题目：上海市出租车订单数据分析报告

姓　　名：　王倩妮

学　　号：　2015112956

班　　级：交通 2015-02 班

任课教师：　谢军

2017 年 12 月 27 日

一、问题描述

现有两日滴滴平台上海市出租车订单数据，需利用 GIS 对订单数据数据进行时间、空间两个层面的分析。首先利用 QGIS 或类似 GIS 软件将订单数据转化成 SHP 文件（点），然后结合路网和小区 shape 分析出租车一天内不同时段，工作日与非工作日的起点和终点的空间分布情况，并解释现象出现的原因。在分析过程中需使用多组图片来进行辅助说明，同时制定定量指标来支撑分析结论。

二、数据预处理

2.1 数据预处理

2.1.1 原始数据描述

本问题提供以下四项原始数据：

- 上海市路网文件
- 数据项：长度、名称等 22 项属性
- 上海市路网划分方格区域文件
- 数据项：区域编号、包含路段数、区域
- 滴滴平台 2016-12-22 星期四出租车订单数据
- 数据项：订单 id、驾驶员 id、车辆 id、起点经度、起点纬度、终点经度、终点纬度、开始时间、结束时间、产品类型、线路
- 滴滴平台 2016-12-24 星期六出租车订单数据
- 数据项：订单 id、驾驶员 id、车辆 id、起点经度、起点纬度、终点经度、终点纬度、开始时间、结束时间、产品类型、线路

2.1.2 数据补全

经验证发现，订单原始数据中“开始时间”数据正常，但“结束时间”数据项存在近 1/3 空数据，即“0000-00-00 00:00:00”数据。时间数据是进行数据分割的重要依据，空数据的存在会使有效的 D 数据大大减小，造成实际 D 数据与有效 D 数据差距悬殊，从而降低分析结果的准确性。为解决这一问题，需对原始数据进行补全操作。

目前处理数据缺失的方法主要有：直接舍弃空值、填充 0 值、填充均值、利用上下数据填充、插值法、利用算法填充等，但由于数据缺失量较大，此处采用如下方法进行补全：

STEP1：针对两天的数据，分别划分正常数据与缺失数据。

STEP2：分别求取周四、周六正常数据的旅行时间平均值。

STEP3：针对缺失数据，利用如下公式补全：

$$e\_time_{i,k} = s\_time_{i,k} + period_k \left(k = Thur / Sat\right)$$

e_time——结束时间

s_time——开始时间

$period$——旅行时间均值

表 1 旅行时间均值对照表

序号	时间	星期	正常数据旅行时间均值
1	2016-12-22	周四	00:21:05.515941
2	2016-12-24	周六	00:18:36.266232

2.2 数据文件分割

本题需对不同时段的订单数据进行分析，因此，选取“早高峰”、“平峰”、“晚高峰”三个特征时段，每段时长 2 小时，进行后续分析。

特征时段定义如下：

- 早高峰：7:30-9:30
- 平峰：14:00-16:00
- 晚高峰：17:30-19:30

按照以上定义分割预处理后的数据，并存储为 csv 格式文件用于后续分析。

三、指标确定

3.1 密度指标

利用 “样方计数法”原理，选取“每个小区内订单数量占该日该时段总订单量的百分比”对每个小区每个时段的订单数进行量化。此指标用于绘制热度分布图像。

$$\sigma = \frac{N_{(i,j)}}{N_{k,p}} \times 100\%$$

(i, j)——小区编号

k——$Thur / Sat$

p——早高峰/平峰/晚高峰

四、订单数据分析

4.1 数据量分析

经预处理与分割后，用于进一步分析的数据条数如下表所示：

表 2 数据条数对照表

序号	数据条目数	对应时间	序号	数据条目数	对应时间
1	3816	周四早高峰 O	2	3588	周四早高峰 D
3	2427	周四平峰 O	4	2519	周四平峰 D
5	2570	周四晚高峰 O	6	2762	周四晚高峰 D
7	2982	周六早高峰 O	8	2986	周六早高峰 D
9	2426	周六平峰 O	10	2472	周六平峰 D
11	2691	周六晚高峰 O	12	3019	周六晚高峰 D

由表 2 数据可以得出，相同时间间隔内，早晚高峰时段订单量大于平峰时段订单量，可知在早晚高峰利用滴滴平台接单的出租车数量大于在平峰利用滴滴平台接单的出租车，从而间接反映出上海市早晚高峰时段交通出行量大于平峰交通出行量。

4.2 按密度指标的热度图像绘制

利用 Nature_Break（fisher_jenks）绘图模式对密度指标进行可视化处理。

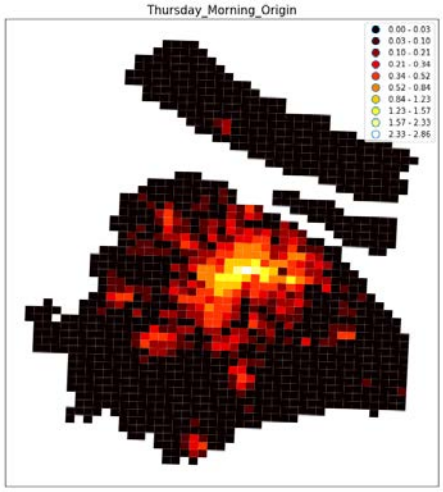


图 1 周四早高峰 O 点分布

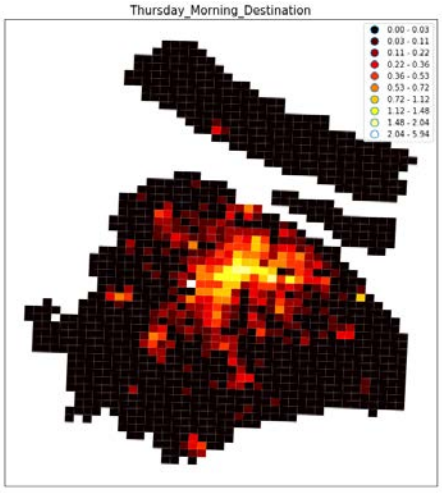


图 2 周四早高峰 D 点分布

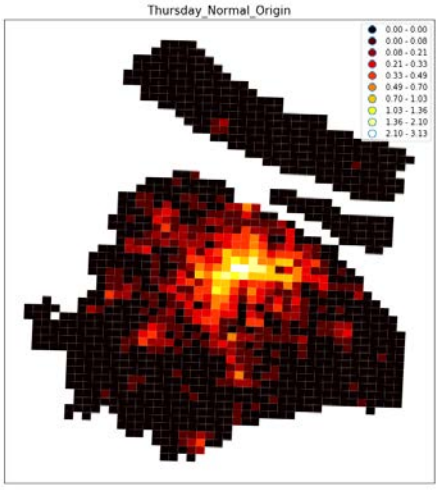


图 3 周四平峰 O 点分布

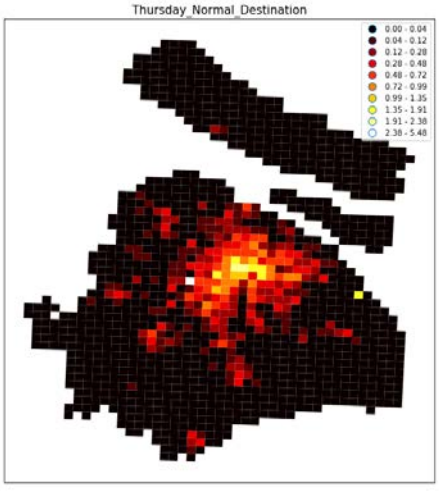


图 4 周四平峰 D 点分布

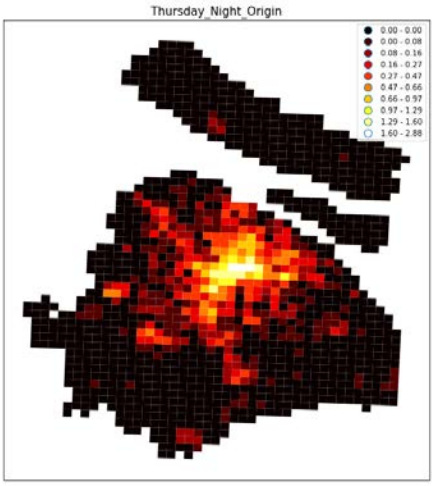


图 5 周四晚高峰 O 点分布

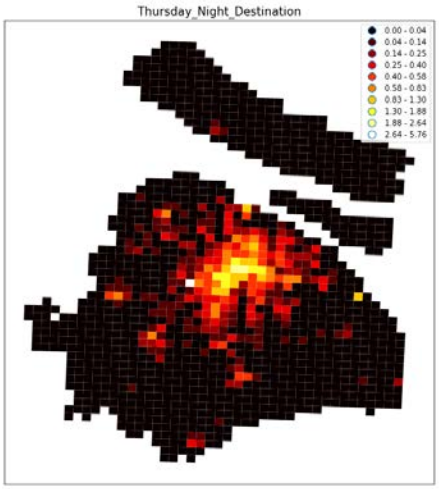


图 6 周四晚高峰 D 点分布

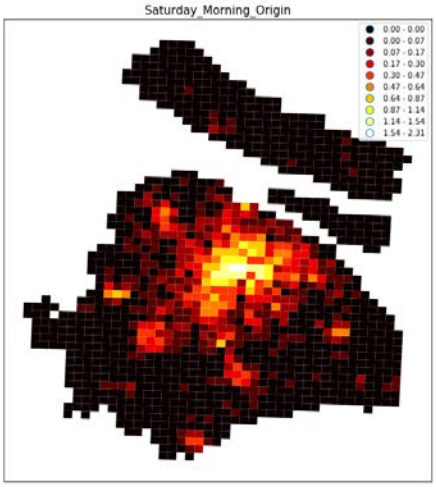


图 7 周六早高峰 O 点分布

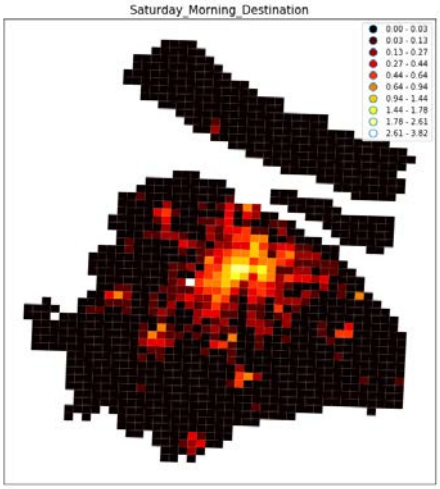


图 8 周六早高峰 D 点分布

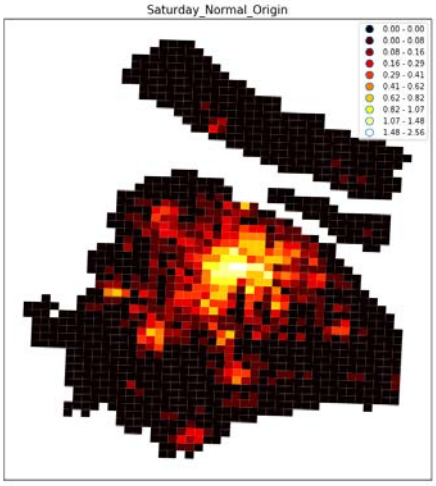


图 9 周六平峰 O 点分布

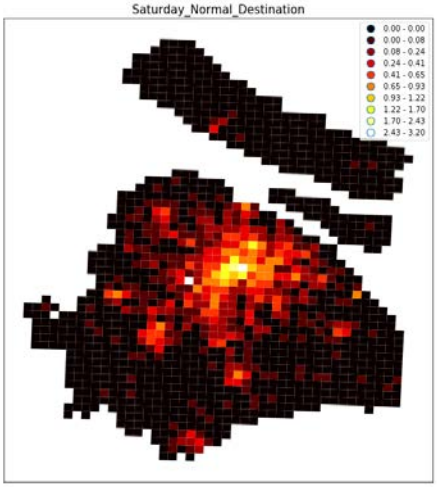


图 10 周六平峰 D 点分布

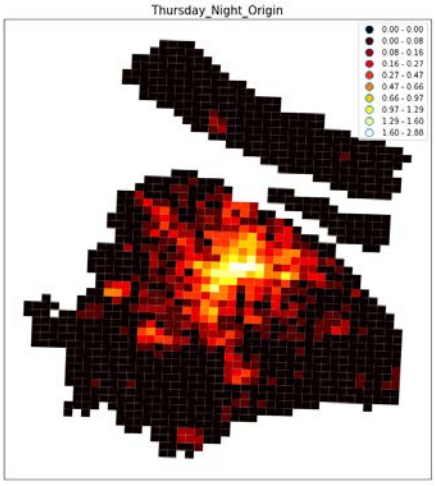


图 11 周六晚高峰 O 点分布

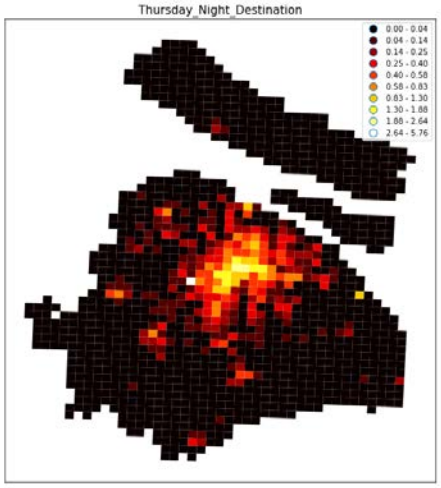


图 12 周六晚高峰 D 点分布

4.3 中心趋势分析

结合以上热度分布图，可以通过订单数据发现上海市城市分布的特点。

图 13 上海市区划图

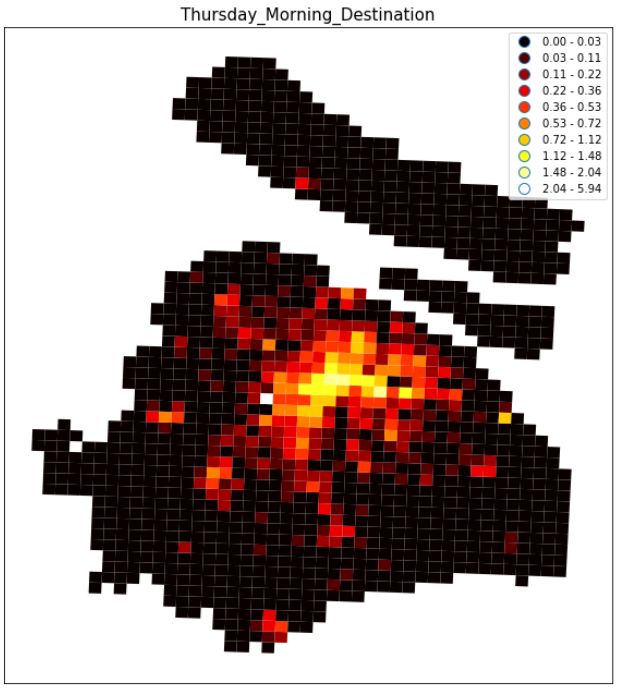


图 14 周四早高峰 D 点分布

以周四早高峰 D 点分布为例，对比上海市区划图，可以推测出上海市的发展模式为“双中心、多组团”的发展。即可以通过热度图像定位城市发展的大中心与区域组团中心。

以整个上海市为研究对象，订单数据量最大的区域为“上海市区”，次之为“浦东新区”与“闵行区”。以以上三个区域为出行起点或终点的订单数据占总订单数据的绝大多数。闵行区与上海市区距离较近且无河流间隔，可以归为一个区域进行研究，因此以整个上海市为研究对象，上海市区与闵行区组成的整体与浦东新区为整个上海地区的

双中心。上海市区整体热度较高，而浦东新区东部地区由于距离上海市区距离较远，热度较低。据此也可以说明，目前浦东新区的发展主要集中在靠近黄浦江区域。未来随着城市的发展，浦东新区也将逐渐朝东方向发展。

再分析其余各个片区，对照以下两图不难发现，订单数据定位这些片区的分中心与规划中相一致。

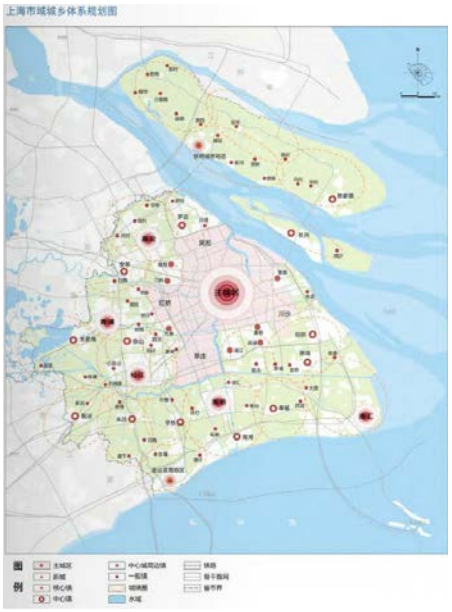


图 15 上海市域城乡体系规划图

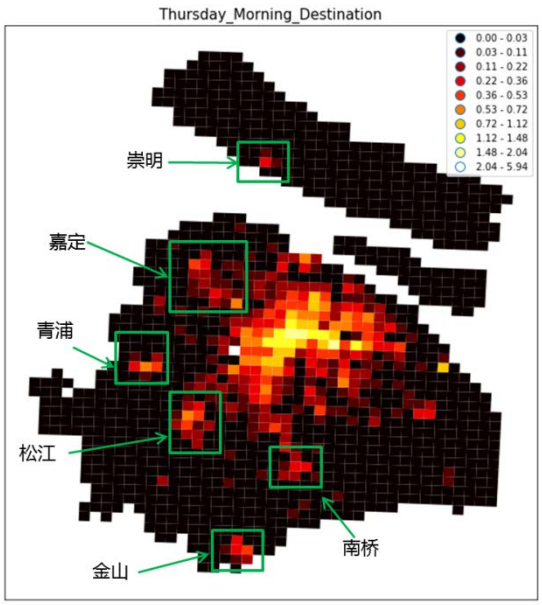


图 16 周四早高峰 D 点分布

4.4 订单数据总体趋势分析

4.4.1 分析步骤

以上分析通过统计点数量赋予面要素以新属性，类似于四阶段法中的“交通生成”步骤，但仅通过这样的方法不利于联系起 O 与 D 进行流向、流量的分析。因此接下来，利用类似四阶段发中的“交通分布”步骤，绘制出行线。

STEP1：依照分割后数据读取 O 点/D 点处于某个时段数据。

STEP2：提取这些数据中 O 点与 D 点所在的方格。

STEP3：以方格中心点代替方格小区。

STEP4：将两个点元素组合为线元素。

STEP5：确定无重合的线元素集合。

STEP6：统计该数据集中线元素集合每个元素出现的次数。

STEP7：绘制 OD 线。

依据绘制出的 OD 线在整体和局部、共性与差异性等角度进行进一步分析。

4.4.2 分析结论

绘制所得的出行线效果如下图所示，可见以此进行示意较为杂乱。结合图 1 至图 12 的热度图像，可以发现，各时段拥有的共性特点主要有以下方面：

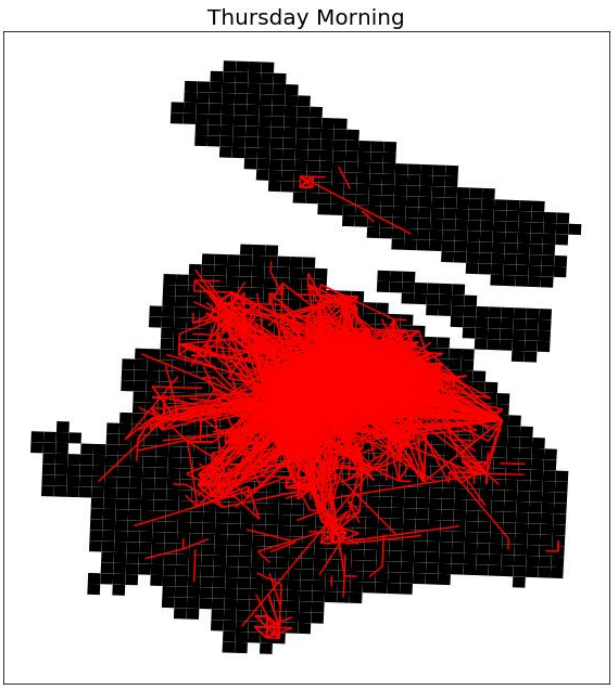


图 17 周四早高峰出行线

- 结论 1：热点区域

不论时段与工作日非工作日与 O 或 D 数据如何变化，编号为 x23_y25、x24_y25、x25_y25、x26_y25 的 4 个方格区域热度始终处于较高水平。经验证发现四个方格区域位于人民广场、上海外滩、陆家嘴附近，可见这些区域为上海市交通最为繁忙的区域。

- 结论 2：人们采用出租出行的距离

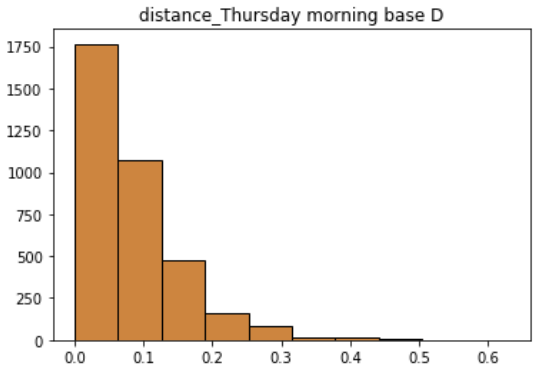


图 18 周四早高峰出行欧式距离直方图

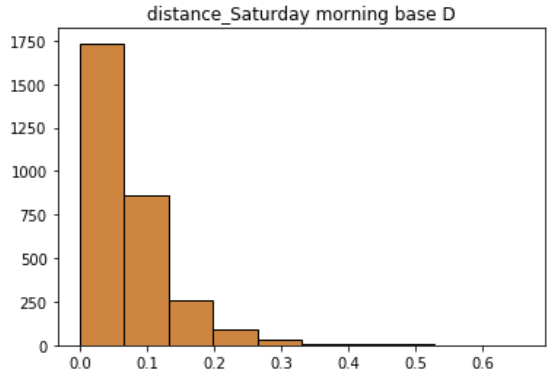
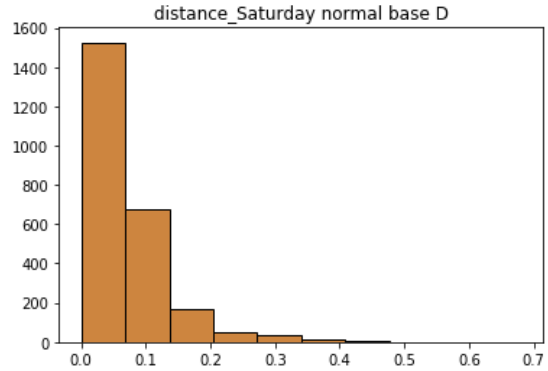


图 19 周六早高峰出行欧式距离直方图



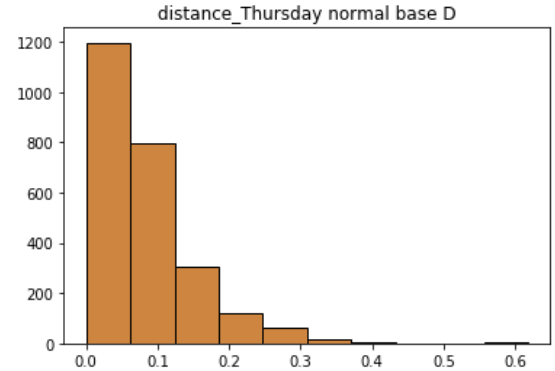


图 20 周四平峰出行欧式距离直方图

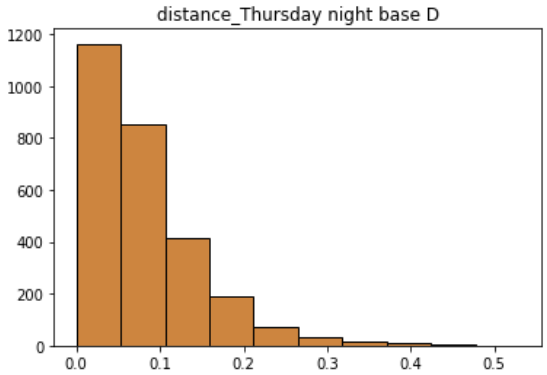


图 21 周六平峰出行欧式距离直方图

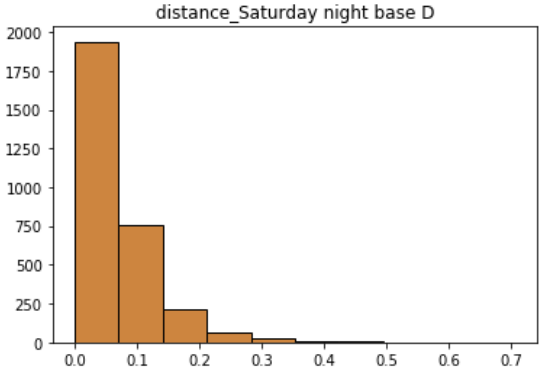


图 22 周四晚高峰出行欧式距离直方图

图 23 周六晚高峰出行欧式距离直方图

通过结合两天共六个时段的出行欧式距离分布直方图，可以发现随着距离的增加，出租车订单数量逐渐减少。而差异性在具体时段、工作日与非工作日之中体现的并不明显。

4.5 局部现象分析

通过观察图 1-12 可以发现以下结论。

- 结论 1：交通枢纽客流显著

经观察图 1 至图 12 中偶数图即各时段 D 数据，不难发现，图中最明显的白色区块，编号：x19_y23，此区域热度σ为所有方格区域中的最大值，经与 OPEN STREET MAP 对比发现此区域内包含虹桥机场与虹桥火车站。由此我们可以看到交通枢纽的集散作用的显著性。但观测各时段此区域的 O 数据，此方格区域热度σ并没有像 D 那样明显，也在侧面反映出人们在前往虹桥机场、火车站时更乐于采用出租出行的方式，而在离开虹桥机场、火车站时可能更乐于采用地铁、机场大巴等交通方式。

应用同样的方法，浦东机场所在方格区域，编号：x38_y22，也可见与上述结论相一致的现象。

以周四晚高峰为例，示意效果如下图所示。

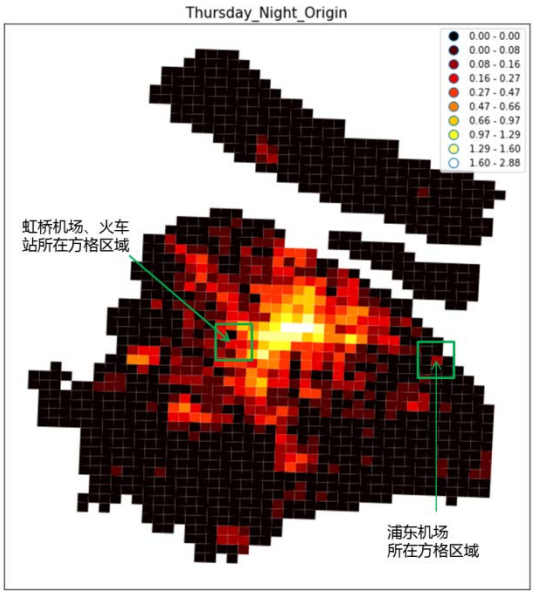


图 24 周四晚高峰 O 机场位置示意

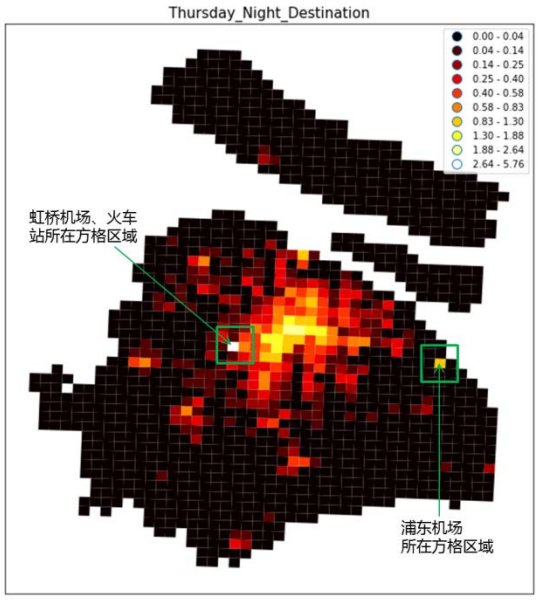


图 25 周四晚高峰 D 机场位置示意

以周四早高峰为例，在早高峰内到达虹桥机场方块区域的订单数量为 213 单，来源示意图如下图所示。将 O 与 D 区域中点的连线距离（仅为经纬度构成的坐标点的欧式距离，与实际直线长度存在差距）绘制直方图，可以看出：去往虹桥机场的出租出行交通主要集中在由上海市区、浦东新区出发的区域，并大致分布在以虹桥机场为中心，以一定长度为半径的范围内，并偏向东方向。在距离方面，由直方图可以发现随着与虹桥区域 D 点距离的增加，订单数量总趋势大致为“先增加后减少”，所以与出租的费用、便利程度等因素均有关，选择出租方式前往虹桥机场与出行距离有着密切的联系。

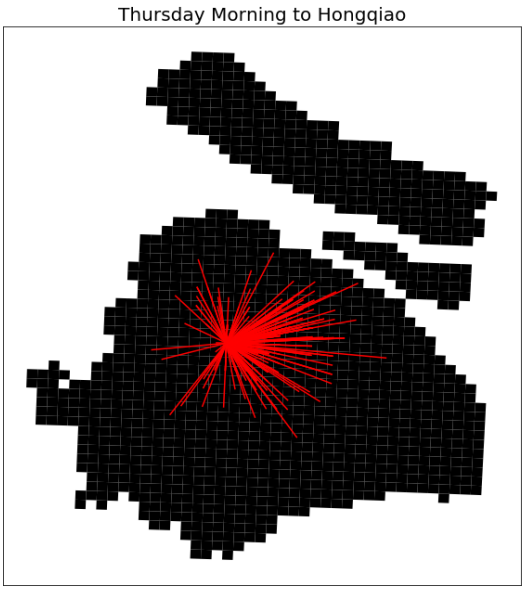


图 26 周四早高峰去往虹桥 OD 线

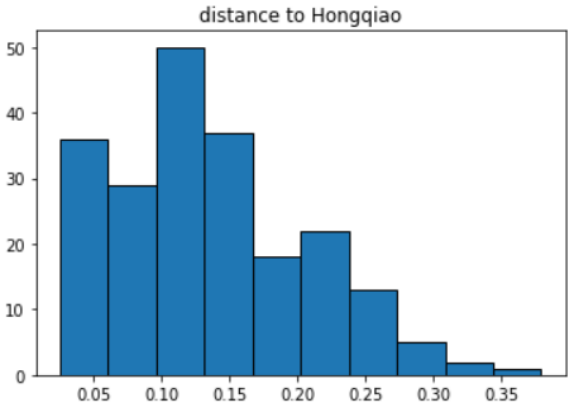


图 27 去往虹桥欧式距离直方图

使用同样的方法绘制早高峰时段到达浦东机场区域的示意图。在周四早高峰时段到达的订单数为 30 单。同样绘制距离分布直方图可以看出：去往浦东机场的出租出行交通也主要集中在上海市区和浦东新区。在距离方面，由于浦东机场位于上海东部，因此

到达浦东机场的订单欧式距离教到达虹桥机场的欧氏距离远，并且随着与浦东机场区域 D 点距离的增加，订单数量总趋势大致为“在某一水平后保持稳定”，当然由于样本量较小，这还需更多数据去验证以上规律。

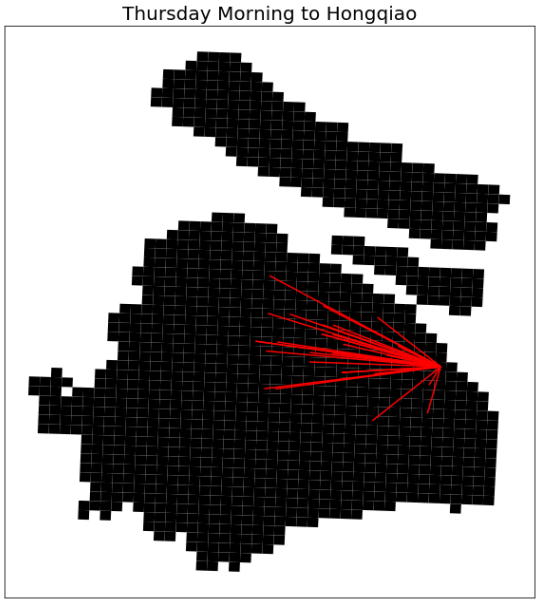


图 28 周四早高峰去往浦东机场 OD 线

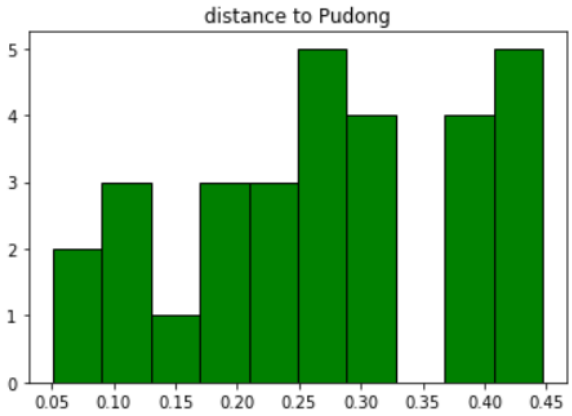


图 29 去往浦东机场欧式距离直方图

- 结论 2：使用出租数据的局限性

由于上海市城市面积较大，轨道交通覆盖面积广且十分便利，因此使用出租车出行的人群的订单数据有些时候存在一定偏差。如依据实际情况上海市浦江镇作为上海市面积最大的大型保障居住区，应该与工作区之间存在一定潮汐现象，但由于出租车消费水平、地铁便利度等原因的限制，人们可能更多采用地铁方式出行，因此在分析订单数据的过程中无法发现此潮汐现象。

五、运行环境

- IDE：jupyter notebook
- 编程语言：Python3
- 依赖第三方库：geopandas、numpy、matplotlib、pysal、time
- 操作系统：Windows 8.1
- 处理器:Intel(R)Core(TM)i5-5200U CPU @ 2.20GHz 2.20GHz
- 安装内存(RAM):4.00GB

心得体会

由于分析数据与分析能力的有限性，部分推论还需进一步使用更多数据进行验证。在本次作业的进展过程中，遇到各类问题，在老师的帮助下有了一定的思路。今后在数据挖掘、数据分析、数据可视化等方面还需更多时间与锻炼机会去进一步提升能力。