基于 POI 聚类的无桩共享单车数据处理过程

Mobike 为例：

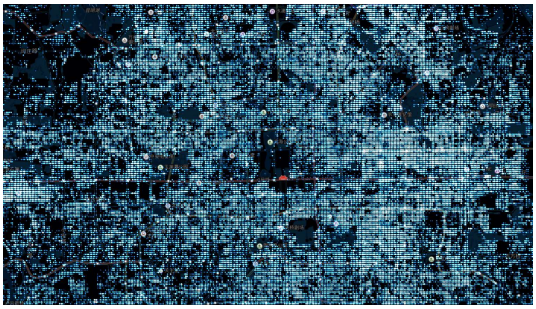
0. 说明

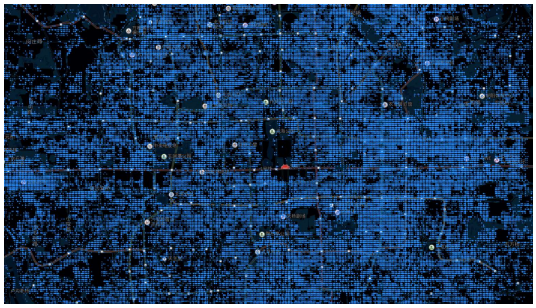
- 全文除特殊说明外，所提到的区域均指 POI 聚类区域

1. 数据分析

1.1 数据可视化

1） 北京单车分布图（初始：0522）





2） 初始区域单车个数热力图（初始）

3） 区域的单车流入流出量之差热力图

4） 区域活跃度热力图

5） 初始区域离散度热力图

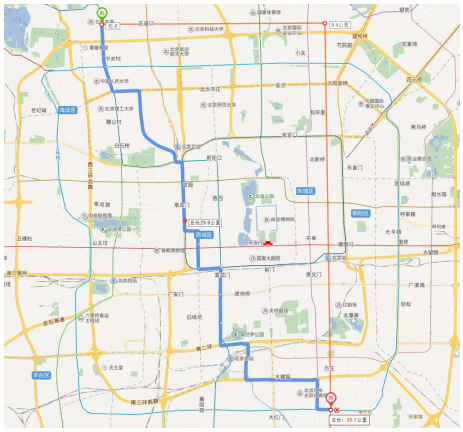
1.2 数据预处理

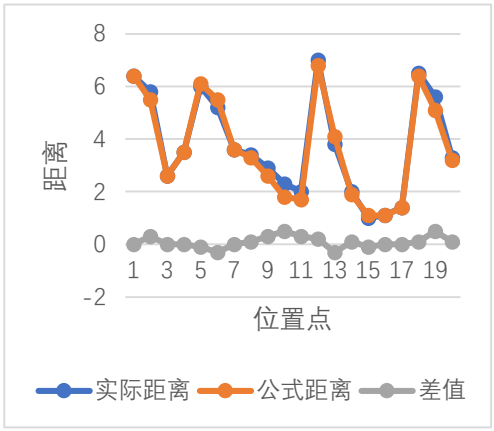
1）轨迹筛选：

a) 时长不小于 2 分钟，距离在 400m 至 20km 之间，速度在 5km/h 至 25km/h 之间（均包含边界）

b) 折线距离（曼哈顿距离）

$$d_{OD}=2R\left(\left|\sin(\frac{\pi}{360}(long_D-long_O))\right|+\left|\sin(\frac{\pi}{360}(lat_D-lat_O))\right|\right)$$

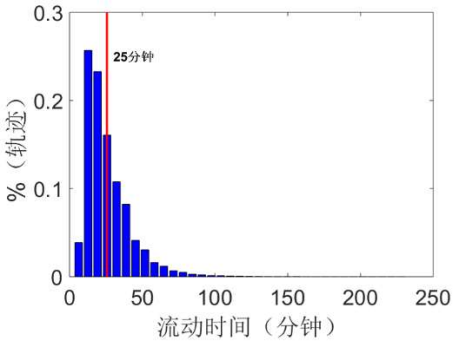


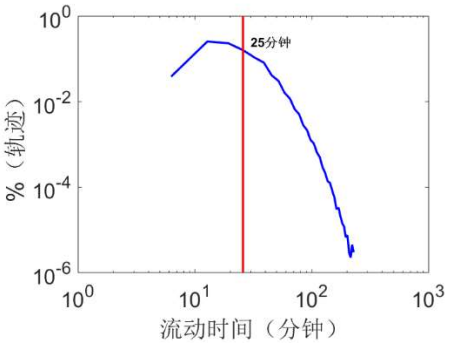


1.3 单车轨迹分析（筛选之后）

1）时间特征

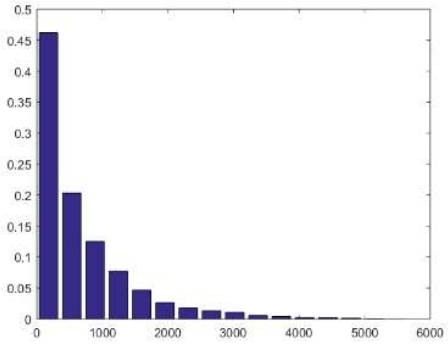
a) 持续时间分布





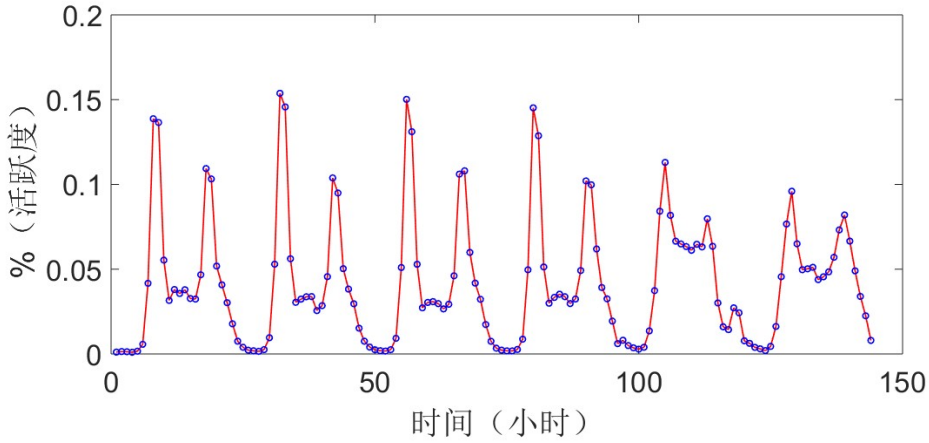
注：工作日与周末

b） 停留时间分布



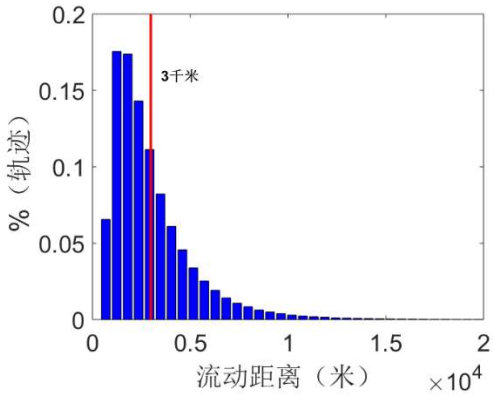
注：工作日，时间单位为分钟，平均停留时间为 10h

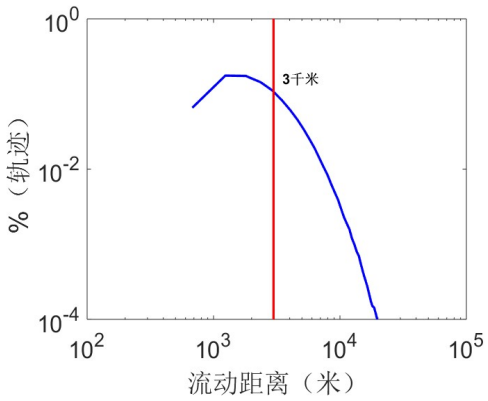
c） 每天不同时间间隔单车的活跃度



2）空间特征

a） 移动步长分布

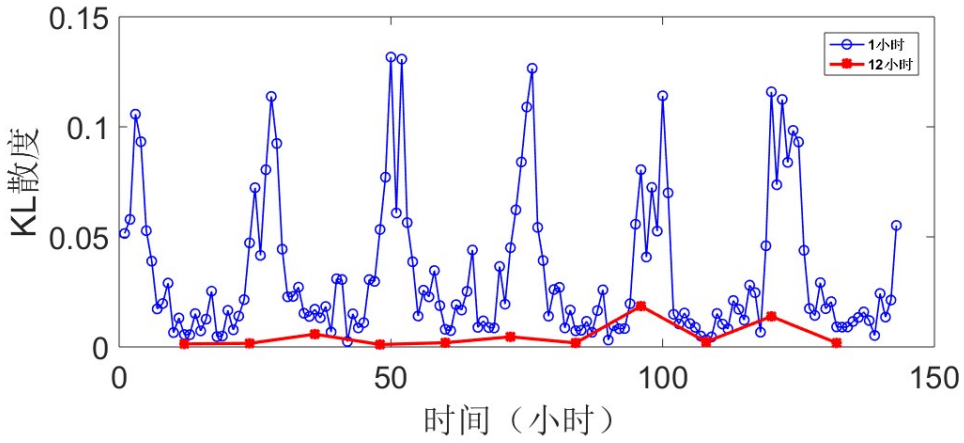




注：工作日与周末

b） 回转半径

c） 每天不同时间间隔单车轨迹距离分布的 KL 散度（1h、12h）（距离范围为 400m 至 20km，按每 400m 等分成 59 份）

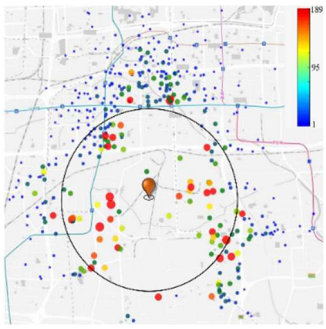


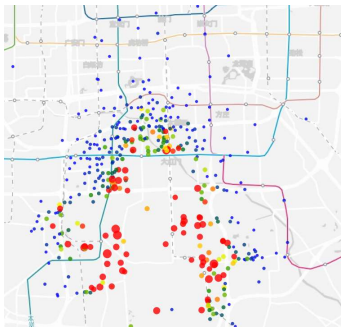
d） 距离影响因素分析

① 区域移动距离分布

② 区域间距离与活跃度的关系

③ 可视化特定区域出发的轨迹





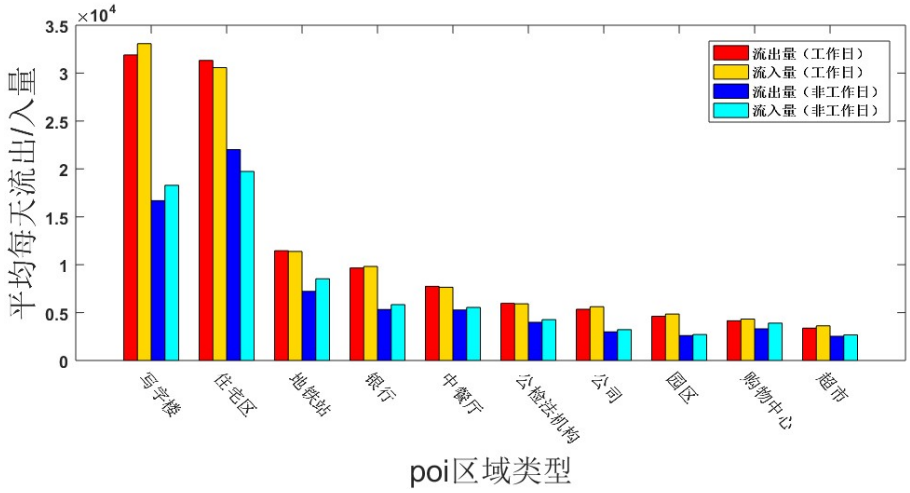
e） 区域活跃度排序

① 将工作日分为四个时间段，非工作日分为三个时间段，统计不同时间段每种 POI 区域的活跃度以及 POI 聚类区域的活跃度（分别统计流出量与流入量）

② 时间间隔 1h，统计每个 POI 区域与 POI 聚类区域的剩余量（流出量与流入量之差），利用 kmeans 方法进行聚类（轮廓系数确定 k 值）

② 统计 POI 区域的平均每天活跃度、流出量、流入量

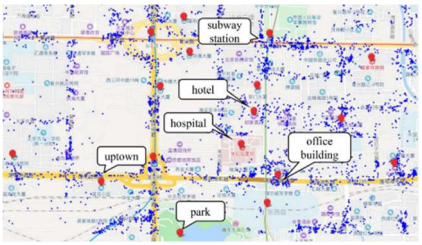
POI 区域类型	POI 区域数量	平均每天活跃度	平均每天每个 POI 区域活跃度
写字楼	4816	64980.50	13.4926
住宅区	7695	61905.50	8.0449
地铁站	306	22843.75	74.6528
银行	1075	19478.50	18.1195
中餐厅	1088	15391.25	14.1464
公检法机构	882	11896.75	13.4884
公司	1034	10956.25	10.5960
园区	757	9469.00	12.5086
购物中心	435	8473.00	19.4782
超市	394	7002.00	17.7716

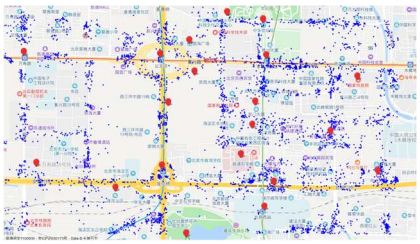


2. 空间聚类方法研究

2.1 基于 POI 聚类

1）原因





2.2 方法比较

1）POI 聚类方法：将单车位置点分配到最近的 POI 区域上，再对 POI 区域进行 kmeans 聚类

2）均匀网格划分

3）方法比较：两种方法分别将整个区域划分成 100、500、1000、5000、10000、20000 份，计算不同的聚类效果评价指标，证明 POI 聚类效果好于网格。

3. 单车空间不平衡原因分析

3.1 容量

1）定义：容量表示某一时刻在某一个区域内的单车数量

2）初始区域容量概率分布

3.2 流通量

1）定义：

①流通量为区域单车的流动过程

②剩余量为区域单车流出量与流入量之差

③活跃度为区域单车流出量与流入量之和

2）区域剩余量概率分布

4）区域活跃度概率分布

5）区域活跃度与容量、POI 种类、POI 个数的关系（散点图、皮尔斯相似度）

3.3 离散度

1）定义：N 为区域单车容量，d_{cp_i} 为区域内第 i 辆单车位置到区域中心的距离，$\overline{d_{cp}}$ 为区域所有单车位置到区域中心的平均距离

$$\varphi = \sqrt{\sum_{i=1}^{N}(d_{cp_i} - \overline{d_{cp}})^2}$$

2）初始区域离散度概率分布

3）不同时间间隔离散度分布的 KL 散度（初始-1d，初始-2d，初始-3d，初始-4d）

4. 可预测性分析

5. 模型

5.1 基于活跃度和距离

思考：

1. 是 POI 区域活跃度还是 POI 聚类区域的活跃度？
2. 统一时间窗口值（1h）
3. 不同时间间隔内的区域活跃度不一致（怎么划分时间间隔：①可以按工作日与非工作日比较明显的几个时间区间进行划分②剩余量聚类），其中区域流出量与流入量也会不同，应该分开考虑。

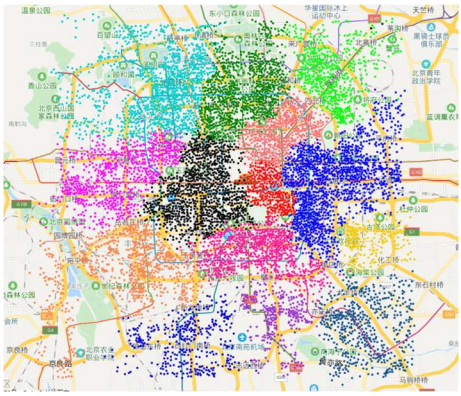
5.2 基于 OD 矩阵

思考：

1. OD 矩阵应该也要分时间段
2. 停留时间服从什么分布，为什么每分钟出行量很平均？

6. 应用

6.1 共享单车的调度

思考：

1. 调度应该分为两种情况：

   ①日常调度：比如时间窗口为一天，一般在人们活动较少的时候进行，比如午夜，将单车过多的区域的单车移动到单车不足的区域，以达到均衡，每个区域可以应对第二天的高峰。

   ②突发情况调度：比如某一个区域即将有大型活动（可预知），该区域的会有大量人群聚集（模型首先应该有考虑这种事件发生的情况，即活动开始前、中、后的人群流向，有多少人多大可能会骑单车过来，有多少人多大可能会骑单车回去，即单车需求的预测，活动前如何调度、活动之后如何调度）。
2. 无论哪一种情况，都需要较精准的需求预测，对于第一种调度，区域判断是否需要调度的方法：设定每一个区域初始单车数（最优），若区域单车数增加（减少）到一定阈值（可以是占初始比例，也可以是区域最大与最小活跃度），单车的调度很难做到实施调度，需要考虑成本，人力等，实时调度也不是很有必要。第二种情况可能更多的考虑疏散人群吧，比较不知道怎么弄。
3. 确定需要调度的区域，考虑成本（简单的根据距离），既要距离最小又要保证调度之后每个区域单车数在正常范围内。
4. 预测区域的状态，过度需求：考虑工作日早晚高峰、周末公园等

6.2 仿真实验

1）变量

①单车损坏率

6.3 实验结果评价

1）复杂网络

①节点、边

②度分布、平均路径长度、网络直径、聚类系数

③网络密度、度中心性、中介中心性、接近中心性