北京大学学报(自然科学版) 第 56 卷 第 3 期 2020 年 5 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 56, No. 3 (May 2020)
doi: 10.13209/j.0479-8023.2020.035

基于手机信令数据的特大城市人口时空分布及其社会经济属性估测
——以北京市为例

海晓东[1] 刘云舒[2,3] 赵鹏军[3,†] 张辉[1]

1. 北京大学经济学院, 北京 100871; 2. 北京大学深圳研究生院, 深圳 518055; 3. 北京大学城市与环境学院, 北京 100871;
† 通信作者, E-mail: pengjun.zhao@pku.edu.cn

摘要 提出应用手机信令数据, 基于空间模式单元(Spatial Pattern Unit)进行人口动态分布估测和人口属性识别的方法, 并以北京为例开展实例研究。以手机信令数据为主, 结合大样本问卷调查数据和腾讯热力图数据, 对人口布局进行分时段估测, 分析人口分布的时空间动态特征; 采用大样本问卷调查数据, 以人口社会经济属性和通勤出行特征等关键指标, 对调查的种子空间单元进行模式分类和识别, 运用机器学习的方法进行全域地域空间的人口属性估测识别, 最后对估测结果进行对比和验证。所提方法和研究结果可以为监测人口布局动态、针对人口属性布局商业服务和合理规划城市设施等提供决策支撑。

关键词 人口时空分布; 人口属性估测; 动态监测; 机器学习; 手机信令数据

Using Mobile Phone Data to Estimate the Temporal-Spatial Distribution and Socioeconomic Attributes of Population in Megacities: A Case Study of Beijing

HAI Xiaodong[1], LIU Yunshu[2,3], ZHAO Pengjun[3,†], ZHANG Hui[1]

1. School of Economics, Peking University, Beijing 100871; 2. Shenzhen Graduate School, Peking University, Shenzhen 518055; 3. College of Urban and Environmental Sciences, Peking University, Beijing 100871;
† Corresponding author, E-mail: pengjun.zhao@pku.edu.cn

Abstract This study proposes a technique to identify the temporal-spatial distribution and socioeconomic attributes of population by using mobile phone data. This technique has a fine geographic scale, which is called as Spatial Pattern Unit. The study uses Beijing as a case and conducts an empirical application of the technique. Firstly, it investigates the temporal-spatial distribution of population in Beijing by using multiple data sources, including mobile phone data, travel survey data and heat map data. Secondly, it classifies the spatial pattern unit into different categories in terms of socioeconomic attributes of population and travel behavior features. Thirdly, it applies machine learning approach to estimate socioeconomic attributes of population for all spatial pattern units. Finally, it compares and verifies the results of analysis. The approaches and findings would be valuable to monitoring population distribution, locating business services and planning urban infrastructure.

Key words temporal-spatial distribution of population; estimation of socioeconomic attributes of population; dynamic monitoring; machine learning; mobile phone data

人口时空分布是城市理论研究的重要内容之一。人口分布特征是社会环境、城市管理以及公共服务设施的直接作用结果, 也是分析人口环境效应的关键要素[1−2]。随着我国城镇化水平的进一步提高, 人口在城市持续聚集, 尤其是特大城市发展较快。人口集聚促进城市扩张与城市改造, 重新塑造

国家自然科学基金(41925003)和北京建筑大学未来城市设计高精尖创新中心项目(udc2018010921)资助
收稿日期: 2019–05–10; 修回日期: 2020–01–16

城市人口时空间格局，为城市规划与管理带来新的挑战，同时给城市公共资源的高效配置带来新的要求。了解城市人口时空间布局特征已成为城市研究领域亟待解决的科学问题[3–4]。

对人口时空分布特征进行动态监测，对于合理安排城市应急系统也至关重要[5–6]。人口动态监测的主要目的，是在给定时间范围内，对存在于特定空间区域中人口的基本信息(性别、年龄)、流动趋向与范围、职业与居住、婚育与家庭结构等进行实时数据获取、识别和监测，并进行人口布局、流动和集聚状态的分析，为城市人口应急管理和安全疏解等决策提供及时的信息支撑。

当前，国内外关于人口时空分布特征的相关研究以普查数据和传统抽样统计数据为主[7–9]。我国现有的周期性人口统计调查制度主要包括国家统计局组织的十年一度的人口普查、五年一度的1%人口抽样调查和每年一度的1‰人口变动抽样调查，主要通过调查员入户的方式，获取各个行政单元内人口数量、结构和分布空间等情况。然而，人口普查和抽样调查的人口动态监测方法存在调查范式难以统一、空间分辨率低和时效性差的先天不足。首先，人口普查和调查必须确定统一的调查模式，在相同的时间节点，在全国范围展开入户调查，人力物力成本高，中间环节多，调查结果的精准度受人为因素的影响较大；其次，普查与调查以全国统一的行政区划范围作为统计单元，各省、市、自治区等行政区域空间差异较大，可比性差，同时，在研究中，CBD及商圈等任意兴趣区的统计数据难以获取；第三，人口普查和抽样调查以年际，甚至十年际为采样周期，具有频率低、时效性差的特点，无法满足细时间粒度的人口动态监控的需求。

近年来，随着互联网和通信技术的发展，社会信息化进程不断加快，大数据方法和技术逐步应用于人口时空分布动态监测研究中。在人口信息获取的数据来源上，学者们逐渐从传统的统计数据(普查、抽样调查和深入访谈等)转向LBS数据的研究(遥感影像数据[10–11]、热力图数据[12–13]和手机信令数据[14–15]等)。在人口动态监测的优化思路方面，主要包括统计人口的空间化研究。针对传统统计数据空间分布不均和分辨率低的问题，相关学者将统计人口数据与LBS数据进行融合，综合考虑人口分布的影响因素，实现统计人口在高尺度空间的精细化分配[11,16–17]。

基于移动通信网络业务的迅速开展，手机信令数据的应用已逐渐被研究者关注。手机等智能终端在为人们提供社交、商务等生活服务的同时，也记录了人们的时空间信息，为人口动态监测带来新的发展机遇。城市空间功能结构与居民时空间活动特征是手机信令数据研究中两个重要的方向，城市环境与居民出行之间相互影响，居民的社会活动也是城市功能区域的直观表征。相关学者基于手机信令用户的时空间分布和出行轨迹特征，对城市空间功能区域进行研究，如城镇体系的划分[18]、建成环境的评价[19]以及空间职能结构的识别与分析[20–22]。在针对城市居民活动特征的研究中，手机信令数据被应用于居民出行模式与出行量的识别[23–24]、人口分布与空间活动的动态监测[15]以及交通调查与规划[25]。

大数据在识别人口的属性方面也存在不足之处。用数据表做比喻，大数据体现为“行数多而列数少”。例如，基于手机信令数据，可以提取用户的时空间驻留和出行特征，然而用户收入、职业和家庭结构等信息却难以识别。同时，由于大数据样本的异质性(如老人和儿童等样本缺失)[26]，可能导致人口属性判别的偏差。

本文提出基于空间模式单元(Spatial Pattern Unit)的多源数据人口总量估测和人口属性识别方法，并以北京为例开展实证研究。本研究拟重点回答两个科学问题：如何对人口分布的动态特征进行刻画？如何借助大数据对人口属性进行识别？这两个问题的研究对于城市理论研究和城乡规划实践均具有一定的意义。

1 人口动态分布估测和人口属性识别方法

1.1 人口动态分布估测

空间模式单元是城市空间模式划分的基本单元，也是人口动态估测的基本统计空间。传统的人口统计数据通常以行政区作为单元，通过逐级加并，汇总得到。但是，受到环境因素(山川或河流等)以及社会经济因素(基础设施和公共服务等)的综合影响[27–28]，人口在行政区范围内分布并不均匀，导致很多本来应该没有人口分布的地区也被“赋予”人口计量。为实现对人口时空间特征更准确的识别，需要更加精细的空间和时间粒度。本文采用1 km×1

km 为基本空间模式单元[11,28], 以栅格数据为数据结构, 研究不同时段的人口分布情况。

本研究的技术路线(图 1)分为 3 个阶段: 第一阶段是应用手机信令数据, 基于基本空间模式单元的人口总量及分时段分布估测; 第二阶段是应用手机信令数据, 采用传统的基于区县行政单元的人口统计分析思路, 进行基于区县行政范围的人口总量及分时段分布估测; 第三阶段是应用腾讯热力图数据, 基于基本空间模式单元的人口总量及分时段估测, 并采用该阶段的结果, 对上述两个阶段的结果进行校核。

在第一阶段, 首先将手机信令数据与北京区县人口统计数据(该数据来自 2015 年全国 1%人口抽样调查)匹配到 1 km 网格空间模式单元中。然后, 通过空间相交计算, 统计得到网格 α 中手机常住用户的数量 L_α 和统计数据中的常住人口数量 W_α, 则居住地在网格 α 中的手机常住用户对应的扩样率 K_α 为

$$K_\alpha = L_\alpha / W_\alpha 。 \tag{1}$$

基于网格编号, 统计任意网格 β 在某一时段 τ 内的手机常住居民的居住地来源。假设来源于网格 α 的手机用户为 $l_{\alpha,\tau}$, 则其所代表的来源于网格 α 的总人口为 $l_{\alpha,\tau}/K_\alpha$。将网格 β 中所有来源的人口总数求和, 即为该时段的估测人口总数 $Z_{\beta,\tau}$:

$$Z_{\beta,\tau} = \sum l_{\alpha,\tau}/K_\alpha 。 \tag{2}$$

在第二阶段, 采用同样的方法, 可以估测出基于区县行政单元的人口时空分布。

第三阶段, 应用腾讯数据进行人口时空分布的对比和分析。基于腾讯热力图数据中的热力值与特定时空间内腾讯产品的活跃用户数成正比, 首先通过热力图数据的夜间(0—6 点)热力值与人口统计数据(来自 2015 年全国 1%人口抽样调查)进行回归拟合, 得到热力值与常住人口之间的定量关系。然后, 根据各个时段的热力值, 预测常住人口在不同时段的空间分布。最后, 选取任意时段基于基本空间单元的人口估测和基于区县单元的人口估测数据, 分别与对应时段的基于热力图数据的人口时空分布估测结果进行配对样本 T 检验分析, 从而实现人口动态分布的对比验证。

1.2 大数据人口属性识别方法

1.2.1 技术路线

如图 2 所示, 首先, 综合考虑空间区位、用地权属和居民规模等特征, 在北京市范围内选择 35 个小区作为种子单元, 即抽样调查单元。第二步, 针对用户属性与交通出行状况等核心问题, 在种子单元展开深入问卷调查。第三步, 根据种子单元的空间属性特征进行分类, 得到 N 种空间模式单元, 从而对空间模式变量进行控制。第四步, 根据空间模式单元的划分标准, 在 1 km 网格的尺度上, 对北京市进行空间模式单元的划分。第五步, 基于 N 种空间模式单元中的调研结果, 分别得到居民个体社

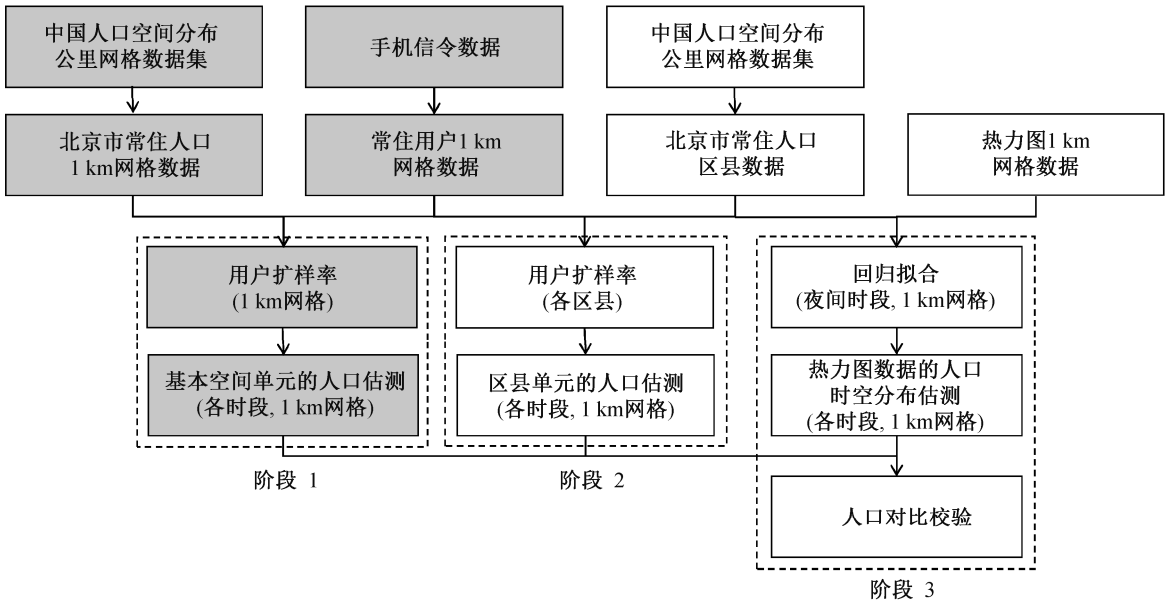

图 1 人口总量估测技术路线

Fig. 1 Total population estimation technology roadmaps

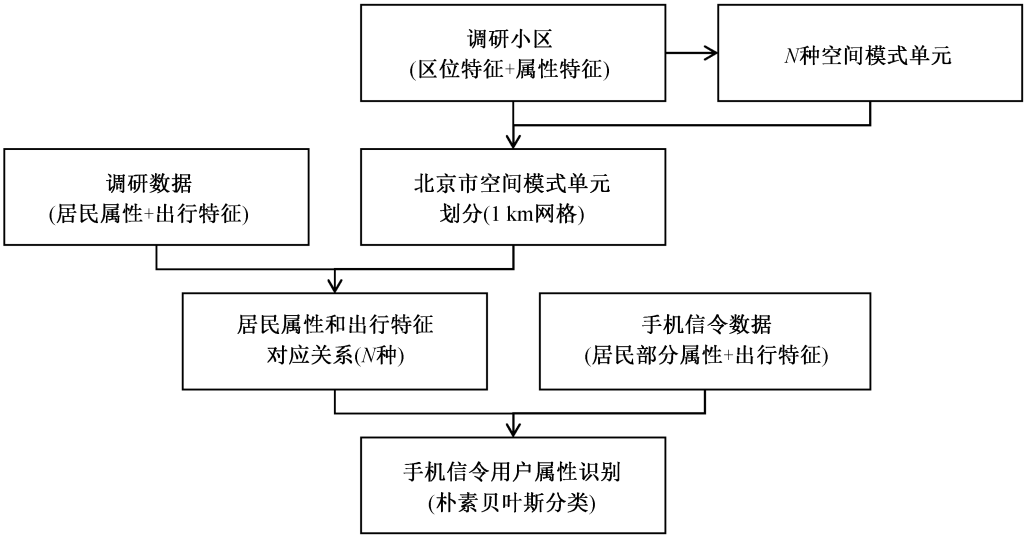

图 2　人口属性识别技术路线

Fig. 2　Socioeconomic attributes identification technology roadmaps

会经济属性和出行特征之间的组合概率。最后，运用机器学习中的朴素贝叶斯分类算法，根据调研数据中部分居民属性和出行特征出现的组合概率(即先验概率)，求解此条件下其他居民属性出现的概率(即后验概率)，并选择最大值对应的属性分类结果作为最终的识别结果。

1.2.2 空间模式单元的属性

城市的空间特征对居民出行和活动具有显著影响[29–31]。研究发现，城市空间对居民出行的影响主要体现在人口密度、土地混合度、目的地可达性和公共交通网络等方面[32–34]。人类活动密度是紧凑型土地开发模式的一种表征，主要通过居住[35–36]、就业[35,37]和人口密度[35]等进行量化表示。相关研究表明活动密度是影响居民出行的潜在因素[36]。土地利用多样性反映特定区域范围内不同土地利用类型的数量及空间配比，不同土地利用类型的交叉可以同时满足居民不同的出行目的，土地利用混合度(LUM)被引入作为表征土地利用多样性的指标[38]，有研究表明，土地利用混合度与居民出行成正向相关关系[27,39]。

目的地可达性用于衡量居住区到目的地之间的时间或空间距离，城市主要就业中心与核心商业设施是构成居民出行的两个重要目的地。在一定时间范围内可以到达的工作机会数量反映工作机会的可达性[40–41]，商业设施的覆盖范围与距离是商业可达性的核心指标[42–43]。交通设施可达性指居住区与公共交通站点之间的距离，便利的公共交通系统对居民出行有显著的促进作用。到最近地铁站/公交车站的距离、公交站点以及线网密度通常作为衡量交通设施可达性的空间指标[44–45]。鉴于此，本研究重点关注城市空间的人类活动密度、土地利用多样性、目的地可达性和交通设施可达性4个方面的属性特征(表1)。另外，房价体现空间区位对居民居住地选择和交通出行的长期影响[46]，也被选为空间模式的主要属性指标。

1.2.3 空间模式单元的分类

系统聚类算法(hierarchical clustering)又称层次聚类法，是最经典和常用的聚类方法之一，通过度量样本点之间的距离和类与类之间的关联程度来进行类别划分。本研究根据不同空间单元的属性特征，使用系统聚类算法中的离差平方和法(Ward's method)，通过方差分析，对不同的空间单元按距离准则进行逐步分类。结合系统聚类结果中的分类树状图，对聚类标准进行归纳和调整。本研究将北京35个调研小区划分为11类，分级标准见表2，聚类结果如表3所示。

1.2.4 居民属性与出行特征提取

社会经济属性是衡量和影响居民社会生活和交通出行特征的重要因素。性别和年龄是刻画居民特征的基本指标，个人月收入、有无子女和就业类型分别反映居民的个人经济水平、家庭结构和工作状况，对其出行方式、出行习惯和出行特征产生直接影响[41,47]。本研究提取的居民属性与出行特征指标如表4所示。

表 1 空间模式属性指标

Table 1 Spatial pattern unit attribute indicators

属性	指标	计算方法
人类活动密度	居住密度	基于腾讯热力图和安居客数据, 对热力值与户数之间的关系进行拟合(R^2=0.9196), 进而计算得到每个网格空间的居住密度
土地利用多样性	土地利用混合度	基于高德地图 POI 数据进行计算: $H=\sum \mathrm{ABS}(\rho_k \log_{10}\rho_k)$, 其中 ρ_k 为网格空间内第 k 类 POI 的密度
目的地可达性	市中心距离	几何中心与城市中心(天安门)之间的直线距离
交通设施可达性	最近地铁站距离	几何中心与最近邻地铁站点之间的直线距离
房价	平均房价	基于链家网房价交易数据进行空间插值(Kriging method), 通过空间统计得到每个网格空间的平均房价

表 2 属性指标分级标准

Table 2 Attribute indicator grading standards

空间模式属性指标	居住密度/(户·km^{-2})	土地利用混合度	市中心距离/km	最近地铁站距离/km	平均房价/(万元·m^{-2})
低	≤6400	≤4480	≤10.6	≤2	≤8
中	–	–	(10.6, 20]	–	–
高	>6400	>4480	>20	>2	>8

表 3 调研小区聚类结果

Table 3 Clustering results of the research communities

小区名称	居住密度	土地利用混合度	市中心距离	最近地铁站距离	平均房价	类别
大河庄苑	低	高	低	低	高	Type 1
三环新城	低	低	低	低	低	Type 2
四方景园 芳城园三区	高	低	低	低	低	Type 3
百万庄中里 大栅栏胡同区 板厂南里	高	低	低	低	高	Type 4
交东小区 朝内小区 光辉里小区 三丰里	高	高	低	低	高	Type 5
天通苑东区 北坞嘉园 卢西嘉园	低	低	中	高	低	Type 6
利泽西园 时代龙和大道	低	高	中	低	低	Type 7
燕北园 上地东里	低	低	中	低	高	Type 8
富锦嘉园 龙泽融泽嘉园 新康家园 上林溪 永乐东区 怡海花园 林肯公园 正阳北里 新华街五里 金隅丽港城	低	低	中	低	低	Type 9
天宫院小区 长虹小区 潞河名苑一期 长阳半岛	低	低	高	低	低	Type 10
建新北区 天时名苑 金隅万科城	低	高	高	低	低	Type 11

表 4　居民属性和出行特征

Table 4　Socioeconomic attributes and travel characteristics

指标分类	指标	指标详情
社会经济属性	性别 G	男、女
	年龄 A(岁)	16~69
	个人月收入 I(万元)	≤0.3; (0.3, 0.8]; (0.8, 1.5]; >1.5
	有无子女 C	有、无
	就业类型 E	基础产业、商业服务业、公共服务业、建筑与制造业、未知
出行特征	周均工作时长 T_w	一周内, 在工作地驻留的总时长
	周均通勤次数 F_c	一周内, 通勤总次数(往返记为一次)
	平均通勤时长 T_c	单次通勤所用的时间(单程耗时)

1.2.5 朴素贝叶斯分类器的构建

朴素贝叶斯分类器(naive Bayes classifier, NBC)以贝叶斯定理为理论基础, 其基本思想是, 通过待分类样本先验概率的计算, 求解出此条件下各个分类类别出现的概率(即后验概率), 并选择最大后验概率所对应的分类结果作为最终的预测结果。

假设 $x=\{x_1, x_2, x_3, \ldots, x_m\}$ 为待分类的单个居民社会经济属性集合, 其中 x_i 为性别、年龄、周均工作时长、周均通勤次数和平均通勤时长的组合, $x_i=(G, A, T_w, F_c, T_c)$。已知类别集合为 $c=\{c_1, c_2, c_3, \ldots, c_m\}$, 由单个居民社会经济属性(个人月收入、有无子女和就业类型)组成, $c_i=(I, C, E)$。基于朴素贝叶斯公式(式(3)), 对单个居民的社会经济属性进行判断(式(4)):

$$P(c \mid x)=\frac{P(c)P(x \mid c)}{P(x)}, \tag{3}$$

$$f(G, A, T_w, F_c, T_c)=(I, C, E)。 \tag{4}$$

2 案例介绍与数据

2.1 北京案例介绍

北京是我国具有代表意义的典型特大城市之一。随着城镇化的发展, 北京市常住人口快速增长, 城市空间迅速扩张, 至2016年末, 全市常住人口达2172.9万人, 建成区面积增加至1445.54 km^2[48]。北京市经济增长速度稳定, 2016年实现地区生产总值24899.3亿元, 同比增长6.7%[49]。同时, 城市交通出行需求与设施建设持续增加, 据统计, 2016年北京市中心城工作日日均出行总量为2666万人次(不含步行), 同期市级交通固定资产投资完成508.2亿元, 同比增长0.14%[50]。

与其他特大城市一样, 北京市存在住房紧张、交通拥堵和环境压力等诸多城市问题[4,51−53]。据高德地图发布的《2017年中国主要城市交通分析报告》(https://report.amap.com/share.do?id=8a38bb8660f9109101610835e79701bf), 2017年全国26%的城市早晚高峰期间交通严重拥堵, 北京市就业者平均每日通勤时间居全国首位, 达到97分钟。

2.2 本文所用数据

2.2.1 手机信令数据

手机信令数据是一款匿名、脱敏的群体性数据产品, 基于手机定位算法或蜂窝小区定位技术, 对手机用户所在的基站小区ID进行定位。基于国内某运营商数据产品, 在研究区域内采集153700199条用户驻留数据和475245320条用户出行数据(2017年9月)。

根据人口总量估测与属性扩充的要求, 对手机信令用户的常住地进行识别。在一个月内, 如果用户在相同行政区内出现超过10天, 则判定为常住用户。对常住用户每天晚上9点到第二天早上8点停留的地点分别进行时长加并排序, 时间最长的地点即为用户居住地。

2.2.2 热力图数据

热力图数据基于腾讯产品活跃用户的街道级位置定位产生, 记录分时段的人口活动强度。通过产品应用程序网络接口, 以1小时为采样周期, 在研究区域内采集2015年7月31日—8月1日共48小时的数据。原始数据包括4个属性, 分别为经度、纬度、时间和人口活动强度。其中, 人口活动强度与相同位置下的人口密度正相关。由于采样数据缺失, 仅得到北京六环内及六环周边范围内的有效数据。

2.2.3 种子单元的社区问卷调查数据

北京大学城乡规划与交通研究中心于2017年4—7月, 在北京市范围内开展居民出行调查活动, 通过问卷调研的方式获取居民的社会经济属性和交通出行信息。该调查抽样采用分层抽样与人口规模比例抽样相结合的方法, 根据小区居住人口规模差异, 对35个调研小区分别进行50~200份问卷发放, 共回收问卷4043份, 其中有效问卷3209份。调研小区的空间分布如图3所示。

2.2.4 人口统计数据

北京区县人口统计数据来自2015年全国1%人口抽样调查, 其网格形式数据来自中国人口空间分布公里网格数据集(2015)(http://www. resdc.cn/data.aspx?DATAID=251)。该数据为栅格数据类型(网格

图 3 调研小区空间分布

Fig. 3 Spatial distribution of the research communities

范围内的人口数), 基于土地利用类型、夜间灯光亮度和居民点密度等因素对人口分布权重进行综合考量, 利用多因子权重分配法, 得到 1 km×1 km 网格的常住人口空间分布。

2.3 人口时空分布估测结果

2.3.1 基于基本空间模式单元的人口总量及分时段分布估测

职住信息是城市问题研究中的重要关注点, 本文分别从休息时段(0—6 点)和工作时段(8—11 和 15—17 点)[12]中随机选取一个时段, 以北京市常住人口 1 km 网格数据为基准, 对市域范围内的人口驻留总量进行估测。根据手机信令用户的数据脱敏标准, 估测结果为 2017 年 9 月在 2 点(2:00—2:59)和 16 点(16:00—16:59)两个时段的人口驻留总量, 结果如图 4 所示。

图 4(a)和(b)分别为 2 点和 16 点手机信令用户及估测人口分布情况。可以看出, 手机信令驻留用户与估测后的驻留人口总量在中心城区的空间分布具有较大的差异。手机信令用户的人口驻留总量具有更强的中心性, 高密度区域(>140000)在五环内分散地分布, 热力特征明显。同时, 对于周边区县(昌平、通州和亦庄等)的中心区有更好的识别度。估测人口的驻留总量呈现较大的差异化特征, 其高密度区域分布较少, 集中在城市二环内, 中等密度(12000~140000)区域具有较高的识别度, 以城市中心为核心, 呈现中心对称和片状分布的特征。在不同的时间节点, 估测人口比信令人口有更好的特征代表性。在 2 点的六环及周边郊区, 估测人口密度明显增加, 这对夜间居住人口的空间分布信息是很好的补充。在 16 点, 估测人口呈现更好的中心对称的圈层结构, 城市的就业核心区(国贸、中关村地区等)具有更好的识别度。

应用手机信令数据, 采用传统的区县行政单元的人口估测思路, 进行基于区县行政范围的人口总量及分时段分布估测, 同样可以得到 2 点和 16 点两个时段的人口驻留总量。

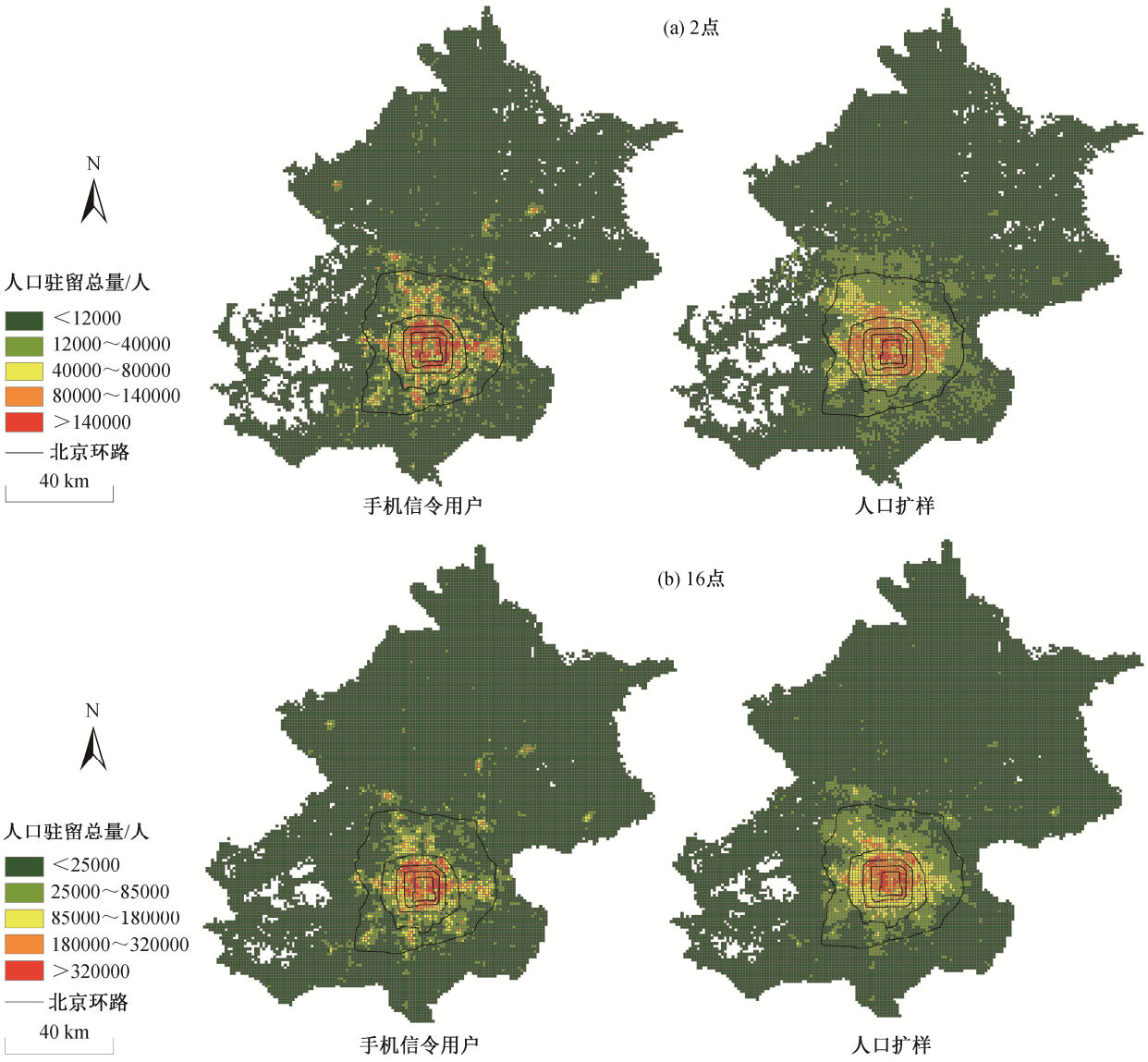

图 4　2017 年 9 月分时段人口驻留总量

Fig. 4　Distribution of total resident population in September 2017

2.3.2 基于热力图数据的人口时空分布估测

腾讯热力图数据的热力值与当前时段内活跃用户的密度成正比，因此，我们用热力图数据对常住人口的全时段样本量进行换算，为手机信令驻留人口的估测结果提供校验。首先，通过热力图数据的夜间(0—6 点)热力值与统计数据中常住人口数的回归拟合(表 5)，得到热力值 P 与常住人口 H 之间的量化关系为

$$P=1.435H+3277.388\ 。\qquad(5)$$

热力图数据为北京市六环及周边范围内的 24

表 5　热力图数据与统计数据的回归关系

Table 5　Regression result between heatmap data and statistical data

模型	非标准化系数		标准系数	t	Sig.	B 的 95.0%置信区间	
	B	标准误差				下限	上限
常量	3277.390	141.47	0.60	23.17	0.00	2999.93	3554.84
heatmap	1.435	0.04		32.54	0.00	1.35	1.52

小时全时段数据，通过任意时段内热力值的空间统计(spatial statistic)，可由式(3)得到该时空间基于热力图数据估测的常住人口数。

2.3.3 人口时空分布估测结果的对比分析

取 2 点和 16 点两个时段，应用基于热力图数据的人口时空分布估测结果，对上述手机信令数据估测结果进行比对。

首先，以热力图数据估测得到的常住人口数为基准，分别对基于区县行政单元和基于基本空间模式单元两种估测方法得到的人口总量进行标准化缩放处理(分别与热力图数据估测的常住人口进行求商计算)。两种估测方法得到的人口总量结果对比如图 5 所示。图 5(a)和(b)分别表示在 2 点和 16 点两个时段内，不同估测结果之间的量化关系，可以看出，两种估测方法得到的结果之间呈现一定的线性关系，16 点时段的线性拟合程度(R^2=0.50843)优于 2 点时段(R^2=0.32805)。同时，在不同的时段，与基于基本空间模式单元的估测结果相比，基于区县行政单元的估测结果均存在明显的高估现象，这与扩样率计算的空间尺度有关，更大的空间尺度导致扩样率的高估。

进一步地，通过配对样本 T 检验进行分析，得到两两样本之间的统计关系，如表 6 所示。可以看出，基于基本空间模式单元的人口总量和基于区县

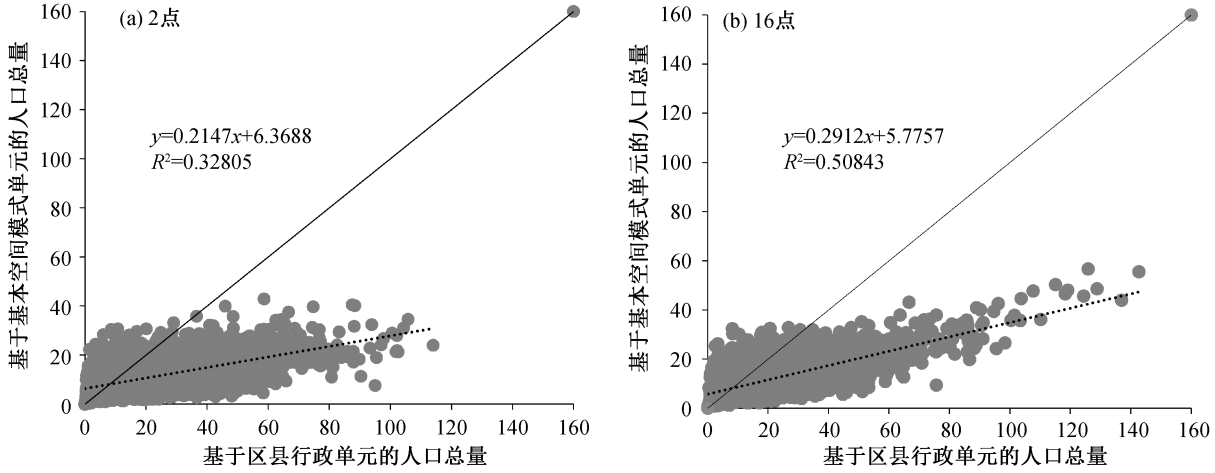

图 5 不同估测方法人口总量对比

Fig. 5 Comparison of population estimation results by two methods

表 6 配对样本 T 检验统计结果

Table 6 Paired-sample T test statistics results

时段	统计量	均值	*N*	标准差	均值标准误差
2 点	基于基本空间模式单元的人口总量	47825.28	1896	39136.33	898.80
	基于区县行政单元的人口总量	101833.42	1896	115922.19	2662.24
	基于热力图数据的常住人口数	4099.26	1896	945.17	21.71
16 点	基于基本空间模式单元的人口总量	86274.09	1896	90121.73	2069.71
	基于区县行政单元的人口总量	181507.58	1896	241893.88	5555.28
	基于热力图数据的常住人口数	6247.36	1896	3802.59	87.33

时段	统计量	成对差分					*t*	df	Sig.(双侧)
		均值	标准差	均值标准误差	差分的 95%置信区间				
					下限	上限			
2 点	基于基本空间模式单元的人口总量	43726.02	38599.52	886.47	41987.46	45464.57	49.33	1895	0.00
	基于区县行政单元的人口总量	97734.15	115229.05	2646.32	92544.14	102924.17	36.93	1895	0.00
16 点	基于基本空间模式单元的人口总量	95233.49	162090.99	3722.54	87932.78	102534.20	25.58	1895	0.00
	基于区县行政单元的人口总量	175260.23	238915.14	5486.87	164499.29	186021.16	31.94	1895	0.00

说明：表格双实线以上是基本的样本统计量，以下是成对差分后的样本检验结果。

行政单元的人口总量均与基于热力图数据的估测结果有显著性差异(Sig.=0.00)。与基于区县得到的人口总量相比，基于1 km网格的估测结果具有更小的均值标准误差，即基于基本空间模式单元的人口总量估测结果更具稳定性。

2.4 人口属性识别结果

根据调研小区空间模式的划分标准，将北京市主城区(六环路以内)按照1 km格网空间进行划分，最终得到10种空间模式，如图6所示。可以看出，城市的空间模式呈现“中心对称”和“线状分布”的特征。各种空间类型围绕着城市中心，呈现中心对称、放射状的特点，同时类别8和11明显地沿地铁线路分布，从城市中心向外延伸，说明目的地可达性(与城市中心的距离)、交通设施可达性(与最近邻地铁站的距离)对城市的模式空间有显著影响。由于调研小区类型的有限性，部分城市空间未划分出特定的空间模式。

本文通过构建朴素贝叶斯分类器，对各个空间模式类别中居民的属性特征进行预测。从调研数据中随机抽取2/3为训练集，其余1/3为测试集。通过训练集，对居民部分属性特征(性别和年龄)、通勤出行特征(周均工作时长、周均通勤次数和平均通勤时长)组合的先验概率进行统计，对测试集属性特征(个人月收入、有无子女和就业类型)出现的后验概率进行计算，从而选择最大后验概率所对应的属性类别作为扩充结果，预测结果的准确度如表7所示。

从表7可以看出，有无子女的预测准确度最高，平均值为73.65%，个人月收入次之，平均预测准确度为29.90%，就业类型的平均准确度最低，仅为23.98%。有无子女作为二分类选项，在调研结果中有更充分的样本集合，故预测效果较好，而个人月收入与就业类型均为四分类选项，样本代表性较差，故平均的预测水平较低。同时，由于调研数据中就业类型项的比例分布不均(基础产业∶商业服务业∶公共服务业∶建筑与制造业=1438∶971∶430∶61)，故出现较多的无法预测项。

3 讨论与结论

本文基于空间模式单元，运用多源数据融合的方法，对人口空间布局进行分时段变化估测，进而刻画人口分布的时空间动态特征，同时运用机器学习的方法，实现人口属性的匹配识别。在此基础上，分别对人口总量估测结果和人口属性识别结果进行

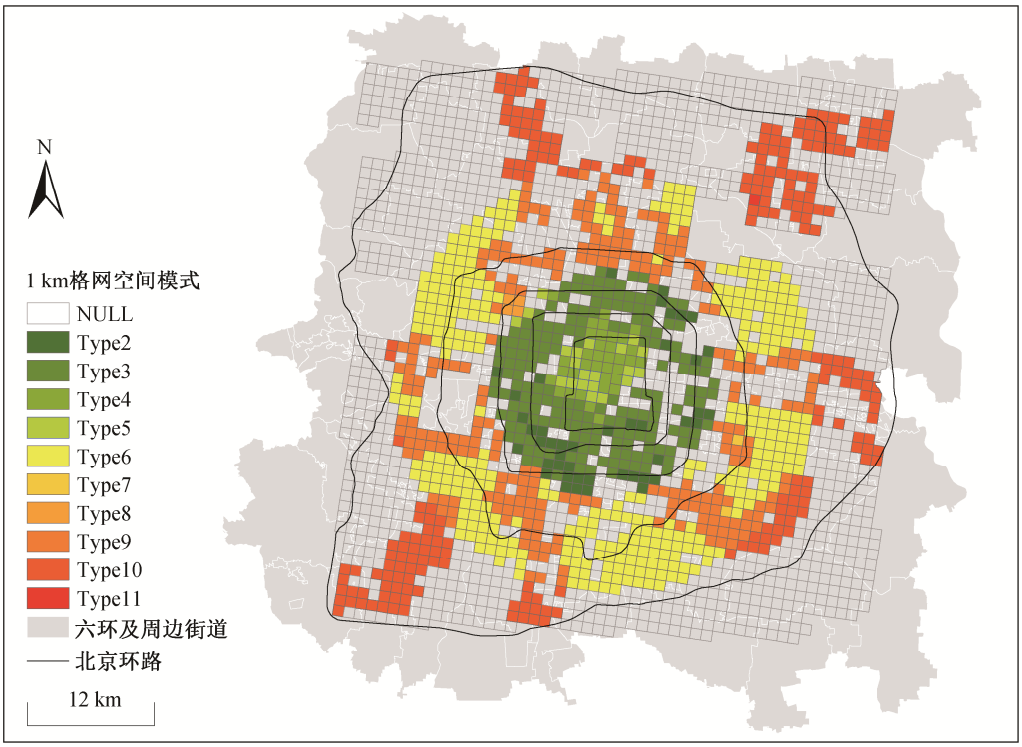

图6　北京市空间模式划分

Fig. 6　Spatial pattern unit division of Beijing

表 7 人口属性预测结果

Table 7 Socioeconomic attribute prediction results

类别	样本量	预测准确度/%		
		个人月收入 I	有无子女 C	就业类型 E
Type 1	58	–	79.00	–
Type 2	102	–	70.59	–
Type 3	184	21.15	71.31	–
Type 4	168	27.14	76.25	–
Type 5	274	28.46	73.74	–
Type 6	364	25.45	78.93	–
Type 7	199	31.51	68.03	–
Type 8	156	29.99	80.58	–
Type 9	983	31.11	74.98	31.61
Type 10	319	33.49	67.92	20.75
Type 11	292	40.82	68.87	19.59
合计	3099	–	–	–
平均	–	29.90	73.65	23.98

验证, 为人口动态监测的测定方法提出建议。

以北京市 1 km空间单元为研究对象, 通过手机信令数据与人口统计数据、问卷调查数据及热力图数据的融合, 对全时段的人口驻留总量进行估测, 实现人口社会经济属性的识别。结果表明: 1) 手机信令用户的驻留总量与基于基本空间单元估测后的人口驻留总量之间存在着明显的空间差异性; 2) 与基于细粒度 1 km空间单元的估测结果相比, 基于区县的传统估测结果, 存在明显的高估现象; 3) 针对大数据属性缺失问题, 可以通过机器学习的方法, 从大样本调研数据中进行学习, 实现预测和补充。

因此, 基于手机信令大数据的应用, 我们为人口动态监测以及人口属性识别的测定方法提出以下建议。

1) 通过空间尺度的细化, 实现人口总量估测方法的优化。在人口总量的估测过程中, 扩样率是指征用户代表性的关键指标。传统的估测方法以区县的行政区划为范围进行扩样率的计算(即假设相同区县范围的用户具有相似的交通出行特征), 具有较大的不确定性和空间变异性。随着人口空间化理论的发展, 人口分布空间可以细化至区县级尺度以下。构建 1 km空间模式单元是值得探讨的研究方向, 可以有效地减小空间尺度, 修正假设偏差, 实现任意时空间范围内人口驻留总量的估测。

2) 通过手机信令大数据与问卷调研数据的结合, 实现人口属性的识别。手机信令等大数据记录了用户大量的时空间信息, 问卷调研数据则包含丰富的社会经济属性信息, 二者的有效结合为人口属性识别研究提供新的思路。应用机器学习的方法, 构建朴素贝叶斯分类器, 可以为二者的结合提供理论基础。通过对调研数据的属性信息和出行特征进行学习和预测, 实现对社会经济属性(家庭结构、收入和职业)的识别。

本文方法也存在不足之处。由于多源数据统计口径不同, 时空间尺度差异较大, 在数据估测计算中存在一定的偏差。在人口动态分布估测中, 根据人口统计数据(2015 年)和手机信令数据(2017 年)进行扩样率的计算, 由于 2017 年手机信令数据用户量比 2015 年有所增加, 可能在一定程度上造成扩样率和估测结果的高估。在人口属性识别过程中, 由于个别调研小区采样数量有限以及样本结构分布并不完全均匀, 导致“个人月收入”和“就业类型”的估测准确度相对较差。在后续的研究中, 可以通过统一数据采集时间、完善样本数量和改善样本结构来进行优化。

人口时空分布特征的动态监测是城市理论研究的重要基础。本文应用大数据在人口样本量与高分辨率时空间识别的优势, 提出基于空间模式单元的人口动态分布估测和人口属性识别技术, 对于开展城市人口分布及其演化的理论研究以及城乡规划设施、商业网点的布局优化具有实践意义。

参考文献

[1] 梁亚婷. 基于遥感和 GIS 的城市人口时空分布研究[D]. 上海: 上海师范大学, 2015

[2] Sun J B, Yuan J, Wang Y, et al. Exploring space–time structure of human mobility in urban space. Physica A: Statistical Mechanics and Its Applications, 2011, 390(5): 929–942

[3] 朱传耿, 顾朝林, 马荣华, 等. 中国流动人口的影响要素与空间分布. 地理学报, 2001, 56(5): 549–560

[4] 陆化普. 大城市交通问题的症结与出路. 城市发展研究, 1997(5): 16–20

[5] 陈楠. 基于 GIS 的人口时空分布特征研究[D]. 青岛: 山东科技大学, 2005

[6] 肖宝仲. 基于信令分析的智慧城市人流监控管理研究[D]. 北京: 北京化工大学, 2013

[7] 李吉塘, 周春山, 杨高. 珠海外来人口分布的时空演变特征研究. 城市学刊, 2018, 39(4): 38–42

[8] 杨振, 雷军. 1982—2010 年乌鲁木齐市主城区人口

时空分布特征及模拟. 中国科学院大学学报, 2018, 35(4): 506–514

[9] Bracken I, Martin D. The generation of spatial population distributions from census centroid data. Environment and Planning A: Economy and Space, 1989, 21 (4): 537–543

[10] 冯甜甜. 基于高分辨率遥感数据的城市精细尺度人口估算研究[D]. 武汉: 武汉大学, 2010

[11] 吴健生, 许多, 谢舞丹, 等. 基于遥感影像的中尺度人口统计数据空间化——以京津冀地区为例. 北京大学学报(自然科学版), 2015, 51(4): 707–717

[12] 刘云舒, 赵鹏军, 梁进社. 基于位置服务数据的城市活力研究——以北京市六环内区域为例. 地域研究与开发, 2018, 37(6): 64–69

[13] 赵鹏军, 曹毓书. 基于多源 LBS 数据的职住平衡对比研究——以北京城区为例. 北京大学学报(自然科学版), 2018, 54(6): 1290–1302

[14] 王德, 朱查松, 谢栋灿. 上海市居民就业地迁移研究——基于手机信令数据的分析. 中国人口科学, 2016(1): 80–89

[15] 钟炜菁, 王德, 谢栋灿, 等. 上海市人口分布与空间活动的动态特征研究——基于手机信令数据的探索. 地理研究, 2017, 36(5): 972–984

[16] 廖顺宝, 孙九林. 基于 GIS 的青藏高原人口统计数据空间化. 地理学报, 2003, 58(1): 25–33

[17] Stevens F R, Gaughan A E, Linard C, et al. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLoS ONE, 2015, 10(2): e0107042

[18] 钮心毅, 王垚, 丁亮. 利用手机信令数据测度城镇体系的等级结构. 规划师, 2017, 33(1): 50–56

[19] 王德, 钟炜菁, 谢栋灿, 等. 手机信令数据在城市建成环境评价中的应用——以上海市宝山区为例. 城市规划学刊, 2015(5): 82–90

[20] Niu Xinyi, Ding Liang, Song Xiaodong. Understaning urban spatial structure of shanghai central city based on mobile phone data. China City Planning Review, 2015, 24(3): 15–23

[21] 王德, 王灿, 谢栋灿, 等. 基于手机信令数据的上海市不同等级商业中心商圈的比较——以南京东路、五角场、鞍山路为例. 城市规划学刊, 2015(3): 50–60

[22] 张天然. 基于手机信令数据的上海市域职住空间分析. 城市交通, 2016, 14(01): 15–23

[23] Wang H, Calabrese F, Lorenzo G D, et al. Transportation mode inference from anonymized and aggregated mobile phone call detail records // 13th International IEEE Conference on Intelligent Transortation Systems. Funchal, 2010: 318–323

[24] Aguilera V, Allio S, Benezech V, et al. Using cell phone data to measure quality of service and passenger flows of Paris transit system. Transportation Research Part C: Emerging Technologies, 2014, 43: 198–211

[25] 冉斌. 手机数据在交通调查和交通规划中的应用. 城市交通, 2013, 11(1): 72–81

[26] 甄峰, 王波. “大数据”热潮下人文地理学研究的再思考. 地理研究, 2015, 34(5): 803–811

[27] Pozzi F, Small C. Modeling the distribution of human population with night-time satellite imagery and gridded population of the world [C/OL] // Pecora 15/Land Satellite Information IV/ISPRS Commission I/FIEOS 2002 Conference Proceedings. [2019–05–04]. https://pdfs.semanticscholar.org/035a/a66794b9958f703e6f620f5c4775adf86285.pdf

[28] 王雪梅, 李新, 马明国. 基于遥感和 GIS 的人口数据空间化研究进展及案例分析. 遥感技术与应用, 2004, 19(5): 320–327

[29] Zhao Pengjun. Car use, commuting and urban form in a rapidly growing city: evidence from Beijing. Transportation Planning and Technology, 2011, 34(6): 509–527

[30] Zhao Pengjun. The impact of the built environment on individual workers’ commuting behavior in Beijing. International Journal of Sustainable Transportation, 2013, 7(5): 389–415

[31] Zhao Pengjun. The impact of the built environment on bicycle commuting: evidence from Beijing. Urban Studies, 2014, 51(5): 1019–1037

[32] Ewing R, Cervero R. Travel and the built environ-ent. Journal of the American Planning Association, 2010, 76(3): 265–294

[33] Li Shengxiao, Zhao Pengjun. Exploring car ownership and car use in neighborhoods near metro stations in Beijing: does the neighborhood built environment matter?. Transportation Research Part D: Transport and Environment, 2017, 56: 1–17

[34] 赵鹏军. 土地集约利用对可持续城市交通的作用:基于国际文献理论分析. 城市发展研究, 2018, 25(9): 108–116

[35] Schwanen T, Deileman F M, Dijst M. The impact of metropolitan structure on commute behavior in the

Netherlands: a multilevel approach. Growth and Change, 2004, 35(3): 304–333

[36] Cervero R, Kockelman K. Travel demand and the 3Ds: density, diversity, and design. Transportation Research Part D: Transport and Environment, 1997, 2(3): 199–219

[37] 塔娜, 柴彦威, 关美宝. 建成环境对北京市郊区居民工作日汽车出行的影响. 地理学报, 2015, 70(10): 1675–1685

[38] 郑红玉, 黄建洪, 卓跃飞, 等. 土地混合利用测度研究进展. 中国土地科学, 2019, 33(3): 95–104

[39] Zhang Mengzhu, Zhao Pengjun. The impact of land-use mix on residents' travel energy consumption: new evidence from Beijing. Transportation Research Part D: Transport and Environment, 2017, 57: 224–236

[40] Cervero R. Built environments and mode choice: toward a normative framework. Transportation Research Part D: Transport and Environment, 2002, 7(4): 265–284

[41] Limtanakool N, Dijst M, Schwanen T. The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium- and longer-distance trips. Journal of Transport Geography, 2006, 14(5): 327–341

[42] Handy S. Regional versus local accessibility: implications for nonwork travel [R]. University of California Transportation Center, Working Papers, 1993: 58–66

[43] Owen N, Humpel N, Leslie E, et al. Understanding environmental influences on walking: review and research agenda. American Journal of Preventive Medicine, 2004, 27(1): 67–76

[44] 王丰龙, 王冬根. 北京市居民汽车使用的特征及其影响因素. 地理学报, 2014, 69(6): 771–781

[45] Olszewski P, Wibowo S S. Using equivalent walking distance to assess pedestrian accessibility to transit stations in Singapore. Transportation Research Record, 2005: 38–45

[46] 郑思齐, 丁文捷, 陆化普. 住房、交通与城市空间规划. 城市问题, 2009(1): 29–34

[47] Georggi N L, Pendyala R M. Analysis of long-distance travel behavior of the elderly and low income [C/OL] // E-C026: Personal Travel: The Long and Short of It: Conference Proceedings 2001 [2019–05–08]. http://onlinepubs.trb.org/onlinepubs/circulars/ec026/02_georggi.pdf

[48] 中华人民共和国住房和城乡建设部. 2017年城乡建设统计年鉴[EB/OL]. (2019–01–24) [2019–05–08]. http://www.mohurd.gov.cn/xytj/tjzljsxytjgb/jstjnj/

[49] 北京市统计局. 北京市2016年国民经济和社会发展统计公报[EB/OL]. (2017–02–25) [2019–05–08]. http://tjj.beijing.gov.cn/zxfbu/202002/t20200216_1634839.html

[50] 北京交通发展研究院. 2017 北京市交通发展年度报告[EB/OL]. (2017) [2019–05–08]. http://www.bjtrc.org.cn/List/index/cid/7.html

[51] 董黎明. 城市化与住房问题. 国外城市规划, 2001(3): 21–24

[52] 吴海瑾. 城市化进程中流动人口的住房保障问题研究——兼谈推行公共租赁住房制度. 城市发展研究, 2009, 16(12): 82–85

[53] Arvin M B, Pradhan R P, Norman N R. Transportation intensity, urbanization, economic growth, and CO_2 emissions in the G-20 countries. Utilities Policy, 2015, 35: 50–66