

復旦大學



本科生课程论文

题 目: 手机基站信息的经济分析

学 院: 大数据学院

专 业: 数据科学与大数据技术

姓 名: 沈嘉伦

学 号: 16307110030

任课老师: 吴立波 刘庆富

作者签名:

一、项目背景

据工信部调查，截止 2020 年 3 月末，全国三家基础电信企业的移动电话用户总数达 15.9 亿户，移动电话基站总数达 852.3 万个¹。在此如此高的普及率下，由于人机交互产生大量数据，其中手机信令数据是比较重要的一部分：基站会定时记录用户的地理位置，如距离用户最近的基站会每隔两小时会进行一次记录；而由于人的移动，在一段时间区间内，我们可能会观测到不同基站记录的频次。例如一个月内，个体 A 在 1 号基站被记录 10 次，在 2 号基站被记录 30 次……在此条件下，可获得关于个体的频次分布，显然不同个体的分布存在显著差异。而由于这一记录的分布来自用户的移动，因此通过数据可挖掘个体移动信息，从而反映例如经济景气、城市人群流动性、居民的居住与出行特征等经济现象。

二、分析的经济逻辑与方法

本研究使用了 78146 个用户在 2016 年 11 月的基站观测信息（共 4093796 条）。原始数据说明如下：

字段名	描述	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 1 mobile.csv 表单内容

字段名	描述	说明
station_id	基站编号	基站标识，不同值表示不同基站
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

表 2 station.csv 表单内容

对基站位置作出散点图不难发现，数据的覆盖范围为上海市。注意到，原始数据中不包含单条信令的时间戳，而是用户月度在单个基站的观测总数，难以从微观上分析用户单日的行为特征，或区分工作地与居住地。但是原始数据的大数据量对于从宏观上分析上海市各个区域的经济特征提供了充分的信息，因此本研究着重于对已有的手机信令数据进行可视化展示，找出手机信令集中的热点地区；以及分区统计，尝试分析手机信令信息与各区经济指标、产业划分的联系。

由于原始数据中不包含基站所属的行政区划信息，首当其冲的一步是需要根据各个基站的经纬度，获得其所属的区划信息。这里使用 R 中的 rgdal 包对互联网上获得的全国县级行政区划边界数据²进行了预处理（代码见 code/shanghai_map.R），运用 python 的 shapely 包中实现的地理围栏算法获得了基站经纬度的对应区划（代码见 code/geo_fencing.py）。

¹ 中华人民共和国工业与信息化部：2020 年一季度通信业经济运行情况，<http://www.miit.gov.cn/n1146312/n1146904/n1648372/c7869521/content.html>

² 中国行政区边界 shp 下载（省，市，县），https://blog.csdn.net/niu_dige/article/details/104856967，数据较大因此没有包含在 /data 中。

本研究主要使用了 R 对数据进行统计分析与可视化展示³（代码见 code/pj.R）。对于用户和基站层面的信息，分别进行描述统计，根据次数划分区间，观察用户与基站对应的观测次数的分布。对于每个行政区，统计基站总数、基站观测总次数，进而计算基站的平均观测次数，以此与各区的生产总值、产业情况等进行对比分析。（所有可视化结果的原图均在/result 文件夹中。）

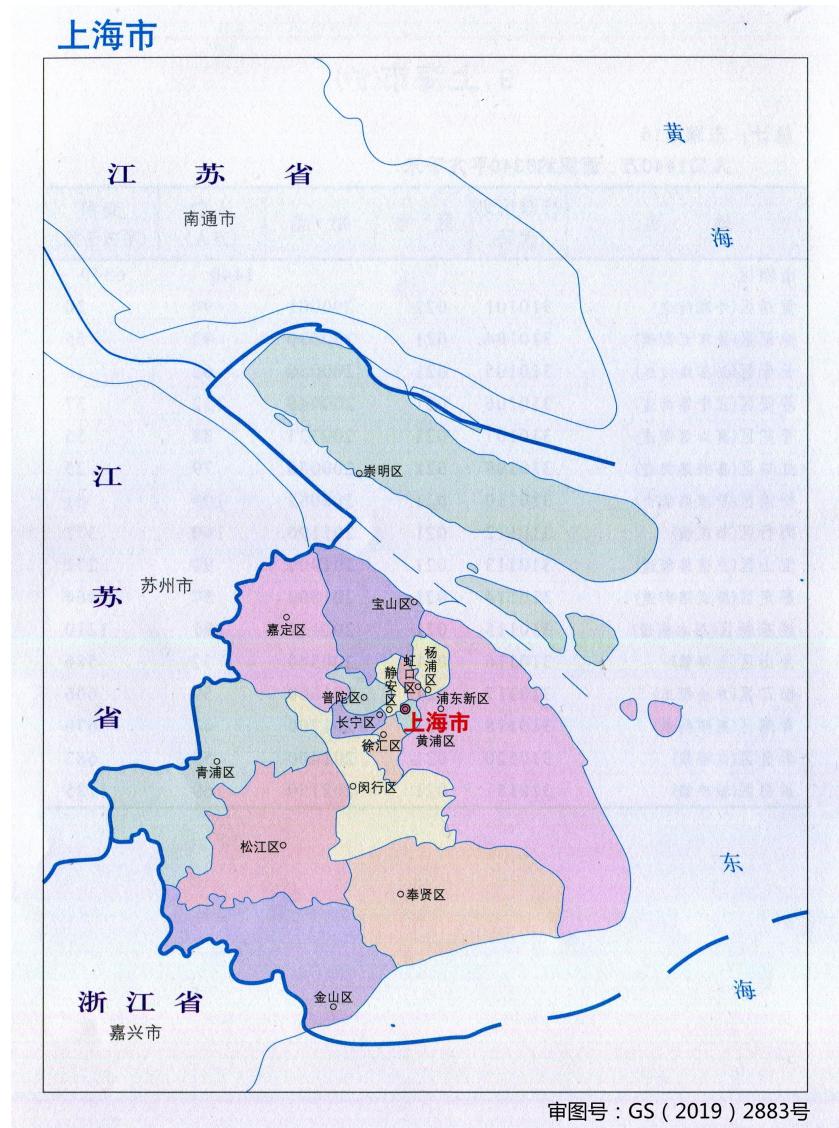


图 1 上海市行政区划

三、分析结果

3.1. 基站层面的分析结果

统计 mobile.csv 中不同基站编号累计的观测频次，得到分区间分布图如图 2 所示，频数分布直方图如图 3 所示。这里，将基站按观测总数进行了不同层次的分组，没有在 mobile.csv 中出现的基站编号对应的观测次数为 0。研究的基站总数为 6160 个（已经去除了不包含在 station.csv 中、没有位置信息的基站），其中观测总数大于 0 的基站

³ 部分可视化结果基于开源框架 kepler.gl (<https://github.com/keplergl/kepler.gl>)

数为 5295 个 (86%)，观测数为 0 的基站数为 865 个 (14%)。从图中可以看出，基站接收的信号总次数呈幂律分布，接收信号越多的基站数越少，基站接收的信号分布很不均匀，存在一些接收信号数很大的“热点”基站——接收信号数大于 4000 次的基站有 183 个，约占基站总数的 3%。

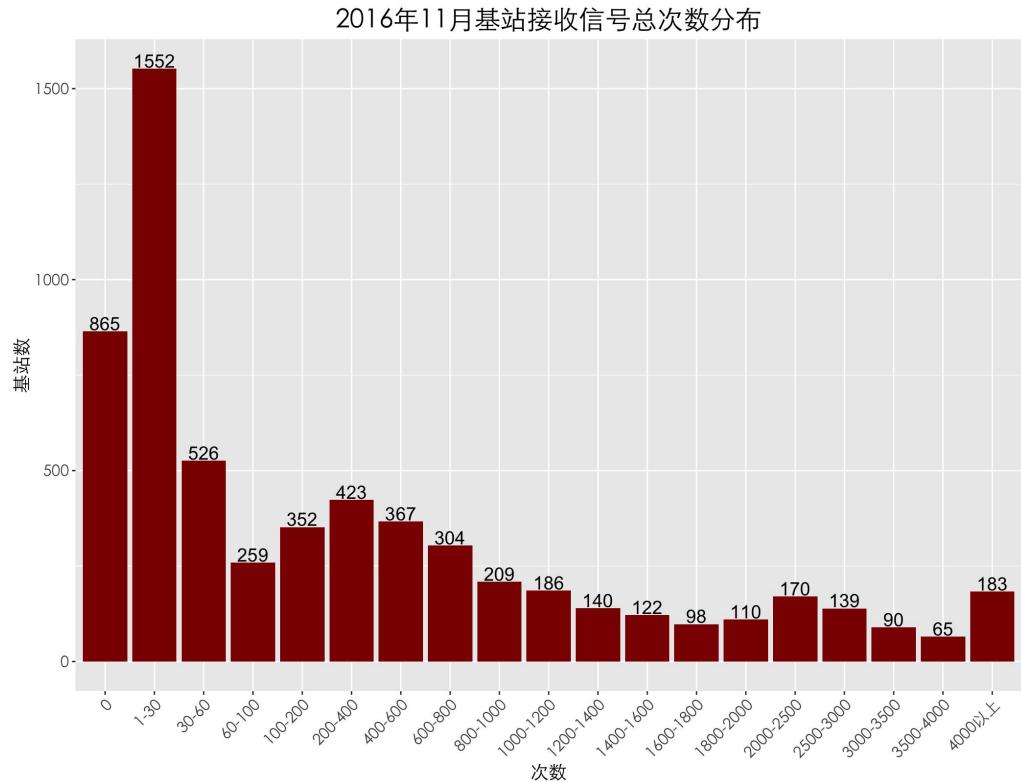


图 2 2016 年 11 月上海市基站接收信号总次数分布 (分区间)

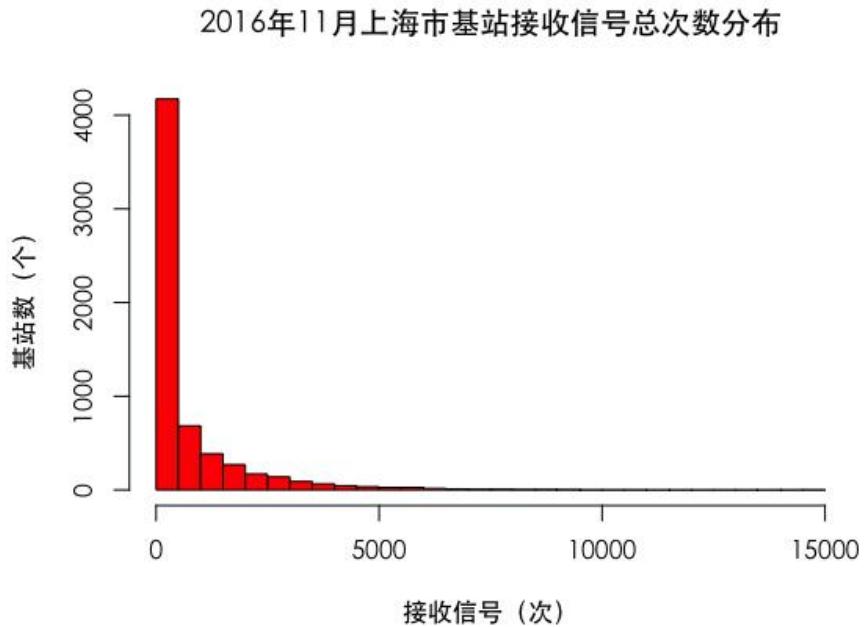


图 3 2016 年 11 月上海市基站接收信号总次数分布

首先注意到，基站本身在上海市范围内的分布就是不均匀的，仅根据基站位置，作出热力图如图 4 所示。我们可以直观地看到，不同地区的基站密度显然不同（如图 4，红/蓝色分别代表基站密度高/低）：市中心地区的基站密度明显高于市郊地区；市郊地区只有区政府所在地和人口相对集中的城镇附近基站密度较高。

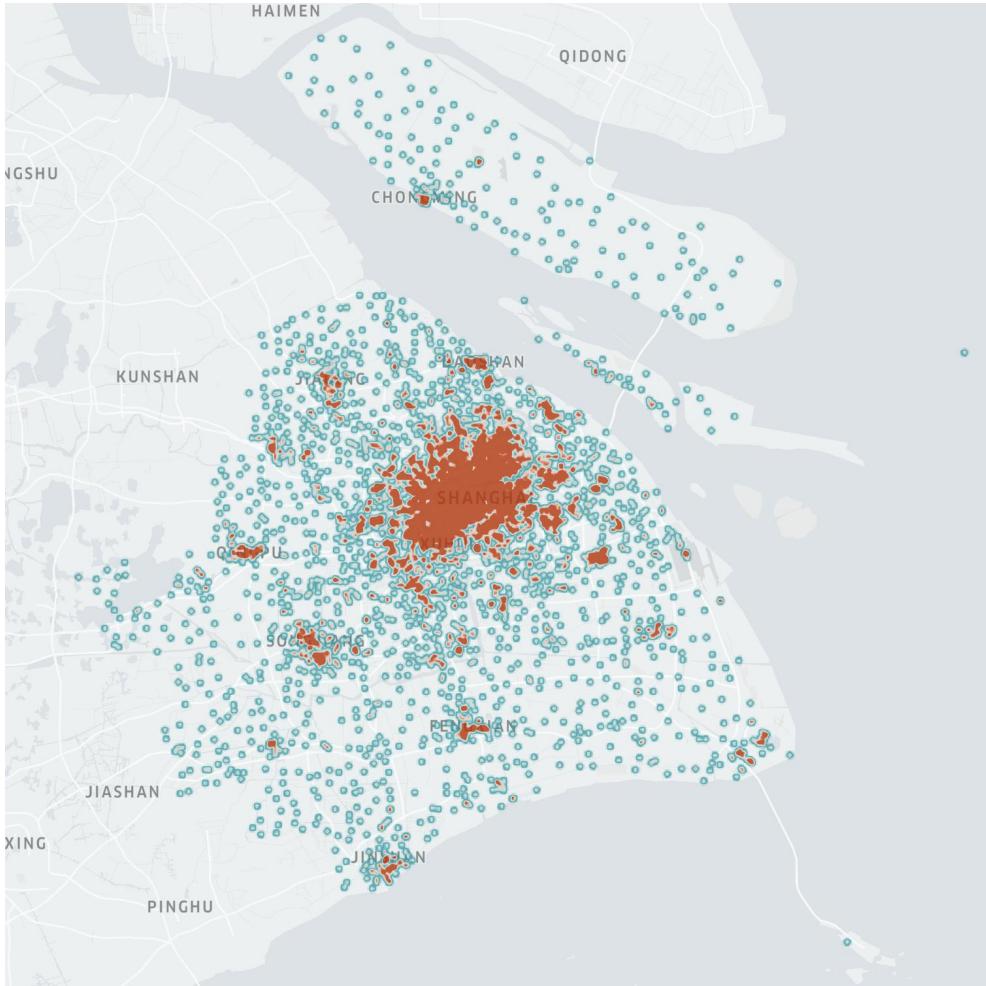


图 4 上海地区基站分布热力图

我们更加关注的是基站观测信号数与地理位置的关系。根据基站的地理位置信息和各基站接收信号总次数，作出如图 5、图 6 的热力图与散点图（图 5 中红/蓝色分别代表基站接收信号数高/低，图 6 中红/绿/蓝色分别代表基站接收信号数高/中/低，数据点越透明代表接收信号数越小）。

由于热力图会将一定范围内的数据叠加进行展示，从图 5 中可以看出基站接收信号次数的总体情况，市中心区域的基站信号数密度明显高于郊区，浦东的基站信号数密度高于浦东。尽管郊区城镇的基站密度与市中心接近（图 4），但是信号数密度普遍低于中心城区。

图 6 的散点图可以更明显地展示个别基站间的数据差异。有趣的是，接收信号数最多的单个基站出现在了市区边缘的闵行区而非市区中心；接收信号较多的一些点除了市中心外，分布在闵行、嘉定、宝山、浦东等郊区的靠近中心城区的边缘位置。

接下来。我们缩小作图范围到近郊（图 7）和市区（图 8、图 9），对基站接收信号数较多的市中心区域进行进一步考察。

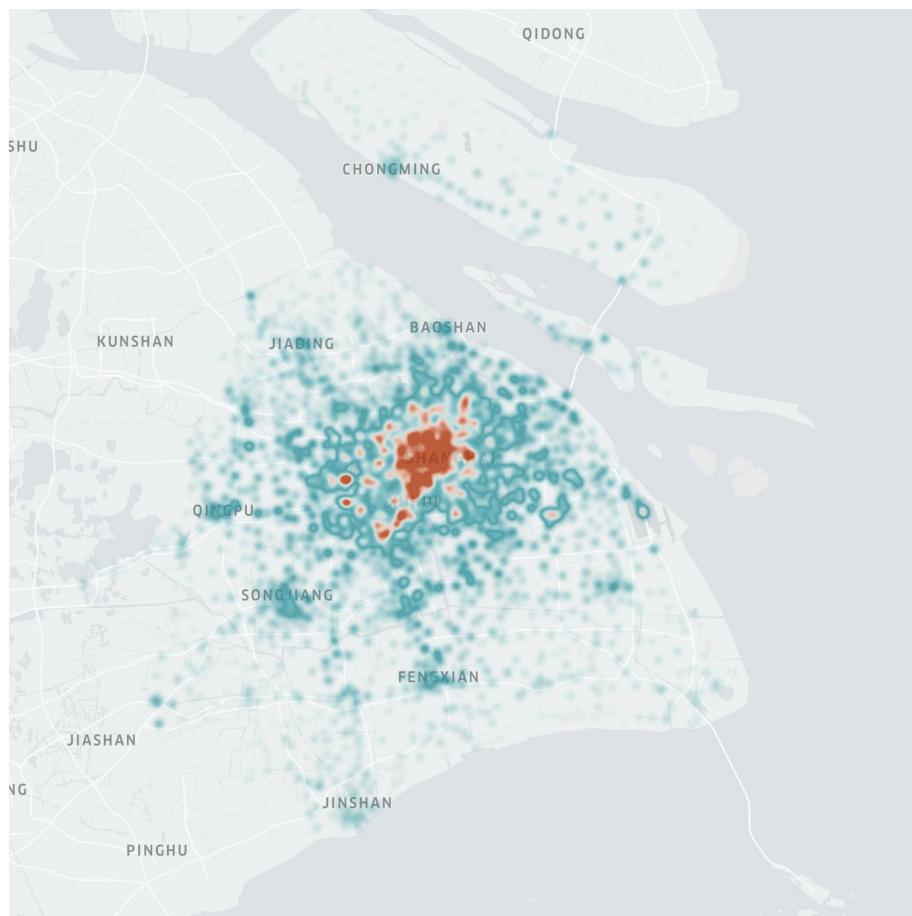


图5 2016年11月上海地区基站接收信号数热力图

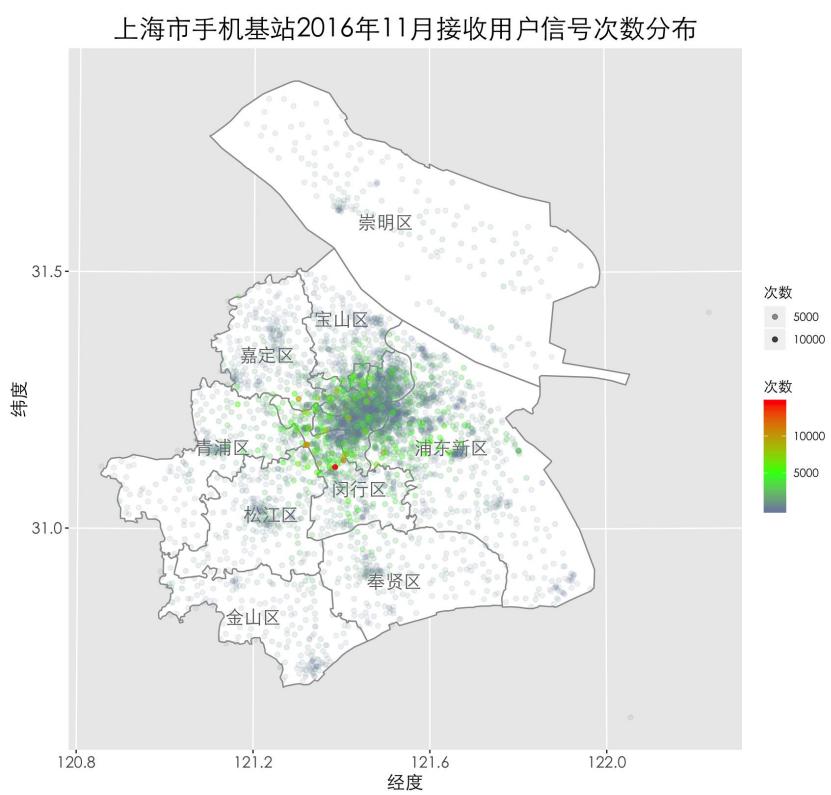


图6 2016年11月上海地区基站接收信号数散点图

图 7 和图 8 采用了暗色背景，使得路网更加清晰可见。可以发现，接收信号较多的基站的分布存在以下特征：

- ①市区外围接收信号数较多的基站主要沿郊区和市区间的交通主干道分布，在主干道的交叉点、以及人流量较大的交通枢纽附近，基站的接收信号数往往较高，例如莘庄立交、虹桥机场、上海火车站等；五角场等城市副中心的信号数也较多；
- ②市区中的信号集中区靠近徐家汇、人民广场、陆家嘴、新天地、静安寺等 CBD 和商圈区域；
- ③注意到，在浦东地区离市区较远的地方，有两个基站信号比较集中的地区，分别是上海迪士尼度假区和浦东机场。

图 9 直观地反映了市区范围内基站的接收信号数的点分布情况。市区内接收信号数量较多的基站普遍接收 5000 条左右的信号（绿色），少于市区外围数个接收 10000 条左右信号的基站（橙色），推测是因为市区内的基站数量较多，分担了信号压力；市区外围，特别是市区西南方的闵行地区的交通要道附近，单座基站需要承担较大压力；相比之下，市区内单座基站的压力小了许多（蓝灰色）。

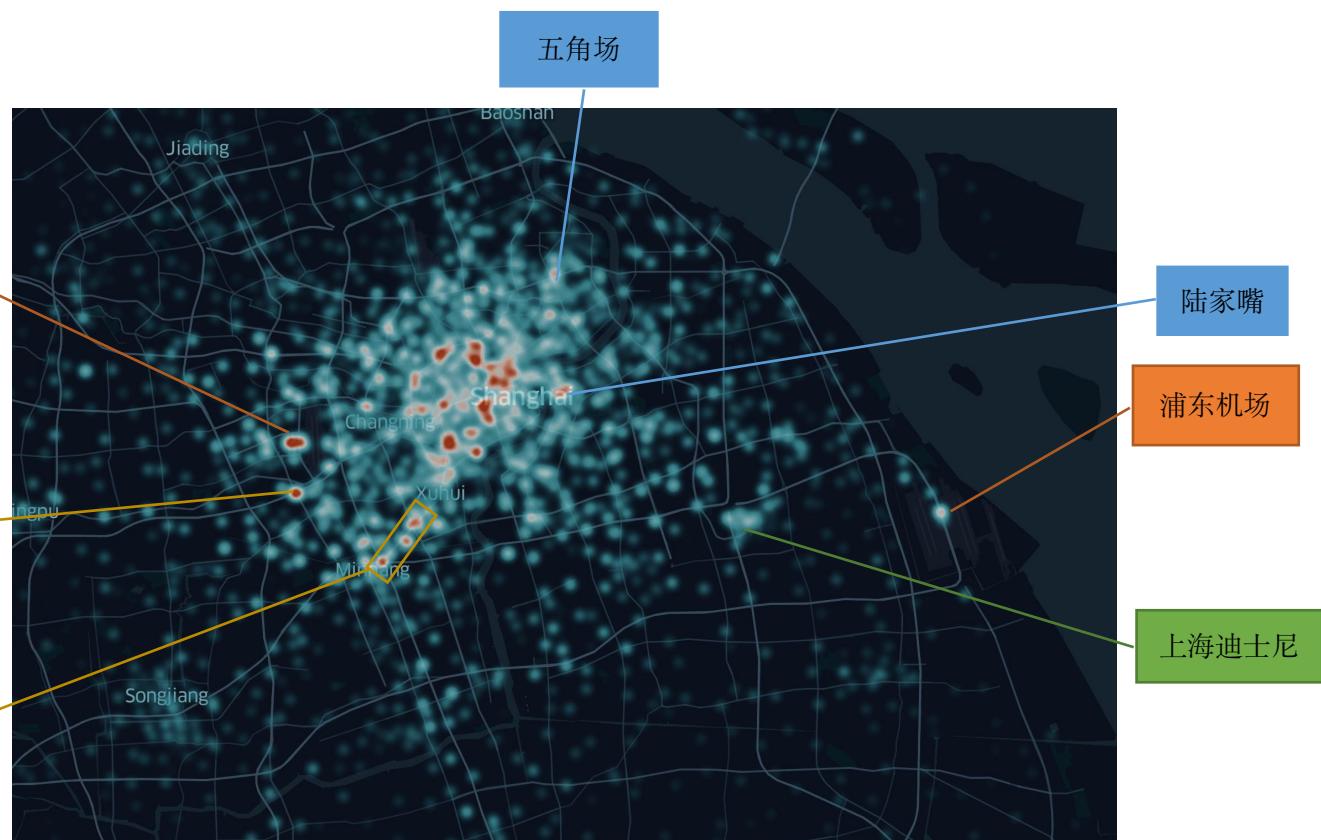
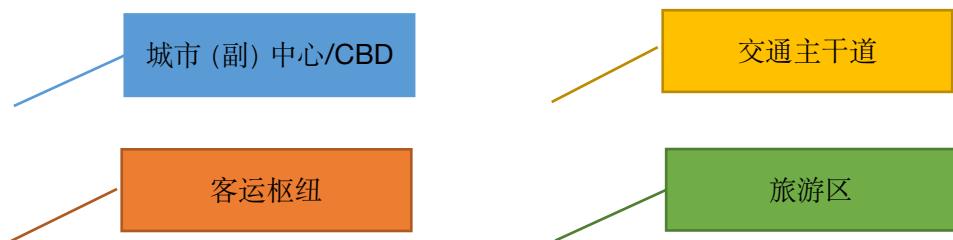


图 7 2016 年 11 月上海地区基站接收信号数热力图（近郊）

图例：



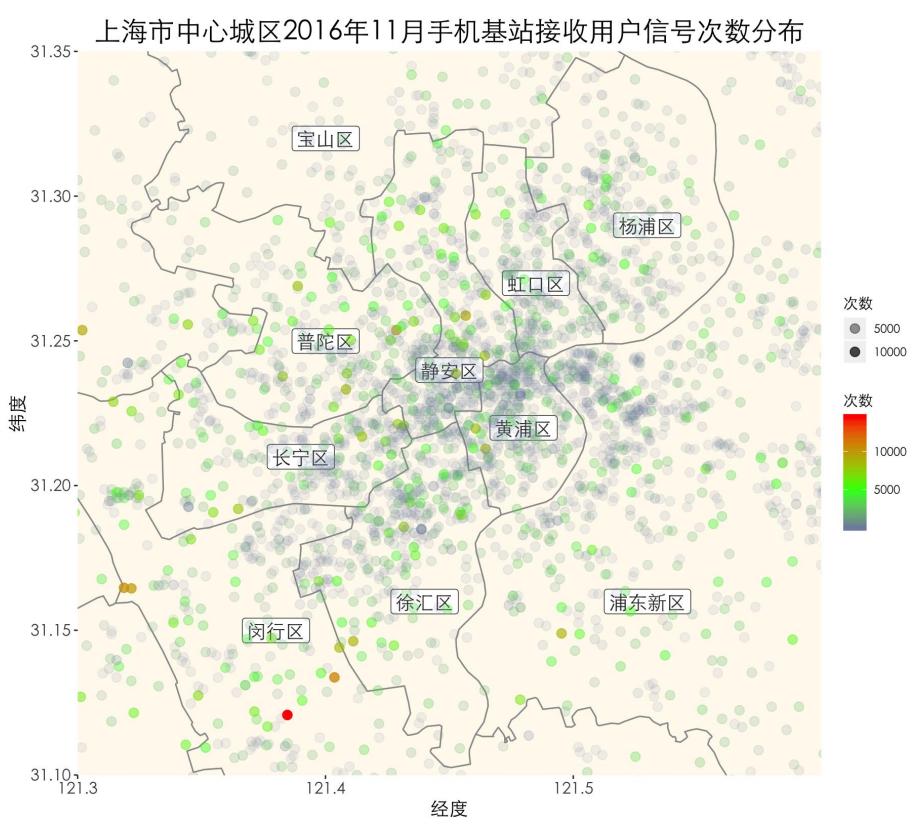
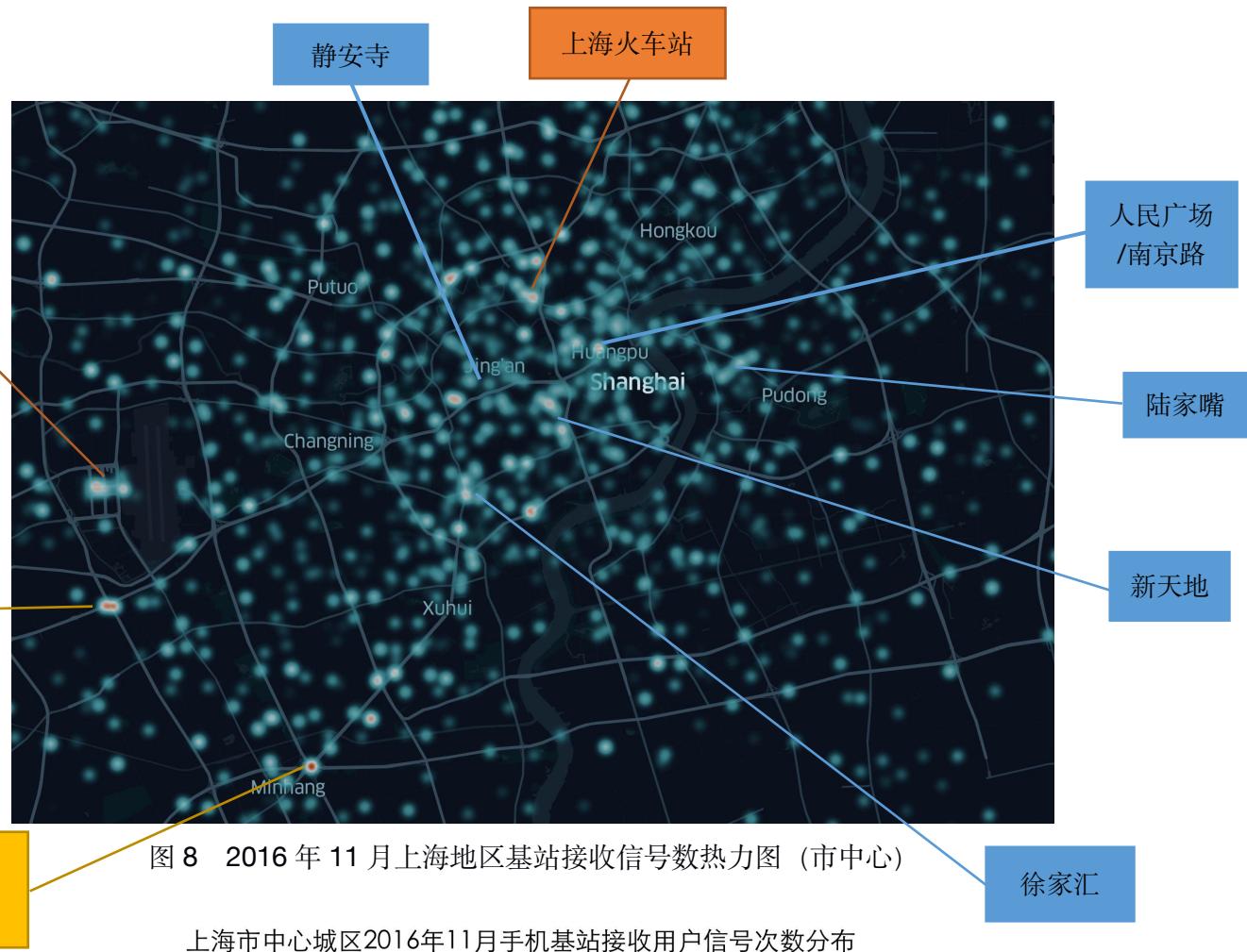


图 9 2016 年 11 月上海地区基站接收信号数散点图（市中心）

基于基站的统计数据，提取出了所研究的时间段内接受用户信号总数大于 5000 次的基站（共 104 座，数量从多到少的区前 5 名为：闵行、浦东、静安、长宁、普陀），作出的位置分布图如图 10 所示。接受信号数最多的前 10 座基站的信息如图 11 和表 3 所示。可以看到，这 10 座基站都分布在主干道旁，说明上海的交通流量确实巨大。

上海市手机基站2016年11月接收用户信号次数>5000次的基站位置

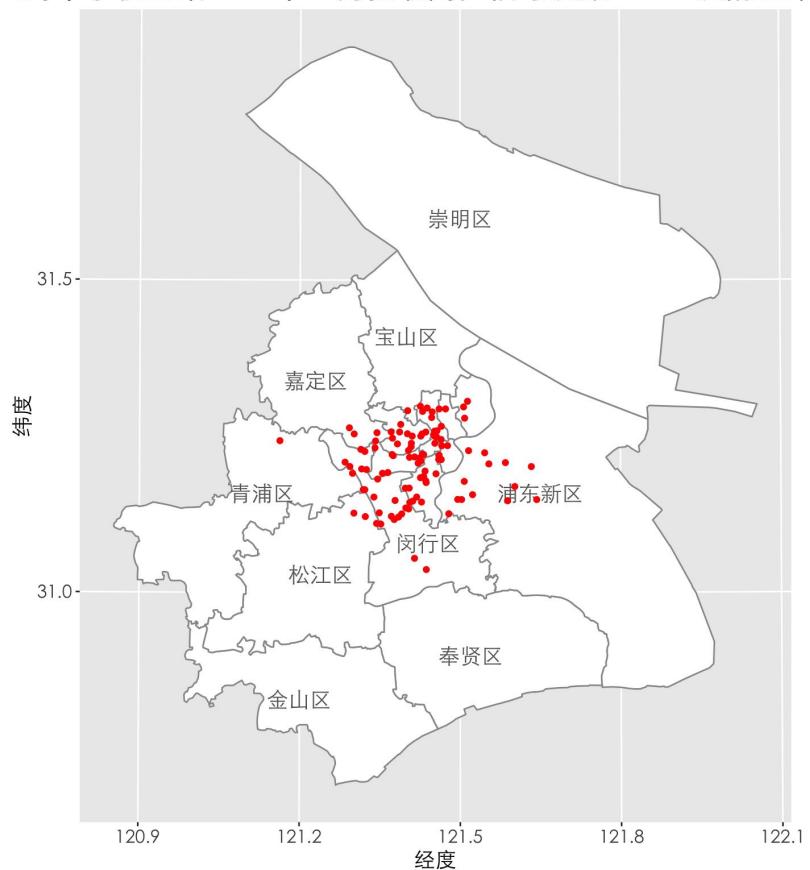


图 10 2016 年 11 月上海地区接收信号数大于 5000 次的基站位置

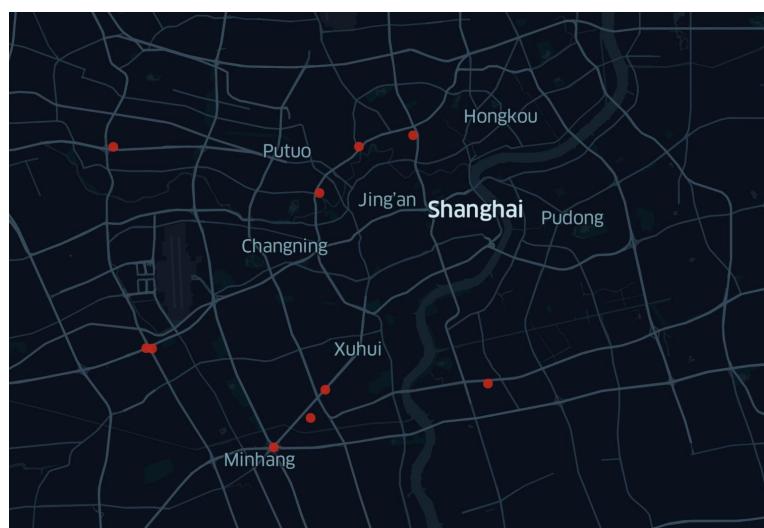


图 11 2016 年 11 月上海地区接收信号数最多的 10 座基站位置

基站编号	接收信号数	所属区	靠近的主要道路
C0947	14528	闵行区	沪昆高速、外环高速、沪闵高架路
C0966	10761	闵行区	沪闵高架路
C8648	10327	闵行区	沪渝高速、嘉闵高架路
C1095	9921	闵行区	沪渝高速、嘉闵高架路
C8347	9360	嘉定区	京沪高速、嘉闵高架路
C1009	9345	普陀区	内环高架路、武宁路
C0930	9324	静安区	内环高架路、沪太路
C0644	8836	浦东新区	中环路、罗山高架路
C0152	8827	徐汇区	中环路、沪闵高架路
C0066	8536	普陀区	内环高架路、金沙江路

表3 2016年11月上海地区接收信号数最多的10座基站信息

通过地理数据，我们可以找到不同基站位置对应的市辖区。统计每个区中的基站接收信号总数、平均接收信号数，作出如图12~图15的各区基站接收信号总次数与平均次数地图（颜色越深表示总或平均接收信号数越多）。从图上可以看出，接收信号总次数最多的是浦东新区，而平均接收信号次数最多的是闵行区。对于浦东新区，虽然信号集中的“热点”较少，但是由于其面积较大，因此接收信号总数最多。对于闵行区，沪闵高架可谓是上海的交通“大动脉”，其沿线的基站接收信号数居于上海之最，而区内基站密度相较于市中心小许多，因此其基站平均接收信号数达到了全市第一。市中心区

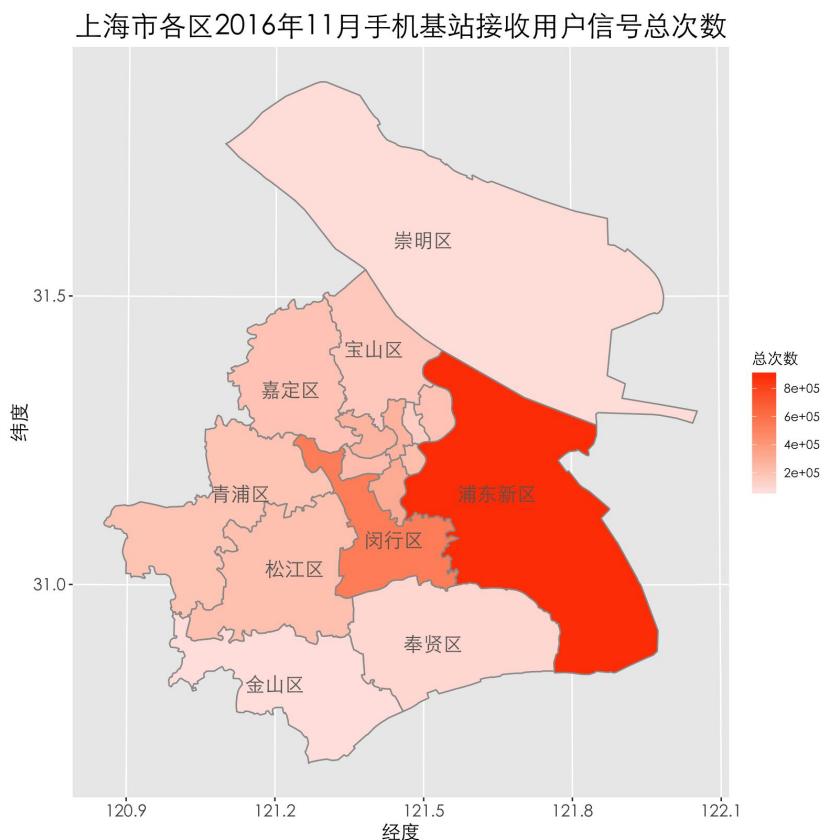


图13 上海市各区2016年11月手机基站接收用户信号总次数

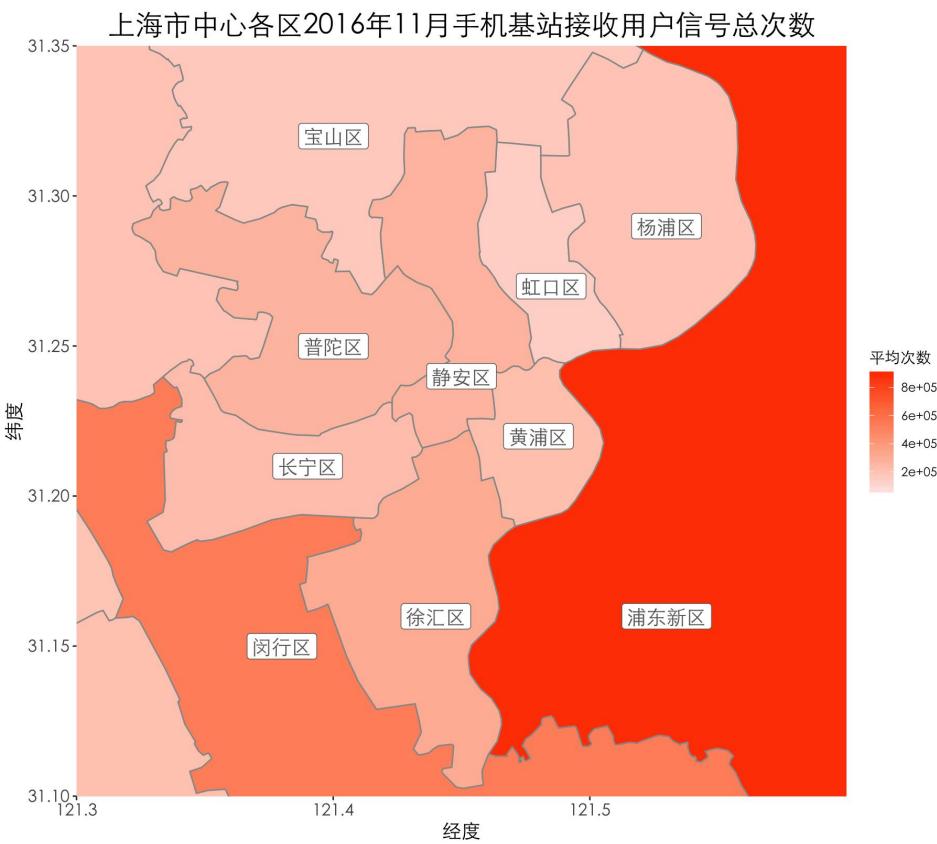


图 14 上海市中心各区 2016 年 11 月手机基站接收用户信号总次数

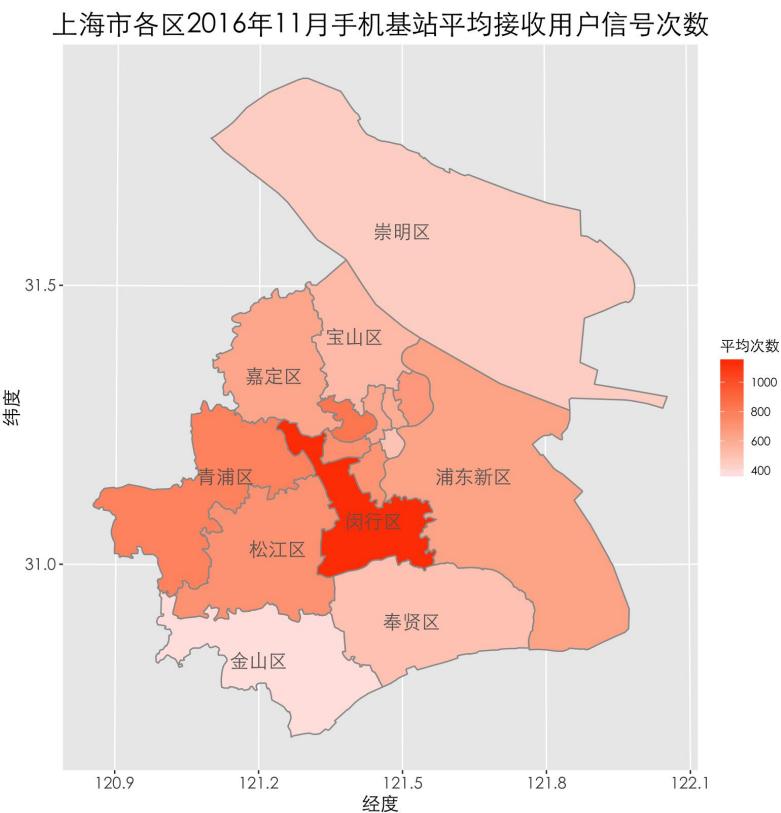


图 15 上海市各区 2016 年 11 月手机基站接收用户信号平均次数

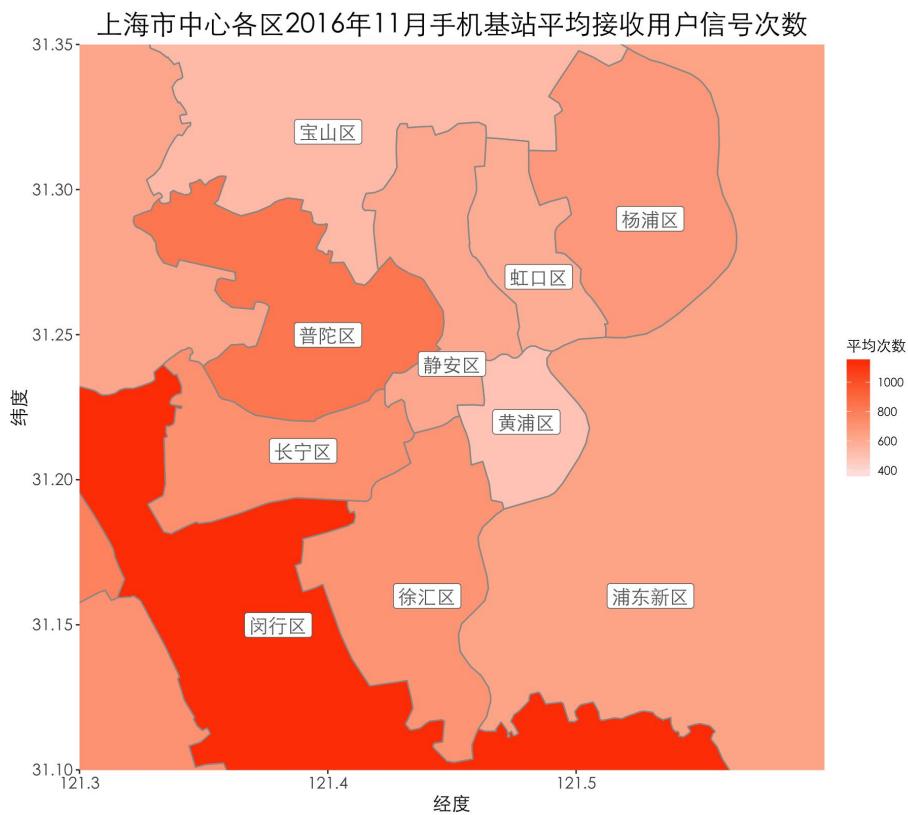


图 16 上海市中心各区 2016 年 11 月手机基站接收用户信号平均次数

结合以上基站数据和上海市各区 2018 年地理、常住人口、生产总值数据⁴，尝试分析基站的总接收信号数与区生产总值、平均接收信号数与区人均生产总值的关系。分别作出散点图如图 17、图 18 所示。由于条件所限，只能得到各区 2018 全年的生产总值数据，因此得到的结果较不具有代表性。

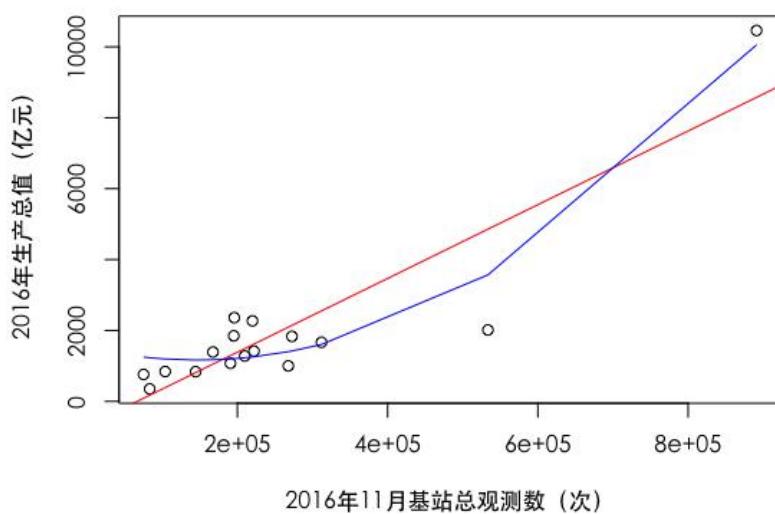


图 17 上海市各区 2016 年生产总值与 2016 年 11 月基站总观测次数关系

⁴ 来源：上海市地方志办公室，<http://www.shtong.gov.cn/node2/index.html>

尝试对上海市各区 2016 年生产总值与 2016 年 11 月基站总观测次数进行了线性与二次线性拟合，得到的拟合曲线分别如图 17 中的红色、蓝色曲线所示。线性拟合结果较差；二次线性拟合结果稍好一些，但是仍然不具有可解释性。人均生产总值与基站平均观测数之间没有发现明显的关系。上海市各区之间的经济与自然条件相差较大，产业布局也不尽相同，生产总值与基站观测数受到人口、资源等多方面影响，要探寻两者间的关系可能还需要添加更多其他变量，以及更多、更完整的手机信令历史数据。

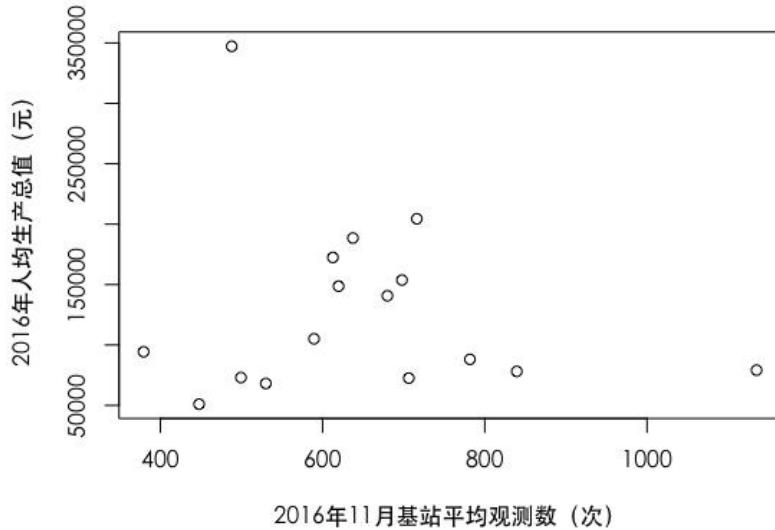


图 18 上海市各区 2016 年人均生产总值与 2016 年 11 月基站平均观测次数关系

3.2. 用户层面的分析结果

统计 `mobile.csv` 中不同用户编号累计的观测频次，得到分区间分布图如图 19 所示，频数分布直方图如图 20 所示。这里，将用户按观测总数进行了不同层次的分组。研究的用户总数为 78122 名（已经去除了不包含在 `station.csv` 中、没有位置信息的基站）。从图中可以看出，用户发送的信号总次数呈幂律分布，发送信号越多的用户数越少，用户发送的信号分布很不均匀，存在一些发送信号数很大的“热点”用户——发送信号数大于 300 次的用户有 918 名，约占用户数的 1.17%。

根据手机信令的产生原理，每名用户每隔一段时间都会向最近的基站发送一次信息，所有用户的发送总数应该接近，但是本数据的实际情况并非如此。由于缺少用户信令的时间戳信息，以及不同信令可能对应用户的不同相关操作等信息，这里没有对用户信令进行更深入的研究。

这里选择了发送信令数最多的前 10 名与前 6 名用户作为“典型用户”，对其发送信号数大于等于 80 次的位置进行了可视化展示，展现了这些用户在一个月内的移动轨迹，结果如图 21、图 22 所示。从图中可以看出，即使“典型用户”相较于普通用户发送的信令数据多了许多，其移动路径大多仍有规律可循。用户发送信息较多的地区大多可以分成两个或三个部分，呈现“家——公司”或“家——学校——公司”的“两点一线”、“三点一线”的特征。

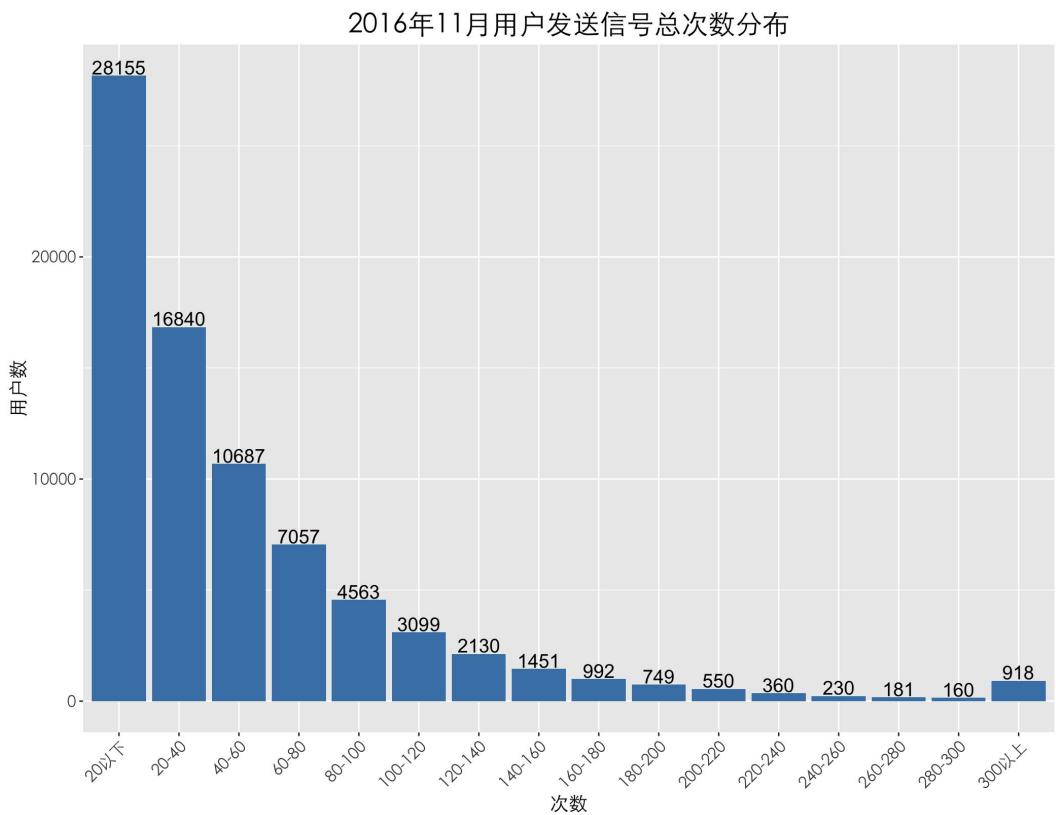


图 19 2016 年 11 月上海市部分用户发送信号总次数分布 (分区间)

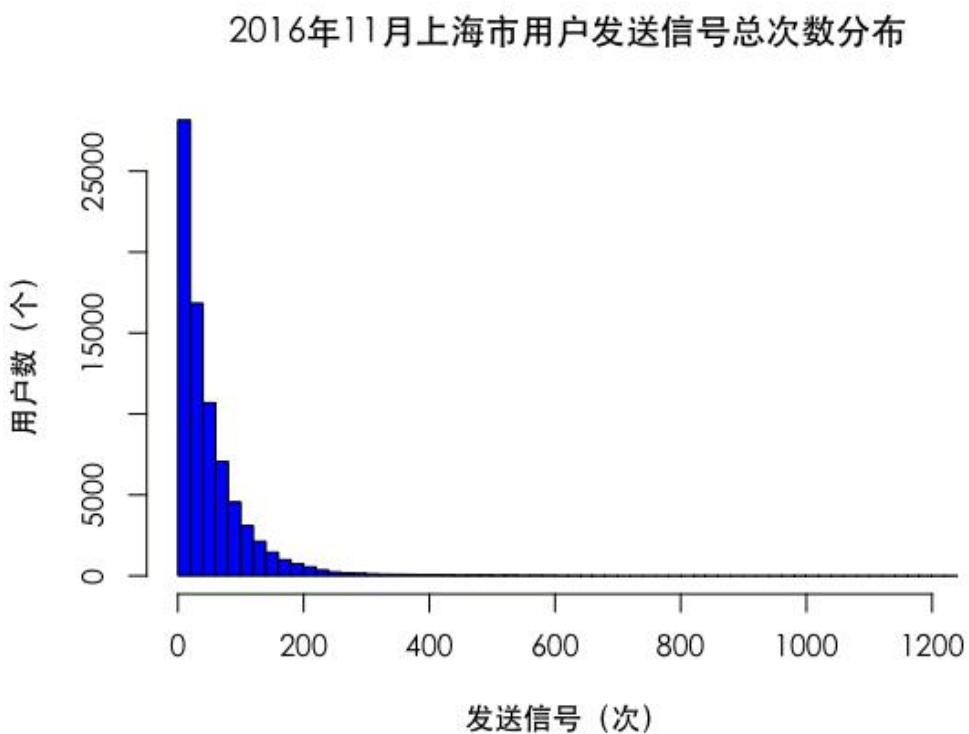


图 20 2016 年 11 月上海市部分用户发送信号总次数分布

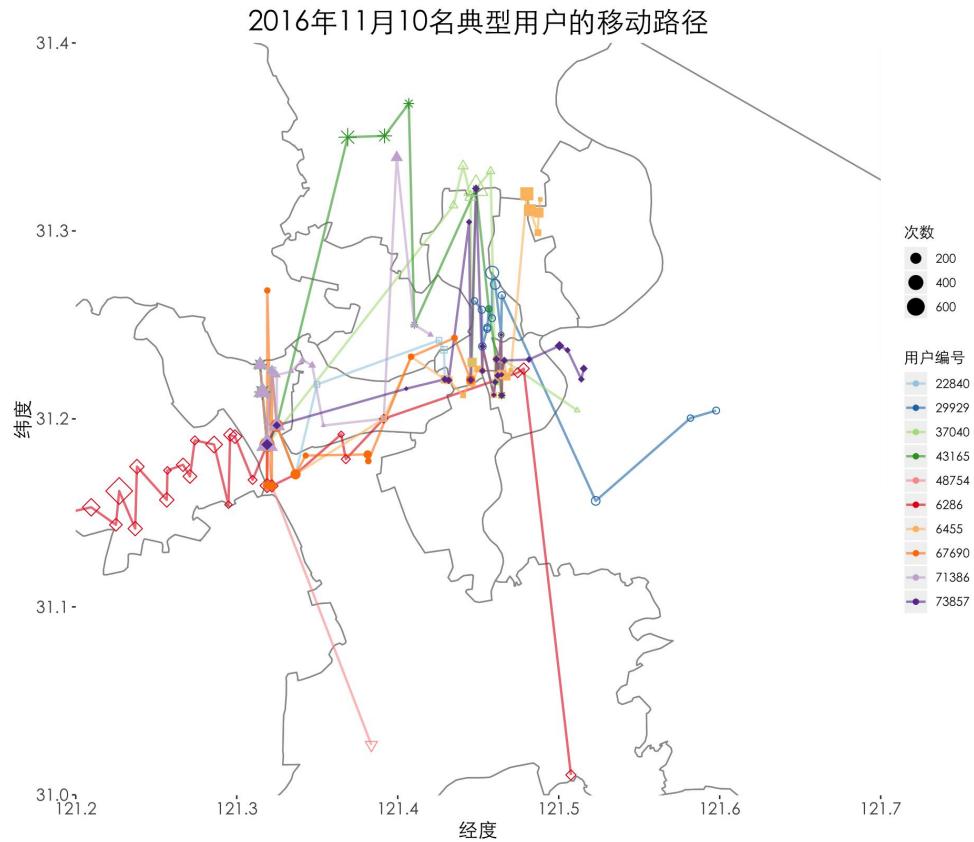


图 21 2016 年 11 月上海市 10 名典型用户的移动路径

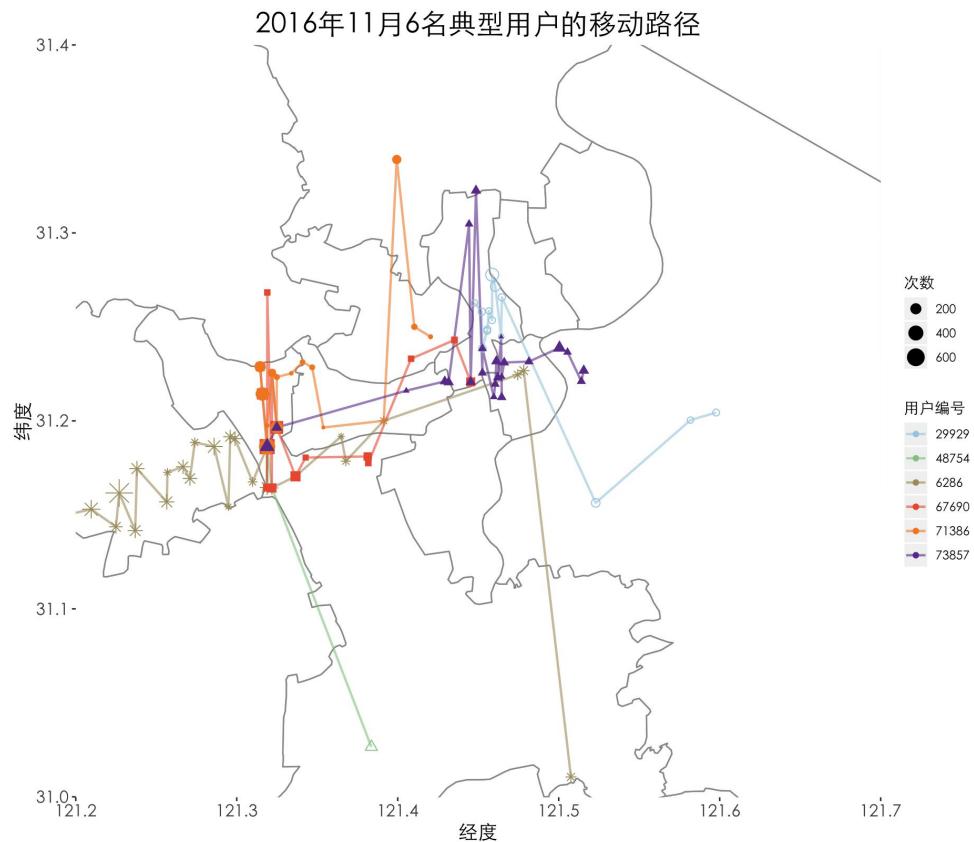


图 22 2016 年 11 月上海市 6 名典型用户的移动路径

四、结论

本研究利用了 2016 年 11 月上海市部分手机用户的信令信息与基站信息，对其进行了统计分析与可视化，发现基站接收的手机信令分布不均匀，接收信息集中的基站靠近交通主干道、交通枢纽、城市商业中心、旅游景点等场所；各区基站接收信令的统计信息存在差异，其与各区经济信息间的联系还需要进一步研究；用户发送手机信令的分布不均匀，可能受到了数据量及研究范围的局限；根据手机信令得到的用户在月间的移动轨迹呈现“两点一线”、“三点一线”特征。