文章编号:1009-6744(2019)02-0238-09 中图分类号:U491.1 文献标志码:A
DOI:10.16097/j.cnki.1009-6744.2019.02.034

基于出租车GPS大数据的城市热点出行路段识别方法

曲昭伟,王 鑫,宋现敏*,夏英集,袁咪莉
(吉林大学 交通学院,长春 130022)

摘 要: 连续两个出租车GPS定位点之间的时空间隔使得乘客上下车的位置必然介于一个线性区间内,据此提出轨迹线密度方法,用于在位置界限模糊的热点出行区域进一步搜索热点路段.利用成都市出租车GPS数据,借助核密度估计分析出租车上下客位置的时空特性;基于轨迹线密度方法,计算了成都市春熙路商圈的路网密度值,划分路段热点强度,识别出了热点路段的位置,结合实际的出行需求分布完成方法有效性的验证.结果表明,本文所采用的方法能够有效识别出行需求旺盛的城市热点路段,不仅可以为出租车司机寻找客源提供重要的参考,还能够在交通相关部门选择出租车停靠站的位置时提供数据支持.

关键词: 城市交通;轨迹线密度;核密度分析;城市热点路段;GPS大数据;出租车停靠站

Urban Hotspot Travel Section Identification Method Based on Taxi GPS Large Data

QU Zhao-wei, WANG Xin, SONG Xian-min, XIA Ying-ji, YUAN Mi-li
(School of Transportation, Jilin University, Changchun 130022, China)

Abstract: Due to the spatio-temporal interval between two consecutive taxi GPS points, the location where passengers get on or off the taxi is in a linear range. A method of trajectory density was proposed for further searching hotspots sections in the hotspots area where the hotspots location is fuzzy. Utilizing taxi GPS data of Chengdu, the spatio-temporal characteristics of the pick-up and drop-off location were analyzed by means of kernel density method; Based on the trajectory density method, the density of road network in Chunxi commercial district of Chengdu was calculated, the hotspots intensity of the section was divided, and the location of hotspots section was identified. The validity of the proposed method is verified by the actual travel demand distribution. The result shows that our method can effectively identify the hotspots sections where travel demand is strong. It can not only provide important reference for taxi drivers to find customers, but also give data support for traffic related departments to locate reasonable taxi stops.

Keywords: urban traffic; trajectory density; kernel density estimation; urban hotspot section; GPS big data; taxi stands

0 引 言

城市出行热点通常是指商业发达、交通流量较大的区域,在某种程度上是人们密集出行的体现,研究其时空特性的分布规律对基础设施的建设和城市交通管理与规划具有重要的现实意义[1-3].由于受到数据采集方式等因素的影响,早期关于城市热点区域的研究较少,主要是通过土地利用类型数据和固定检测器采集的车辆信息对城市交

收稿日期:2018-09-10 修回日期:2018-12-13 录用日期:2019-01-14
基金项目:国家自然科学基金/National Natural Science Foundation of China(51278220);吉林省教育厅“十三五”科学技术项目/ Jilin Provincial Department of Education “13th Five-year” Science and Technology Project (JJKH20190153KJ).
作者简介:曲昭伟(1962-),男,吉林长春人,教授,博士. *通信作者:songxm@ jlu.edu.cn

通状态及空间分布模式进行分析[4-5].近年来,随着定位技术和无线通信的发展,大部分城市的出租车均安装了车载GPS,越来越多的研究致力于将出租车作为浮动车,基于海量GPS数据分析居民日常出行模式及交通状态规律,以便更加准确高效地识别出城市热点出行区域[6-7].

关于利用GPS数据挖掘城市热区的研究方法主要集中在聚类分析和空间统计.Lee[8]从出租车运行历史数据中提取乘客上车位置信息,分析其时空分布特性,并通过对乘客上车点进行聚类分析,提取了城市热点区域分布特性,为空驶出租车推荐合理的寻客位置;Gui等[9]将密集出租车轨迹数据视为停车点,并利用基于DBSCAN算法对停止点进行聚类提取了交通热点;Zou[10]针对城市交通网络建立了空间分析模型,并通过空间自相关统计量和核密度估计方法分析了城市交通状态的空间分布特征,发现城市交通状态存在显著的空间依赖性和异质性;还有一些学者利用贝叶斯网络、格网划分、建立概率分布模型等方法对城市热点区域进行分析[11-13].以上研究虽然各自采用的技术、方法不尽相同,但本质上都是对出租车运行轨迹中上下客近似点的分析.事实上,出租车GPS轨迹中的上下客事件并不是单纯的点事件,而是由于采样间隔的存在所导致的线性事件,根据GPS数据的点状特征挖掘城市热点位置必然会在一定程度上存在偏差.综上,本文在核密度分析方法的基础上,利用成都市出租车GPS数据,分析了出租车上下客位置分布的时空特性,确定成都市热点出行区域;依据乘客上下车前后时刻出租车运动轨迹的0-1线性特性,提出一种出租车上下客轨迹线密度方法,实现更为精确地识别城市热点路段.本文的研究框架如图1所示.

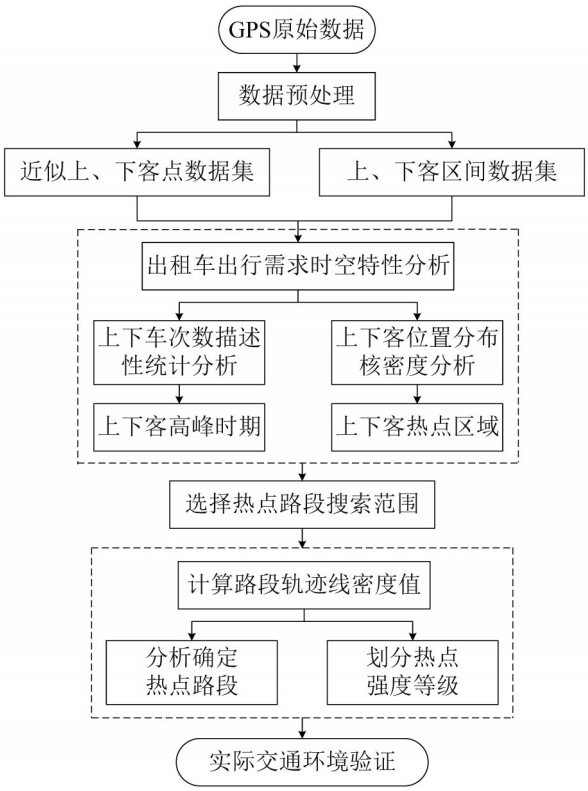


图1　研究框架

Fig. 1　Framework of this research

1　数　据

出租车GPS轨迹数据通常包括:出租车编号、经纬度、车载状态、定位时间及速度等信息.本文选取成都市1.3万余辆出租车在2014年8月4~10日1周7天内6:00-24:00的GPS轨迹数据,通过对数据的一系列预处理最终得到3亿余条有效数据,其中部分样本数据如表1所示.

表1　出租车GPS数据集样本

Table 1　Sample of taxi GPS data set

id	gps time	lng	lat	status
5884	2014-08-15 11:33:05.000	104.066 810 60	30.618 652 14	0
5884	2014-08-15 11:33:35.000	104.066 886 90	30.617 563 24	0
5884	2014-08-15 11:34:06.000	104.066 963 19	30.615 430 83	0
5884	2014-08-15 11:34:37.000	104.067 039 48	30.614 063 26	1
5884	2014-08-15 11:35:08.000	104.067 199 70	30.610 328 67	1
…	…	…	…	…

出租车GPS数据的车载状态包含载客(1)与空载(0)两种.现有的研究方法均是利用车载状态0→1中的点1和1→0中的点0近似地代替上下客位置点.实际上,车载GPS上传数据存在一定的时间间

隔,乘客上下车不是发生在某个状态位置的点事件,而是车载状态0→1或1→0变化过程中的线性事件,即上下车位置介于出租车GPS轨迹的[0,1]或[1,0]线性区间内.本文在Microsoft SQL Sever 2008数据库环境下分别从预处理后的GPS数据中提取了上客区间数据集$S[0,1]$、下客区间数据集$S[1,0]$,以及近似上客点数据集$S^{1}[0,1]$和近似下客点数据集$S^{0}[1,0]$,为后续城市热点区域的时空特性分析提供数据支持.

2 出租车出行需求的时空特性分析

2.1 时间特征分析

选取成都市出租车GPS近似上下客点数据集,对乘客的上下客近似点进行统计,分析出租车上下客次数在1周及1天内的变化规律,分别如图2和图3所示.

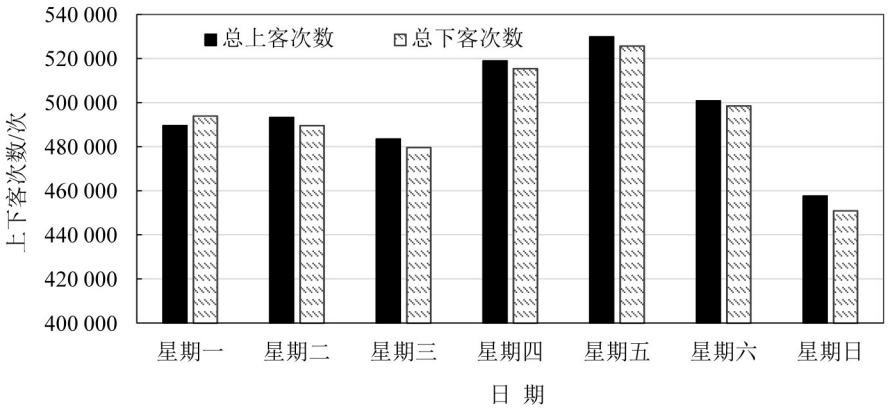


图2 出行需求周变化

Fig. 2 Weekly variation of travel demand

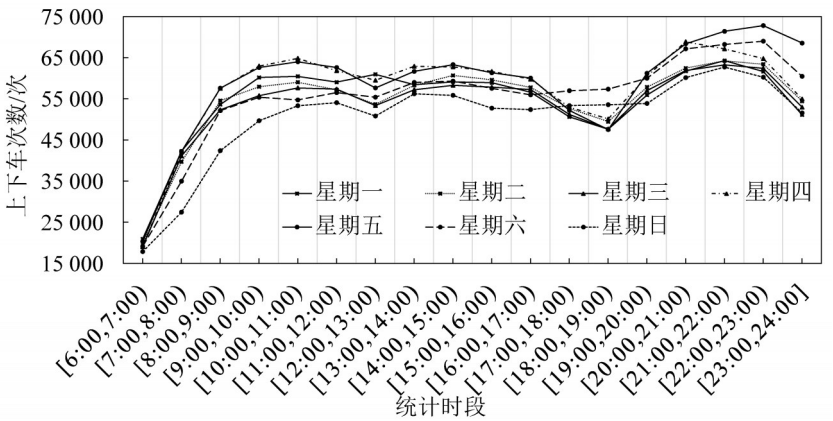


图3 出行需求日变化

Fig. 3 Daily variation of travel demand

由图2中上下客次数在不同时间内的对比可以发现,每日的上客次数整体要略高于下客次数,1周内乘车出行次数最大和最小值分别出现在星期五和星期日.图3显示工作日和休息日的上下客次数随时间的变化规律趋势大体一致,但在局部时间段内有所差异:

(1) 工作日的上下客次数呈现出早晚两个高峰期,分别是上午9:00-11:00和下午20:00-22:00;而星期五的晚高峰则相对延后了1 h,出现在21:00-23:00时段,此时的上下客次数也是1周各个时段中的最高值,这是因为星期五是工作日中的最后1天,人们夜间休闲娱乐活动大量增加,乘车出行次数随之升高.

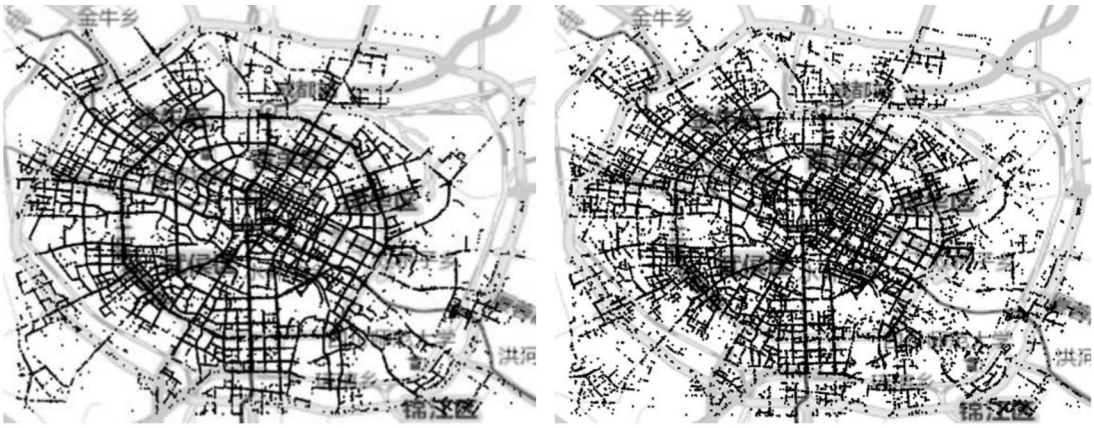
(2) 休息日上下客次数随时间变化的大体趋势基本一致,白天变化较为平缓,并无明显的激增或锐减.星期六的晚高峰与星期五相同,出现在21:00-23:00时段;而星期日同工作日类似,出现在20:00-22:00时段;此外,星期日各个时段的上下

客次数都明显低于星期六.其原因是星期日为休息日最后1天,夜生活结束较早,人们大多居家休息,出行次数明显下降.

综上所述,1周内出租车出行次数的峰值出现在星期五,且星期五21:00-23:00时段的上下客次数也是各个时段中的最高值.

2.2 空间特征分析

选取出租车GPS轨迹的近似上下客数据点,将其与研究范围内的路网地图进行匹配,并借助ArcGIS地理信息系统平台进行可视化表达,如图4所示.



(a) 近似上客点　　(b) 近似下客点

图4　数据点可视化

Fig. 4　Visual map of data points

$$\lambda(x,y)=\frac{1}{\pi r^2}\sum_{i=1}^{n}k\left(\frac{d_{is}}{r}\right) \tag{1}$$

式中:$\lambda(x,y)$是圆心点$S(x,y)$处的核密度值;r为搜索半径;d_{is}为点i到点S的距离;$k(\ )$为核函数,以Silverman著作中描述的二次核函数为基础.

2.2.2 城市出行热点区域空间分布特征

选择星期五各时段对出租车上下客近似点进行核密度估计,其中21:00-23:00的可视化结果如图6和图7所示.空间特征分析可以发现:

在上下客晚高峰(21:00-23:00),上客热点区域数量较多,主要分布在车站、春熙路商圈、武侯祠和宽窄巷子等热门旅游景点附近;下客热点相对上客来说较少,主要集中在酒吧等休闲娱乐场所附近.通过多时段对比发现,成都火车站和春熙路商圈始终是上下客的热点区域.火车站由于其作业的特殊性,呈现出点状的辐射态.春熙路是成都最繁华的商业街,商圈覆盖多条街道,热点区域呈现网状特征.

2.2.1 核密度分析原理

在传统的城市与区域研究中,核密度分析方法由于其易于实现,且能够较好地反映地理空间分布中的距离衰减效应等优点,已作为一种可视化工具被广泛使用,同时也成为了最常用的热点分析方法[14-15].

在核密度分析中,每个点的上方均覆盖着一个平滑曲面,如图5所示,在该点的水平切面的核函数值最大,随着与该点距离的增大核函数值逐渐减小至零.每个输出栅格像元的核密度均为叠加在栅格像元中心的所有核表面值之和.事件点i处的核密度值可以用式(1)表示[16].

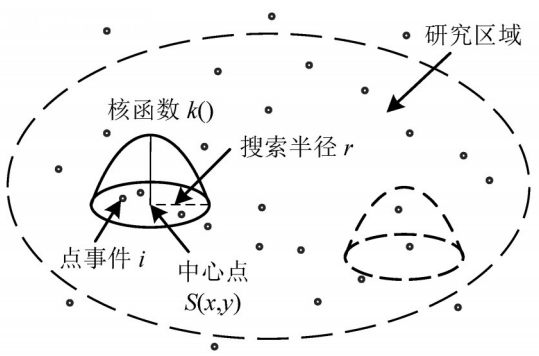


图5　核密度原理示意图

Fig. 5　Diagram of kernel density method

图6 出租车上客近似点核密度图(21:00-23:00)

Fig. 6 Kernel density of taxi pick-up approximate points (21:00-23:00)

图7 出租车下客近似点核密度图(21:00-23:00)

Fig. 7 Kernel density of taxi drop-off approximate points(21:00-23:00)

3 城市热点区域的上下客轨迹线密度

乘客上下车是发生在出租车轨迹[0,1]或[1,0]区间内的线性事件,基于近似上下客数据的点分布特征,对出租车出行需求进行空间分析会在热点区域的各条道路上存在一定的偏差.因此,利用线性事件对应的0→1或1→0轨迹线,对其线性轨迹特征进行分析能够更加合理地表现出上下客位置在路段上准确的空间分布.

3.1 轨迹线密度建模

轨迹线密度用于计算在一定搜索长度内上下客事件发生时所形成的线状轨迹的密度值,等价于在一定时间内单位长度上发生的上下客事件的累积概率值,计算方法如式(2)所示.其中概率的计算是将上下客轨迹线投影在搜索路段内的长度与

此条上下客轨迹线自身的投影长度作比.

$$\mathrm{TD}=\left(\sum_{i=1}^{n}P_i\right)\Big/L=\left(\sum_{i=1}^{n}\frac{d_i}{D_i}\right)\Big/L \tag{2}$$

式中:L为搜索路段长度(m);n为搜索路段L内可能发生的上下客事件的总数量(次);D_i表示第i个上下客事件所应对的轨迹线投影在道路边缘线上的自身长度(m);d_i表示第i个上下客事件所对应的轨迹线投影在某个搜索路段L内的部分长度(m).

如图8所示,用箭头端代表车载状态为1的点,另一端代表车载状态为0的点,则此条带箭头的黑色实线代表了发生1次上客行为的轨迹线;确定搜索长度为L,在某个指定的时间段内,某个搜索路段上一共可能发生了4次上下客事件$n=4$,这4条上下客轨迹线自身投影在道路上的长度分别为D_1,D_2,D_3,D_4,而投影在该搜索路段L内的长度分别是d_1,d_2,d_3,d_4,那么经过计算,该搜索路段上的轨迹线密度为

$$\mathrm{TD}=\left(\frac{d_1}{D_1}+\frac{d_2}{D_2}+\frac{d_3}{D_3}+\frac{d_4}{D_4}\right)\Big/L \tag{3}$$

通过对轨迹线密度定义的理解,可以知道某条路段的轨迹线密度值越高,说明在此路段上发生上下客的次数越多,概率越大.

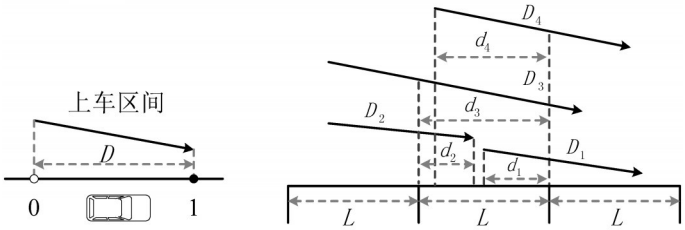


图8　轨迹线密度示意图

Fig. 8　Definition of trajectory density

3.2 实验结果分析

基于成都市出行热点区域的时空分布特性,本节选择春熙路商圈作为研究区域,进行轨迹线密度的计算.如图9所示,将此区域内的交叉口按照顺时针方向编号,顺城大街相较其他道路来说核密度值较小,不在本次实验范围内.调查得知,为了遏制出租车随意停靠载客的现象,成都市在本文选取的4条研究道路上设置了一定数量的出租车停靠站,顺时针方向8个,逆时针方向4个.如果市民和出租车司机按照规定在此乘车和载客,那么其实际位置应包含在利用本文的方法所识别出的热点路段内.

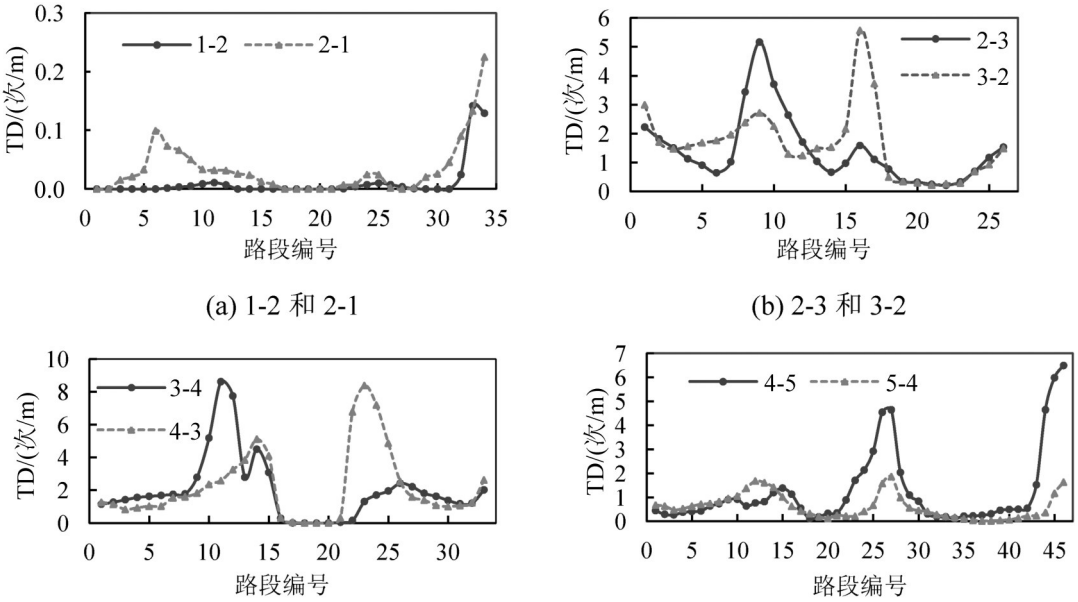
根据城市道路交通标志和标线设置规范[17],出租车专用上下客停车位长度应为6 m,高峰小时上客人次大于50的站点停车位不超过3个.本文中将搜索长度设定为3个停车位的长度18 m.由于道路上均存在对向车流,在整个路网上将其视为顺时针和逆时针两个方向进行轨迹线密度的计算,结果如图10所示,其中4条道路轨迹线密度的部分描述性统计特征量如表2所示.

图9　研究区域路网

Fig. 9　The road network of study area

通过对轨迹线密度图表的分析可知,各条道路的密度值不在同一水平范围内,1-2道路整体偏小.结合实际交通环境分析可知,由于春熙路商圈路网内部多存在禁止车辆通行的步行街,连接到主干路的进出口数目不尽相同,乘客更倾向于选择在离目的地较近的路段上乘车,因此不同道路的上下客次数相差较大.将各条道路视为独立的研究单元,分析可知道路双向上均存在数量不等的峰值点,这与热点路段上发生上下客事件的概率

大,其轨迹线密度必然明显高于其他路段的实际情况相吻合,可以判定峰值点所属路段即为热点路段,因此利用轨迹线密度方法识别出该区域内存在31个热点路段.



(a) 1-2 和 2-1 (b) 2-3 和 3-2

(c) 3-4 和 4-3 (d) 4-5 和 5-4

图 10 研究区域4条道路的轨迹线密度

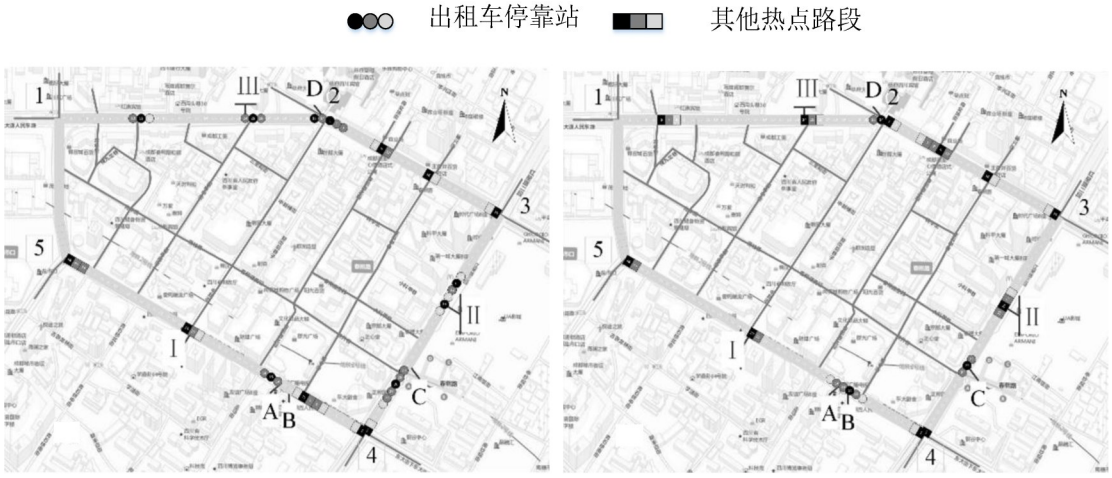
Fig. 10 Trajectory density of four roads in study area

表2 轨迹线密度描述性统计特征量

Table 2 Statistical characteristic quantity of trajectory density

道路编号	1-2		2-3		3-4		4-5	
最小值	0.000 00		0.212 94		0.000 00		0.123 90	
最大值	0.142 09		5.166 19		8.622 32		6.495 72	
均 值	0.010 79		1.421 60		2.013 91		1.225 38	
线密度峰值	11	0.010 85	1	2.226 59	11	8.622 32	1	0.450 61
	25	0.010 20	9	5.166 19	14	4.509 15	10	0.924 51
	33	0.142 09	16	1.589 34	26	2.428 57	15	1.389 79
			26	1.535 34	33	2.026 67	27	4.650 80
							46	6.495 72
道路编号	2-1		3-2		4-3		5-4	
最小值	0.000 00		0.216 48		0.000 00		0.015 37	
最大值	0.225 17		5.564 68		8.382 16		1.862 17	
均 值	0.033 35		1.634 82		2.207 24		0.615 97	
线密度峰值	6	0.099 34	1	2.989 78	1	1.262 86	1	0.676 08
	24	0.024 43	9	2.713 34	14	5.122 36	12	1.683 19
	34	0.225 17	16	5.564 68	23	8.382 16	27	1.862 17
			26	1.479 64	33	2.627 20	46	1.635 19

为了更加直观的了解研究区域内热点路段的空间分布情况,分析热点路段与出租车停靠站数目不符的原因,本文将各路段按照峰值点、峰值的70%分位点,峰值的60%分位点将热度强弱划分为3个等级,并借助1组深浅渐变色将其在地图上进行可视化表达,其中,与出租车停靠站位置相符的热点路段用圆形表示,其他用矩形表示,结果如图11所示.



(a) 顺时针方向　　(b) 逆时针方向

图 11　轨迹线密度可视化

Fig. 11　Visual map of trajectory density

观察热点路段分布图可以发现：顺时针方向上，利用轨迹线密度方法对出租车停靠站所属热点路段的识别准确率为100%；其他热点路段位置均在交叉口附近，呈现出相似的空间分布特征.这与城市道路交叉口处交通流量大、打车成功概率高，居民更倾向于在此处候车的常识相符.逆时针方向上，识别出了B、C、D这3处出租车停靠站所属路段，A、B实际距离约为25 m，且停车位数量均为1个，线密度方法将其识别成为1个热点，但未识别出的停靠站A仍属热度较高的路段，在密度峰值的70%分位点内.此外，热点路段中的Ⅰ、Ⅲ临近蜀都大厦和新良大酒店等来往人流量较大的建筑出入口处，故将其识别为热点路段是合理的.由于3-4道路的中段禁止机动车辆通行，Ⅱ位于允许通行的路段尽头，对向位置是3-4方向上的出租车停靠站，因此Ⅱ号路段落在对向出租车停靠站上下客的轨迹区间内，也被识别为热点路段.综上，利用轨迹线密度方法能够准确地检测出热点路段的空间分布位置，对整个路网上的出租车停靠站识别的准确率为91.67%；由于居民和出租车司机并不完全遵守在出租车停靠区乘车和候客的交通规则，热点路段除出租车停靠站位置外，多位于交叉口和人流量较大的建筑出入口处.

4　结　论

(1) 利用成都市出租车GPS轨迹数据挖掘城市热点路段分布信息，借助核密度分析实现了城市热点区域的可视化表达，分析了出行需求的时空分布特性.

(2) 根据出租车乘客上下车事件的线性轨迹特征，提出轨迹线密度的方法在热点区域内进一步确定热点路段的位置.实验结果表明，轨迹线密度方法能够在热点区域内更加精确地识别出热点路段，避免了以往研究中利用上下客近似点分析确定热点位置所带来的误差.

(3) 城市出行热点路段的时空分布信息对于城市的交通规划与管理具有非常重要的意义，这些信息在一定程度上能够有效改善打车难与出租车空驶率高并存的矛盾现状.此外，热点路段的分布特征在城市出租车停靠区的选址问题方面具有重要的参考价值.

参考文献：

[1] ZHAO P X, QIN K, ZHOU Q, et al. Detecting hotspots from taxi trajectory data using spatial cluster analysis[J]. Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015(II-4/W2): 131-135.

[2] YU Z, ZHANG L Z, XIE X, et al. Mining interesting locations and travel sequences from GPS trajectories [C]// Quemada J, León G, Maarek Y, Nejdl W. Proceedings of the 18th international conference on World wide web. New York: ACM, 2009: 791-800.

[3] CHANG H W, TAI Y C, HSU Y. Context-aware taxi demand hotspots prediction[J]. International Journal of Business Intelligence & Data Mining, 2010, 5(1): 3-18.

[4] ELGAMMAL A, DURISWAMI R, HARWOOD D, et al. Background and foreground modelling using nonparametric kernel density estimation for visual surveil[J]. Proceedings of the IEEE, 2002, 90(7): 1151-1163.

[5] XIE Z X, YAN J. Kernel density estimation of traffic accidents in a network space[J]. Computers Environment & Urban Systems, 2008, 32(5): 396-406.

[6] CAO X, CONG G, Jensen C S. Mining significant semantic locations from GPS data[J]. Proceedings of the VLDB Endowment, 2010, 3(1): 1009-1020.

[7] XU X L, DOU W C, ZHANG X Y, et al. A traffic hotline discovery method over cloud of things using big taxi GPS data[J]. Software Practice & Experience, 2017, 47 (3): 361-377.

[8] LEE J, SHIN I, PARK G L. Analysis of the passenger pick-up pattern for taxi location recommendation[C]// IEEE. 4th International Conference on Networked Computing and Advanced Information Management, LOS ALAMITOS: IEEE, 2008: 199-204.

[9] GUI Z M, YU H P. Mining traffic hot spots from massive taxi trace[J]. Journal of Computational Information Systems, 2014, 10(7): 2751-2760.

[10] ZOU H X, YUE Y, LI Q Q, et al. A spatial analysis approach for describing spatial pattern of urban traffic state[C]// IEEE. 13th International IEEE Conference on Intelligent Transportation Systems, Piscataway: IEEE , 2010: 557-562.

[11] WESTGATE B S, WOODARD D B, MATTESON D S, et al. Travel time estimation for ambulances using Bayesian data augmentation[J]. Annals of Applied Statistics, 2013, 7(2): 1139-1161.

[12] LI B, ZHANG D Q, SUN L, et al. Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset[C]// IEEE. 2011 IEEE International Conference on Pervasive Computing and Communications Workshops, Piscataway: IEEE, 2011: 63-68.

[13] YUAN J, ZHENG Y, ZHANG L H, et al. Where to find my next passenger[C]// Landay J, SHI Yuan-chun, Patterson D, Rogers Y, XIE Xing. Proceedings of the 13th International Conference on Ubiquitous Computing, New York: ACM, 2011: 109-118.

[14] SHEATHER S J, JONES M C. A reliable data-based bandwidth selection method for kernel density estimation[J]. Journal of the Royal Statistical Society, 1991, 53(3): 683-690.

[15] BORRUSO G. Network density estimation: A GIS approach for analysing point patterns in a network space[J]. Transactions in Gis, 2008, 12(3): 377-402.

[16] ANDERSON T K. Kernel density estimation and K-means clustering to profile road accident hotspots[J]. Accident Analysis & Prevention, 2009, 41(3): 359-364.

[17] GB51038-2015. 城市道路交通标志和标线设置规范[S]. 北京:中国计划出版社, 2015. [GB51038-2015. Specification for layout of urban road traffic signs and markings[S]. Beijing: China Planning Press, 2015.]