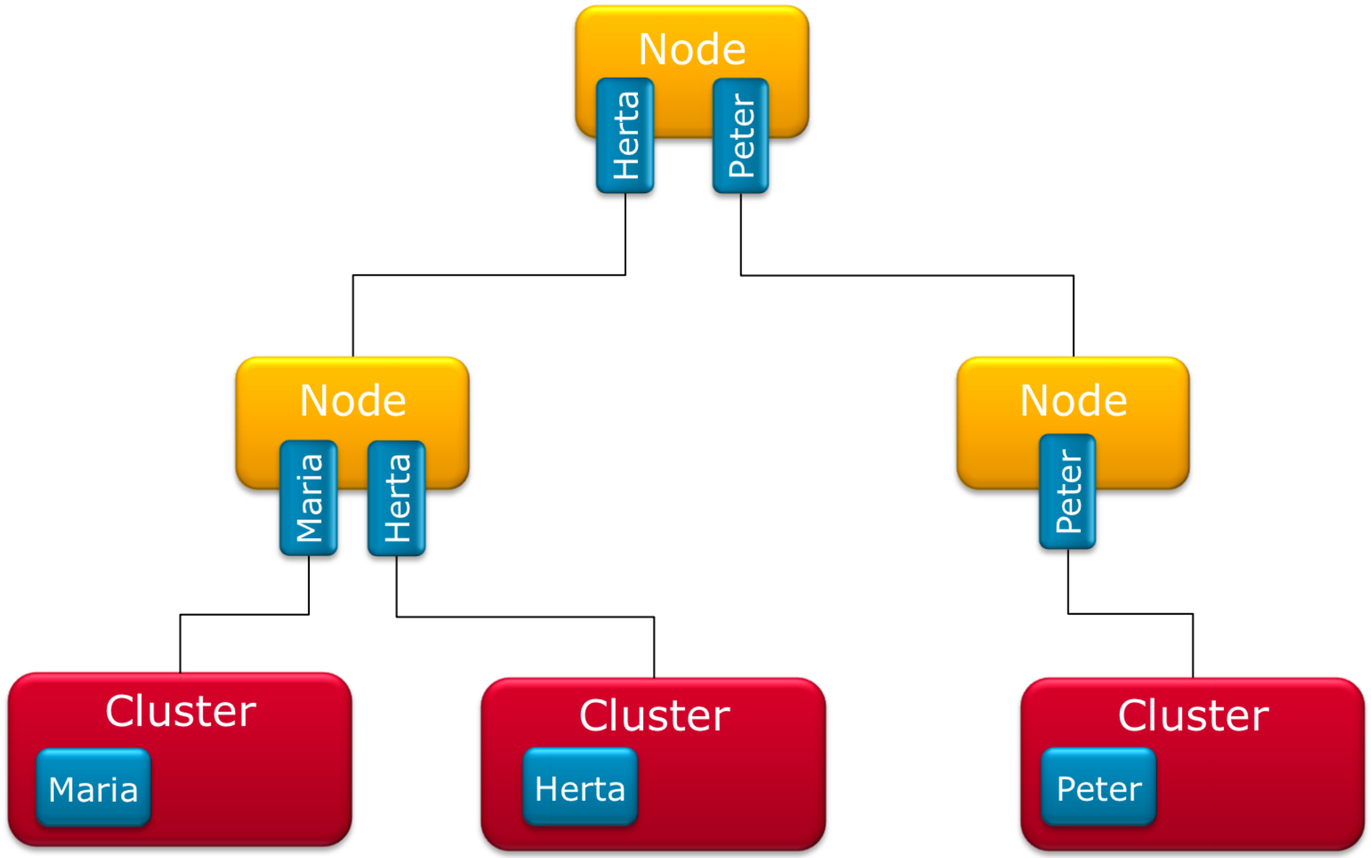
Агломеративная кластеризация

- Недостатки:
  - Некоторые кластеры содержат мало путей
  - Мало похожие пути собираются в один кластер — в нем слишком много путей (правый кластер)



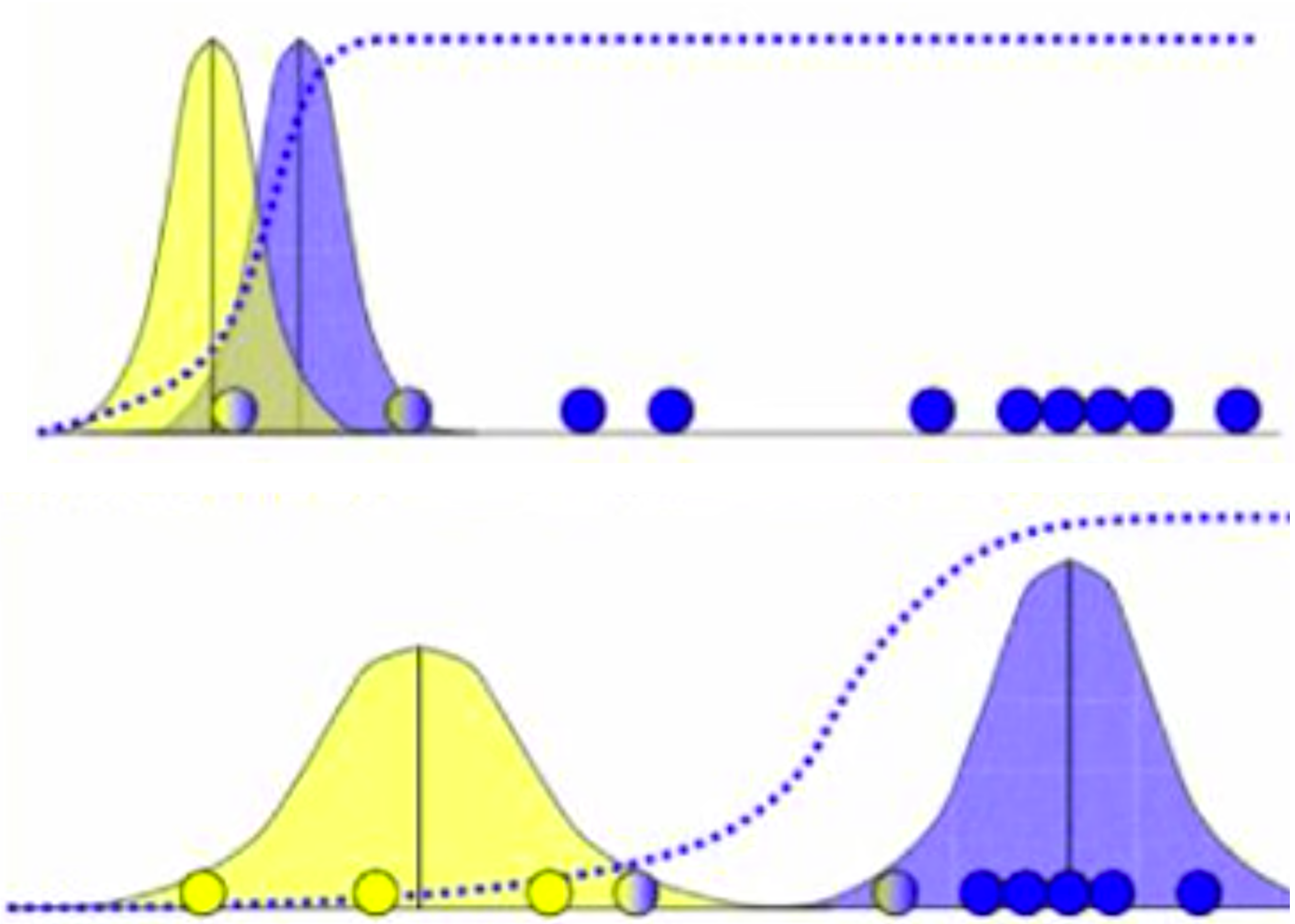
GRGPF

- Специализирован для многомерных неевклидовых пространств
- Полных открытых реализаций не было — реализован самостоятельно
- Оптимизирован для очень больших данных — может работать, если данные не помещаются в основную память

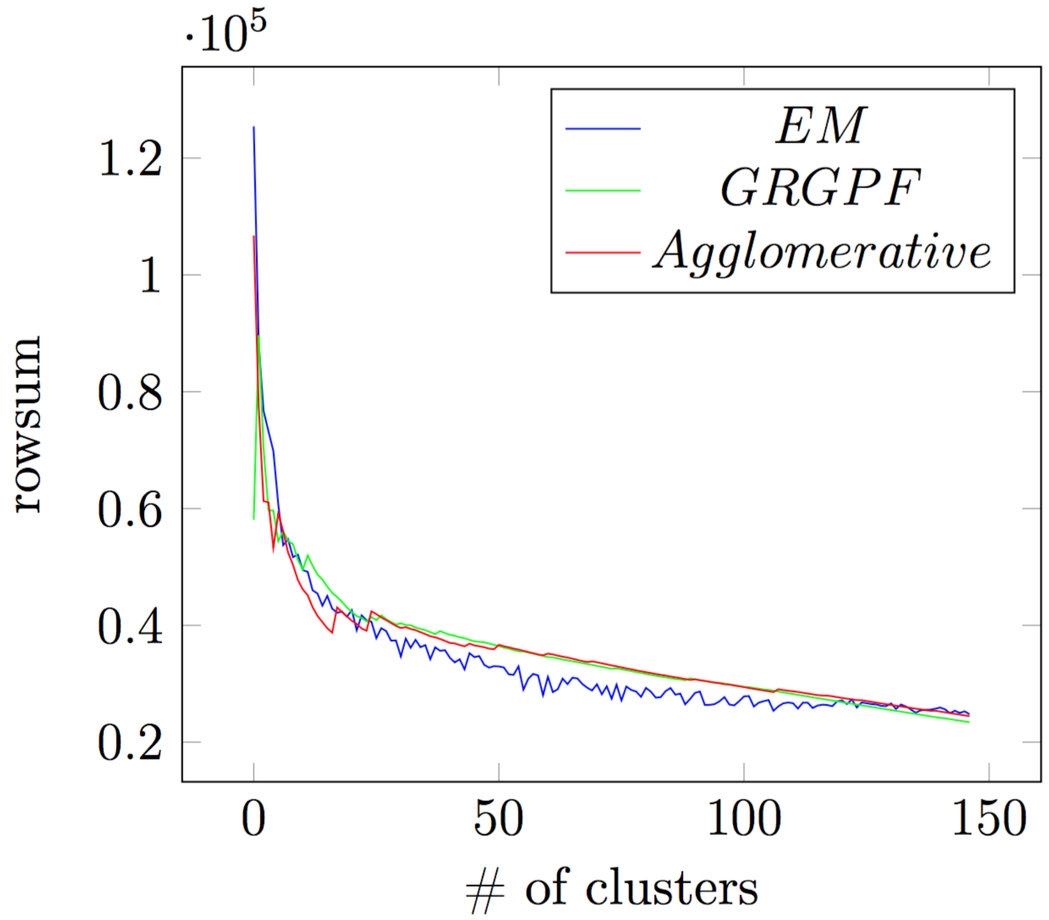


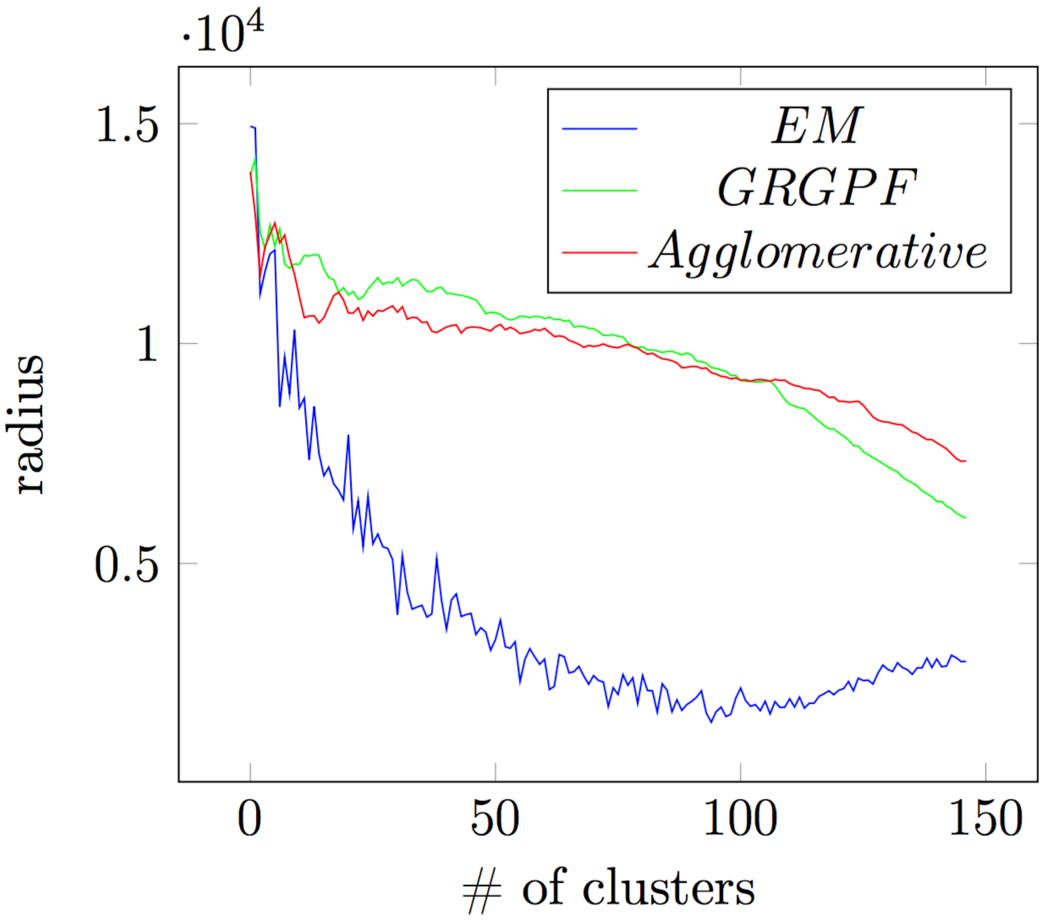
Expectation-maximization

- Статистический алгоритм, считает оценку максимального правдоподобия
- Использует вероятности вместо жесткого присвоения кластеру
- Обладает проблемой локального максимума



Оценка кластеризации

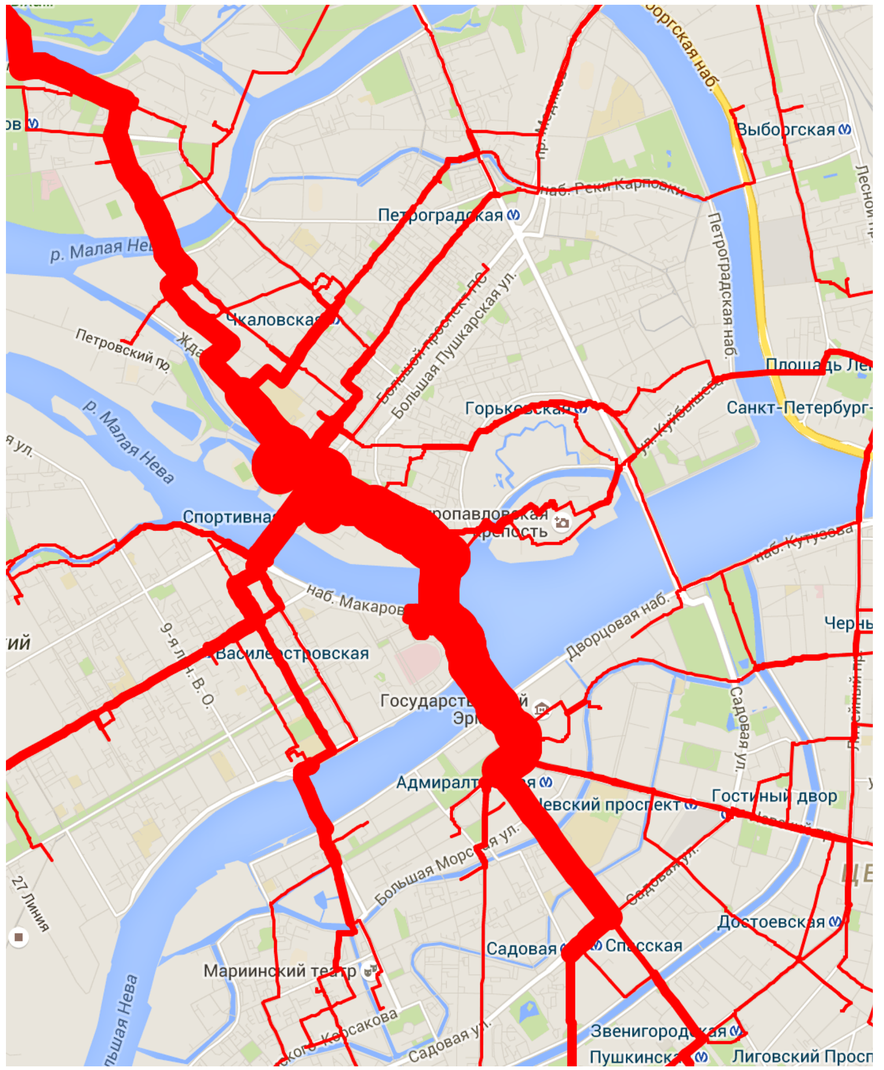




Результаты

Кластеры

Популярные пути



Результаты

- Данные подготовлены для кластеризации маршрутов
- Исследованы различные неевклидовы алгоритмы кластеризации
- Реализован алгоритм GRGPF
- Построены кластеры путей, построена статистика популярных путей

# top clusters	03.10	20.10	24.10	31.10	21.11	24.11
top 5	0.32	0.35	0.32	0.34	0.27	0.31
top 10	0.57	0.50	0.52	0.55	0.52	0.56
top 15	0.73	0.64	0.69	0.71	0.75	0.75
top 27	1.0	1.0	1.0	1.0	1.0	1.0