

利用公交刷卡数据分析北京职住关系和通勤出行

龙瀛, 张宇, 崔承印

(北京市城市规划设计研究院, 北京 100045)

摘要: 基于位置服务 (Location Based Service, LBS) 技术为研究城市系统的时空动态规律提供了新的视角, 已往多基于移动通讯 (GSM)、全球定位系统 (GPS)、社会化网络 (SNS) 和无线宽带热点 (Wi-Fi) 数据开展研究, 但少有研究利用公交IC卡刷卡数据进行城市系统分析。普遍存在的LBS数据虽然具有丰富的时间和空间信息, 但缺乏社会维度信息, 使其应用范围受到一定限制。本文基于2008年北京市连续一周的公交IC卡 (Smart Card Data, SCD) 刷卡数据, 结合2005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, 并将识别结果在公交站点和交通分析小区 (TAZ) 尺度上汇总: ① 将识别的通勤出行分别从通勤时间和距离角度, 与居民出行调查数据和其他已有北京相关研究进行对比, 显示较好的吻合性; ② 对来自3大典型居住区和去往6大典型办公区的通勤出行进行可视化并对比分析; ③ 对全市基于公交的通勤出行进行可视化, 并识别主要交通流方向。本研究初步提出了从传统的居民出行调查和城市GIS数据建立规则, 用于SCD数据挖掘的方法, 具有较好的可靠性。

关键词: 公交IC卡刷卡数据; 职住关系; 通勤时间和距离; 空间错位; 北京

1 引言

本文旨在利用海量公交IC卡 (Smart Card Data, SCD) 的刷卡数据, 识别北京市的居住—就业 (Housing-jobs) 区位分布二者之间的关系, 并对公交通勤出行进行识别、评价和可视化。以往的职住分析主要基于问卷调查数据开展。目前, 随着基于位置服务 (Location Based Services, LBS), 如移动通讯 (GSM)、全球定位系统 (GPS)、社会化网络 (SNS) 和无线宽带热点等技术的进步和广泛应用, 大规模、高质量的个体时空数据的获取正不断成为可能, 这些遍布 (Ubiquitous) 的个人贡献 (Volunteered) 的数据对描述和理解城市空间结构提供了新的渠道^[1-2], 例如, Goodchild^[3]突破了传统的光学传感器思想, 较早地提出将人作为传感器 (Human as Sensors) 以获取空间和社会信息的设想。具体地, GPS数据应用方面, Newhaus^[4]使用定位数据记录了城市居民的出行日志, 并进行可视化; Jiang等^[5]利用瑞典出租车的GPS轨迹, 分析人们的出行特征, 进而分析城市空间结构对其影响; Liu等^[6]通过出租车司机日常出行的轨迹, 揭示出租车司机的出行特征。GSM方面^①, Ratti等^[7]利用意大利米兰的手机位置信息数据来表达城市活动的强度和时空分布特征; Calabrese和Ratti^[8]利用罗马的手机网络数据对城市活动进行实时监测; Reades等^[9]也利用手机网络数据来表征城市活动。SNS方面, Milton^[11]利用Twitter数据来制作英国降雪深度的地图; Jones等^[12]利用网络博客 (blog) 进行经济地理分析具有较强优势。Wi-Fi方面, Rekimoto等^[13]通过Wi-Fi发射塔信号数据研究装置持有者所在位置及其室内外活动; Torrens^[14]在美国的盐湖城分析了Wi-Fi热点的地理特征。总体上, 从Shavol^[15]提出人类感知 (Human Sensing) 以来, LBS技术已经被越来越多地应用于城市研究中, 但是还在起步。在更多的

收稿日期: 2011-12-01; 修订日期: 2012-04-15

基金项目: 国家自然科学基金项目(51078213) [Foundation: National Natural Science Foundation of China, No.51078213]

作者简介: 龙瀛 (1980-), 男, 博士, 高级工程师, 中国地理学会会员 (S110007674M), 主要研究方向为规划支持系统和城市系统微观模拟。E-mail: longying1980@gmail.com

① 更多请参考综述性论文Steenbruggen等^[9]

城市空间结构研究,多采用典型调查的方法获取个体数据^[16],获取大容量微观样本还存在一定的障碍^[17]。

SCD作为一种大规模的具有地理标识(Geo-tagged)和时间标签的数据,可以用于城市空间结构的研究与分析。以往这类研究主要基于居民出行调查数据开展,但是这种调查成本高,每隔几年才能开展一次,而SCD具有连续性好、覆盖面广、信息全面且动态更新等优点,因此在城市研究中具有较好的可应用性。在国外,有研究建议使用SCD进行公共交通的规划设计决策^[18];Joh和Hwang^[19]利用首尔大都市区400万人次的1000万次SCD,分析了公交卡持有者的出行轨迹与都市区的土地利用特征;Jang^[20]利用首尔100万次SCD,对公交出行时间和换乘信息进行估计;Roth等^[21]基于伦敦实时的Oyster卡数据库,得到伦敦地铁乘客移动特征,用来分析多中心的城市结构。在国内,截至2007年,有超过100个城市启动了公交IC卡收费系统^[22],利用SCD已有部分研究,如董晓晶等^[23]开发了一套北京SCD处理工具,为公交线网优化和交通管理提供依据;彭晗等^[24]利用长春市的公交IC卡数据,辅助支持城市公交换乘枢纽选址;杨智伟等^[25]基于大连SCD进行客流预测。总体上,这些研究主要侧重于利用SCD支持交通系统的规划设计和交通系统特征,少有研究利用SCD进行城市结构分析,如职住关系、通勤出行时空分布等。

本文拟基于2008年北京市连续一周的SCD分析北京的职住关系和通勤出行特征^②。北京自2006年4月1日起,全面推行公交IC卡^[26],截至2007年4月2日,全市公交IC卡销售量已接近1000万张,目前乘用公交车的乘客刷卡比例已超过90%^③。每张公交IC卡都对应一个唯一的ID,是一种不记名卡,购买时不需要提供个人信息^④。2008年北京的公交出行比例已经超过30%^[27],因此公交出行数据可以很好的反映城市社会空间的变化情况。本文主要探讨如何对SCD进行预处理,并基于传统的居民出行调查和城市GIS数据建立SCD数据挖掘的规则,并将结果与已有研究进行对比。

2 数据

2.1 公交线路、公交站点和交通分析小区(TAZ)

公交线路和公交站点GIS图层用于将SCD进行空间化(Geocoding)。北京市^⑤2008年4月共有600多条公交线路(按上下行计算共计1287条,其中一票制566条,分段计价线路721条,图1a)。共有约3.7万个公交站点^⑥(图1b),为分析方便,北京市域划分为1118个交通分析小区(Traffic Analysis Zone, TAZ)(图1c)。

2.2 土地使用现状图

本研究引入地块(Parcel)级别的土地使用现状图(2007年,图1d)作为识别居住地和就业地的依据之一,该图层包含每个地块的土地使用功能和建筑面积属性。本文假设居住地来自居住类地块,而就业地则来自商业、公共设施和工业等类型地块。在总计133503个地块中,有29112个居住地块和57285个就业地块^[28]。该图层将用于计算每个公交站点作为居住或工作地的概率,以处理使用一周数据得到的重复性结果(具体见“3.3 基于一周SCD识别职住地”)。

② 暂不包括轨道交通刷卡数据。

③ 北京市公交IC卡由北京市政交通一卡通有限公司管理运作,官方网站为www.bjsuperpass.com,在该网站上可以输入卡片正面的ID查询数周内该持卡人的公交出行记录。

④ 本研究最后在中观和宏观层面进行分析,同时考虑到北京公交IC卡为匿名卡,因此本研究不涉及持卡人的个人隐私问题。

⑤ 如无特殊说明,北京市指市域范围,共16410 km²。

⑥ 该数字为所有公交线路站点数的总和,而不是站台数量。

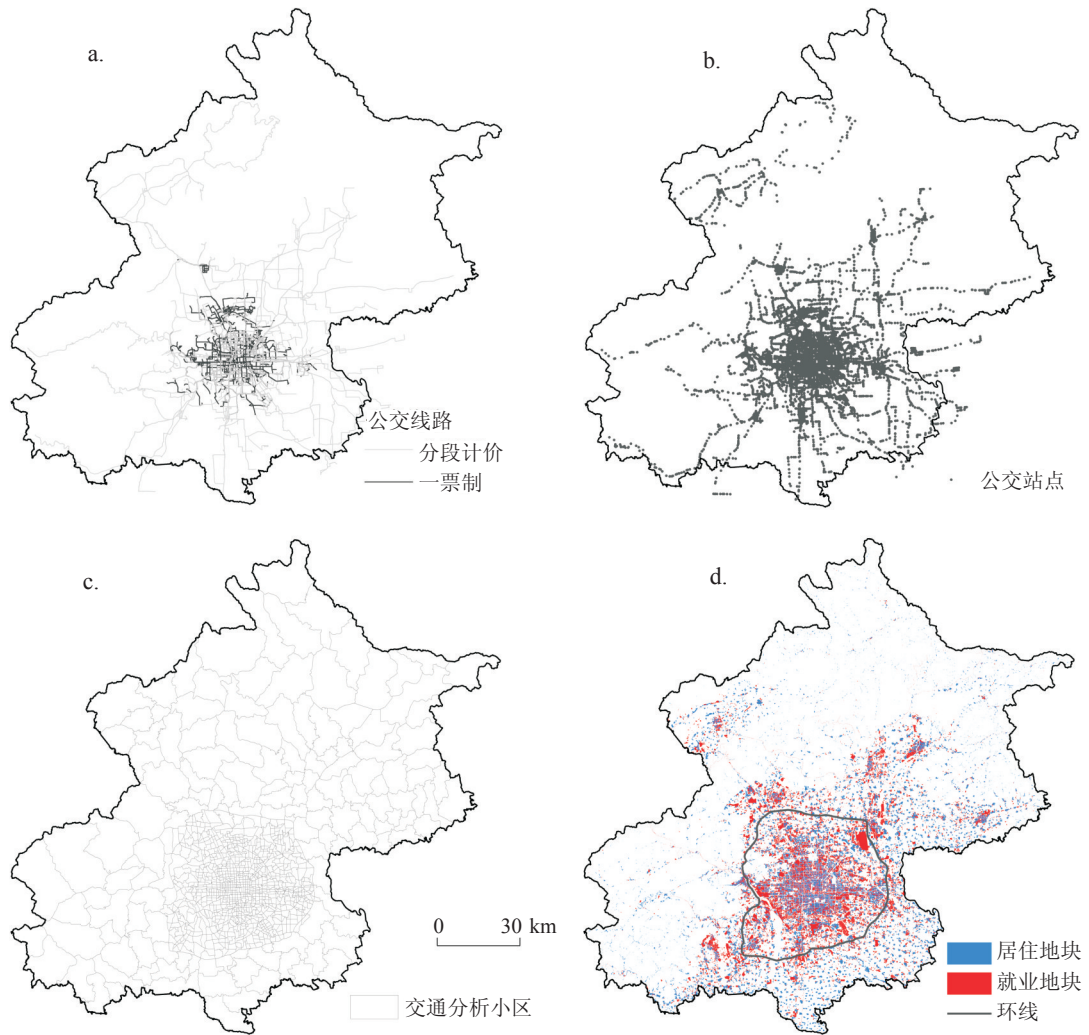


图1 北京市公交线路 (a), 公交站点 (b), 交通分析小区 (TAZ) (c) 和土地使用现状图 (d)
 Fig. 1 The bus lines (a), bus stops (b), traffic analysis zones (TAZs) (c), and land use pattern (d) of the Beijing Metropolitan Area (BMA)

2.3 公交IC卡刷卡数据 (SCD)

本文所利用的SCD为2008年4月北京市完整一周(周一至周日)的刷卡记录, 鉴于非技术原因不包括祥龙公司的运营线路和轨道交通数据。记录涵盖的基本信息包括: 每个持卡人刷卡的时间和地点(其中地点以线路号和站点号表示)、卡类型(普通卡、学生卡或工作人员卡等)、交易序号(表示持卡人累计刷卡次数)、司机编号和车辆编号等。该一周数据共有77976010次刷卡记录, 对应8549072张一卡通, 因此可估算每卡日均使用约1.30次。

北京公交线路按照计价方式分为两种, ①短距离的一票制线路, 主要位于中心城地区, 每次应收票价1元, 刷卡四折优惠, 实收0.4元。对于此类刷卡记录, SCD仅记录上车刷卡时间, 上车刷卡站点虽有记录但位置信息不准确^⑦。②分段计价线路, 这一类线路的路线一般较长, 一段或两段位于五环外, 计价方式为分段计价, 一般为1元起价(刷卡

⑦ 一般为线路首站。。

实收 0.4 元), 根据乘坐距离长短, 10 km 后每增加 5 或 10 km 增加 1 元。这类线路的 SCD 则记录了持卡人的完整刷卡时空信息。一票制线路只有部分出行信息, 难以识别职住地, 但一票制线路不会影响两次分段计价线路之间乘坐一票制线路这种情况的识别结果。

将 SCD 从时间和空间两个维度展示。对于时间维度, 一周内每天的总刷卡次数如图 2a 所示, 工作日 (周二) 和周末 (周六) 各时段刷卡次数如图 2b 所示 (以上车时间计), 二者差别较明显, 平日早晚高峰突出, 周末早晚高峰与中间时段的平峰刷卡量较为接近。多数公交出行分布在 6:00-22:00。对于空间维度, 中心城的公交出行密度显著高于周边地区 (图 2c)。

2.4 居民出行调查数据

北京市在 1986 年、2000 年、2005 年和 2010 年分别进行了 4 次居民出行调查^⑧。本文利用 2005 年的居民出行调查数据 (以下简称“2005 年调查”) 用于支持 SCD 的数据挖掘。该调查涵盖了北京市 18 个区 (县) 共 1118 个 TAZ 的基础地理数据 (和图 2 中的 TAZ 一致)^[29]。调查规模为 74839 户, 被调查人数为 191835 人, 抽样率为 1.36%。调查采用出行日志形式, 对于每次出行, 调查数据包括出发时间/地点, 到达时间/地点, 出行目的和方式, 还有其他诸如出行距离、目的地类型和公交出行线路编号等信息。调查中也包含住户和个人信息。住户信息包括家庭规模、户籍情况、居住地和家庭收入等, 个人信息包括性别、年龄、工作类型和地点、是否有驾照或者公交月卡等。

3 方法

3.1 数据预处理和数据模型

SCD 存储了刷卡的原始数据, 要基于该数据分析职住关系, 需要进行必要的预处理。首先利用公交站点 GIS 图层, 基于刷卡记录对应的线路和站点信息, 对 SCD 进行空间化; 然后将每个持卡人连续一周的刷卡记录进行合并, 得到每个持卡人一周的公交出行日志, 记录了所有公交出行的起始时间、起始地点和卡类型等信息。

⑧ 本文成稿时, 还没有获得 2010 版调查数据。

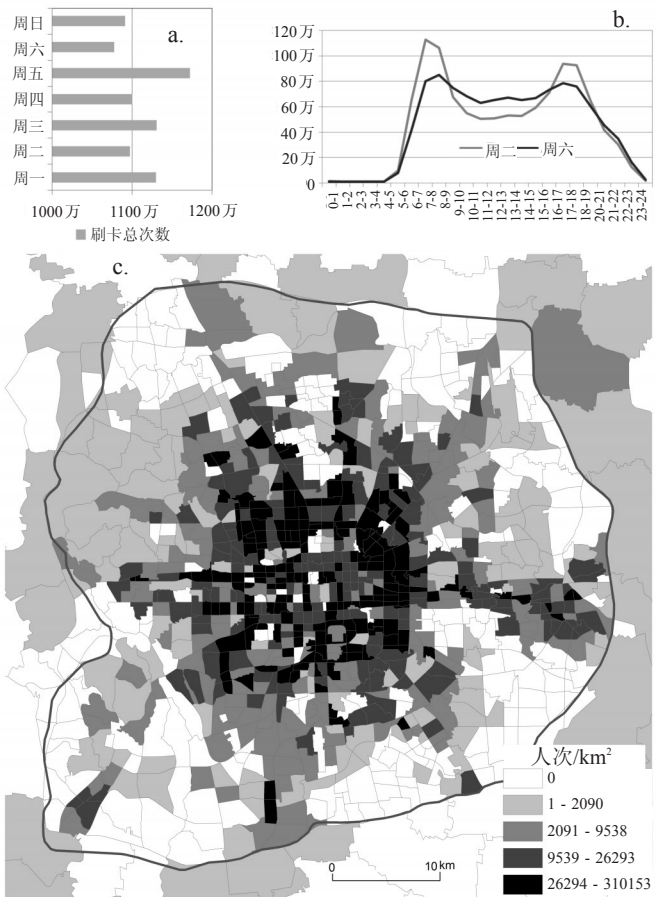


图2 北京市SCD的时空分布特征:(a)一周内每日刷卡总次数;(b)周二和周六不同出发时间的刷卡次数;(c)六环内各TAZ的出行密度

Fig. 2 The spatiotemporal characteristics of bus trips in the SCD. (a) The total count of bus trips in each day of a week; (b) The trip count in various departure hours of Tuesday and Saturday; (c) Trip count densities for each TAZ within the 6th Ring Road of the BMA

本文构建了两种数据模型用于表达预处理后的SCD: 出行 (TRIP) 数据模型和“地点—时间—时长 (Position-Time-Duration, PTD)”数据模型。对于TRIP数据模型, SCD中的一次出行, 代表了持卡人一次上车和下车的乘车过程, 一次出行可以表达为出发地点 (OP)、出发时间 (OT)、到达地点 (DP) 和到达时间 (DT) 的集合, $TRIP = \{OP, OT, DP, DT\}$ 。而PTD数据模型可由TRIP数据模型转换得到, $PTD = \{P, t, D\}$, 其中P代表一个公交站点, t代表在地点P的开始时间, D代表地点P的持续时间。与TRIP相比, PTD数据模型更容易与时间地理学结合。关于如何将TRIP转换为PTD, 这里以一个例子进行说明, 假设一位持卡人离开居住地 (公交站ID为H0) 时间为7:00, 到达就业地 (公交站ID为J0) 时间为8:00, 一天工作结束后, 持卡人乘公交车于17:00离开就业地J0并于18:00到达居住地H0。TRIP数据模型为 $\{H0, 7:00, J0, 8:00\}$ 和 $\{J0, 17:00, H0, 18:00\}$, 转换的PTD数据模型则为 $\{H0, 18:00 (-1), 13\text{ h}\}$ 和 $\{J0, 8:00, 9\text{ h}\}$, 分别表示为以家庭为基点的活动和以工作为基点的活动。以家庭为基点的活动从前一天的18:00开始, 持续13个小时, 到次日7:00结束, 而就业活动从8:00开始持续9 h, 到17:00结束。

3.2 基于一日SCD识别职住地

使用PTD数据模型识别每位持卡人的职住地, 首先利用一日SCD识别, 最后综合每日识别结果进行综合, 给出最终的职住地。居住地和就业地的识别过程是相互独立的。对于基于一日SCD的居住地识别, 假定首次出行的出发站点为持卡者的居住地。在2005年调查中, 99.5%居民的首次出行的出发地点是居住地, 这和本文设定居住地的规则是一致的。需要指出, 对于首次出行乘坐一票制公交车的持卡人, 是无法识别居住地的。

为了识别就业地, 需要识别乘坐公交车的就业出行。假设全职工作是一天中时间最长的活动, 如果满足下面的条件, 第 k 个地点的 P_k 地可以视为持卡人的就业地。条件一: 卡片类型不是学生卡; 条件二: $D_k \geq 360$; 条件三: $k < 1$ 。

如果持卡人在某地停留的时间超过360 min (6 h) (首个地点即居住地除外), 可认为该地是持卡人的就业地。根据2005年调查, 27550个被调查者 (其中210人次中午回家休息, 不计入) 的平均工作时间为9 h19 min (标准差为1 h41min), 96%的被调查者每天工作时间超过6 h, 据此以6 h为基准识别就业地是可行的。

需要强调的是, 鉴于轨道交通数据的缺失, 为了将其对分析结果的影响降到最低, 采取了如下方法: 在持卡人每次刷卡的纪录中, 都记录了每次对应的交易序号 (顺序增加), 因此如果持卡人某两次出行之间乘坐了轨道交通, 则从交易序号可以识别出来 (如两次公交刷卡之间的序号相差不是1, 而是2或者3等)。利用该信息, 在识别居住地和就业地的过程中, 从技术上进行处理, 可以保证目前的识别结果去掉了轨道交通的影响。但不足之处在于, 轨道交通数据的缺失, 会造成识别的有效结果数量下降。

3.3 基于一周SCD识别职住地

持卡人在一周7天的每日内的识别结果差别较大, 为了利用一周数据确定最终的居住地, 本文使用基于规则的方法 (Rule-based Approach)和决策树 (Decision Tree) 的方法对每日的结果进行综合 (图3), 该过程同时考虑了每个识别地的频率 (即一周内识别的次数)和空间分布。具体地, 如果只有一日有识别的居住地, 那么没有充分理由认为该站点为最终居住地。如果超过一个, 且所有的居住地相同, 认为该站点为持卡人最终的居住地。如果日居住地不同, 将这些站点根据之间的距离进行聚类, 如果两地距离小于某个阈值, 则可视这两个站点为一个, 即属于同一个集群 (Cluster)。本文将该阈值设为500 m, 大约等于两临近公交站的平均距离 ($231 \times 2 = 462\text{ m}$), Zhao等^[30]在其研究中也把阈值设为500 m。需要指出的是, 如果两个站点出现频率一样, 为了确定最佳的最终居住地, 引入“居住潜力 (Residential Potential)”和“就业潜力 (Job Potential)”概念, 分别表示一个站点属于居住或就业地的概率, 二者是基于土地使用现状数据计算得到的, 具体公式为:

$$p_h^k = \frac{S_h^k}{S_t^k}; p_j^k = \frac{S_j^k}{S_t^k} \quad (1)$$

式中： p_h^k 为站点 k 的居住潜力； p_j^k 为站点 k 的就业潜力，基于公交站点图层生成泰森(Thiessen)多边形； S_h^k 为地块 k 邻近居住地块的建筑面积； S_j^k 为地块 k 邻近就业地块的建筑面积； S_t^k 为地块 k 邻近所有地块的建筑面积；站点 k 的邻近地块是质心在其泰森多边形内的地块，最后将潜力指标进行归一化。

如果存在多个集群且每个集群内只有一个站点，无法确定最终的居住地。如果存在唯一的集群包含数量最多的站点(最大集群)，可认为该集群的最高频率站点为最终的居住地。其他情形详见图3。

同理，识别最终就业地的方法与最终居住地的识别基本相同，只需将决策树中的“居住”变为“就业”即可。需要指出的是，在确定最终居住和就业地的过程中，使用的是完整一周的数据，而非仅仅考虑工作日。

3.4 基于职住地识别通勤出行

基于识别的最终职住地，使用TRIP数据模型识别从居住地到就业地的通勤出行^⑨。通勤距离可以通过公交站点GIS图层生成的公交线路网络距离计算，同样也可以计算职住地之间的欧氏距离。通勤时间是指在居住地上车和在就业地下车之间的时长。对于一位持卡人，如果满足如下3个条件，①一天中首次出行的上车地点为居住地；②就业地出现在一日出行中；③居住地和就业地在同一天(在同一集群中的站点视为相同)，则可以成功识别通勤出行。需要指出的是，如果某持卡人一周内可识别的通勤出行超过一次，则一周内的通勤时间可能有所不同，则将平均通勤时间作为最终的通勤时间。

4 结果分析

使用两种工具进行数据挖掘和可视化：①由于原始SCD存储在MS SQL Server中，因此使用结构化查询语言(SQL)进行数据预处理和数据模型生成，以提高运算效率；②

⑨ 本文中的通勤时间和通勤距离都是单向的，即从居住地到就业地。

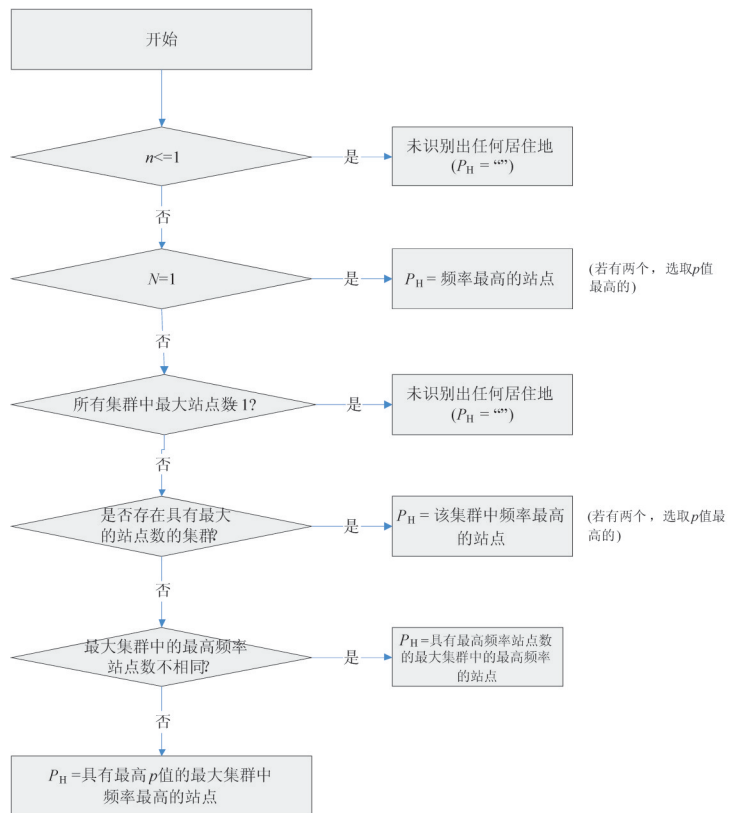


图3 基于一周SCD识别最终居住地的决策树

Fig. 3 The decision tree diagram for identifying housing places based on one-day results

注： n 代表一周内每位持卡人的识别居住地数量； N 是每个集群的数量； P_H 是居住地，决策树中未考虑识别居住地的顺序， p 是一个居住地的居住潜力。

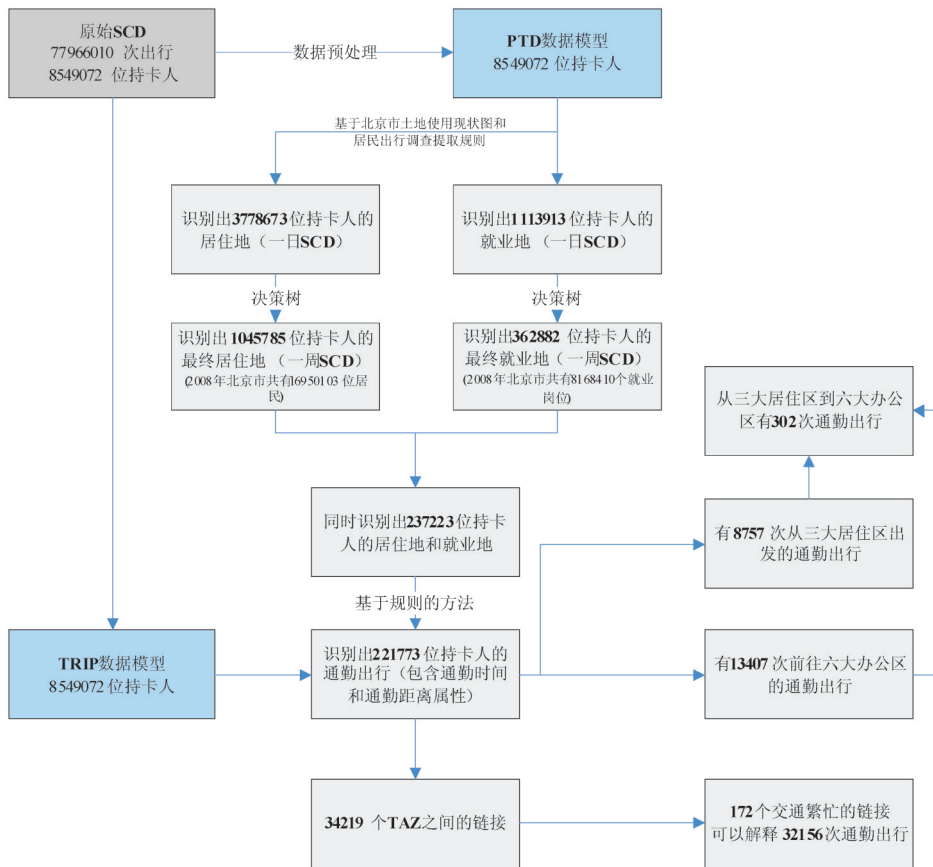


图4 基于SCD识别居住地、就业地和通勤出行的过程和结果概览

Fig. 4 The identification results for the whole paper

将处理过的SCD和2005年调查、GIS图层统一存储在ESRI ArcGIS的空间数据库(Personal Geodatabase)中,并基于ESRI Geoprocessing模块采用Python脚本语言进行开发,用于识别职住地和通勤出行,并对通勤出行进行可视化。

基于上述工具,分析过程和研究结果一览如图4所示。

4.1 基于公交站点和TAZ尺度的职住地识别

使用一周数据的职住地识别是基于每日的识别结果开展的。图5给出了具有不同天数的持卡人数量,显示随着识别天数的增加,识别出居住地或者就业地的持卡人数量显著下降。根据3.3部分的方法,只有大于等于2天具有识别结果的持卡人有望识别出最终的居住地或者就业地。

使用周数据最终分别识别出1045785位持卡人(全部8549072位持卡人的12.2%)的居住地和362882位持卡人(占全部持卡人的4.2%)的就业地。由于居住地和就业地的识别过程是相互独立的,因此共有237223位持卡人(占比2.8%)的居住地和就业地均被识别。最

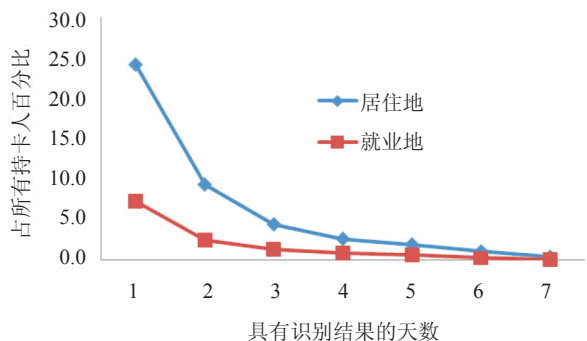


图5 具有不同天数识别结果的持卡人数量

Fig. 5 The cardholder count for various days with identification results

后将识别结果在 TAZ 尺度上汇总 (图 6)。

4.2 通勤出行识别

在 237223 位既有居住地和就业地的持卡人中, 识别了 221773 位持卡人的通勤出行。平均通勤时间是 36 min, 标准差是 24.2 min。平均通勤距离 (欧氏距离) 为 8.2 km, 标准差为 7.0 km。将识别的通勤出行根据居住地在 TAZ 尺度进行汇总, 得到不同 TAZ 的平均通勤时间和距离 (图 7)。同外围地区相比, 中心地区平均通勤时间和距离较短。通勤距离分布的圈层结构印证了北京市的单中心城市结构。

与 2005 年调查的对比情况见表 1。在 2005 年调查中, 6651 被调查者 (占识别结果的 3%) 乘坐公交车出行, 进行了如下 3 方面的对比分析: ① 2005 年调查显示, 平均通勤时间为 40.5 min (标准差为 23.1 min), 平均通勤距离为 8.4 km (标准差为 8.3 km), 与本文研究结果基本吻合; ② 对比通勤时间和距离的累积分布函数 (CDF, 图 8), 由于 2005 年调

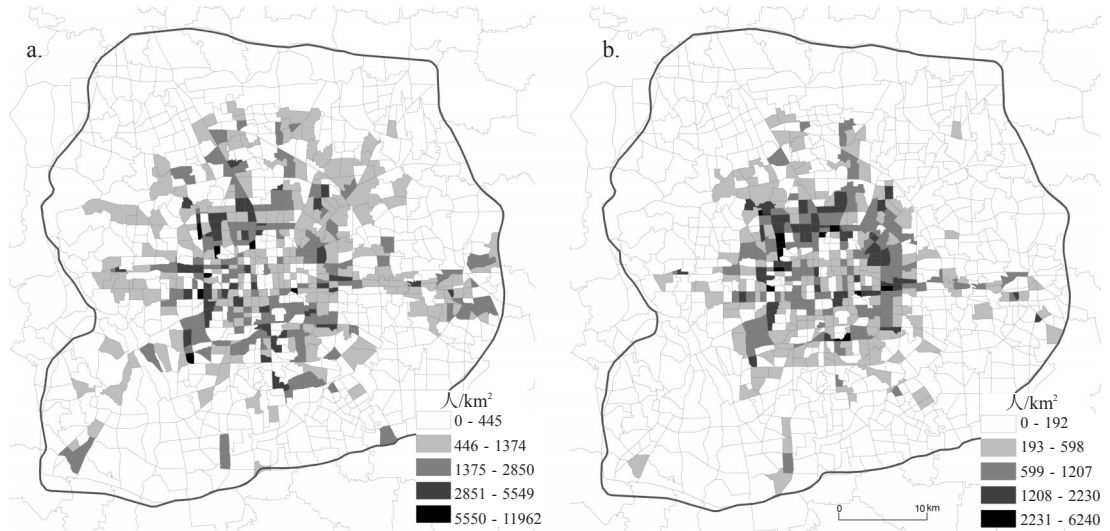


图 6 北京市中心区识别的居住 (a) 和就业 (b) 密度图

Fig. 6 The identified housing (a) and job (b) density maps (a and b) in the central area of the BMA

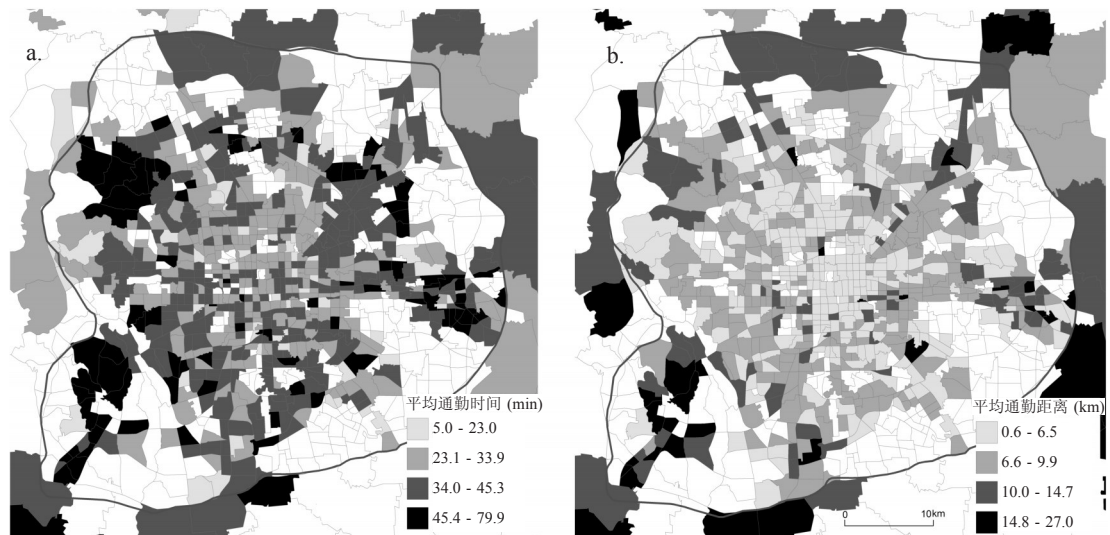


图 7 北京市中心区各 TAZ 的平均通勤时间 (a) 和距离 (b)

Fig. 7 The average trip time (a) and distance (b) of cardholders for each TAZ in the central area of the BMA

表1 对比本研究 and 2005年调查中的通勤出行的时间和距离(区县尺度)

Tab. 1 The comparison for commuting time and distance of commuting trips in our results and the 2005 survey

区域	本文结果			2005年调查			本文结果/2005调查		
	数量	时间 (min)	距离 (km)	数量	时间 (min)	距离 (km)	时间比率	距离比率	
中心城区	东城区	4179	35.1	6.5	317	37.7	5.8	0.93	1.12
	西城区	9145	33.7	7.1	467	35.2	6.3	0.96	1.13
	崇文区	3762	39.8	7.6	276	37.6	5.8	1.06	1.31
	宣武区	4377	36.6	8.2	432	40.3	6.9	0.91	1.19
	朝阳区	66918	37.2	7.5	2031	42.7	8.7	0.87	0.87
	海淀区	48888	35.7	7.3	1277	39.8	8.0	0.90	0.92
	丰台区	32170	38.6	9.0	678	46.6	9.9	0.83	0.91
	石景山区	4561	34.3	7.6	313	30.3	6.2	1.13	1.21
	昌平区	13035	36.5	8.8	202	47.4	11.1	0.77	0.79
	通州区	10400	38.4	10.1	181	40.9	12.8	0.94	0.79
近郊区	大兴区	9455	38.9	9.1	94	40.1	10.1	0.97	0.91
	房山区	3057	47.4	15.7	157	31.7	11.5	1.49	1.37
	门头沟区	1196	31.1	9.9	113	36.7	9.1	0.85	1.08
	怀柔区	299	44.3	12.5	8	28.8	11.6	1.54	1.08
远郊区	密云区	149	43.7	13.1	7	34.6	16.1	1.26	0.82
	平谷区	730	43.8	15.7	8	42.5	23.8	1.03	0.66
	顺义区	5497	34.3	10.0	80	39.5	14.1	0.87	0.71
	延庆区	254	36.8	12.1	10	56.0	41.9	0.66	0.29

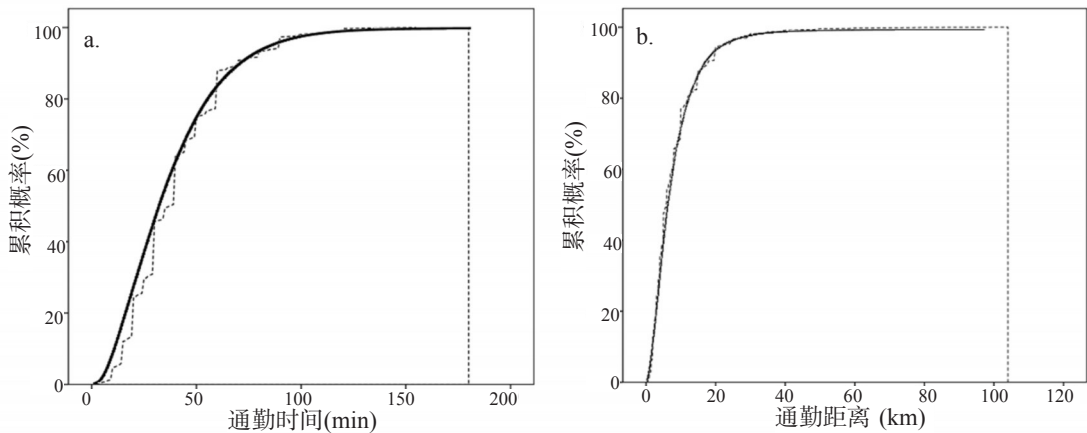


图8 本文的识别结果与2005年调查的通勤出行的累积概率分布函数对比:(a)通勤时间;(b)通勤距离。虚线:2005年调查;实线:本文的识别结果

Fig. 8 CDF comparison of commuting time (a) and distance (b) between commuting trips in the 2005 survey (dash lines) and our results (solid lines)

查是凭被调查者的记忆获取的,因此数据是离散的,导致CDF非平滑,但是和本文的研究结果也基本重合。③由于2005年调查的通勤人数较少,不适合进行TAZ尺度上的比较,因此在区县尺度上与本文研究结果进行对比,如图1d所示(本文仍沿用18个区的建制),包括8个中心城区,5个近郊区和5个远郊区。对比结果显示,本文的结果同2005年

⑩ 为了缓解交通拥堵问题,北京市政府分别于2011年4月26日和6月14日,开通了从回龙观(HLG)和天通苑(TTY)到金融街(JRJ),从通州(TZH)到CBD的联运通勤车(参考:<http://news.sina.com.cn/c/2011-04-26/013522356172.shtml>, <http://news.dichan.sina.com.cn/bj/2011/06/14/333255.html>)。图9b显示的交通繁忙的链接标识,可以为此项政策提供支持。

表 2 其他研究中的通勤时间和距离

Tab. 2 Commuting time and distance of Beijing by various existing researches

名称	通勤方式及年份	样本数量	平均通勤时间 (min)	平均通勤距离 (km)
本研究	公交车, 2008	221773	36.0 (24.2)	8.2 (7.0)
2005 年调查	公交车, 2005	6651	40.5 (23.1)	8.4 (8.3)
刘志林和王茂军 ^[16]	公交车, 2007	307	46.3 (N/A)	N/A
Wang 和 Chai ^[31]	公交车, 2001	227	55.1 (30.4)	N/A
Zhao 等 ^[30]	公交车和轨道交通, 2001	220	52.4 (26.6)	N/A

注: 括号里的数值为平均通勤时间和距离的标准差。除去本研究, 其他研究中的乘坐公交出行的样本都是从小规模调查中提取的。

调查相似度很高, 特别是通勤密集的中心城区。两者也存在一些差异, 例如怀柔区的通勤时间比率为 1.54, 可能是由于 2005 年调查中该区域样本数量有限造成。

同时也将本文研究结果与其他类似研究进行对比 (表 2)。2001 年和 2007 年的通勤时间显著高于本研究结果, 可能由于北京公共交通系统在 2001-2008 年间的改善。本研究中的通勤时间不考虑从居住地步行和骑自行车到公交车站和从公交站到就业地的时间^⑧, 可能也是造成结果差异的原因之一。

4.3 通勤出行的可视化

为了更好地展现北京市的通勤出行形态, 将识别的通勤出行进行空间化, 每条线代表一个通勤出行, 通勤时间、通勤距离和持卡人的 ID 记录在 GIS 图层的属性中。超长和超短时间的通勤出行标注在图 9a 中, 超过 90 min 的通勤主要来自于新城, 如平谷、密云和怀柔, 而小于 10 min 的主要来自五环内区域。为了识别北京市域的主导通勤出行方向, 进一步将通勤出行在 TAZ 尺度上汇总, 计算两个 TAZ 之间的通勤出行数量 (本文将两个 TAZ 之间的出行定义为“链接 link”, 共有 34219 个链接)。借鉴 Jiang 和 Liu^[32]研究中的首/尾划分方法 (head-tail division), 将出行数量分为六级 (图 9b), 级别越高表示通勤出行数量越多, 4-6 级对应的 175 个链接 (占总链接数的 0.5%) 包括 32156 次通勤出行 (占比 14.8%), 表示少数的路段承载了较多的出行, 其主要发生于六环路内, 跨越六环的情况很少, 这些链接建议提供直达的快速公共交通服务。

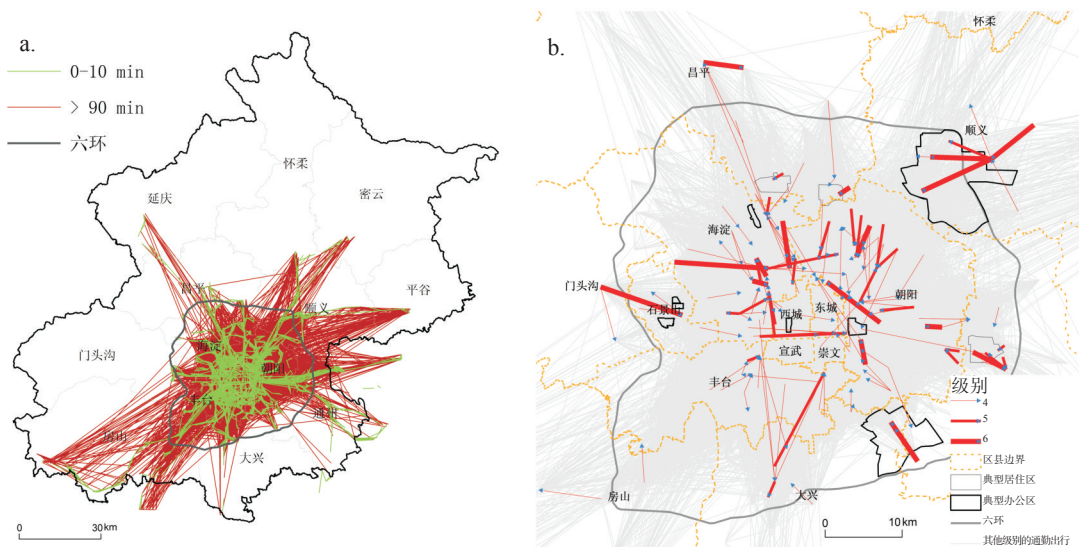


图 9 北京市中心区的通勤形态 (a) 极端出行时间的通勤出行; (b) TAZ 尺度的通勤链接

注: 图中箭头方向代表从居住地到就业地的方向。

Fig. 9 The identified commuting trips in the BMA with least time and greatest travel time (a) and commuting trip count aggregated in the TAZ scale

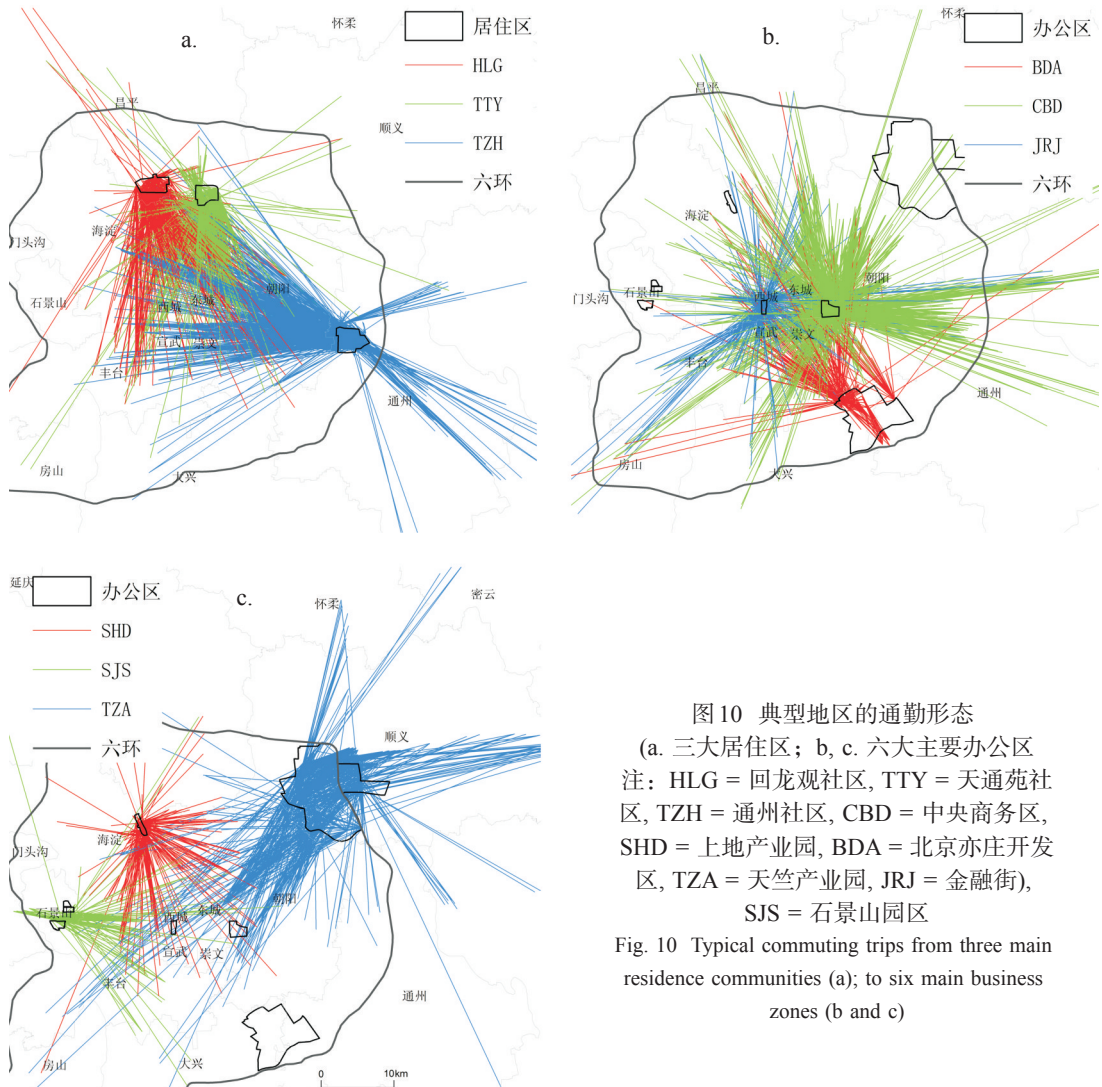


图 10 典型地区的通勤形态

(a. 三大居住区; b, c. 六大主要办公区
注: HLG = 回龙观社区, TTY = 天通苑社区, TZH = 通州社区, CBD = 中央商务区, SHD = 上地产业园, BDA = 北京亦庄开发区, TZA = 天竺产业园, JRJ = 金融街, SJS = 石景山园区

Fig. 10 Typical commuting trips from three main residence communities (a); to six main business zones (b and c)

由于大型居住区和过度集聚办公区的存在, 北京的交通拥堵情况和长距离出行受到多方关注。为此, 从识别的所有通勤出行中识别来自回龙观、天通苑和通州这3大居住区的出行(图10a)。前两个社区是20世纪90年代建设的北京北部最大的两个居住区, 通州地区集中了北京东部多个新建居住社区。类似地, 识别去往6个主要办公区的通勤出行(图10bc), 这6个办公区包括CBD(中央商务区)、上地(IT产业园)、亦庄(北京最大的国家级工业园区)、天竺(航空港)、石景山(北京西部办公区)和金融街(金融、银行和保险业集聚的区域)。

下面将典型地区的通勤出行从时间和距离角度进行汇总并对比(表3)。对于来自居住区的通勤出行, TTY居民的通勤距离比TZH短,

表 3 来自三大居住区和去往六大办公区的通勤时间和距离对比

Tab. 3 Commuting time and distance of various residence communities and business zones of Beijing

区域	通勤时间 (min)	通勤距离 (km)	占有识别的通勤出行的比例 (%)
来自三大居住区的通勤出行			3.9
TZH	45.1	10.0	1.4
HLG	39.4	7.0	1.0
TTY	36.2	6.1	1.5
去往六大办公区的通勤出行			6.0
CBD	41.4	9.4	2.7
SHD	40.4	6.7	0.3
JRJ	34.9	7.1	0.5
TZA	31.6	10.0	1.3
SJS	28.4	6.9	0.3
BDA	26.6	6.4	0.8

只有少数居民在南部地区工作；部分TZH居民在新城内部就业，少数在北京西部地区就业。对于去往办公区的通勤出行，CBD从不同地区吸引了分布广泛的就业者，使其通勤距离在所有办公区中是最长的；在BDA的就业者通勤时间和距离最短，可能是由于该区域是地方就业中心。除此之外，我们还发现仅有302次通勤出行(占比0.14%)是从三大居住区到六大办公区，和预期有很大差距，图9b可以说明这一点，一个可能的原因是往来于这两类区域的多数通勤者多使用私家车出行，而非乘坐公交车。

5 结论与讨论

本文使用公交IC卡刷卡数据(SCD)对城市职住关系和通勤出行进行了分析。首先，为了对SCD进行预处理并进行数据挖掘，构建了两类数据模型，原始的TRIP和PTD数据模型，前者用于识别通勤出行，后者用于识别职住地。其次，提出了一种利用一日SCD和土地使用图识别职住地的算法，并利用决策树对一周7天的识别结果进行综合，得到最终的职住地。识别结果包括1045785位持卡人的居住地和362882位持卡人的就业地，其中有237223位持卡人同时具有识别的职住地。再次，对通勤出行进行识别并可视化，共识别221773位持卡人的通勤出行。为了更好地展示北京市的总体通勤形态，将通勤时间和通勤距离在TAZ尺度进行汇总，一方面，特别分析了3大典型居住区和6个典型办公区的通勤出行特征；另一方面，根据TAZ间的通勤出行数量，识别通勤出行的主导方向。最后，从3个方面(平均值和标准差、累积分布函数、区(县)尺度的空间分布)将本研究结果与2005年居民出行调查的通勤时间和距离进行对比，结果显示二者具有较强的相似性，此外也与其他已有研究进行了对比分析。

本文的贡献主要体现在3方面：①使用传统调查和GIS数据建立规则，对SCD进行数据挖掘和模式识别，进而用于研究城市空间动态的方法，可以充分利用传统数据与“新”数据之间的长项，共同用于城市研究。②采用连续一周的SCD研究北京市的通勤形态，利用决策树的方法对分日识别结果进行了综合，相较于仅使用一日数据进行分析结果更为可靠。③与传统的问卷调查和居民出行调查相比，更详细地展示了北京市通勤出行的特征，与其他调研结果相比，本文所获得的通勤出行具有样本量大、时空信息准确的优势。总体上，本文提出的SCD分析方法在城市空间动态研究，尤其是通勤特征分析中有很好的适用性，为大都市区通勤研究提供了一种新的分析思路。

未来将从以下几方面开展后续研究。①北京市轨道交通系统也广泛应用公交IC卡，SCD信息与本文一致，随着轨道交通线路的建设，轨道交通出行比例上升较快(从2008年的8.0%增加到2009年的10.0%)，因此将通过进一步获得轨道交通SCD得到更完善的乘客出行信息。②居民出行调查数据中包括出行目的信息，应用本文的研究方法，利用PTD数据模型可以识别其他城市活动(如购物、医疗和娱乐)，进而形成人的一日活动链，对人们的活动行为可以有更深入的了解。

致谢：北京大学城市与环境学院人文地理专业硕士研究生卫晓对本文进行了文献搜索和稿件校对，在此表示感谢。

参考文献 (References)

- [1] Batty M. Invisible cities. *Environment and Planning B: Planning and Design*. 1990, 17: 127-130, doi: 10.1068/b170127.
- [2] Ahas R, Mark Uë. Location based services: New challenges for planning and public administration? *Futures*, 2005, 37: 547-561, doi: 10.1016/j.futures.2004.10.012.
- [3] Goodchild M F. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 2007, 69(4): 211-221, doi: 10.1007/s10708-007-9111-y.

- [4] Newhaus F. Urban diary: A tracking project//UCL Working Paper Series. Paper 151.
Available on line: <http://discovery.ucl.ac.uk/19245/>
- [5] Jiang B, Yin J, Zhao S. Characterizing human mobility patterns in a large street network, *Physical Review*, 2009, E80 021136, doi: 10.1103/PhysRevE.80.021136.
- [6] Liu L, Andris C, Ratti C. Uncovering cabdrivers' behavior patterns from their digital traces *Computers. Environment and Urban Systems*, 2010, 34(6): 541-548, doi: 10.1016/j.compenvurbsys.2010.07.004.
- [7] Ratti C, Pulselli R M, Williams S et al. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 2006, 33(5): 727-748, doi: 10.1068/b32047.
- [8] Calabrese F, Ratti C. Real time Rome. *Networks and Communication Studies*, 2006, 20: 247-258.
- [9] Steenbruggen J, Borzacchiello M T, Nijkamp P et al. Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: A review of applications and opportunities. *GeoJournal*, 2011, doi: 10.1007/s10708-011-9413-y.
- [10] Reades J, Calabrese F, Ratti C. Eigenplaces: Analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 2009, 36(5): 824-836, doi: 10.1068/b34133t.
- [11] Milton R. Twitter tags and real-time visualisation of complex geographic data with MapTube. *Advances in Spatial Analysis & e-Social Science*, 2010. A One Day CASA Conference, UCL, UK.
Available on line: <http://www.casa.ucl.ac.uk/conference/programme.html?abstract=a1500>
- [12] Jones B W, Spigel B, Malecki E J. Blog links as pipelines to buzz elsewhere: The case of New York theater blogs. *Environment and Planning B: Planning and Design*, 2010, 37(1): 99-111, doi: 10.1068/b35026.
- [13] Rekimoto J, Miyaki T, Ishizawa T. LifeTag: WiFi-based continuous location logging for life pattern analysis. 3rd International Symposium on Location- and Context-Awareness (LOCA2007), 2007: 35-49.
- [14] Torrens P M. Wi-Fi geographies. *Annals of the Association of American Geographers*, 2008, 98(1): 59-84, doi: 10.1080/00045600701734133.
- [15] Shoval N. Sensing human society. *Environment and Planning B: Planning and Design*, 2007, 34(2): 191-195, doi: 10.1068/b3402com.
- [16] Liu Zhilin, Wang Maojun. Job accessibility and its impacts on commuting time of urban residents in Beijing: From a spatial mismatch perspective. *Acta Geographica Sinica*, 2011, 66(4): 457-467. [刘志林, 王茂军. 北京市职住空间错位对居民通勤行为的影响分析: 基于就业可达性与通勤时间的讨论. *地理学报*, 2011, 66(4): 457-467.]
- [17] Long Y, Shen Z, Mao Q et al. Retrieving individuals' attributes from aggregated dataset for urban micro-simulation: A primary exploration. *Geoinformatics*, 2010, doi: 10.1109/GEOINFORMATICS.2010.5567738.
- [18] Bagchi M, White P. What role for smart-card data from bus systems? *Municipal Engineer* 2004, 157(1): 39-46, doi: 10.1680/muen.2004.157.1.39.
- [19] Joh C H, Hwang C. A Time-geographic analysis of trip trajectories and land use characteristics in Seoul metropolitan area by using multidimensional sequence alignment and spatial analysis, 2010 AAG Annual Meeting, Washington, DC. Abstract available on line: <http://meridian.aag.org/callforpapers/program/AbstractDetail.cfm?AbstractID=30812>
- [20] Jang W. Travel time and transfer analysis using transit smart card data: Transportation Research Record. *Journal of the Transportation Research Board*, 2010, 2144: 142-149, doi: 10.3141/2144-16.
- [21] Roth C, Kang S M, Batty M et al. Structure of urban movements: Polycentric activity and entangled hierarchical flows *PLoS ONE*, 2011, 6(1): e15923, doi: 10.1371/journal.pone.0015923.
- [22] Zhou Tao, Zhai Changxu, Gao Zhigang. Approaching bus OD matrices based on data reduced from bus IC cards. *Urban Transport of China*, 2007, 5(3): 48-52. [周涛, 翟长旭, 高志刚. 基于公交IC卡数据的OD推算技术研究. *城市交通*, 2007, 5(3): 48-52.]
- [23] Dong Xiaojing, Yu Zhiwei, Fu Weiwei et al. Data processing and analyzing system for bus IC card based on GIS. *Geospatial Information*, 2009, 7(5): 124-126. [董晓晶, 余志伟, 伏伟伟 等. 基于GIS的公交IC卡数据处理及分析系统. *地理空间信息*, 2009, 7(5): 124-126.]
- [24] Peng Han, Han Xiuhua, Tian Zhenzhong et al. Transfer matrix construction method based on bus IC card data processing. *Computer and Communications*, 2007, (4): 32-34. [彭晗, 韩秀华, 田振中 等. 公交IC卡数据处理的换乘矩阵构造方法研究. *交通与计算机*, 2007, (4): 32-34.]
- [25] Yang Zhiwei, Zhao Qian, Zhao Shengchuan et al. Passenger flow volume forecasting method based on public transit intelligent card (IC) survey data. *Transport Standardization*, 2009, (9): 115-119. [杨智伟, 赵骞, 赵胜川 等. 基于公交IC卡数据信息的客流预测方法研究. *交通标准化*, 2009, (9): 115-119.]
- [26] Liu Xiaoming. Bus smart cards will be extensively adopted in Beijing on April 1, 2006, 2009, <http://www.enet.com.cn/article/2005/1229/A20051229488080.shtml>. accessed on 2010-05-25. [刘小明. 2006年4月1日起北京将全面推行公交IC卡. <http://www.enet.com.cn/article/2005/1229/A20051229488080.shtml>]

- [27] Beijing Transportation Research Center. Beijing Transportation Annual Report 2009. 2009. [北京市交通发展研究中心. 2009北京市交通发展年度报告. 2009.]
- [28] Beijing Institute of City Planning. Beijing Job Position Allocation for 2008. 2010. [北京市城市规划设计研究院. 2008年北京市就业空间分布研究报告. 2010.]
- [29] Beijing Municipal Commission of Transport, Beijing Transportation Research Center. The 3rd Transportation Comprehensive Survey Report of Beijing Working Report. 2007. [北京市交通委员会, 北京市交通发展研究中心. 北京市第三次交通综合调查总报告. <http://210.75.218.99/InfoCenter/Func/OpenNews.asp?NewsID=INN20090706001>. 2007.]
- [30] Zhao P, Lv B, Roo G de. Impact of the jobs-housing balance on urban commuting in Beijing in the transformation era. *Journal of Transport Geography*, 2011, 19: 59-69, doi: 10.1016/j.jtrangeo.2009.09.008.
- [31] Wang D, Chai Y. The jobs-housing relationship and commuting in Beijing, China: The legacy of Danwei. *Journal of Transport Geography*, 2009, 17: 30-38, doi: 10.1016/j.jtrangeo.2008.04.005.
- [32] Jiang B, Liu X. Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information. *International Journal of Geographical Information Science* 2012, 26(2): 215-229, doi: 10.1080/13658816.2011.575074.

Identifying Commuting Pattern of Beijing Using Bus Smart Card Data

LONG Ying, ZHANG Yu, CUI Chengyin
(*Beijing Institute of City Planning, Beijing 100045, China*)

Abstract: This paper combines the one-week bus smart card data (SCD) and one-day household travel survey as well as the parcel-level land use map for identifying jobs-housing places and commuting trips in the Beijing Metropolitan Area with an area of 16,410 square kilometers. The identification result is aggregated in the bus stop and traffic analysis zone (TAZ) levels, respectively. In particular, commuting trips with commuting time and distance attached from three typical residence communities and those to five typical business zones are mapped and compared with each other to analyze commuting patterns of Beijing. The identified commuting trips are compared with those in the household travel survey in terms of commuting time and distance, indicating that our results are coincident with the survey significantly. Our approach is proved to have its potential in identifying more solid identification result based on rules extracted from existing surveys or censuses.

Key words: bus smart card data; jobs-housing; commuting time and distance; spatial mismatch; Beijing