

基于移动网络流量日志的城市空间行为分析

强思维¹⁾, 陈夏明¹⁾, 姜开达²⁾, 金耀辉^{1,2)}

¹⁾区域光纤通信网与新型光通信系统国家重点实验室, 上海交通大学, 上海, 200240

²⁾网络信息中心, 上海交通大学, 上海, 200240

摘 要 城市的空间组织和居民行为研究是城市地理学研究的重点, 随着信息技术的快速发展, 居民的时空行为对城市空间的组织和结构的影响呈现出日益增加的趋势, 因此, 对**城市空间以及居民时空行为**的深入分析具有很高的研究价值。本文通过采集杭州市区移动 3G 网络流量日志, 首先采用空间点模式的分析方法研究了城市居民的聚集模式, 居民移动的距离、方向等方面的特征, 之后采用网格的方法对城市空间进行分块, 并以区块为主体, 研究了热点区块出现的时空点, 区块人流的更迭速率, 工作日人流的潮汐效应, 提出了区块差异指数的概念, 并利用其对区块进行聚类, 分析了区块间的相关性和区块间距离之间的关系。由于本文研究的研究数据来源于移动 3G 网络流量日志, 因此具有覆盖面广、数据量大等特点, 非常适合从大时空尺度层面研究居民和城市空间行为。

关键词 时空分析, 空间点模式, 移动性分析, 更迭速率, 潮汐效应, 差异指数

Urban Spatio-temporal Behavior Analysis Based on Mobile Network Traffic Logs

Qiang Siwei¹⁾, Chen Xiaming¹⁾, Jiang Kaida²⁾, Jin Yaohui^{1,2)}

¹⁾State Key Laboratory of Advanced Optical Communication Systems and Networks, Shanghai Jiao Tong University, Shanghai, 200240

²⁾Network & Information Center, Shanghai Jiao Tong University, Shanghai, 200240

Abstract City organization and residents behavior is one of the key research in urban geography. With the rapid development of information technology, the impact of research on residents spatial and temporal behavior on urban spatial organization and structure shows a growing trend, therefore in-depth analysis of the spatio-temporal behavior of city space and urban residents have high research value. After acquisition of Hangzhou mobile 3G network traffic logs, using spatial point pattern analysis method, we study the gathering patterns of urban residents, and features including moving distance and direction. Using grid approach, we divided the urban space into blocks, and focus on the emergence of hotspot point, the change rate of human flow, tidal effects on weekdays, and present the concept of blocks difference index, and use it to cluster blocks and analyze the relationship between the similarity of blocks the their distance. Since our research data comes from mobile 3G network traffic logs, it has a wide coverage and large volume, which is ideal for search on residents and city behavior from large spatio-temporal scales.

Key words spatio-temporal analysis, spatial point pattern, mobility analysis, change rate, tidal effects, differences index

城市和居民的时空行为研究一直以来都是城市地理学研究的重点,主要包括:城市交通研究^[1,2]、城市功能分区研究^[3,4]、城市等级体系研究^[5-7]等。随着信息技术的快速发展,居民的时空行为对城市空间组织和结构的影响呈现出日益增加的趋势。因此,对居民时空行为深入分析具有很高的研究价值。

时空数据的采集与处理通常是研究城市和居民时空行为的关键点和难点。传统的数据获取手段以人口普查和活动日志为最常用^[8]。然而,这类途径获取的数据通常具有一些弊端,例如现有的人口普查方法并不能获知居民在物理空间中移动的实时情况,活动日志数据的精确性也因调查者的回忆、填写态度等影响而有不同程度的失真。在信息时代,除了利用以上传统方法,城市和居民的行为数据还能通过智能手机、出租车、智能卡等方式获取,例如手机通话流量和位置数据、出租车位置信息、公交或地铁站点的刷卡数据^[8]。通过这一类途径获取的数据具有很好的实时性,并能够准确反应居民的真实行为和状态,其收集数据对应的用户群体,定位精度、时间,适合于研究的问题对比如下。

表1 不同数据源之间的比较

	手机流量	出租车 GPS	公交刷卡
用户群体	手机用户	出租车乘客	公交乘客
定位精度	数百米	数到数十米	公交站点
定位时间	上网时段	乘车时段	进出站时刻
研究问题	大尺度城市、居民时空行为	用户轨迹、城市街道	居民通勤、公交设施

本文通过采集杭州市移动 3G 网络流量日志,能够近乎实时的获取杭州城区约十分之一的居民的时空位置信息,因而能够较好的还原出城市空间和居民时空行为,然而也存在一些缺点,例如定位精度不如 GPS,同时,原始数据量较为庞大,单日的平均网络日志存储量约为 20G,因此给针对以周计或者更长时间跨度的挖掘带来一些挑战。针对海量数据,本文采用大数据的方法和工具,数据的存储采用了 Hadoop 分布式存储系统 HDFS,挖掘过程则运用了 Hadoop MapReduce 通用的并行计算框架 Spark^[9]等主流大数据处理工具。大数据工具的使用极大的提升了处理的速度和便捷性,并能做出近实时的分析。

本文主要的研究问题包括:杭州市区居民的时空行为的特点,包括了总人口的空间聚集性,居民的移动距离、方向;利用网格的方法将城市

空间进行分块,主要了研究包括热点区块出现的时空点,不同区块人流的更迭速率,工作日人流的潮汐效应,提出了区块差异指数的概念,并利用其对区块进行了聚类,并分析了区块间的相关性和距离之间的关系。

1. 相关工作

随着定位技术的不断发展与广泛应用,基于 GPS 等定位技术的居民行为时空数据已被广泛应用于居民行为的研究中,研究的主体是城市居民,研究方向主要围绕城市居民通勤的特征,例如:通勤模式、不同群体的通勤差异、职住关系与通勤行为、城市空间与通勤行为、过量通勤等^[10-14],研究方法主要运用描述性统计、回归分析等分析方法。文献[15]通过调查获取了典型街区实体空间信息、被调查者的社会属性、通勤行为空间和对通勤沿路相关实体要素感知信息等对居民的通勤行为进行分析与模拟。文献[16]以北京市天通苑与亦庄两个郊区巨型社区为案例,采用 GPS 定位数据和活动日志相结合的为期一周的居民时空行为数据,分别利用传统方法和通勤弹性视角研究了居民的通勤特征。文献[17]基于 2008 年北京市连续一周的公交 IC 卡刷卡数据,结合 05 年居民出行调查、地块级别的土地利用图,识别公交持卡人的居住地、就业地和通勤出行,分析了 3 大典型居住区和 6 个典型办公区的通勤出行特征,研究了通勤出行的主导方向以及 08 和 05 年通勤时间和距离的相似性。前人的研究多就居民为研究对象展开研究,对于城市不同区块所承载的居民时空行为的异同等的研究较为有限,同时,受限于数据来源,研究对象的数量相对较少,实时性相对较弱。

2. 数据来源和基本研究方法

本文使用的数据来源于杭州市区移动 3G 网络流量日志,包含 2012 年 8 月 16-28 日大约两周的数据,数据包含移动用户识别码(IMSI)、用户上网时连接的基站的位置区编码(LAC)和小区标识(CI)、HTTP 请求的时间戳等。本文主要通过 IMSI 字段区分单个用户,并将 LAC 和 CI 转换成基站的经纬度坐标,结合 HTTP 请求的时间戳标定用户出现的时空点。

本文主要研究杭州市区，地理位置的经纬度范围在 120.02° – 120.48° 、 30.15° – 30.42° 的区域，包括了上城区、下城区、西湖区、江干区、拱墅区、滨江区、萧山区等主要区块。据统计该经纬度范围内的基站总共有约 2×10^4 个，单天网络日志条数大约为 6×10^7 条，原始数据量约为 20G，用户数约为 3×10^5 人，约占核心城区总人口的 $1/10$ 。



图1 基站密度统计

图1是研究区域范围内的基站分布的密度统计图。可见市区中心的基站分布的密度较大而周边较小，但考虑到3G基站的覆盖范围最大可达数千米，因此，研究范围内的全部基站基本能够完全覆盖研究范围。

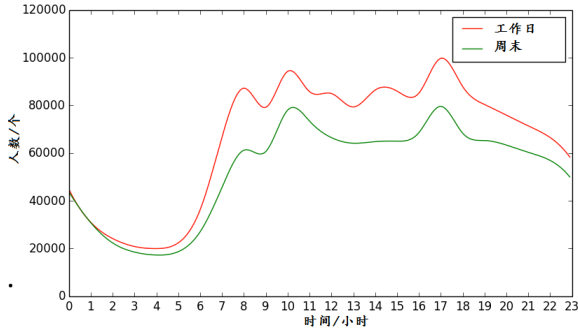


图2 每小时连接基站的总用户数统计

图2是按一小时为时间窗口统计的工作日和周末的平均总接入用户数。可见，白天接入的用户较多，夜间较少，且工作日大于周末，白天每小时的接入用户数大约为 7×10^4 。

基于本文数据源，挖掘的难点主要体现在，其一，定位误差相对GPS定位误差较大，通常为数百米，无法将用户位置准确还原到城市的道路或建筑物中，缺少与之相关的语义信息，其二，由于只有用户上网的时间段才能获取用户的空间位置信息，因而，用户在时间轴上具有很大的跳跃性，难以从中推测出用户的轨迹信息。数据的来源限制了无法针对单个用户的挖掘，同时常用的轨迹挖掘方法也难以奏效。

由于难以准确获取个体用户连续时空位置，轨迹挖掘的方法失效，本文并不针对个体用户的行为进行挖掘，而是针对用户群体或利用网格对城市空间进行分块（将地理空间切分成均等大小的方块，经实验，粒度选为200~500m较为适宜，一方面与基站定位精度的尺度相当，同时保证了具有一定的用户量以满足统计特性），研究大尺度范围内的城市空间和居民的时空行为。由于用户上网行为符合统计学规律，因此针对群体行为的统计能够符合真实情况。

文本的基本研究方法包括数据的存储，数据的处理和计算。由于数据量庞大，利用传统的关系型数据库进行管理和挖掘几乎不可能，因此，本文采用了Hadoop分布式文件系统HDFS对数据进行存储；采用UC Berkeley AMP lab所开源的Hadoop MapReduce通用的并行计算框架Spark，进行计算和分析，由于中间输出结果可以直接保存在内存中不再需要读写HDFS，因此Spark能够更好地适用于数据挖掘与机器学习等需要迭代的MapReduce算法。

3. 空间点模式和移动性分析

3.1. 空间点模式分析

平面空间点模式分析方法通常可被分为两类：检测空间过程的一阶影响，主要研究点事件的均值随着空间变化而变化的过程，常用方法是核密度估计法；检测空间过程的二阶影响，主要考察点过程之间的空间依赖性，其中K函数使用了研究范围内的全部点之间距离来对集聚模式进行度量，因而更加稳定可靠^[18]。采用空间点模式的分析方法进行分析能够揭示出用户空间分布的基本规律。

核密度估计根据观测值获得概率密度的平滑估计值。对区域内任意点上的事件密度进行估计，最常用的方法就是以该点为圆心，统计以该点为圆心，以 r 为半径范围之内事件发生的数量，然后除以圆的面积，因此 s 处的密度可以表示如下。

$$\lambda(s) = \sum_{i=1}^n \frac{1}{\pi r^2} k\left(\frac{d_{is}}{r}\right) \quad (1)$$

其中半径 r 为带宽， $k(\cdot)$ 为核函数，常为对称密度函数，可理解为点 i 处的权重，常用核函数有均匀核、二次核、三角核、高斯核、余弦核等； d_{is} 为点 i 到点 s 的距离。带宽 r 决定了密度曲面的平滑程度。

文本首先利用网格的方法对空间进行划分，统计出单日连接过区块内基站的用户数，再进行核密度估计。通常选择何种核函数不是密度估计中最关键的因素，因为任何核函数都可以保证对密度估计具有稳定性，本文选取采用了高斯核，而带宽选择对密度曲线边界的影响很大，带宽较小时边界光滑性较差，反之，当带较大时边界光滑性较好，但曲线拟合度较差，经实验比较区块和带宽的粒度选为500m较为适宜。图3是平均单日用户数统计的核密度估计图，显示了用户分布情况，可见，西湖周边区域人数较为密集，并向四周扩散呈现递减的趋势。



图3 单日用户数核密度估计

K 函数的基本思想是：假定依次在各个事件中心设置半径为 d 的圆，统计出落入圆内其他事件的数量，计算出其期望并除以总事件密度，并对于不同距离值不断重复这个过程。 K 函数定义如下。

$$K(d) = A \sum_{i=1}^n \sum_{j=1}^n \frac{\delta_{ij}(d)}{n^2} \quad (2)$$

其中， n 为事件的个数， $i \neq j$ ， d 为空间距离， d_{ij} 为事件 i 、 j 之间的距离， $\delta_{ij}(d) = \begin{cases} 1 & (d_{ij} \leq d) \\ 0 & (d_{ij} > d) \end{cases}$ ， A 为研究区域的面积。

L 函数是 K 函数的一个变形，用来衡量点分布模式随尺度的变化规律，保持了方差的稳定。

$$L(d) = \sqrt{\frac{K(d)}{\pi}} - d \quad (3)$$

在随机分布的假设下， $L(d)$ 的期望值等于0， $L(d)$ 与 d 的关系图可用于检验依赖于尺度 d 的事件的空间分布模式。 $L(d)$ 第一个峰值对应的 d 值表示了事件空间集聚的特征空间尺度。

本文借助 L 函数，并重新定义 $\delta_{ij}(d) = \begin{cases} n_i \times n_j & (d_{ij} \leq d) \\ 0 & (d_{ij} > d) \end{cases}$ ， n_i 是区块 i 内用户数， d_{ij} 是区块 i 、 j 的中心点之间的距离。

图4是工作日和周末的 L 函数分析图，横坐标为研究尺度 d ，纵坐标为 $L(d)$ ，由函数的曲线可见用户在空间分布上呈现出明显的聚集性，工作日更为明显，由函数的峰值可见用户分布的特征空间尺度约为6200m。

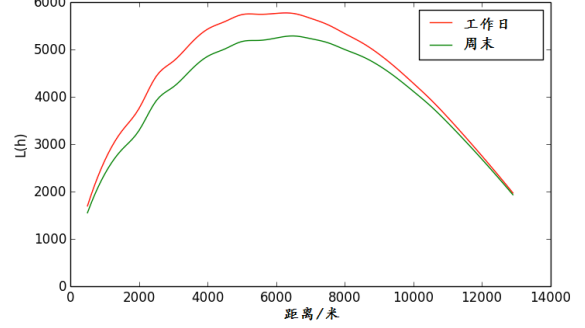


图4 L函数分析图

3.2. 移动性分析

用户移动性分析主要分析运动用户移动的距离和方向，包括用户平均移动距离和时间之间的关系，以及位于不同区域的用户到达下一地点的方向和距离的倾向性，来揭示出用户移动的基本时空规律。

用户的移动距离近似为某时间窗口内用户所有连接基站之间的最大距离。计算公式如下。

$$u_{t_i}.dist = \max(\sum_{i=1}^n \sum_{j=1}^n distance(p_i, p_j)) \quad (4)$$

其中， $u_{t_i}.dist$ 是用户 u 在时间窗口 t_i 内移动的最大距离，在时间窗口 t_i 连接的基站数为 n ， $distance(p_i, p_j)$ 是基站 p_i 、 p_j 之间的距离。

图5是用户移动距离的统计和时间的关系，时间窗口为一小时，箱线图包含了数据中的五个主要统计量：最小值、第一四分位数、中位数、第三四分位数以及最大值，能够显示数据的基本分布情况。由图5可见，工作日相对于周末的平均移动距离较大，并且在早、晚高峰显示出较为明显的通勤特征。

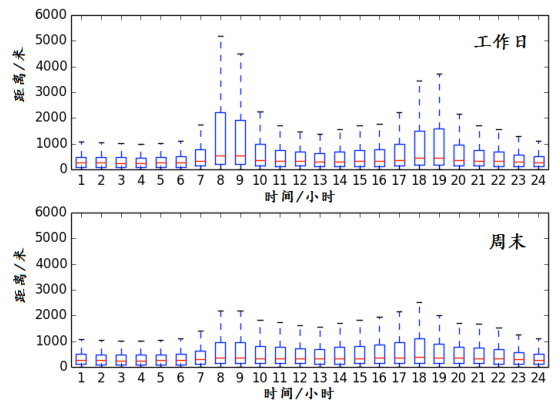


图5 移动距离和时间的关系

用户移动距离和方向的倾向性关于空间位置的关系，通过计算相邻两个时间窗口用户位置的变化来显示。其中，用户在时间窗口 t_i 的位置通过用户在时间窗口 t_i 连接的所有基站的位置的期望来近似，公式如下。

$$\begin{cases} u_{t_i}.lng = \sum_{j=1}^n p_j.lng & (5) \\ u_{t_i}.lat = \sum_{j=1}^n p_j.lat & (6) \end{cases}$$

其中， $u_{t_i}.lng$ 、 $u_{t_i}.lat$ 分别对应了用户 u 在时间窗口 t_i 的经度和纬度，在时间窗口 t_i 连接的基站数为 n ， $p_j.lng$ 、 $p_j.lat$ 分别是第 j 个基站的经度和纬度。

图 6 所示的用户单日移动方向和距离图通过将用户在两个相邻的时间窗口的位置进行连线，按照东、南、西、北、东北、东南、西北、西南八个方向进行归类，计算出每个方位移动距离的期望。在图 5 中分别选取了位于市中心东（C）、南（D）、西（A）、北（B）、中（E）五个地点进行观测，每个方向的半径分别表示了单日内用户向各个方向移动的期望距离。由图 6 可见，从区位关系上看，从各个地点到下一地点的移动方向大多指向中心城区，表明中心城区也是城市职能的强中心。

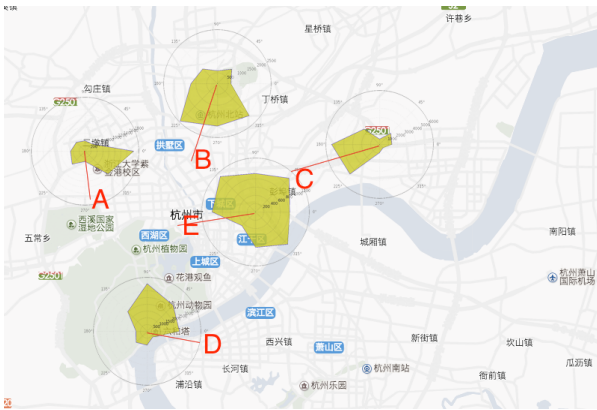


图 6 不同区域用户单日移动方向和距离

4. 网格分块的区块行为分析

4.1. 热点区块

定义 1. 热点区块. 热点区块是某时间窗口内总人数统计中相对最多的 $x\%$ 区块。

研究热点区块（即人数最为稠密的区域）的时空行为对城市的规划，包括城市空间的组织和结构的调整，以及相关公共资源，例如城市交通设施的建设，网络基站的部署，具有重要的指导意义。

文本首先利用网格的方法对空间进行划分，粒度选为 $500m$ ，总共有 33750 个区块，之后统计出各时间窗口内各区块所包含的总用户数，并对其排序，取前 0.1% （前 34）个区块标记为该时间窗口内的热点区块。

图 7 是分别对工作日、周末的白天、晚上，各个时段统计出的，单天内热点区块出现的空间位置，以及出现的小时数，时间窗口为 1 小时。柱状图的高度正比于热点出现的总时长。由图 7 可见，与西湖相邻的东北区域是白天热点的高发区域，其南北相邻的两块区域为晚上热点的高发区域，并且周末的热点与工作日相比较为分散，由此可以推测出，西湖相邻的城区中心区域可能为商业区，而两侧可能多为住宅区，工作日人口活动区域相对比较集中，而周末则比较发散。

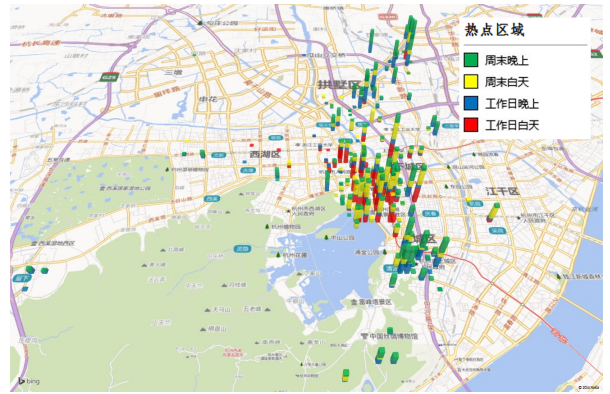


图 7 单日内热点区块出现的位置和总时长

图 8 显示了一周内出现过热点的区块中热点出现的时间，有色部分表示该区域是该时间段的热点区域，并以区块每小时热点是否出现作为特征进行 kmeans ($k=3$) 聚类得到的结果。由图 8 可见，热点出现可以分为 3 种模式，即白天出现（推测为商业区），晚上出现（推测为住宅区），突发出现（可能与突发事件相关）。

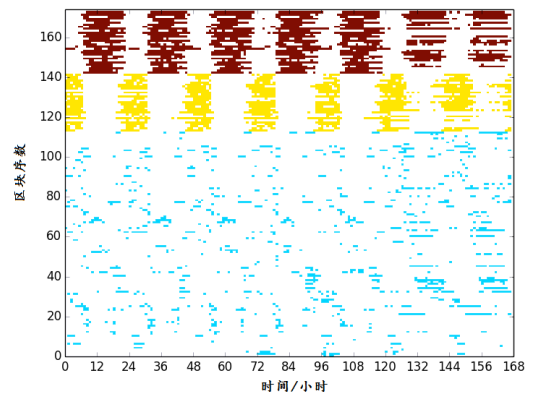


图 8 一周内热点区块出现的时间

4.2. 区块人流更迭速率

定义 2. 更迭速率. 区块的更迭速率是区块在相邻时间窗口内所包含的相异用户的数量和相同用户的数量的比值, 计算公式如下。

$$r = \frac{\#[s_{t_{i+1}} - s_{t_i}] + \#[s_{t_i} - s_{t_{i+1}}]}{\#[s_{t_i} \cap s_{t_{i+1}}]} \quad (7)$$

其中, r 是更迭速率, s_{t_i} 和 $s_{t_{i+1}}$ 分别是 t_i 和 t_{i+1} 时刻的区块用户的集合, $\#[.]$ 表示集合大小。更迭速率的大小可用来说明区块内用户群体的稳定性或流动性。 r 越小表明用户集合越趋于稳定, 反之 r 越大表明用户集合变化越快, 流动性越强。

图 9 是工作日上午 8 时区块人流的更迭速率正规化后的结果, 即某区块 i 的更迭速率为:

$$\hat{r}_i = \frac{r_i - \min_{1 \leq j \leq n} \{r_j\}}{\max_{1 \leq j \leq n} \{r_j\} - \min_{1 \leq j \leq n} \{r_j\}} \quad (8)$$

其中, n 为所有区块的数量。

由图 9 可见, 工作日早高峰时段, 区块更迭速率与区块的功能有着密切联系, 更迭速率最大的一部分区块恰好位于主要道路。

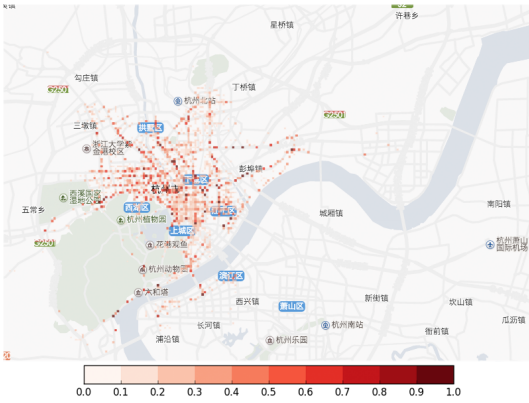


图 9 工作日上午 8 时区块人流更迭速率

图 10 显示了各个区块的平均更迭速率和时间的关系。可见白天人口的流动性高于夜间, 早晚高峰最强, 除早晚高峰, 周末人口的流动性强于工作日, 早晚高峰时段工作日人口流动性最强。

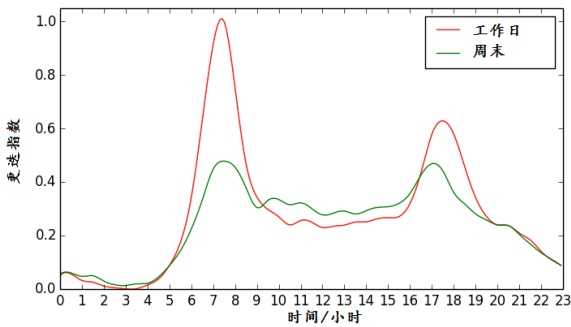


图 10 区块平均更迭速率和时间的关系

4.3. 差异指数和区块相似性

定义 3. 差异指数. 区块的差异指数是区块人数在一天内的分布和全部区域人数在一天内的分布的差异, 计算公式如下。

$$\begin{cases} d_{t_i}^{(k)} = c_{t_i}^{(k)} - c_{t_i} & (9) \\ c_{t_i}^{(k)} = \frac{n_{t_i}^{(k)}}{\sum_{i=1}^T n_{t_i}^{(k)}} & (10) \end{cases}$$

其中, 一天被分成 T 个等长的时间窗口, 区块 k 在时间窗口 t_i 内的人数为 $n_{t_i}^{(k)}$, 占一天总人数的比例为 $c_{t_i}^{(k)}$, c_{t_i} 则是针对全部区域计算的结果, $d_{t_i}^{(k)}$ 是差异指数, 表示了区块内人口数量的变化趋势和总人口数量的变化趋势的差异。 $d_{t_i}^{(k)} = 0$ 表明区块 k 的人口在时间窗口 t_i 内和整体总人口数量相一致, $d_{t_i}^{(k)} > 0$ 表明区块 k 的人口数量相对于自身在一天中的分布较为密集, $d_{t_i}^{(k)} < 0$ 则表明较为稀疏。

图 11 是工作日不同时刻差异指数经过正则化后在空间中的分布, 红色和蓝色分别表示人口相对的密集和稀疏区域。可见, 早、晚人口在杭州市周边区域较为密集, 而白天则在中心区域较为密集, 显示出工作日人流的潮汐效应, 即早高峰人流流向市中心, 晚高峰则相反。

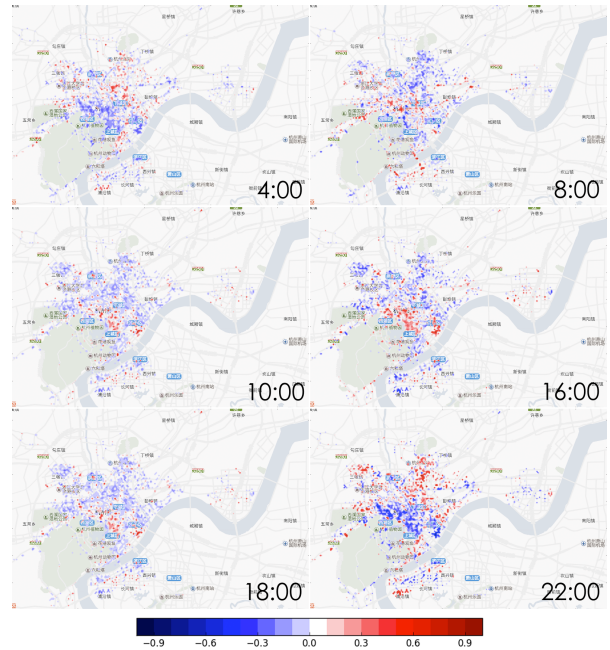


图 11 工作日人流的潮汐效应

图 12 是选取到的 3 个特殊地点计算出的差异指数随时间的变化趋势，图中 A、B、C 分别对应了艮秋立交桥，黄龙时代广场，三塘北村东区。艮秋立交桥位于杭州市交通主干道的十字路口，黄龙时代广场位于商业区，三塘北村东区则位于住宅区，分别代表了城市的三种主要功能区域。由差异指数可见以艮秋立交桥为代表的交通路段在早晚高峰呈现出明显的密集性，而以黄龙时代广场为代表的商业区和以三塘北村东区为代表的住宅区则分别在白天和晚上呈现出密集性。

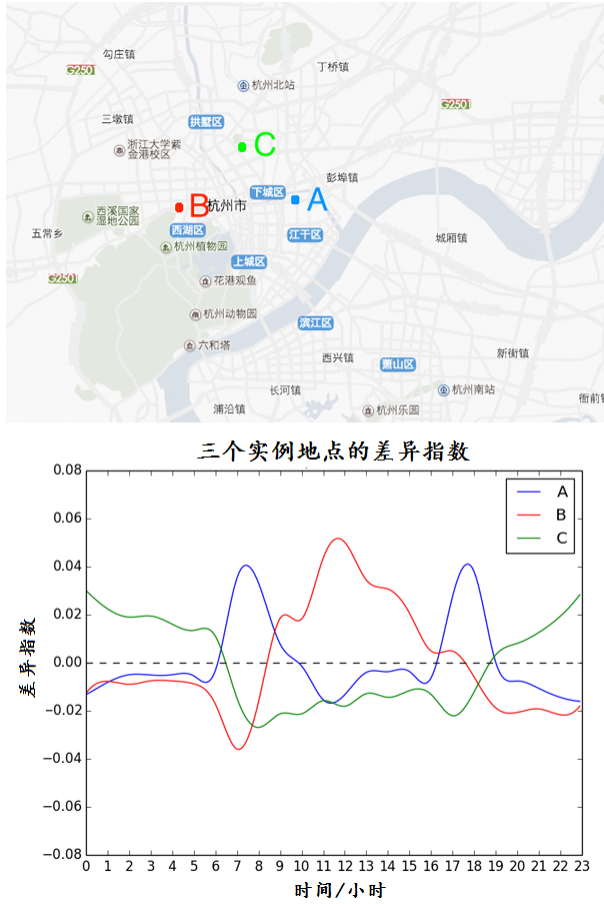


图 12 三个实例地点的差异指数随时间的变化

差异指数的分析有助于分析城市不同区域的功能特征，并能够利用其对城市的不同区块进行聚类，获知不同功能区块在空间中的分布情况。图 13 是提取各个功能区块的差异指数的时间序列作为特征，对其进行 kmeans (k=3) 聚类，得到的结果，上图展示了 3 种不同类别在空间中的分布情况，而下图展示了其聚类中心的差异指数随时间的变化，可以推测出蓝色、红色、绿色区域分别对应了交通路段、商业区、住宅区，三种典型的的城市功能区域。

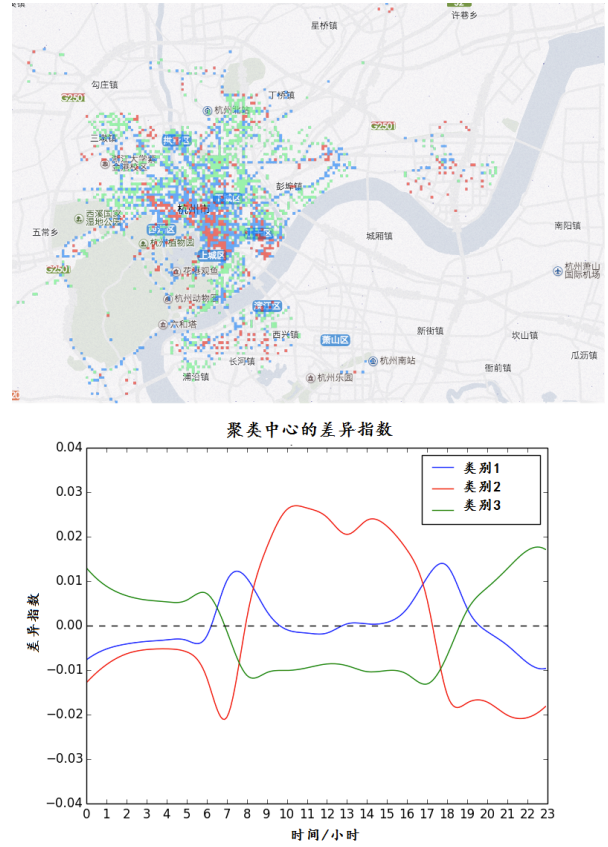


图 13 聚类的空间分布及聚类中心差异指数随时间的变化

利用差异指数还能计算区块之间的相关性，其计算公式如下。

$$c(r) = \frac{1}{\sigma^2} \frac{\sum_{i,j,t} (d_t^{(i)} - \bar{d}_t)(d_t^{(j)} - \bar{d}_t) \delta(r_{ij} - r)}{\sum_{i,j} \delta(r_{ij} - r)} \quad (11)$$

其中 $d_t^{(i)}$ 是区块 i 在时间窗口 t 内的差异指数，

$\bar{d}_t$ 是 $d_t^{(i)}$ 的期望， r_{ij} 是区块 i 、 j 中心之间的距离，

$\delta(\cdot)$ 是脉冲函数，即 $\delta(r_{ij} - r) = \begin{cases} 1 & (r_{ij} \approx r) \\ 0 & \end{cases}$ ，

$\sigma^2 = \sum_{i,j,t} (d_t^{(i)} - \bar{d}_t)^2 / N$ 为方差， N 为所有区块的数量， $c(r)$ 即为关联性。易推出 $r = 0$ 时 $c(r) = 1$ 。若 $c(r) > 0$ 则说明区块在改距离尺度上呈现正相关关系。

图 14 是区域间的相关性和距离之间的关系，横坐标为对数坐标，可见，区域间相关性与距离近似符如下关系。

$$c(r) \sim -\gamma \lg(r) \quad (12)$$

γ 表征了区域间相关性关于距离衰减的程度。同时可见在 4000m 空间尺度范围内区域间呈现正相关关系，大于 4000m 则呈现出弱负相关性。

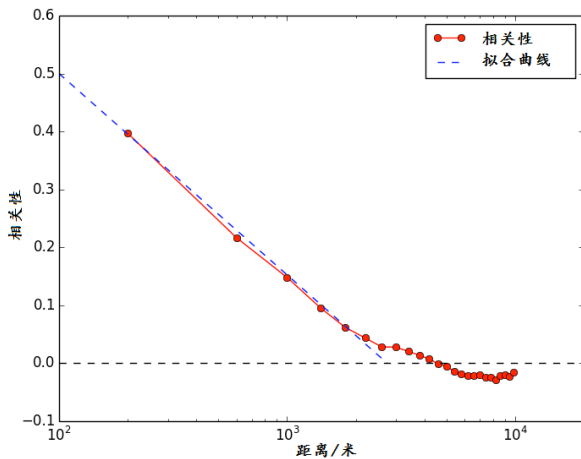


图 14 区间间相关性和距离的关系

5. 结束语

深入分析城市和居民时空行为对城市规划，包括城市空间的组织和结构的调整，具有很高的研究价值。本文通过采集杭州市区的移动 3G 网络流量日志，采用统计学的方法，分析了杭州市区居民时空行为的特点。由于数据具有覆盖面广、数据量大等特点，处理过程中运用了 Spark 等大数据的处理工具和方法。研究内容主要包括了两方面：基于用户群体的空间点模式和移动性分析和基于网格分块的区块行为分析。前者采用空间点模式的分析方法，研究了城市居民的聚集模式以及居民移动距离、方向等方面的特征。后者则基于网格分块的区块行为分析，采用网格的方法对城市空间进行分块，以区块为研究对象，重点研究了热点区块出现的时空点，区块人流的更迭速率，工作日人流的潮汐效应，提出了区块差异指数的概念，并利用其对区块进行聚类，其中的每一个类别都对应了一种具有典型功能的区域，最后分析了区块间的相关性和距离之间的关系。

致谢 感谢 OMNILab 的所有老师和同学的讨论和帮助。

参考文献

[1] Becker R A, Caceres R, Hanson K, et al. 2011. A tale of one city: Using cellular network data for urban planning. *Pervasive Computing, IEEE*, 10(4): 18-26.

[2] Liu Y, Kang C, Gao S, et al. 2012a. Understanding intra-urban trip

patterns from taxi trajectory data. *Journal of Geographical Systems*, 14(4): 463-483.

[3] Hollenstein L, Purves R. 2013. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, (1): 21-48.

[4] Cranshaw J, Schwartz R, Hong J, et al. 2012. The livelihoods project: Utilizing social media to understand the dynamics of a city. 6th International AAAI Conference on Weblogs and Social Media (ICWSM-12), Dublin, Ireland: June 4-7.

[5] Naaman M, Zhang A X, Brody S, et al. 2012. On the study of diurnal urban routines on Twitter. 6th International AAAI Conference on Weblogs and Social Media (ICWSM-12), Dublin, Ireland: June 4-7.

[6] Krings G, Calabrese F, Ratti C, et al. 2009. Urban gravity: A model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, L07003: 1-8.

[7] Kang C, Zhang Y, Ma X, et al. 2013. Inferring properties and revealing geographical impacts of intercity mobile communication network of China using a subnet data set. *International Journal of Geographical Information Science*, 27(3): 431-448.

[8] 秦萧, 甄峰, 熊丽芳, 等. 2013. 大数据时代城市时空行为研究方法[J]. *地理科学进展*. 32 (9): 1352-1361

[9] <https://spark.apache.org/>

[10] Sohn J. 2005. Are commuting patterns a good indicator of urban spatial structure. *Journal of Transport Geography*, 13(4): 306-317.

[11] Lee B S, McDonald J F. 2003. Determinants of commuting time and distance for Seoul residents: The impact of family status on the commuting of women. *Urban Studies*, 40(7): 1283-1302.

[12] Schwanen T, Mokhtarian P L. 2005. What affects commute mode choice: Neighborhood physical structure or preferences toward neighborhoods. *Journal of Transport Geography*, 13: 83-99.

[13] Cao X, Mokhtarian P L. 2005. How do individuals adapt their personal travel? A concept exploration of the consideration of travel-related strategies. *Transport Policy*, (12): 199-206.

[14] Horner M. 2002. Extensions to the concept of excess commuting. *Environment and Planning A*, 34(3): 543-566.

[15] 周素红, 闫小培. 2006. 基于居民通勤行为分析的城市空间解读——以广州市典型街区为案例. *地理学报*, 61(2): 179-189.

[16] 申悦, 柴彦威. 2012. 基于 GPS 数据的城市居民通勤弹性研究: 以北京市郊区巨型社区为例. *地理学报*, 67(6): 733-744.

[17] 龙瀛, 张宇, 崔承印. 2012. 利用公交刷卡数据分析北京职住关系和通勤出行. *地理学报*, 67(10): 1-14.

[18] Gatrell A, Bailey T, Diggle P, et al. 1996. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21(1): 256-274.