Tinkoff Data Science Challenge
Восстановление позиции мерчанта

Иван Бендына

Постановка задачи

Мерчант — место, принимающее платежи с использованием банковской карты.

Банк хочет знать геолокацию мерчантов.

Постановка задачи

Даны чекины пользователей.

Чекин это отправка своих геокоординат в небольшой промежуток времени от покупки.

Постановка задачи

Данные очень шумные.

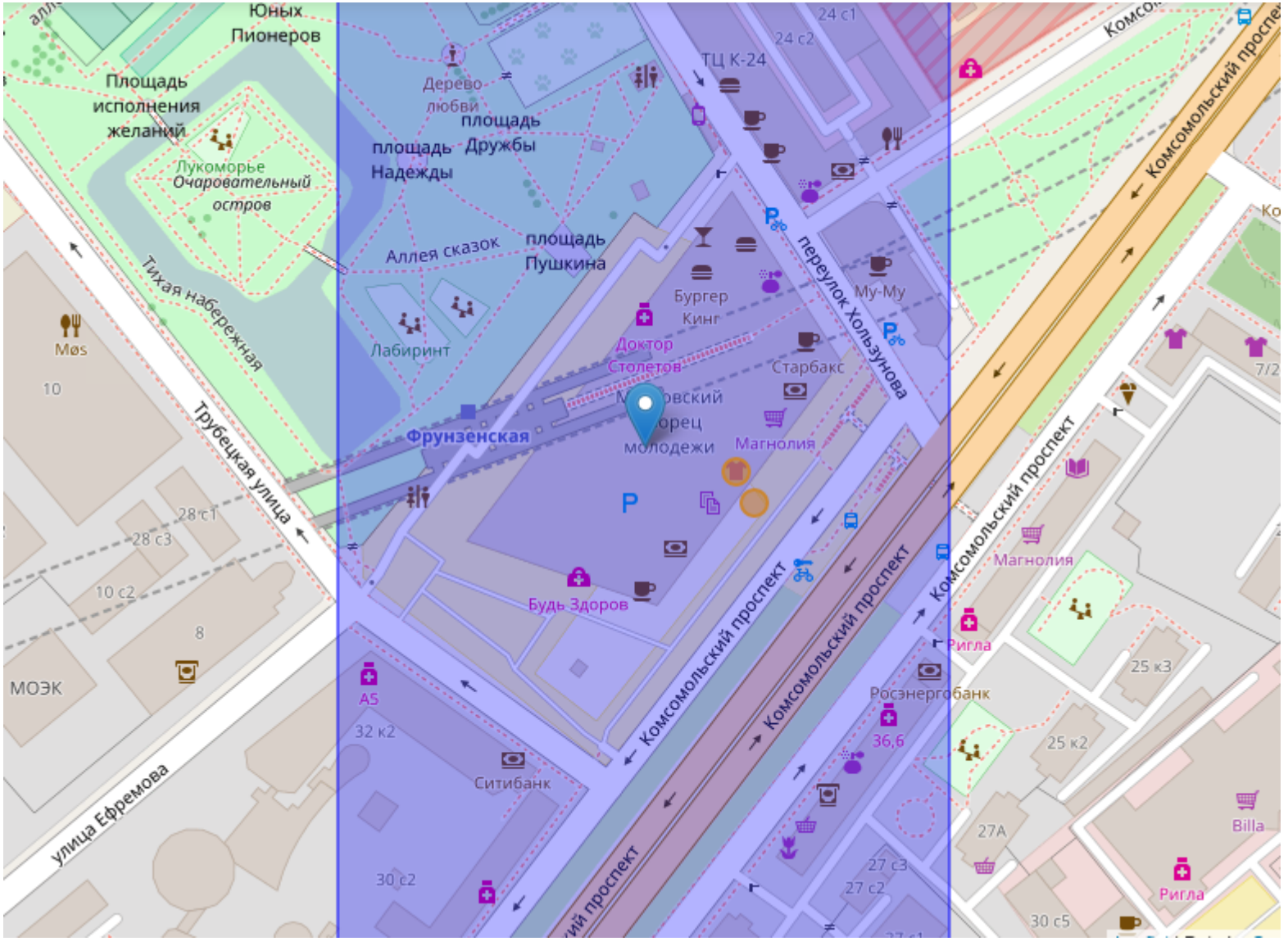
Возможные причины:

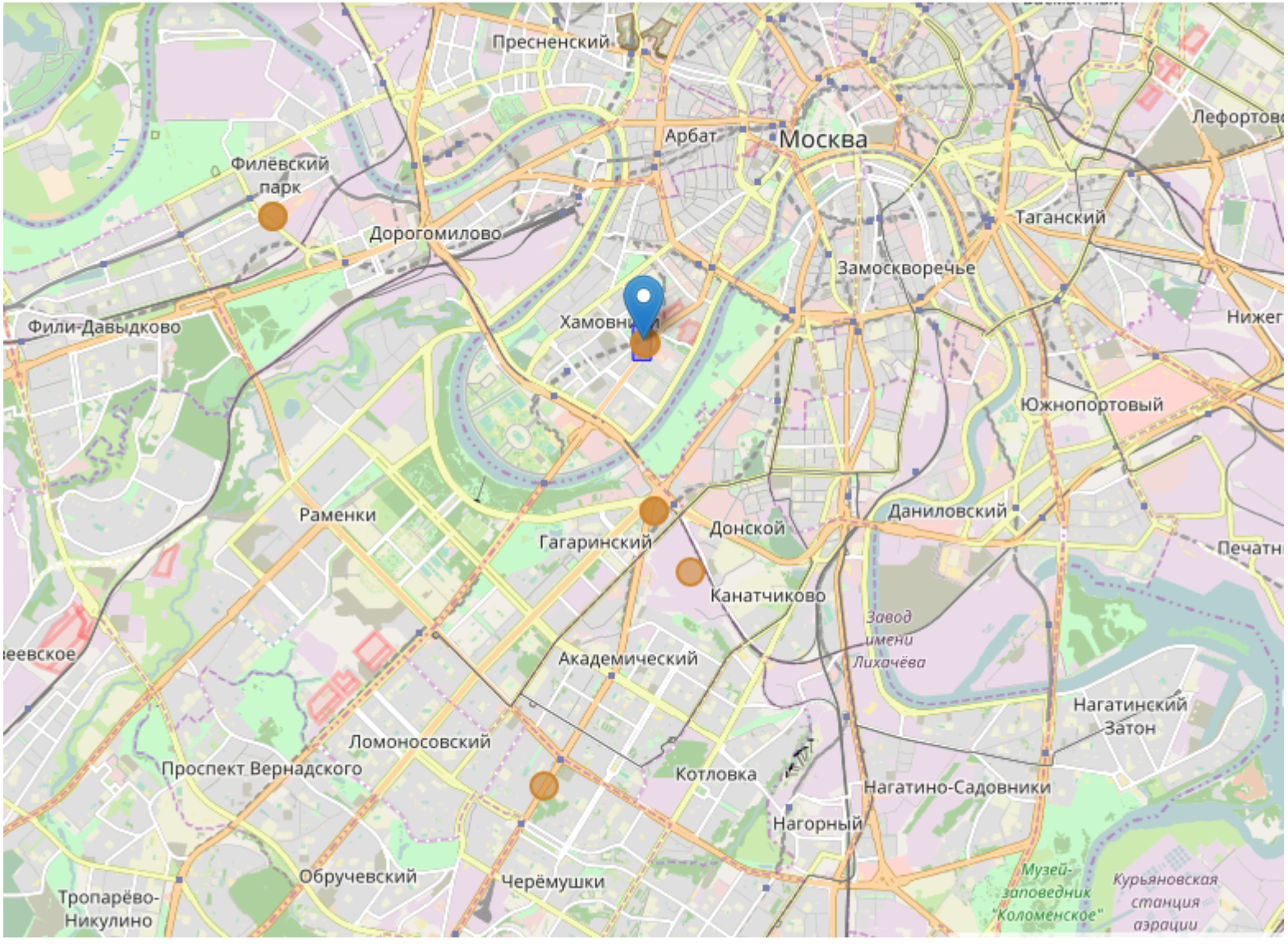
- не обновились GPS координаты на телефоне
- неточности геолокационного алгоритма (wi-fi и вышки)
- мерчант может быть только онлайн
- пользователь успел далеко отъехать
- ошибки ввода, и.т.д

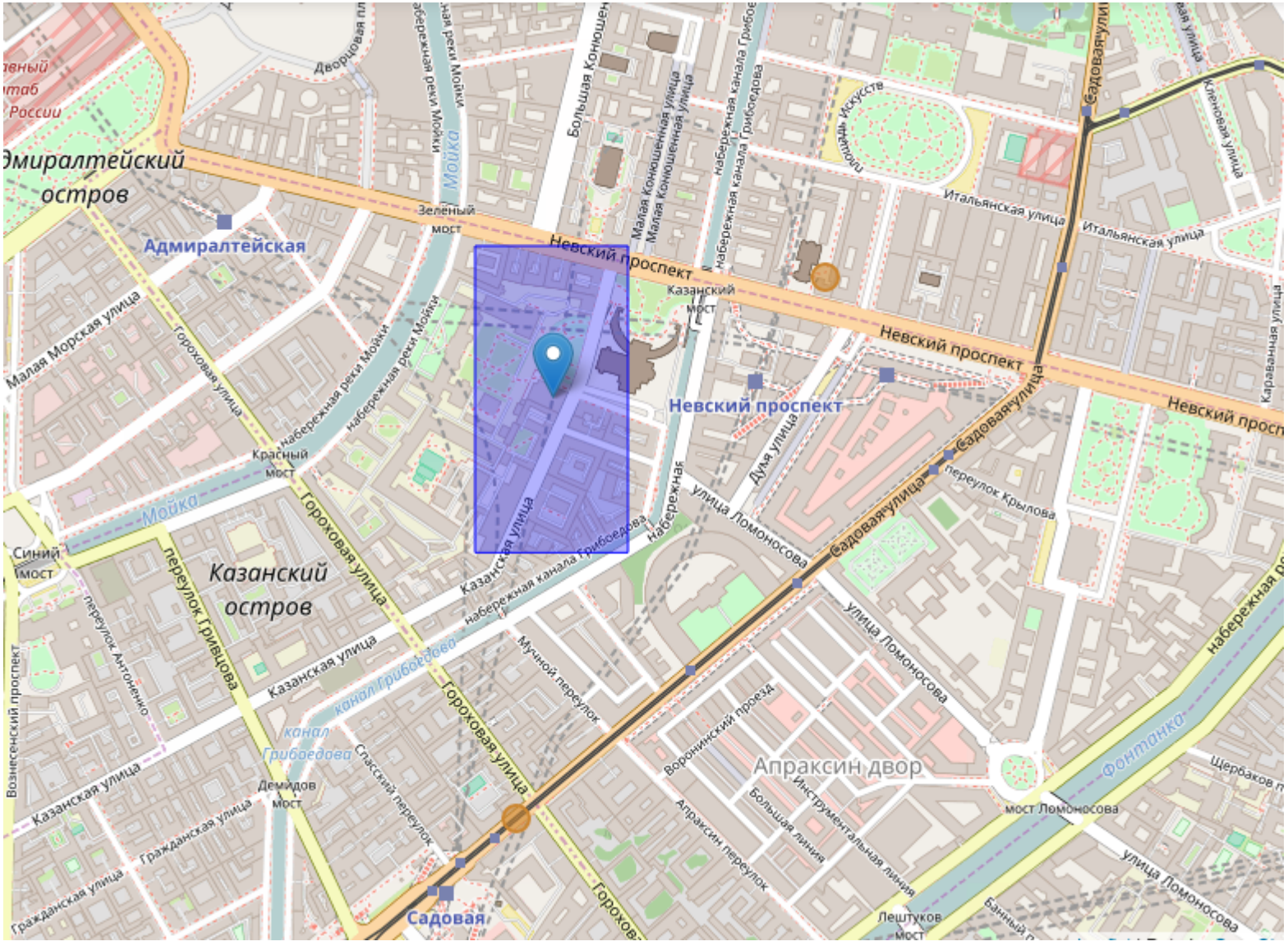
Постановка задачи

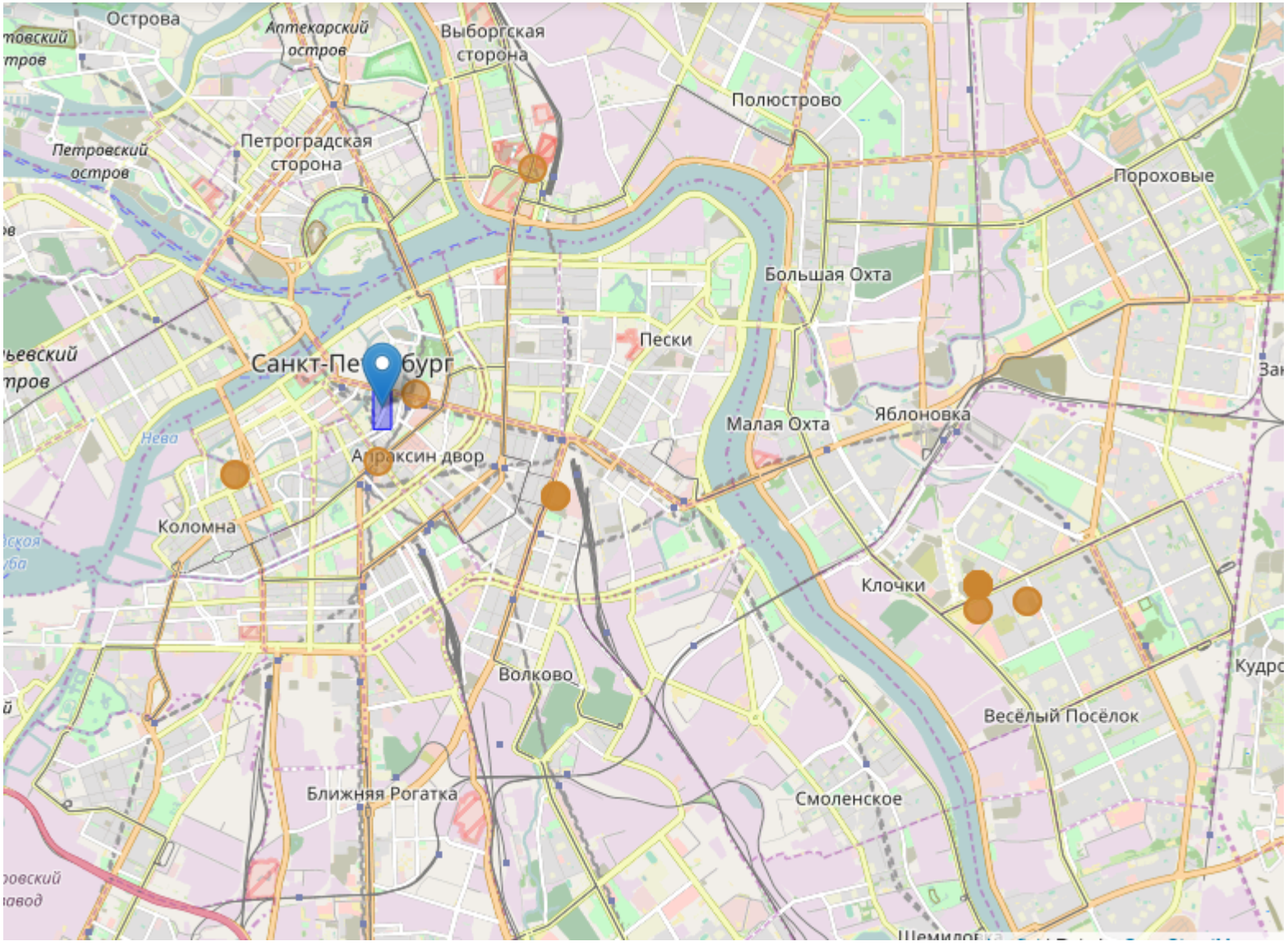
Метрика: accuracy.

Ответ правильный, если координаты отличаются не более чем на 0.002 от настоящих по широте и долготе.


Юных Пионеров
Площадь исполнения желаний
Лукоморье
Очаровательный остров
Дерево любви
площадь Дружбы
площадь Надежды
Аллея сказок
площадь Пушкина
Лабиринт
Тихая набережная
Трубецкая улица
Møs
10
28 с1
28 с3
10 с2
8
МОЭК
улица Ефремова
32 к2
А5
Ситибанк
30 с2
Фрунзенская
ТЦ К-24
24 с1
24 с2
переулок Хользунова
Бургер Кинг
Доктор Столетов
Старбакс
Магнолия
Будь Здоров
Комсомольский проспект
Му-Му
Ригла
Росэнергобанк
36,6
25 к3
25 к2
27А
27 с3
27 с2
30 с5
Billa


Пресненский
Арбат
Москва
Лефортов
Филёвский парк
Дорогомилово
Таганский
Замоскворечье
Фили-Давыдково
Хамовники
Нижег
Южнопортовый
Раменки
Гагаринский
Донской
Даниловский
Печатн
Канатчиково
Завод имени Лихачёва
Академический
Нагатинский Затон
Ломоносовский
Проспект Вернадского
Котловка
Нагатино-Садовники
Нагорный
Обручевский
Черёмушки
Тропарёво-Никулино
Музей-заповедник "Коломенское"
Курьяновская станция аэрации


Эмиралтейский остров
Адмиралтейская
Невский проспект
Казанский мост
Невский проспект
Невский проспект
Малая Морская улица
Гороховая улица
набережная реки Мойки
Красный мост
Мойка
Синий мост
Казанский остров
переулок Гривцова
Казанская улица
набережная канала Грибоедова
Казанская улица
Вознесенский проспект
переулок Антоненко
канал Грибоедова
Гороховая улица
Мучной переулок
Спасский переулок
Демидов мост
Гражданская улица
Гражданская улица
Садовая
Садовая улица
Дума улица
улица Ломоносова
переулок Крылова
Воронинский проезд
Апраксин двор
Инструментальная линия
Большая линия
Апраксин переулок
мост Ломоносова
Фонтанка
набережная
Лештуков мост
улица Ломоносова
Большая Конюшенная улица
Малая Конюшенная улица
Итальянская улица
Итальянская улица
Караванная улица
Кленовая улица
Садовая улица
Зеленый мост


Острова
Аптекарский остров
Выборгская сторона
Полюстрово
Петроградская сторона
Петровский остров
Пороховые
Большая Охта
Пески
Санкт-Петербург
Нева
Малая Охта
Яблоновка
Апраксин двор
Коломна
Клочки
Волково
Весёлый Посёлок
Ближняя Рогатка
Смоленское

Фильтрация

Частые точки чекинов:

55.75034704,	37.62385111	(27955)
55.75034704,	37.6235111	(722)
55.750347,	37.623851	(471)
55.750347043,	37.623851113	(150)
0.0,	0.0	(25018)
51.17889992,	-1.82639994	(5420)
51.1789,	-1.8264	(228)
51.172,	-1.2634	(177)

Фильтрация

Фильтруем точки чекинов не в России.

Геоданные OSM

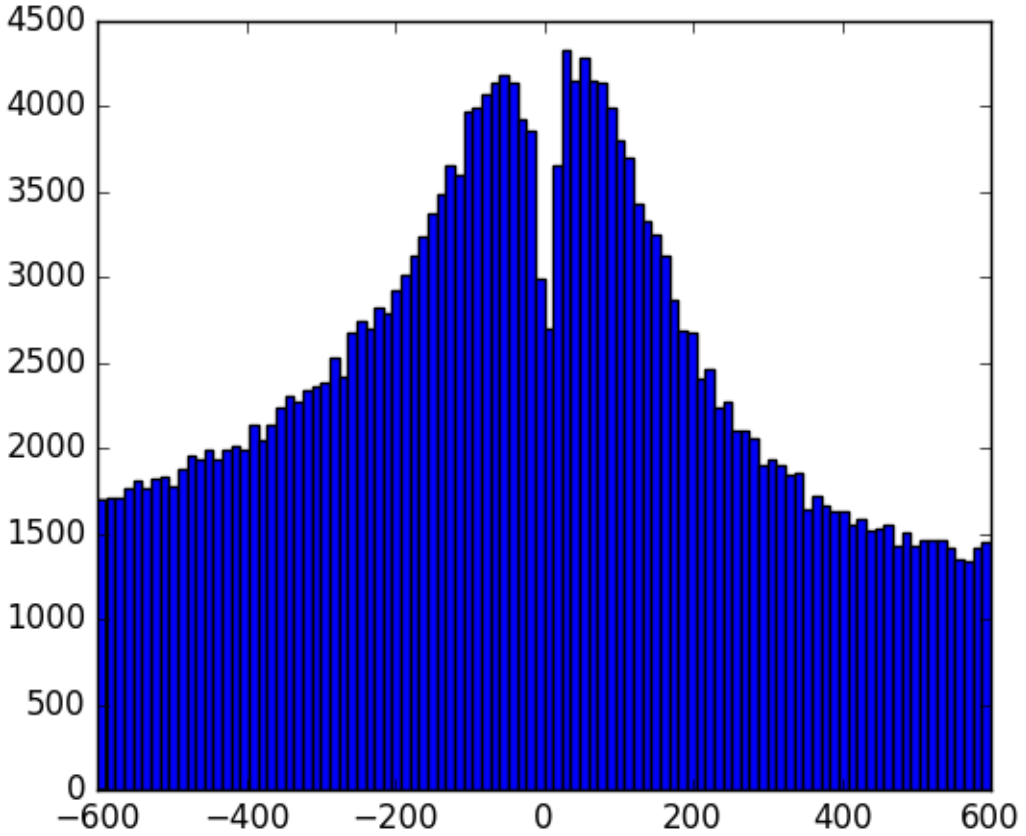
Nominatim reverse geocoder – адрес по координатам.

Можно запустить локально.

Данные в формате osm.pbf http://download.geofabrik.de

Россия 2Gb, развернуть на ноутбуке около 10 часов.

Распределение разницы времени транзакции и чекина



Решение

Создаем матрицу объекты-признаки, где объекты это чекины.
Таргет – это попадает ли чекин в прямоугольник.

Оптимальное решение: 0.47

Признаки

Признаки основанные на расстоянии от этого чекина до других чекинов этого же мерчанта.

- среднее расстояние
- среднее расстояние до уникальных
- перцентиль по среднему расстоянию

Признаки

Признаки основанные на расположенных рядом известных мерчантах.

- минимальное расстояние
- количество в прямоугольнике 0.002 (0.004, 0.006)

Признаки

Признаки основанные на зданиях.

- площадь здания
- количество магазинов
- суммарная/максимальная площадь зданий в радиусе 200 метров

Признаки

Признаки основанные на времени.

- разница между чекином и транзакцией
- была ли одна транзакция с разными координатами чекинов?

Модель

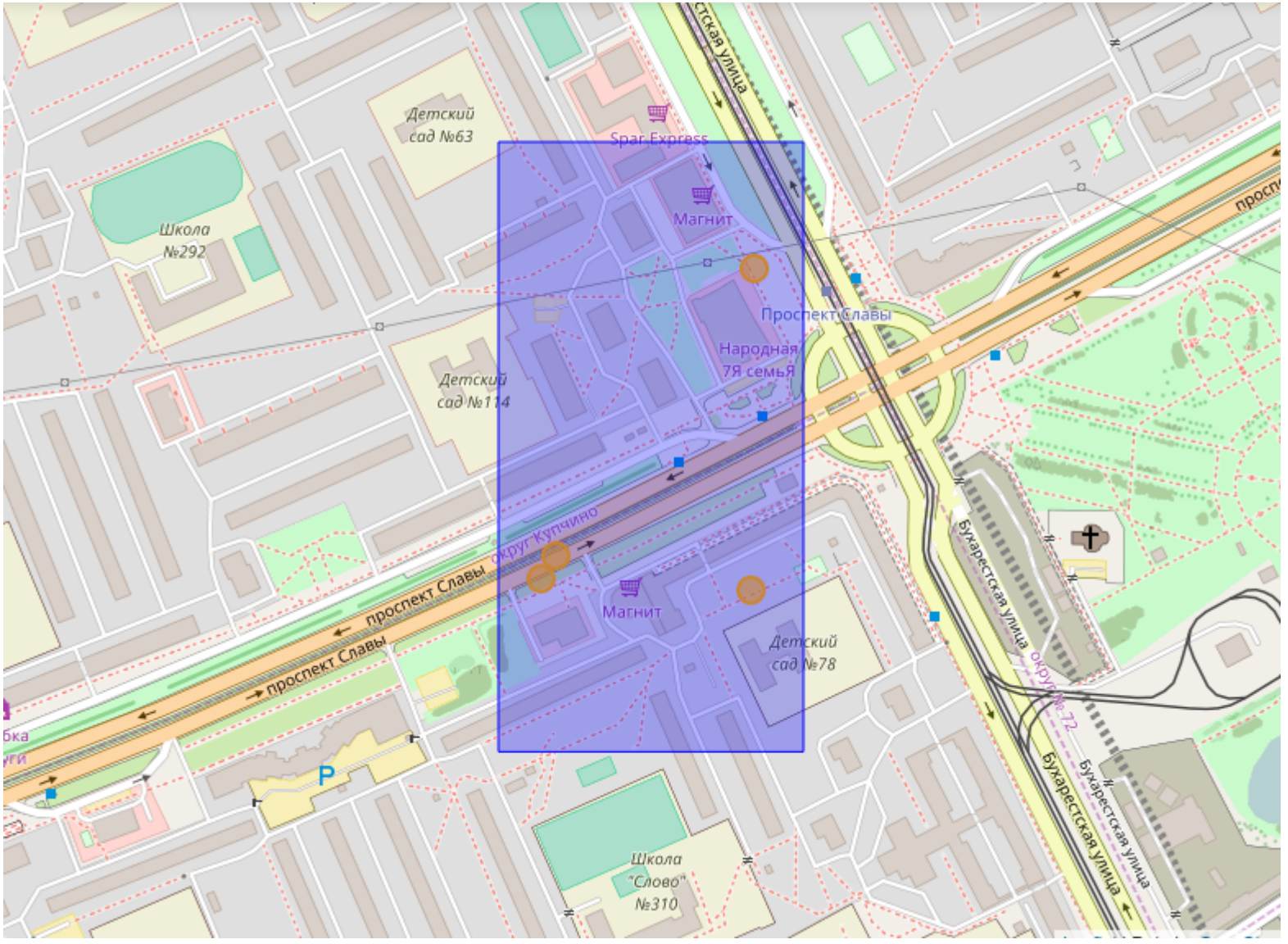
Все эти признаки в xgboost – в качестве предсказания чекин с наибольшей вероятностью.

LB score **0.360**

Объединение транзакций

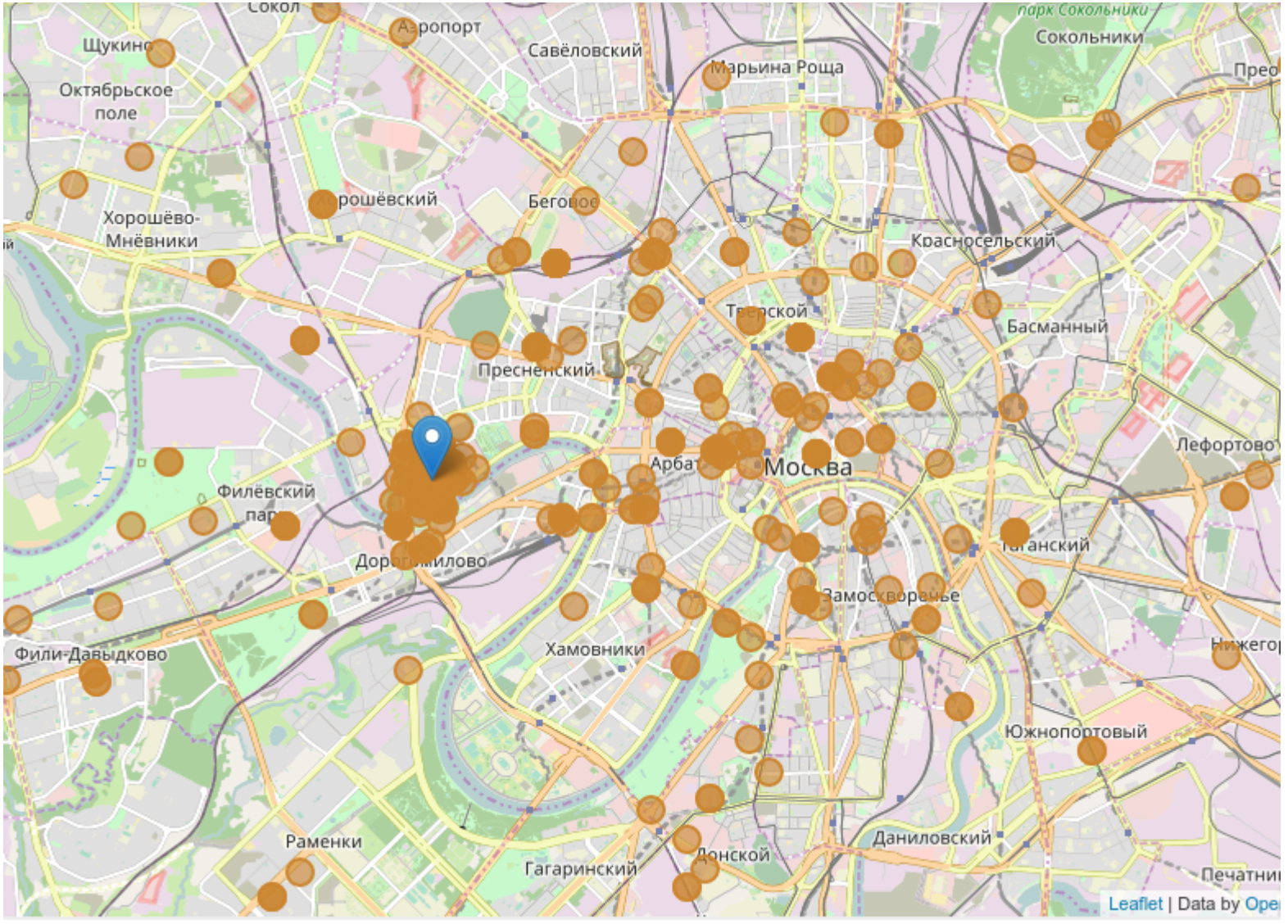
Если вероятности нескольких транзакций достаточно большие и они находятся недалеко, то в качестве предсказания – центр описанного прямоугольника этих транзакций.

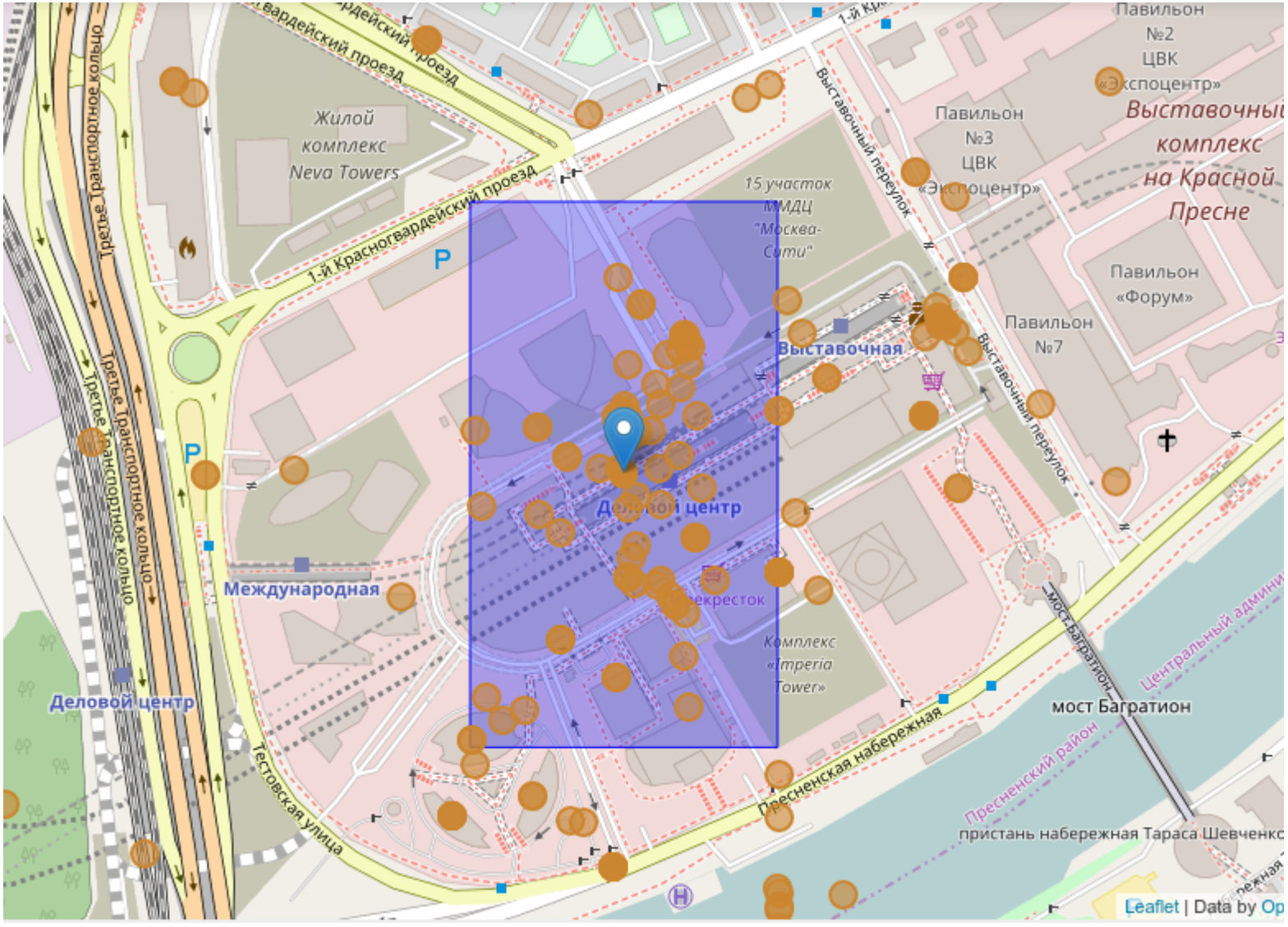
LB score **0.365**

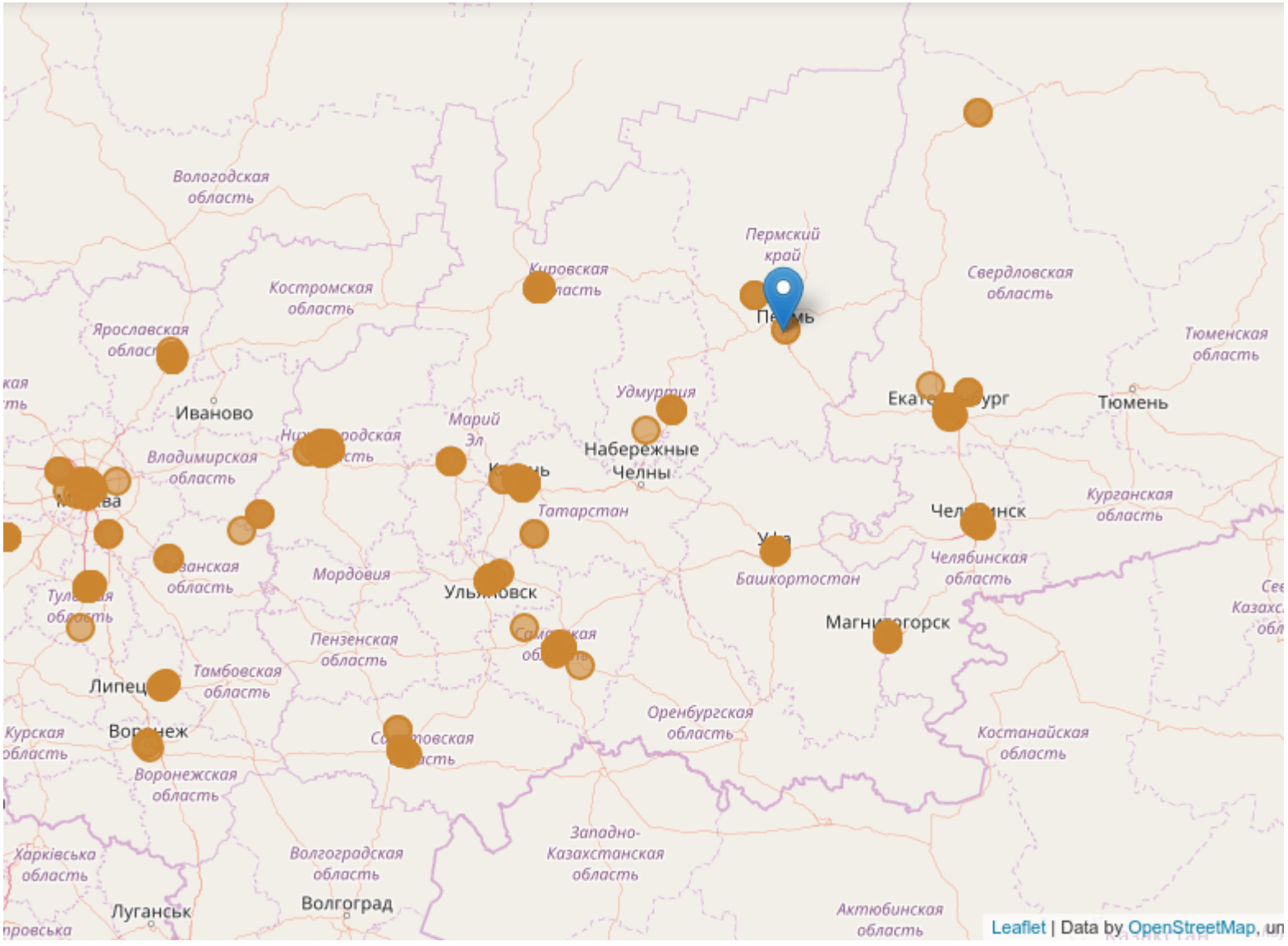

Детский
сад №63
Spar Express
Магнит
Школа
№292
Проспект Славы
Народная
7Я семьЯ
Детский
сад №114
округ Купчино
проспект Славы
проспект Славы
Магнит
Детский
сад №78
Бухарестская улица
округ №72
Бухарестская улица
Бухарестская улица
Школа
"Слово"
№310

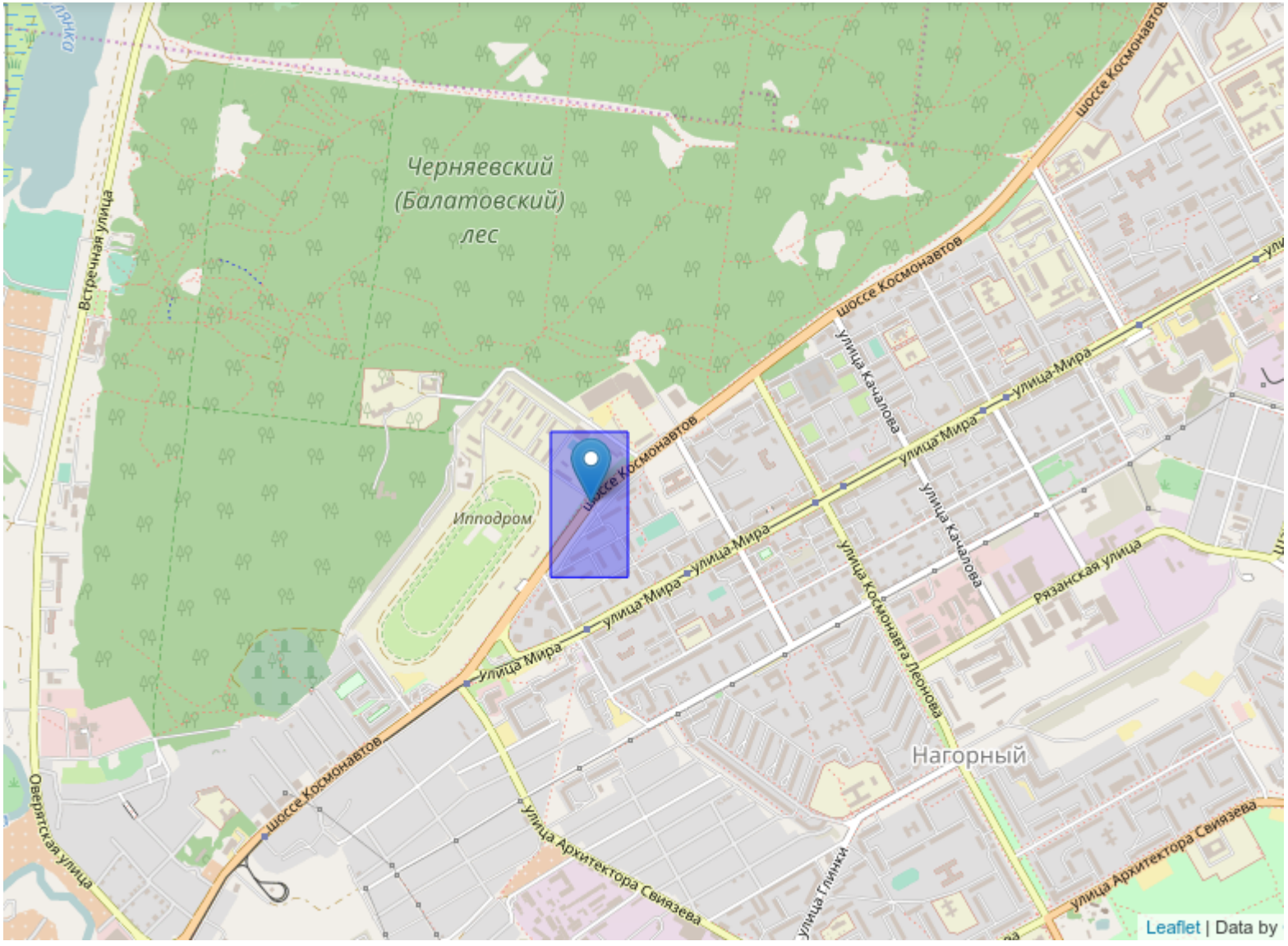
Смещение предсказания

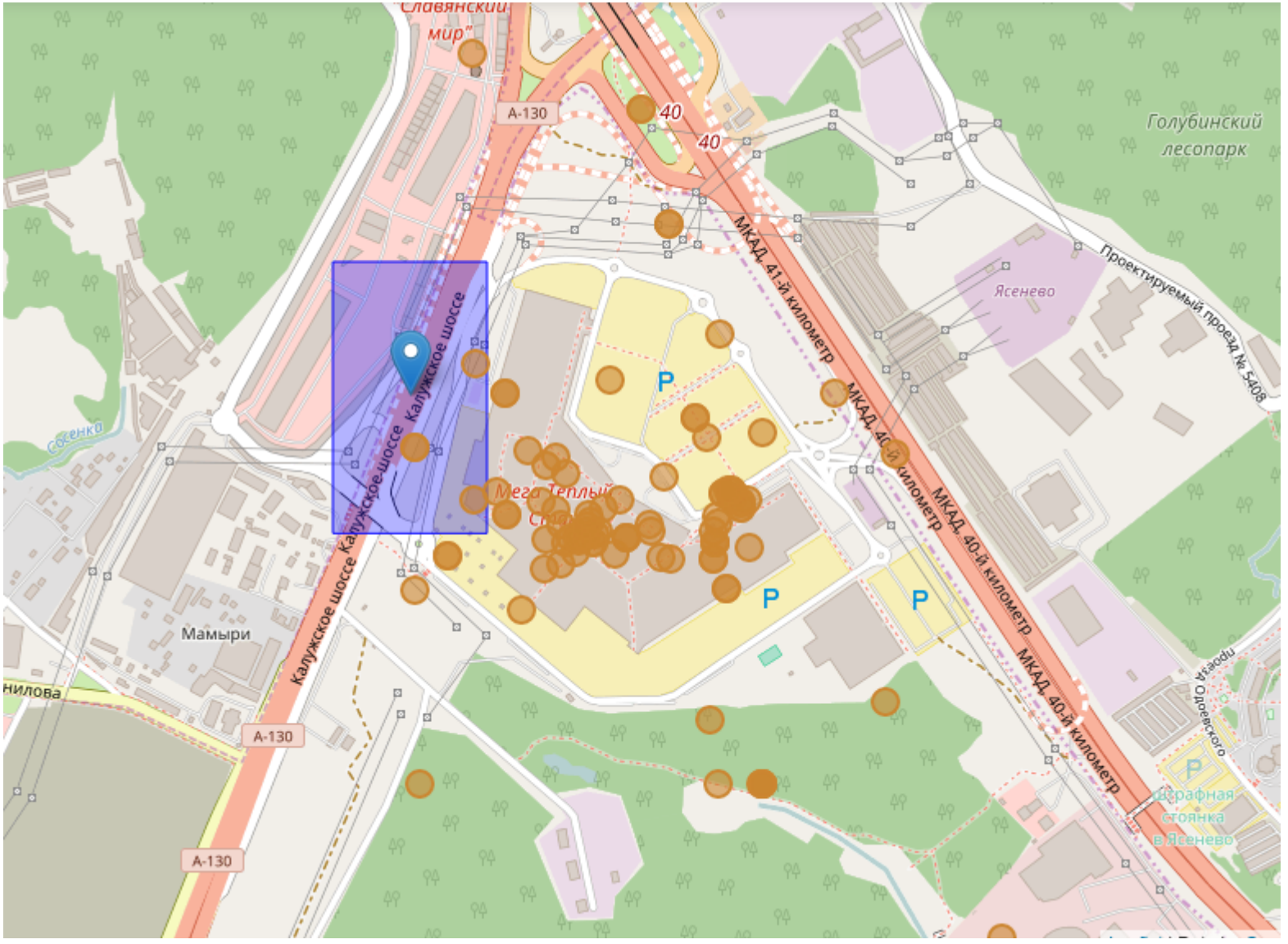
Если на кросс-валидации проверить точность для прямоугольника 0.0025, то она возрастает на 0.02. То есть часто мы не угадываем совсем немного.

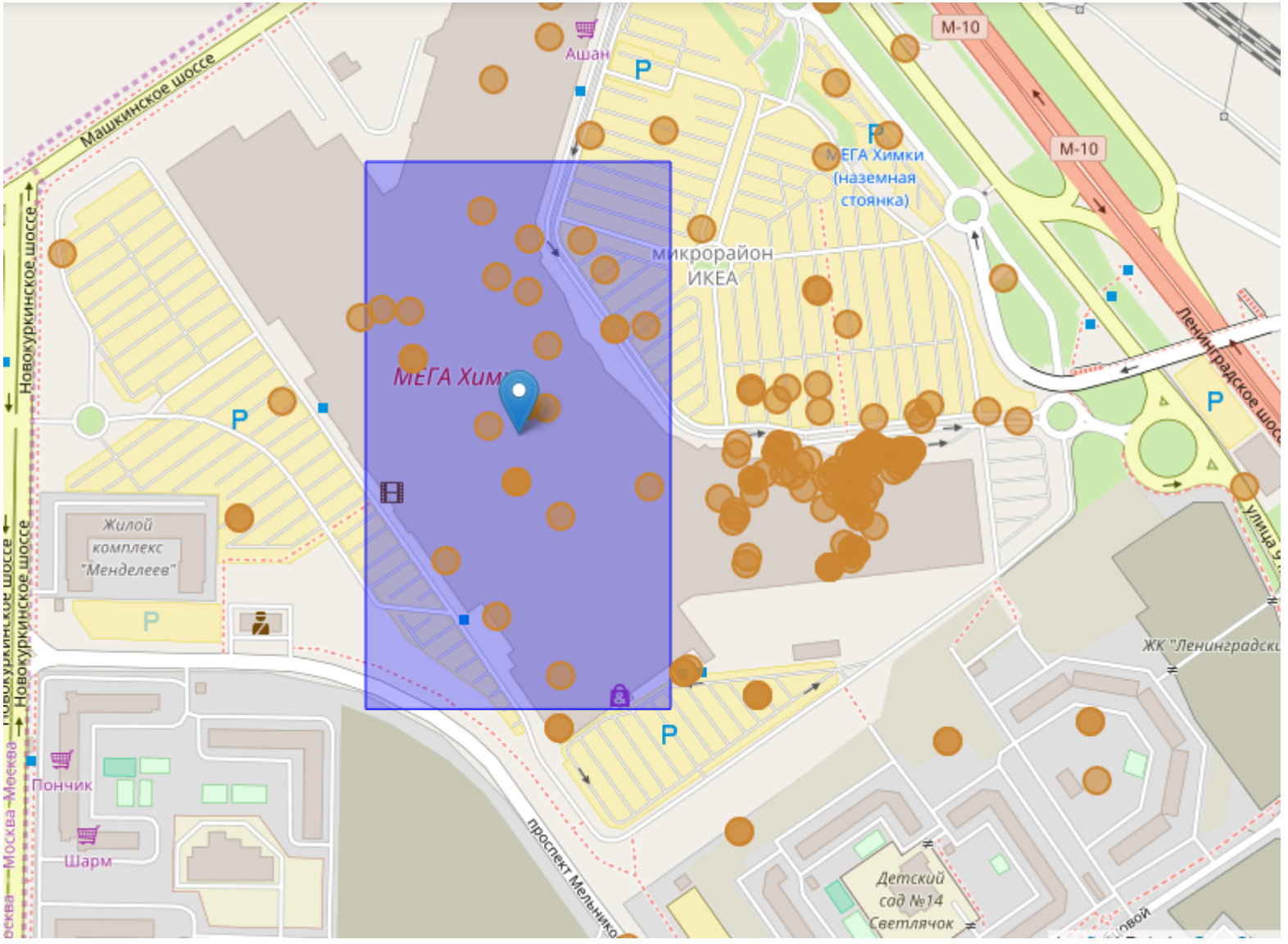

Сокол
Аэропорт
Щукино
Савёловский
Марьина Роща
Сокольники
Октябрьское поле
Хорошёвский
Беговое
Хорошёво-Мнёвники
Красносельский
Тверской
Басманный
Пресненский
Арбат
Москва
Лефортово
Филёвский парк
Таганский
Замоскворечье
Фили-Давыдково
Хамовники
Южнопортовый
Раменки
Гагаринский
Донской
Даниловский
Leaflet | Data by


Жилой комплекс Neva Towers
15 участок ММДЦ "Москва-Сити"
Павильон №2 ЦВК «Экспоцентр»
Павильон №3 ЦВК «Экспоцентр»
Выставочный комплекс на Красной Пресне
Павильон «Форум»
Павильон №7
1-й Красногвардейский проезд
Выставочный переулок
Третье Транспортное кольцо
Выставочная
Международная
Деловой центр
Комплекс «Imperia Tower»
Пресненская набережная
мост Багратион
Пресненский район
Центральный административный округ
Тестовская улица
пристань набережная Тараса Шевченко
Leaflet | Data by Op


Вологодская область
Костромская область
Ярославская область
Иваново
Владимирская область
Пермский край
Кировская область
Свердловская область
Тюменская область
Удмуртия
Набережные Челны
Марий Эл
Татарстан
Тюмень
Курганская область
Челябинская область
Башкортостан
Мордовия
Ульяновск
Пензенская область
Тамбовская область
Липецк
Воронежская область
Оренбургская область
Костанайская область
Западно-Казахстанская область
Волгоградская область
Волгоград
Луганськ
Актюбинская область
Leaflet | Data by OpenStreetMap, un


Черняевский
(Балатовский)
лес
Ипподром
Нагорный
шоссе Космонавтов
Встречная улица
улица Мира
улица Качалова
улица Космонавта Леонова
Рязанская улица
Оверятская улица
улица Архитектора Свиязева
улица Глинки
Leaflet | Data by


Славянский мир"
А-130
40
40
Голубинский лесопарк
МКАД, 41-й километр
Проектируемый проезд № 5408
Ясенево
Калужское шоссе
Калужское шоссе
Мега Теплый Стан
Мамыри
МКАД, 40-й километр
МКАД, 40-й километр
МКАД, 40-й километр
Одоевского
Одоевского
Штрафная стоянка в Ясенево
А-130
А-130


Машкинское шоссе
Новокуркинское шоссе
Ашан
М-10
МЕГА Химки
(наземная
стоянка)
микрорайон
ИКЕА
Ленинградское шоссе
Жилой
комплекс
"Менделеев"
Пончик
Шарм
проспект Мельников
Детский
сад №14
Светлячок


Ашан
Мега Дыбенко
P
A-118
КАД
Р-21
Мурманское шоссе
Свердловское городское поселение

Смещение предсказания

Добавляем вручную три последних.

LB score **+0.002**

Смещение предсказания

Делаем смещение к известным мерчантам, чтобы в прямоугольнике было наибольшее количество мерчантов.

Смещение не более чем на 0.001.

LB score **0.394**

Что можно попробовать

- регрессия, например, расстояние от чекина до мерчанта
- ранжирование
- смещение с учетом магазинов, больших ТЦ

Ссылки

- https://boosters.pro/champ_3 страница соревнования
- https://github.com/CapitanKK/tinkoff-challenge-2 мое решение

Вопросы