

厦门出租车运行特征分析研究解决方案

数据清洗

本赛题提供了两套数据

- 订单数据（巡游车、网约车）时间范围：**工作日、非工作日、五一假期**
- GPS数据（巡游车、网约车）时间范围：**五一假期**

订单数据的清洗

- 根据以下规则对无效数据进行剔除，并进行坐标系转换，遇到空值用-1进行填充

思考和想法

一、数据清洗

- 1.为了保证 GPS 数据的精确性，需要剔除上下车经纬度位置相差小于 50m 的行程。
- 2.若出现下车时间不在上车时间之后的情况，则以下车时间为为准，按里程数/行程总时间计算上车实际时间，确保上车时间和下车时间的合理性。
- 3.剔除重复记录，同一辆出租车在同一时刻出现多条数据记录，需要剔除重复数据。
- 4.通过行驶速度=行驶路程/行程时间，通过对城市出租车运营数据的分析，一般取城市出租车行驶的平均速度为80km/h，则中间变量车速不在 0-80km/h 范围内的数据属于异常数据。
- 5.订单载客里程、载客时间分别在40 km 和 80 min 以内，将这 2 个值作为载客里程、载客时间的阈值，将超出阈值的出租车计价器与网约车记录剔除。

18/21

- 计算额外基础指标 (eg.每次行程的时间、平均速度)

◦ 关于数据清洗时上下车经纬度间距离的计算方法：参考使用基于经纬度的曼哈顿距离，具体可参考[这篇文章](#)。

- 将网约车订单和巡游车订单数据进行融合，最终数据表的字段如下：

字段名	中文解释	例子	备注
VEHICLE_NO	车牌号	5811	
DEP_TIME	上车时间	2019/7/6 11:51:00 PM	
DEP_LONGITUDE	上车经度	118.10624	CGCS2000坐标系
DEP_LATITUDE	上车纬度	24.590694	CGCS2000坐标系
DEP_AREA	上车地点	同集南路-大润发(集美嘉庚体育馆店)-上车点	
DEST_TIME	下车时间	2019/7/7 12:01:00 AM	
DEST_LONGITUDE	下车经度	118.09434	CGCS2000坐标系
DEST_LATITUDE	下车纬度	24.612946	CGCS2000坐标系
DEST_AREA	下车地点	集美区-爱车义族汽车服务有限公司	
DRIVE_MILE	行驶里程	3.4	单位:km
NOPASS_MILE	空驶公里	1.8	单位:km
CAR_TYPE	车辆类型	1	1-巡游车, 0-网约车

字段名	中文解释	例子	备注
WAIT_TIME	等待时间	206	单位: 秒(猜的)
AVERAGE_V	平均速度	22.666666667	单位:km/h
DRIVE_TIME	行驶时间	540	单位:秒

GPS数据清洗

- 网约车和巡游车数据不做融合处理（袁学长说的原因）

数据清洗补充说明

- 在对五一假期GPS数据进行清洗时，需根据五一期间的订单数据进行配对，将配对不上的剔除
- 主要关注载客和空驶状态，其他状态剔除，我们想利用GPS数据提升五一假期出租车运营指标的计算精度

出行起终点识别技术构建

- 数据源：融合后的五一/工作日/双休日订单数据(巡游+网约车)
- 使用工具：Pandas, 地图Web服务API
- 步骤：
 - 对于每一条有效行程，利用地图API进行匹配起终点，对于没有起终点的重点关注，有起终点的进行纠偏
 - 最后将匹配结果输出成表格
- 期望结果：

字段名	中文解释	例子	备注
VEHICLE_NO	车牌号	5811	
DEP_TIME	上车时间	2019/7/6 11:51:00 PM	
DEP_LONGITUDE	上车经度	118.10624	GCJ-02坐标系
DEP_LATITUDE	上车纬度	24.590694	GCJ-02坐标系
DEP_AREA	上车地点	同集南路-大润发(集美嘉庚体育馆店)-上车点	
DEST_TIME	下车时间	2019/7/7 12:01:00 AM	
DEST_LONGITUDE	下车经度	118.09434	GCJ-02坐标系
DEST_LATITUDE	下车纬度	24.612946	GCJ-02坐标系
DEST_AREA	下车地点	集美区-爱车义族汽车服务有限公司	

出租车运行特征分析

- 数据源：融合后的五一/工作日/双休日订单数据(巡游+网约车)、五一GPS数据(巡游+网约车)
- 使用工具：Pandas、Sklearn、ArcGIS、Matplotlib、Folium
- 思路：

提取载客点空间特征->确定热区->计算整体出租车运营指标->计算热区运营指标->时空对比/整体-热区对比(作图)->结合政策/土地用途->得出结论

- 载客点空间特征分析（载客点）可考取不同时间对比分析，得到时空特征；抓典型
 - 上车点的地图匹配与DBSCAN聚类：分析乘客上车点的分布情况，得知总体趋势
 - 基于信息熵的出租车载客点分布均衡情况分析：按照厦门行政区划的划分方式对每一块区域进行信息熵计算，将得到的每个信息熵归一化处理，得到 J_i , i 表示区域编号， J 值越小，说明载客点分布越不均匀，反之，则说明该区域载客点分布均衡
 - 中位数中心：由于中位数中心不容易受到极值影响，它可表示最“热”的载客区域
 - 核密度分析：根据核密度函数，分析载客区域中的密度分布，进而得出上车热点区域
 - 动态热力图分析：可分析载客热点区域的类型，是偶发性热点还是持续性热点
 - 标准差椭圆：根据椭圆的长轴的旋转情况，可分析载客点在空间的分布倾向性，即载客点是否是东西/南北向分布，亦或者是否沿着某一条街道分布等
 - 凸壳：可分析乘客的出行范围
 - 关于热点区域提取说明：由于ArcGIS只能提取计算区域的边界，这并不是热点区域边界。所以可考虑使用渔网或者手工框选热区的方法(软件右下角有鼠标当前位置的经纬度显示)
 - 基于信息熵的出租车载客点分布均衡情况分析-补充说明：将厦门市分成 r 个区域，设变量 μ_i 表示载客点随机落入第*i*个区域，其概率为 $p(\mu_i)$ ， i 的取值为0, 1, ..., $r - 1$ 。则出租车载客点的信息熵 H 为：

$H = \sum_{i=0}^{r-1} p(\mu_i)I(\mu_i) = -\sum_{i=0}^{r-1} p(\mu_i)\log_b p(\mu_i)$ 。为了比较方便，将信息熵做归一化处理 $J = \frac{H}{H_m}$ ，其中 H_m 为最大信息熵，即所有载客点平均落入各个区域，也就是说，它们落入各个区域的概率相等。

- 上车点的地图匹配与DBSCAN聚类补充说明：minPts由噪声点和聚类的肘部法则确定，eps由k近邻距离确定。

- 出租车通用指标分析 可考虑选取不同时间对比分析，得到每个指标的时变特征；抓典型

- 厦门市出租车整体运营情况 需要分工作日/双休日/五一假期进行对比分析

- 日均载客时间：出租车司机单日平均劳动强度

$$T = \frac{\sum t_i}{m}$$

t_i 表示第 i 辆车单日总载客时间； m 为单日所有运营车辆的数量。 $i=0,1,2,\dots,m$

- 日均载客里程：运营车辆对道路资源的占用情况

$$L = \frac{\sum l_i}{m}$$

l_i 表示第 i 辆车单日总载客里程； m 为单日所有运营车辆的数量。 $i=0,1,2,\dots,m$

- 日均载客次数：司机劳动强度+区域居民出行需求

$$C = \frac{\sum N_i^{on}}{m}$$

N_i^{on} 表示第 i 辆车单日载客次数； m 为单日所有运营车辆的数量。 $i=0,1,2,\dots,m$

- 小时订单变化：反映居民在不同时间段的出行需求 (可分巡游车/网约车/整体)
- 平均速度：反映整体/不同时段道路拥堵情况
- 空驶里程率：反映城市出租车运行效率。若该值偏高，说明车辆运营效果不理想；若该值偏低，说明车辆运营效果较为理想。（可能要分巡游车/网约车，老学长说的原因）

$$\text{空驶里程率} = \frac{\text{车均空驶里程}}{\text{车均总行驶里程}}$$

- 厦门市出租车载客热点区域特征提取

- 小时订单变化：反映热区居民出行需求随时间变化特征
- 热区订单占总订单比：反映热区整体出行需求
- 热区载客出行OD分布：什么地区的人经常去热区，热区的人又喜欢去哪里（进出热区都要算）
- 热区载客出行平均时间/里程：对热区载客出行OD分布的量化计算，反映以热区为中心的居民活动范围（计算从热区出去的）
- 平均速度：热区路面拥堵情况。

- 对于平均速度的说明：由于工作日/双休日没有车辆GPS数据，故在计算热区平均速度时，将订单OD均在热区范围或者O在热区范围，D在热区附近的订单纳入计算，做个近似。对于五一假期，由于有GPS数据，可以根据轨迹和热区范围进行截取。
- 热区载客出行平均时间/里程：对于平均载客时间：从热区出发的所有订单的载客时间求平均；对于平均载客里程：从热区出发的所有订单的载客里程求平均

景区周边出租车运行特征分析

- 数据源：融合后的五一/工作日/双休日订单数据(巡游+网约车)

- 使用工具：pandas、ArcGIS

- 思路：

- 计算景区周边载客点聚集的范围和强度

- 最大点密度：以景区为圆心，载客点密度最大的圆。反映载客点分布范围
- 最大点密度半径：最大点密度圆的半径。反映载客点聚集强度

- 最大点密度计算方法：密度的计算参考了人口密度的计算方法，放到这里也就变成了：载客点数/圆面积。根据每次半径增大时所纳入的载客点数，如果跟上一次比，纳入的载客点下降了，并且少于10个，说明现在圆的边界载客比较稀疏了，可以停止计算了。如果少于10个后又再次增大，少于10个影响不大，还是稀疏的；如果大于10个，说明可能触及另外一个载客热点了

- 出租车泊位设置方案

- 目标：最大覆盖载客需求点
- 步骤：根据最大点密度半径框选矩形区域，建立渔网 -> 建立泊车位选择模型 -> 求最优解（利用遗传算法/粒子群算法等优化方法）
- 模型：

$$\max Z = \sum_{i \in I} \sum_{j \in J} h_i y_{ij} \quad (1)$$

参数约束

$$\sum_{j \in J} c_j k_j x_j \geq \sum_{i \in I} h_i y_{ij} \quad (2)$$

$$\sum_{j \in J} c_j k_j x_j \leq \sum_{i \in I} h_i \quad (3)$$

$$\sum_{j \in J} x_j = P \quad (4)$$

$$x_j = 0, 1 \quad \forall j \in J \quad (5)$$

$$y_{ij} = \begin{cases} 1 & d_{ij} < S_{min} \quad \forall i \in I, j \in J \\ f(d_{ij}) & S_{min} \leq d_{ij} \leq S_{max} \quad \forall i \in I, j \in J \\ 0 & d_{ij} > S_{max} \forall i \in I, j \in J \end{cases} \quad (6)$$

$$y_{ij} \leq x_j \quad \forall i \in I, j \in J \quad (7)$$

$$\sum_{\forall j \in J} y_{ij} \leq 1 \quad (8)$$

- 式(1)表示停靠站覆盖的出行需求最大
- 式(2)表示各停靠站所提供的运力能够满足高峰小时乘客前往候车点候车乘车的需求
- 式(3)表示各停靠站运力不会超过高峰小时产生的出行需求
- 式(4)表示建设的停靠站数目等于预期数目
- 式(5)表示 x_j 为决策变量。0-未建设；1-已建设
- 式(6)表示某个需求点*i*前往泊车点*j*乘车的比例跟泊车点*j*的覆盖范围有关。如果需求点与泊车点的距离小于泊车点覆盖范围，则一定会去；如果在最小覆盖范围和最大覆盖范围之间，则有可能去；如果超过最大覆盖范围，则不会去。其中 $f(d_{ij}) = \frac{S_{max}-d_{ij}}{S_{max}-S_{min}}$
- 式(7)表示出行者若处于停靠站服务范围内，则考虑前往停靠站乘坐出租车，受到停靠站建设位置的约束。
- 式(8)表示若处于停靠站服务范围内，每一个需求点的出行者全部或者部分选择前往停靠站乘坐出租车。

参数说明和取值

参数名	解释	取值
I	出行需求点集合，每一个需求点 $i \in I$	将渔网的每个中心点作为需求点 <i>i</i> （作为代表）。I为出行需求点 <i>i</i> 的集合。将每块渔网的需求总量看作出行需求合集。
J	停靠站候选点集，每一个停靠站 $j \in J$	根据实际需求确定个数和位置，需要考虑用地/路面拥堵情况等，即结合出租车运行特征
x_j	是否在 <i>j</i> 建造了停靠站？	0-没有建造；1-已经建造了
k_j	停靠站 <i>j</i> 的高峰小时泊位周转率	根据实际情况确定
h_i	需求点 <i>i</i> 在高峰小时的出行需求	<i>i</i> 所在的渔网载客点计数。
y_{ij}	表示需求点 <i>i</i> 产生的高峰小时出行需求，选择前往停靠站 <i>j</i> 的比例	与停靠站 <i>j</i> 与需求点 <i>i</i> 的距离有关
P	计划修建的停靠站数量	根据实际情况确定
d_{ij}	需求点 <i>i</i> 到停靠站 <i>j</i> 的距离	可考虑曼哈顿距离
S_{min}	停靠站 <i>j</i> 的最小服务范围	取300m
S_{max}	停靠站 <i>j</i> 的最大服务范围	取500m
c_j	停靠站 <i>j</i> 的泊位数量	一般取1-5个

期望的效果(举例)

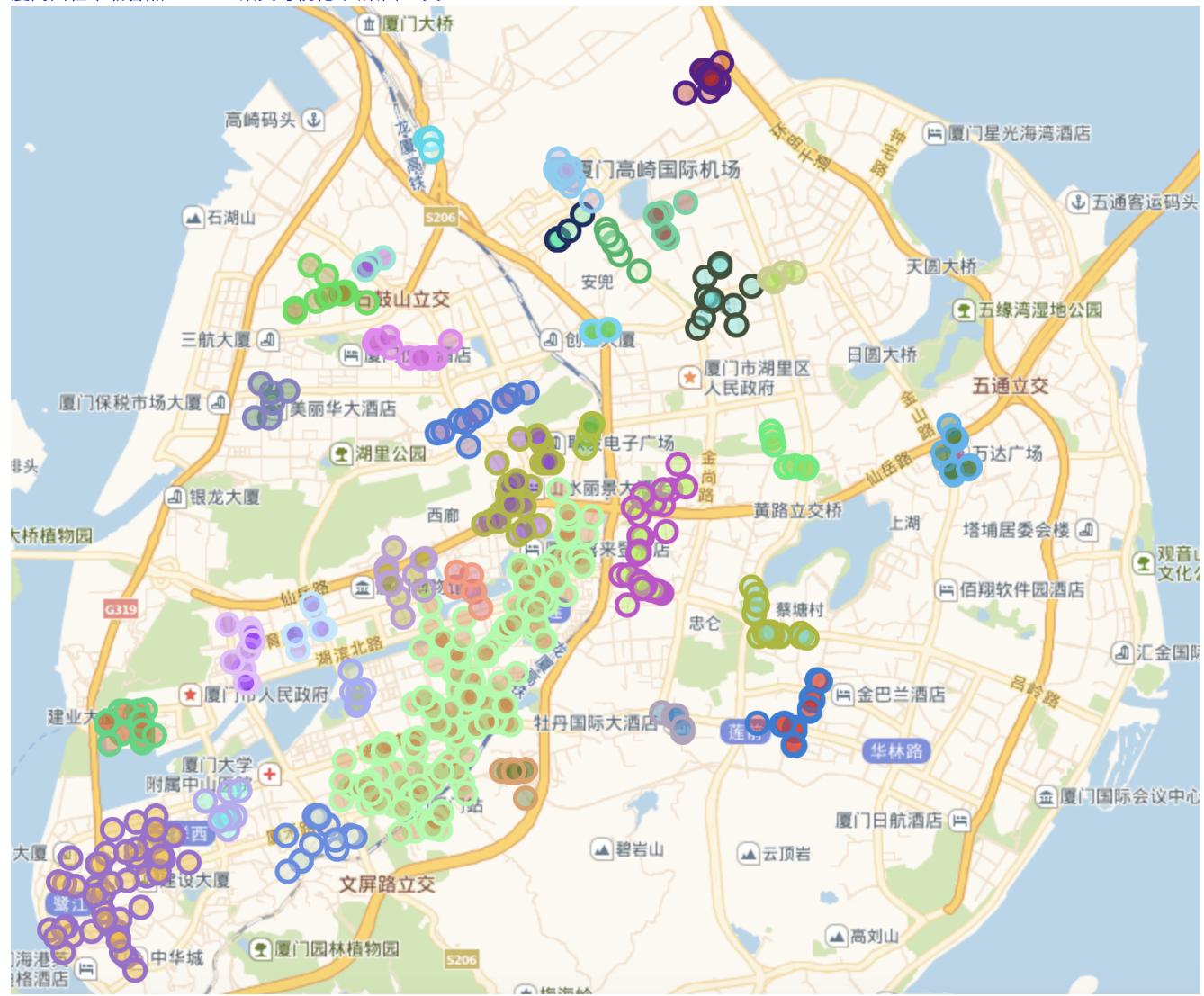
序号	停靠站建设数量	泊位数量/个	服务能力(目标函数值)	占全部出行需求比例/%
1	20	54	2160	31.8
2	40	104	4160	51.9
...

可视化演示网址

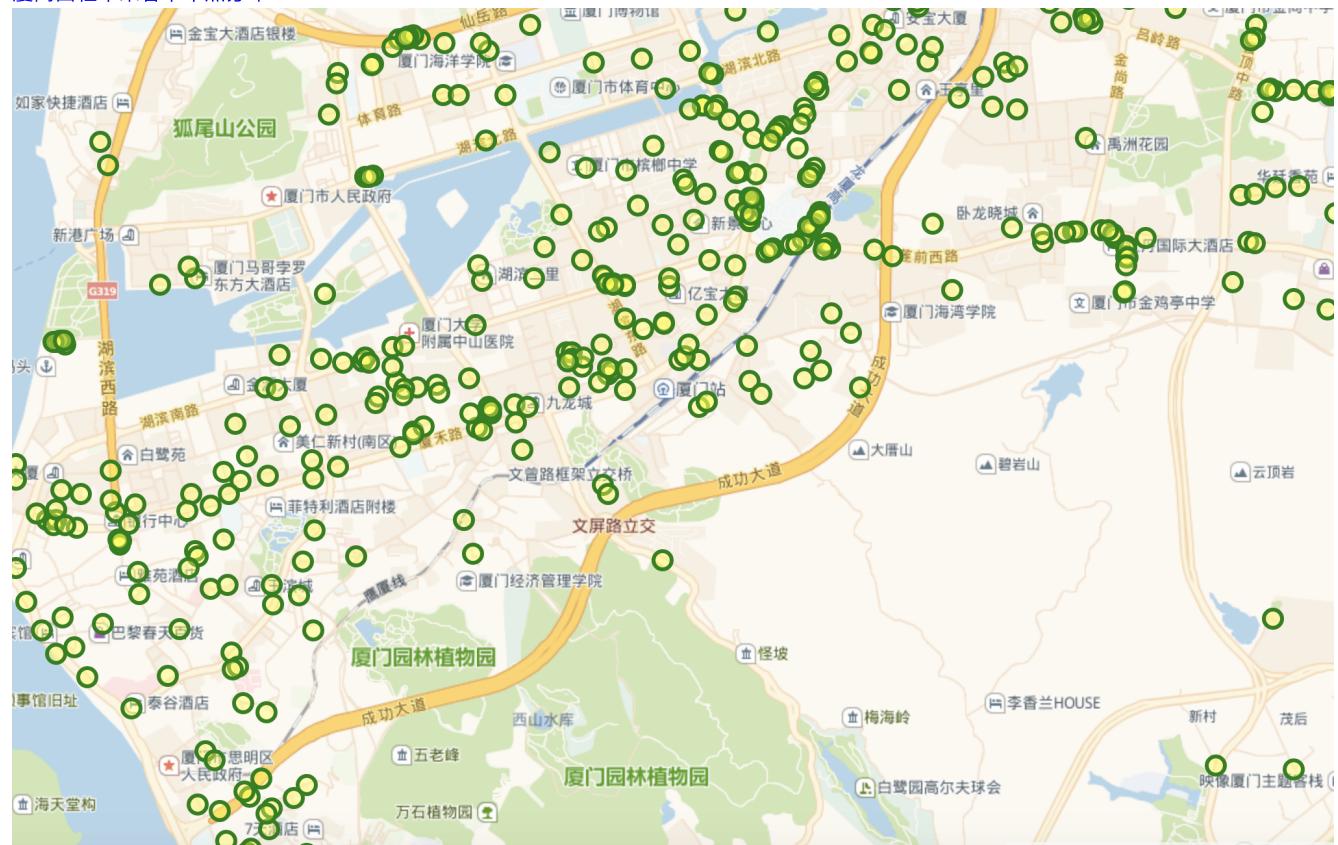
- 厦门出租车乘客分区域计数



- 厦门出租车载客点DBSCAN聚类可视化-共聚出35类



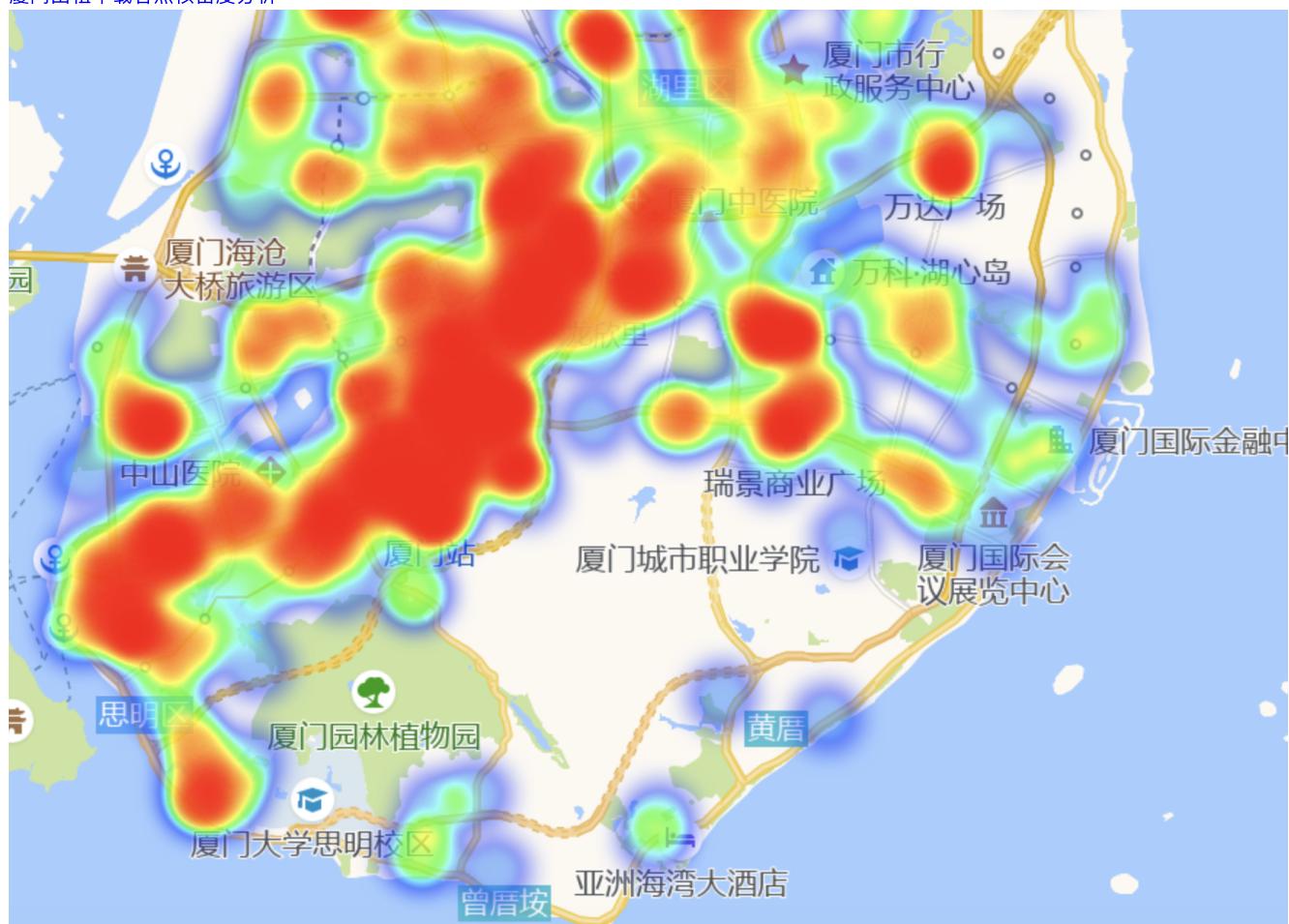
- 厦门出租车乘客下车点分布



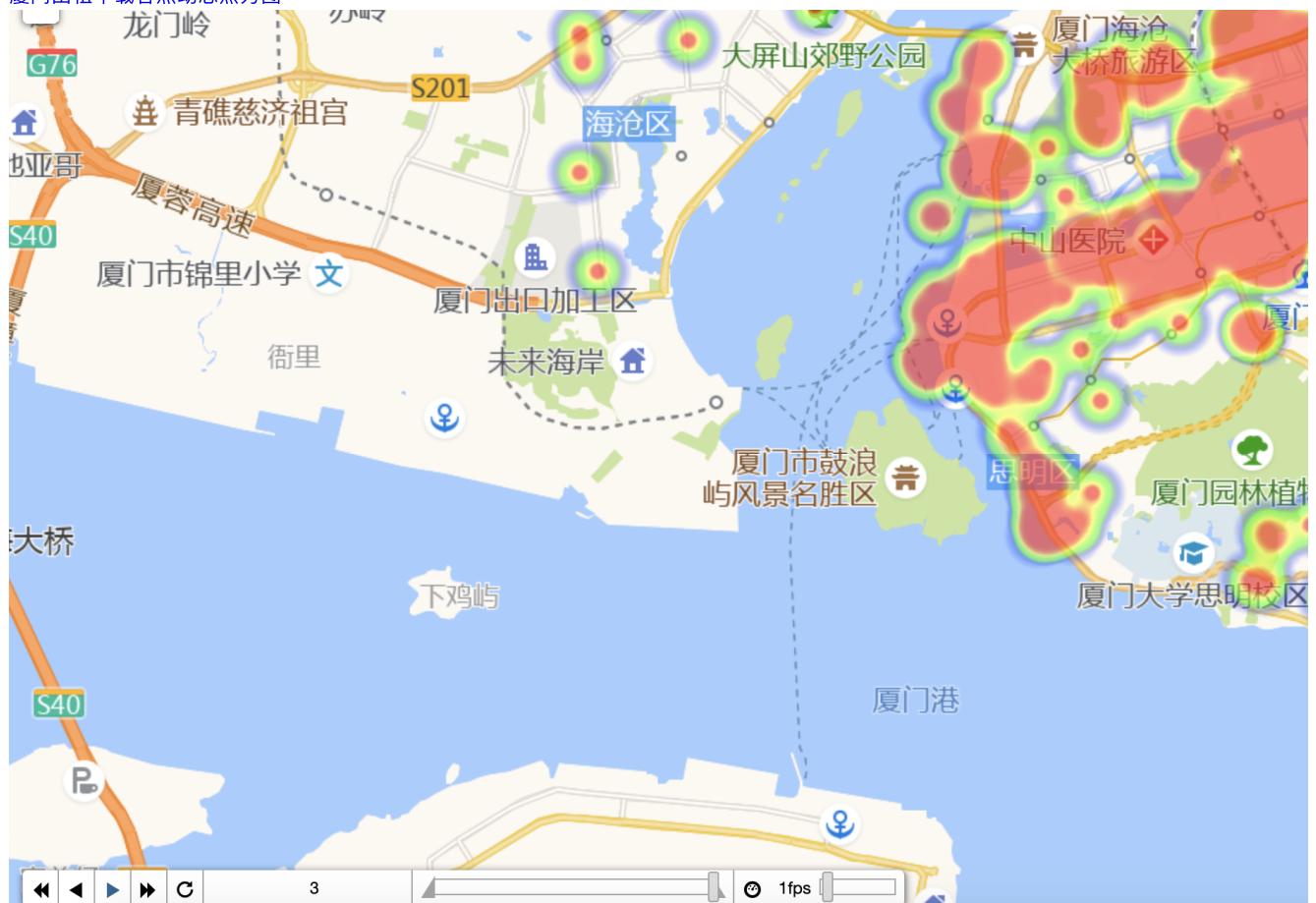
- 厦门出租车乘客上车点分布



- 厦门出租车载客点核密度分析



- 厦门出租车载客点动态热力图



主要参考文献

- [1] 叶臻, 王海洋, 贺明光, 等. 基于出行需求时空分布的出租车停靠站选址方法[J]. 华东交通大学学报, 2017, 34(06): 97–103.
- [2] 祁文田. 基于GPS数据的出租车载客点空间特征分析[D]. 吉林大学, 2013.
- [3] 林鹏飞, 翁剑成, 刘文韬, 等. 基于多源数据的网络约租车与出租车运营特征分析[J]. 交通工程, 2020, 20(01): 26–33.
- [4] 周静一, 李晔. 基于浮动车技术的杭州市出租汽车运行特征分析[J]. 科技信息, 2011(03): 120–122.
- [5] 韩勇, 樊顺, 周林, 等. 基于聚类算法的出租载客点时空分布特征研究[J]. 中国海洋大学学报(自然科学版), 2019(S1 vo 49): 155–162.
- [6] 金雷, 谢秉磊. 基于上落客时空特征的出租车停靠站选址模型[J]. 交通运输系统工程与信息, 2015, 15(02): 182–188+194.
- [7] 胡兰兰. 基于GPS出租车高收益热点区域推荐[D]. 温州大学, 2019.