# [DAM 小组文档]

数据分析与数据挖掘

1552729 李依璇
1552761 韩乐桐
1652677 吴桐欣

# 目录

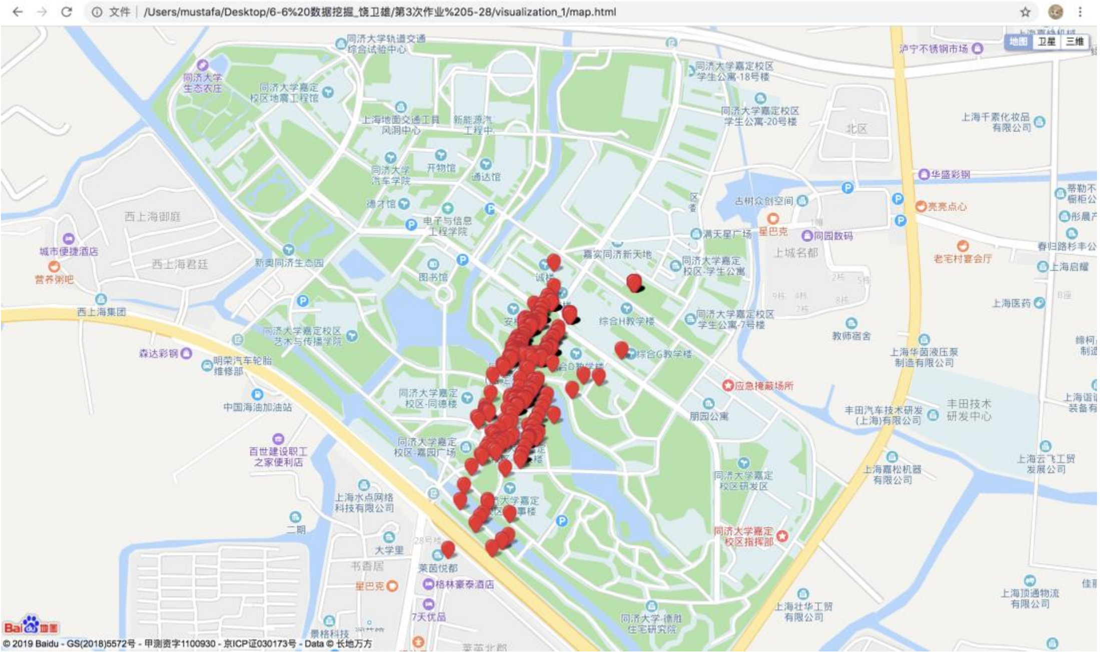
# 一、5 种模型的误差对比

- 组内各模型最优误差列表如下:

| 编号 | 模型 | 中位误差 | 90%误差 | 平均误差 |
|---|---|---|---|---|
| 1 | CNN 回归 | 297 | 681 | 340.57 |
| 2 | CNN 多分类 | 13 | 57 | 31.08 |
| 3 | LSTM 回归 | 424 | 1587 | 822.51 |
| 4 | LSTM 多分类 | 23 | 254 | 91.27 |
| 5 | CNN/LSTM 混合模型 | 11.25 | 30 | 18.44 |
| 6 | Autoencoder/LSTM 混合模型 | 126 | 866.5 | 302.5 |

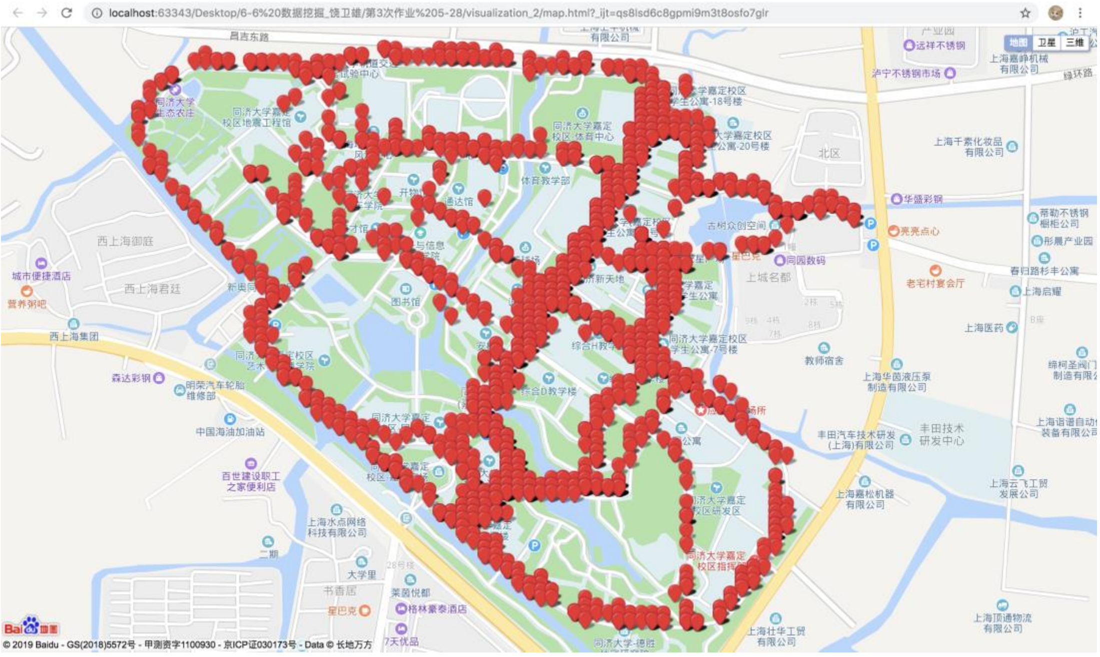
## （1） CNN 回归模型

| 模型 | 中位误差 | 90%误差 | 平均误差 |
|---|---|---|---|
| CNN 回归 | 297 | 681 | 340.57 |



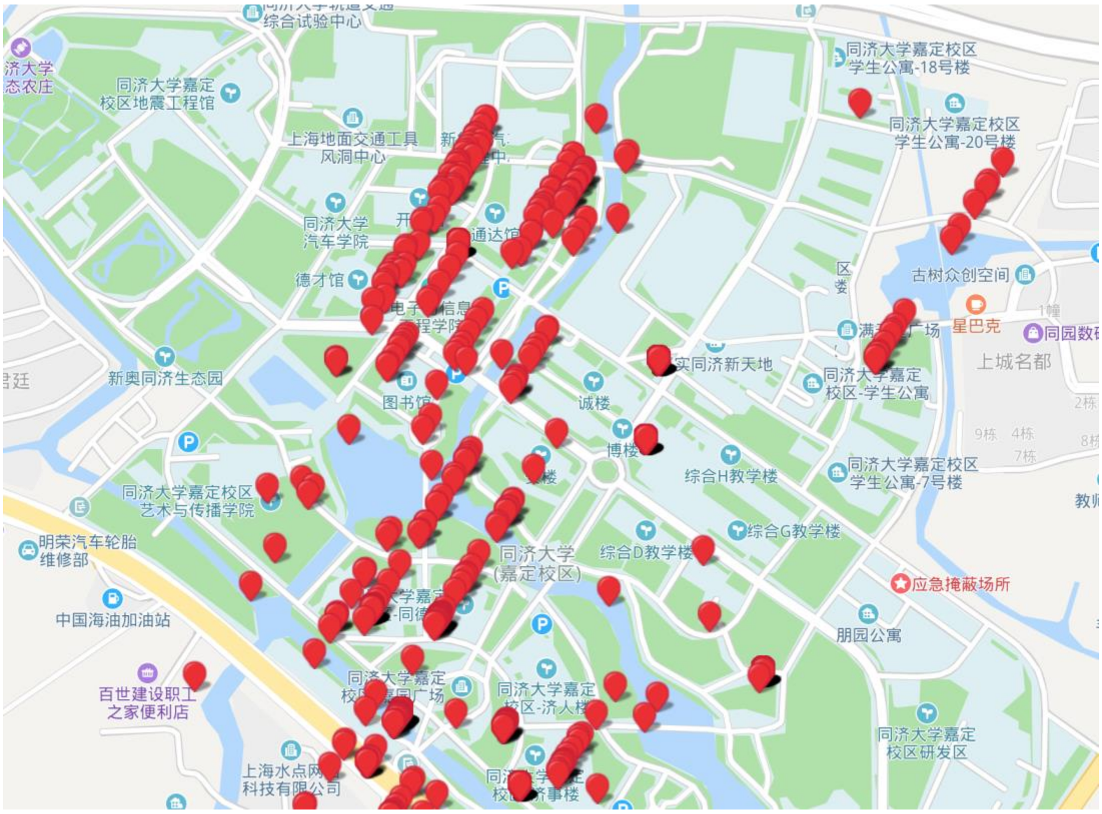
## （2）CNN 多分类模型

| 模型 | 中位误差 | 90%误差 | 平均误差 |
|---|---|---|---|
| CNN 多分类 | 13 | 57 | 31.08 |



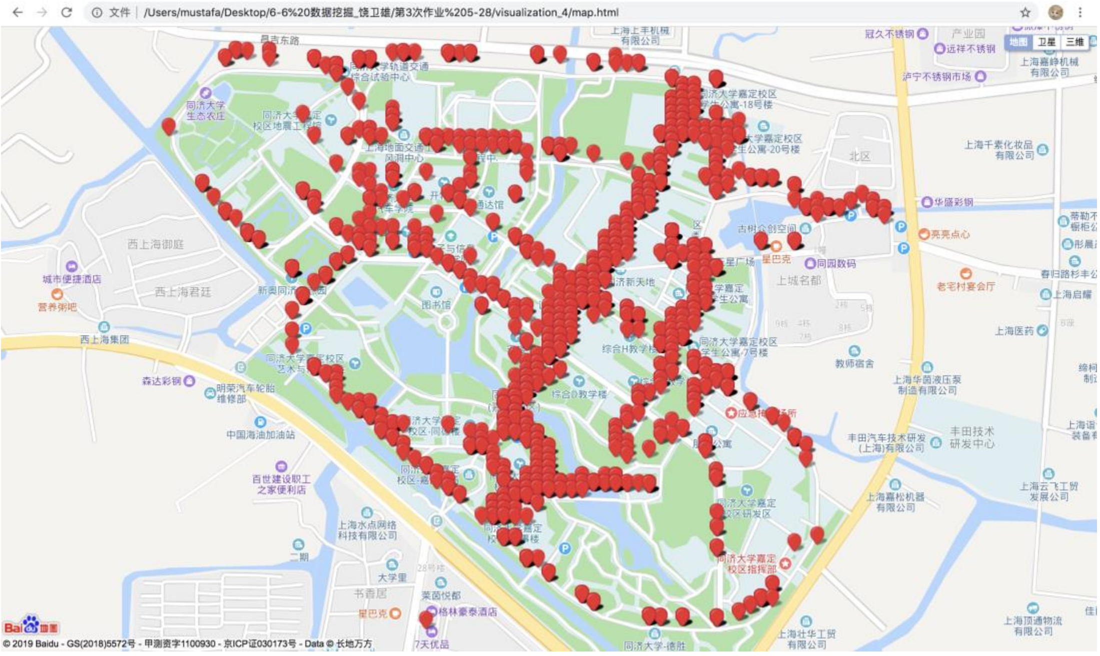
## （3）LSTM 回归模型

| 模型 | 中位误差 | 90%误差 | 平均误差 |
|---|---|---|---|
| LSTM 回归 | 424 | 1587 | 822.51 |



## （4）LSTM 多分类模型

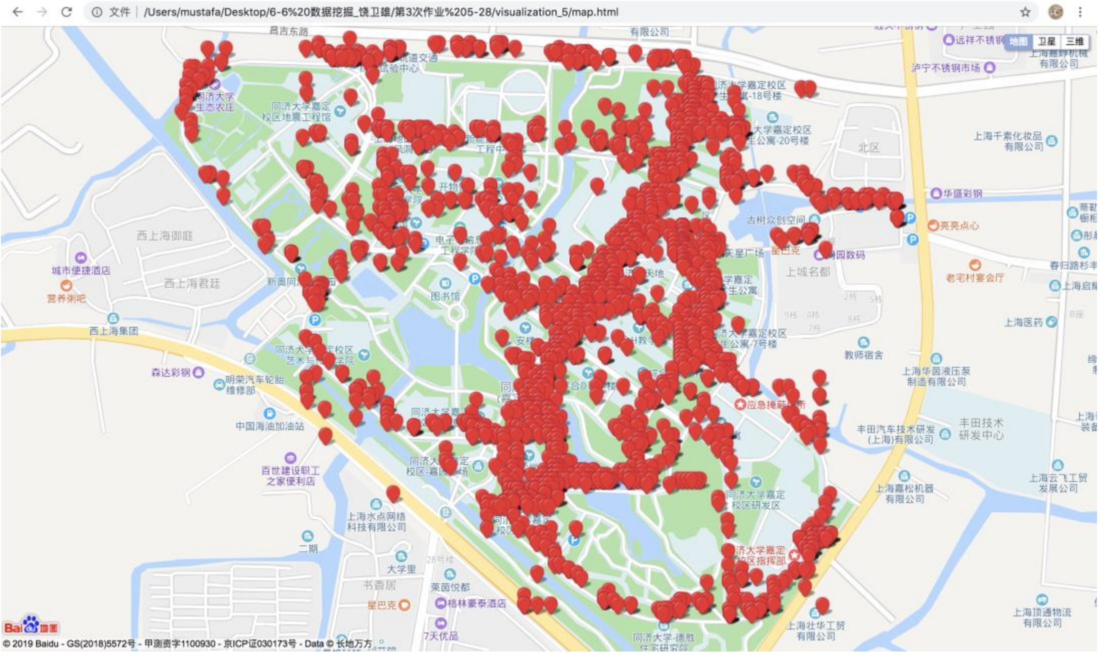
| 模型 | 中位误差 | 90%误差 | 平均误差 |
|---|---|---|---|
| LSTM 多分类 | 23 | 254 | 91.27 |



文件 | /Users/mustafa/Desktop/6-6%20数据挖掘_饶卫雄/第3次作业%205-28/visualization_4/map.html

## （5）CNN/LSTM 混合模型

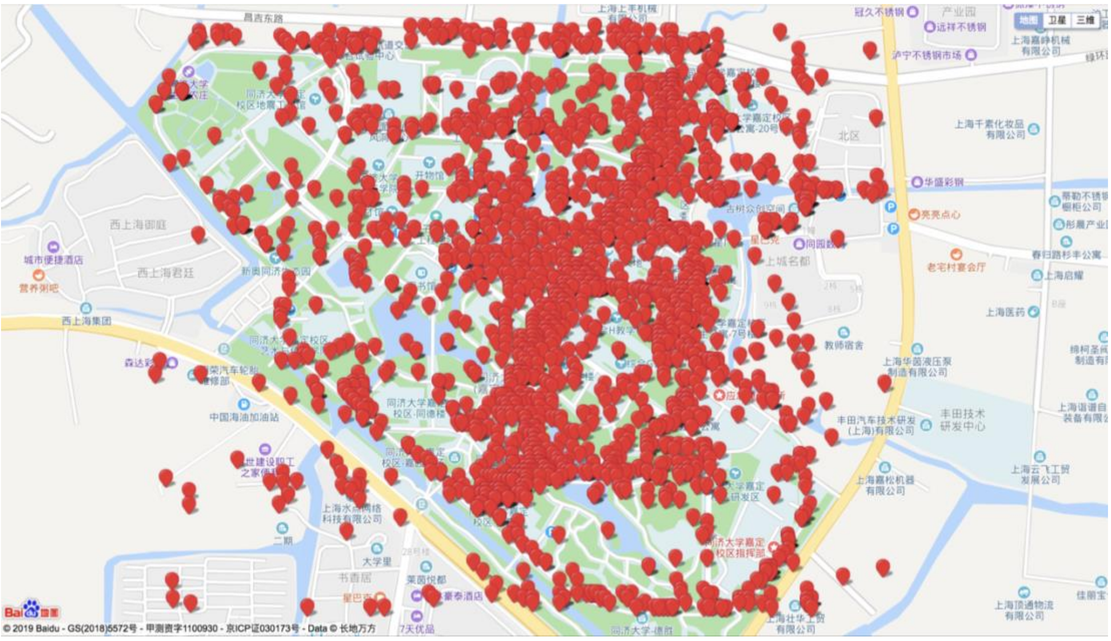
| 模型 | 中位误差 | 90%误差 | 平均误差 |
|---|---|---|---|
| CNN/LSTM 混合模型 | 11.25 | 30 | 18.44 |



文件 | /Users/mustafa/Desktop/6-6%20数据挖掘_饶卫雄/第3次作业%205-28/visualization_5/map.html

## （6）AUTOENCODER/LSTM 混合模型

| 模型 | 中位误差 | 90%误差 | 平均误差 |
|---|---|---|---|
| Autoencoder/LSTM 混合模型 | 126 | 866.5 | 302.5 |



# 二、数据处理方式

通过横向对比组内所有成员的数据处理方式，我们发现不同特征、标签向量格式会在一定程度上影响模型的预测结果。

在 CNN 多分类模型中，若将基站经纬度映射到地图上并使用栅格进行划分，并通过 MR 数据对应的基站进行预测，则 90%误差只有 57m，中位误差能够达到 13m。

在 LSTM 模型中，我们组内分为两种数据处理方法：按照轨迹（traj_id）划分以及按照手机（IMSI）划分，其中共有 70 条有效轨迹，每条轨迹百余条 MR 数据；而共有 4 部手机，每部手机千余条 MR 数据。

定义栅格标签的方式也有两种。一种是直接对栅格排序，例如在给定区域内划分 5000 个格子，则有 5000 个标签，对应的多分类模型有 5000 个分类；另一种是用(x, y)来定义定义一个栅格，x 标签约 130 个，y 标签约 160 个，对应的多分类模型有两个输出层，一个输出 x 标签的概率向量，一个输出 y 标签的概率向量。

由于经纬度数据之间差距特别小，直接使用经纬度数据进行训练会带来很大误差。一个方法是把经纬度转换成 utm 坐标，还有一个方法是只取经纬度的小数部分（因为整数部分都是一样的），并将小数部分放大。

# 三、模型构建

## （1）批标准化 BATCH NORMALIZATION

使用 Batch Normalization 的层通过规范化与线性变换使得每一层网络的输入数据的均值与方差都在一定范围内，使得后一层网络不必不断去适应底层网络中输入的变化，从而实现了网络中层与层之间的解耦，允许每一层进行独立学习，有利于提高整个神经网络的学习速度。

## （2）初始化器 INITIALIZER

normal：正态分布的初始化器

uniform：均匀分布的初始化器

truncated normal：截尾的正态分布的初始化器

- 初始化对网络训练有巨大的影响。(避免在某一层的 forward/backward 中进入饱和区域，拖慢网络训练进程)

## （3）优化器 OPTIMIZER

经过实践，在不同的模型中使用以下几种优化器，可以有比较好的学习效果，收敛更快：Adadelta，Adam，SGD，RMSprop。

## （4）调参经验

- 神经元多并不是很好，会极大降低模型构建速度，并且若向量维度较大，则不会带来相应程度的准确率提高，若向量维度较小反而会使准确率降低；
- batch size 也要适中，太小会导致模型学习速度慢，并且可能有偏重；太大会导致模型泛化

# 四、LS 第 6 题： TM 自编码

## （1）LSTM 自编码模型

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 10, 32)            0
_________________________________________________________________
model_1 (Model)              (None, 10, 2)             4932
_________________________________________________________________
model_2 (Model)              (None, 10, 32)            10444
=================================================================
Total params: 15,376
Trainable params: 15,328
Non-trainable params: 48
_________________________________________________________________
```

## （2）LSTM 编码模型

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 10, 32)            0
_________________________________________________________________
```

```
dense_1 (Dense)            (None, 10, 28)          924
_________________________________________________________________
lstm_1 (LSTM)              (None, 10, 16)          2880
_________________________________________________________________
batch_normalization_1 (Batch (None, 10, 16)           64
_________________________________________________________________
lstm_2 (LSTM)              (None, 10, 8)           800
_________________________________________________________________
lstm_3 (LSTM)              (None, 10, 4)           208
_________________________________________________________________
lstm_4 (LSTM)              (None, 10, 2)           56
=================================================================
Total params: 4,932
Trainable params: 4,900
Non-trainable params: 32
_________________________________________________________________
```

## （3）LSTM 译码模型

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         (None, 10, 2)             0
_________________________________________________________________
```

```
lstm_5 (LSTM)            (None, 10, 4)          112
_________________________________________________________________
lstm_6 (LSTM)            (None, 10, 8)          416
_________________________________________________________________
batch_normalization_2 (Batch (None, 10, 8)          32
_________________________________________________________________
lstm_7 (LSTM)            (None, 10, 16)         1600
_________________________________________________________________
dense_2 (Dense)            (None, 10, 28)         476
_________________________________________________________________
lstm_8 (LSTM)            (None, 10, 32)         7808
=================================================================
Total params: 10,444

Trainable params: 10,428

Non-trainable params: 16
_________________________________________________________________
```

## (4) LSTM 模型

- 手机一

```
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            (None, 10, 2)        0
__________________________________________________________________________________________________
```

```
dense_1 (Dense)                (None, 10, 10)      30          input_1[0][0]
__________________________________________________________________________________________________
batch_normalization_1 (BatchNor   (None, 10, 10)      40          dense_1[0][0]
__________________________________________________________________________________________________
dense_2 (Dense)       (None, 10, 100)       1100        batch_normalization_1[0][0]
__________________________________________________________________________________________________
batch_normalization_2 (BatchNor    (None, 10,
100)       400          dense_2[0][0]
__________________________________________________________________________________________________
dense_3 (Dense)        (None, 10, 256)       25856       batch_normalization_2[0][0]
__________________________________________________________________________________________________
batch_normalization_3 (BatchNor (None, 10, 256)       1024        dense_3[0][0]
__________________________________________________________________________________________________
lstm_1 (LSTM)            (None, 500)          1514000     batch_normalization_3[0][0]
__________________________________________________________________________________________________
batch_normalization_4 (BatchNor (None, 500)          2000        lstm_1[0][0]
__________________________________________________________________________________________________
dense_4 (Dense)          (None, 100)          50100       batch_normalization_4[0][0]
__________________________________________________________________________________________________
batch_normalization_5 (BatchNor (None, 100)          400         dense_4[0][0]
__________________________________________________________________________________________________
dense_5 (Dense)          (None, 256)          25856       batch_normalization_5[0][0]
__________________________________________________________________________________________________
batch_normalization_6 (BatchNor (None, 256)          1024        dense_5[0][0]
__________________________________________________________________________________________________
dense_6 (Dense)       (None, 512)          131584      batch_normalization_6[0][0]
__________________________________________________________________________________________________
```

```
row (Dense)                  (None, 131)        67203       dense_6[0][0]
__________________________________________________________________________________________________
col (Dense)                  (None, 166)        85158       dense_6[0][0]
==================================================================================================
Total params: 1,905,775
Trainable params: 1,903,331
Non-trainable params: 2,444
__________________________________________________________________________________________________
```

- 手机二

```
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            (None, 10, 2)        0
__________________________________________________________________________________________________
dense_1 (Dense)                 (None, 10, 10)       30          input_1[0][0]
__________________________________________________________________________________________________
batch_normalization_1 (BatchNor (None, 10, 10)       40          dense_1[0][0]
__________________________________________________________________________________________________
dense_2 (Dense)      (None, 10, 100)      1100        batch_normalization_1[0][0]
__________________________________________________________________________________________________
batch_normalization_2 (BatchNor (None, 10, 100)      400         dense_2[0][0]
__________________________________________________________________________________________________
dense_3 (Dense)      (None, 10, 256)      25856       batch_normalization_2[0][0]
__________________________________________________________________________________________________
batch_normalization_3 (BatchNor (None, 10, 256)      1024        dense_3[0][0]
__________________________________________________________________________________________________

lstm_1 (LSTM)          (None, 200)          365600      batch_normalization_3[0][0]

__________________________________________________________________________________________________

batch_normalization_4 (BatchNor (None, 200)          800         lstm_1[0][0]

__________________________________________________________________________________________________

dense_4 (Dense)        (None, 30)           6030        batch_normalization_4[0][0]

__________________________________________________________________________________________________

batch_normalization_5 (BatchNor (None, 30)           120         dense_4[0][0]

__________________________________________________________________________________________________

dense_5 (Dense)        (None, 256)          7936        batch_normalization_5[0][0]

__________________________________________________________________________________________________

batch_normalization_6 (BatchNor (None, 256)          1024        dense_5[0][0]

__________________________________________________________________________________________________

dense_6 (Dense)        (None, 512)          131584      batch_normalization_6[0][0]

__________________________________________________________________________________________________

row (Dense)                     (None, 131)          67203       dense_6[0][0]

__________________________________________________________________________________________________col
(Dense)                         (None, 166)          85158       dense_6[0][0]

==================================================================================================

Total params: 693,905

Trainable params: 692,201

Non-trainable params: 1,704

__________________________________________________________________________________________________
```

- 手机三

```
__________________________________________________________________________________________________

Layer (type)                    Output Shape         Param #     Connected to
```

```
==================================================================================================
input_1 (InputLayer)            (None, 10, 2)        0
__________________________________________________________________________________________________
dense_1 (Dense)                 (None, 10, 10)       30          input_1[0][0]
__________________________________________________________________________________________________
batch_normalization_1 (BatchNor (None, 10, 10)       40          dense_1[0][0]
__________________________________________________________________________________________________
dense_2 (Dense)         (None, 10, 100)      1100        batch_normalization_1[0][0]
__________________________________________________________________________________________________
batch_normalization_2 (BatchNor (None, 10, 100)      400         dense_2[0][0]
__________________________________________________________________________________________________
dense_3 (Dense)        (None, 10, 256)      25856       batch_normalization_2[0][0]
__________________________________________________________________________________________________
batch_normalization_3 (BatchNor (None, 10, 256)      1024        dense_3[0][0]
__________________________________________________________________________________________________
lstm_1 (LSTM)        (None, 500)          1514000     batch_normalization_3[0][0]
__________________________________________________________________________________________________
batch_normalization_4 (BatchNor (None, 500)          2000        lstm_1[0][0]
__________________________________________________________________________________________________
dense_4 (Dense)      (None, 100)          50100       batch_normalization_4[0][0]
__________________________________________________________________________________________________
batch_normalization_5 (BatchNor (None, 100)          400         dense_4[0][0]
__________________________________________________________________________________________________
dense_5 (Dense)       (None, 256)          25856       batch_normalization_5[0][0]
__________________________________________________________________________________________________
batch_normalization_6 (BatchNor (None, 256)          1024        dense_5[0][0]
______________________________________________________________
dense_6 (Dense)      (None, 512)        131584      batch_normalization_6[0][0]
______________________________________________________________
row (Dense)                   (None, 131)        67203       dense_6[0][0]
______________________________________________________________
col (Dense)                   (None, 166)        85158       dense_6[0][0]
==============================================================
Total params: 1,905,775
Trainable params: 1,903,331
Non-trainable params: 2,444
______________________________________________________________
```

- 手机四

```
______________________________________________________________
Layor (type)                  Output Shape         Param #     Connected to
==============================================================
input_1 (InputLayer)          (None, 10, 2)        0
______________________________________________________________
dense_1 (Dense)               (None, 10, 10)       30          input_1[0][0]
______________________________________________________________
batch_normalization_1 (BatchNor (None, 10, 10)       40          dense_1[0][0]
______________________________________________________________
dense_2 (Dense)          (None, 10, 100)      1100        batch_normalization_1[0][0]
______________________________________________________________
batch_normalization_2 (BatchNor (None, 10, 100)      400         dense_2[0][0]
______________________________________________________________
```

```
dense_3 (Dense)       (None, 10, 250)      25250       batch_normalization_2[0][0]

____________________________________________________________________________

batch_normalization_3 (BatchNor (None, 10, 250)      1000        dense_3[0][0]

____________________________________________________________________________

lstm_1 (LSTM)       (None, 10, 100)      140400      batch_normalization_3[0][0]

____________________________________________________________________________

batch_normalization_4 (BatchNor (None, 10, 100)      400         lstm_1[0][0]

____________________________________________________________________________

reshape_1 (Reshape)     (None, 1000)         0           batch_normalization_4[0][0]

____________________________________________________________________________

batch_normalization_5 (BatchNor (None, 1000)         4000        reshape_1[0][0]

____________________________________________________________________________

dense_4 (Dense)          (None, 200)          200200      batch_normalization_5[0][0]

____________________________________________________________________________

batch_normalization_6 (BatchNor (None, 200)          800         dense_4[0][0]

____________________________________________________________________________

dense_5 (Dense)        (None, 256)          51456       batch_normalization_6[0][0]

____________________________________________________________________________

batch_normalization_7 (BatchNor (None, 256)          1024        dense_5[0][0]

____________________________________________________________________________

dense_6 (Dense)     (None, 512)          131584      batch_normalization_7[0][0]

____________________________________________________________________________

row (Dense)                    (None, 131)          67203       dense_6[0][0]

____________________________________________________________________________

col (Dense)                    (None, 166)          85158       dense_6[0][0]
==================================================
```

Total params: 710,045

Trainable params: 706,213

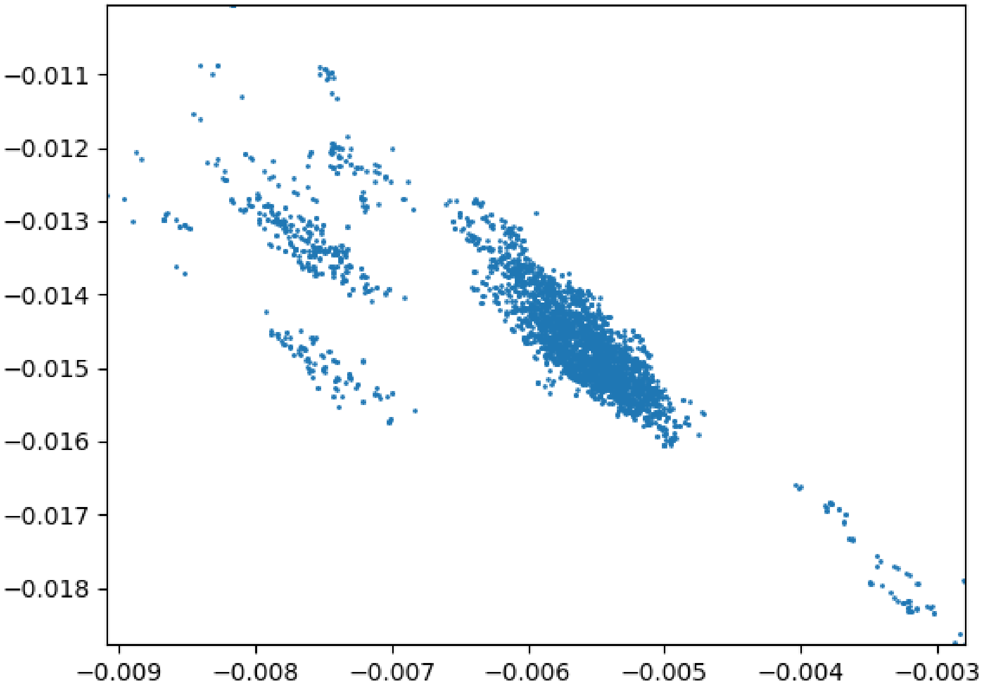
Non-trainable params: 3,832

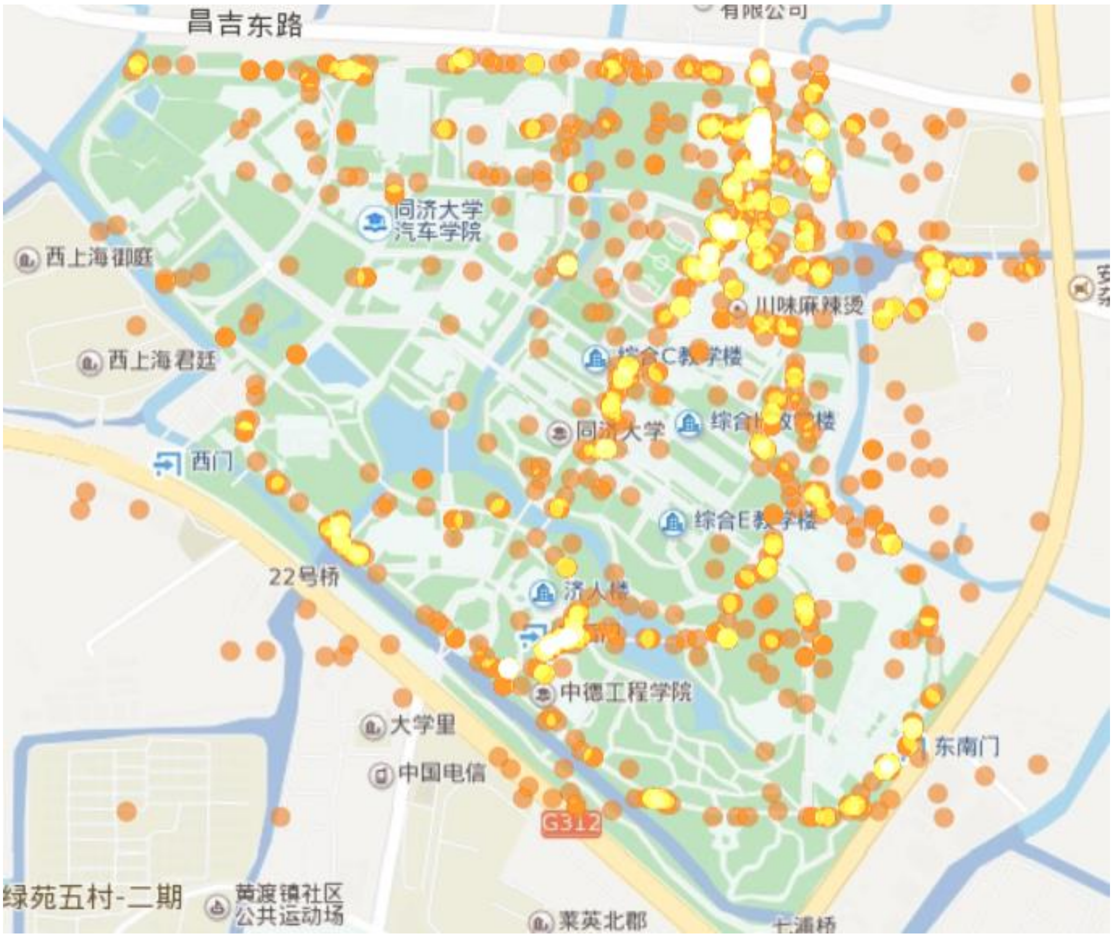
---

## （5）编码的结果可视化

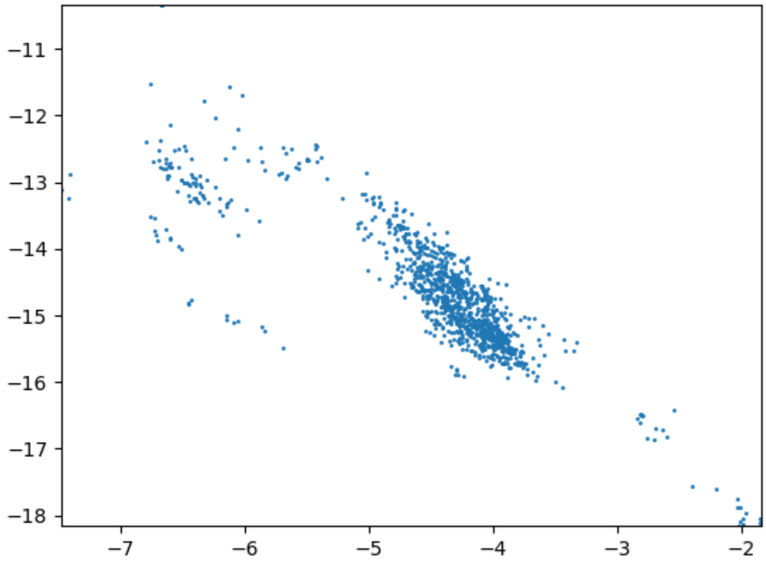
原始的 mr 特征有 32 个维度，经训练后的编码器，降至 2 维进行可视化，结果如下。

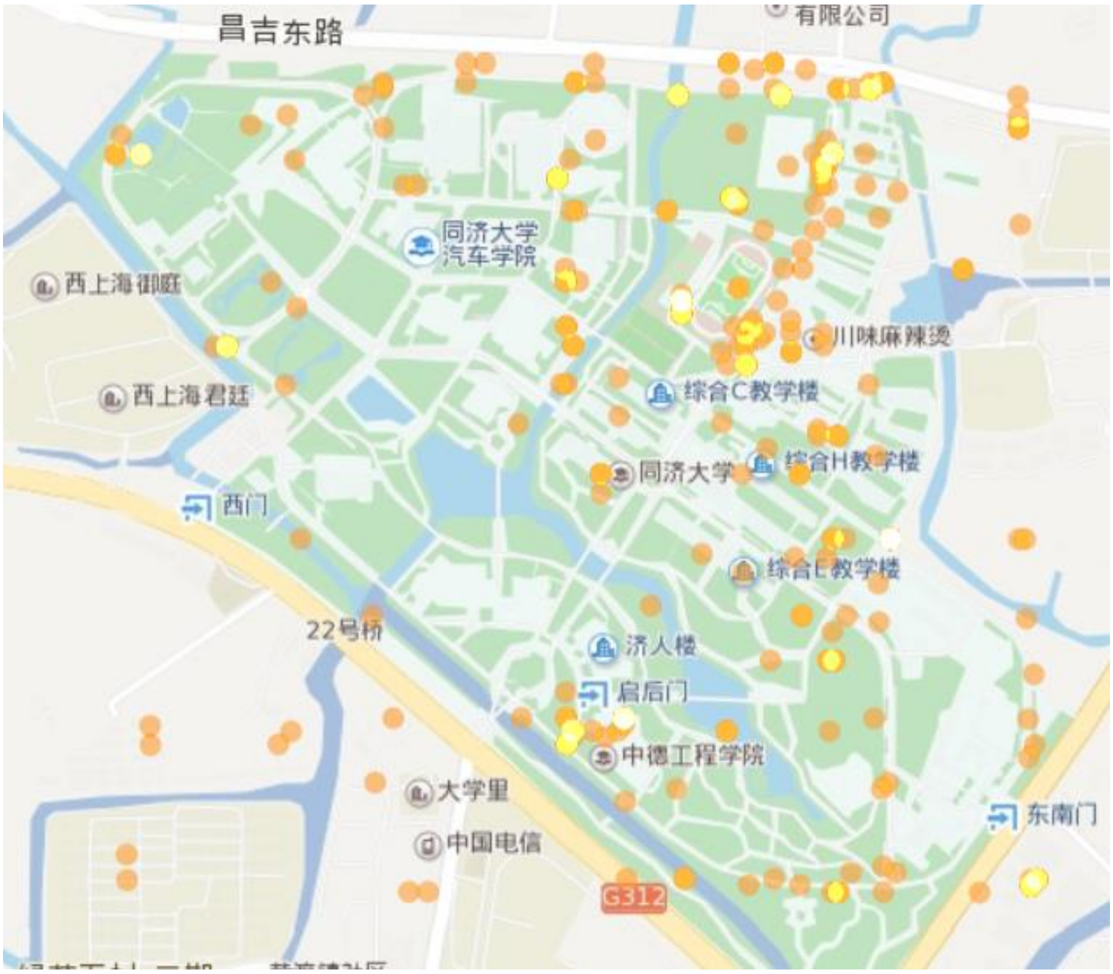
- **手机一**



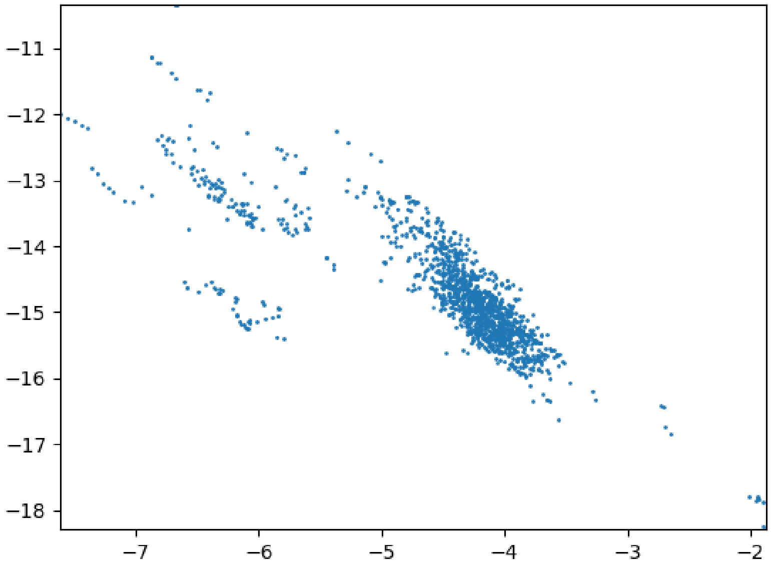

昌吉东路
有限公司
同济大学
汽车学院
西上海御庭
西上海君廷
西门
22号桥
川味麻辣烫
综合C教学楼
同济大学
综合K教学楼
综合E教学楼
济人楼
中德工程学院
大学里
中国电信
G312
东南门
绿苑五村-二期
黄渡镇社区
公共运动场
莱英北郡
七浦桥

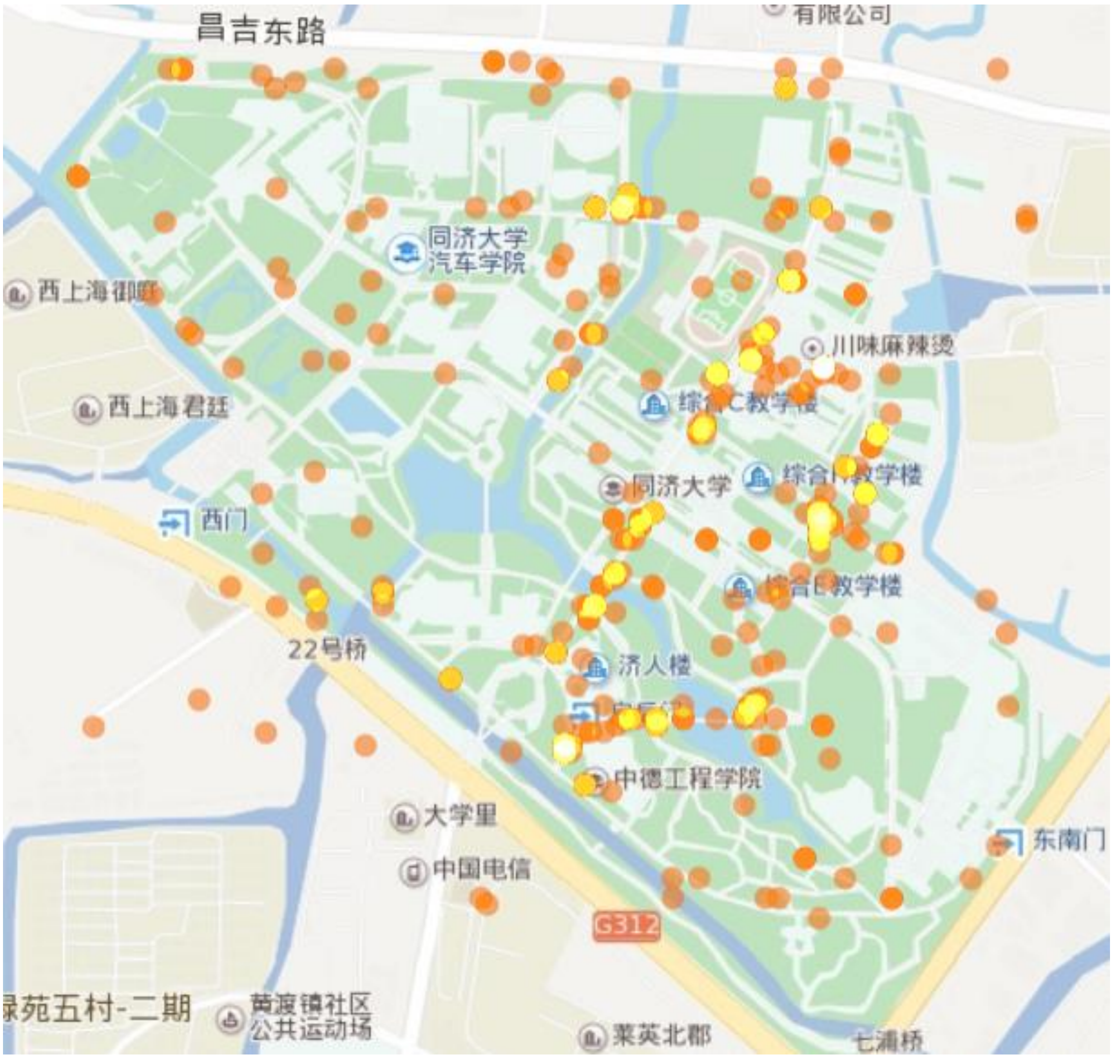
- 手机二


−11
−12
−13
−14
−15
−16
−17
−18
−7
−6
−5
−4
−3
−2

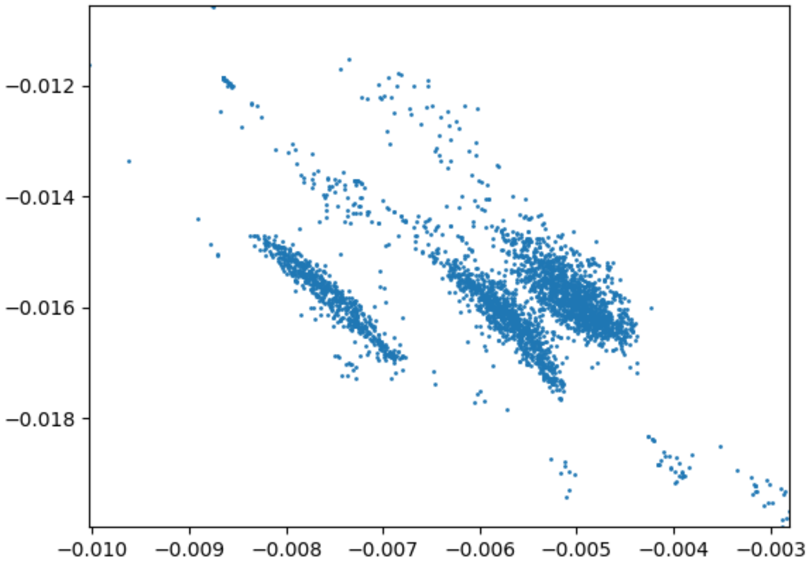

昌吉东路
有限公司
同济大学
汽车学院
西上海御庭
西上海君廷
西门
22号桥
川味麻辣烫
综合C教学楼
同济大学
综合H教学楼
综合E教学楼
济人楼
启后门
中德工程学院
大学里
中国电信
G312
东南门

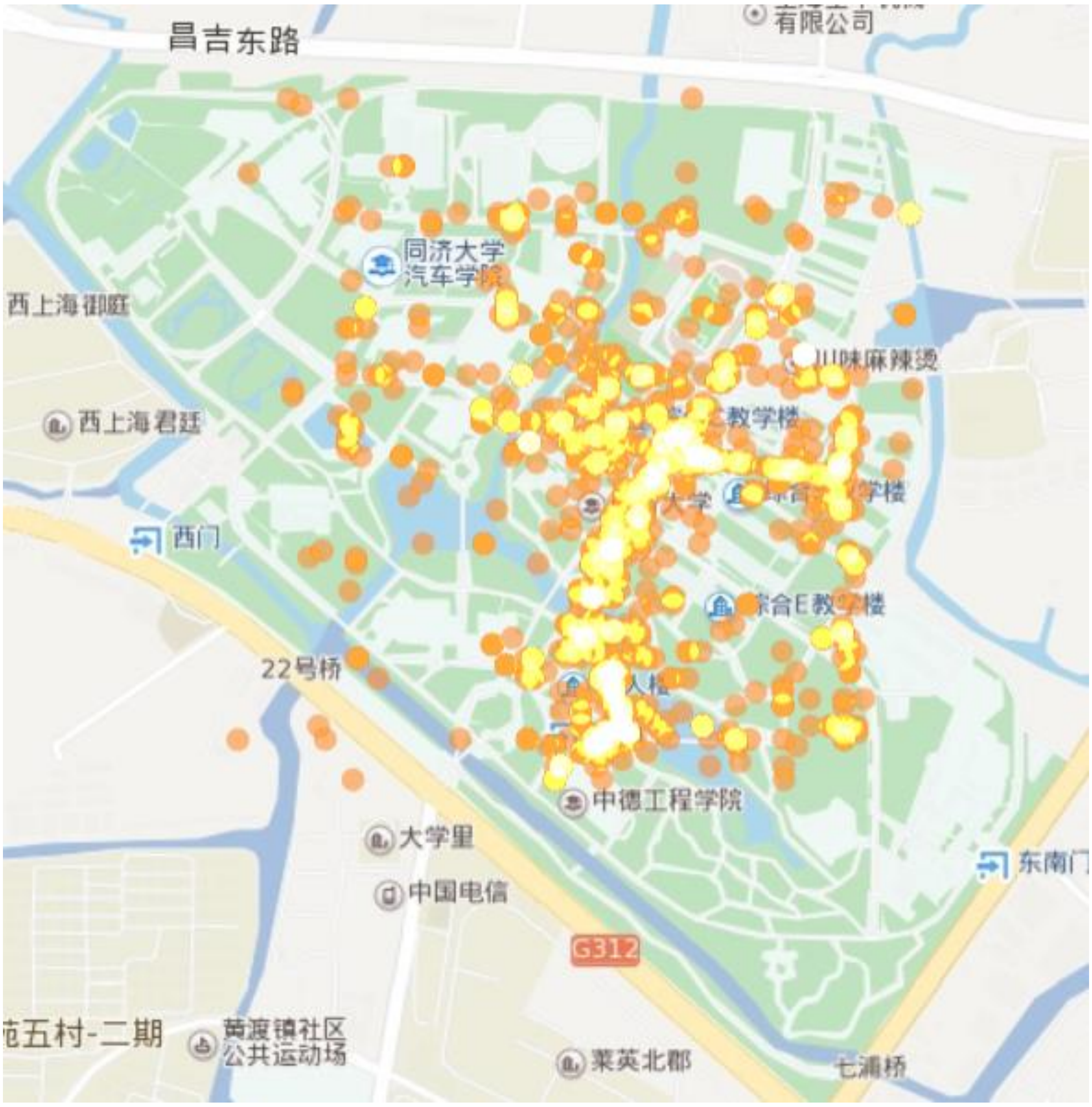
- 手机三


−11
−12
−13
−14
−15
−16
−17
−18
−7
−6
−5
−4
−3
−2


昌吉东路
有限公司
同济大学
汽车学院
西上海御庭
西上海君廷
西门
22号桥
川味麻辣烫
综合C教学楼
同济大学
综合 教学楼
综合E教学楼
济人楼
中德工程学院
大学里
中国电信
G312
东南门
录苑五村-二期
黄渡镇社区
公共运动场
莱英北郡
七浦桥

- 手机四


−0.012
−0.014
−0.016
−0.018
−0.010
−0.009
−0.008
−0.007
−0.006
−0.005
−0.004
−0.003


有限公司
昌吉东路
同济大学
汽车学院
西上海御庭
西上海君廷
西门
川味麻辣烫
教学楼
大学
学楼
综合E教学楼
22号桥
中德工程学院
大学里
中国电信
G312
东南门
五村-二期
黄渡镇社区
公共运动场
莱英北郡
七浦桥