DOI:10.13203/j.whugis20150646　　文章编号:1671-8860(2017)01-0049-07

顾及手机基站分布的核密度估计城市人群时空停留分布

杨喜平[1]　方志祥[1]　赵志远[1]　萧世伦[1]　尹　凌[2]

1　武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079
2　中国科学院深圳先进技术研究院,广东 深圳,518055

摘　要:为了减小人群在连续空间上停留分布的估计误差,结合手机基站的空间的分布特点,根据基站间的邻近性来计算带宽控制参数,使搜索带宽随着基站的分布而变化;利用最小二乘交叉验证和对数概率两种方法来评价其估计效果,结果表明变化带宽比固定带宽的核密度估计效果更优。以深圳市手机位置数据为例,利用改进方法估计了几个典型时段城市人群停留的时空分布差异,反映了城市人群对城市不同区域的使用情况及其随时间变化情况。

关键词:手机数据;核密度估计;人群停留;时空分析

中图法分类号:P208　　**文献标志码**:A

城市人群的时空停留分布可以反映人群对城市不同空间的使用规律,因此,详细掌握城市人群的时空停留分布可以帮助指导城市总体规划、基础设施建设、优化资源配置、城市应急管理(如自然灾害估计受灾人口)、商业选址优化以及交通流预测等[1]。最近,手机位置数据为详细研究城市人群时空分布提供了新的机遇和挑战[2-3]。如利用通讯 Erlang 值对城市人群进行实时监测[3],动态估计人群分布[4],分析动态人群分布[5]与 Erlang 值、通话个数、用户数的关系等[6]。

手机位置数据是通过基站进行定位的,只能从数据中提取出基于基站的人群停留分布,并不能得到整个城市连续空间上的停留分布。目前,大多数研究采用等值面法来表示人群在连续空间上的分布,假设基站的覆盖范围为其对应的泰森多边形,人群在多边形内是均匀分布的。这种表示方法存在一些缺点:(1)泰森多边形内的土地利用并不是均质的,如存在水系、山地等,将人群平均分配到这些区域并不合理。(2)采用规则的边对基站信号进行切割不符合现实情况,这会导致在多边形内人口密度是一致的,而在相邻多边形间出现阶梯状不连续的变化,忽略了空间现象发生的连续性[7-8]。

针对等值面法的缺陷,一些学者提出采用核密度法作为估计人口分布的空间连续模型,从而得到连续空间上的人群分布。核密度方法可以将样本点数据转化成一个平滑的表面,已被广泛用来估计人群的连续空间分布[7-10],因此利用该模型可以从基于基站的人群停留分布生成连续空间上的人群停留分布。但要注意的是,得到的密度值只能相对地代表人群的多少而不是真实的停留人口密度。目前采用核密度估计人群分布的研究采用统一固定的带宽进行估计,而在现实中遇到的空间数据多数是异质的,分布不均匀,选择固定带宽会给估计带来误差,尤其是在人群密度很高的城市,细小的变化会带来很大的估计误差。因此,采用核密度估计人群在连续空间上的停留分布时,带宽的选择至关重要,并且要根据基站空间分布和特征属性来决定带宽[11]。

深圳市是全国人口密度最高的城市,平均人口密度为 5 545 人/km^2,在市中心商业区人口密度达到 10 万以上[1],采用核密度估计人群分布时更加要注重带宽的选择。本文以深圳市手机位置数据为例,结合手机基站的空间分布,根据基站间

收稿日期:2015-10-30

项目资助:国家自然科学基金(41231171,41371420);武汉大学自主科研项目拔尖创新人才类项目(2042015KF0167);中国科学院资源与环境信息系统国家重点实验室 2013 年开放基金。

第一作者:杨喜平,博士生,主要从事时空数据分析与挖掘研究。0yangxiping0@163.com

通讯作者:方志祥,博士,教授。zxfang@whu.edu.cn

的邻近性来计算带宽尺度参数，以控制不同基站的搜索带宽进行核密度估计，从而提高核密度估计人群分布的精度。最后从手机数据中提取出几个典型时段的基站停留人数，采用改进的核密度法来估计人群分布，通过时段间作差来分析人群在这几个时段间停留的空间分布差异。

1 数据描述

本文采用的数据是深圳市某工作日的手机位置数据，约1 600万用户，该数据采样间隔为0.5 h或1 h，通讯公司为了检测故障或其他目的会在一定的时间间隔内主动记录一次用户所在服务基站的位置，即无论用户是否进行通讯活动(通话、发短信或上网)都会记录。如表1所示，每条记录包括用户的身份标识((identification，ID)、基站的经纬度以及记录时间。其中，为了保护用户的隐私，运营商已经对用户的ID进行了加密处理，从数据中提取出5 940个基站并对每个基站进行唯一编号。基站分布如图1所示，手机基站的分布不均匀，在市中心基站的服务范围较小，而在郊区服务范围较大。在所有基站对中，最小的基站对间距离为1.03 m，最大值为87 499.87 m，平均距离为22 281.36 m，其中，距离小于100 m的基站对有1 367对。

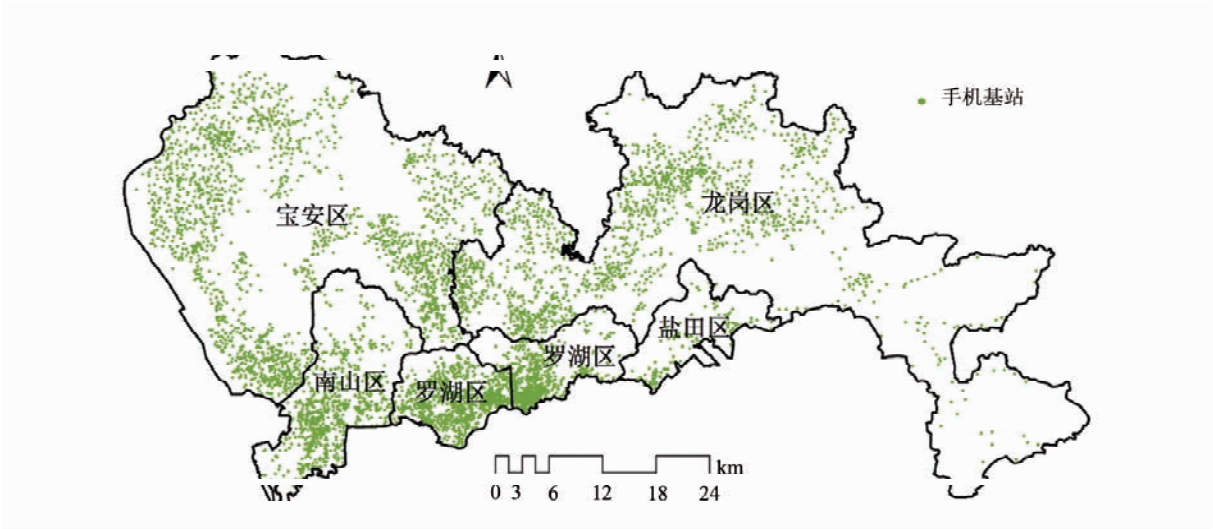

图1 手机基站空间分布

Fig. 1 Spatial Distribution of Mobile Phone Towers

表1 手机位置数据记录实例

Tab. 1 Examples of Mobile Phone Location Data

用户ID	经度	纬度	记录时间
* * * ffc5d851d * * *	113. xxx	22. xxx	00:20:15
* * * 8a5eaa5eb * * *	113. xxx	22. xxx	09:36:40
⋮			⋮
* * * 4b770d2bb * * *	113. xxx	22. xxx	22:50:09

2 顾及基站分布的核密度估计法

在空间分析中，核密度方法可以将样本点数据转换成连续平滑的面，已经被广泛地应用在交通、犯罪和流行病领域[8]。核密度$\hat{f}(x)$的计算公式为：

$$\hat{f}(x)=\sum_{i=1}^{n}\frac{1}{h^2}\cdot k\left(\frac{x-x_i}{h}\right) \tag{1}$$

式中，h为搜索带宽；n为与待估点x的距离小于或等于h的样本点数；$k(\cdot)$为核函数；x_i为第i个已知点。

式(1)中带宽h对所有样本点是固定不变的，而手机基站在城市中的分布是不均匀的，基站的覆盖范围随着人群的分布而变化，采用固定的带宽会给估计带来误差。

本文提出采用变化的带宽代替固定带宽，根据基站与其相邻基站间的距离计算距离搜索带宽尺度参数来控制不同基站的带宽，变化带宽的核密度估计公式为：

$$\hat{f}(x)=\sum_{i=1}^{n}\frac{1}{h_i^2}\cdot w_i\cdot k\left(\frac{x-x_i}{h_i}\right)$$
$$h_i=h_0\lambda_i \tag{2}$$

式中，h_0为初始的距离带宽；w_i为基站i的权重，本文中权重为基站的停留人数；本文采用高斯核函数$k(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$；$n$为所有基站中与$x$距离小于其各自带宽的基站数；$\lambda_i$为基站$i$的带宽尺度参数，其计算步骤为：

1) 首先利用基站点生成 Voronoi 多边形,Voronoi 多边形可以帮助识别基站 i 的相邻基站[12];

2) 计算基站 i 与其相邻基站的平均距离 $d_i = \sum_{j=1}^{m} d_{ij}/m$,$d_{ij}$ 为基站 i 与基站 j 的欧氏距离,m 为与基站 i 相邻的基站个数;

3) 城市中所有基站的 d_i 的均值 $\overline{d} = \sum_{i=1}^{N} d_i/N$,$N$ 为所有基站个数;

4) 计算每个基站的带宽尺度参数 $\lambda_i = d_i/\overline{d}$。

由于该尺度参数考虑了基站在城市的分布情况,每个基站根据其周围相邻基站间的距离来控制其距离带宽,使得带宽在基站高密度区域变小,低密度区域增大,并且随着基站与其邻近基站的分布变化。

3　实验与分析

3.1　初始带宽选择和效果比较

这部分选取两种方法来比较在权重 $w_i=1$ 的情况下,即只考虑基站点空间分布时,固定带宽和加入尺度参数的核密度方法的估计效果。在后面的分析中,下标 fix 代表固定带宽,var 表示变化带宽。首先采用最小二乘交叉验证法来选择初始带宽 h_0,计算基站的均方误差 $\mathrm{MISE}(h) = \int (\hat{f}(x_i) - f(x_i))^2 \mathrm{d}x$,其中 $\hat{f}(x_i)$ 为基站 i 密度估计值,$f(x_i)$ 为真值,而密度真值并不知道,在该方法中是用 $\hat{f}_{-i}(x_i)$ 代替,$\hat{f}_{-i}(x_i)$ 表示利用剩余的样本点来估计 x_i 的密度值。当均方误差越小时,估计值和真值越接近,表示效果越好,以此来寻找最优带宽[13]。根据文献[13]的公式推导,MISE 的最后计算公式为:

$$\mathrm{MISE}(h) = \int (\hat{f}(x_i) - f(x_i))^2 \mathrm{d}x = \sum_i \sum_j w_i w_j \frac{1}{\sqrt{2}h} k\left(\frac{x_i - x_j}{\sqrt{2}h}\right) - \frac{2}{n}\sum_i \left[\frac{(\hat{f}(x_i) - w_i/\sqrt{2\pi}}{1 - w_i}\right] \tag{3}$$

根据式(3),本文分别计算了固定带宽和引入带宽尺度因子核密度估计的 MISE 值,其中 h 从 100 m 起始,以间隔 50 m 增长到 2 000 m,结果如图 2 所示。当 $h<200$ m 时,采用固定带宽核密度估计效果优于变化带宽的估计;当 $h=200$ m 时,固定带宽的估计达到最优 $\mathrm{MISE}_{\mathrm{fix}}^{\min}=20$;当 $h>200$ m 时,变化带宽的效果优于固定带宽;当 $h=350$ m 时,变化带宽的估计效果达到最优 $\mathrm{MISE}_{\mathrm{var}}^{\min}=16.6$,其中 $\mathrm{MISE}_{\mathrm{fix}}^{\min} < \mathrm{MISE}_{\mathrm{var}}^{\min}$,这表明使用本文所提出的方法引入带宽尺度参数后,核密度的估计效果得到了进一步提高。

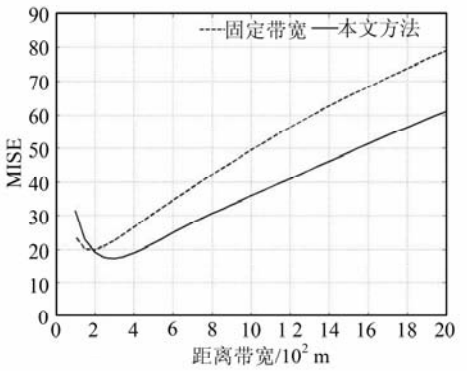

图 2　不同搜索带宽下的 MISE

Fig. 2　MISE of Different Searching Bandwidth

当最优带宽选定后,为了进一步比较固定带宽和引入带宽尺度因子的核密度方法的估计效果,本文借鉴文献[10]提出的对数概率方法[10]。该方法同样是一种交叉验证的方法,将样本点分为训练数据集和测试数据集,用训练数据集来构建核密度估计,然后计算测试数据集的对数概率,具体的计算公式为:

$$L = \frac{1}{n_t}\sum_{r=1}^{n_t} \lg \hat{f}(x_t) \tag{4}$$

式中,x_t 为测试数据集中元素;n_t 为测试数据集的个数;$\hat{f}(x_t)$ 为用训练数据集来估计 x_t 的核密度值;L 表示测试点被分配的平均概率值,越大表示测试集被分配的概率值越高,核密度估计效果也越好。

本文每次从手机基站中随机选取 500 个基站作为测试集,剩余基站为训练数据集,分别采用固定带宽和引入带宽尺度参数的核密度估计测试集的密度值,利用式(4)来计算测试集的平均对数概率值,共进行 50 次实验,结果如图 3 所示。所有的实验中 L_{var} 都要大于 L_{fix},引入带宽尺度参数后对数概率值平均增大了 0.35 左右,这表明与固定带宽核密度估计相比,本文所提出的变化带宽方法进一步减小了核密度估计的误差。

采用上述两种方法来比较固定带宽和变化带宽的核密度估计效果,结果表明,变化的带宽会进一步减少固定带宽核密度估计的误差。

3.2　人群时空停留分布差异

本文采用改进的方法来分析城市人群时空停留分布差异。首先根据城市人群的日常生活规律将一天切割成 5 个典型的时段。如表 2 所示,在

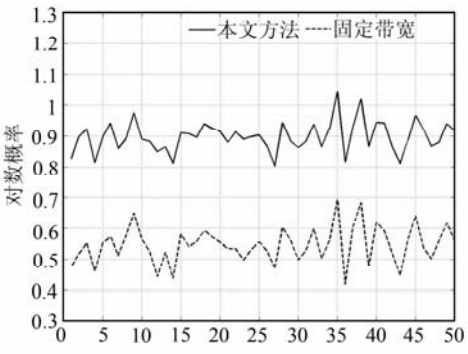

图3 两种方法的对数概率值

Fig. 3 Log-Probability of Two Methods

T_1 时段城市中大部分人群在家进行睡觉休息，T_2 和 T_4 是城市人群的主要工作时间段，T_3 和 T_5 是人群时间比较自由的时段，可以参加一些其他活动时间(如吃饭、购物、娱乐等活动)。本文忽略了早晚两个通勤时间段06:00～09:00和17:00～19:00，在这两个时间段有大量的人群在移动，对于研究城市人群的移动模式或交通同样非常重要，但不是本文的研究重点。

表2 5个典型的时段

Tab. 2 Five Typical Time Intervals

时段	时间
T_1	00:00～06:00
T_2	09:00～12:00
T_3	12:00～14:00
T_4	14:00～17:00
T_5	19:00～22:00

从数据中分别提取以上5个时段各基站上停留人数，停留时间阈值 ΔT 设置为30 min，将各时段基站的停留人数作为式(2)权重，初始带宽 h_0 选择350 m，进行核密度估计得到 f_{T_1}、f_{T_2}、f_{T_3}、f_{T_4}、f_{T_5}。采用各时段间的密度差值分析人群在时空分布的差异，定义 $f_{ij}=f_{T_i}-f_{T_j}$ 为时段 T_i 和 T_j 的密度差。对单个格网来讲，$f_{ij}>0$ 表示与 T_i 时段相比，T_j 时段人群增加，$f_{ij}=0$ 表示人群没有变化，$f_{ij}<0$ 表示人群减少。本文分析了 f_{21}、f_{32}、f_{42}、f_{54}、f_{51}，为了突出显示人群变化较显著的区域，选取 $f_{ij}>500$ 和 $f_{ij}<-500$ 的区域，结合百度地图来分析不同时段人群停留区域的功能特点。

f_{21} 给出了早上工作时段与晚上睡觉时段人群停留的空间分布差异，可以帮助了解城市中主要的工作区和居住区。如图4所示，在工作时段 T_2，人群主要集中在福田区的车公庙、市民中心和华强北商业区，罗湖区的老街和国贸商业区，南山区的深圳大学和科技园区，富士康和华为工业区以及一些位于宝安区和龙岗区的工业园区。这些区域基本聚集了城市大多数工作岗位，在工作时段吸引了大量的人群。在 T_1 时段，人群的分布区域较广，这些区域覆盖了城市中一些主要的居住社区，包括南山区的前海和后海，福田区的沙头和赤尾，罗湖区的清水河和黄贝岭，宝安区的西乡、民治，龙岗区的布吉，以及分布在宝安区和龙岗区工业园附近的居住区。

f_{32} 可以帮助分析在中午休息时间段哪些区域的人群较早上工作时间段有明显增加，这些区域在中午的时候人群比较活跃。图5给出了 $f_{32}>500$ 的区域，通过与规划图进行比较，发现这些区域主要分布在一些商业区、居住区和旅游区，因为人群在中午有短暂的自由时间，可以选择在这段时间离开工作地去商业区就餐、购物等，离家近的人群可以选择回家休息等。在一些旅游区如世界之窗和欢乐谷、梧桐山和碧海湾风景区等，在中午旅游的人群明显增加，因为景区在非节假日一般是早上10:00左右才对外营业，可能一些人群吃过午饭后才去景区游玩。

图6给出了 f_{42} 的空间分布，它可以帮助分析下午工作时段和早上工作时间人群的空间分布差异，可以看到市中心一些商业区的人群在下午还会继续增加，如华强北商业区和国贸商业区，而一些居住区尤其是位于福田区和罗湖区商业中心周围的居住区在下午上班时间人群还会继续减少。

f_{54} 可以帮助分析下午下班后人群都流向哪些区域，如图7所示，可以看出在 T_5 时段，人群主要分布在城市的居住区，与图4(b)的空间分布大致相同，但覆盖范围较小，表明一些人群在下班后(19:00～22:00之间)并没有立即回家，而是在其他区域参与一些活动。为了分析人群在这段时间主要在哪些区域活动，并且排除居住区的干扰，用时段 T_5 和时段 T_1 做密度差得到 f_{51}，如图8所示，可以看出晚上人群的活动区域主要分布在深南大道及沿线(图8中粗线道路)两侧的区域以及位于宝安区和龙岗区的一些零星区域，这些区域主要是一些市中心的商业区，聚集着大量的餐馆、购物广场和娱乐场所等，因此一些人群下班后可能在这些区域进行就餐、购物或参加一些娱乐等活动后才回家。还包括一些工作区如位于IT科技园的腾讯大厦及其附近区域、华为工业园等，在这些地方工作的人群可能在下午下班后还需继续加班。

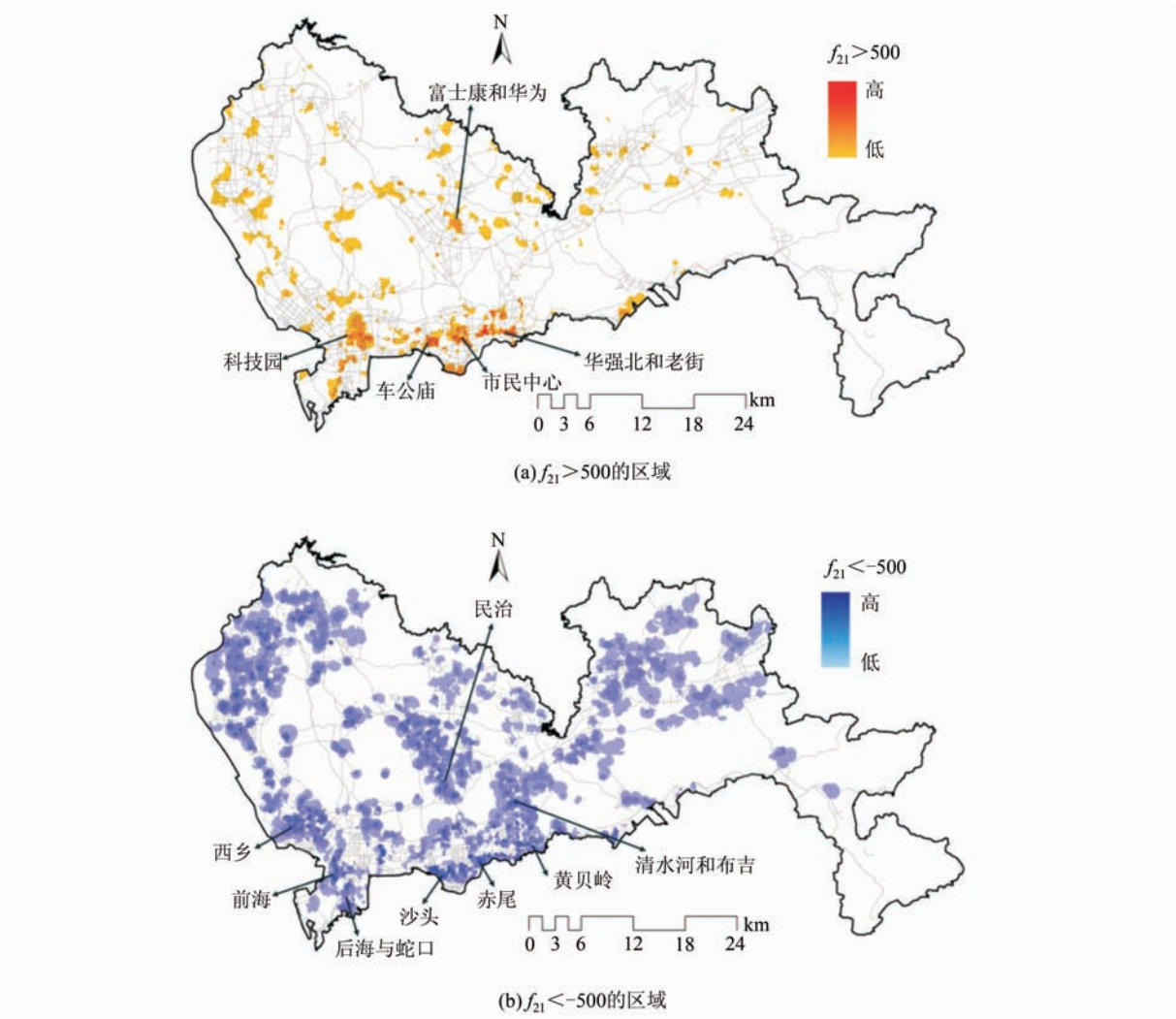

图 4　$f_{21}>500$ 和 $f_{21}<-500$ 的区域

Fig. 4　The Area of $f_{21}>500$ and $f_{21}<-500$

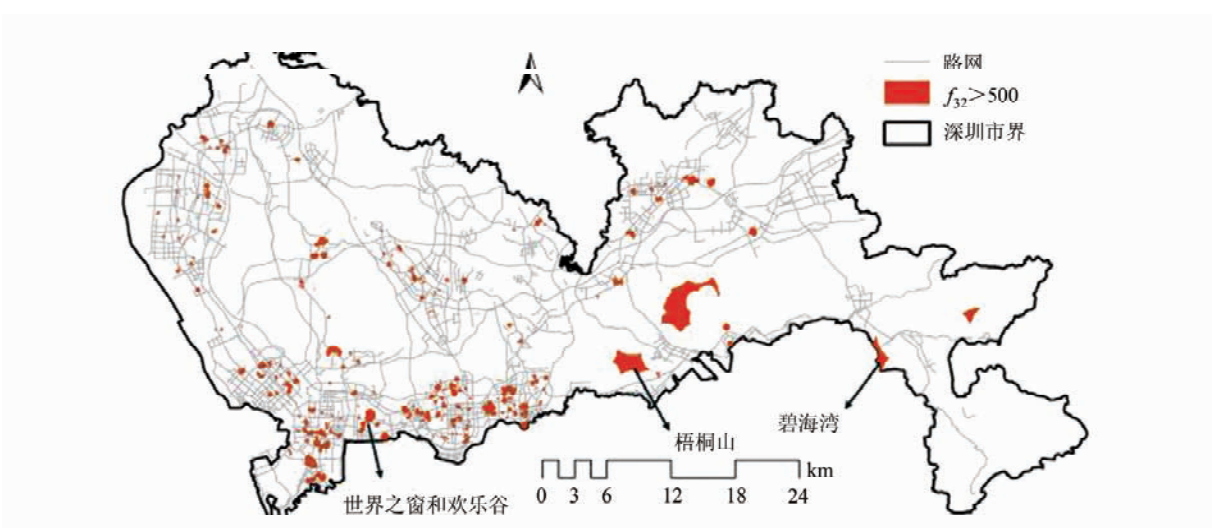

图 5　$f_{32}>500$ 的区域

Fig. 5　Area of $f_{32}>500$

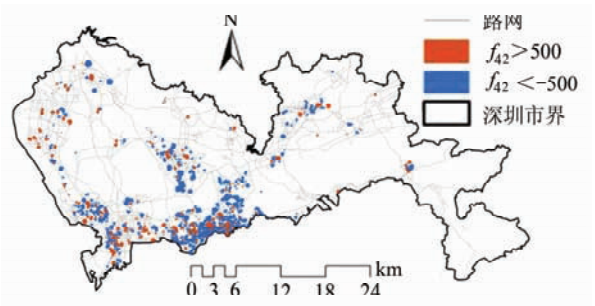

图6 $f_{42}>500$ 和 $f_{42}<-500$ 的区域

Fig. 6 Area of $f_{42}>500$ and $f_{42}<-500$

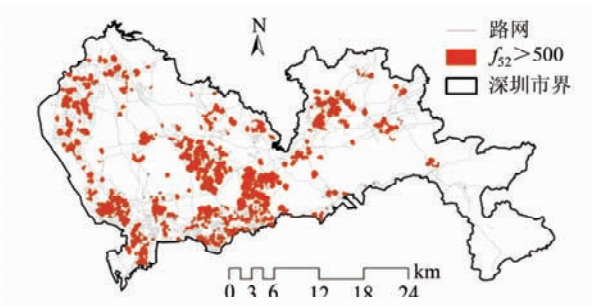

图7 $f_{54}>500$ 的区域

Fig. 7 Area of $f_{54}>500$

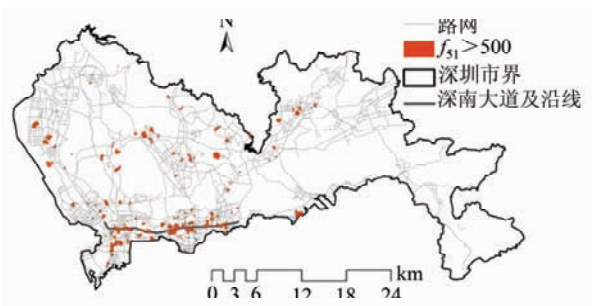

图8 $f_{51}>500$ 的区域

Fig. 8 Area of $f_{51}>500$

4 结 语

手机位置数据为研究时空高分辨率的城市人群活动提供了机遇和挑战，其中一个挑战就是手机位置数据是采用基站进行定位的，并不是人群的具体位置，因此需要估计人群在连续空间上的分布。核密度法已经被用来作为人口分布连续模型，但传统的核密度采用固定带宽而不考虑样本点的空间分布和属性，这会给人口分布估计带来误差，尤其是在人群高密度的区域。针对这个缺陷，本文根据深圳市手机数据，结合手机基站的空间分布特点，在计算核密度时加入带宽控制参数，使得搜索带宽随着基站的分布变化，通过与固定带宽核密度估计进行对比分析，发现变化的带宽可以减少核密度估计带来的误差。最后，从手机数据中提取出几个典型时段基站的停留人数，然后采用改进的核密度方法估计人群分布，通过时段间的密度差来分析城市人群在不同时段的停留分布差异。这些高分辨率的人群时空停留分布可以帮助理解人群使用城市空间的规律，从而帮助指导城市规划，根据人群停留推测土地利用、商业设施选址以及建立基于人群时空停留的城市交通流预测模型等。

参 考 文 献

[1] Mao Xia, Xu Rongrong, Li Xinshuo, et al. Fine Grid Dynamic Features of Population Distribution in Shenzhen[J]. *Acta Geographica Sinica*, 2010, 65(4):443-453(毛夏，徐蓉蓉，李新硕，等. 深圳市人口分布的细网格动态特征[J]. 地理学报，2010，65(4)：443-453)

[2] Liu Yu, Xiao Yu, Gao Song, et al. A Review of Human Mobility Research Based on Location Aware Devices[J]. *Geography and GeoInformation Science*, 2011, 27(4):8-13 (刘瑜，肖昱，高松，等. 基于位置感知设备的人类移动研究综述[J]. 地理与地理信息科学，2011，27(4)：8-13)

[3] Shaw Shihlung, Fang Zhixiang. Rethinking Human Behavior Research from the Perspective of Space-time GIS[J]. *Geomatics and Informationg Science of Wuhan University*, 2014, 39(6):667-670(萧世伦，方志祥. 从时空GIS视野来定量分析人类行为的思考[J]. 武汉大学学报·信息科学版，2014，39(6):667-670)

[4] Ratti C, Williams S, Frenchman D, et al. Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis[J]. *Environment and Planning B Planning and Design*, 2006, 33(5): 727-732

[5] Deville P, Linard C, Martin S, et al. Dynamic Population Mapping Using Mobile Phone Data[J]. *Proceedings of the National Academy of Sciences*, 2014, 111(45): 15 888-15 893

[6] Kang C, Liu Y, Ma X, et al. Towards Estimating Urban Population Distributions from Mobile Call Data[J]. *Journal of Urban Technology*, 2012, 19(4): 3-21

[7] Lu Anmin, Li Chengming, Lin Zongjian, et al. Spatial Continuous Surface Model of Population Density[J]. *Acta Geodaetica et Cartographica Sinica*, 2003, 32(4):344-348 (吕安民，李成名，林宗坚，等. 人口密度的空间连续分布模型[J]. 测绘学报，2003，32(4)：344-348)

[8] Yu Wenhao, Ai Tinghua. The Visualization and Analysis of POI Features Under Network Space Supported by Kernel Density Estimation[J]. *Acta Geodaetica et Cartographica Sinica*, 2015, 44(1): 82-90(禹文豪，艾廷华. 核密度估计法支持下的网络

空间 POI 点可视化与分析[J]. 测绘学报，2015，44(1)：82-90)

[9] Yan Qingwu, Bian Zhengfu, Zhang Ping, et al. Census Spatialization Based on Settlements Density [J]. *Geography and Geo-Information Science*, 2011, 27(5): 95-98 (闫庆武，卞正富，张萍，等. 基于居民点密度的人口密度空间化[J]. 地理与地理信息科学，2011，27(5)：95-98)

[10] Lichman M, Smyth P. Modeling Human Location Data with Mixtures of Kernel Densities[C]. The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2014

[11] Carlos H A, Shi X, Sargent J, et al. Density Estimation and Adaptive Bandwidths: A Primer for Public Health Practitioners[J]. *International Journal of health geographics*, 2010, 9(1): 1-8

[12] Tu Wei, Li Qingquan, Fang Zhixiang. A Heuristic Algorithm for Large Scale Vehicle Routing Problem [J]. *Geomatics and Information Science of Wuhan University*, 2013, 38(3): 307-310 (涂伟，李清泉，方志祥. 一种大规模车辆路径问题的启发式算法[J]. 武汉大学学报·信息科学版，2013，38(3)：307-310)

[13] Wang B, Wang X. Bandwidth Selection for Weighted Kernel Density Estimation[J]. *The Electronic Journal of Statistics*, 2007, DOI: 10.1214/154957804100000000

Analyzing Space-Time Variation of Urban Human Stay Using Kernel Density Estimation by Considering Spatial Distribution of Mobile Phone Towers

YANG Xiping[1]　*FANG Zhixiang*[1]　*ZHAO Zhiyuan*[1]　*SHAW Shihlung*[1]　*YIN Ling*[2]

1　State Key Laboratory of Information Engineering in Surveying, Mapping, Remote and Sensing, Wuhan University, Wuhan 430079, China

2　Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

Abstract: The spatial-temporal distribution of urban human concentrationscan reflect the law of use of urban spatial structures, so understanding spatial-temporal patterns of human concentrationsis helpful for optimizing urban structures, selecting public facility locations and predicting traffic flows. In recent years, the availability of mobile phone location data provides an opportunity and challenge for studying human concentrations patterns. Therefore, we only can extract these patterns based on base stations in mobile phone datasets, to produce a continuous population distribution. Kernel density estimate (KDE) can generate a continuous surface and has been widely used to estimate population distribution, but the traditional KDE assumes that the sample data points are homogeneous and uses fixed bandwidth to estimate densityfrom all data points, however, the service area of base stations in the city varies with the distribution of population distribution, so fixed bandwidth will create error. In order to eliminate errors, we introduce a search bandwidth control parameter to make the bandwidth vary with the spatial distribution of mobile phone towers. Least-squares cross validation (LSCV) and log-probability methods were used to test the proposed approach. Experimentalresultsdemonstrate that this improvement can make the estimation better than fixed bandwidths. Taking mobile location data of Shenzhen as an example, we extracted urban human concentrationsfor five typical time intervals, and the improved KDE was used to analyze the distribution difference of the five time intervals, providing deep understanding of condition of urban different areas as used by humans and how it varies over time.

Key words: mobile phone data; kernel density estimation; human stay; space-time analysis

First author: YANG Xiping, PhD candidate, specializes in spatial-temporal data analysis and mining. E-mail: 0yangxiping0@163.com

Corresponding author: FANG Zhixiang, PhD, professor. E-mail: zxfang@whu.edu.cn

Foundation support: The National Natural Science Foundation of China, Nos. 41231171, 41371420; The Independent Research Program of Wuhan University, No. 2042015KF0167; The Open Research Fund Program of State Key Laboratory of Resources and Environmental Information System, No. 2013.