引文格式：高蕴灵,李英冰,何阳,等. 出租车OD序列聚类的城市功能区识别算法研究[J]. 地理空间信息, 2024, 22(2): 8-12. doi:10. 3969/j.issn. 1672-4623.2024.02.002

出租车OD序列聚类的城市功能区识别算法研究

高蕴灵[1]，李英冰[1*]，何　阳[2]，栾梦杰[1]，李欣然[1]
（1. 武汉大学 测绘学院，湖北 武汉 430070；2. 中铁第一勘察设计研究院集团有限公司，陕西 西安 710043）

摘　要：提出一种由添加窗口约束的规整路径距离改进的K中心聚类算法（LDTW-K-medoids），通过构造时间序列、动态时间规整、时间序列聚类、聚类结果解读的流程来识别城市功能区。将算法应用在纽约市城市功能区识别中，对50万条出租车OD数据进行处理，将265个交通小区划分为五大功能区，并结合出租车全局流动模式、建筑物面积指数和富集指数来定量识别功能性质。与谷歌地图和卫星影像对比的结果表明，算法能够有效识别城市功能区，总体精度为83.8%。

关键词：城市功能区；OD序列；动态时间规整；规整路径距离；K中心聚类；建筑物数据
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2024）02-0008-05

Urban Functional Regions Recognition Algorithm for Taxi OD Series Clustering

GAO Yunling[1], LI Yingbing[1], HE Yang[2], LUAN Mengjie[1], LI Xinran[1]
(1. School of Geodesy and Geomatics, Wuhan University, Wuhan 430070, China;
2. China Railway First Survey and Design Institute Group Co., Ltd., Xi'an 710043, China)

Abstract: In this paper, we provided a method that using warp path distance with window limits to improve K-medoids clustering algorithm (LDTW-K-medoids), which could identify urban functional regions by constructing orientation-destination (OD) series, using limited dynamic time warping, time series clustering and interpretation of clustering results. We applied the algorithm to identify urban functional regions in New York City, used 506 625 taxi OD records to divide 265 traffic districts into five functional regions, and combined the global flow pattern of taxis, building area index and enrichment index to quantitatively identify the functional properties. Compared with Google maps and satellite images, the result shows that the algorithm can effectively identify urban functional regions, with overall accuracy of 83.8%.

Key words: urban functional region, OD series, dynamic time warp, warp path distance, K-medoids clustering, building data

出租车上下车（OD）数据记录了居民在城市不同区域的流动信息[1]，常用于城市交通运行分析、居民出行行为分析[2-5]、城市土地利用识别[1,6]、城市功能区划分[7]。规整路径距离（warp path distance，WPD）是利用动态时间规整（dfynamic time warping，DTW）得到的2条时间序列间的最优匹配路径距离，能够衡量OD序列间的相似度，被应用到城市功能区识别的研究中[7]。

本文以交通小区为研究单元，利用出租车OD数据构建各交通小区OD时间序列，通过添加窗口约束的动态时间规整算法（LDTW）来求解不同交通小区序列间的规整路径距离，结合K中心聚类算法划分城市功能区，识别在出租车不同流动模式下的城市功能区属性。

1　研究方法

针对DTW算法在进行高维序列运算中出现的时间复杂度高和病态对齐的现象，本文通过关键时间节点对齐来添加窗口约束，从而求解出租车OD序列间的规整路径距离，并用该距离作为K中心聚类算法的规则来进行城市功能区的划分。由于城市的功能并不总是稳定的，城市区域在不同时段的活动性可能导致其社会功能发生阶段性转变[6]，因此区分了出租车在工作日和非工作日不同的全局流动模式，并综合建筑物分布情况来定量识别功能区属性。最后对比谷歌地图和卫星影像来检验算法在城市功能识别中的有效性（图1）。

1.1　基于LDTW-K-medoids的功能区划分

DTW算法通过有界性、连续性、单调性这3个限制条件来实现全局最优匹配，但没有兼顾序列的局部结构信息，在高维序列计算中容易导致对齐路径长度过大，出现病态对齐现象[8]。添加窗口约束的动态时

收稿日期：2022-09-11。
项目来源：国家重点研发计划项目（2018YFC0807000）。
第一作者简介：高蕴灵（2000—），硕士研究生，研究方向为时空大数据分析，E-mail：ybli@whu.edu.cn。
通信作者：李英冰（1977—），博士研究生，副教授，研究方向为空间大数据分析，E-mail：ybli@whu.edu.cn。

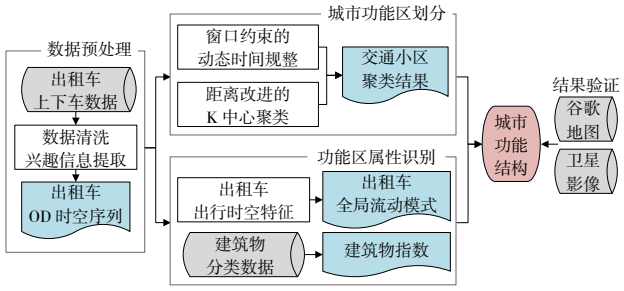


图1　算法流程图

间规整算法规定了序列中的数据点参与对齐的次数以及与其对齐数据点的下标范围，不仅有效减少了病态对齐的现象，还降低了时间成本[9]。因此，本文提出了一种适用于出租车OD序列的LDTW算法来求解不同交通小区序列间的规整路径距离。

针对每个交通小区，区分工作日与非工作日，统计一定时间间隔 t 小时内各小区的上车乘客数（O）和下车乘客数（D），组合得到 m 个交通小区的 $4n$ 维等时间间距OD序列：

$$\mathrm{OD}_j=\{O_W^{1j},\cdots,O_W^{ij},\cdots,O_W^{nj},O_R^{1j},\cdots,O_R^{ij},\cdots,O_W^{nj},D_W^{1j},\cdots,D_W^{ij},\cdots,D_W^{nj},D_R^{1j},\cdots,D_R^{ij},\cdots,D_R^{nj}\} \tag{1}$$

式中，$n=24/t$，OD_j 为第 j 个交通小区的等时间间距OD序列；O_W^{ij} 和 D_W^{ij} 分别为第 j 个交通小区工作日第 i 个时间段的上车乘客数和下车乘客数；O_R^{ij} 和 D_R^{ij} 分别为第 j 个交通小区非工作日第 i 个时间段的上车乘客数和下车乘客数。本文取 $t=1$。

在原有限制条件的基础上，添加如下的窗口约束来限制对齐路径长度：

1）区分工作日和非工作日，即对齐2条序列的 $\{O_W^{nj},O_R^{nj},D_W^{nj},D_R^{nj}\}$ 4个节点；

2）关键时间点（每天的6:00、12:00、18:00）需要对应，即对齐2条序列的 $\{O_W^{\frac{n}{4}j},O_W^{\frac{n}{2}j},O_W^{\frac{3n}{4}j},O_R^{\frac{n}{4}j},O_R^{\frac{n}{2}j},O_R^{\frac{3n}{4}j},D_W^{\frac{n}{4}j},D_W^{\frac{n}{2}j},D_W^{\frac{3n}{4}j},D_R^{\frac{n}{4}j},D_R^{\frac{n}{2}j},D_R^{\frac{3n}{4}j}\}$ 12个节点。

利用上述方法计算 m 个交通小区的OD时间序列两两之间的规整路径距离，最终得到 m 阶规整路径距离矩阵如式（2）。

聚类作为挖掘大规模时空数据的重要手段而被广泛应用于城市功能分区中[10]。K中心聚类是围绕中心点的聚类算法之一，相比于常用的K均值算法，该算法的簇中心点是实际对象而不是均值，因此消除了对孤立点的敏感性。一般K中心聚类过程中用欧氏距离

$$\mathrm{WPD_matrix}=\begin{bmatrix}0&\mathrm{WPD}_{1,2}&\mathrm{WPD}_{1,3}&\cdots&\mathrm{WPD}_{1,m-1}&\mathrm{WPD}_{1,m}\\ \mathrm{WPD}_{2,1}&0&\mathrm{WPD}_{2,3}&\cdots&\mathrm{WPD}_{2,m-1}&\mathrm{WPD}_{2,m}\\ \mathrm{WPD}_{3,1}&\mathrm{WPD}_{3,2}&0&\cdots&\mathrm{WPD}_{3,m-1}&\mathrm{WPD}_{3,m}\\ \vdots&\vdots&\vdots&&\vdots&\vdots\\ \mathrm{WPD}_{m-1,1}&\mathrm{WPD}_{m-1,2}&\mathrm{WPD}_{m-1,3}&\cdots&0&\mathrm{WPD}_{m-1,m}\\ \mathrm{WPD}_{m,1}&\mathrm{WPD}_{m,2}&\mathrm{WPD}_{m,3}&\cdots&\mathrm{WPD}_{m,m-1}&0\end{bmatrix} \tag{2}$$

来衡量对象间的距离，但本文旨在通过比较不同小区OD序列间的相似度来达到功能分区的目的，因此用规整路径距离来代替欧氏距离进行计算。

聚类结果评价采用轮廓系数。轮廓系数能够反映聚类结果的内聚度和分离度，越接近1表示聚类效果越好[11]，其计算公式为：

$$S=\frac{1}{n}\sum_{i=1}^{n}\frac{y(i)-x(i)}{\max\{x(i),y(i)\}} \tag{3}$$

式中，$x(i)$ 为对象 i 到它所在类中其他对象的平均距离；$y(i)$ 为对象 i 到其他类中所有对象的平均距离。

1.2　不同流动模式下的功能属性识别

出租车上下车事件在不同时段具有不同的分布规律[3]，而居民出行的变化导致功能区功能也发生改变[12]。因此结合出租车不同时段的全局流动模式来讨论功能区属性变化。净流量比是指在一定时期内，其他区域居民净流入某一区域的比例，能够反映出租车出行的全局流动模式[2]。其计算公式为式（4）。

$$\mathrm{NFR}_{ij}=\frac{D_i-O_i}{O_i+D_i} \tag{4}$$

式中，NFR_{ij} 为第 j 个交通小区在时段 i 的净流量比；O_i 为时段 i 内的流入量；D_i 为时段 i 内的流出量；$\mathrm{NFR}_{ij}>0$ 为区域热度增强，$\mathrm{NFR}_{ij}<0$ 为区域热度降低。

城市中分布密集的建筑物体现了城市不同区域的功能性质，构造建筑物面积指数和富集指数能够定量识别城市功能性质。建筑物面积指数反映了单元中主要分布的建筑物类别，建筑物富集指数反映了该建筑物类别在某单元中的聚集程度。功能区的功能由建筑物面积指数最大的2~3个类别决定，建筑物富集指数用于对比不同功能区的功能差异。对于聚类得到的每个功能区单元，计算公式分别为：

$$F_{i,c}=\frac{n_{i,c}}{S_C} \tag{5}$$

$$R_{i,c}=\frac{n_{i,c}/N_c}{N_i/N} \tag{6}$$

式中，$F_{i,c}$ 为聚类单元 c 中建筑物类别 i 的面积指数；$R_{i,c}$ 为聚类单元 c 中建筑物类别 i 的富集指数；$n_{i,c}$ 为聚类单元 c 中建筑物类别 i 的总面积；N_c 为聚类单元 c 中所有建筑物的总面积；N_i 为研究区域建筑物类别 i 的总面积；N 为研究区域所有建筑物的总面积。

2 纽约市实例计算

以纽约市265个交通小区为研究单元，将LDTW-K-medoids算法应用于纽约市功能分区，利用工作日和非工作日约50万条出租车上下客数据将交通小区划分为5类；结合建筑物分类数据和出租车不同时段的全局流动模式来识别功能区属性。

2.1 研究区概况与数据来源

纽约市是美国第一大城市，位于美国纽约州东南部大西洋沿岸。纽约市作为典型发达城市，其城市功能发展完善，且具有丰富的开放数据，适合用于城市功能结构的研究。研究区包括纽约市的265个交通小区，总面积为785.11 km^2，包括布朗克斯区（Bronx）、布鲁克林区（Brooklyn）、曼哈顿（Manhattan）、皇后区（Queens）、斯泰登岛（Staten Island）5个行政区。

2.1.1 出租车OD数据

出租车OD数据下载于纽约市出租车与豪华轿车委员会（Taxi & Limousine Commission，TLC）官方网站。对数据集进行预处理，包括数据清洗和兴趣信息提取。数据清洗包括删除上车时间晚于下车时间、行程距离小于零、乘客数量缺失等不合理的数据；兴趣信息包括上下车时间、上下车交通小区编号、乘客数。

2.1.2 建筑物分类数据

建筑物分类数据下载于纽约城市规划部（department of city planning，DCP）官方网站。DCP依据建筑物用途将建筑物分为11个类别：独栋或双拼别墅、多户无电梯住房、多户电梯住房、商务办公建筑、商住两用建筑、工业制造建筑、广场及户外游憩、公共设施和机构、公共交通设施、停车设施、闲置地。

2.2 探索性数据分析

出租车区域交互模式如图2a所示。出租车交互集中在布鲁克林区、曼哈顿区、皇后区。皇后区和曼哈顿区之间频繁的流动主要来源于肯尼迪国际机场和拉瓜迪亚机场。

出租车出行量年月分布如图2b所示。自2010年以来的12 a间，出租车出行量呈下降趋势。2010—2015年，出租车出行量年间变化较为稳定；2015年后，受Uber打车加入市场的影响，出租车每年出行量逐渐减少；此外，2020年初受到新冠疫情的影响，出行量急剧下降，而后保持较低水平。

出租车出行量的周内日变化特征和日内时变化特征分别如图2c和2d所示，统计范围为2019年3月1日至2019年3月31日。在一周内，周四、周五的出行量较大，周日最少。而在1 d天内，出行量在1:00~6:00持续下降，在7:00~9:00快速增长，而后持续缓慢增长，在17:00时略微下降，在18:00~20:00出现高峰，夜间行程量仍然较多，到后半夜逐渐回落。

选取2019年3月7日（周四）和2019年3月10日（周日）2 d共计506 652条出行记录，计算每天24个时间段内各交通小区的净流量比，可视化如图3所示，编号1表示0:00~1:00时段，编号2表示1:00~2:00时段，以此类推。红色表示小区热度增强，呈流入状态；蓝色表示小区热度减弱，呈流出状态。

在同一时段内，非工作日交通小区的热度往往比工作日更低。一些交通小区在工作日的7:00~9:00表现出明显的高流出状态，18:00~20:00表现出明显的高流入状态，而这一现象在非工作日表现不明显。推测7:00~9:00为上班高峰期，18:00~20:00为下班高峰期。

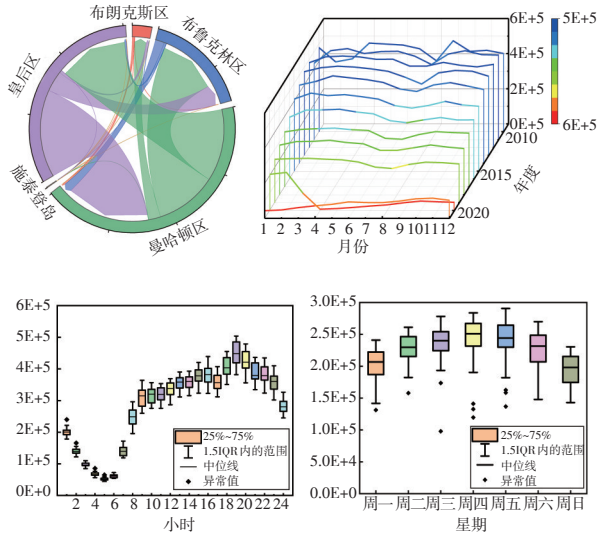


图2 出租车出行时空特征

2.3 城市功能识别

利用LDTW算法得到265个交通小区OD序列间的规整路径距离矩阵，将其作为K中心聚类的距离矩阵，可视化如图4所示。聚类数目为2、3、4、5、

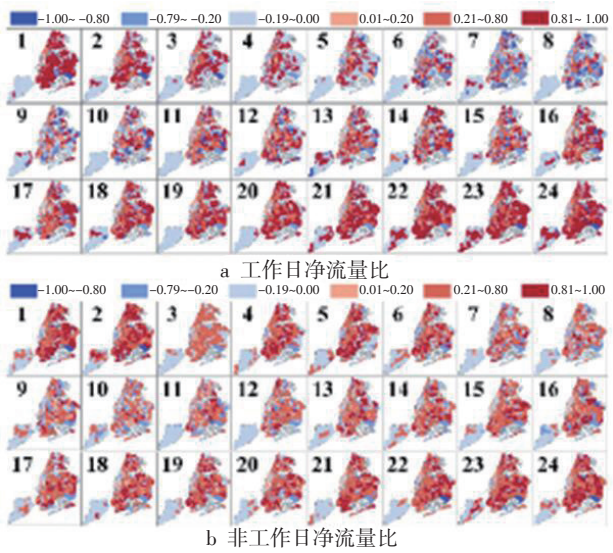


图3　出租车不同时段的全局流动模式

6、7、8、9时，对应的轮廓系数分别为0.850、0.733、0.641、0.405、0.215、0.176、0.325、0.130。

随着聚类数目的增大，轮廓系数呈降低趋势；而聚类数目过少时，城市功能结构划分不明确[7]。综合考虑轮廓系数、聚类合理性以及建筑物分布情况，取$k=5$时的聚类结果进行功能识别，可视化如图5所示。

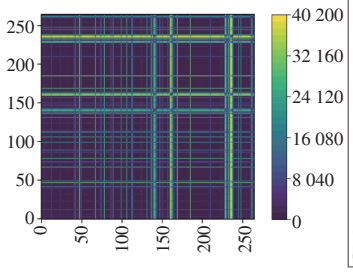


图4　OD序列间的规整路径距离

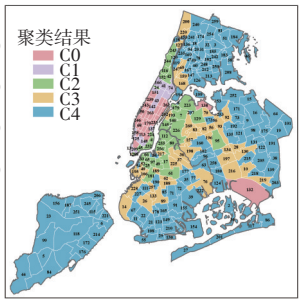


图5　城市功能区划分结果

5个类别的面积占研究区面积的比例分别为5.58%、2.01%、9.11%、17.07%、66.23%，每个类别中不同用途的建筑物的面积指数和富集指数的计算结果如表1所示。依据不同流动模式下的功能属性识别方法进行分析，整体来看，纽约市城市功能结构表现为以曼哈顿为中心的商业–工作–居住–游憩的圈层结构，城市功能完善，交通发达，商业发展好，且生活服务、休闲娱乐等基本设施分布广泛。

C0是中心商务区和交通枢纽，该区域交通发达，人流量最大。在工作日，流入集中在上班高峰期，流出集中在下班高峰期，区域表现为工作功能；在非工作日，流入集中在上午10:00时以后，流出集中在晚上10:00时以后，区域表现为商业功能。

C1是以居住为主、商业为辅的市中心居住区，位于市中心周围，区域流量仅次于C0，少量分布有大学、医院、博物馆、教堂等建筑。区域中多户电梯住房和商住两用建筑的富集指数最高。在工作日，区域表现为居住功能；在非工作日，白天和夜间都有一定的流入，区域表现为商业功能。

C2是工作性质和居住性质的混合功能区，包含大量的工业制造建筑和停车设施。作为工作区和居住区的混合区，该区域热度一直较高，白天略高于深夜，非工作日的热度明显低于工作日。

C3是以居住为主的功能区。区域内住房多为独栋或双拼别墅，混合少量的无电梯住房。区域还包含森林公园、体育馆等户外游憩地点。该区域在工作日上班高峰期有大量流出，下班高峰期大量流入，表现出明显的居住区特性。

C4是以游憩和居住为主的功能区，靠近城市边缘。C4和C3有相似的人员流动特性，但C4中与商业、工业相关建筑的富集指数都远低于C3。区域内商用建筑和交通设施的富集指数最低，住房更偏向于独栋或双拼别墅。

3　结果分析与评价

为了检验算法应用于城市功能区识别的效果，结合谷歌地图和卫星影像数据，将人工判别得到的城市功能作为真值，与算法识别结果进行比较，验证研究得到的纽约市功能分区结果的准确性，对照示例如图6所示。

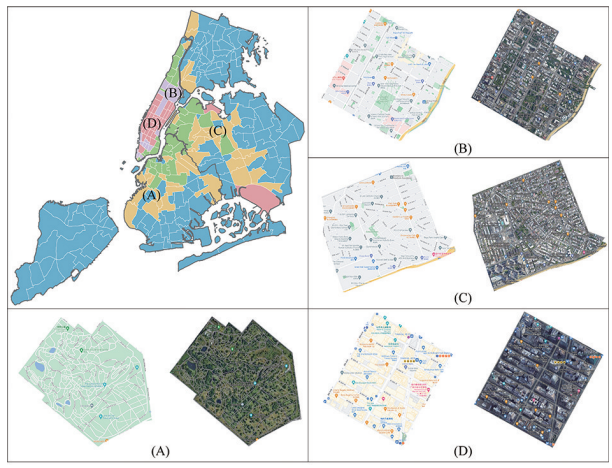


图6　识别结果对照示例

通过属性信息查询及目视解译，建立混淆矩阵定量评价识别精度，混淆矩阵如表2所示。算法的总体精度为83.8%，识别精度高，其中部分C1被错分到C0，原因可能是C0和C1都属于城市高度发达地区，功能混合度高，人流量大。综合考虑纽约市高度混合的用地现状，算法有效识别了纽约市城市功能区。

表1 建筑物面积指数和富集指数

建筑物用途	建筑物面积指数/%					建筑物富集指数				
	C0	C1	C2	C3	C4	C0	C1	C2	C3	C4
独栋或双拼别墅	1.20	0.95	15.36	28.46	35.31	0.04	0.03	0.52	0.96	1.19
多户无电梯住房	3.59	5.64	16.28	11.39	6.32	0.46	0.72	2.07	1.45	0.81
多户电梯住房	7.91	16.90	11.56	7.89	3.84	1.39	2.97	2.03	1.39	0.67
商住两用建筑	10.23	13.60	10.82	5.98	2.02	2.45	3.26	2.59	1.43	0.49
商务办公建筑	9.40	8.33	5.95	4.77	3.66	2.10	1.86	1.33	1.06	0.82
工业制造建筑	0.26	0.44	9.69	6.54	2.34	0.08	0.12	2.76	1.86	0.67
公共交通设施	57.61	3.74	9.67	7.17	5.36	6.28	0.41	1.05	0.78	0.58
公共设施和机构	4.90	10.94	7.82	8.45	8.57	0.59	1.32	0.94	1.02	1.03
广场及户外游憩	2.98	36.57	7.31	14.40	24.70	0.15	1.79	0.36	0.71	1.21
停车设施	0.67	1.30	2.60	2.41	1.03	0.49	0.94	1.88	1.74	0.75
闲置地	1.24	1.58	2.94	2.54	6.85	0.23	0.30	0.55	0.48	1.28

表2 功能区识别混淆矩阵/%

真实功能	识别功能				
	C0	C1	C2	C3	C4
C0	81.8	42.1	9.1	0.0	2.2
C1	18.2	52.6	0.0	0.0	0.0
C2	0.0	0.0	84.8	24.4	3.7
C3	0.0	5.3	3.0	71.1	1.5
C4	0.0	0.0	3.0	4.4	92.6

4 结 语

本文利用出租车OD数据和建筑分类数据，识别了纽约市的城市功能结构，并结合谷歌地图和卫星影像数据，验证了结果的准确性。文章的主要工作与创新为：

1）提出一种添加窗口约束的动态时间规整改进的K中心聚类算法来划分城市功能区，并区分两工作日和非工作日，识别了在出租车不同流动模式下的城市功能区主题。

2）将算法应用于纽约市，对50万条出租车OD数据进行处理，发现了纽约市以曼哈顿为中心的商业–工作–居住–游憩的城市功能圈层结构，算法识别精度高达92.6%。

算法准确性受到原始交通小区划分的影响，并且除了出租车以外，还有公交、地铁等诸多交通工具承载了居民日常出行，后续研究可以考虑结合多源数据，细化交通小区地块，进一步挖掘城市功能分布特征。

参考文献

[1] 葛盼盼. 基于出租车轨迹数据的城市土地利用类型识别方法研究[D]. 南京:南京大学,2019

[2] 张岩,李英冰,郑翔,等. 基于出租车数据的轨迹网络与流动模式研究[J]. 地理空间信息,2021,19(7):1-5

[3] 唐炉亮,郑文斌,王志强,等. 城市出租车上下客的GPS轨迹时空分布探测方法[J]. 地球信息科学学报,2015,17(10):1 179-1 186

[4] 赖桂君. 城市功能区单元尺度下的深圳市居民出租车出行时空特征研究[D]. 广州:广州大学,2022

[5] Yang J S. Understanding Collective Human Mobility Spatiotemporal Patterns on Weekdays from Taxi Origin-destination Point Data[J]. Sensors, 2019, 19(12): 2 812-2 815

[6] Pang G, Qi G, Wu Z, et al. Land-use Classification Using Taxi GPS Traces[J]. IEEE Transactions on Intelligent Transportation Systems,2013,14(1):113-123

[7] 姚尧,张亚涛,关庆锋,等. 使用时序出租车轨迹识别多层次城市功能结构[J]. 武汉大学学报(信息科学版),2019,44(6):875-884

[8] 郝石磊,王志海,刘海洋. 基于局部梯度和二进制模式的时间序列分类算法[J]. 软件学报,2022,33(5):1 817-1 832

[9] Sakoe H, Chiba S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition[J]. IEEE Trans. on Acoustics,Speech and Signal Processing,1978,26(1): 43-49

[10] 宁鹏飞,万幼,沈怡然,等. 基于签到数据的城市热点功能区识别研究[J]. 测绘地理信息,2018,43(2):110-114

[11] Rousseeuw P J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis[J]. Journal of Computational and Applied Mathematics, 1987(20): 53-65

[12] 王润泽,周鹏,潘悦,等. 基于大数据的城市功能区人口时空聚散模式研究[J]. 地理与地理信息科学,2022,38(1):45-50