

基于手机数据的城市人口分布感知

徐仲之，曲迎春，孙黎，王璞

(中南大学交通运输工程学院 长沙 410075)

【摘要】提出了一种基于手机数据、考虑手机市场占有率为城市人口分布感知方法，并将该方法应用于美国旧金山湾区；介绍了使用手机数据动态感知城市人口分布的潜在可能，并计算了湾区各个小区白天与夜间的人口数量差。研究结果表明，基于手机数据的城市人口分布感知，对城市突发事件预警、城市交通管控、城市公共资源配置等方面都有着重大意义。

关键词 人类动力学；手机数据分析；人口分布感知；城市交通

中图分类号 N94 文献标志码 A doi:10.3969/j.issn.1001-0548.2017.01.018

Urban Population Sensing via Mobile Phone Data

XU Zhong-zhi, QU Ying-chun, SUN Li, and WANG Pu

(School of Traffic and Transportation Engineering, Central South University Changsha 410075)

Abstract This paper presents an urban population sensing method based on mobile phone dataset and applies it to San Francisco Bay Area. The difference of mobile phone market shares in different tracts is considered. We introduce the potential application of dynamic population sensing using mobile phone data and calculate the relative difference of daytime population and nighttime population in different tracts in Bay Area. The knowledge of urban population distribution has great importance of the early-warning of city emergency, urban traffic control, and the allocation of city public resources.

Key words human dynamics; mobile phone data; population sensing; urban transportation

区域人口数量和区域人口分布对于国家政策的制定^[1-2]、区域经营决策的制定^[3]、人类行为的定量化分析^[4]等方面都具有着重要作用^[5-7]。在过去的数十年间，人口分布的相关研究进展迅速，出现了很多新模型、新方法。从最初的依靠人力进行人口普查的方式，发展到依靠遥感卫星进行感知的方法、使用地理信息系统进行建模的方法等^[8]。部分发达国家亦开展了一些国家层级的人口分布测量项目，取得了相关成果。然而，多数人口分布测量方法较复杂，实施难度较高，数据获取较困难，导致世界许多地区的人口分布信息时效性差，更新较慢，甚至缺乏。

20世纪90年代，人口分布感知的相关研究逐渐兴起，文献[9]回顾了20世纪主要的人口分布感知技术，重点回顾了地理信息系统(geographic information system, GIS)技术的进步对人口分布感知技术发展所起到的推动作用。文献[10]介绍了一种人口分布信息栅格化方法，人口分布信息的栅格化虽然提高了人口分布数据的精度，但削弱了同地理语

义的结合。文献[11]介绍了全球人口分布感知项目LandScan，该项目在提升人口分布感知精度的同时，保留了地理语义，能够更精确地应用于地质灾害预测、疾病管控等相关研究。文献[12]提出一种采用了数据融合技术的人口分布感知方法，融合了人口普查数据和卫星数据，提高了非洲人口分布测量的分辨率，并且基于此测量结果，分析了人口空间分布中心性和偏远地区对于人口稠密区的可达性。研究发现绝大部分人分布在极少的地方(21%的土地含盖了90%的人口)，这为通讯设备的建立，生活服务设施的建立提供了相关指导。覆盖整个东南亚地区的人口分布数据主要来自2000年的人口普查数据，空间分辨率不足。文献[13]结合人口普查数据、卫星数据和土地使用率数据，将分辨率提高至100 m左右。以上相关研究虽然从不同方面改进了人口分布感知技术，但是由于其采用的数据(遥感数据或普查数据)获取困难，导致时效性较差。

如上文所述，人口分布在空间分辨率上的研究非常丰富，但在时间分辨率上并未得到足够的重视。

收稿日期：2016-01-06；修回日期：2016-08-23

基金项目：国家自然科学基金面上项目(61473320)；霍英东青年教师基金基础研究课题(141075)；湖南省科技计划项目(2015RS4011)

作者简介：徐仲之(1991-)，男，主要从事数据挖掘、人类动力学和复杂网络方面的研究。

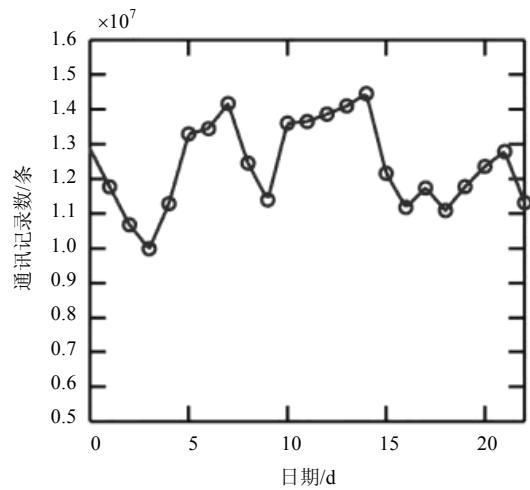
文献[14]提出了一种动态测量人口分布的方法, 以旧金山为例动态测量了旧金山的夜间和白天人口。为了达到动态测量的目的, 该方法结合了卫星数据、土地使用率数据、用地类型数据、路网数据等。文献[15]以土地利用类型作为建模的媒介, 融合人口普查数据、土地利用空间数据和建筑物空间数据, 建立了“人口-昼夜-土地利用”关系模型, 动态感知北京市昼夜人口分布变化。文献[14-15]提出的方法虽然能够动态测量城市人口分布, 但使用数据多样且难以获取, 建模方法复杂。

手机是一种良好的信息采集器。随着全球范围内手机普及率的提高, 有大量手机信令数据可以被采集和利用^[16-17]。手机数据海量、实时、易获取的特性, 使其越来越广泛地应用于交通工程^[18-19]、城市规划^[17,20]等研究领域, 也为城市人口分布感知提供了新的方向, 使动态感知人口分布成为可能。文献[21]利用葡萄牙和法国某通讯公司数个月的手机数据, 动态感知了葡萄牙和法国人口分布情况。文献[22]简要介绍了基于移动基站的人口分布动态监测系统, 该系统可提供较精细时间分辨率的人口分布信息。然而, 文献[21-22]所使用的数据量庞大, 在手机信令数据不十分丰富的区域难以展开, 且当应用于全网实时信令分析时, 交换机和相关信令链路的负荷过高, 并需要巨量的成本投入; 小样本数据问题和数据分布不均匀性问题未得到解决。

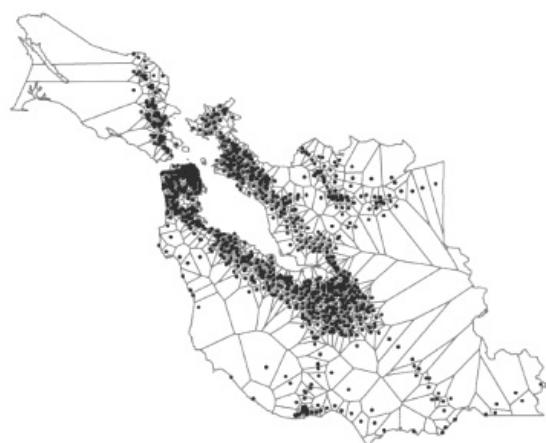
1 旧金山湾区手机数据和人口分布信息

1.1 湾区手机数据信息

本文使用湾区手机数据感知湾区人口分布信息。数据来自美国某通信公司, 记录了2010年中连续的21天, 共429 595个手机用户的通话详单(call detail records, CDR)数据信息, 平均每天约有1 200 000条手机CDR数据, 如图1a所示。当手机使用者进行通话或发送短信时, 其通信时刻及通信基站会被记录下来。如图1b所示, 根据泰森多边形算法^[23], 将湾区按基站位置划分为若干多边形(即基站小区), 使得基站和小区一一对应。通过每个小区包含的基站信息, 可以确定一条手机CDR信息的发生小区。本文把手机用户在20:00到第二天7:00间手机CDR记录次数最多的小区定义为其住址小区^[24]。被选取的手机用户在所有21天的20:00到第二天7:00时间段中需至少有一条CDR信息, 用于确定其住址小区。据此从中选取了360 612个手机用户的CDR信息来进行人口分布感知研究。



a. 手机CDR数据量

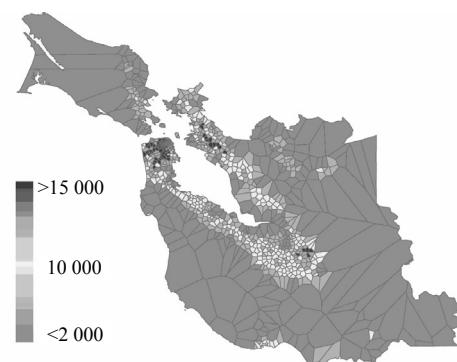


b. 湾区手机基站及基站小区示意图

图1 手机数据量信息及基站分布示意图

1.2 湾区夜间人口分布信息

图2a展示了湾区夜间人口密度分布情况, 数据为人口普查数据, 旧金山、奥克兰、圣荷西和尤宁城等大城市人口密度较大。郊区多为山林, 人口密度较小。湾区各小区面积分布如图2b所示, 可以看出, 绝大部分的小区面积都较小, 人口分布感知结果的精度将会较高。



a. 湾区夜间人口分布

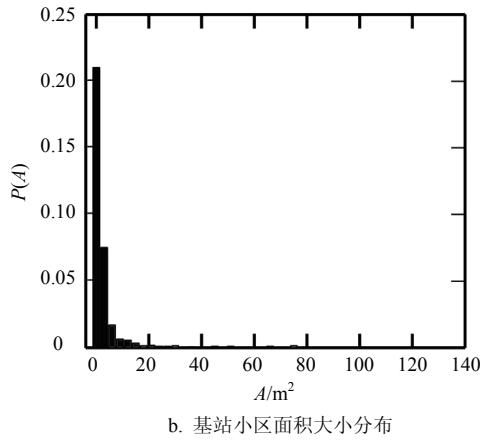


图2 湾区夜间人口分布

2 基于手机数据的城市夜间人口分布感知

文献[25-28]的结果表明，人口密度与人口活跃程度存在超线性关系。文献[21]使用非线性方程 $\rho_c = \alpha(\sigma_c)^\beta$ 表征人口密度与手机用户活跃度的关系。其中， σ_c 表示 c 小区夜间手机用户密度， ρ_c 表示 c 小区的夜间人口密度。研究表明，非线性方程 $\rho_c = \alpha(\sigma_c)^\beta$ 有着很好的拟合效果。

在湾区，当手机用户使用手机进行通信(通话/短信/上网)时，距离其最近的基站会被选择使用。其通信的起始时间和所用基站会被记录下来。因此，对于湾区中的 c 小区，可以计算出其夜间手机用户数 N_c ，夜间手机用户数密度即为 N_c / A_c (A_c 为 c 小区的面积)。根据文献[21]中人口密度与手机用户活跃度的关系表达式， c 小区夜间人口密度与夜间手机用户密度的关系可表示为：

$$\rho_c = \alpha(N_c / A_c)^\beta \quad (1)$$

式中， α 反映了人口密度与手机用户密度的大致比例关系； β 反映了在人口密度较大的小区，人口活跃度和人口密度的超线性关系。文献[25-28]显示， β 略低于 1。基站小区 c 的人口密度 ρ_c 采用公式 $\rho_c = \sum_v \omega_v A_{(c \cap v)} / A_c$ 计算，其中， ω_v 指行政区 v 的人口密度，数据来自 2010 年美国人口普查局人口普查数据； A_c 指基站小区 c 的面积； $A_{(c \cap v)}$ 指基站小区 c 与行政区 v 的重合面积。由此，每个基站小区人口密度 ρ_c 是已知的，可以使用线性回归的方法来标定式(1)中的 α 值和 β 值。

手机市场占有率为一个区域作为研究对象的手机用户的占比，受手机在该区域的普及度和所使用数据通讯商的市场份额等影响。在城市的不同区域，手机市场占有率是不同的。国家级别的人口分

布感知，由于研究区域较大，精度需求较低，手机市场占有率为影响较小^[21]；城市级别的人口分布感知要求更高的精度，需要考虑手机市场占有率为因素。本文使用公式 $P_c = R_c / U_c$ 计算小区 c 的手机市场占有率为^[29]，其中， R_c 表示 c 小区的人口普查数据， U_c 表示整个研究时间段(即 21 天)中所侦测到的以 c 小区为住址小区的手机用户数量。 c 小区手机市场占有率为 P_c 和扩样系数 λ_c 之间的关系为 $P_c = 1 / \lambda_c$ 。式(1)修改为：

$$\rho_c = \alpha(\lambda_c N_c / A_c)^\beta \quad (2)$$

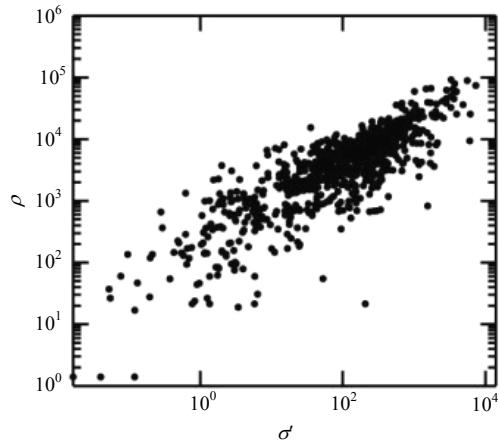
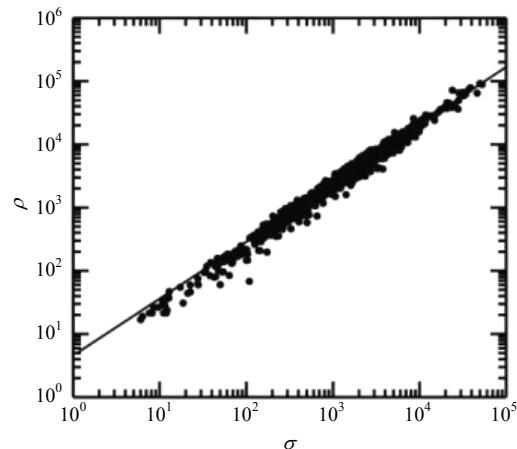
a. 未考虑手机市场占有率为因素的夜间人口密度 ρ 与手机用户密度 σ 的关系b. 考虑手机市场占有率为因素夜间人口密度 ρ 与手机用户密度 σ 的关系

图3 夜间人口密度与手机用户密度关系图

图3a展示了未考虑手机市场占有率为因素时夜间人口密度与手机用户密度的关系，图3b展示了考虑手机市场占有率为因素的情况下夜间人口密度与手机用户密度的关系，可见考虑手机市场占有率为因素在本研究中是必要的。本文将建立多元回归分析模型，求解参数 α 和 β 。

回归式(2)中，其回归参数是非线性的。但将回

归式(2)改写为 $\log(\rho_c) = \log(\alpha) + \beta \log(\lambda_c N_c / A_c)$ 的形式, 便可将非线性回归方程转变为线性回归方程, 进而求解回归参数为:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \varepsilon \sim \sigma^2 \quad (3)$$

线性回归模型式(3)用来确定参数 β_0 和 β_1 的值, 在本问题中, $\beta_0 = \log(\alpha)$, $\beta_1 = \beta$, 数据集 (x_i, y_i) ($i = 1, 2, \dots, n$, $n = 892$) 中, $x_i = \log(\lambda_c N_c / A_c)$, $y_i = \log(\rho_i)$ 。线性回归模型式(3)中, 参数 β_0 和参数 β_1 均服从正态分布, 回归模型的回归度越高, β_0 的均值越接近于0, β_1 的均值越接近1。本文使用最常用到的最小二乘法来解决此线性回归问题。结果显示, $\beta_0 = 0.585$, 其置信度为95%的置信区间为(0.563, 0.601), $\beta_1 = 0.929$, 其置信度为95%的置信区间为(0.922, 0.936), 回归模型相关系数 $R^2 = 0.98$, 均方根误差 $rmse = 0.1$, 说明变量 X 与变量 Y 有着高度相关性。

根据式(2)中参数 α 和 β 与回归式(3)中参数 β_0 和 β_1 之间的关系, 可以计算出 $\alpha = 3.848$, $\beta = 0.929$ 。由此可以得出湾区夜间人口密度与手机用户密度的关系 $\rho_c = 3.848 (\lambda_c N_c / A_c)^{0.929}$ (图3b)。最后, 依据人口普查数据湾区总人口数 P , 对小区 c 的人口密度进行调整, 即 $\rho_c = P \alpha (\lambda_c N_c / A_c)^\beta / \hat{P}$, 使湾区总人口感知结果与普查数据相一致。

根据湾区手机CDR数据信息, 可以得到任一小区夜间手机用户数 N , 由式(3)便可以计算出任一小区夜间人口密度, 进而完成湾区夜间人口分布感知。如图4所示。图4a展示了根据湾区人口普查数据绘制的湾区人口密度分布图, 用以表征湾区夜间人口分布, 用作基础数据。图4b为基于手机数据的湾区夜间人口密度分布感知结果。本文以各小区感知人口数量为横坐标, 普查人口数量为纵坐标进行线性拟合, 拟合结果如图4c所示, 可以看出, 基于本文提出的城市人口分布感知方法所得到的湾区夜间人口密度感知结果和人口普查结果一致性很高。

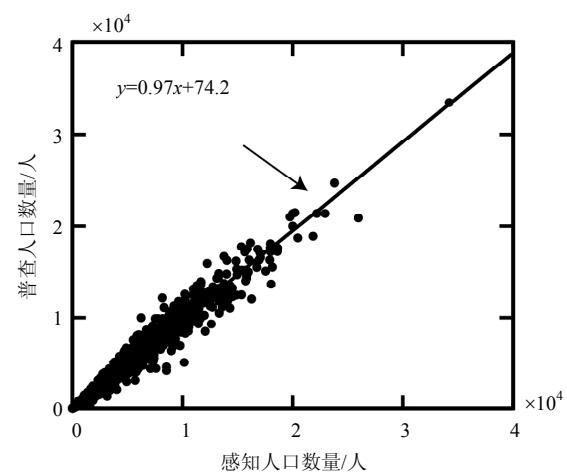
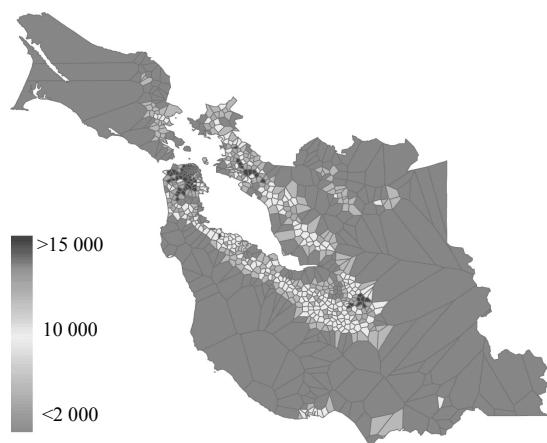
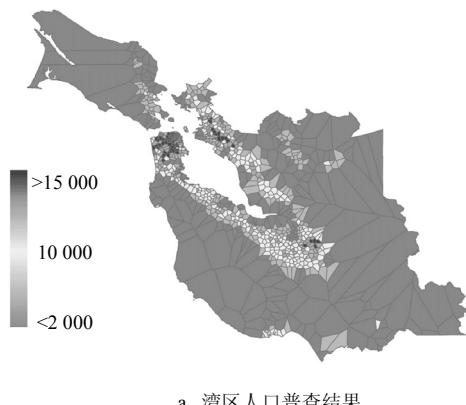


图4 湾区人口普查结果和人口感知结果关系图

3 基于手机数据的城市白天人口分布感知

一个手机用户的扩样系数取决于其所在住址小区的扩样系数。夜间, 在同一个小区 c 侦测到的手机用户, 自然具有相同的扩样系数 λ_c ; 然而在白天, 由于人口移动行为, 在同一个小区 c 侦测到的手机用户, 其住址小区可能属于不同的小区, 其扩样系数亦不相同。如图5所示, 小区 A 的扩样系数 $\lambda_A = 4$, 住址小区为小区 A 的手机用户, 若侦测到其白天产生了小区 A 至小区 C 的出行, 应将其等效为一次4人次出行, 以消除手机市场占有率的影响。因此, 本文使用式(4)求取白天某时段的湾区人口分布感知情况:

$$\rho_c = \alpha \left(\sum_{i=1}^N \lambda_i / A_i \right)^\beta \quad (4)$$

式中, N 是此时段内小区 c 的手机用户数量; λ_i 是用户 i 的扩样系数。

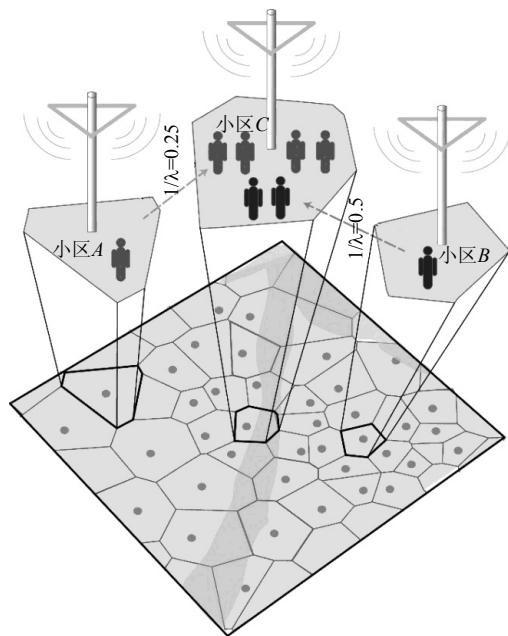


图5 考虑手机市场占有率的等效出行人次示意图

4 基于手机数据的城市动态人口分布感知

手机数据数量大, 获取速度快, 采集途径简单, 具有即时性、高效性等优点。这些特征为动态感知城市人口分布提供了可能。本研究使用的湾区手机数

据, 当CDR信息被记录时, 其触发时间戳亦被记录下来, 这是时序动态研究的基础。不同于传统的人口普查方式等所获取的人口分布数据, 基于包含时间戳的手机数据, 可以研究白天和夜间, 周末与周中, 夏季与秋季等不同时间段城市人口分布的差异, 动态地了解和认识城市人口分布和流动规律。由于所研究数据限制(时间戳未记录日期), 本文仅探究白天与夜间湾区人口密度分布的相对差, 以展示基于手机数据的城市人口分布感知方法在动态感知上的可行性。

本文将21天的手机CDR数据划分为白天部分(7:00~20:00)和夜间部分(20:00~7:00), 以此展开白天与夜间湾区动态人口分布研究。式(4)作为两个时段的人口密度计算式, α 和 β 均采用上文中的线性回归最优拟合值(即: $\alpha = 3.848$, $\beta = 0.929$)。

计算了白天与夜间湾区人口分布密度相对差, 如图6所示。可以看出明显的空间分布特征差异, 白天人口密度较大的小区基本沿高速路分布, 和文献[21]的研究结果相一致。研究时间变化对城市人口空间分布的影响, 对城市管控、交通规划等具有重要意义; 同时, 城市动态人口分布感知方法也可能对城市动态人口分布的预测以及城市人口聚集区域管控等相关热点领域的研究提供帮助和启发。

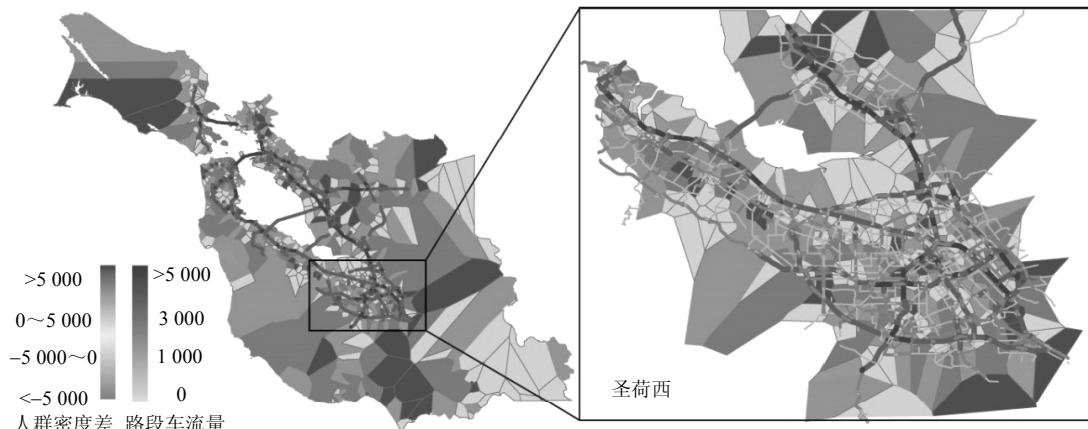


图6 湾区白天与夜间的人口相对差分布图

5 结束语

近年来, 手机在全球范围, 尤其是偏远地区的迅速普及, 移动通信覆盖范围的迅速增长, 带来了海量的手机数据信息。当手机用户使用手机进行通讯时, 通过手机和基站之间的信息交换, 其通信发生时间、所使用基站编号等均被记录下来, 这些信息使得人们能够使用手机数据, 探究手机用户活跃度与区域人口密度的关系。同时, 计算机技术发展

迅速, 存储和计算成本大幅下降, 快速处理大量数据信息成为可能。这些均为城市人口分布感知提供了新方向和新思路。不同于传统的人口分布感知方法, 基于手机数据的人口分布感知方法, 实施简单、成本低, 更利于在其他数据信息相对匮乏的偏远地区展开; 且手机数据时效性强, 相较于传统的人口分布感知方法, 基于手机数据的人口分布感知方法能够动态感知城市人口分布, 这为城市管控、城市人口流动预测等相关方向的研究提供了基础。

本文的研究结果表明, 在旧金山湾区, 区域人口密度和区域手机用户数存在超线性关系; 在小区级别的人口分布感知中, 考虑手机市场占有率因素是必要的, 考虑手机市场占有率的区域人口密度和区域手机用户数, 其关系可用表达式 $\rho_c = \alpha \left(\sum_{i=1}^N \lambda_i / A_i \right)^\beta$ 表示。

本文提出的城市人口分布感知方法, 无论在城市静态人口分布感知中, 还是在城市动态人口分布感知中, 都取得了很好的结果, 为该领域和相关领域的后续研究提供了一定的基础。

参 考 文 献

- [1] BONGAARTS J, SINDING S. Population policy in transition in the developing world[J]. Science, 2011, 333(6042): 574-576.
- [2] TATEM A J, GARCIA A J, SNOW R W, et al. Millennium development health metrics: Where do Africa's children and women of childbearing age live?[J]. Population Health Metrics, 2013, 11(1): 1-11.
- [3] CHECCHI F, STEWART B T, PALMER J J, et al. Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations[J]. International Journal of Health Geographics, 2013, 12(1): 347-360.
- [4] 周涛, 韩筱璞, 闫小勇, 等. 人类行为时空特性的统计力学[J]. 电子科技大学学报, 2013, 42(4): 481-540.
ZHOU Tao, HAN Xiao-pu, YAN Xiao-yong, et al. Statistical mechanics on temporal and spatial activities of human[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(4): 481-540.
- [5] LINARD C, TATEM A J. Large-scale spatial population databases in infectious disease research[J]. International Journal of Health Geographics, 2012, 11(1): 1-13.
- [6] O'NEILL B C, DALTON M, FUCHS R, et al. Global demographic trends and future carbon emissions[J]. Proceedings of the National Academy of Sciences, 2010, 107(41): 17521-17526.
- [7] O'LOUGHLIN J, WITMER F D W, LINKE A M, et al. Climate variability and conflict risk in East Africa, 1990-2009[J]. Proceedings of the National Academy of Sciences, 2012, 109(45): 18344-18349.
- [8] 卓莉, 黄信锐, 陶海燕, 等. 基于多智能体模型与建筑物信息的高空间分辨率人口分布模拟[J]. 地理研究, 2014, 33(3): 520-531.
ZHUO Li, HUANG Xin-rui, TAO Hai-yan, et al. The simulation of high spatial resolution population distribution based on multi-agent model and construction information[J]. Geographical Research, 2014, 33(3): 520-531.
- [9] DEICHMANN U. A review of spatial population database design and modeling[M]. California, USA: National Center for Geographic Information and Analysis, 1996.
- [10] TOBLER W, DEICHMANN U, GOTTSSEGGEN J, et al. World population in a grid of spherical quadrilaterals[J]. International Journal of Population Geography, 1997, 3(3): 203-225.
- [11] DOBSON J E, BRIGHT E A, COLEMAN P R, et al. LandScan: a global population database for estimating populations at risk[J]. Photogrammetric Engineering and Remote Sensing, 2000, 66(7): 849-857.
- [12] LINARD C, GILBERT M, SNOW R W, et al. Population distribution, settlement patterns and accessibility across Africa in 2010[J]. PloS One, 2012, 7(2): e31743.
- [13] GAUGHAN A E, STEVENS F R, LINARD C, et al. High resolution population distribution maps for Southeast Asia in 2010 and 2015[J]. PloS One, 2013, 8(2): e55882.
- [14] BHADURI B, BRIGHT E, COLEMAN P, et al. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics[J]. GeoJournal, 2007, 69(1-2): 103-117.
- [15] 戚伟, 李颖, 刘盛和, 等. 城市昼夜人口空间分布的估算及其特征——以北京市海淀区为例[J]. 地理学报, 2013, 68(10): 1344-1356.
QI Wei, LI Ying, LIU Sheng-he, et al. The estimation and characteristics of urban population distribution in daytime and nighttime—an example of Haidian, Beijing[J]. Acta Geographica Sinica, 2013, 68(10): 1344-1356.
- [16] 唐小勇, 周涛. 手机信令数据在交通规划中的应用及思考[C]//协同发展与交通实践——2015年中国城市交通规划年会暨第28次学术研讨会论文集. 杭州: [s.n.]. 2015: 25-30.
TANG Xiao-yong, ZHOU Tao. The implication and muse of mobile phone data in traffic design[C]//Coordinate Development and Traffic Practice—Proceeding of Chinese Urban Traffic Design. Hangzhou: [s.n.]. 2015: 25-30.
- [17] 郭璨, 甄峰, 朱寿佳. 智能手机定位数据应用于城市研究的进展与展望[J]. 人文地理, 2014, 29(6): 18-23.
GUO Can, ZHEN Feng, ZHU Shou-jia. The progress and anticipation of the smart phone location data in urban research[J]. Human Geography, 2014, 29(6): 18-23.
- [18] 王璞, 黄智仁, 龚航. 大数据时代的交通工程[J]. 电子科技大学学报, 2013, 42(6): 806-816.
WANG Pu, HUANG Zhi-ren, GONG Hang. Transportation engineering in the big data era[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(6): 806-816.
- [19] 冉斌. 手机数据在交通调查和交通规划中的应用[J]. 城市交通, 2013(1): 72-81.
RAN Bin. Use of cellphone data in travel survey and transportation planning[J]. Urban Transport of China, 2013 (1): 72-81.
- [20] 丁亮, 钮心毅, 宋小冬. 利用手机数据识别上海中心城区的通勤区[J]. 城市规划, 2015(9): 100-106.
DING Liang, NIU Xin-yi, SONG Xiao-dong. Identifying the commuting area of Shanghai central city using mobile phone data[J]. City Planning Review, 2015(9): 100-106.
- [21] 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): 15888-15893.
- [22] 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.
- MAO Xia, XU Rong-rong, LI Xin-shuo, et al. The dynamic network characteristics of Shenzhen population distribution[J]. *Acta Geographica Sinica*, 2010, 65(4): 443-453.
- [23] FU T, YIN X, ZHANG Y. Voronoi algorithm model and the realization of its program[J]. Computer Simulation, 2006, 23: 89-91.
- [24] WANG P, HUNTER T, BAYEN A M, et al. Understanding road usage patterns in urban areas[J]. *Scientific Reports*, 2012, 2: 1001.
- [25] TATEM A J, HUANG Z, NARIB C, et al. Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning[J]. *Malaria Journal*, 2014, 13(1): 1-16.
- [26] SCHLAPFER M, BETTENCOURT L M A, GRAUWIN S, et al. The scaling of human interactions with city size[J]. *Journal of the Royal Society Interface*, 2014, 11(98): 20130789.
- [27] GOMEZ-LIEVANO A, YOUN H J, BETTENCOURT L M A. The statistics of urban scaling and their connection to Zipf's law[J]. *PLoS One*, 2012, 7(7): e40393.
- [28] KRINGS G, KARSAI M, BERNHARDSSON S, et al. Effects of time window size and placement on the structure of an aggregated communication network[J]. *EPJ Data Science*, 2012, 1(4): 1-16.
- [29] WANG P, GONZÁLEZ M C, HIDALGO C A, et al. Understanding the spreading patterns of mobile phone viruses[J]. *Science*, 2009, 324(5930): 1071-1076.

编 辑 蒋 晓