文章编号:1671-6833(2024)03-0046-09

基于手机信令数据的城市居民动态 OD 矩阵提取方法

田 钊[1], 张乾钟[1], 赵 轩[1], 陈 斌[2], 佘 维[1], 杨艳芳[3,4]

(1.郑州大学 网络空间安全学院,河南 郑州 450001;2.郑州大学 计算机与人工智能学院,河南 郑州 450001;3.交通运输部科学研究院,北京 100029;4.综合交通运输大数据应用技术交通运输行业重点实验室,北京 100029)

摘 要:现有的城市居民出行调查周期较长,交通小区划分粒度粗糙,导致调查不能及时准确地获取居民出行信息。针对该问题,提出了一种基于手机信令数据的城市居民动态 OD 矩阵提取方法。首先,针对信令数据中的两种复杂噪声:乒乓切换和漂移数据,提出了基于窗口阈值的检测与等效位置替换方法,以及复杂漂移点的检测和标记处理方法;然后,提出一种改进的 ST-DBSCAN 聚类方法,引入一种等时化方法将时间信息与空间信息相结合,识别出行过程中的驻留点;最后,基于地理信息系统构建含有道路关键节点的路网,将居民出行 OD 与路网节点相匹配,有效推导出城市居民动态 OD 矩阵。实验结果表明:与 ST-DBSCAN 算法相比,所提改进的 ST-DBSCAN 算法在聚类效果和识别速度上分别提升了 6.10%和 5.26%;与统计方法和二阶统计量方法相比,基于改进的 ST-DBSCAN 算法的动态 OD 矩阵提取方法在均方误差(*MSE*)上分别降低了 16.98%和 21.55%。以北京市为例,运用提出的动态 OD 矩阵提取方法,能够及时有效地分析城市居民日常与高峰时段的出行特征。

关键词:城市出行; 智能交通系统; 手机信令数据; 动态 OD 矩阵; 驻留点识别; 时空特征分析

中图分类号: TU998;U491 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.03.006

传统交通调查方法在获取居民出行信息方面取得了显著进展,但其调查周期长、成本高和抽样率有限等问题限制了其在现代智能交通系统对快速获取动态出行起讫点(origin-destination, OD)矩阵方面的应用[1-3],这促使研究者寻找更有效的数据收集方法。21 世纪出现了新兴的交通信息获取方法,尽管它们在数据采集速度方面有所改进,但面临多源异构交通数据的成本高、覆盖率低等问题,GPS 虽然能提供高精度位置数据,但在大规模应用中仍存在限制[4]。

近年来,随着移动通信网络的迅速发展和智能手机的广泛应用,基于手机定位的大数据为动态 OD 矩阵获取带来了新的活力[5-6]。手机信令数据成本低、覆盖广泛,为动态 OD 矩阵分析提供了强大的数据支持。此外,动态交通分配的合理路径集合算法研究为交通分配提供了新的视角[7]。White 等[8]调查了在英国肯特地区使用电话账单数据获取 OD 信息的可行性并验证了从手机数据中获取 OD 信息的可行性。Caceres 等[9]评估了使用手机位置数据库提取 OD 矩阵的可行性,同时使用手机通信网络模拟器从电话网络中模拟和提取 OD 矩阵。Zhang 等[10]使用模拟的蜂窝探测器轨迹信息估计每日 OD 需求,并通过 VISSIM 模拟测试该方法。这些结果显示了使用手机数据作为分析 OD 矩阵的巨大潜力[11]。尽管 White 等[8]和 Caceres 等[9]的研究验证了从手机数据中获取 OD 信息的可行性,但本研究主要关注如何从手机数据中精确、高效地提取 OD 信息。

城市交通流在宏观和微观层面的流动性有着潜在的联系[12]。宏观层面的研究通常将停留点与特定大小的研究区域相连接,以确定区域级别的 OD 对,通常为基于交通小区的划分或基于基站群聚类

收稿日期:2023-10-20;**修订日期**:2023-11-28

基金项目:河南省重点研发与推广专项基金资助项目(212102310039);河南省高校科技创新人才支持计划基金资助项目(21HASTIT031);综合交通运输大数据应用技术交通运输行业重点实验室开放课题(2022B1201)

作者简介:田钊(1985—),男,河南荥阳人,郑州大学副教授,博士,主要从事智能交通、信息安全等研究,E-mail:tianzhao@zzu.edu.cn。

通信作者:杨艳芳(1985—),女,广西百色人,交通运输部科学研究院副研究员,博士,主要从事交通运输信息化研究,E-mail:yangyf@motcats.ac.cn。

引用本文:田钊,张乾钟,赵轩,等. 基于手机信令数据的城市居民动态 OD 矩阵提取方法[J]. 郑州大学学报(工学版),2024,45(3):46-54. (TIAN Z, ZHANG Q Z, ZHAO X, et al. Dynamic OD matrix extraction method of urban residents based on cell phone signaling data[J]. Journal of Zhengzhou University (Engineering Science),2024,45(3):46-54.)

的划分粒度。微观层面则更关注细致的分析,直接使用最小粒度的 OD 对进行交通流分析。黄美灵等[13]利用 CDMA 通信网络手机数据,在划分的交通小区下使用面积判断法求得出行者 OD 信息。魏玉萍等[14]、蔡军等[15]进一步将所得出行者的 OD 位置信息与已有交通小区相结合获取居民出行动态 OD 矩阵。这些研究为理解城市交通流提供了有价值的见解,但也存在一些局限性。例如,黄美灵等[13]的方法可能受限于数据的覆盖范围和精确度,而魏玉萍等[14]、蔡军等[15]的方法可能导致信息的丢失或模糊。此外,现有研究可能过于依赖特定大小的交通小区或基站群聚类的划分粒度,无法捕捉到复杂的交通流和动态的交通流。

城市基础道路网络作为城市的基本骨架,承担着大部分交通流。现有研究通常集中在宏观或微观的某一个方面,虽然基于区域或点的 OD 在先前研究中常用,但未充分利用城市基础道路网络的信息,也未充分考虑复杂和动态的交通流[16-18]。本文旨在结合宏观和微观视角,从基础路网节点视角出发,将从手机数据中提取的出行端点分配给道路关键节点。附近的出行数据可以用道路节点值表示为共同的起点/终点,进而通过分析手机信令这一连续动态大数据,探讨城市居民活动的时空变化特征。

1　动态 OD 矩阵提取方法

区别于传统的 OD 调查方法,本文提出了针对手机信令数据特征的动态 OD 矩阵提取方法。该方法包括数据预处理、驻留点识别以及动态 OD 矩阵构建,技术框架如图 1 所示。

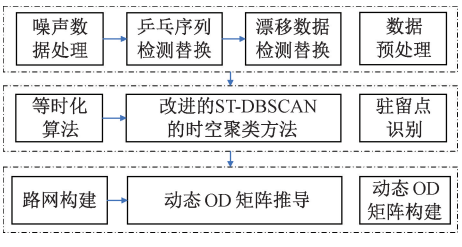

图 1　本文算法的技术框架

Figure 1　Technical framework of the algorithm in this study

1.1　数据预处理

基站生成位置记录的触发机制分为事件驱动机制和网络更新机制。事件驱动机制产生的数据包括通话详单和网络协议数据,而网络更新机制产生的数据则为保证通信服务而生成,包括手机位置和更新信息。网络更新数据的记录频率显然高于事件驱动数据,因此本文采用数据生成频率较高的网络更新机制产生的手机信令数据(CDR)。

手机信令数据中的噪声数据处理不当,会严重影响出行轨迹段的提取和识别精度。本文涉及的噪声数据有 4 种:关键数据缺失、重复冗余数据、乒乓切换、漂移数据。

1.1.1　关键数据缺失

关键数据的缺失会影响后续数据清洗和数据对齐操作,故需删除关键字段值为空的记录。

1.1.2　重复冗余数据

用户在同一基站范围内活动会生成大量冗余数据。这类数据包含完全相同的字段。重复冗余数据过滤的方法:信令数据按照 TIMESTAPMP 字段顺序排列;删除连续出现的同一 CELLID 字段的冗余数据。

1.1.3　乒乓切换

手机处于多个基站的信号覆盖范围内时,手机信号会在短时间内连接多个不同的基站,这种现象称为乒乓切换。本文采用基于时间窗的乒乓序列检测和等效位置替换算法。

乒乓序列检测示意图如图 2 所示,L_1、L_2、L_3、L_4、L_5、L_6、L_7 分别标识不同的基站。算法步骤如下。

步骤 1　设置时间窗口阈值,将原始序列中的第 1 个点设置为基准点。

步骤 2　从基准点开始的时间窗口内的轨迹点提取为序列 S。

步骤 3　判断序列 S 中基准点是否重新出现,是则转步骤 4,否则转步骤 5。

步骤 4　若基准点重新出现,则将序列 S 中基准点和其最后一次出现之间的信令数据作为乒乓切换序列,该序列的下一个信令点作为新的基准点。针对乒乓切换序列,寻找等效位置:①计算乒乓切换序列中每个点的停留时间比例,信令点的停留时间比例见式(1)、(2);②停留比例最高的信令点作为等效位置替换乒乓序列。

$$lo_{\text{mean}} = \frac{\sum Tk_i \cdot lo_i}{\sum Tk_i}; \tag{1}$$

$$la_{\text{mean}} = \frac{\sum Tk_i \cdot la_i}{\sum Tk_i}。 \tag{2}$$

式中:lo_i、la_i 为信令点的经纬度;Tk_i 为该点的停留时间。

步骤 5　若序列 S 中基准点没有再次出现,则将下一点作为基准点。

步骤 6　重复步骤 3~5,直至所有信令点遍历结束。

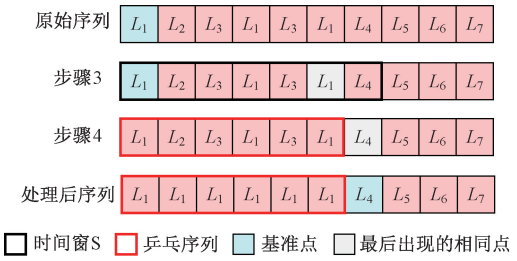

图 2 乒乓序列检测示意图

Figure 2 Ping-pong switching detect diagram

1.1.4 *漂移数据*

漂移数据是指在无线通信过程中,由于信号传播特性和环境因素的影响,导致记录的手机位置与实际位置存在一定偏差而产生的数据。本文将漂移点定义为轨迹列表中进行长距离高速移动的信号点。

漂移数据可分为简单漂移和复杂漂移。简单漂移为较容易识别的偏离点,复杂漂移指手机在移动过程中连接多个距离实际位置远超预期的基站。

本文提出一种复杂漂移点检测算法,利用高频正常点对信令数据进行判别。通过引入高频正常点以保证部分正常点不会被标记为漂移点,同时将被误判为正常点的高频正常点重新认定为漂移点。

复杂漂移点检测算法步骤如下。

步骤 1 设定漂移距离阈值、速度阈值、高频正常点频率阈值,对每个轨迹点出现的频率进行统计,若频率超过设定阈值,则将其识别为高频正常点。

步骤 2 计算基准轨迹点 n_2 与下一未被确认轨迹点 u_1 之间的距离 Dis、速度 $\overline{V}$;判断是否均大于步骤 1 中相应阈值。若大于转步骤 3,否则转步骤 4;两轨迹点间距离和速度的计算见式(3)~(6)。

$$Dis = 2r\arcsin\sqrt{h}; \tag{3}$$

$$h = \sqrt{\varphi(la_2,la_1) + \cos(la_1)\cos(la_2)\varphi(lo_2,lo_1)}; \tag{4}$$

$$\varphi(lo_1,lo_2) = \sin^2\left(\frac{lo_2 - lo_1}{2}\right); \tag{5}$$

$$\overline{V} = \frac{Dis}{Tp_{i+1} - Tp_i}。 \tag{6}$$

式中:r 为地球半径;la_i、$lo_i(i \in \{1,2\})$ 分别为轨迹点的纬度与经度;Tp_i 为轨迹点的时间戳。式(3)中距离公式采用 Haversine 公式,式(5)为 $\varphi(lo_1, lo_2)$ 计算,$\varphi(la_1,la_2)$ 同理。

步骤 3 若基准点与未被确认点间的参数值均大于预设阈值,则继续判断点 u_1 是否在高频正常点中:①若下一个点 u_1 不在高频正常点中,则该点为漂移点 $d_{2,2}$,将未确认点 u_2 作为下一识别点;②若该值属于高频正常点,则认定该正常点 n_2 为漂移点,因间隔时间较长而被判定为正常点,将正常点 n_2 修改为漂移点 $d_{1,4}$,并且重新设置上一个正常点 n_1 为基准点,最后将点 n_2 为基准点所识别的漂移点 $d_{2,i}$ 依次设置为未被确认点 u_i。

步骤 4 若基准点与未被确认点间的参数值不都大于预设阈值,则未被确认点 u_1 将被识别为正常点 n_3,并设置为基准点。

步骤 5 重复步骤 2~4,直至所有轨迹点识别完成。

漂移序列检测示意图如图 3 所示,n_i、$d_{i,j}$、u_i 为用户出行轨迹中的序列点,其中 n_1 为已被识别的正常轨迹点;$d_{1,j}$ 为以 n_1 为基准点识别成漂移点的漂移序列;u_1 为未被确认的正常点。

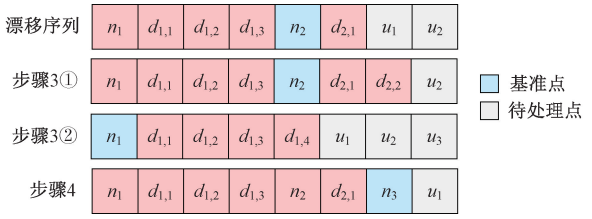

图 3 漂移序列检测示意图

Figure 3 Drift data detect diagram

1.2 驻留点识别

本文提出了一种改进的 ST-DBSCAN 聚类算法识别轨迹点中的停留点。该算法在 ST-DBSCAN 聚类算法的基础上引进等时化算法,对出行中的时间维度进行语义扩充,从而达到在时间和空间上同时聚类的效果。本文方法能够更好地处理时间序列数据。驻留点识别可视化流程如图 4 所示。

1.2.1 *数据等时化*

为了对传统的 ST-DBSCAN 算法在时间维度上进行扩充,需要对手机信令数据等时化处理。本文使用线性插值法,首先将时间间隔设定为 T',24 h 共 1 440 min,故分为 1 440/ T' 个时间片。数据等时化要求每个时间片都要有轨迹点,因此对于时间片不连续的情况,需要插入轨迹点作为补充。

等时化处理方法如下:若轨迹点 P_n 位于时间片 t_n 内,轨迹点 P_{n+1} 位于时间片 t_{n+k} 内,则需要插入 $k-1$ 个轨迹点;若相邻两轨迹点同处一时间片中,则保留前者。式(7)为待插入的轨迹点的时间,式(8)、(9)为经、纬度表达式。

$$p_m^{\text{time}} = (\lfloor p_n^{time}/T' \rfloor + m) \cdot T'; \tag{7}$$

$$p_m^{\text{lon}} = p_n^{\text{lon}} + \frac{p_{n+1}^{\text{lon}} - p_n^{\text{lon}}}{p_{n+1}^{\text{inte}} - p_n^{\text{inte}}} \cdot m; \tag{8}$$

$$p_m^{\text{lat}} = p_n^{\text{lat}} + \frac{p_{n+1}^{\text{lat}} - p_n^{\text{lat}}}{p_{n+1}^{\text{inte}} - p_n^{\text{inte}}} \cdot m。 \tag{9}$$

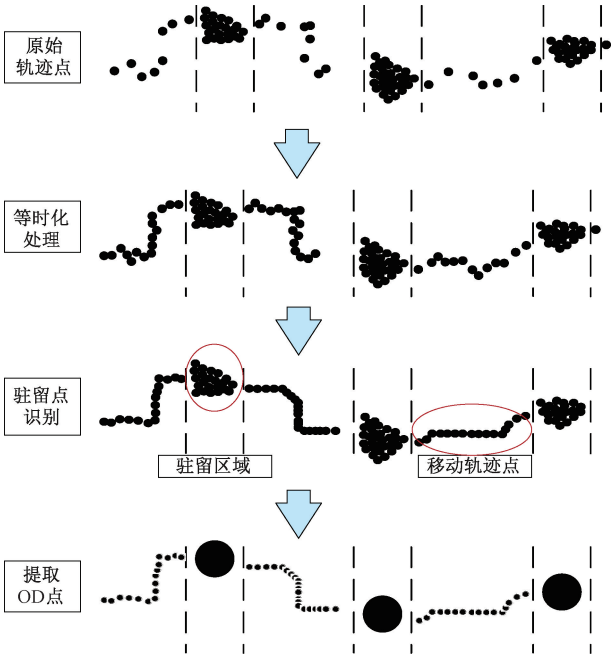

图 4　驻留点识别示意图

Figure 4　Schematic diagram of dwell point identification

式中:$m,n \in \{1,2,\cdots,k-1\}$;$p_m^{\text{time}}$ 为第 m 个待插入的轨迹点的时间;p_m^{lon} 为第 m 个待插入的轨迹点的经度;p_m^{lat} 为第 m 个待插入的轨迹点的纬度;T' 为时间片粒度。

1.2.2　基于改进的 ST-DBSCAN 的时空聚类方法

等时化处理使得数据在时间维度上具有连续性,从而将居民出行过程中的时间信息与空间信息相融合,形成更可靠的驻留点识别结果。本文的改进算法在处理数据时,不仅考虑了空间距离,还考虑了时间距离。这样,本文的算法能够更准确地识别出居民的停留点,尤其是在时间跨度较大的情况下。

本文对于居民出行驻留点的定义:以轨迹点 P_i 为圆心,驻留点定义中出行距离阈值 ε 为半径的 ε-邻域内,若包含某轨迹点在内的轨迹点数量 M 大于驻留点时间判断阈值 $minPts$,那么称该 ε-邻域内的所有轨迹点组成的点的集合 S 为驻留区域,同时该圆的中心点为停留点,到达时间为该圆内的第 1 个轨迹点的到达时间。等时化处理信令数据集后,居民出行过程中表达为驻留时间的时间维度信息转换成空间维度的轨迹点的密度大小,则 $minPts$ 即为所表达的值。改进的 ST-DBSCAN 聚类算法伪代码如下。

输入:清洗后的信令数据集 cleaned_CDR、空间邻域半径阈值 ε、时间邻域半径阈值,核心对象判断阈值 $minPts$;

输出:聚类后的 OD 轨迹。

① 初始化空间邻域半径阈值 ε、时间邻域半径阈值,核心对象判断阈值 $minPts$;

② for 未被标记的对象 p in 数据源 cleaned_CDR:

　　等时化处理(p);

　　if (p 的 ε-邻域内轨迹点数量<$minPts$):

　　标记 p 为噪声对象;

　　else:

　　建立新簇,找到所有与 p 密度相连的对象并将其加入新簇;

　　找出新簇内所有核心对象的密度相连的对象并将它们加入新簇;

　　标记新簇内的所有对象;

③ END FOR

④ END

1.3　动态 OD 矩阵构建

OD 信息提取之后,要从这些海量的 OD 信息中抽取交通研究所需的特征信息。需要把微观的 OD 数据聚合以相对宏观化的形式呈现,因此本文将对研究区域内的 OD 信息与城市道路网络关键节点进行匹配,从而构建动态 OD 矩阵。

在城市交通研究中,选择合适的 OD 矩阵构建策略至关重要。基于统计方法的 OD 矩阵求解依赖传统交通流量数据能准确估计基础流量,但在处理复杂噪声和动态 OD 矩阵方面存在局限,更适用于初步流量估计。基于路段流量二阶统计量的 OD 矩阵估计提高了模型准确性,但在捕获个体出行行为和处理复杂噪声方面仍有限制,适用于需要准确反映流量特性的情景。而本文提出的构建方法基于数据去噪和改进的聚类算法,适用于大规模研究,但准确性受到模型假设和参数选择的影响。

1.3.1　路网构建

现代城市中路网结构复杂,各种道路纵横交错,然而城市道路的分布并不是完全无序的。城市道路建设通常有着明确的规划,这使得路网有着明确的分区和清晰的结构。

首先,通过 OSM(open street map)API 获取北京市道路网络数据,然后根据道路等级属性进行筛选,包括国道、县道、城市快速路、高速路和其他低等级道路。本文的关注重点是道路交叉口和端点。因此,在获取详细道路网络数据后,首先,筛选排除低等级道路,得到城市主干道的矢量图;其次,利用地理信息系统(GIS)工具 ArcMap 进行拓扑检查,以判断道路之间的关系并检测连接错误或者悬挂的线,经过拓扑检查,多余的错误路段将被剪断;最后,将路网的交叉口或端点定义为节点,而路段则是两节点间的道路。此时整个基础路网形成了节点-路段-

节点的拓扑结构 $Net=(N,R)$,其中 $N=\{n_1,n_2,\cdots,n_m\}$,$R=\{(n_i,n_j)\mid n_i,n_j\in N\}$。

1.3.2 动态 OD 矩阵

利用 1.2 节提出的驻留点识别算法,提取研究区域内居民的出行起讫点集合 $OD=\{(o_1,d_1),(o_2,d_2),\cdots,(o_n,d_n)\}$,结合 1.3.1 节中路网拓扑关键节点集合 $N=\{n_1,n_2,\cdots,n_m\}$,计算出行动态 OD 矩阵,计算步骤如下。

步骤 1 初始化一个二维 OD 矩阵 $\boldsymbol{D}$,$\boldsymbol{D}$ 为 $m\times m$ 的矩阵,存储分配至节点的 OD 信息。

步骤 2 从出行起讫点集合 OD 中依次取出 (o_k,d_k),$k=1,2,\cdots,n$,并分别将起点 o_k、终点 d_k 匹配至最近邻的路网节点,o_k 的最近邻节点为 α,d_k 的最近邻节点为 β,每个起讫点都会被分配至最近邻节点。

步骤 3 若 $\alpha\neq\beta$,认定 $\{o_k\rightarrow d_k\}$ 为一次有效出行,则矩阵 $\boldsymbol{D}$ 中 $D_{\alpha\beta}$ 的值:$D_{\alpha\beta}=D_{\alpha\beta}+1$,关注从不同起始节点到不同结束节点之间的出行次数;若 $\alpha=\beta$,表明出行仅发生在单个节点内部,不计入统计数据,则忽略此次节点内部出行。将重点放在跨节点的出行上,可以更好地了解不同区域之间的出行需求和行为。

步骤 4 统计每个路网拓扑节点的发生量和吸引量,即分别对矩阵第 1~m 行和第 1~m 列求和得到每个道路节点发生量 G 和吸引量 A。

$$G=(G_1,G_2,\cdots,G_m); \tag{10}$$

$$G_i=\sum_{j=1}^{m}D_{ij}; \tag{11}$$

$$A=(A_1,A_2,\cdots,A_m); \tag{12}$$

$$A_j=\sum_{i=1}^{m}D_{ij}。 \tag{13}$$

式中:G_i 为单个道路节点发生量;A_i 为单个道路节点吸引量;$i,j\in\{1,2,\cdots,m\}$。

2 实验

2.1 实验数据

本研究使用的 CDR 数据由北京市某电信运营商提供,数据采集时间为 2012 年 8 月 10 日 0 时至 2012 年 8 月 10 日 24 时。信令数据属性包括国际移动用户识别码 IMSI、信令产生时间戳 TIMESTAMP、位置区码 LAC、蜂窝区码 CELLID、事件类型 EVENTID,数据类型如表 1 所示。

2.1.1 手机基站数据

基站的信息包括 LAC、CELLID 以及基站的经纬度(LON、LAT)。基站数据类型如表 2 所示。

表 1 信令数据格式表

Table 1 CDR data format table

属性	描述	所属类型
IMSI	国际移动用户标识,经过单向加密,唯一标识用户	Text
TIMESTAMP	信令产生时间戳,精确到秒,共 14 位	Datetime
LAC	位置区码,是为寻呼而设置的一个区域,覆盖一片地理区域	UINT32
CELLID	蜂窝区码	UINT32
EVENTID	事件类型	UINT8

表 2 基站数据格式表

Table 2 Base station data format table

基站信息	描述	所属类型
LAC	位置区码,是为寻呼而设置的一个区域,覆盖一片地理区域	UINT32
CELLID	蜂窝区码	UINT32
LON	基站基于大地坐标系下的经度	UINT32
LAT	基站基于大地坐标系下的纬度	UINT32

手机信令数据集的总信令数达到 376 923 931 条,这些数据的平均采样间隔为 213 s,每位用户每天平均产生 23 条信令。每个用户每天平均 CDR 数据量分布直方图如图 5 所示。

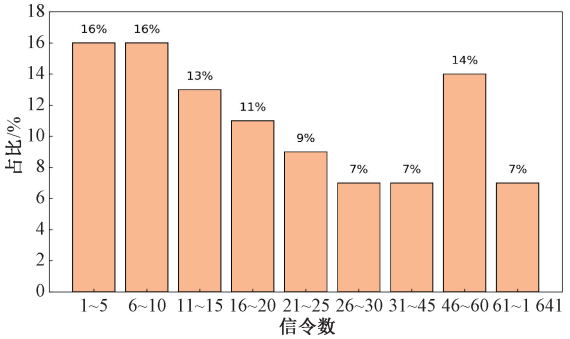

图 5 每人每天产生的信令数

Figure 5 Number of signals generated per person per day

现代通信网络中的拓扑结构与数据结构中的树状模型显著相似。这种结构的一个关键组成部分是蜂窝小区,它在逻辑结构上可以被认为是最基本的通信单元。由于蜂窝区码在所研究的区域内维持其唯一性,本文选择蜂窝区码作为基站的唯一标识。通过对表 1 和表 2 的 UNION 操作,提取了本文所需的相关字段,信令数据的部分示例值如表 3 所示。北京市基站分布热力图如图 6 所示。

2.2 数据预处理

数据预处理 4 个步骤为关键数据缺失去除、重复冗余数据去除、乒乓切换序列检测、漂移数据检测。关键数据缺失和重复冗余数据的去除通过较为简单的逻辑便可完成。乒乓切换序列检测算法中涉

表 3　部分信令数据示例

Table 3　Partial signaling data example

用户编号	基站编号	时间	纬度/(°)	经度/(°)
3701 ******22443	194	2012-08-10T00:56	116.184 991	39.911 211
3701 ******34553	14	2012-08-10T01:48	116.357 778	39.971 667
3701 ******33672	2448	2012-08-10T02:24	116.438 201	39.928 201

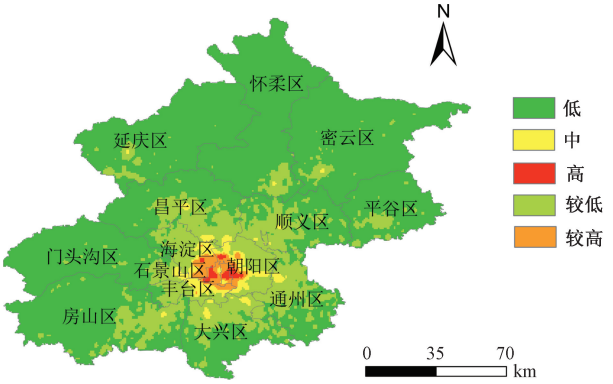

图 6　北京市基站分布热力图

Figure 6　Heatmap of base station distribution in Beijing

及时间窗口阈值,通过分析乒乓序列切换的时间特征,将时间窗阈值设置为 300 s。

漂移数据的清洗过程涉及 3 个预设阈值,分别是漂移距离判定阈值 $Dis=2$ km、漂移速度判定阈值 $V'=120$ km/h 以及高频正常点频率阈值 $Freq'=3$。根据城市出行的特点,高频正常点的频率阈值设定与居民出行时间跨度有关,出行时间越长,该频率越高。考虑到该数据集大部分轨迹较为稀疏,因此将频率阈值设置为 3。

2.3　驻留点识别

在交通出行领域,单次出行的行为定义:行程距离超过 300 m 且行程时间超过 15 min。本文根据信令数据集的实际情况,将驻留邻域半径 ε 取值为 300,驻留时间阈值 $minPts$ 取值为 15。

本文使用轮廓系数来评估本文改进的 ST-DBSCAN 算法的聚类效果。轮廓系数通过结合聚类的凝聚度和分离度来度量聚类结果的质量。轮廓系数的值介于-1 和 1 之间,值越大聚类效果越好。对于每个样本点 i,其轮廓系数 $s(i)=(b(i)-a(i))/\max\{a(i),b(i)\}$,整个数据集的轮廓系数为所有样本点的轮廓系数的平均值。其中 $a(i)$ 为样本点 i 的凝聚度,表示样本点 i 与同一聚类中所有其他点的平均距离;$b(i)$ 为样本点 i 的分离度,表示样本点 i 与其他聚类中所有点的平均距离。

高效的聚类结果意味着算法能够更准确地识别出行者的 OD 位置信息,这对于获取精确的动态 OD 矩阵至关重要。同时,聚类效率的提升也表明本文算法在处理复杂交通数据时具有更强的鲁棒性,这对于应对复杂和不确定的实际交通情况具有重要意义。在识别速度上,尽管改进的 ST-DBSCAN 不如 *K*-Means 和 DBSCAN 快速,但相比原始的 ST-DBSCAN,其识别速度有所提升,可以更好地满足现代智能交通系统对于快速获取动态 OD 矩阵的需求。

与聚类效率的提升相比,识别速度的提升不明显。这是因为 *K*-Means 和 DBSCAN 在速度上通常优于基于密度的聚类算法(如 ST-DBSCAN),而 ST-DBSCAN 更擅长处理各种形状和大小的簇,比如在复杂的空间或时间模式,ST-DBSCAN 有更高的聚类效率。改进的 ST-DBSCAN、ST-DBSCAN、*K*-Means、DBSCAN 在识别速度与聚类效率上的对比如图 7 所示。与传统 ST-DBSCAN 算法相比,本文算法在聚类效率上提升了 6.10%,识别速度提高了 5.26%。

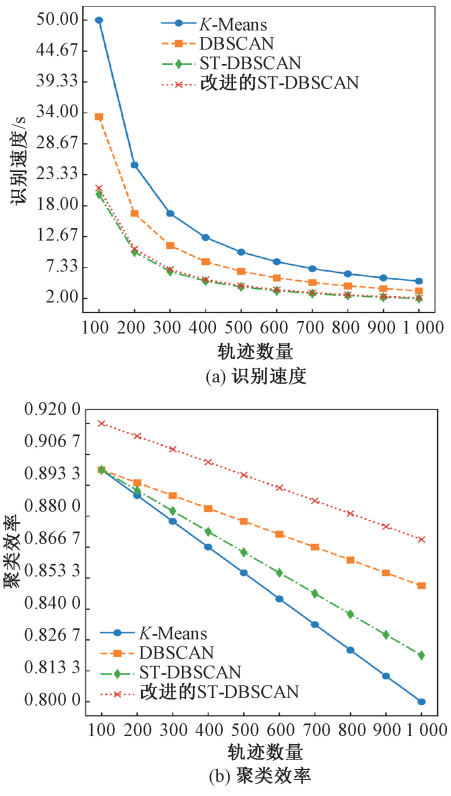

图 7　不同聚类算法的识别速度与聚类效率对比

Figure 7　Comparison of recognition speed and clustering efficiency across different clustering algorithms

2.4　动态 OD 矩阵构建

按照 1.3.1 节的路网构建方法对北京市路网进行拓扑建模,提取路网中关键交叉路口及路段端点,

并利用 ArcGis 进行标序,标序原则上以经纬度顺序从左到右、自下而上进行标注。北京市道路关键节点如图 8 所示。

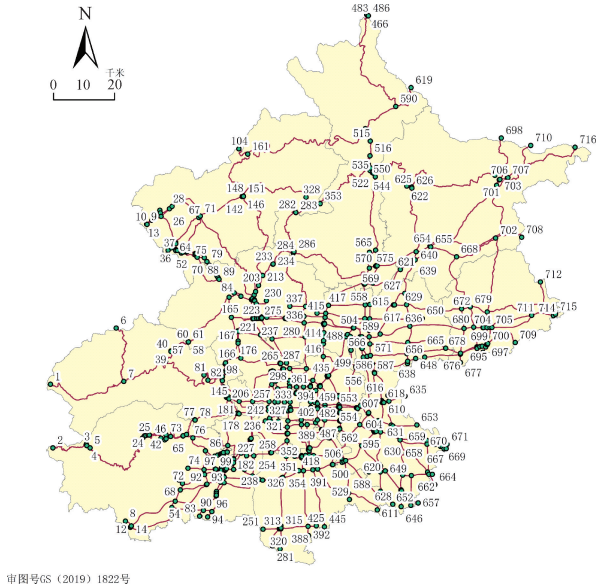

图 8 北京市路网关键节点图

Figure 8 Key nodes of the road network in Beijing

图 9 为研究区域内的居民每 2 h 出行时变动态 OD 热力图。城市居民出行时变动态 OD 热力图反映了居民在城市关键交通节点的出行需求,也反映了北京市居民通勤状况的变化情况。

从时间上看,OD 流在 00:00 至 04:00 未发生明显变化;从 04:00 起,大量居民进入城市中心,流量开始逐渐产生;从 06:00 至 24:00 一直处于较高的状态。周一至周四午高峰一般为 12:00 至 14:00,图 9 模拟日期为周五,14:00 至 16:00 的流量强度明显高于午高峰,而晚高峰则从 18:00 一直持续到 24:00,这也符合人们的一般活动规律。从空间上看,节点间 OD 流向大致相同,与较外围的行政区相比,处于北京市中心的行政区(如朝阳区与中心城区东北方向的顺义区、怀柔区,西北方向的昌平区,西南方向的房山区等)都有明显的出行趋势。

图 10 为北京市全天候居民出行 OD 发生量与吸引量的变化趋势。出行产生量高峰为 00:00 至 02:00,于 03:00 骤减,00:00 至 03:00 对应的 OD 吸引量无明显波动,可知 00:00 至 02:00 出发的 OD 流主要以长途旅程为主。在这段时间内,货车出行可能是导致 OD 流滞后的一个重要因素。由于货车通常需要在夜间或凌晨避开高峰时段进行长途运输,因此在这一时段的 OD 流中占据较大比例。由于路况、交通管制、休息时间等因素,货车在长途旅程中到达时间相对出发时间有一定的滞后性,因此长途旅程的 OD 流量会在图 10 中吸引量上升阶段逐渐被接收。

在城市交通流分析中,动态 OD 矩阵的获取是一项关键任务,其性能评价指标直接影响了交通管理和规划的有效性。在对比实验中,相对于基于统计方法的 OD 矩阵求解和基于路段流量二阶统计量

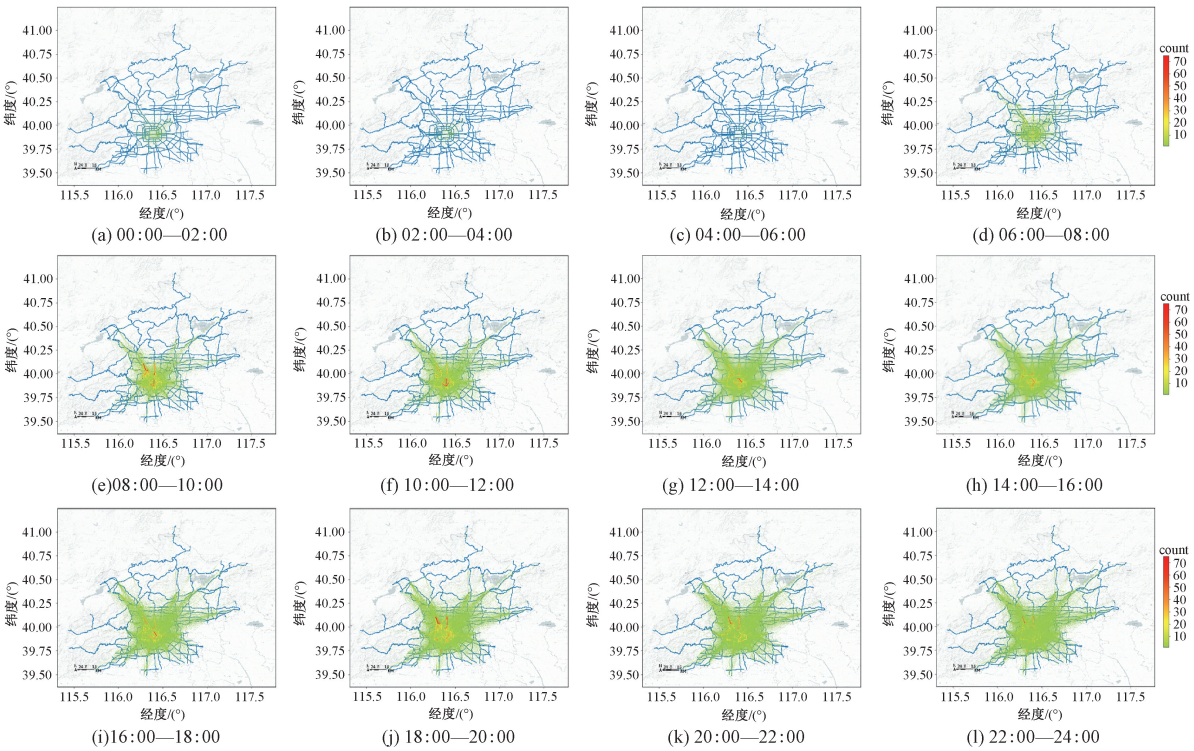

图 9 北京市分时动态 OD 热力图

Figure 9 Time-sharing dynamic OD heat map of Beijing

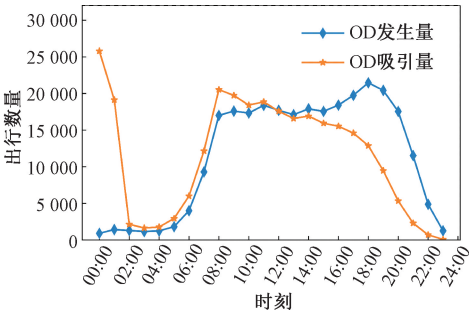

图 10　交通发生量与吸引量

Figure 10　Traffic generation and attraction

的 OD 矩阵估计,本文方法的均方误差 *MSE* 分别降低了 16.98% 和 21.55%。这表明本文方法不仅能更精确地捕捉 OD 矩阵,而且在大规模和复杂数据集上的适应能力较强。

图 11 为各方法的均方误差 *MSE* 对比。由图 11 可知,本文方法在精确捕获交通流动特性方面具有优势,在大规模和复杂数据集上的适应能力较强。

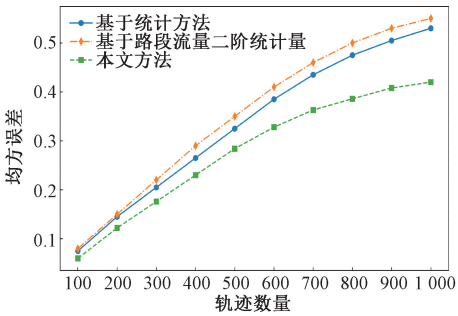

图 11　不同 OD 矩阵构建方法的 *MSE* 对比

Figure 11　Comparison of *MSE* for different OD matrix estimation methods

3　结论

本文针对交通调查数据获取时间跨度长、研究区域节点粒度粗糙等问题,提出了一种基于手机信令数据的城市居民动态 OD 提取方法。该方法从大量的 CDR 数据中提取居民出行轨迹的驻留点信息,并将 OD 信息与路网关键节点匹配,推导出动态 OD 矩阵。与传统方法相比,本文方法展现了显著的优势,为交通规划与管理提供了新的视角和工具。

下一步工作可以探索更有效的算法实现方式、并行计算等策略,以进一步提升算法性能。未来的研究方向还包括考虑城市交通运输系统的复杂性与多样性,探讨不同交通工具之间的出行特征差异,如公共交通、私家车、共享单车等。这将有助于进一步揭示城市居民出行偏好及其影响因素,可以使交通规划部门在制定策略时更全面地考虑居民出行需求,优化交通资源配置,提高城市交通效率。

参考文献:

[1] WU L L, YANG B, JING P. Travel mode detection based on GPS raw data collected by smartphones: a systematic review of the existing methodologies[J]. Information, 2016, 7(4): 67.

[2] PRELIPCEAN A C, GIDÓFALVI G, SUSILO Y O. Transportation mode detection: an in-depth review of applicability and reliability[J]. Transport Reviews, 2017, 37(4): 442-464.

[3] SHEN L, STOPHER P R. Review of GPS travel survey and GPS data-processing methods[J]. Transport Reviews, 2014, 34(3): 316-334.

[4] 刘华斌. 手机信令数据背景下城市交通出行方式选择辨识方法研究[D]. 北京: 北京交通大学, 2019.
LIU H B. Urban transportation modes recognition based on mobile signaling data[D]. Beijing: Beijing Jiaotong University, 2019.

[5] 苗壮. 基于手机信令数据的数据清洗挖掘与居民职住空间分析[D]. 成都: 西南交通大学, 2017.
MIAO Z. Research on data cleaning, mining, jobs and residential locations based on mobile phone signaling data[D]. Chengdu: Southwest Jiaotong University, 2017.

[6] 余锦斌. 基于手机信令的数据分析引擎设计与实现[D]. 南京: 东南大学, 2018.
YU J B. Design and implementation of analysis engine based on mobile phone data[D]. Nanjing: Southeast University, 2018.

[7] 李曙光. 用于动态交通分配的合理路径集合算法研究[J]. 郑州大学学报(工学版), 2009, 30(2): 125-128.
LI S G. Reasonable path set algorithm for dynamic traffic assignment[J]. Journal of Zhengzhou University (Engineering Science), 2009, 30(2): 125-128.

[8] WHITE J, WELLS I. Extracting origin destination information from mobile phone data[C]//Eleventh International Conference on Road Transport Information and Control. London: IET, 2002: 30-34.

[9] CACERES N, WIDEBERG J P, BENITEZ F G. Deriving origin-destination data from a mobile phone network[J]. IET Intelligent Transport Systems, 2007, 1(1): 15.

[10] ZHANG Y, QIN X, DONG S, et al. Daily O-D matrix estimation using cellular probe data[C]// Transportation Research Board 89th Annual Meeting. Washington DC: TRB, 2010.

[11] MAMEI M, BICOCCHI N, LIPPI M, et al. Evaluating origin-destination matrices obtained from CDR data[J]. Sensors, 2019, 19(20): 4470.

[12] FEKIH M, BONNETAIN L, FURNO A, et al. Potential of cellular signaling data for time-of-day estimation and

spatial classification of travel demand: a large-scale comparative study with travel survey and land use data[J]. Transportation Letters, 2022, 14(7): 787-805.

[13] 黄美灵, 陆百川. 基于手机定位的交通OD数据获取技术[J]. 重庆交通大学学报(自然科学版), 2010, 29(1): 162-166.
HUANG M L, LU B C. Traffic OD data collection technology based on mobile phone location[J]. Journal of Chongqing Jiaotong University (Natural Science), 2010, 29(1): 162-166.

[14] 魏玉萍, 韩印. 基于手机定位的交通OD获取技术[J]. 交通与运输(学术版), 2011(2): 33-36.
WEI Y P, HAN Y. Traffic OD date collection technology based on mobile phone[J]. Traffic & Transportation, 2011(2): 33-36.

[15] 蔡军, 刘锴, 刘涟涟. 基于VISUM模型的公交OD反推研究: 以西宁市为例[J]. 交通运输系统工程与信息, 2013, 13(1): 49-56.
CAI J, LIU K, LIU L L. Bus OD matrix estimation by VISUM model: case of Xining of China[J]. Journal of Transportation Systems Engineering and Information Technology, 2013, 13(1): 49-56.

[16] YU C, HE Z C. Analysing the spatial-temporal characteristics of bus travel demand using the heat map[J]. Journal of Transport Geography, 2017, 58: 247-255.

[17] GUO D S, ZHU X, JIN H, et al. Discovering spatial patterns in origin-destination mobility data[J]. Transactions in GIS, 2012, 16(3): 411-429.

[18] GONZÁLEZ M C, HIDALGO C A, BARABÁSI A L. Understanding individual human mobility patterns[J]. Nature, 2008, 453(7196): 779-782.

Dynamic OD Matrix Extraction Method of Urban Residents Based on Cell Phone Signaling Data

TIAN Zhao[1], ZHANG Qianzhong[1], ZHAO Xuan[1], CHEN Bin[2], SHE Wei[1], YANG Yanfang[3,4]

(1. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China; 3. China Academy of Transportation Sciences, Beijing 100029, China; 4. Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing 100029, China)

Abstract: Previous surveys of urban resident travel were hindered by prolonged durations and insufficient granularity in traffic zone divisions, which impeded the timely and accurate acquisition of travel data. To address this issue, this study proposed a method for extracting the dynamic origin-destination (OD) matrix of urban residents based on mobile phone signaling data. Firstly, methods to address two complex types of noise inherent in the signaling data: ping-pong switching data and drifting data were proposed. Specifically, a window threshold-based detection and equivalent location replacement method for ping-pong switching data was proposed, as well as a complex drift point detection and marking method for drifting data. Secondly, an enhanced ST-DBSCAN clustering algorithm was proposed, which incorporated a temporal isochronization method to integrate temporal and spatial information, enabling the identification of dwell points during travel. Finally, a road network with key nodes was established using geographic information system (GIS), aligning resident travel OD with the network nodes to effectively derive the dynamic OD matrix of urban residents. Experimental results showed that the enhanced ST-DBSCAN clustering algorithm outperformed the traditional ST-DBSCAN, improving clustering efficiency by 6.10% and identification speed by 5.26%. Furthermore, the dynamic OD matrix extraction method based on the enhanced ST-DBSCAN clustering algorithm achieved approximately 16.98% and 21.55% reductions in mean squared error compared to the conventional statistical methods and the second-order statistical methods, respectively. By applying the proposed dynamic OD matrix extraction method to the case of Beijing, this study was able to conduct timely and effective analyses of daily and peak travel patterns of urban residents.

Keywords: city travel; intelligent traffic system; cell phone signaling data; dynamic origin-destination matrix; dwell point identification; spatial-temporal characteristics analysis