О возможности применения данных сотового оператора для построения карт дорожных пробок.

Дерендяев А. ИППИ РАН

Входные данные

A. Выгрузка из базы данных CDR (Call Data Record). Включает в себя следующие поля:

- Момент времени.
- ID телефона –хэш от MSISDN.
- CID - Номер БС.
- LAC – номер подсети. Любая БС идентифицируется по паре CID/LAC.
- Тип события.
- Географическая долгота БС.
- Географическая широта БС.
- Тип станции (OUTDOOR, INDOOR, METRO).
- Начальный угол направления антенны БС(облучаемый сектор).
- Конечный угол направления антенны БС(облучаемый сектор).

Пример:
2012-07-02 00:00:00.006;5068feb0;1612;50031;8;37.56268044;55.39734577;OUTDOOR;210;178.25;

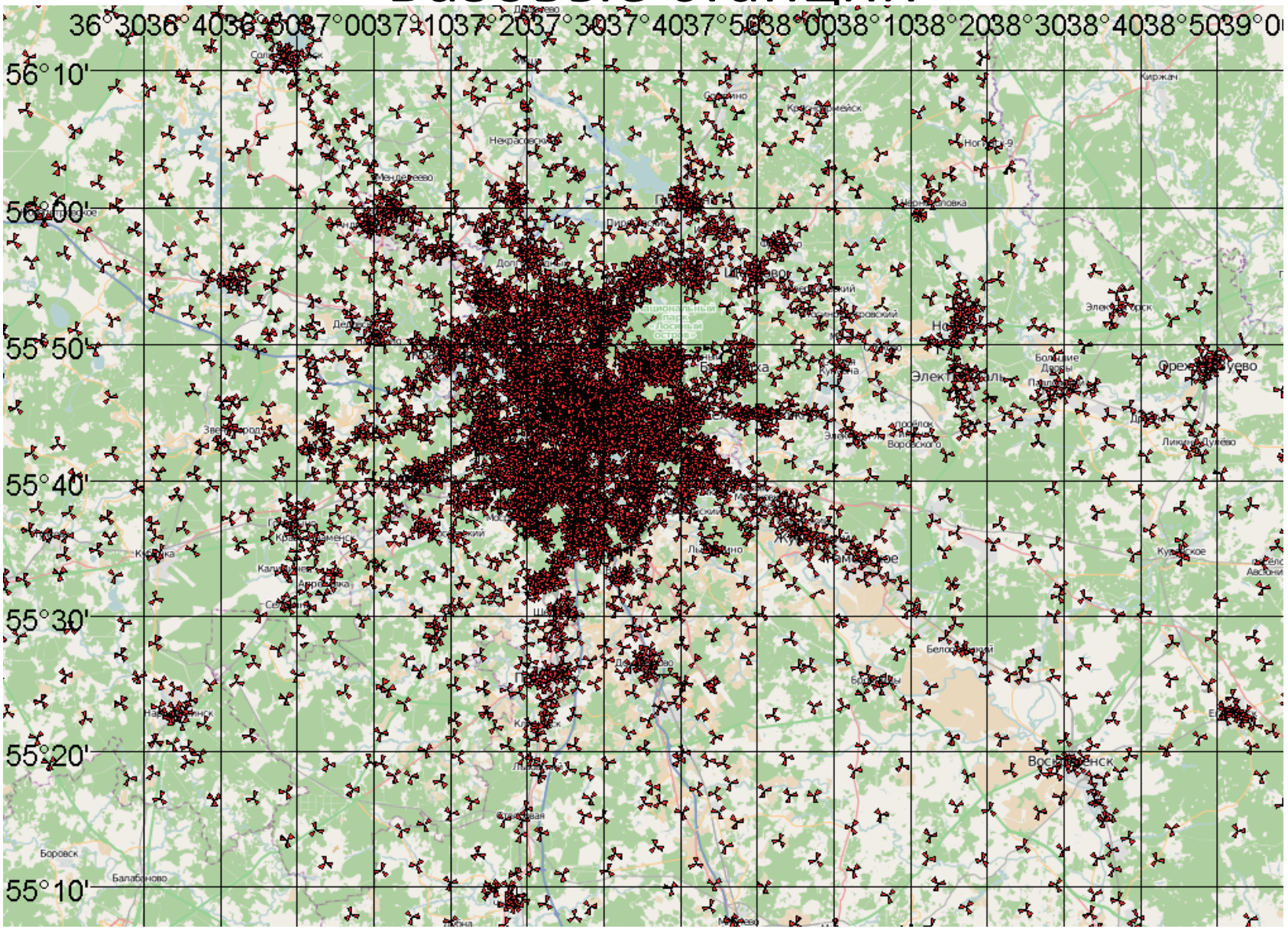
Таких событий происходит от 44 тысяч в минуту (ночью) до 311 тысяч (вечером).

Возможные типы событий

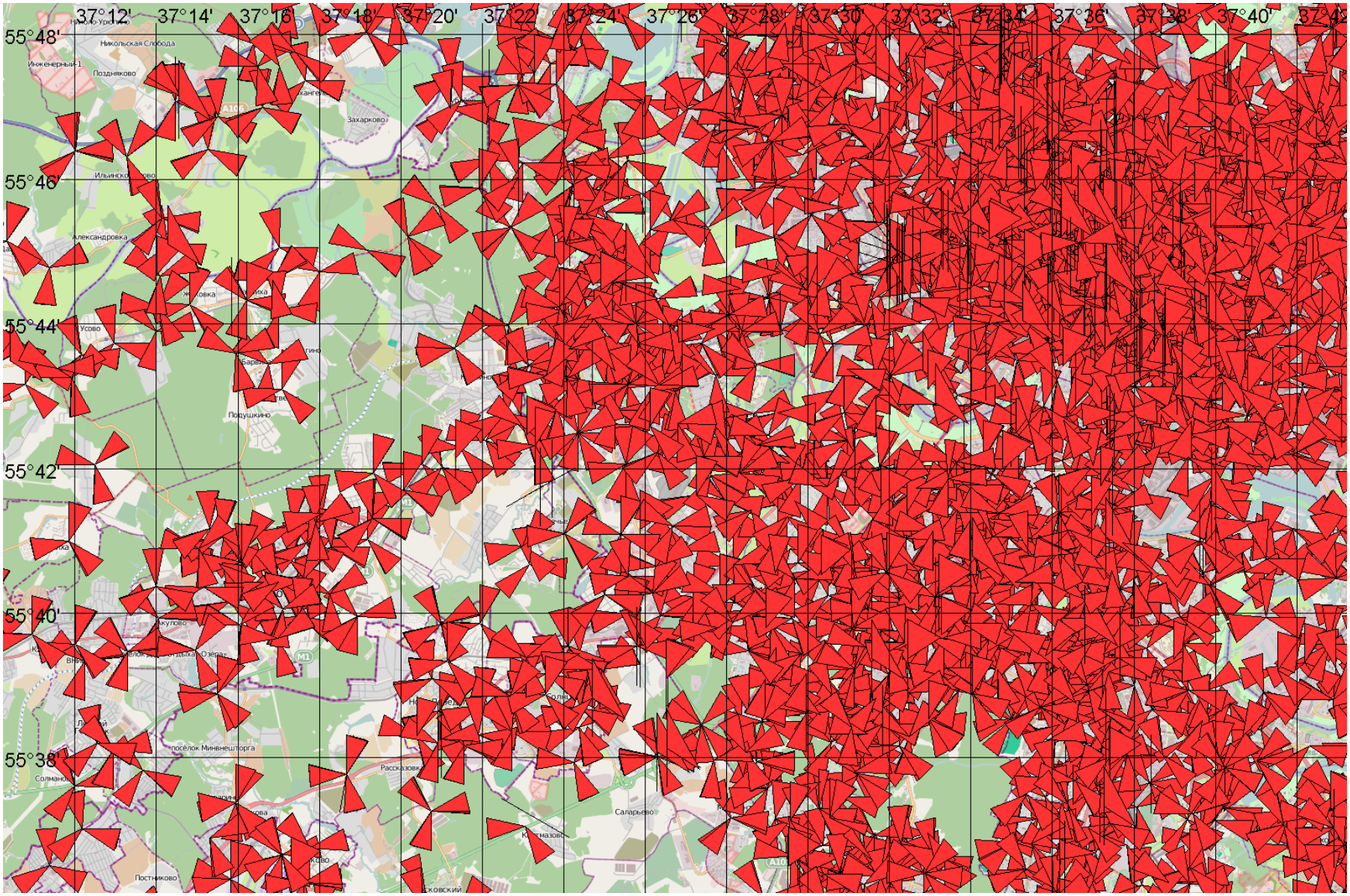
1. LocationUpdate (включение/выключение/смена соты/таймаут).
2. IMSI Attach. Включение телефона, в той же области (LA), в которой абонент выключил свой телефон. Если абонент включил телефон в другой области (LA), коммутатор сгенерирует событие типа LU.
3. Изменение местоположения (соты) абонента во время разговора или SMS
4. Абонент удален из VLR (Visitor Location Register)
5. Пометка абонента как «неактивный».
6. Абонент выключил свой телефон.
7. Определение абонента, как «недоступный». (Не было ни одного periodical LU, в течение определенного времени).
8. Изменение местоположения абонента (соты) во время передачи данных (GPRS)
9. Изменение местоположения абонента (соты) после входящего звонка или SMS.
10. Активность генерируется в том случае, если VLR узнает об изменении местоположения абонента от SGSN во время PSI процедуры.
11. Изменение местоположения абонента (соты) по запросу обновления абонентских данных, в течение голосового вызова.

Предположение: Будем использовать только события 1 типа, т.е. только собственно события регистрации в соте (они составляют 69% всех событий).

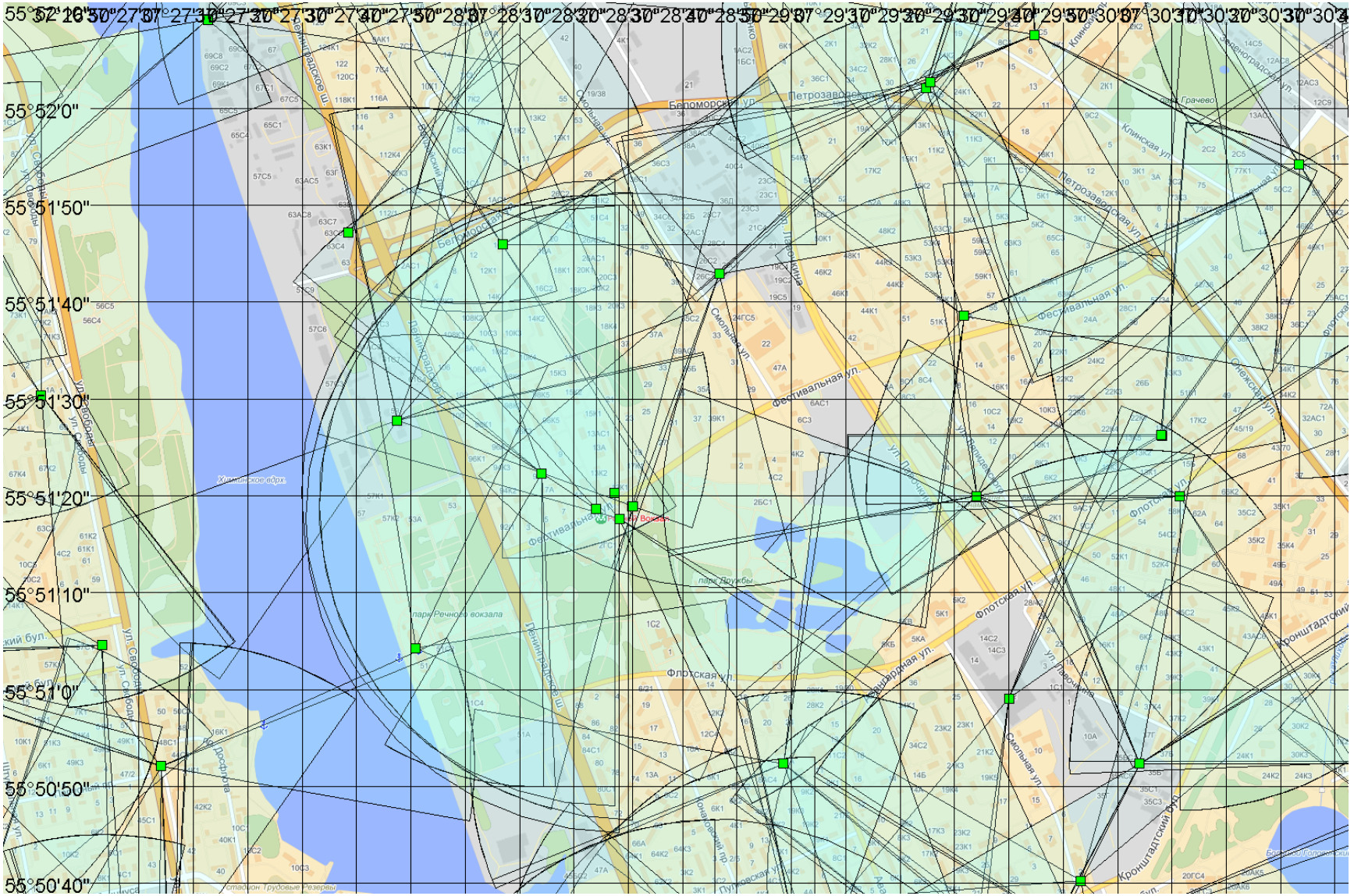
Базовые станции



Базовые станции



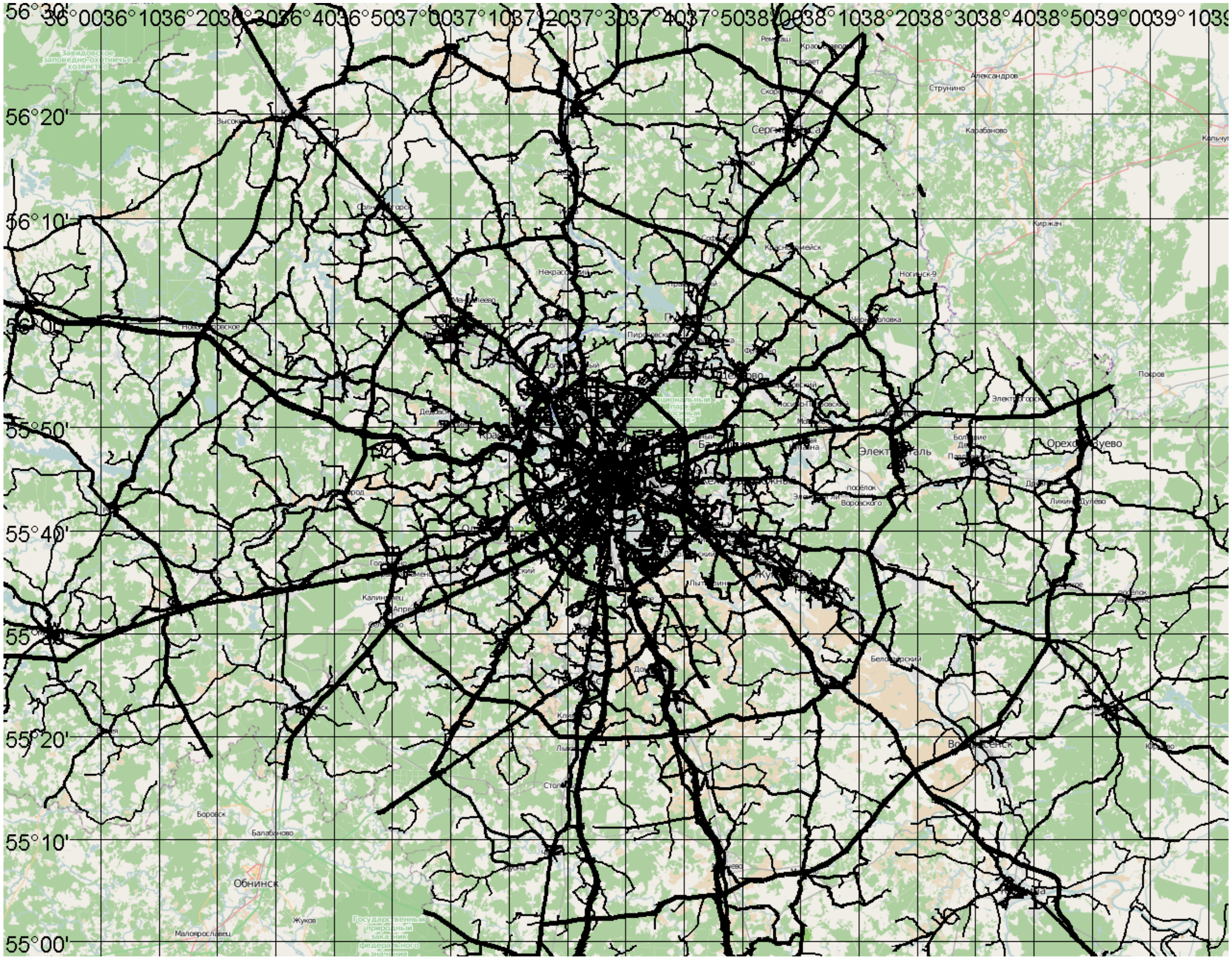
Базовые станции



Б. Локализованный граф дорог, полученный с сайта OpenStreetMap(OSM). Включает в себя пару shape-файлов (отдельно для Москвы и Московской области), с атрибутивной информацией, в том числе:

- ID на сайте OSM.
- Направление движения (однополосная или нет).
- Тип дороги (highway, 1 класс, 2 класс, пешеходная тропа, сервисный проезд, соединения дорог, и т.д.)
- Максимальная скорость.
- Число полос.
- Некоторые другие атрибуты.

Граф дорог


56°30'
56°20'
56°10'
55°50'
55°40'
55°20'
55°10'
55°00'
Обнинск
Балабаново
Боровск
Малоярославец
Жуков
Струнино
Александров
Карабаново
Киржач
Покров

Общая идея

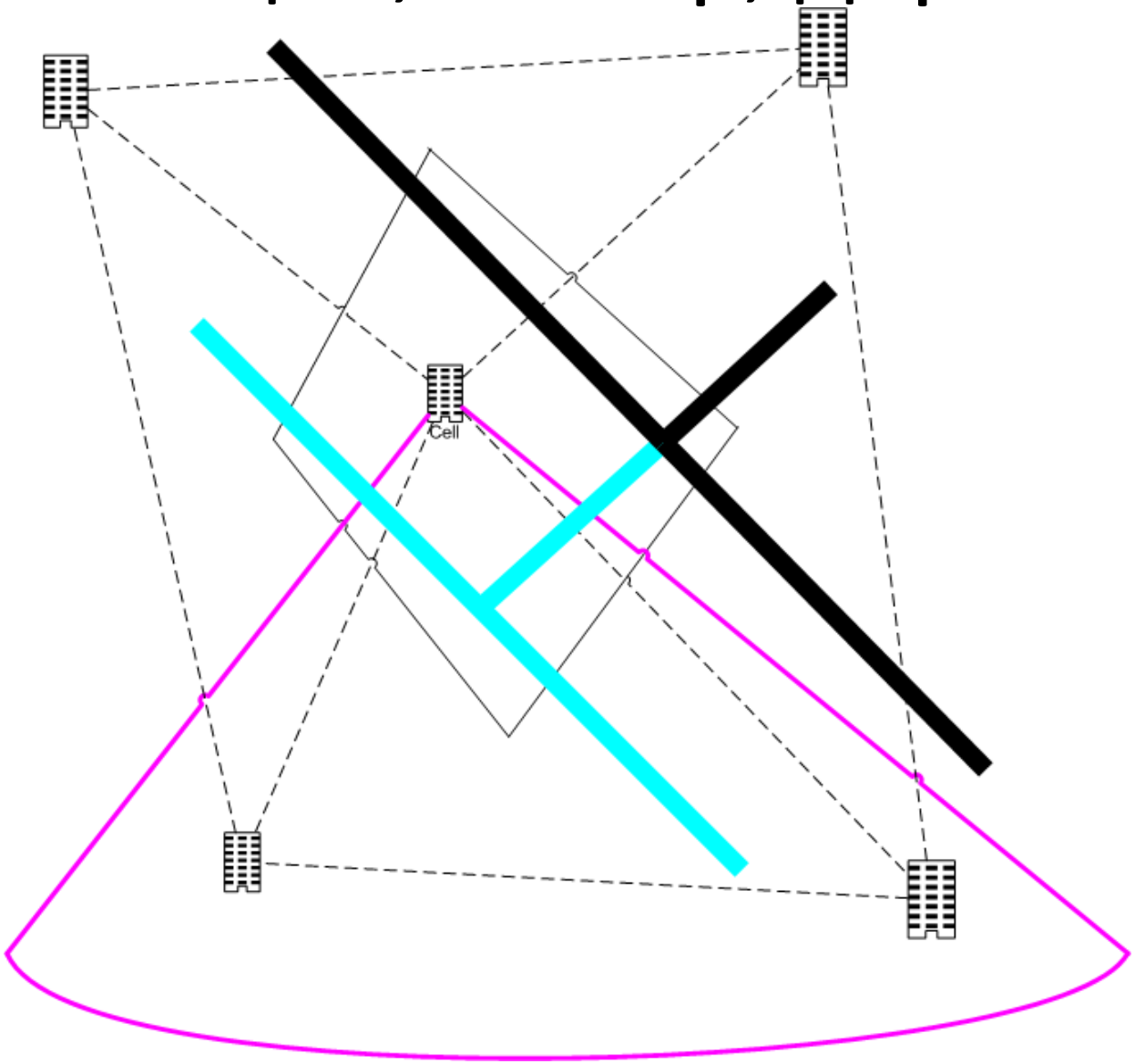
Построить треки (последовательность БС) для всех абонентов, оценить возможные участки дорог и скорость прохождения по ним. Выбрать минимальный (в географическом смысле) путь и для всех участков дорог на данном пути оценить скорость передвижения абонентов.

Предварительная обработка

- Из-за того, что данные из OSM идут в двух shape-файлах, их надо«сшить», с учетом их ID в базе OSM. Кроме того, так как в OSM дорога с одним ID может иметь несколько пересечений (в том числе и не в конечных точках), требуется регуляризация данного графа, т.е. разделение таких дорог на независимые участки. Каждому такому участку выдается новый ID. Из всего массива полученных дорог мы убираем излишние (в том числе пешеходные тропы, сервисные подъезды и внутредворовые дороги)
- Разделение БД CDR – из всего потока событий можно извлечь информацию, касающуюся только БС: ее номер, номер подсети, широту, долготу, тип станции, углы направления антенны. Эти данные для одной и той же станции одинаковы для всех событий. Любая БС идентифицируется по паре номер БС/номер подсети, но в рамках московского региона для идентификации достаточно номера БС.
- Привязка к БС возможных участков дорог. Разделяя всю территория диаграммой Вороного с центрами в БС, будем считать, что те участки дорог, которые попали в область с центром какой-нить БС, будут возможными участками дорог для данной БС. Кроме того, после такого разбиения учтем углы направления антенн БС и выкинем непопадающие участки дорог. Данные предположения не сильно отрываются от реальности, так как углы не сильно соблюдаются.

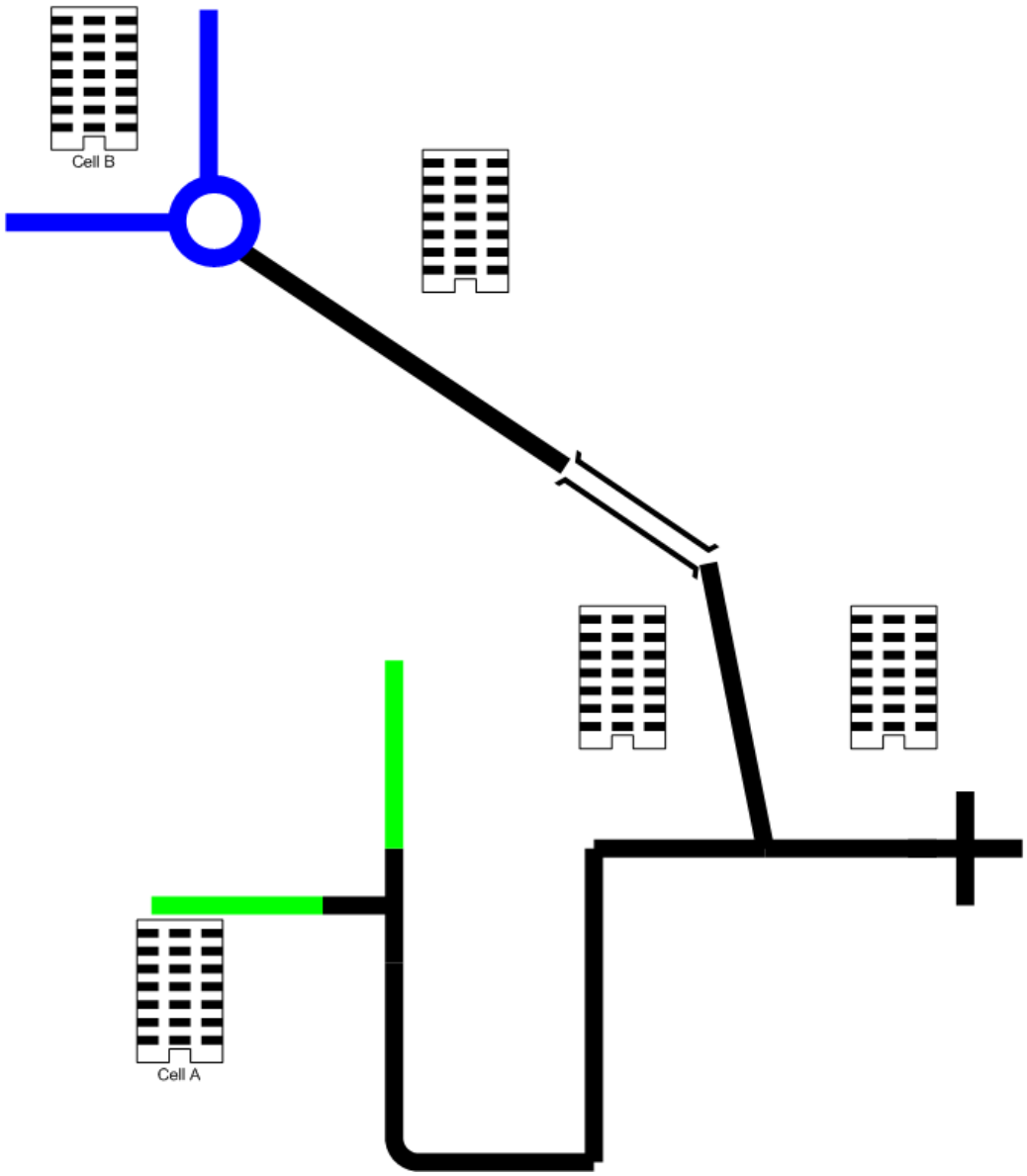
  Самым правильным методом была бы карта распределения приема БС, полученная экспериментальным путем. Это позволит учесть сразу все возможные поправки (на рельеф, переотражения, некорректно проставленные углы и прочее).

Станции, сектор, дороги

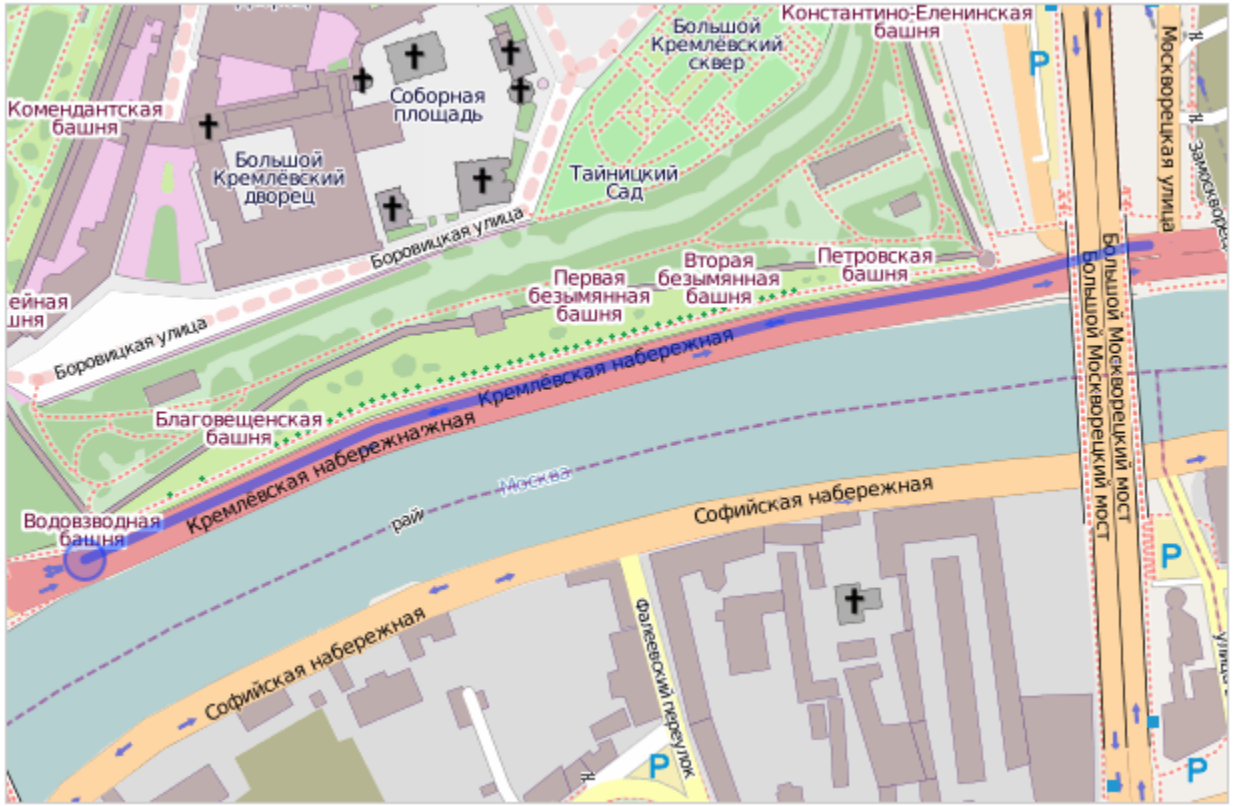


Основной алгоритм

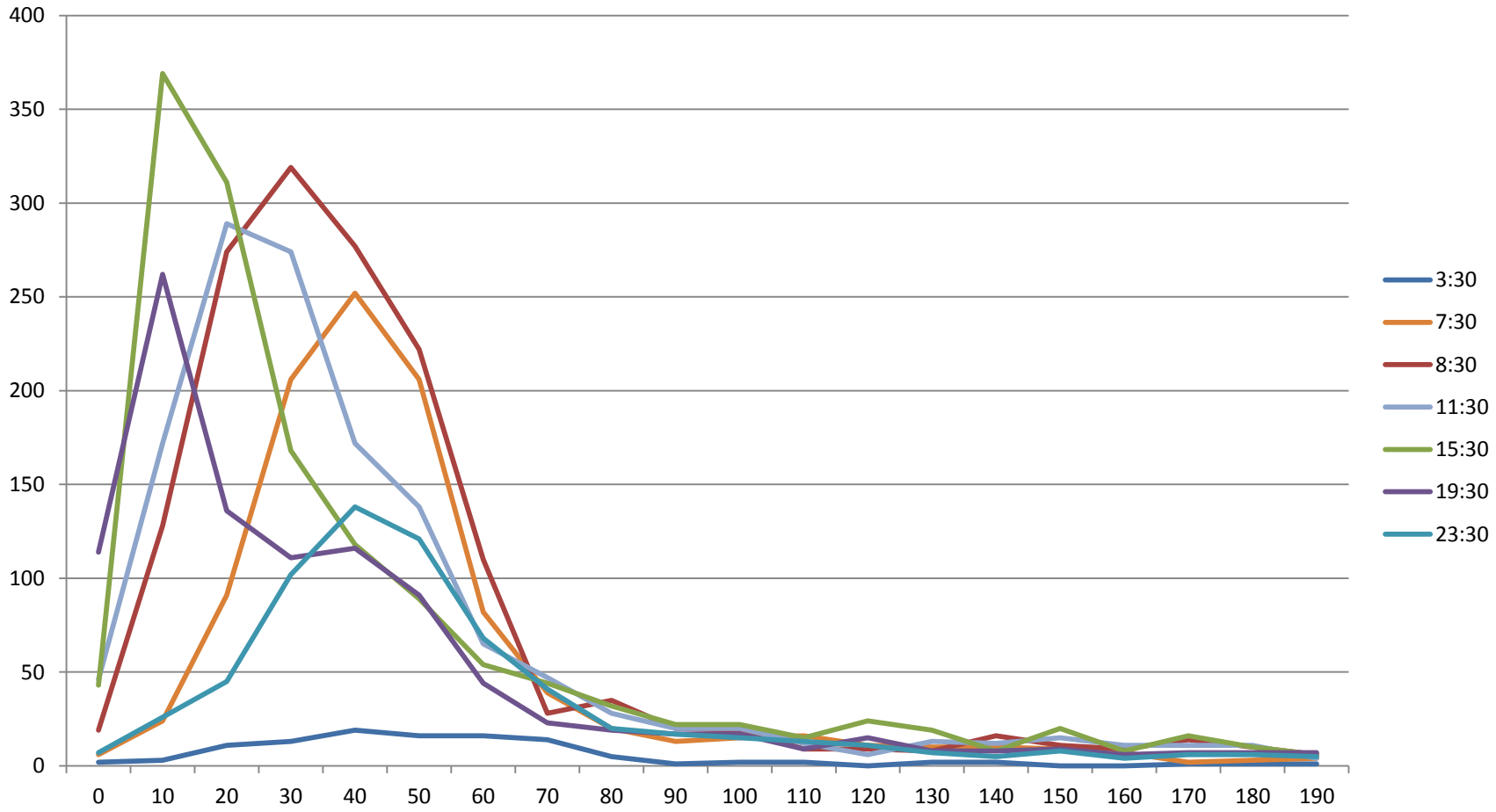
- Для каждого вновь поступающего события (по порядку из БД CDR), определяем наличие трека для данного абоненты. Если трека нет, то начинаем его строить.
- Если трек есть и последняя пара событий из него удовлетворяет некоторым условиям (например, время между событиями меньше получаса, расстояние между БС меньше 20 км, нет возвратных прыжков, и, м.б., некоторым другим), то тогда жадным алгоритмом Дейкстры по графу дорог (вес ребер – длина дорог) строится минимальный путь между всеми дорогами, привязанными к первой БС и всеми дорогами, привязанными к второй БС (с учетом односторонности).
- Высчитывается длина этого минимального пути и, с учетом разницы времен между событиями, можно оценить среднюю путевую скорость на этом маршруте. Эта скорость записывается в статистику для каждого участка дорог полученного маршрута.
- Анализируя статистику для всех участков дорог, мы выбираем минимальную моду этого распределения, квантуя по 10 км/ч. При этом существуют предположения: не имеет смысл анализировать участки дороги с малым числом событий (менее пяти) и со скоростями, превышающим здравый смысл (т.е., более 200 км/ч).


Cell B
Cell A

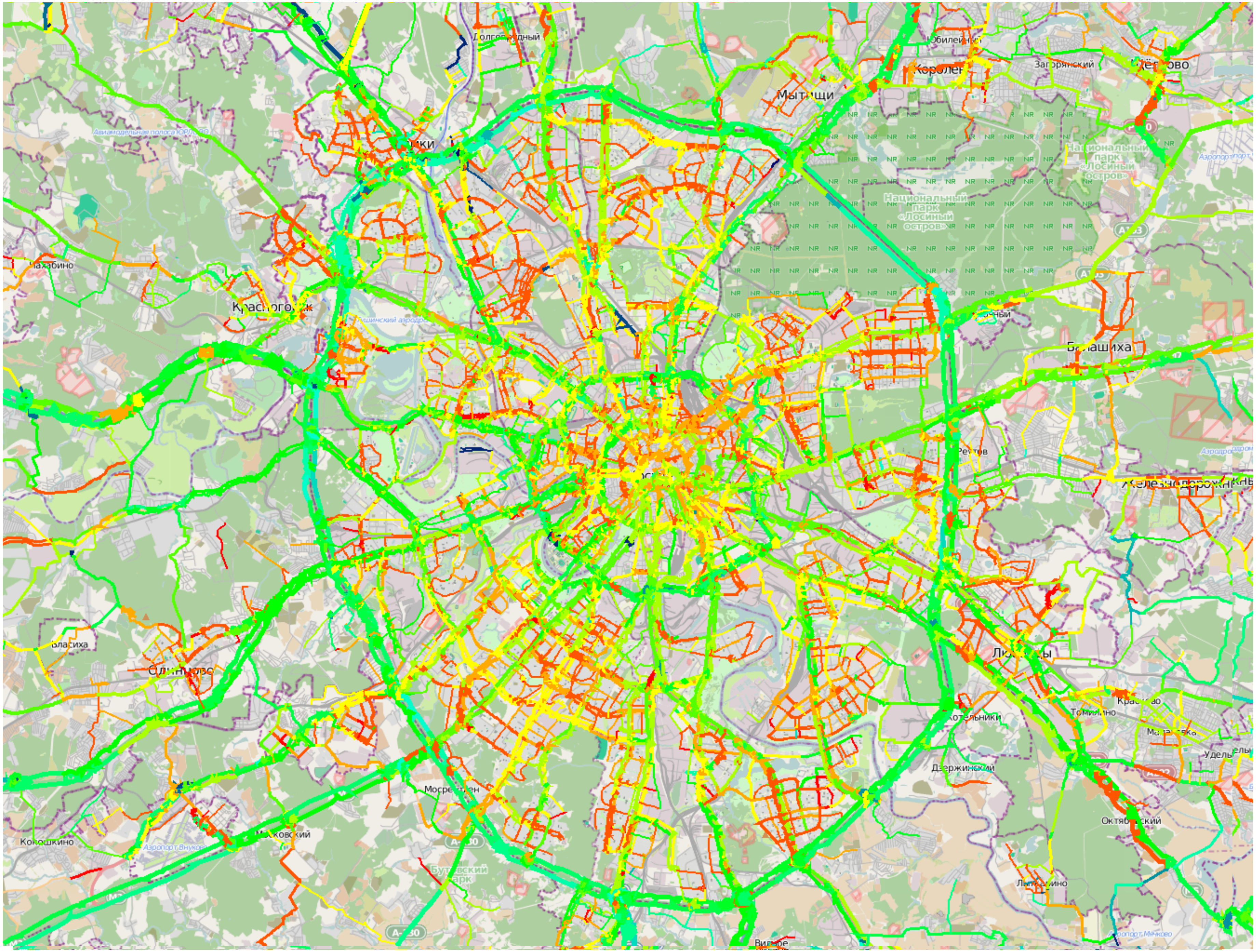
Кремлевская набережная



Кремлевская набережная



Число предположительных абонентов на Кремлевской набережной в зависимости от скорости в км/ч в разные моменты времени


Мытищи
Загорянский
Юбилейный
Национальный парк «Лосиный остров»
Балашиха
Реутов
Дзержинский
Власиха
Мосрентген
Московский
Октябрьский
Бутовский парк

Возможные дополнения

1. Исследовать не только последнюю пару событий. Позволит устойчиво оценить границы попадания в зону действия станции.
2. Можно учитывать другие типы событий, в первую очередь – 4 типа, т.е. удаления пользователя из списка абонентов для данной станции.
3. Оценивать вес ребер не по их длине, а по времени, с учетом уже вычисленных скоростей.
4. Оценивать не только один минимальный путь, а несколько. Учитывать их веса при принятии решения о скорости.
5. Использовать не моду распределения, а более правдивые оценки скорости. Например, 70% квантиль. Учитывать вид распределения – он разный для разных дорог в разные моменты времени.

Спасибо за внимание!