基于居民出行特征的职住地精细化识别

黄兴如，李奕萱，刘中亮，冯瀚斌，王希昭，闫　龙，胡博文，李炫孜，李大中
（联通数字科技有限公司　数据智能事业部，北京　100010）

摘　要：为了解决传统职住模型测算规则的单一性和局限性，降低各区域居民用户因作息规律差异或临时性变化而造成的职住地识别误差，创新性提出一种基于不同区域居民出行特征的职住地精细化识别方法。首先，采用“3 min 切片”和“角度＋驻留时间＋连接次数”等多种方式对手机信令数据进行降噪提炼；然后，基于时空约束密度聚类进行驻留点识别分析；最后，根据各城市居民日常出行特征，通过引入加权驻留时长动态更新各城市区域居民用户职住地测算规则，进而精细化识别不同城市用户职住地分布。实验结果表明，所提方法涉及的过程均合理有效，且最终的职住地识别效果要明显优于传统单一职住模型测算规则，适用于同时批量处理多个区域职住地问题，尤其对因突发状况而产生作息时间变化的城市效果更为显著。

关键词：信令数据；出行特征；密度聚类；加权驻留时长；职住地识别
中图分类号：TP311；F299. 2　　**文献标识码：**A　　**DOI：**10. 19358/j. issn. 2097 – 1788. 2024. 08. 008
引用格式：黄兴如，李奕萱，刘中亮，等. 基于居民出行特征的职住地精细化识别［J］. 网络安全与数据治理，2024，43(8)：44 – 48.

Fine-grained identification method of home-work location based on travel characteristics of residents

Huang Xingru, Li Yixuan, Liu Zhongliang, Feng Hanbin, Wang Xizhao,
Yan Long, Hu Bowen, Li Xuanzi, Li Dazhong
(Data Intelligence Division, Unicom Digital Technology Co., Ltd., Beijing 100010, China)

Abstract: To address the simplicity and limitations of the traditional home-work model calculation rules and reduce the identification errors caused by differences in the daily routines of residents in various regions or temporary changes, this study proposed a fine-grained identification method of home-work location based on the travel characteristics of residents in different regions. Firstly, various methods such as "3-minute slicing" and "angle + stay time + connection frequency" are used to denoise and refine the mobile phone signaling data. Then, based on spatiotemporal constrained density clustering, stay points are identified and analyzed. Finally, according to the daily travel characteristics of residents in various cities, weighted stay duration is introduced to dynamically update the home-work calculation rules for residents in different city areas, thereby refining the identification of home-work distribution for users in different cities. Experimental results show that the processes involved in this method are reasonable and effective, and the final home-work identification results are significantly better than those of traditional single home-work model calculation rules. This method is suitable for batch processing of home-work problems in multiple regions simultaneously, particularly for cities where changes in routines are caused by unexpected events.

Key words: cellular signaling data; travel characteristics; DBSCAN; weighted stay duration; home-work location identification

0　引言

精准有效识别不同区域居民职住地以及挖掘居民处于职住地的时空规律可为城市职住规划、经济发展布局、公共资源分配和交通管理决策提供数据支持。手机信令数据具有覆盖广、延迟低、时效高、周期长等特点，因此借助手机信令位置数据进行居民活动分析研究具有良好的基础和开端，能够从大规模时空轨迹信息中挖掘居民的活动范围、出行时长、驻留兴趣点和出行方式等重

要时空属性特征[1-3]。

由于设备测量、计算方法、数据传输等因素影响，致使获得的轨迹数据多存有误差，而研究表明利用空间聚类算法将邻近的位置点进行聚合形成累计停留时间可减少该影响[4-5]。在此基础上，通过设置多日夜间和多日日间的驻留日长以及每日的最短驻留时长等指标，可对用户的居住地、工作地和惯常性活动点进行识别[6]。Zang 等[7]依据手机用户在自定义的职住时间段内分别产生的业务频繁程度来确定职住地。Isaacman 等[8]基于手机通话定位，通过空间聚类识别用户的重要活动地点，进而通过时间分析确定职住地。唐小勇等[9]提出一种职住计算框架，识别用户在一天内的多日稳定点和综合工作日与节假日稳定点，基于此来判断用户的职住地。张天然[10]利用每日 20：00 至次日 8：00 和工作日 9：00～18：00 的手机数据训练识别，将出现概率最高且超过 60% 的区域作为用户的职住地。可见当前职住地测算方法的基本原理是采用某种规则对居住、工作行为的时间、空间特征进行归纳测算。然而，上述方法中所设定的时间规则具有一定的局限性，并未兼顾到不同区域因地理位置、经济条件，甚至重大事件造成的各种作息时间差异，进而导致识别的用户职住地可能存在误差；尤其是疫情期间各区域居民职住地会存在不规律性变化[11]。

职住地测算的关键问题是如何解决手机信令数据的时间连续性（用户信令事件记录的时间间隔不固定），以时间特征作为识别规则的相关方法可以分类四种：累积时间法、特征时间法、信息熵法、时间阈值法[12]。现基于累积时间法、特征时间法和时间阈值法，提出一种适用于全国不同区域的职住地精细化识别方法，以消除不同区域用户因作息时间差异造成的职住地识别误差。

1 研究方法

1.1 手机信令数据预处理

用户信令位置数据具有存量大、离散化、噪声多、不完整等特点，为提高数据的利用率，准确提炼用户关键的停留和出行信息，需对其进行预处理。

（1）数据清洗：过滤原始数据中的空值、错误值、重复值以及移动终端识别码（MSID）不可获取的信令数据。

（2）数据整理：将每一个 MSID 的信令记录根据其发生时间戳进行升序排序。

（3）漂移数据处理[3]：如果用户的某个记录点在短时间内突然切换至较远的基站，则认为该点为数据漂移点。因此，若相邻记录的距离大于 5 km，且平均速度超过汽车行驶的最高车速（180 km/h），则剔除该记录。

（4）冗余数据处理：通常情况下，用户 3 min 内会产生多个信令数据，然而该时间段内活动轨迹有限，导致 3 min时间内出现多条重复数据。据统计，用户在 3 min 时间内连接过的基站通常不会超过 5 个，因此获取其在 3 min内连接次数最多的 3 个基站位置，根据此数据判定用户的大致所在位置。“3 min 切片”也在很大程度上保证了数据的时间和区域连续性。

（5）噪声数据处理：由于乒乓切换或漂移等误差，以及定位技术自身的局限性，导致了手机信令数据包含大量噪声。基于此，采用“角度 + 驻留时间 + 连接次数”方法进行该类噪声过滤（其中角度的计算借助余弦定理）。噪声数据通常会导致轨迹形成的角度很小，在每一个点的驻留时间很短，而且连接次数通常比较小，因此采用这三个条件相结合的方式可以更精准地过滤掉噪声数据。

1.2 驻留点识别

在轨迹点研究中，通常采用聚类的方法对轨迹点数据进行分析研究，一是可以降低数据自身的误差影响，二是可以获得直接分析不易得出的信息，如移动对象的常驻区域、活动时段等信息。对于轨迹点的聚类，通常有基于密度的聚类、基于网格的聚类以及基于模型的聚类等方法[13-14]。由于基于密度的聚类方法不仅能够识别出噪声点，还可以发现任意形状的簇类，不需要事先知道要形成的簇类的数量，故而被广泛应用于大规模数据聚类[15]。本研究采用的是具有代表性的基于密度聚类方法 DBSCAN（Density-Based Spatial Clustering of Applications with Noise），该聚类算法主要根据数据分布的紧密程度进行聚类，其涵盖的两个参数邻域样本阈值 MinPts 和距离阈值 Eps 决定了聚类效果。

驻留点聚类识别过程中，综合考虑移动轨迹时间和空间属性，维度涵盖经度、纬度和时间序列。由于信令数据分布离散、不均匀，且运动状态产生的信令数据明显多于静止状态，因此时间维度仅考虑顺序性，不考虑实际的时间间隔，将 Index 列按比例归一化，生成 Time_order 列。考虑到数据量较少情况下，容易导致经纬度距离很近的位置点因为时间维度的原因不能被聚类在一起，因此针对不同数据类进行差异处理：

（1）若数据量总条数大于等于 150 条，那么归一化方式采取 x/x_{max}，其中 x 为 Index 列数值，x_{max} 为最大数值。

（2）若数据量总条数小于 150 条，那么归一化的最小值为 $1-x_{max}\cdot\theta$，将归一化的值归于 $[1-x_{max}\cdot\theta, 1]$，其中 θ 设置为 0.006 6。

基于时空约束密度聚类的驻留点识别分析，使结果

更具合理性、科学性和准确性，避免由于移动过程中多次经过一个点而造成将该点标识为驻留点的情况。此外，考虑到移动中短暂驻留易被识别为驻留点，则结合连接次数和总驻留时长，针对上述聚类结果进行再次判断，将连接次数较小且总驻留时长较短的驻留点视为移动轨迹点。

1.3 不同区域职住地精细化推演

传统职住模型测算是依据工作日常规的作息规律，假定在日间8：00～18：00为工作时间，就业者主要驻留的位置即为工作地，其余两个时段0：00～8：00和18：00～24：00的主要驻留位置则被视作居住地[3]。而我国横跨5个时区，东西时差相差4小时，如若统一采用传统职住模型测算规则进行职住地识别，则无法反映出各个城市、时区居民职住作息时间的差异，具有一定的局限性。本研究基于各城市居民用户每小时平均移动轨迹长度，结合已有职住地测算方法累积时间法、特征时间法和时间阈值法，提出一种新的用户职住地识别方法，精细化识别不同城市用户职住地分布，以提高居民用户职住地识别的准确性。

根据各城市居民日常作息规律，计算每日居民用户每小时平均移动轨迹长度，将其归一化处理后得出城市居民当日每小时出行活跃度。基于城市当日24小时出行活跃度不难发现存在早高峰时段和晚高峰时段，将早晚高峰时段作为工作地和居住地的分界点，即早高峰～晚高峰期间为职地驻留时段，晚高峰～24：00和00：00～早高峰为住地驻留时段；此外，为兼顾到城市多数居民用户出行规律的差异性和一致性，降低职住地识别误差，现基于24小时出行活跃度进一步加工出每小时驻留权重系数（驻留权重系数 = 1 - 出行活跃度），通过引入加权驻留时长精准识别用户当日职住地。

根据上述职住时段划分规则和驻留权重系数，结合用户每日驻留点数据，针对当日职住地进行测算：工作日白天职地驻留时段期间，最大累计加权驻留时长且驻留时间大于某个阈值的驻留点位即为当日工作地；夜晚住地驻留时段期间，最大累计加权驻留时长且驻留时间大于某个阈值的驻留点位即为当日居住地。

2 实验与结果分析

2.1 实验数据集及预处理

为验证模型结果的有效性和准确性，抽取某运营商两个不同时区省份（新疆维吾尔自治区和北京市）的1 700万用户，利用该用户集2023年12月4日～2023年12月10日七天内的移动终端位置数据，其中每条记录包括用户唯一标识、事件发生省份/地市/区县、日期、时间、位置区域码（LAC）、基站编号（CID）、经度和纬度字段。由于移动手机终端与基站间信息交互过程中存在无线信号强度不稳定现象，致使系统在传输和存储等过程存有偏差，因此需对原始终端位置数据进行相应预处理，涉及操作步骤涵盖基础数据异常值清洗、事件发生时间戳排序、漂移数据剔除、冗余数据提炼和噪声数据处理，详情请见1.1节。

2.2 实验结果与分析

（1）时空停留点聚类

针对新疆维吾尔自治区和北京市两地每个用户每日的轨迹数据集合，执行基于密度的DBSCAN时空停留点聚类过程，通过轨迹停留点与其所属类簇中心点的距离分布，验证聚类结果的合理性。由图1可以看到，91.96%的停留点与聚类中心点距离小于1 km；同时对比分析常规轨迹数据预处理方法[3]，应用本研究数据，密度聚类结果表明85.24%的停留点与聚类中心点距离小于1 km，进一步证明了当前轨迹数据预处理方法的合理性和有效性。

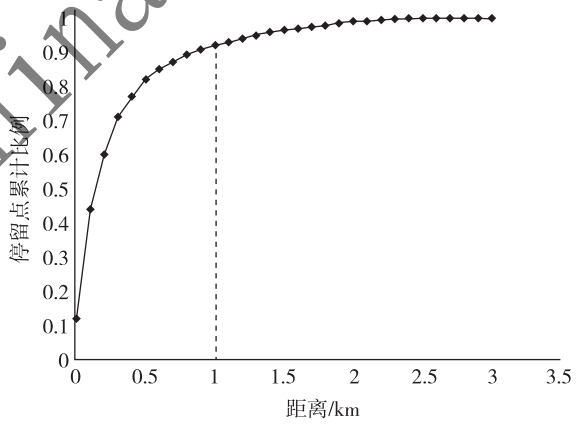


图1 停留点和类簇中心点距离累积分布图

（2）职住地识别

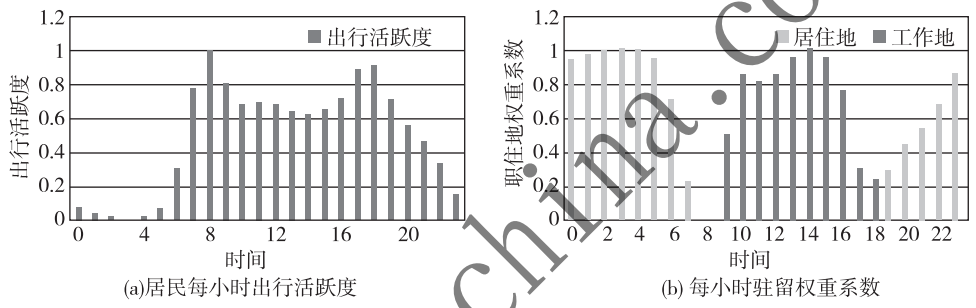
针对上述重要停留点聚类过程中生成的每个用户的每个重要停留点簇，以簇中心点代表该类簇所属停留点，其停留时间为簇中停留点的累计。基于此，计算用户每日各簇中心点（驻留点）累计加权驻留时间：首先，基于某城市当日居民用户每小时平均移动轨迹长度，通过1.3节中方法加工出该城市当日每小时驻留权重系数（图2（a）为某城市居民当日每小时出行活跃度，图2（b）为某城市居民当日每小时驻留权重系数）；然后，统计出该城市每个用户当日在各驻留点每小时驻留时间，基于各驻留点每小时驻留时间和对应的权重系数计算出每个用户在各驻留点每小时加权驻留时间；最后，通过驻留点每小时加权驻留时间统计出该驻留点累计加权驻留时间。通过上述各驻留点累计加权驻留时间，结合累积时

间法、特征时间法和时间阈值法测算用户当日职住地，即工作日白天职地驻留时段期间，最大累计加权驻留时长且驻留时间大于 2 h 的驻留点位即为当日工作地；夜晚住地驻留时段期间，最大累计加权驻留时长且驻留时间大于 4 h 的驻留点位即为当日居住地。

基于上述不同区域居民用户职住地精细化测算结果，通过用户在职住地每小时累计停留时间，对比分析传统单一职住模型测算规则[3]，图 3 展示了不同方法职住地停留时间分布情况。从图中可以看出，本研究方法的用户职住地整体停留时间分布明显要优于传统单一职住地测算规则模型，尤其是新疆维吾尔自治区，更加符合居民用户的日常作息规律。

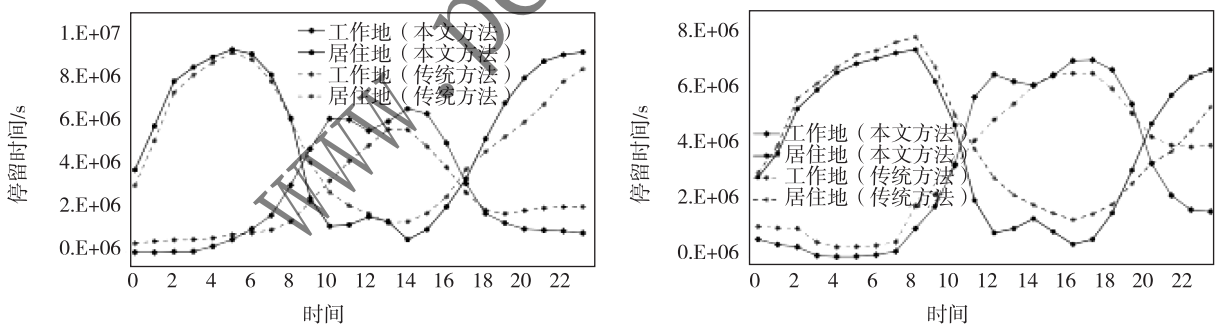
此外，对比分析运营商通过无线网络定位技术解析的部分用户职住地结果，发现 90.43% 的用户职地和 93.62% 的用户住地分别与本文精细化测算结果（职地/住地）距离小于 1 km，而 81.64% 的用户职地和 83.22% 的用户住地分别与传统方法测算结果（职地/住地）距离小于 1 km，进一步证实了本研究方法的合理有效性。

传统职住地测算方法较为单一，无法同时批量处理全国各区域（如新疆、青海、北京、黑龙江等不同时区省市）职住地问题；尤其是疫情期间，部分城市由于区域性临时政策出台造成城市作息时间的差异性变化，致使职住地识别更加困难。本研究方法基于各城市居民用户出行特征，通过引入加权驻留时长动态更新各城市区域居民用户职住地测算规则，进而精细化识别不同城市用户职住地分布，相较于传统职住地测算方法更具灵活性和准确性，尤其对于突发状况而产生作息时间变化的城市效果更为显著。



(a)居民每小时出行活跃度　(b) 每小时驻留权重系数

图 2　某城市居民当日每小时出行活跃度和权重系数



(a)北京市　(b)新疆维吾尔自治区

图 3　不同时区省市用户职住地整体停留时间分布

3　结论

依据各城市居民用户作息规律，结合已有职住地测算方法累积时间法、特征时间法和时间阈值法，创新性提出一种基于不同区域居民用户出行特征的职住地精细化识别方法，过程内容包括多方式数据预处理、基于时空约束密度聚类的驻留点识别分析和不同区域职住地精细化推演等。首先，采用多种方式对手机信令数据进行降噪提炼，涉及漂移数据处理、3 min 切片冗余数据剔除以及利用“角度 + 驻留时间 + 连接次数”方法进行噪声过滤等步骤，以确保数据准确率，提高数据的利用率。然后，基于时空约束密度聚类进行驻留点识别分析，过程中综合考虑移动轨迹时间和空间属性，维度涵盖经度、纬度、时间序列和停留时间，促使结果更具合理性、科学性和准确性，避免由于移动过程中多次经过一个点或短暂驻留而造成将该点标识为驻留点的情况。最后，基于各城市居民用户出行特征，通过引入加权驻留时长精细化识别不同城市用户职住地分布，以提高居民用户职住地识别的准确性。实验结果表明，文中方法的每个过

程都是合理有效的，并且最终的职住地识别效果要优于传统单一职住模型测算规则，适用于同时批量处理多个区域（如新疆、青海、北京、黑龙江等不同时区省市）职住地问题；且能够动态更新职住地模型测算规则，尤其对于因突发状况而产生作息时间变化的城市效果更为显著。

参考文献

[1] WANG Y, YOU Y, HUANG J, et al. Differences in urban daytime and night block vitality based on mobile phone signaling data: a case study of Kunming's urban district [J]. Open Geosciences, 2024, 16 (1): 116 - 127.

[2] 孙世超，吕豪. 大数据环境下基于职住地识别的公交通勤行为判断与特征分析 [J]. 上海海事大学学报，2023，44 (4)：45 - 50.

[3] 刘鹏，林航飞. 基于手机信令数据的职住地识别方法 [J]. 综合运输，2022，44 (5)：14 - 17，33.

[4] 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.

[5] 吴子啸. 基于手机数据的出行链推演算法 [J]. 城市交通，2019，17 (3)：11 - 18，83.

[6] 宋少飞，李玮峰，杨东援. 基于移动通信数据的居民居住地识别方法研究 [J]. 综合运输，2015，37 (12)：72 - 76.

[7] ZANG H, BOLOT J. Anonymization of location data does not work: a large-scale measurement study [C]// Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, 2011: 145 - 156.

[8] ISAACMAN S, BECKER R, CÁCERES R, et al. Identifying important places in people's lives from cellular network data [C]// Pervasive Computing: 9th International Conference, 2011. Springer, Berlin, Heidelberg, 2011: 133 - 151.

[9] 唐小勇，周涛，陆百川，等. 一种基于手机信令的通勤OD训练方法 [J]. 交通运输系统工程与信息，2016，16 (5)：64 - 70.

[10] 张天然. 基于手机信令数据的上海市域职住空间分析 [J]. 城市交通，2016，14 (1)：15 - 23.

[11] 李阳阳. 疫情背景下通勤者职住地选择变化及影响因素分析 [D]. 北京：北京交通大学，2023.

[12] 钮心毅，谢琛. 手机信令数据识别职住地的时空因素及其影响 [J]. 城市交通，2019，17 (3)：19 - 29.

[13] 夏璠. 轨迹大数据驱动的人类移动行为挖掘与建模 [D]. 济南：山东大学，2023.

[14] 王兆丰，吴杨. 基于改进k-均值算法的未知协议比特流聚类 [J]. 计算机应用，2016，36 (S1)：5 - 8.

[15] 李建邺. 基于手机信令数据的职住地获取研究 [D]. 南京：东南大学，2019.

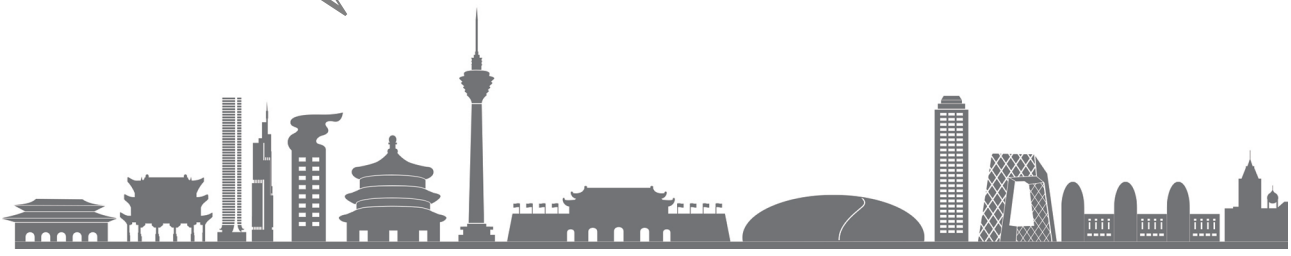
（收稿日期：2024 - 06 - 19）

作者简介：

黄兴如（1988 - ），男，硕士，主要研究方向：数据治理。

李奕萱（1993 - ），男，硕士，主要研究方向：数据科学、数据治理、人工智能。

闫龙（1994 - ），通信作者，男，博士，主要研究方向：数据科学、数据治理、数据安全及相关领域应用。E-mail：yanl17@chinaunicom. cn。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部