# *基于手机基站信息的城市空间特征及基础设施分布的经济学研究*

陈潇 周璐鹿 张言健

## 一. 研究背景

随着个人手机终端的普及，出行群体中手机拥有率和使用率已经达到相当高的比例，手机移动网络也基本上实现了城乡空间区域的全覆盖。手机基站记录等数据成为了一种新的研究城市与人口时空特征的有利工具。例如，对手机定位数据的挖掘和分析可以帮助调整交通政策以及基础设施的建设，使得城市的居民能获得更好的出行体验。在交通调查以及交通规划中，通过手机话单定位数据和手机信令定位数据进行去噪、扩样等预处理，最终能够获得居民出行特征数据。根据这些特征数据，可以分析人口就业分布、通勤出行特征，还可以进一步分析城市人口的时空动态分布等。

很可惜的是，我们没有手机基站数据的时间特征，但是仅利用空间数据与一些额外数据，我们仍然可以挖掘出关于人口及城市的部分特征。

当然，要挖掘基站所反映出来的信息，我们还需要了解基站是如何记录数据的。移动通信网络的信号覆盖逻辑上被设计成由若干个六边形的基站小区相互邻接而构成的蜂窝网络面状服务区，手机终端总是与其中某一个基站小区保持联系，移动通信网络的控制中心会定期或者不定期地主动或被动地记录每个手机终端时间序列的基站小区编号信息。在本次课题中，如若用户一直呆在某个基站小区内，则此基站会每隔两小时会进行一次记录；如若发生了小区间移动则相关的基站则立即记录。一般而言，基站的覆盖范围与设备数量与用量有关，设备越多的地方，为了维持信号的稳定，基站会布置得越多，因此基站连接用户的统计数据可以显示出人在空间上的分布。

## 二. 数据说明

1. 地理位置信息数据

① 手机基站经纬度

② 基站周围 POI 信息，利用百度地图地点检索服务 API 爬取，我们逐个读取 station.csv 中的经纬度，使用 python 发起 http 请求获取基站周围 500 米范围内的不同类型的建筑，并统计它们的数量。由于检索服务 API

是以地址名字的检索的方式进行的，所以我们将各种类型的建筑的搜索关键字设为“写字楼 大厦”“小吃 餐厅”,“商场 购物”,“学校 教育”,“景区”,“公寓 小区”,“博物馆 纪念馆” ,“工厂”, “市场”共 9 种类型的地址类型。

| 字段名 | 描述 | 补充说明 |
| --- | --- | --- |
| station_id | 基站编号 | 基站标识，不通知表示不同基站； |
| lng | 经度 | 百度地图的基站地理位置经度标识 |
| lat | 纬度 | 百度地图的基站地理位置纬度标识 |
| poi_1 | 附近办公楼数 | 从百度地图查询爬取 |
| poi_2 | 附近餐厅数 | 从百度地图查询爬取 |
| poi_3 | 附近商场数 | 从百度地图查询爬取 |
| poi_4 | 附近学校数 | 从百度地图查询爬取 |
| poi_5 | 附近景区数 | 从百度地图查询爬取 |
| poi_6 | 附近小区数 | 从百度地图查询爬取 |
| poi_7 | 附近博物馆数 | 从百度地图查询爬取 |
| poi_8 | 附近工厂数 | 从百度地图查询爬取 |
| poi_9 | 附近市场数 | 从百度地图查询爬取 |
| poi_10 | 基站地址描述 | 从百度地图查询爬取 |

**表 1 POI 数据**

2. 用户手机基站信息统计数据

| 字段名 | 描述 | 补充说明 |
| --- | --- | --- |
| user_id | 用户编号 | 用户标识，不同值表示不同用户； |
| station_id | 基站编号 | 基站标识，不同值表示不同基站； |
| count | 观测频次 | 2016 年 11 月对应用户的该基站观测次数； |

表 2 用户统计数据

## 三. 描述性分析

首先查看基站本身的地理分布情况：

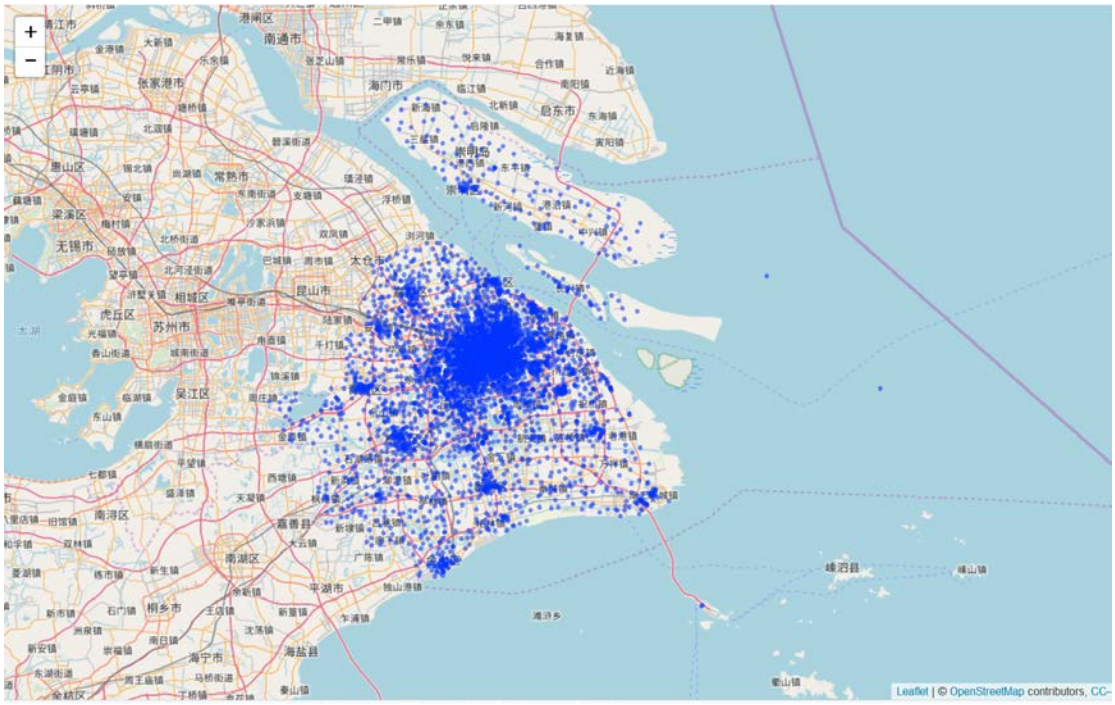
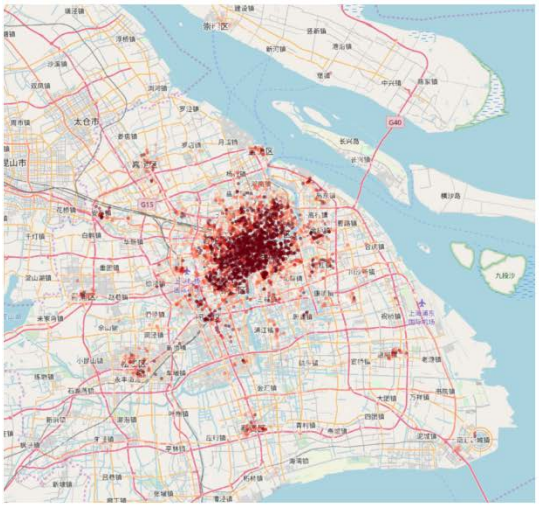


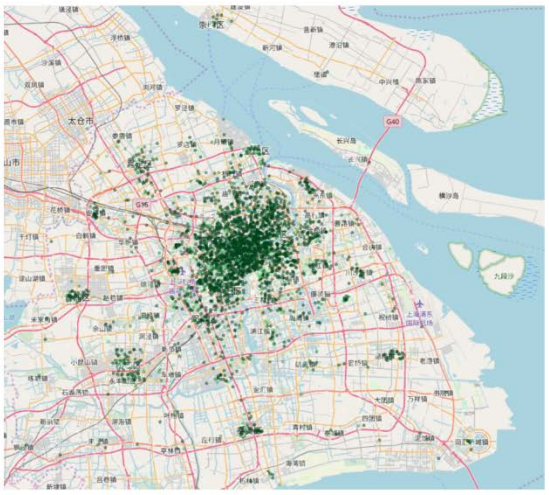

图 3 手机基站地理分布

可以看出，基站本身的地理位置分布和我们认知中的上海区域繁华程度接近，中心城区的分布最为集中，在图上几乎连成一块，其他区域以小聚落，大分散的形式存在。

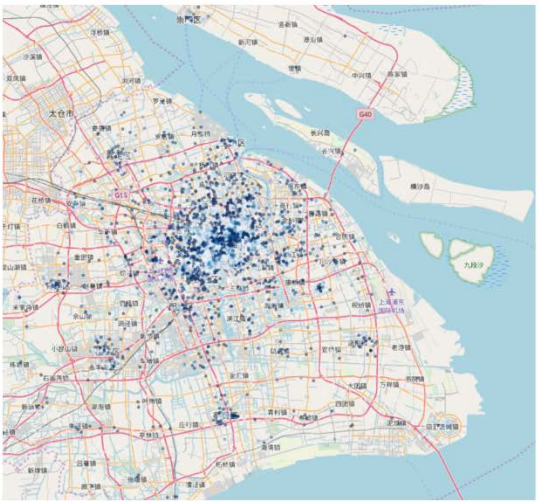
接着再查看手机基站附近的设施：



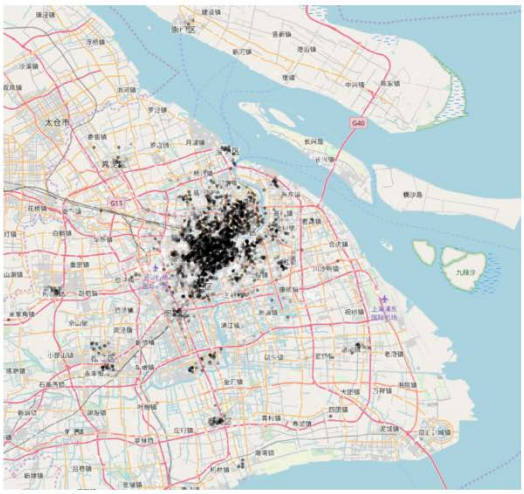
写字楼密度图



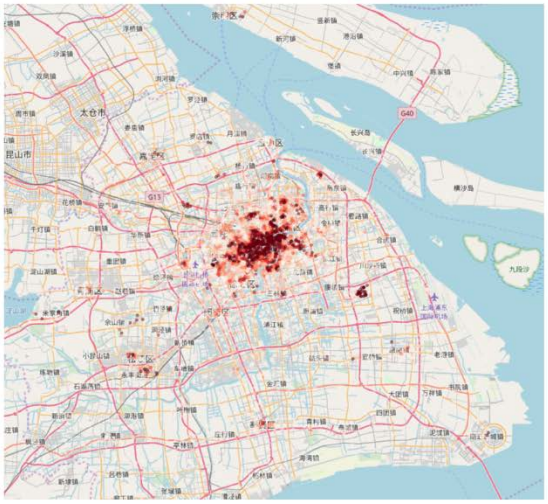
餐厅饮食密度图



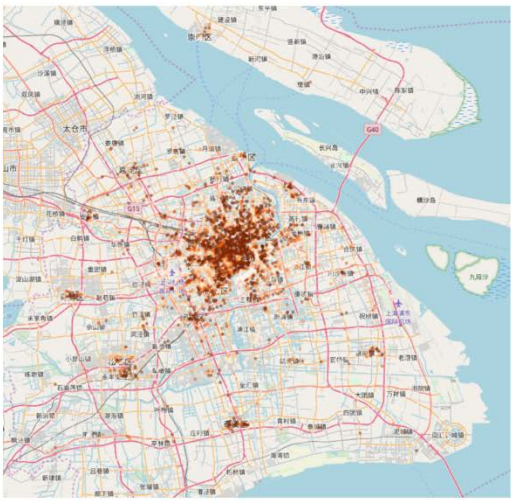
商铺商场密度图



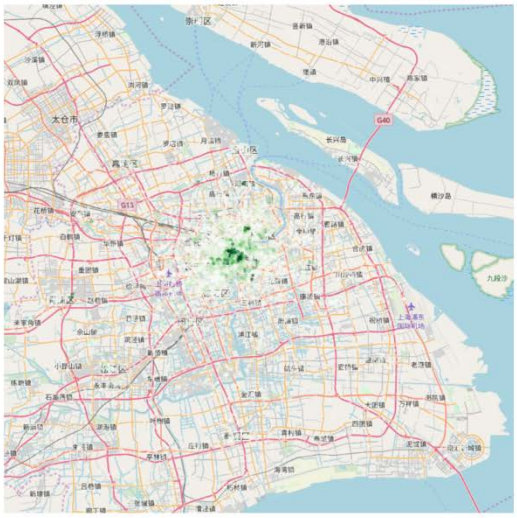
学校教育密度图



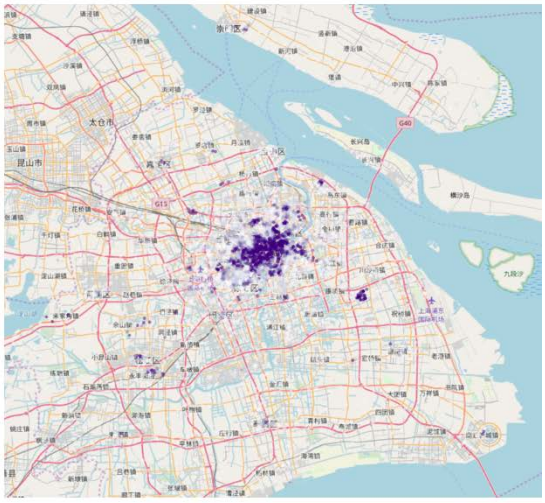
住宅小区密度图



市场市集密度图



博物馆纪念馆密度图



景区密度图

在这 8 幅地理位置分布图中，颜色越深代表周围相应设施越密集，可以看到很明显的地区差异，并且基本符合我们的判断。根据基站记录的信息来进行基础设施的分析可行。

在查看了基站和周边设施的地理位置分布信息后，我们对其余数据信息进行统计：

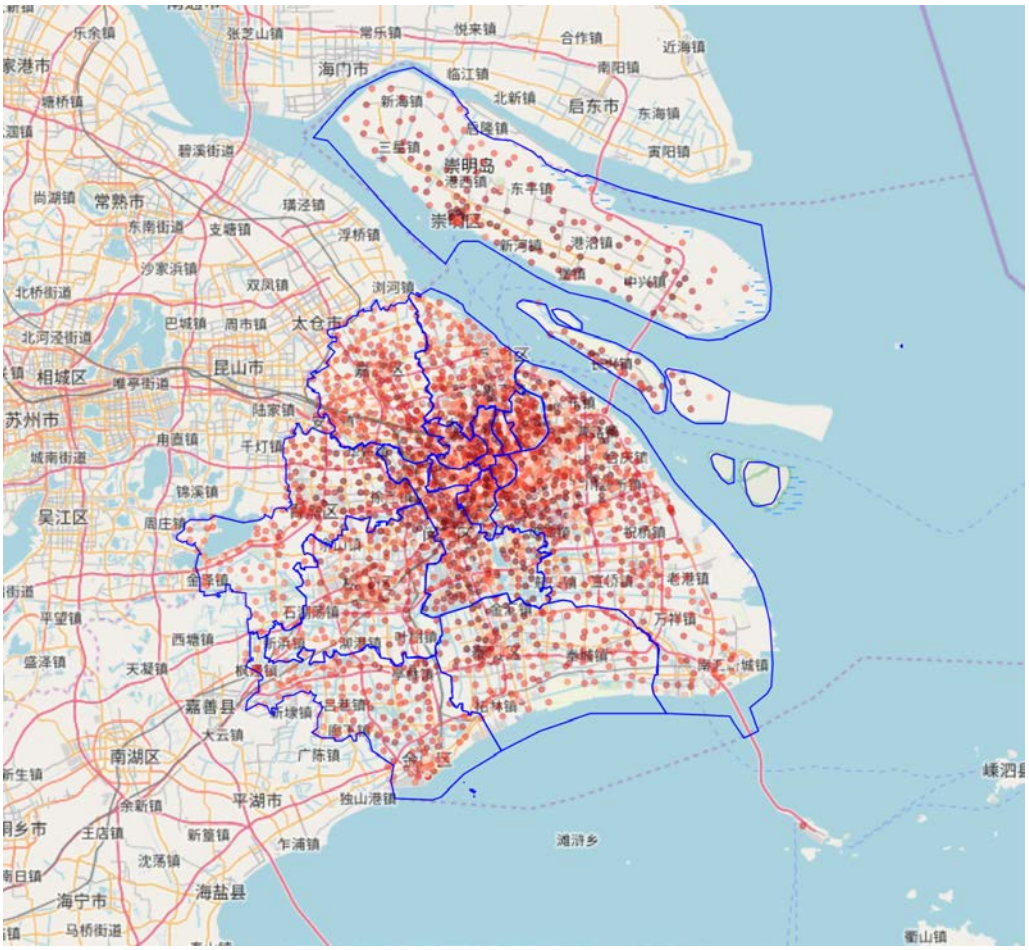



图 4 单个基站所记录的总次数

每个点代表一个基站，颜色深浅代表基站记录数据的多少，从图中可以看出，基站的所记录的数据是比较均匀的，以下的分布图也说明了这一点：

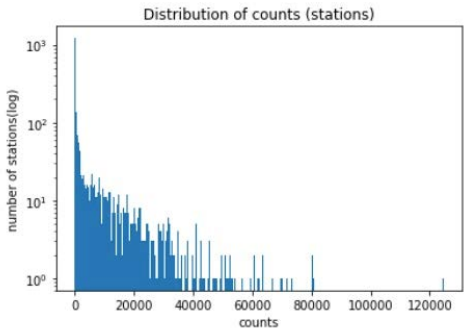
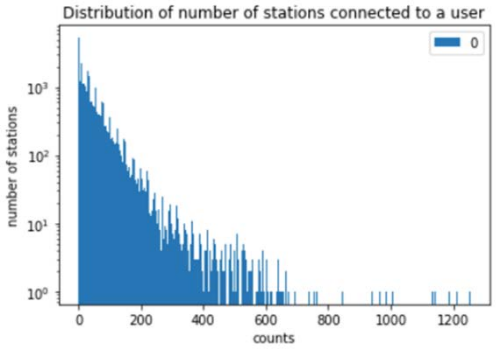



图 5 基站记录次数分布图

在研究完了基站的数据特点之后，我们来研究基站所记录的个体数据的特点，下图为个体被记录的基站个数分布：




个体数量太多，我们抽样来查看单个个体的数据特点。

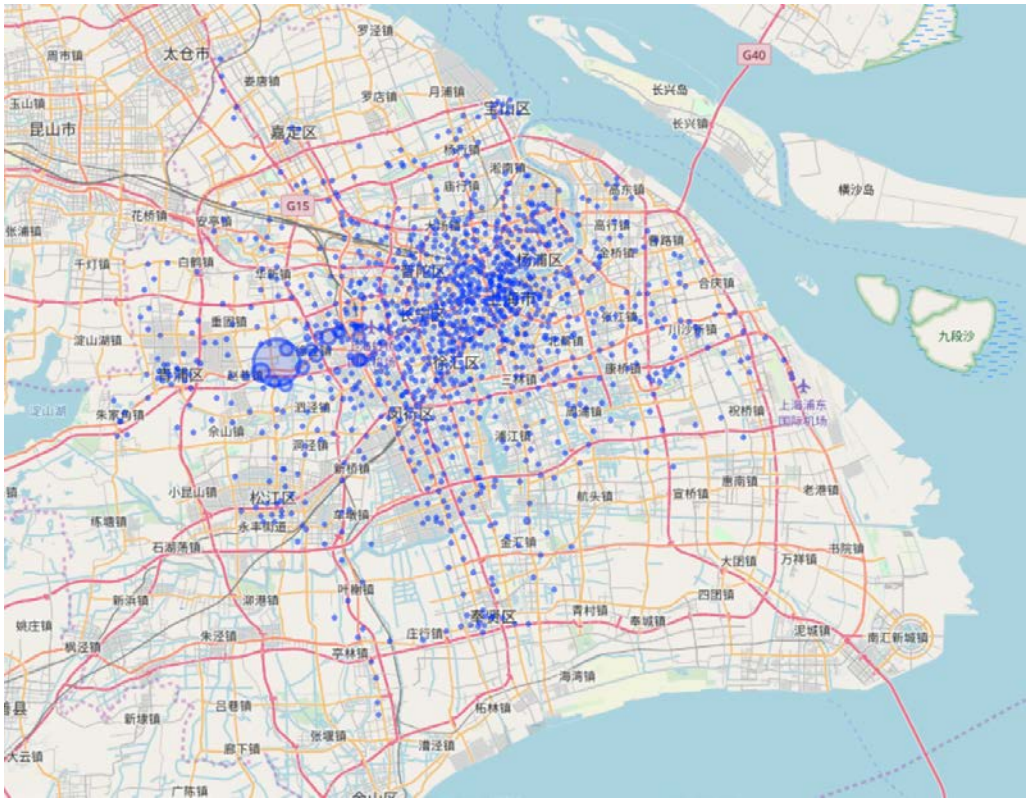



**图 6 抽样 1：个体基站记录活动**

抽样 1 中可以很清晰地看到有一个信息测度数量最大的点（图中的大圆圈处），但是欠缺更多的移动顺序信息，初步推测可能是此个体的居住地。此外，该个体的记录数异常的多，超出常理。很有可能是滴滴司机或调研员，再考虑到基站信息为短期记录，可能是游客。

抽样 2 中可以很清晰地看到有多个信息测度数量较大的点并集中在一起，初步推断为工作地与居住地，类似的这种行为模式分布可以拿来进行聚类和进行实体识别，接着再归纳出社区特点。

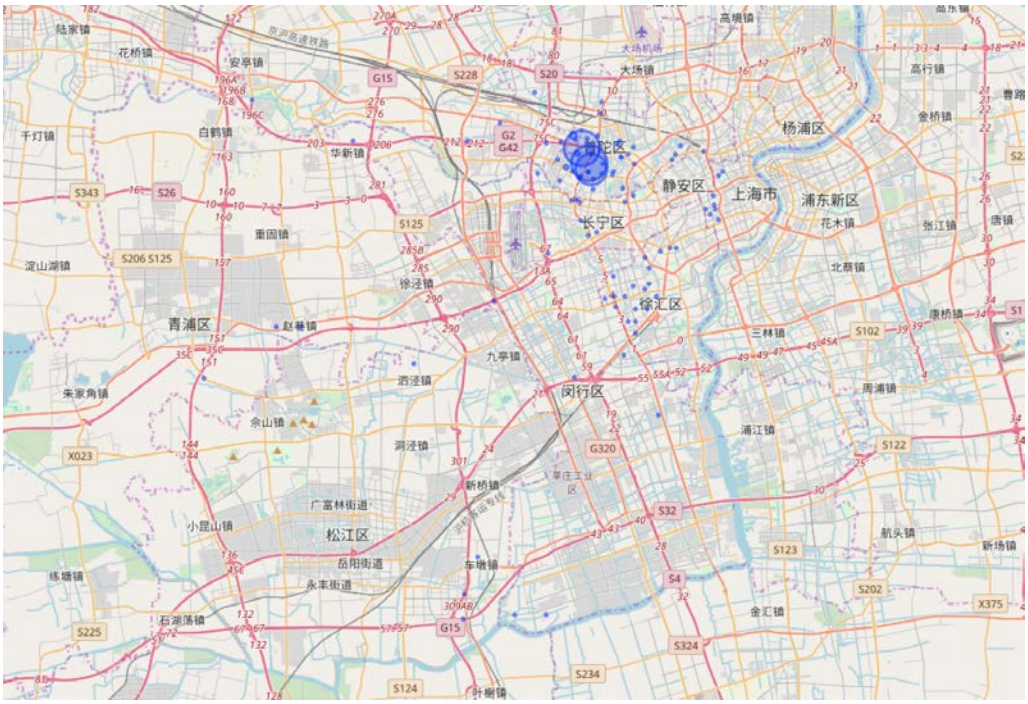



图 7 抽样 2：个体基站记录活动

# 四. 模型探究与结果分析

## *A. 基于区域熵的地点研究*

区域熵是生态学内用来描述区域内物种多样性的概念。区域熵越大，说明这个地点的生物多样性越好；区域熵越小，说明这个地点的生物种类越单一。

在这里，我们利用每个用户在一个基站周围被统计到的次数来计算区域熵。具体计算公式如下：假设$C_i$是基站$i$所对所有$n$个用户计数的总和，$q_{ik}$是用户$k$在基站$i$被记录的次数，

$$C_i = \sum_{k=1}^{n} q_{ik}$$

$$w_k = \frac{q_{ik}}{C_i}$$

$$S_i = \sum_{k=1}^{n} w_i \cdot log(w_i)$$

首先，我们计算出所有地点的区域信息熵分布：

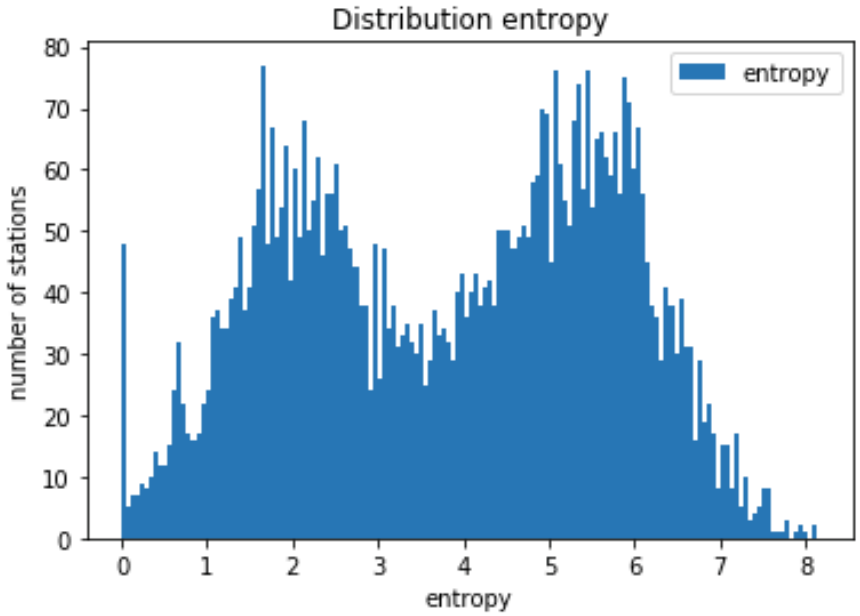
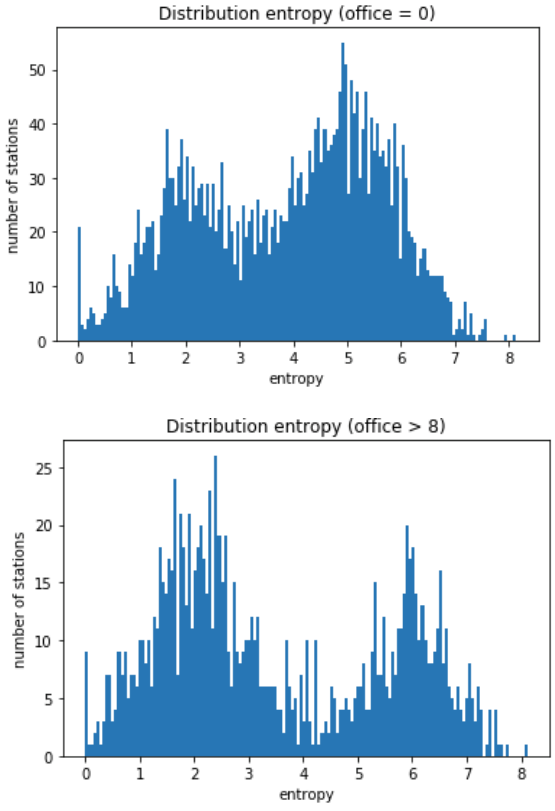



图 8 所有地点的区域信息熵分布

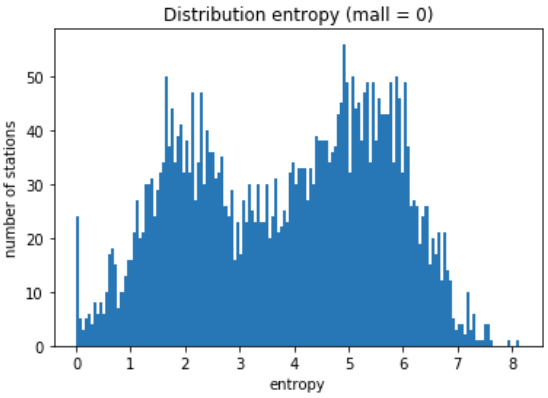
可以看出，在熵值为 2 和熵值为 5.5 附近分别有 2 个峰值。于是我们推测，不同类型的地点在熵的分布上有差别，大致可以分为商业性地区和居住型地区。

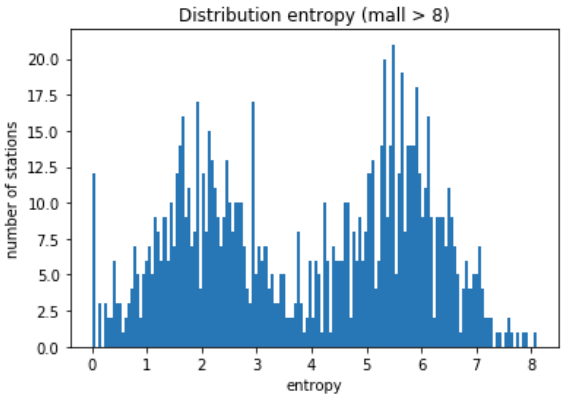
于是我们根据周围设施进行计算，分别计算出周围无办公楼和周围办公楼数量>8 的楼的基站附近熵的分布：



可以看出，在办公楼数量等于 0 的时候，熵值约为 5 时出现最高峰；而办公楼数量大于 8 的地点，熵值最高处约为 2。说明办公场所的人员较为固定。

从以下两图中可以看出，当商场数为 0 时，熵值分布于全部节点的分布几乎一致；而当商场数量增加时，地点在熵值更高时处聚集。




不过，两个峰的现象依然很明显，我们推测是因为一个基站的覆盖面积较大，里面各种功能的设施都比较齐全，所以很难将不同种类的地点完全分开。因此，我们得出结论：上海市各地的基础设施分布较为均匀，在同一个区域内能比较好地完成商业、居住和休闲等功能。如果有粒度更小的数据，还可以对不同类型地点区域熵进行更进一步的研究。

### *B. 利用网络结构进行城市空间结构研究*

上海各个地区的发展程度不一样，城中心较发达，人口集中，基础设施密集，而在郊区，人口稀少，基础设施也较少。随着上海经济的发展和人口资源的流动，城市化建设也在改变上海的市中心与郊区的分布。

行政区划是政府部分根据地区发展和人力、物力资源的空间分布对城市的划分。然而，由于上海的快速发展以及市场经济的影响，当前的人群活动区域与真实的行政区划未必一致，如何快速、准确地获取人群活动区域是新型城市化建设面临的另一挑战。

为此，我们可以使用移动互联网用户手机连接基站的记录对城市空间结构进行感知。为了度量上海城市空间网络结构特性，我们引入了模块度这一度量指标。

模块度的定义是：

$$Q = \frac{1}{2m}\sum_{i,j}\left[w_{ij} - \frac{s_i s_j}{2m}\right]\delta(c_i, c_j)$$

其中$\delta(c_i, c_j) = \begin{cases} 1, u = v \\ 0, u \neq v \end{cases}$，$w_{ij}$是节点 i 和节点 j 之间的权重，$s_i = \sum_j w_{ij}$是节点 i 的强度表示所有与节点 i 相连的边的权重之和；$c_i$表示节点 i 所属的社区；$m = \frac{1}{2}\sum_{ij} w_{ij}$为网络中全体边的权重之和。模块度表示的物理含义是社区内的节点连边数与随机情况下的边数之差，它是知识复杂网络汇总社区结构特征的重要指标，在真实的具有社区结构的网络中，模块度大约在 0.2 至 0.7 之间。同时，模块度也是复杂网络汇总评估社区划分好坏的度量方法。

首先我们利用现有数据构建复杂网络，以基站为节点（nodes），为了减轻算力负担，我们将基站之间如若有记录超过 10 位相同用户定义为一条边（edge），并将共同记录的用户数量作为边的权重。这样我们的复杂网络中包含了 2923 个节点与 296985 条边，经过计算可知图密度为 0.07，并且度分布如图：

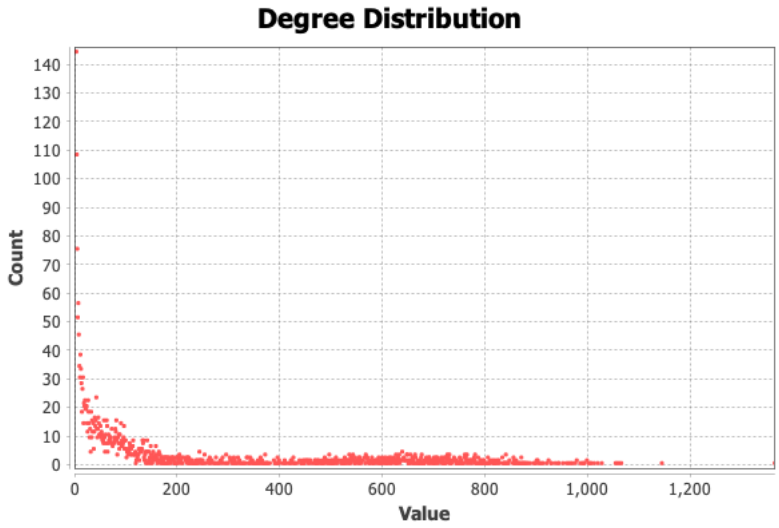



图 9 网络度分布

平均度为 203.206，平均聚类系数为 0.763。下图展示了该网络的拓扑结构。

**图 10 网络拓扑结构**

从网络中我们可以发现，呈现中间集聚四周飘散的特点，这与我们描述性分析中的分布趋势一致，向外伸展出去的簇可以被视为剔除了边缘地区的连接较少的基站数据后的连接趋势视图。整体来讲上海的城市化是从中心地带逐步向外扩张的。

接下来我们对该网络进行社区划分来研究人口活动与空间特征，采取计算模块度的方式，我们将该网络分成了 20 个社区，其拓扑结构如图所示：

图 11 模块化社区结构

基本与我们的猜想符合，因此，上海市的城市空间特征可以被归纳为以中心城区为核心向四周散步影响力的城市特征结构，尽管没有时间数据，但是人口活动特征区域也随此空间特征而固定

接下来我们将其可视化，不同的颜色代表不同的特征分块，同一颜色的区域说明此区块中的基站构成一个稳定的社区，即拥有较为稳定的人口活动与城市连接。并且我们发现其划分特征与上海的行政分区极其相似，这也在从侧面论证了我们利用网络结构对城市空间特征进行描述的可行性与实用性。

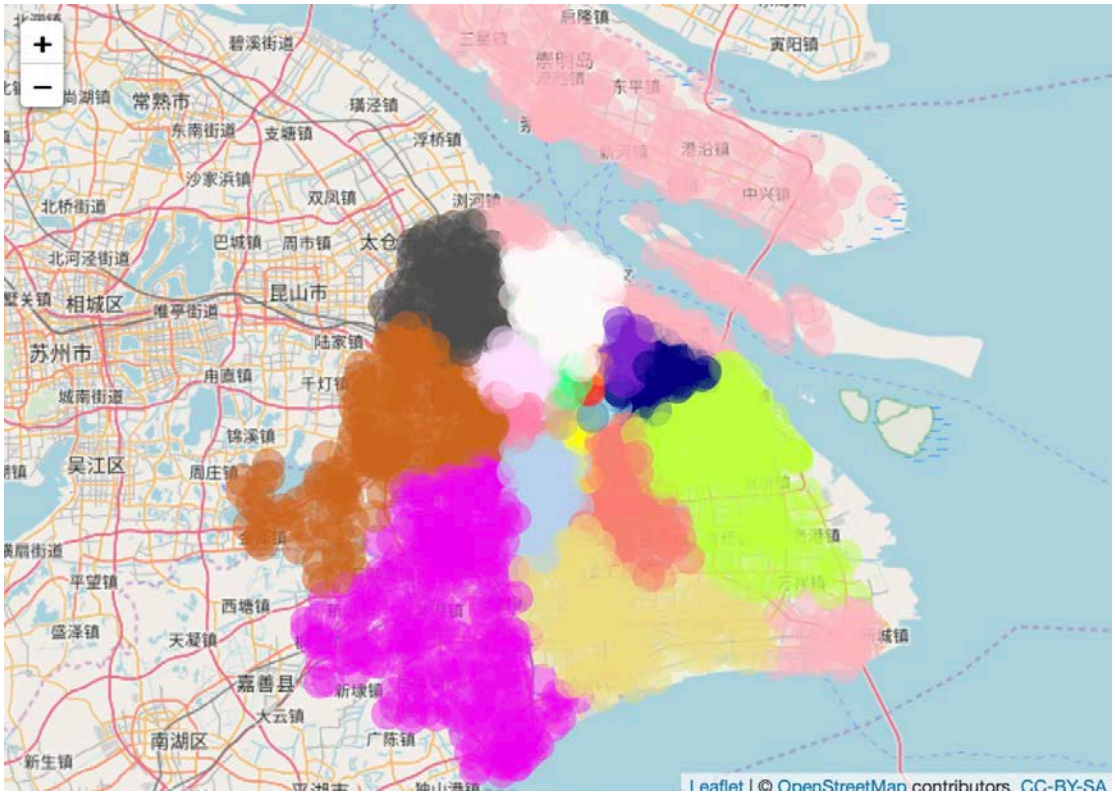



图 12 上海市区块划分

接着，我们统计不同区块中所出现的基础设施，分别用相应颜色表示对应区块。可以发现，核心区域社区的设施总数量超过了边缘社区的数量，证明城区之间的设施配备仍存在一定的差异，这对于我们的城市基建规划有一定的辅导作用。

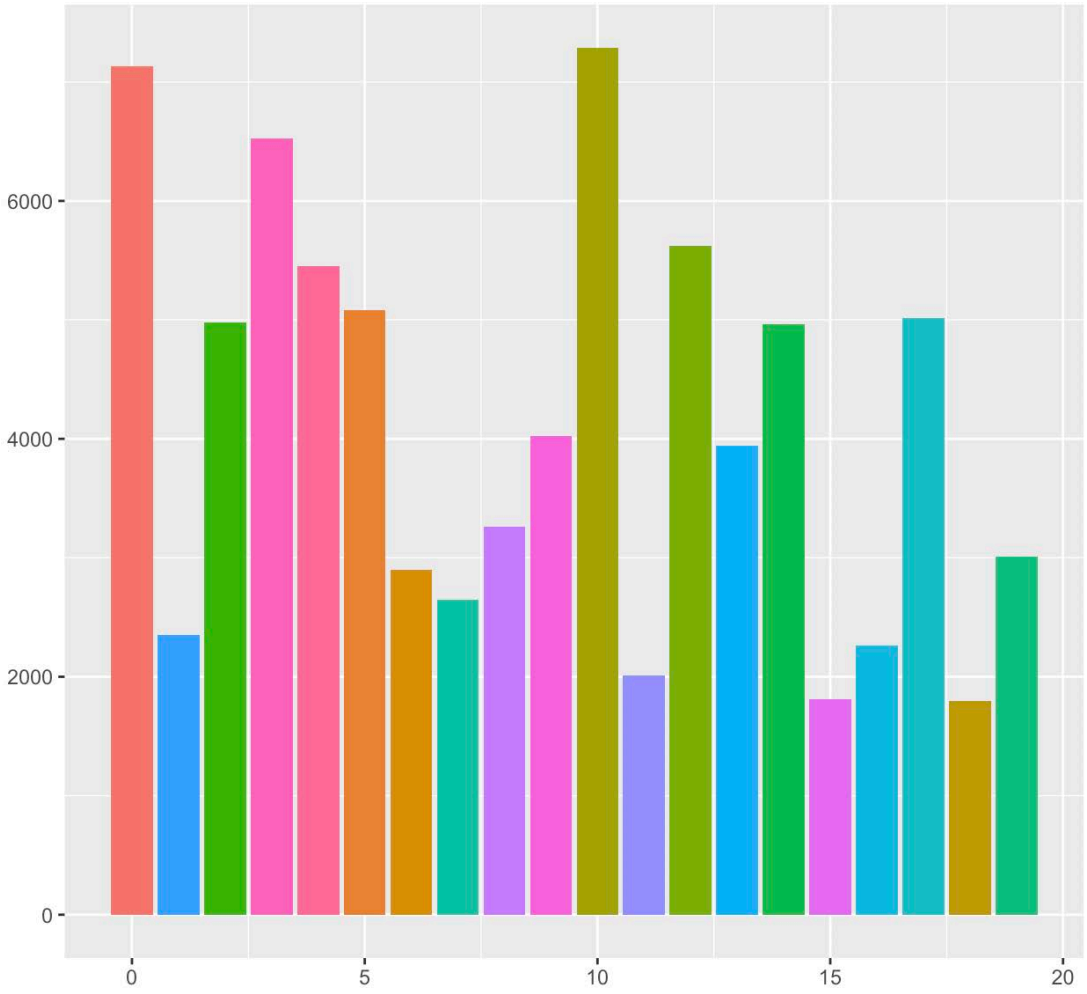



图 13 对应基础设施数量

## 五. 不足与改进方向

我们受限于算力仅截取了权较显著的边进行研究，进行全局的考虑显然更加有助于我们更精准的发现边缘地带与发达地带，同时由于缺少时空完整数据，在具体的人员流动上面无法作出有效的建模，因此城市空间特征并不完善，希望能够有更加完整的时空数据；此外，在模型方面，有了时空数据，我们能够建立的模型将更加精确。