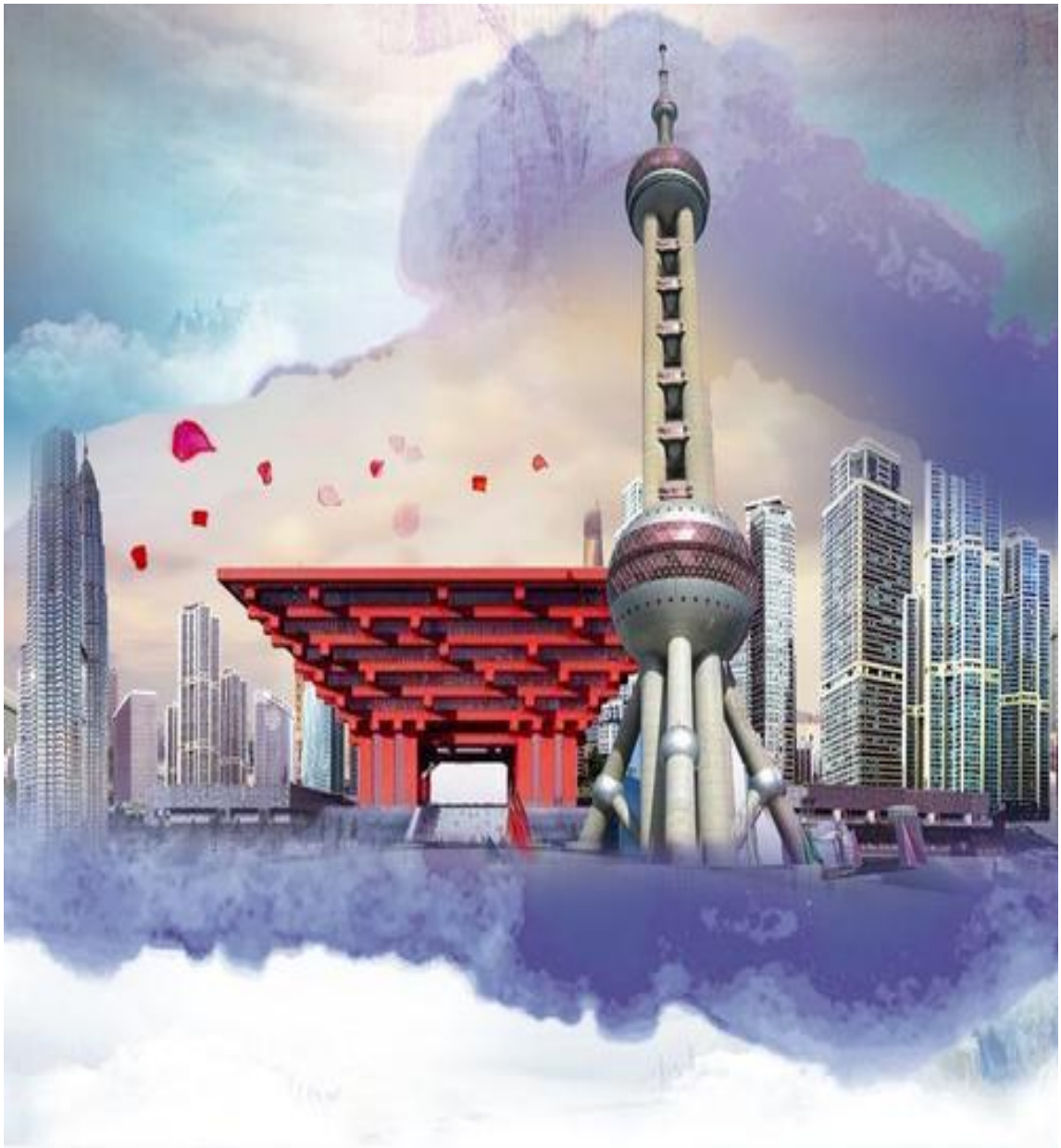


基于手机基站信息的上海市 POI 及网络流分析

钟诚 16307110259

复旦大学大数据学院

2020.6.22

0 前言

道路在人类经济活动中一直扮演着非常重要的角色，城市公路网络往往能引导居民的走向和人群的聚集，从而产生城市中不同的功能区和产业的聚集区。另一方面，城市交通网络是否科学通畅还决定了居民通勤时间的长短和工作效率的高低。在城市路网规划建设中，如何满足居民的日常交通需求是评价路网建设质量的重要指标，传统的研究方式多采用四阶段模型和非集聚模型分析对居民交通需求进行研究，而许多城市中五年一次的人口出行调查数据已经无法满足日新月异的城市规划需要。

信息与通讯技术的广泛应用产生了大量时空数据，而其中的交通和位置数据能很好的反映用户在区域内的移动特征。近几年来，移动式交通信息采集技术和智能交通系统开始成为研究热点，手机这一与现代人日常生活密切相关的通讯工具逐渐成为人口移动信息的重要来源。根据工信部公布的最新数据 [1]，截至 2018 年底，全年净增移动电话用户达到 1.49 亿户，总数达到 15.7 亿户，即目前中国人均拥有 1.12 张手机卡，而其中，上海市的人均手机拥有量位居全国第二，达到 153.9 部/百人。由于手机定位数据覆盖的地理位置广且数据量大，手机定位数据被广泛用于研究居民出行特征，交通状态探测，城市空间结构分析等领域。

本文中，作者将以上海市为例，从上海市路网结构和手机基站数据出发，尝试对上海市城市布局和路网分布的合理性进行分析，同时利用路网数据对上海市现有的经济中心和规划中的经济开发区位置进行预测。

1 研究方法

1.1 网络中心性测度指标

在本课题中，我们将路网结构视作一张无向连通图，道路可以看作图的边，道路的交点可以看作图的节点，通过对这张网络计算有关图中心性的指标，我们便可以得出哪些道路的中心性更强，从而可以推断出哪些道路的道路等级（及对拥堵的承受能力）更高。一般来说，中心性更强的道路往往代表它可以联通的道路更多，承受了更大的交通压力，疏导交通流量的能力也相应更强。以下，我们将介绍与网络中心性相关的一些测度指标，这些指标在后文中也会用到。

1.1.1 中介中心性 (Betweenness Centrality)

中介中心性由美国社会学家林顿·弗里曼(Freeman，1979)教授提出的测度指标 [2]，它测量的是一个点在多大程度上位于图中其他“点对”的“中间”。在网络中，我们通过计算网络中任意两点间最短路径经过该点的数量占两点间最短路径总数之比来计算点的中介中心性，公式如下，其中 e 为所求节点，$\sigma(s,t|e)$ 为 s,t 中最短路径经过 e 的条数

$$BC_e = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)}$$

通过这一测度指标，我们可以判断出那些在不同结构中扮演桥梁作用的节点，量化节点在网络中的控制能力。

1.1.2 接近中心性 (Closeness Centrality)

接近中心性度量的是节点到其他节点平均最短距离之和的倒数，该值越大说明该点在网络中能越快的到达其他点，相比于中介中心性，接近中心性能度量该点到其他非直连道路的

接近程度，指标值越大，说明该点的影响力越广，传递流量的能力越强。求值的公式如下：

$$CC_i = 1/\sum_{j=1,j\neq i}^{n} N_{ij}$$

1.1.3 PageRank 系数

PageRank 算法由 Sergey Brin 和 Lawrence Page 提出，最初被谷歌用于对网页重要性进行排序。之后人们对 PageRank 方法进行了各种改动，在推荐、社会网络分析、自然语言处理等领域提出了非常多实用的解决方案，是一种用于对网络中节点的重要性进行排序的算法。该算法认为，一个节点对系统施加影响的结果会作用在与它相连的节点上，即通过权重转移矩阵和权重转移的马尔可夫性实现权值在每一个节点上的收敛，收敛后节点上的权值即可以反映节点的重要性。

1.2 权重选择

本文中，我们采用基于相互关系准则的标准重要性方法（CRITIC）[3] 对模型中的参数进行估计，该方法公式如下：

$$E_k = \sigma_k \times \sum_{j=1}^{m}(1 - r_{kj})$$

其中，σ_k为第 k 个指标的标准差，即单个指标的信息量，r_{kj}为指标 k 和 j 之间的相关系数，我们将E_k归一化后，即可得到不同指标的权重。

1.3 通过基站记录数据模拟交通流量

在用户使用手机的过程中，基站会记录使用者所在的位置，形成一次计数。但需要注意的是，我们所记录的只是基站的经纬度数据，而不是手机使用者的实际位置。因此，我们需要先将基站的记录位置匹配到距离其最近的道路上，鉴于所获得的基站用户数据并不按照时间序列的顺序排列而是按照计数排列，我们使用最小生成树算法，生成了这些道路节点之间的最小生成树，对用户日常的交通移动状况进行模拟。

1.3.1 Prim 算法

最小生成树，即在联通网的所有生成树中，所有边的代价和最小的生成树。在本文中，该代价即为两点间的最短行驶距离，通过生成最小生成树，我们可以近似模拟用户在这些点上的交通移动倾向。

在 Prim 算法中，我们从某一个顶点 s 开始，在每次迭代中选择一个代价最小的边对应的点加入到最小生成树中，直至覆盖完整个连通图的所有顶点。

另一方面，值得注意的是，该算法的时间复杂度会随着需要连接的顶点数量线性上升，局限于时间和计算机的算力，本文中我们将以每个用户被记录次数排名前五的基站位置进行对道路节点的匹配和最小生成树的生成。

2 数据介绍

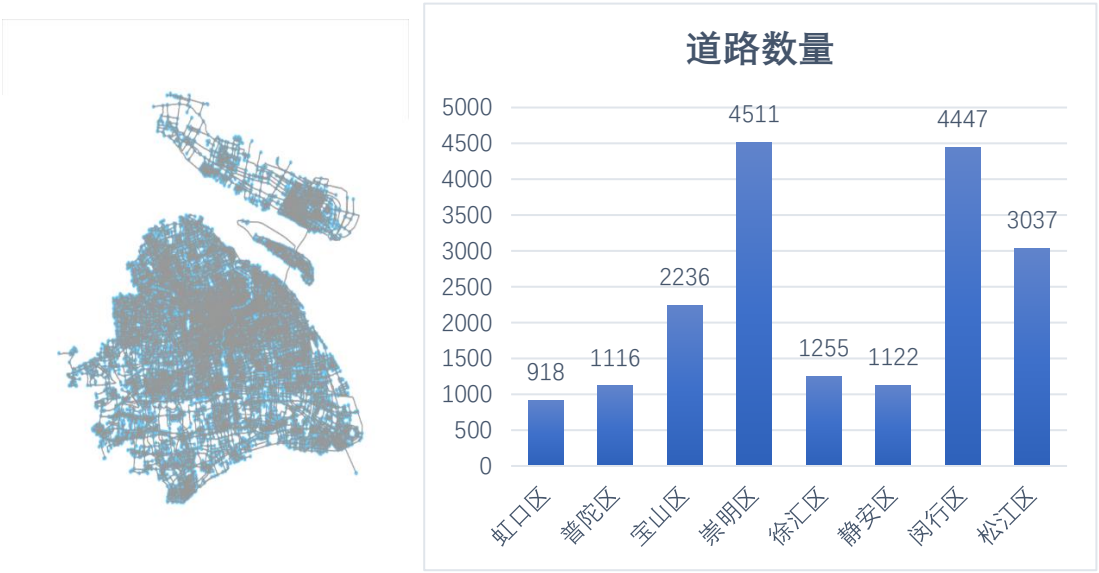
2.1 研究区域

上海地处中国东部、长江入海口、东临东中国海，北、西与江苏、浙江两省相接，界于东经 120°52′-122°12′，北纬 30°40′-31°53′之间。是中华人民共和国省级行政区、直辖市、国家中心城市、超大城市，中国国际经济、金融、贸易、航运、科技创新中心，国家物流枢纽。全市下辖 16 个区，总面积 6340.5 平方千米，2019 年常住人口 2428.14 万人，户籍常住人口 1450.43 万人，外来常住人口 977.71 万人。

鉴于上海市占地面积大，路网状况复杂，如果对全市所有的路网数据计算中心性耗时巨大，所以本文以区为单位进行中心性测度计算，并且舍弃了占地面积大且包含基站数目较少的浦东新区以简化计算。

2.2 路网数据

本文采用了来自 OpenStreetMap (www.openstreetmap.org) [4]的路网数据，采用 python 中的 api—OSMNX 进行 api 调用。通过检索和提取驾驶路网信息，共生成 18642 条路网节点数据，路网结构图如下图所示。可见，上海市路网整体呈现由中心向四周逐渐稀疏的特征，从全市来看，市中心的路网最为密集，向城市边缘逐渐稀疏；从各区来看，区中心路网向四周逐渐发散。市区，闵行区，崇明区的道路较多，虹口区、普陀区的道路较少。



图一： 上海市路网数据（左）与各区所含道路数量直方图（右）

2.3 基站数据

本文中，我们通过基站的位置数据和用户被基站所记录的次数数据进行模拟，所使用的数据集包括基站的位置数据和用户被基站记录的数值数据。位置数据中，我们得到 6161 个基站的 ID,及相对应的经纬度数据；在用户数据中，我们得到了 78144 名用户被基站记录的共计 4246793 条计数数据。其中包含了用户 ID，基站 ID 以及用户被该基站记录的次数。

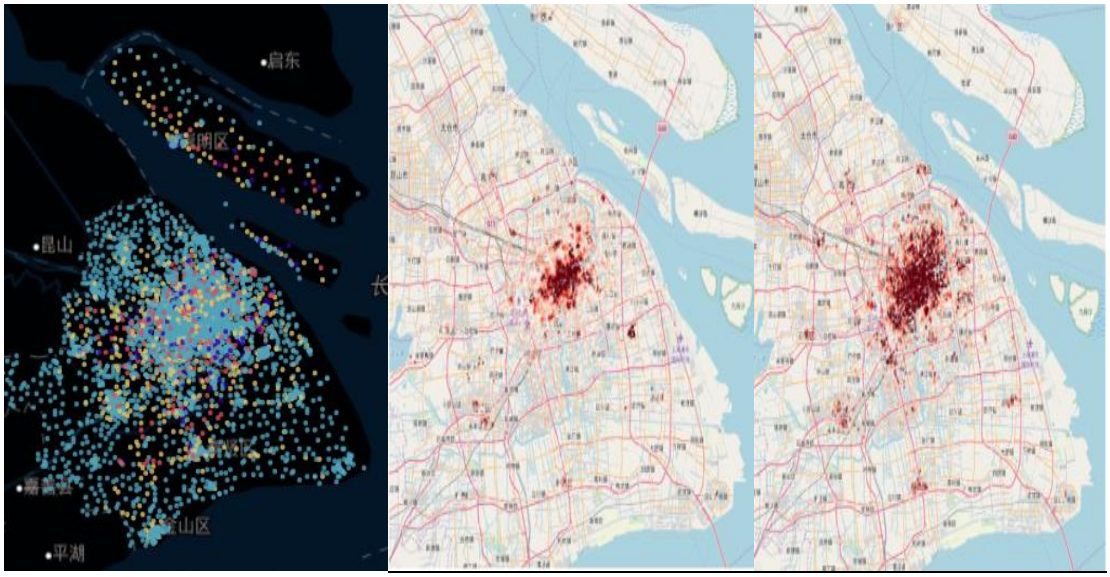
通过这些数据，我们可以对城市居民的出行情况和交通流量进行模拟。

3 基站信息挖掘

3.1 基站分布规律

首先，我们对基站分布的规律进行探究。通过百度地图的 python api 接口 pycharts，我们能够可视化各个基站在上海市的分布情况，并根据每个基站记录数据的总次数对其上色。以 1000 次计数作为分界点，从少至多赋予从浅至深的颜色。结果如下图（图二）所示，可以看到，上海市基站的分布也呈现明显的聚落特征，集聚于市中心和各个区中心，并从这些聚落向外围辐射。从基站记录数据次数的分布中也可以看出，位于城市中心相较于城市边缘的基站记录到的手机用户信息更多。

与此同时，我们通过查询上海的住宅密度分布情况和写字楼分布情况，可以推断出上海市城市分区中住宅区和商业区的分布状况，三者综合，我们可以推断出，基站的密度和住宅区、商业区的密度高度相关，在这些地区的人口密度也相对较高。



图二： 上海市基站密度图（左）、上海市住宅密度图（中）与上海市写字楼密度图（右）

3.2 基站分布规律与道路的关系

在 3.1 中我们提到，基站分布的密度与人口分布的密度高度相关，而人口密度分布与城市交通网络息息相关，通过城市交通网络，可以指引人口在区域内的流动和聚集，从而实现城市的功能分区。我们将基站频率分布图中的道路信息（省道）着重标出，并只显示计数在 1000 以上的基站位置，结果如下（图三），可以看出，上海市省道主要呈环形分布，与用户交互次数较多的基站多半分布在上海市外环高速公路（S-20）内，集聚于内环，中环，南北高架附近，外环高速公路之外则分布较少，集聚于沈海高速（G15，G1501），陈海公路（S128）沿线。而在国道和省道的交叉点则是基站和人口的集聚点，符合我们对道路和城市的一般认知。

图三： 上海市基站密度路网图

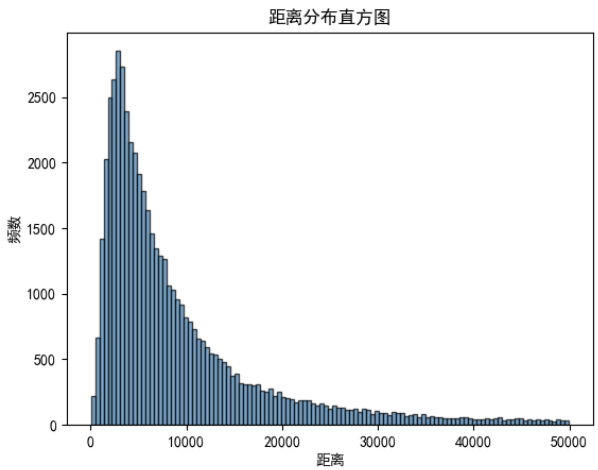
4 道路信息挖掘

由 3 中结果，可以推断，基站记录数据较多的点往往位于住宅区和工作区。换而言之，对手机用户来说，被基站数据记录最频繁的点往往是该用户的住宅和工作单位，因此，我们只需要生成连接这些点的道路连线，即可以推测出该用户从住宅到工作单位的通勤距离和线路。由常识可知，通勤需要在居民日常出行需要中占主要地位，城市的交流流量和压力也大部分来自于通勤路程，如早高峰和晚高峰，便是整个城市都要面对的交通课题。

因此，本文提取了每个用户被记录次数大于 10 且最频繁的五个基站位置，形成最小生成树以探究用户最频繁经过的道路。

4.1 活动路程

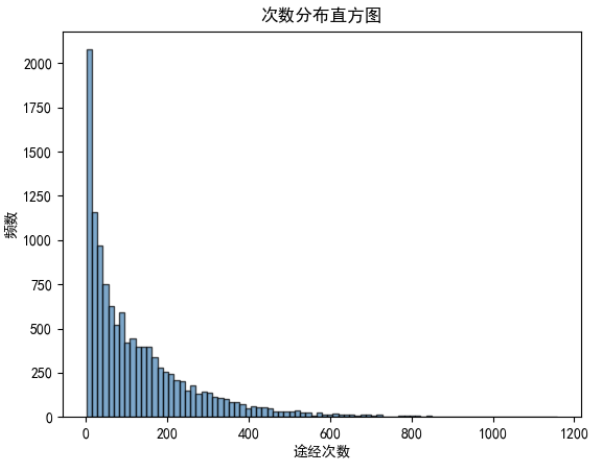
我们首先从每个用户的活动路程开始研究，通过求取每个用户最小生成树的路程，我们可以作出相应的直方图，从直方图（图四）中可以看出，用户通勤距离呈现为长尾分布，为作图清晰对大于 50km 的数据（约占整体数据的 4%）进行截断后，直方图显示大部分用户的通勤距离在 20km 以内。其中，最短的通勤距离仅 18m，可能由于基站记录数据不足所致，而最长的活动距离为近 55km，推断该用户在数据记录期间进行了一次市内长途旅行，导致距离异常增加。



图四： 用户活动距离分布直方图

4.2 活动途径道路

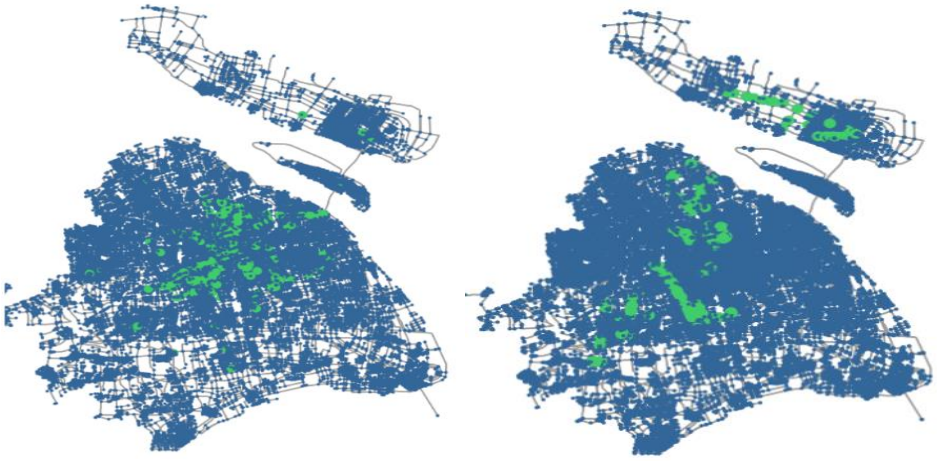
接下来，我们将探究用户活动轨迹上所经过的道路，同样的，我们先以直方图的形式对道路在用户活动轨迹上出现的频率进行呈现。如图五所示，我们共录得 12296 条在用户轨迹上的道路，在出现的频次上满足“二八定律”，即百分之二十的道路承担了百分之八十的交通流量，百分之八十的道路承担了百分之二十的交通流量，而在此之中的百分之二十的道路几乎不承担交通流量。在被记录的道路中，最多在用户的日常路径上出现了 1159 次，为淮海中路淮海国际广场一段（即与东福路交界处一段）。这段路位于淮海路商圈



图五： 路段途径次数分布直方图

内，淮海路更是上海最有名的商业街之一，云集了上海市众多顶级商场的旗舰店。毗邻上海音乐学院和上海理工大学，位于瑞金医院、复旦大学耳鼻喉科医院和上海市第一妇幼保健院之间。更重要的是，这段道路位于上海市两条高架心脏——延安高架和南北高架的交界处延安东路立交桥附近，因此这段道路的流量也相应的成为了我们研究区域内流量最大的道路。

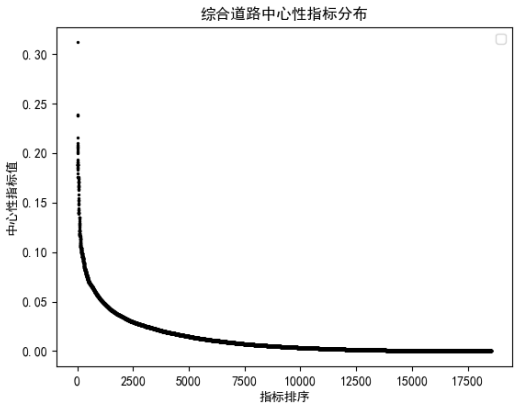
与此同时，我们还对道路节点中被记录次数位于整体前 2% 的节点进行绘图，如图六（左）所示，图中绿色节点则为被特殊标注出的节点，通过放大并将其匹配到道路地图上，我们发现，这些节点主要位于延安高架路，内环高架路、沪闵高架路沿线，外环高速---沪陕高速沿线，基站节点分布走向基本与高架路，高速路走向一致，另一方面，上海市几个主要商圈和高新技术园区，如陆家嘴 CBD、国家级高新技术园区张江高科和经济开发区漕河泾等均出现了高频记录节点的聚集，这一结果与我们对上海城市交通状况现状的认知相符，交通压力主要存在于城市高架路、高速路，高新技术园区和商圈附近的道路，少量的道路承担了大量的交通流量。



图六： 道路模拟流量前 2%路段标识图（左）与道路中心性标度值前 2%标识图（右）

5 道路中心性挖掘

在根据基站记录数据进行道路流量预测的同时，我们也可以通过前文介绍的网络连通性指标对道路的中心性进行预测，从而推断在城市规划过程中设计者们对道路流量的预估。通过 1 中所述道路中心性测度指标计算方法，我们可以计算每个点的中心性综合指标，并根据该指标从高到低进行作图，如图七所示。



图七： 道路中心性指标排序

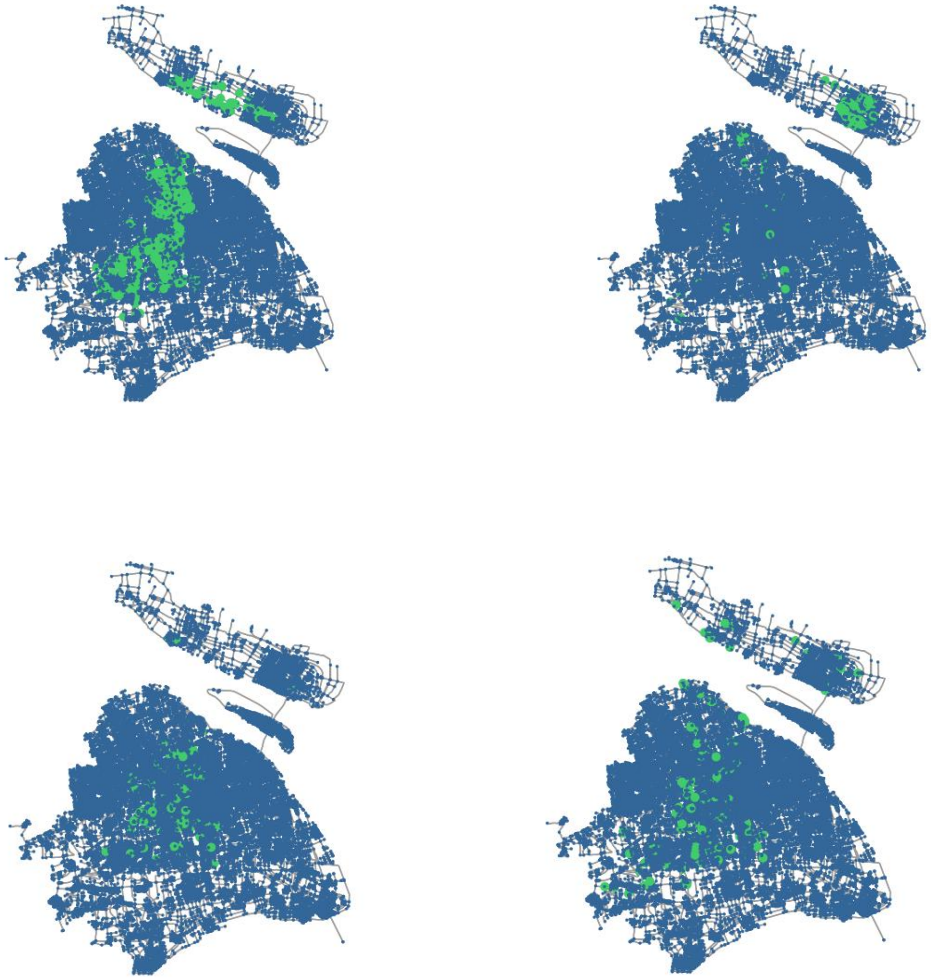
从图中可见，中心性指标分布也呈现长尾趋势，只有少部分道路中心性较高（大于 0.1），而大部分道路中心性较低，两者无论是中心性值和数量上均相差悬殊，其中，中心性高的道路主要是城市高架路、高速路等交通枢纽，而中心性低的道路多半是小区间的街道，主干道旁延伸出的辅路等等，往往路况不佳且利用率较低。道路中心性最高的节点位于静安区新闸路和石门二路交界处，恒丰路桥的延长线上。中心测度超过 0.3，远远领先于其他道路节点。通过查询资料可知，该地区位于南京西路商圈和人民广场商圈之间，通过苏州河即可到达上海火车站，是上海市交通枢纽之一，附近写字楼、商店、公司集聚，也造就了该路段的高中心性。

同样的，我们对中心性值位于前 2%的节点在地图上进行标注（图六右），并将其和 5 中得到的地图进行比较。如图右，可见，两者所标定的区域有所重叠也有所不同，在标记处南北高架之余，崇明区的陈海公路、外滩宝山路、徐汇滨江、松江大学城等区域。相比于通过 5 中通过基站计数进行交通流量推断，通过图结构直接进行中心性测度除了对主干道进行考量之外，也能探测出区域性的交通枢纽和商圈。

6 中心性测度指标和基站数据的综合推断

根据 4，5 所述，通过基站收集到的手机定位数据可以通过最小生成树进行交通流量的模拟，从而近似为城市间的实际交通流量；另一方面，通过图模型计算出的道路中心性指标可以近似为城市规划者在规划城市道路时设计的道路交通流量。根据廖薇薇等人在 2018 年的相应研究思路 [5]，综合两者，我们便可以综合出四种道路类型，在规划中道路流量高，实际道路流量也高的“高—高”型道路往往指代着已经发展较好的商圈和城市交通枢纽，在规划中道路流量低，实际道路流量低的“低—低”型道路则指代辅路和城市小道，交通流量较少；而“低—高”型道路指代在设计时道路流量小，但在实际生活中承担了较大交通流量的道路，这种道路往往容易产生拥堵，是城市拥堵治理的瓶颈环节；“高—低”道路则代表了这种道路节点的临近区域设计时期望让其承担较大的交通流量，但实际上并没能如期实现。而这种中心性较高的道路往往会导向人群的聚集和商业的发展，实际道路流量较低说明该区域现在还没有发展起来，可以预见，出现“高—低”类型的道路节点附近有成为下一个商业中心或人口集聚地的潜力。

从 4.2 中的结论可以看出，道路流量大致遵循“二八定律”，百分之二十的道路承担了百分之八十的交通流量，故此，本文中也使用百分之二十作为道路中心性的分界线，即：前百分之二十的道路被标记为“高”，后百分之二十的道路被标记为“低”。我们将四种状态的道路分别标记于上海市地图中，呈现如下：



图八： 上海市“高—高”路段标识图（左上）、“高—低”路段标识图（右上）
“低—高”路段标识图（左下）、“低—低”路段标识图（右下）

由图八可以看出，“高—高”型道路主要分布于外环高速内及崇明区的陈海公路沿线，包含南京东路商业区，人民广场商业区，淮海路 CBD，漕河泾开发区、滨江开发区、松江开发区等一系列上海市已经发展的较好的经济开发区和商业区。这些地区在城市设计的时候便作为城市功能分区中的经济区存在，故此路网设计上四通八达，以承担商业区带来的巨大车流量和人流量。而实际上，这些商业区也常常人满为患，如果没有通畅的路网系统加以支持，非常容易造成交通堵塞。同时该结果也从另一个方面证明，通过手机基站数据和城市交通路网的中心性指标，可以较精准地对城市商业中心和开发区等繁华地段进行定位。“低—低”型道路较为分散，这里不多加分析说明。

“高—低”型道路主要分布于崇明区陈海公路沿线，宝山工业园区，虹桥机场附近。在 2018 年 1 月 4 日获得国务院同意批复的《上海市城市总体规划（2017-2035 年）》中，明确提出将把上海主城区从外环内，扩大为中心城区+主城片区+高桥镇、高东镇的规划，而主城片区为虹桥、宝山、川沙和民航片区，这四个主城片区的的城市副中心地带，便是潜力无限的空白建设重地。对于虹桥机场来说，借助机场的高吞吐量进行空港园区的建设，也是上海未来发展的重点之一。对于新片区的规划和城市的发展，产业、人口、交通三者缺一不可，而

三者都离不开顺畅交通系统的支撑，因此，通过分析城市路网来预测城市在建设上的发展远景也是合理且可行的。

“低—高”型道路主要分布于上海南站附近，外滩的北京路沿线，五角场附近、虹桥枢纽到嘉敏内圈。道路拥堵一般有两个方面的原因，一是该路段附近存在知名景点或工业园区，导致车流量居高不下，另一方面这些道路由于城市规划时的局限性，周边都是路幅较小的支流马路，无法及时分流，因此造成道路积压。如火车站、外滩、五角场周边都存在不同程度的设计不合理之处，从路网中心性值上便可见一斑。如果要改善这种拥堵状况，从城市设计者的角度，可以在这些道路附近新建分流马路，即提高道路的中心性；从居民的角度，可以推荐车主错峰出行，采用别的出行道路或出行方式通勤，即减少通过该道路的次数。开源节流并举，便可以有效地改善拥堵状况。

7 结论

本文利用基站信息和爬取的路网信息，对上海市路网的中心性测度和模拟道路流量进行了综合分析，同时通过基站分布与基站附近基础设施的 POI 分析尝试通过路网对城市功能分区、道路拥挤状况和新城市中心的位置进行分析和解释。研究发现，对用户来说，大部分手机用户的通勤距离在 10km 以内，而对于路网承载力而言，大部分的道路流量压力由少部分的道路所承载，总体规律遵循“二八定律”。在使用基站数据和路网中心性测度数据对城市分区进行预测时，我们发现基于基站数据的预测能准确定位到城市的道路枢纽，而基于路网中心性测度指标的预测能够定位到城市的商圈和 CBD 等发达地区。

在综合两者指标进行分析时，我们发现，两者指标均较高的路段主要为城市中已经建设完成的开发区或经济园区，测度指标较高，交通流量指标较低的路段为城市未来规划中计划发展的区域，测度指标较低，交通流量指标较高的路段容易出现拥堵的状况，道路设计不够合理。而两者均较低的路段为城市道路中的辅路或分岔路段，几乎不承担什么交通流量。通过这些指标，我们可以对城市的商业中心和未来的发展方向进行预测，进行合理的商业规划；另一方面，在城市路网规划时，也可以根据手机的基站数据对路网的交通流量进行模拟，避免因设计不合理进行的拥堵状况。

同时，如果要基于以上结论进行进一步研究，有以下方向可以继续推进：

（1）由于算力和时间的局限，路网中心性测度指标难以针对整个上海市进行研究，为减少计算的复杂度，本文仅对上海市各个辖区进行了分析，且没有对占地面积较广的浦东新区进行计算。因此，基于本文的结论，还可以对浦东新区的数据和上海全市的路网进行全局性的计算，从而更准确的判断上海各辖区路网交通状况的相互影响。

（2）在本文中，所使用的手机基站数据并不具有时间特征，本文通过提取每位用户被记录次数最高的五个基站，以连接最小生成树的方式进行流量模拟。因此，具有时间特性的基站数据可能能使预测更加准确。

（3）由于数据源的问题，本文只考虑了基于驾车出行方式的路网研究，而在日常生活中，公共交通方式逐渐成为上班族的首选出行方式。因此，如果将公共交通模式纳入考量范围，能够更准确的对交通流量进行模拟。

（4）虽然手机基站数据覆盖面广，但是定位精确性差，只能得到基站的位置数据，如果能考虑更多的交通大数据来源，如出租车位置信息，地铁人流量信息等等，能大大提高数据对实际交通流量模拟的准确性。

8 参考文献

[1]《中国无线电管理年度报告(2018 年)》，工业和信息化部无线电管理局(国家无线电办公室)

[2] Freeman L C. Centrality in social networks conceptual clarification[J]. Social Network, 1979, 1(3):215-239.

[3] DIAKOULAKI，MAVROTAS G，PAPAYANNAKIS L. Determining objective weights in multiple criteria problems：The critic method[J]．Computers 8L Operations Research，1995，22(7)：763—770．

[4] Haklay M, Weber P. OpenStreetMap: User-Generated Street Maps[J]. IEEE Pervasive Computing, 2008, 7(4): p.12-18.

[5]廖薇薇，何家律，李秋萍,等. 城市道路结构等级与基于手机定位数据的交通流量匹配度分析[J]. 地理与地理信息科学, 2018, 034(002):58-65,前插 2.