部分内容·基于层次聚类的站点时域特性分析

引言

站点的时域特性基于站点在不同时段的上下车量所得出，不同的站点存在不同的高峰时段，这与站点本身的属性有着密切联系，比如有的站点位于居民区，那较为合理的猜想是在早、晚高峰时段存在对应的上、下车高峰；工作区域则应有相反的情况。本章节正是基于相关数据，对站点的这些特性进行研究。

基于单日订单流的数据分析

对于站点数据，只保留位于上海地区的站点，并且删除了只允许同站点取还车的这类站点信息，最终保留 3296 处站点数据。

对于订单数据，选择 2016 年 8 月 1 日(周一)进行分析，当日订单数据共计 338146 条。通过订单的上下车时刻，计算时长(分钟)；通过上下车经纬度，计算直线距离(千米)。最终保留时长在 40～80 分钟，距离在 8～40 千米范围内的数据，共计 18553 条。

由于共享车辆的上下车都在站点中进行，因而需要对出租车数据中的上下车地点进行站点匹配，以进行后续分析。通过 pandas 来读取数据，通过 geopy 计算每条订单数据中的上车、下车经纬度与站点数据中所有站点经纬度的距离，选择其中距离最小的站点作为上车、下车的对应站点。保留上下车所对应站点距离均在 1 千米内的数据，共计 13814 条。

站点及订单连线数据可视化

通过 pyecharts 将站点数据和站点间订单连线显示在上海市地图上，如图 1-1、图 1-2 所示。

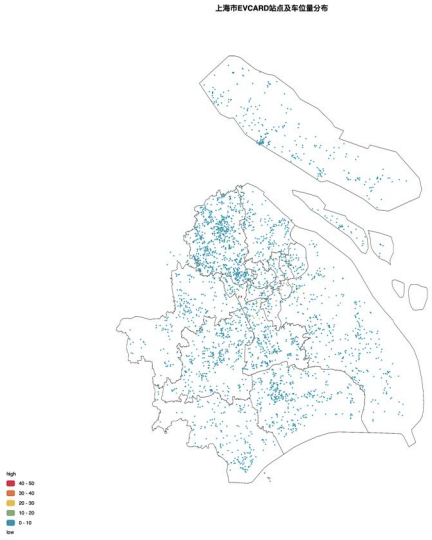

图 1-1 上海市站点及车位量分布

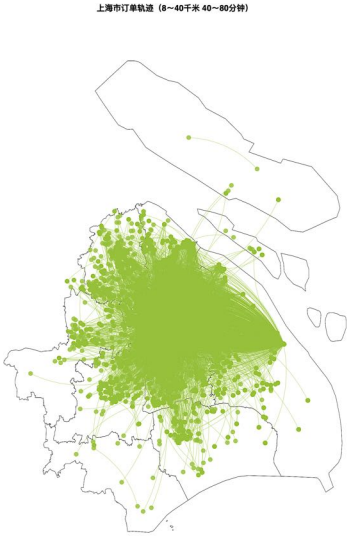

图 1-2 站点间订单连线图

上下车数据热力图

将订单数据分时段划为 0～6, 6～12, 12～18, 18～24 四组，统计每组时间段内各站点的上、下车次数。去除交通枢纽(虹桥)后的上下车次数热力图如图 1-3、图 1-4 所示。

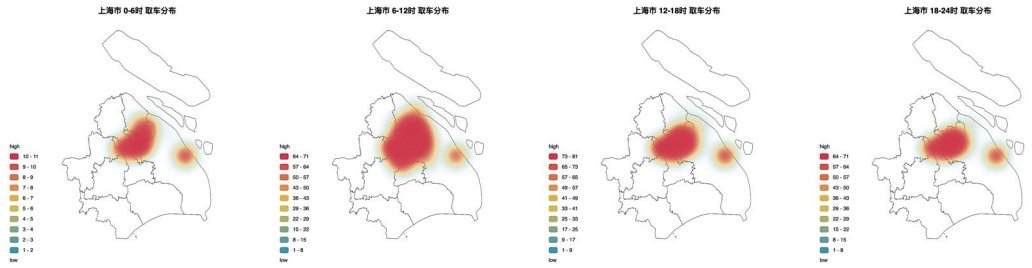

图 1-3 上车次数热力图

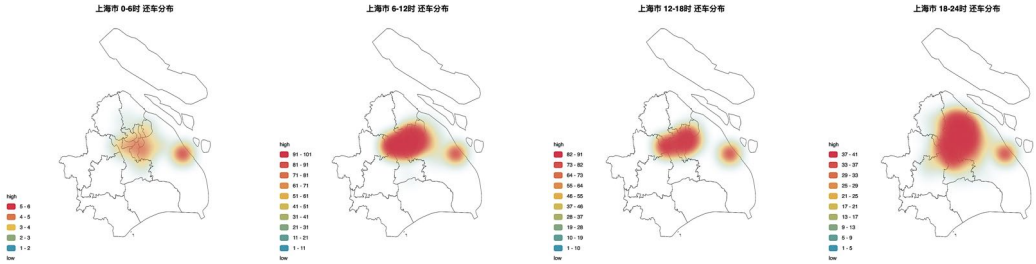

图 1-4 下车次数热力图

热力图反应了各时段不同区域上、下车的繁忙程度。去除虹桥枢纽是因为此站点的上下车量过大，对热力图的分布有较大影响。由于不同时段的车量范围不一致，因而不能横向由颜色深浅来进行比较。由热力图可以发现，上车在 6～12 时、下车在 18～24 时覆盖范围最广，这符合上班时段人流向市区集中、下班时段向外分散的特点。

站点的层次聚类

通过对站点进行层次聚类来划分具有相似时域特性的站点群组。

每个站点取 48 维特征向量，即各站点 24 个时段的上、下车量。通过 SciPy 实现层次聚类，聚类方法为 ward，相应距离选择为欧氏距离。分类后树状图如图 1-5 所示，图 1-5 中第 1 类（图中所示含 1107 个站点）又被细分为类 1 至类 3。选择有代表性的几类站点：类 1 含 357 个, 类 2 含 75 个, 类 3 含 675 个, 类 4 含 349 个, 类 5 含 60 个, 类 7 含 74 个。绘制全部站点的上、下车量折线图、堆叠图及各类站点地理位置分布图，如图 1-6 至 1-8 所示。

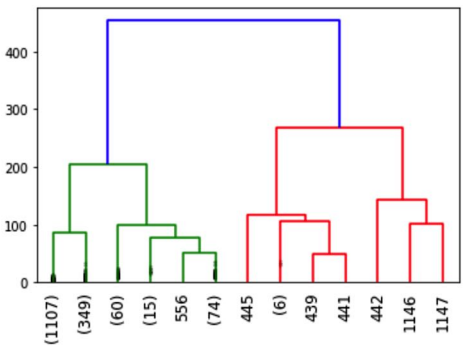

图 1-5 层次聚类树状图

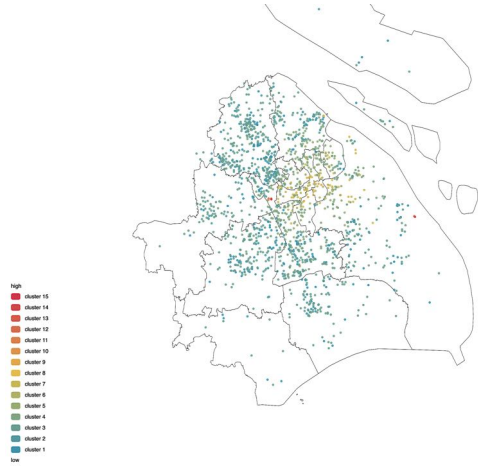

图 1-6 各类别站点地理分布

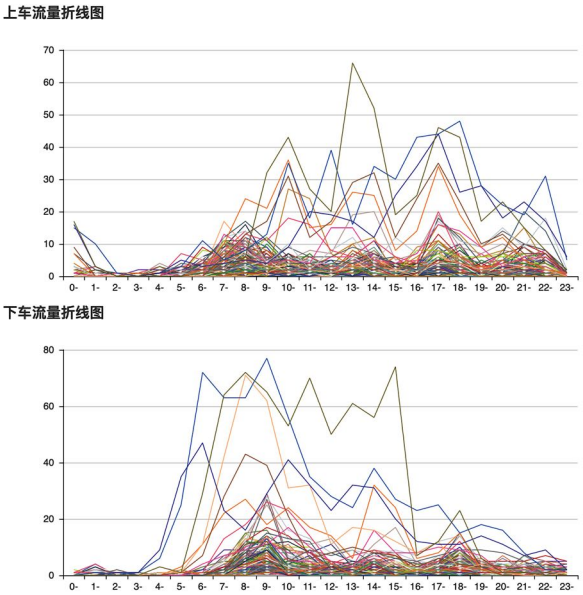

图 1-7 全部站点上下车量折线图

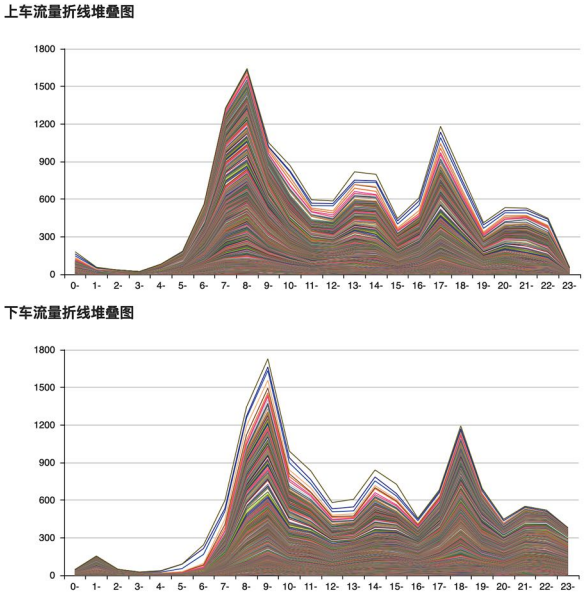

图 1-8 全部站点上下车量堆叠图

从图 1-7、1-8 中可以发现，多数站点各时段内的上、下车量并不大(5 辆左右)。在 5 时前是明显的用车低谷，而 7～10 时、13～15 时、17～19 时用车较频繁。全天在早高峰时段用车量达到最大。

代表性站点概况

对于有代表性的几类站点，其上下车量堆叠图及地理分布如图 1-9 至 1-16 所示。

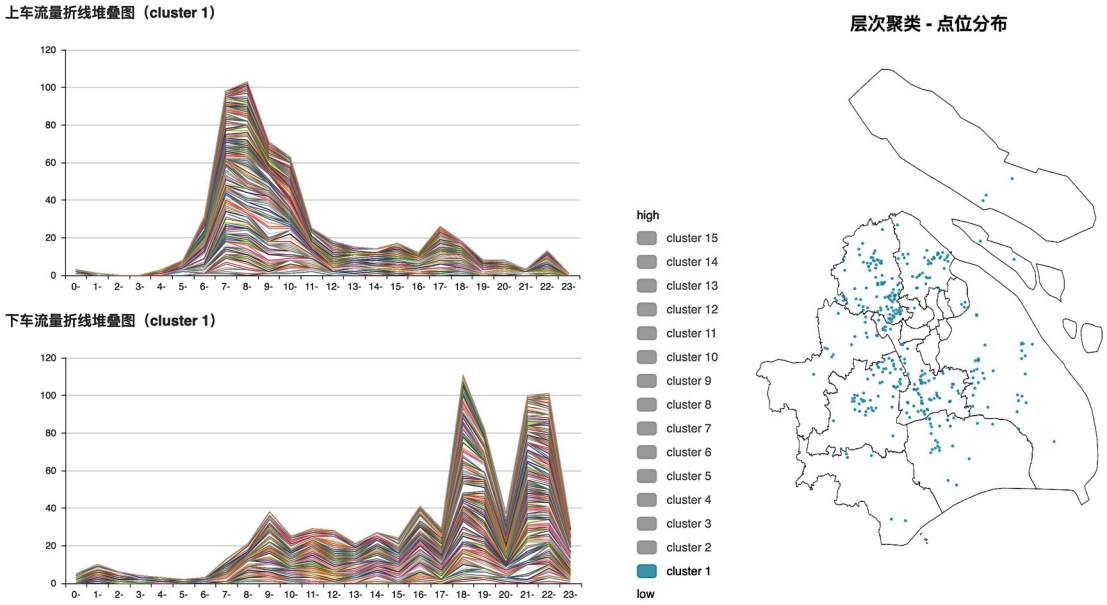

图 1-9 类别 1 站点(357 个)概况

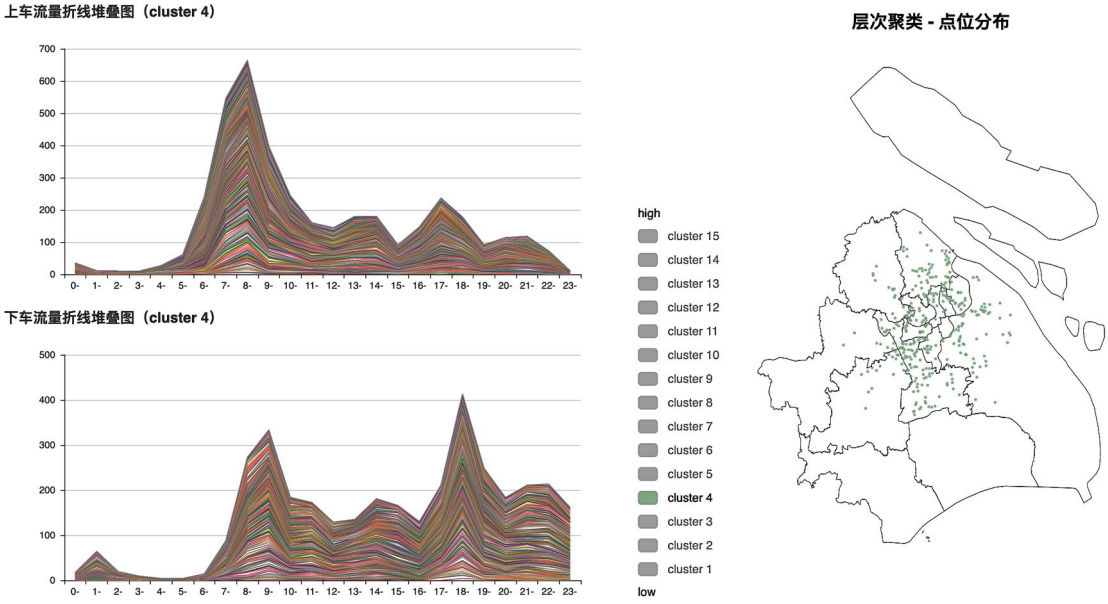

图 1-10 类别 4 站点（349 个）概况

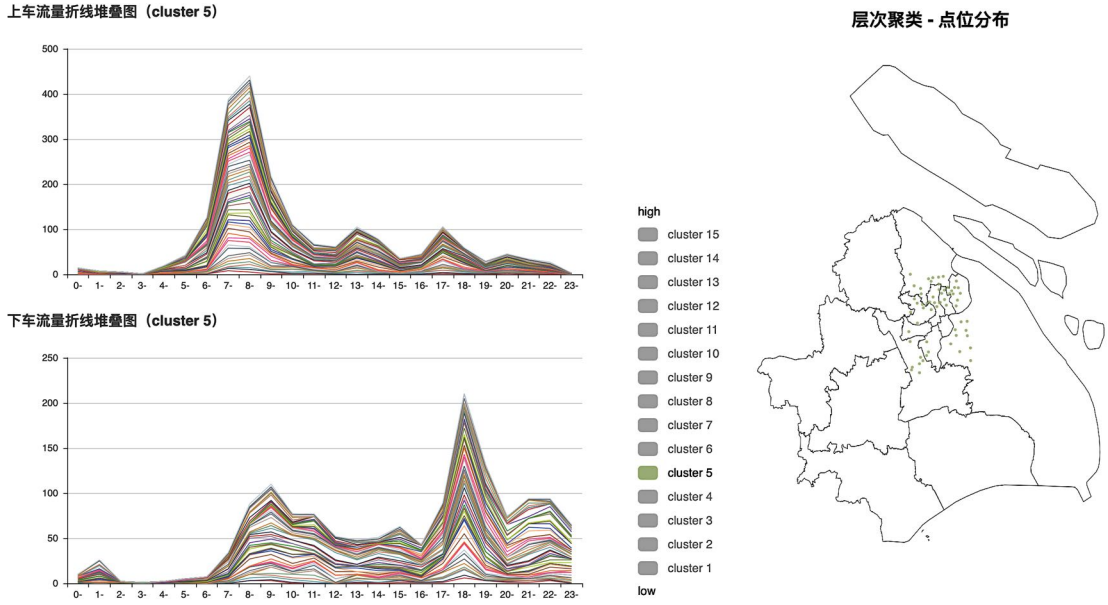

图 1-11 类别 5 站点（60 个）概况

类别 1、4 各站点的上下车量都较小（各时段上限分别为 3 辆、8 辆）。三类站点都属于早高峰上车量大，晚高峰下车量大，同时类别 4 在早高峰时的下车量也比较突出。

这几类站点较多为小区、旅店等居住场所。在站点名称中，这几类站点在关键词“小区”中的占比为 9/12，“酒店”中的占比为 75/124，“宾馆”中的占比为 12/23。

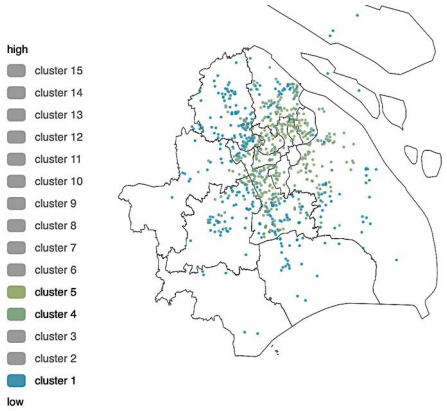

图 1-12 类别 1、4、5 站点分布

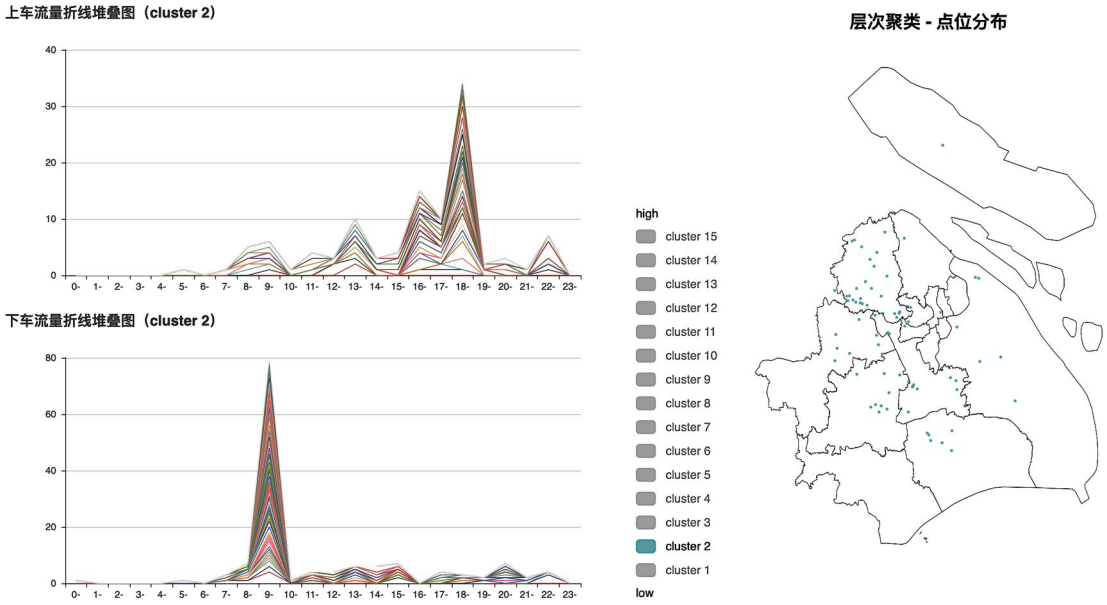

图 1-13 类别 2 站点（75 个）概况

类别 2 的站点各时段上下车量都在 5 辆以内，与类别 1、4、5 相反，这类站点的特征是早高峰下车量大，晚高峰上车量大。站点多为园区、公司、局一类的办公场所，且地理分布上多位于郊区，如“青浦区气象局”、“嘉定都市工业园”等。并非所有的办公场所都满足这种特征，这与其所处的地理位置有关。

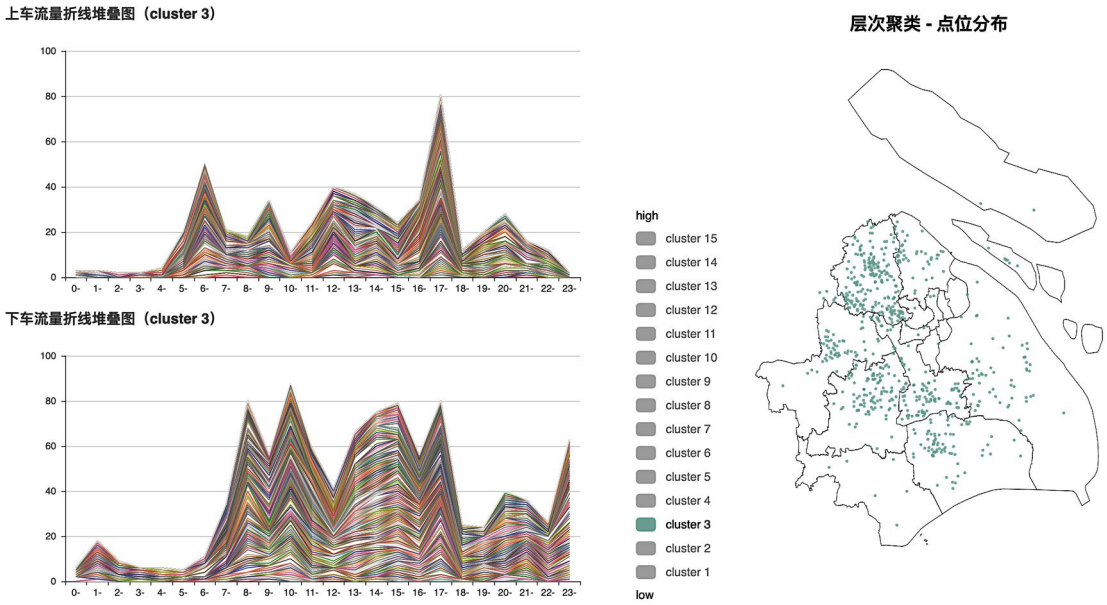

图 1-14 类别 3 站点（675 个）概况

类别 3 的站点各时段上下车量都在 3 辆以内，此类站点仅数量较多，没有什么特征。

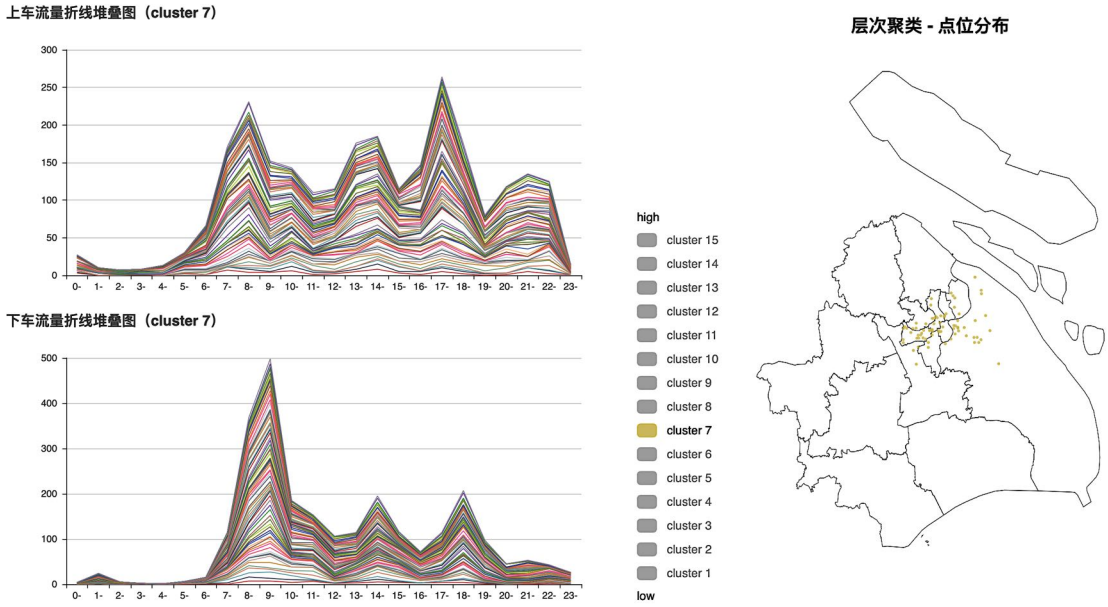

图 1-15 类别 7 站点（74 个）概况

类别 7 的站点各时段上下车量较大，此类站点的特征是早高峰下车量大，而上车量在全天各时段的分布比较平均。此类站点多为一些大型的工作区域，例如张江、五角场等地区，且地理上多分布于市区。

剩余的类别 6、9 为市区内个别工作场所，各时段上下车量大，同时满足早高峰下车量大，晚高峰上车量大的特征。类别 8、10～15 为交通枢纽：吴淞、虹桥、浦东，这些地方是港口、火车站、机场等特殊区域。这些类别的站点地理分布如图 1-16 所示。

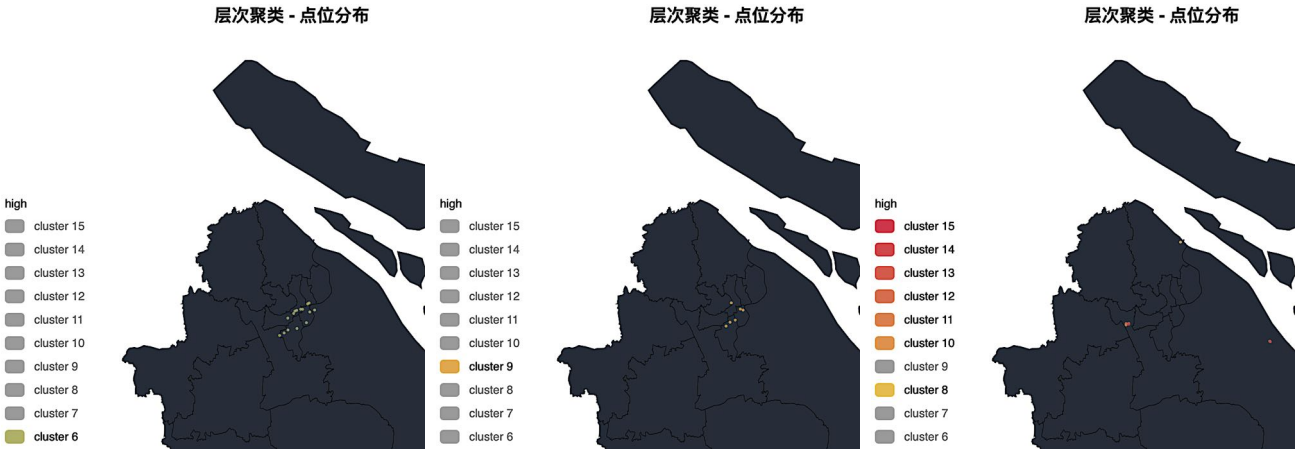

图 1-16 类别 6、9 及三交通枢纽地理分布

由上述分析可以发现，基于单日数据获得的站点群组，其不同的时域特性往往与站点的属性（公司、居民区、交通枢纽等）以及站点所处的地理位置（市区、郊区）有关。

部分内容·基于社团检测的站点地域特性分析

引言

以各站点为网络节点，站点间的每条订单表示节点间存在一条连边，通过 NetworkX 构建有多重边的无向网络。使用基于 Louvain 算法的 community.best_partition 对无向网络中的社团进行检测，从而得到稳定的地域社团，作为站点的地域特性。

基于单日订单流的社团检测

基于较长行程数据的社团检测

数据(时长在 40～80 分钟，距离在 8～40 千米范围内的订单共计 13814 条，涉及站点 1618 个)中订单用时、行程距离较长，因而检测出的社团中各站点的地理分布较为分散，且社团总数也较多。这符合长距离出行的特征，但无法直观展现各站点间区域分布与互相间联系的关系。社团检测结果如图 2-1 所示，各社团并未在地理上有很明显的集中。

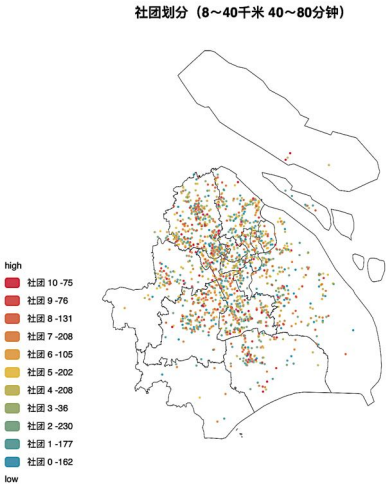

图 2-1 基于较长行程数据的社团检测

基于较短行程数据的社团检测

对原始的数据再次进行处理。选取时长在 20～40 分钟，距离在 6～16 千米范围内的数据，经处理后保留订单数据共计 39150 条，涉及站点 1488 个。数据量相比此前有明显增大，且因为行程距离缩短，在地理范围上更集中。为了比较社团划分的稳定性，进行了两次独立的检测，结果并未有明显不同，站点主要被划分为四部分，如图 2-2 所示。

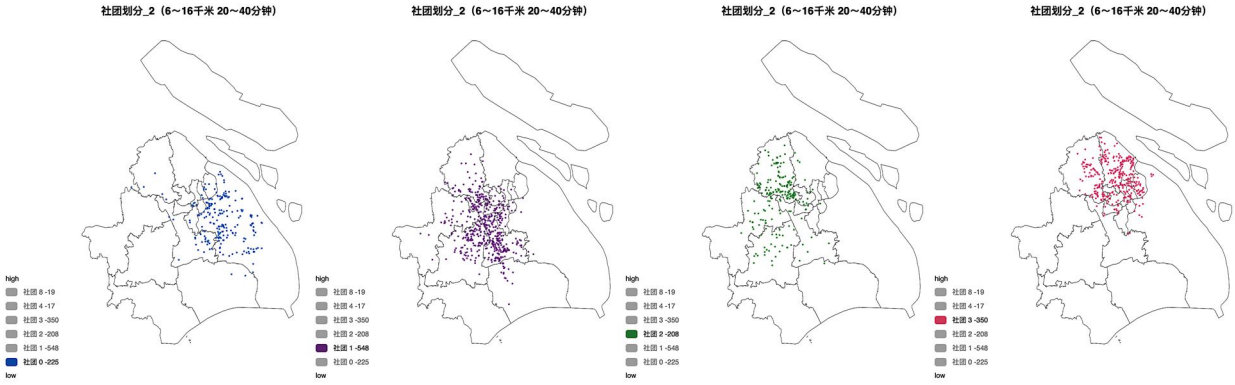

图 2-2 四主要社团站点地理分布

在图 2-2 中，社团 0 中的站点基本分布在浦东，虽然有一小部分站点在闵行区境内，但仍是以黄浦江为界，这表明江确实对两岸间的交通有一定阻隔。社团 1 中的站点主要位于松江、闵行与徐汇。社团 2 的站点主要位于嘉定南。社团 3 的站点主要位于嘉定东、宝山与市区北部。社团划分的结果符合居住地与工作地就近的特点。围绕市中心原点的这四部分区域互相间比较独立，这也符合现实情况。

基于长短行程合并数据的社团检测

将上述长短行程两份数据合并后，得到订单数据 52964 条，涉及站点 1869 个。对其进行社团检测的结果如图 2-3 所示，并未发生前述较长行程数据下社团内站点分散的问题。这说明在进行地域分析时更需要短行程数据的支持。

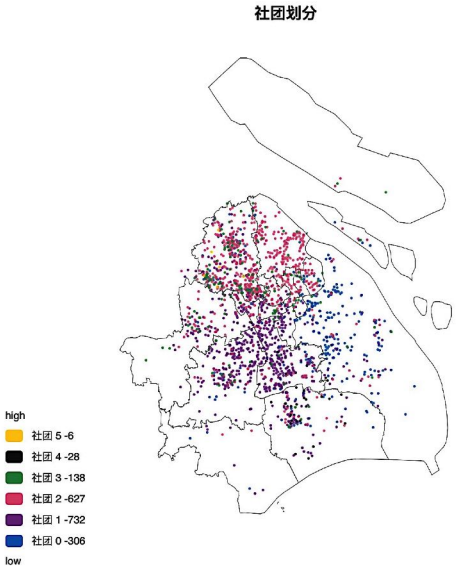

图 2-3 基于长短行程合并数据的社团检测

数据合并后，相比图 2-2，社团的总体布局并无太大变化，主要是增加了一些零散的离社团核心较远的孤立站点，如图 2-4 所示。这符合在出行上少部分人较远，大部分人就近的特点。

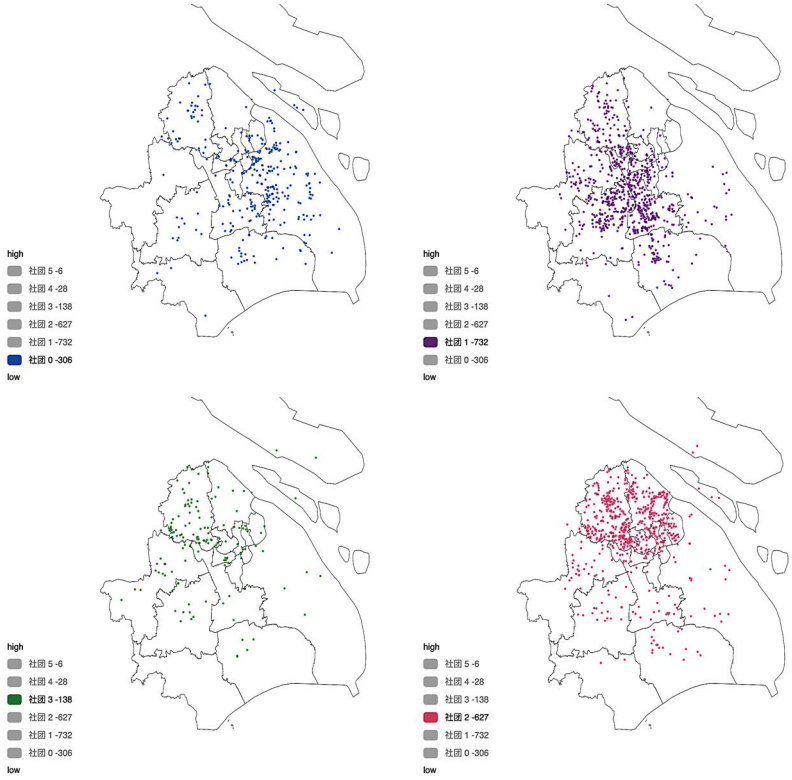

图 2-4 四主要社团站点地理分布