【交通物流 / Transportation Logistics】

下车信息对公交乘客精细化分类影响研究

李军，区静怡，赵文婷

中山大学智能工程学院，广东广州 510006

摘　要：目前广泛采用的公交一票制缺乏下车信息，利用一票制数据得到的乘客分类需要进行精细化分类效果评估．本研究利用包含下车站点的大规模公交票务数据，对比分析了有无下车信息下的分类模型，探究下车信息缺失对乘客分类的影响．采用分类指标和组合聚类方法对公交乘客进行精细化分类，选取中国北京市路面公交连续28 d的乘客上下车数据作为实例，分析下车信息及分类指标对分类结果的影响．结果表明，有无下车信息的2个模型均能实现乘客的有效分类，分类结果都能体现每类乘客的时空活动规律．其中，包含下车信息的分类模型能够识别具有明显出行距离特征的小样本群体，如占比0.25%的长距离出行乘客等特殊群体；而缺少下车信息的分类模型类别平均占比标准差远小于包含下车信息的分类模型，分类结果相对均衡，且各类别乘客在多维度上的规律差异显著，更能体现分类的宏观特征．

关键词：交通运输工程；公共交通；乘客分类；下车信息；组合聚类；出行特征

中图分类号：U121　　**文献标志码：**A　　**doi：**10.3724/SP.J.1249.2023.01109

Impact of bus alighting information on fine classification of transit riders

LI Jun, OU Jingyi, and ZHAO Wenting

School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong Province, P. R. China

Abstract: The widely used bus one ticket system lacks off boarding information, and the passenger classification based on the dataset of one-ticket bus system needs to be refined to evaluate the classification effect. This study uses the large-scale bus ticket data with alighting information to compare and analyze the transit rider classification models with and without the alighting information to explore the impact of the missing alighting information on the fine classification of transit riders. The classification indexes and combined clustering method are adopted to refine the classification of bus passengers. A case study of Beijing conventional buses which contain boarding and alighting data of 28 consecutive days is presented and used to analyze the impact of bus alighting data and classification indexes on the transit rider classification results. The result indicates that both models successfully complete the fine classification of the riders and reflect the spatio-temporal characteristics of each type of riders. The model with alighting information is able to identify the distance-concerned type with few samples such as the long-distance transit riders accounting for 0.25%. The standard deviation of the average proportion of categories in the model without alighting information is much smaller than that in the model with alighting information. The results obtained from the model without alighting information are more balanced, and the regularity of various categories of passengers in multiple dimensions is significantly different, which can better reflect the macro characteristics of classification.

Received: 2022-05-23; **Accepted:** 2022-06-25 **Online** (CNKI)**:** 2022-10-10
Foundation: Research and Development Project in Key Areas of Guangdong Province (2019B090913001)
Corresponding author: Associate professor LI Jun. E-mail: stslijun@mail.sysu.edu.cn
Citation: LI Jun, OU Jingyi, ZHAO Wenting. Impact of bus alighting information on fine classification of transit riders [J]. Journal of Shenzhen University Science and Engineering, 2023, 40(1): 109-117. (in Chinese)

Key words: transport engineering; public transit; rider classification; bus alighting information; combined clustering; travel characteristics

随着移动支付方式的多样化发展，公共交通乘客数据得以有效记录，不同计费方式记录着不同的乘客数据[1]．针对记录上下车数据开展的乘客分类已有充分研究，由于公交一票制中仅记录乘客的上车信息而不采集下车信息，相关乘客分类研究具有一定难度．因此，作为乘客分类研究的基础工作，评估下车信息缺失对乘客精细化分类的影响对使用一票制的公交部门分析乘客特征、提高服务水平至关重要．

乘客分类研究中，首先需要通过出行行为体现量化的乘客出行特征[2-3]，因此，分类需要基于乘客的多维属性．其中，乘客的个人属性是主观分类的重要参考信息[4-9]．乘客使用的卡片类型包含其个人社会信息，存在预先对乘客社会属性进行假设的情况．可通过采用客观分类避免乘客个人信息强主观性对分类结果的影响[10]，如围绕出行规律、时间和空间规律等维度选取分类指标．指标选取应包含刻画出行规律的出行次数及出行天数[11-12]、刻画时间规律的首次出发时间及乘车耗时[12]，以及刻画空间规律的乘车距离、出行起讫点(origin-destination，OD)占比、站点相似性[13-14]、站点距离和出行距离阈值等指标[15-16]；通勤特征识别研究结合上述维度提出高峰时段下的通勤稳定性和出行频率等指标[17-20]．对于缺少下车信息的乘客分类，通常采用出行链匹配或站点吸引权的方法补全下车站点[21-22]，推断率可达90%[16]．然而，在长观察周期的乘客分类研究中，多次推断所累积的重复误差会降低数据补全的正确率，影响乘客分类效果，仅通过上车信息对乘客作出精准分类结果的有效性需进一步论证．

乘客分类算法是更深入的关键问题，其中，时空规律研究多采用k-means[23-24]或基于密度的噪声应用空间聚类(density-based spatial clustering of applications with noise，DBSCAN)算法[14]；多类型指标研究采用二阶聚类算法[12]；依托初始聚类中心的优选研究采用k-prototypes聚类算法[13]．然而，以上研究的乘客分类指标相差较大，且都针对包括上下车信息的数据集，其乘客类别数在5类以内，各研究侧重不同的乘客特征，尚无对全群体的精细化分类方法．

为探究下车信息缺失对公交一票制乘客分类的影响，本研究分别构建无下车信息的分类模型O和有下车信息的分类模型W，从各维度的客观因素中选取乘客分类的通用指标和替代指标，针对模型W补充附加指标，采用组合聚类算法精细挖掘乘客类别．选取中国北京市路面公交的大规模乘客上下车刷卡数据作为研究对象，基于所建立模型分析下车信息存在与否对公交乘客精细化分类的影响．研究结果可为精细化乘客行为分析、个性化公交服务政策和多层次公交需求预测提供决策参考．

1 分类模型构建

1.1 分类指标选取

从多维度客观属性提取乘客的乘车特征指标，具体分为出行规律、时间规律、空间规律及高峰规律．为探究有无下车信息对于乘客精细化客观分类效果的影响，基于上车信息选取分类指标组构建模型O，基于上下车信息选取分类指标组构建模型W．围绕模型W和O分别构建有无下车信息的2组分类指标，包含8项两个模型共享的通用指标，4项两个模型对比的替代指标．此外，包含下车站点信息的模型W在时间规律和空间规律通过下车信息补充3项附加指标．

1.1.1 通用指标构建

构建通用指标时主要考虑乘客上车信息所呈现的特性，可通过刷卡次数和刷卡时间反映乘客的相关行为特性．乘客的出行规律反映了其对公共交通的依赖性，出行时间和出行天数反映其使用公共交通的时间规律．以下从出行规律和时间规律维度提出适用于模型O和W的通用指标，见表1．其中，$s_{\max}$为日均出行次数最大值；$\eta_{\max}$为每周出行天数标准差最大值；$\sigma_{\max}$为日初次出行时间标准差最大值；下标k表示第k位乘客．

1.1.2 替代指标构建

由于一票制下的乘客数据通常包括上车信息，对下车信息的采集并不统一，不包含下车信息时难以刻画乘客出行的空间规律．选取出行日初次乘车和末次乘车的站点关联性揭示空间规律，选取周中工作日的早高峰和晚高峰时段初次乘车关联性刻画高峰规律．在空间规律和高峰规律维度下，针对模型O和W提出替代指标．

表1　模型O和模型W的通用指标

Table 1　Shared classification indexes of transit riders between model O and model W

	通用指标	变量名	值域
出行规律	日均出行次数/(次·d^{-1})	s_k	$(0,s_{\max}]$
	出行周比例	w_k	$(0,1]$
	每周平均出行天数/(d·周$^{-1}$)	A_k	$(0,7]$
	每周出行天数标准差/(d·周$^{-1}$)	η_k	$[0,\eta_{\max}]$
时间规律	日均初次出行时间/min	T_k	$[0,1\,440]$
	日初次出行时间标准差/min	σ_k	$[0,\sigma_{\max}]$
	周中平均出行天数/(d·周$^{-1}$)	R_k	$(0,5]$
	周末平均出行天数/(d·周$^{-1}$)	N_k	$(0,2]$

在替代指标中，针对模型O和W的数据差异，分别提出上车站点邻近度和常OD对比例，刻画乘客多次出行的上车站点关联性和上下车站点关联性．定义指标范围为出行日初末次出行和周中工作日早晚高峰出行，分别表征乘客出行的空间规律和高峰规律．

上车站点邻近度表征乘客多次出行间的站点间距邻近程度和出行集中程度．乘客k的上车站点邻近度指标包括：日初次上车站点邻近度P_k^{F}、日末次上车站点邻近度P_k^{L}、周中早高峰初次上车站点邻近度P_k^{M}和周中晚高峰初次上车站点邻近度P_k^{E}．

上车站点邻近度计算中，若乘客在目标范围内任意两个乘车站点间距在预设的判断值内，可视作乘客的出行站点选择集中．站点邻近度为

$$P_k^X = \frac{\sum_{i=1}^{D_k^X}\sum_{j=1}^{D_k^X}\gamma_{ij}^k}{D_k^X\left(D_k^X-1\right)} \tag{1}$$

$$\gamma_{ij}^k = \begin{cases} 0, & S_{k_{ij}}^X > \varepsilon \text{或} i=j \\ 1, & S_{k_{ij}}^X \leqslant \varepsilon \text{且} i \neq j \end{cases} \tag{2}$$

其中，X为目标指标范围，包括日初次出行F、日末次出行L、周中早高峰初次出行M和周中晚高峰初次出行E；D_k^X为乘客k在目标指标范围X下的总出行天数(单位：d)，且$D_k^X>1$；$S_{k_{ij}}^X$为乘客k在第i和j个目标指标范围内的站点间距(单位：m)；γ_{ij}^k为乘客k的站点间距判断0-1变量；ε为给定的站点间距判断值(单位：m)．

常OD对比例为乘客多次出行的OD对，表征乘客出行起止点的专一性，其指标包括日初次出行常OD对比例V_k^{F}、日末次出行常OD对比例V_k^{L}、周中早高峰初次出行常OD对比例V_k^{M}和周中晚高峰初次出行常OD对比例V_k^{E}．

常OD对比例计算以出行的上车站点和下车站点作为1个OD对．乘客k在目标指标范围X下的总出行天数为D_k^X，目标指标范围出行OD集合中出现频率最高的OD对出现次数记为C_k，常OD对比例计算为

$$V_k^X = C_k/D_k^X \tag{3}$$

当乘客k在X下的总出行天数为0 d或1 d，上车站点邻近度和常OD对比例指标需作额外标定，分别记为−2和−1．总出行天数＞1 d时，上车站点邻近度和常OD对比例指标取值范围为$[0,1]$．因此，上车站点邻近度和常OD对比例指标为分类型和数值型结合的组合型指标．

模型O和模型W中，乘客在空间规律和高峰规律下所选取的上车站点邻近度和常OD对比例等对比分类指标见表2．

表2　模型O和模型W的对比指标

Table 2　Subsitude classification indexes of transit riders between model O and model W

	模型O	模型W
空间规律	日初次上车站点邻近度P_k^{F}	日初次出行常OD对比例V_k^{F}
	日末次上车站点邻近度P_k^{L}	日末次出行常OD对比例V_k^{L}
高峰规律	周中早高峰初次上车站点邻近度P_k^{M}	周中早高峰初次出行常OD对比例V_k^{M}
	周中晚高峰初次上车站点邻近度P_k^{E}	周中晚高峰初次出行常OD对比例V_k^{E}

1.1.3　附加指标构建

针对模型O和W选取替代指标、模型W选取附加指标，考虑到下车的时间和地点信息可呈现一定的乘客特征，有利于刻画乘客的日常活动时长和活动距离广度．因此，在时间规律和空间规律中围绕下车信息，针对性提出模型W的附加指标，包括日均出行时长、日出行时长标准差及日均出行距离，如表3．其中，$H_{\max}$为日均出行时长最大值；$\theta_{\max}$为日出行时长标准差最大值；$L_{\max}$为日均出行距离最大值．

表3 模型W的附加指标

Table 3 Additional classification indexes of the transit rider in model W

	附加指标	变量名	值域
时间规律	日均出行时长/min	H_k	$(0, H_{\max}]$
	日出行时长标准差/min	θ_k	$(0, \theta_{\max}]$
空间规律	日均出行距离/km	L_k	$(0, L_{\max}]$

1.2 乘客分类方法

1.2.1 指标筛选

为提高分类效率，降低计算冗余，在选取分类指标后，需对已有指标进行相关性分析．考虑到模型的输入指标包含数值型和组合型指标，采用斯皮尔曼相关性系数进行计算筛选．对于相关性系数落在[−0.6, 0.6]以外的指标对，可认为其相关性较强，予以剔除．

1.2.2 组合聚类

为深入挖掘乘客在不同维度下的规律与特征，考虑到同时存在组合型指标和数值型指标，本研究采用二阶聚类和k-means聚类相结合的方法，分别对模型O和模型W的乘客指标组进行精细化分类，具体步骤如下．

步骤1 采用二阶聚类对乘客进行客观分类，采用聚类特征(clustering feature, CF)树提高聚类效率并根据贝叶斯信息准则(Bayesian information criterion, BIC)选取最优聚类数．由于空间规律和高峰规律下存在组合型指标，首先将该类组合型指标视为分类型，类型包括无出行、单日出行和多日出行．类别中心在该类指标的表现为单一类型时，视作表现清晰；反之则为非清晰．根据分类型指标表现的一致性与否，将乘客划分为清晰组乘客和非清晰组乘客．

步骤2 对在空间规律和高峰规律下清晰度不明的非清晰组乘客类别应用二次二阶聚类，并同样选取CF树和BIC，得到细化类别；对在空间规律和高峰规律下的清晰组乘客类别应用k-means聚类($k = 2, 3, 4, \cdots, 10$)，并计算不同k值下的聚类误差平方和作为k值选取参考，选取误差平方和变化幅度大的点作为k值．

2 案例分析

2.1 案例概况

中国北京市的路面公交采取分段计费制，通过自动售检票系统(automatic fare collection, AFC)进行数据采集，票务信息包括乘客出行的上下车时空数据．选取2018-04-02至2018-04-29的4个完整且连续的星期作为研究周期．将数据中存在的数据空值、重复记录和不合理数据予以清洗．在经过刷卡数据清洗和空间信息匹配后，采取等比例分层抽样方法，选取周期内乘客票务数据共27 192 981条，包含付费乘客共2 245 561位．票务数据包括乘客身份标识号(identity document, ID)、上车时间及站点、下车时间及站点、线路编号等，采用数据库管理系统Structured Query Language Sever(SQL Server)进行存储管理，作为乘客分类模型的数据应用基础．

2.2 乘客分类结果

结合乘客票务数据，计算乘客在各模型指标组下的表现，利用SPSS软件工具予以分析分类指标的相关性．考虑到北京市面积辽阔，为了尽可能覆盖相邻站点，并保证公交站点服务半径500 m覆盖90%人口的要求，指标计算过程中将站点间距判断值取1 000 m，分别以06:00—10:00和17:00—21:00作为早晚高峰时段．

进行斯皮尔曼相关性分析并经过指标筛选后剔除冗余指标．其中，模型O指标中，剔除乘车周数和周均乘车天数；模型W指标中，剔除乘车周数、周均乘车天数及日乘车时长标准差．模型O和W中，各分类指标间的相关系数见图1，以颜色深浅表示相关系数大小．

分类过程中，模型O和W分别输入10项和12项分类指标后，对数据集进行聚类处理．先采用二阶聚类对目标乘客进行客观分类，得到多类规律清晰的乘客和规律较不清晰的乘客．将清晰度较低的非清晰组乘客类别统一应用二次二阶聚类，得到细化分类；对清晰组乘客类别应用k-means聚类，选取聚类误差平方和下降幅度变化最大值作为k值．聚类后模型O采用缺少下车信息的数据集，可将乘客分为10类．模型W采用包含下车信息的数据集，可将乘客分为12类．

分析模型O和W下的乘客分类情况，可将出行规律初步分为高频用户、中频用户和低频用户．其中，模型O下的高频、中频和低频用户又可各自细分为4类、3类和3类，共计10类；模型W下的高频、中频和低频用户可各细分为4类，共计12类．

各类别乘客的分类特征中，以类别的聚类中心

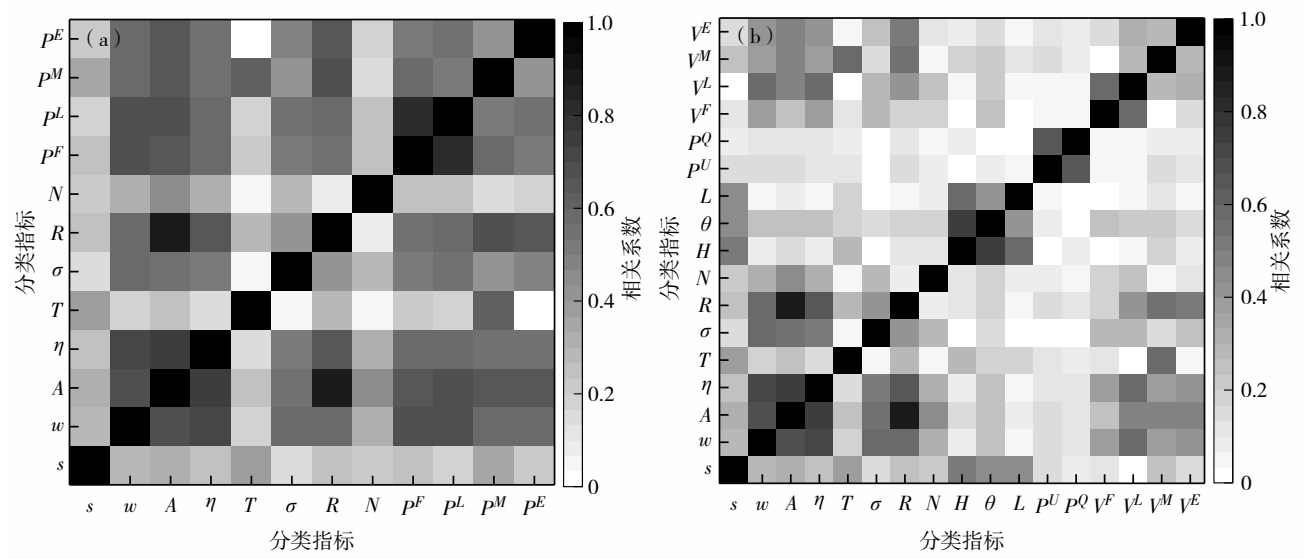

图1　(a)模型O和(b)模型W的指标相关系数

Fig. 1　Correlation coefficients of (a) model O and (b) model W.

表现对类别特征进行描述．其中，出行规律分为高频、中频和低频，分别对应日均乘车次数$s_k \geqslant 1.7$，$1.5 \leqslant s_k < 1.7$及$s_k < 1.5$．时间规律分为晨、早、午、午后和晚，分别对应日均初次出行时间08∶00前、08∶00—10∶00、10∶00—12∶00、12∶00—16∶00及16∶00后．模型W中的日均出行距离以短距、中距与长距表征，分别对应日均出行距离[0, 20)、[20, 60)和[60, ∞] km．以是否存在空间规律和高峰规律表征乘客在这两项规律下的表现．表4为模型O和W的乘客分类特征及其占比．

表4　模型O和模型W的乘客分类特征及其占比

Table 4　Passenger classification results by model O and model W

分类乘客	模型O			模型W		
	类别序号	分类特征	占比/%	类别序号	分类特征	占比/%
高频用户	1	高频午后出行高峰规律	3. 36	1	高频早长距出行高峰规律	0. 25
	2	高频早出行高峰规律	7. 02	2	高频早短距出行高峰规律	5. 87
	3	高频午出行高峰规律	6. 78	3	高频午短距出行高峰规律	13. 71
	4	高频晨出行高峰规律	3. 99	4	高频早中距出行高峰规律	1. 33
中频用户	5	中频午出行	18. 05	5	中频午后短距出行空间规律	17. 49
	6	中频午后出行	20. 86	6	中频午短距出行空间规律	11. 54
	7	中频早出行	13. 73	7	中频早短距出行	13. 60
				8	中频晚短距出行	10. 00
低频用户	8	低频早出行	10. 38	9	低频午后短距出行	20. 85
	9	低频午出行	8. 73	10	低频早中距出行	1. 30
	10	低频午后出行	7. 10	11	低频晨长距出行	0. 30
				12	低频午短距出行	3. 76

2. 3　分类效果讨论

乘客分类中包含高频用户、中频用户和低频用户，其中，高频用户在各维度的分类指标均有表现，类别聚类中心的出行规律更明显．本研究以高频用户为例，分析模型W的小样本乘客分类原因，探讨模型W中下车信息对于乘客分类的影响，并通过雷达图分析模型O和W在分类区分度方面的表现．

不含下车信息的模型O和包含下车信息的模型W中各类高频乘客在各指标下的聚类中心见表5．

由表5可见，相较于模型O，模型W在分类时包括含下车信息的2项附加指标．为探究两种模型的分类结果差异是否由附加指标引起，分别分析模型W中仅隐去日均乘车时长和仅隐去日均乘车距离的分类情况差异．在模型W的分类变量中仅隐去日均出行时长并经过再次聚类，高频用户分类结果与原模型W的分类结果较为接近，仅0.01%的乘客类别发生变动；而仅隐去日均出行距离再次聚类后，高频用户分类结果与原模型W的分类结果产生较大差异，引起65.36%的乘客类别变动，而与模型O的分类效果十分接近，乘客相同类别重复率为98.80%．由此可见，乘客的日均出行距离是模型W与O在分类结果中产生差异的重要指标．

表5 模型O和模型W中高频用户的聚类中心

Table 5 Clustering centers of high-frequency transit riders by model O and model W

模型O					模型W				
指标	第1类	第2类	第3类	第4类	指标	第1类	第2类	第3类	第4类
日均出行次数/(次·d^{-1})	1.70	2.29	2.01	2.55	日均出行次数/(次·d^{-1})	2.91	2.53	1.90	2.90
每周出行天数标准差/(d·周$^{-1}$)	1.03	1.06	1.06	0.90	每周出行天数标准差/(d·周$^{-1}$)	0.94	1.03	1.03	1.00
日均初次出行时间/min	839	556	680	469	日均初次出行时间/min	535	596	644	567
日初次出行时间标准差/min	287	191	256	62	日初次出行时间标准差/min	166	187	212	178
周中平均出行天数/(d·周$^{-1}$)	2.66	3.52	2.91	3.62	周中平均出行天数/(d·周$^{-1}$)	3.16	3.29	3.17	3.21
周末平均出行天数/(d·周$^{-1}$)	0.72	0.92	0.81	0.83	周末平均出行天数/(d·周$^{-1}$)	0.73	0.92	0.80	0.88
日初次上车站点邻近度	0.19	0.34	0.23	0.48	日初次出行常OD对比例	0.53	0.35	0.33	0.41
日末次上车站点邻近度	0.19	0.22	0.18	0.29	日末次出行常OD对比例	0.40	0.25	0.20	0.31
周中早高峰初次上车站点邻近度	0.36	0.44	0.39	0.50	周中早高峰初次出行常OD对比例	0.61	0.47	0.47	0.52
周中晚高峰初次上车站点邻近度	0.28	0.32	0.27	0.39	周中晚高峰初次出行常OD对比例	0.56	0.44	0.44	0.48
					日均出行时长/min	150	70	41	106
					日均出行距离/km	72	13	4	31

高频用户中，模型W的乘客各类别占比相差较大，由表4可见，类别1和类别4的乘客占比仅0.25%和1.33%．对其聚类中心分析可见，由于北京市跨区公交出行行为高发，这部分乘客的日均乘车距离远高于其他2个类别的高频乘客，此指标取值的强差异性使样本量极低的乘客在聚类中被识别．而模型O中高频乘客分类占比相对均衡，类别平均占比标准差相对模型W仅占30.76%，模型O中类别最高占比(20.86%)是最低占比(3.36%)的6.21倍，而模型W中类别最高占比(20.86%)是类别最低占比(0.25%)的83.44倍．模型O中高频用户的4个类别中，均有较大比例的乘客可在模型W中被分至出行距离更短的类别3中，而模型W的高频用户各个类别在模型O中分布相对均匀，两个模型的乘客分类结果仅有34.45%维持一致．图2给出模型O和W中各高频用户类别在模型W和O中的高频乘客数占比情况，横坐标O-1表示模型O中的类型1乘客，其余同理．

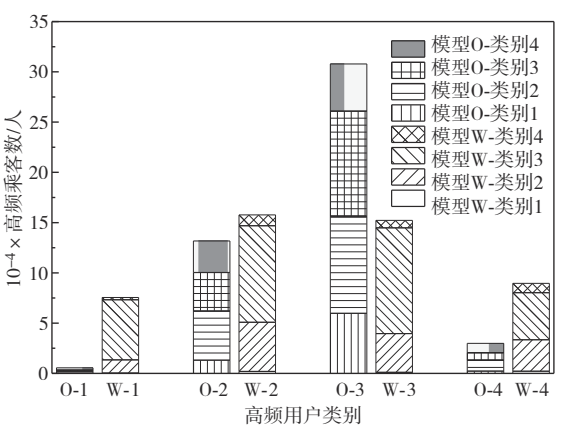

图2 高频用户中模型O和模型W的类别和占比情况

Fig. 2 Clustering results of high-frequency transit riders by categories of model O and model W.

提取高频用户中各类别乘客在出行规律、时间规律、空间规律及高峰规律的聚类中心表现并进行标准化处理，同时选取初次出行时间作为补充表现，各类别的聚类中心表现通过雷达图进行可视化．图3为模型O和模型W的高频用户类别聚类中心．可见，基于模型O和W的高频用户在出行规

律、时间规律、初次出行时间、空间规律和高峰规律上均表现出较强的出行特性．模型 W 中包含下车信息，加入了出行时长和距离指标，各类别间仅在出行规律上具有强差异性，更能有效区分特殊群体，如长距离出行乘客，但在时间规律、空间规律和高峰规律等维度上的区分度不及模型 O．

因此，加入下车信息的模型 W 虽然在小样本群体识别上具有一定优势，但是长距离出行等特殊群体在全群体中的比例较小，影响其他群体的人数分布及挖掘区分．模型 O 中不考虑下车信息的乘客分类分布较为均等，本研究提出的空间规律和高峰规律指标可有效呈现乘客规律；在时间规律、空间规律和高峰规律等维度上，各类别具有显著强差异性，能有效区分不同群体乘客在各维度上的特性．此案例有效验证了在仅采用上车信息等数据的情况下，模型 O 可实现对乘客的精细化分类．

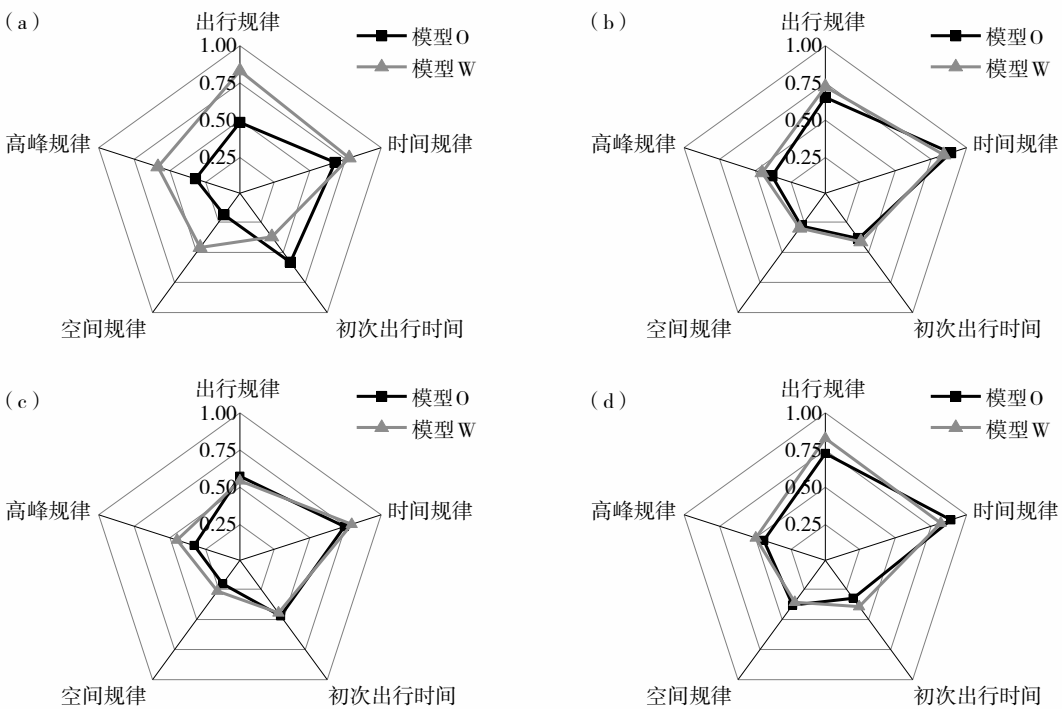

图 3　模型 O 和模型 W 的高频用户类别聚类中心　(a)至(d)分别为用户类别 1 至类别 4

Fig. 3　Clustering center of high-frequency transit riders by (a) category 1, (b) category 2, (c) category 3 and (d) category 4 of model O (black rectangular) and model W (grey triangular).

结　语

为研究下车信息缺失对公交乘客分类的影响，利用公交数据特征，从多维度构建无下车信息模型 O 与有下车信息模型 W 的分类指标组，包含通用指标、替代指标和附加指标．根据模型 O 和 W 的乘客分类指标体系，采用组合聚类并分别从 2 种数据集对乘客分类进行精细化挖掘．北京市公交因其采用分段计费而采集下车信息，成为挖掘有无下车信息对乘客精细化分类影响的理想对象．以北京市 2018 年 4 月中 28 d 的分段计费制公交用户刷卡数据作为研究对象，分别以隐藏及包含下车信息的数据作为模型 O 和 W 的数据集，通过分类指标体系进行乘客精细化分类并进行对比研究．

结果表明，① 针对有无下车信息数据所构建的分类指标组可有效刻画乘客多个维度的出行特性，结合两种聚类方法的乘客分类模型在有无下车信息的情况下均可实现乘客的精细化分类；② 应用包含下车信息的模型，乘客分类在各类别的比例及乘客出行距离上具有较大差异，小样本的长距离出行乘客群体易于被挖掘得到，类别间的乘客出行规律特征差异更明显；③ 应用缺少下车信息的模型，其分类结果占比相对均衡，在时间规律、空间规律和高峰规律上均在类别间呈现出显著的强差异性，分类效果更体现宏观特征．

因此，包含下车信息的乘客分类模型能识别低占比的小样本群体，缺少下车信息的乘客分类模型则在乘客特征差异性上表现更优．对于广泛采用公交一票制的城市，在不采集下车信息的情况下，直

接利用票务信息对公交乘客进行分类，也可获得全面宏观的分类效果，为乘客行为分析、公交服务优化及公交需求预测等提供参考．后续研究可考虑以多个不同规模的城市公交作为研究案例，并以多源数据拓展数据集，改进分类方法与聚类模型，在乘客分类的基础上探究乘车数据生成．

基金项目：广东省重点领域研发计划资助项目（2019B090913001）
作者简介：李军（1968—），中山大学副教授、博士生导师．研究方向：交通运输规划与管理．
E-mail: stslijun@mail.sysu.edu.cn
引　　文：李军，区静怡，赵文婷．下车信息对公交乘客精细化分类影响研究[J]．深圳大学学报理工版，2023，40(1)：109-117.

参考文献 / References：

[1] 刘敏，赵磊，刘祥锋，等．中小城市公交票价体系改革策略[J]．交通与运输，2020，36(1)：79-83.
LIU Min, ZHAO Lei, LIU Xiangfeng, et al. Reform strategy of bus ticket system in small and medium-sized cities [J]. Traffic and Transportation, 2020, 36(1): 79-83. (in Chinese)

[2] 焦朋朋，赵霞，张勇，等．基于交通大数据的移动模式分析综述[J]．中国公路学报，2021，34(12)：175-202.
JIAO Pengpeng, ZHAO Xia, ZHANG Yong, et al. Review of human mobility pattern analysis based on big transportation data [J]. China Journal of Highway and Transport, 2021, 34(12): 175-202. (in Chinese)

[3] CHAPLEAU R, TRÉPANIER M, CHU K K A. The ultimate survey for transit planning: complete information with smartcard data and GIS [C]// Conference of the IEEE Industrial Electronics Society. Lac d′Annecy, France: IEEE, 2008.

[4] CHOO S, MOKHTARIAN P L. What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice [J]. Transportation Research Part A: Policy and Practice, 2004, 38(3): 201-222.

[5] 杨敏，王炜，陈学武，等．工作者通勤出行活动模式的选择行为[J]．西南交通大学学报，2009，44(2)：274-279.
YANG Min, WANG Wei, CHEN Xuewu, et al. Activity pattern choice of work commuting trip by workers [J]. Journal of Southwest Jiaotong University, 2009, 44(2): 274-279. (in Chinese)

[6] 吴静娴，杨敏，陈学武，等．基于Nested Logit模型和蒙特卡罗法的通勤者活动链模拟[J]．交通运输工程与信息学报，2016，14(2)：76-82.
WU Jingxian, YANG Min, CHEN Xuewu, et al. Simulation of commuter′s activity chain based on Nested Logit model and Monte Carlo method [J]. Journal of Transportation Enginerring and Information, 2016, 14(2): 76-82. (in Chinese)

[7] 狄迪，杨东援．基于人群分类的城市公交走廊客流分配模型[J]．同济大学学报自然科学版，2016，44(2)：235-241，275.
DI Di, YANG Dongyuan. A passenger-classification transportation assignment model for urban public traffic corridor [J]. Journal of Tongji University Natural Science, 2016, 44(2): 235-241, 275. (in Chinese)

[8] 陈伯阳，蒋明清，四兵锋，等．基于乘客个体属性的地铁客流分配算法及实证研究[J]．北京交通大学学报，2015，39(6)：39-47.
CHEN Boyang, JIANG Mingqing, SI Bingfeng, et al. Individual attributes based assignment model and empirical research for urban subway network [J]. Journal of Beijing Jiaotong University, 2015, 39(6): 39-47. (in Chinese)

[9] 彭昌溆，周雪梅，张道智，等．基于乘客感知的公交服务质量影响因素分析[J]．交通信息与安全，2013，31(4)：40-44.
PENG Changxu, ZHOU Xuemei, ZHANG Daozhi, et al. Factors affecting bus service quality based on passenger perception [J]. Journal of Transport Information and Safety, 2013, 31(4): 40-44. (in Chinese)

[10] TSAI C Y, CHIU C C. A purchase-based market segmentation methodology [J]. Expert Systems with Applications, 2004, 27(2): 265-276.

[11] MAHRSI M, CÔME E, OUKHELLOU L, et al. Clustering smart card data for urban mobility analysis [J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(3): 712-728.

[12] 邹庆茹，赵鹏，姚向明．基于售检票数据的城市轨道交通乘客分类[J]．交通运输系统工程与信息，2018，18(1)：223-230.
ZOU Qingru, ZHAO Peng, YAO Xiangming. Passenger classification for urban rail transit by mining smart card data [J]. Journal of Transportation Systems Enginerring and Information Technology, 2018, 18(1): 223-230. (in Chinese)

[13] 李飞羽．城市轨道交通乘客行为特征分析及出行预测[D]．广州：华南理工大学，2020.
LI Feiyu. Feature analysis and travel forecast of passenger behavior in urban rail transit [D]. Guangzhou: South

China University of Technology, 2020. (in Chinese)

[14] MA Xiaolei, WU Yaojian, WANG Yinhai, et al. Mining smart card data for transit riders′ travel patterns [J]. Transportation Research Part C: Emerging Technologies, 2013, 36: 1-12.

[15] 林鹏飞，翁剑成，胡松，等. 公共交通乘客个体活动链的日相似性研究[J]. 交通运输系统工程与信息，2020，20(6)：178-183，204.
LIN Pengfei, WENG Jiancheng, HU Song, et al. Day-to-day similarity of individual activity chain of public transport passengers [J]. Journal of Transportation Systems Enginerring and Information Technology, 2020, 20(6): 178-183, 204. (in Chinese)

[16] 李军，邓红平. 基于公交IC卡数据的乘客出行分类研究[J]. 重庆交通大学学报自然科学版，2016，35(6)：109-114.
LI Jun，DENG Hongping. Classification of passenger′s travel behavior based on IC card data [J]. Journal of Chongqing Jiaotong University Natural Science，2016，35(6)：109-114.(in Chinese)

[17] 梁泉，翁剑成，周伟，等. 基于关联规则的公共交通通勤稳定性人群辨识[J]. 吉林大学学报工学版，2019，49(5)：1484-1491.
LIANG Quan，WENG Jiancheng，ZHOU Wei，et al. Stability identification of public transport commute passengersbased on association rules [J]. Journal of Jilin University Engineering and Technology Edition，2019，49(5)：1484-1491.(in Chinese)

[18] CHU K K A，CHAPLEAU R，TREPANIER M. Driver-assisted bus interview passive transit travel survey with smart card automatic fare collection system and applications [J]. Transportation Research Record，2009，2105(1)：1-10.

[19] 陈君，杨东援. 基于APTS数据的公交卡乘客通勤OD分布估计方法[J]. 交通运输系统工程与信息，2013，13(4)：47-53.
CHEN Jun, YANG Dongyuan. Estimating smart card commuters origin-destination distribution based on APTS data [J]. Journal of Transportation Systems Enginerring and Information Technology, 2013. 13(4): 47-53. (in Chinese)

[20] 刘靓. 普适计算环境下居民交通行为特征研究[D]. 上海：同济大学，2008.
LIU Liang. Research on inhabitant travel behavior under pervasive computing environment [D]. Shanghai: Tongji University, 2008. (in Chinese)

[21] 陈修远. 基于出行特性的公交乘客分类研究[D]. 成都：西南交通大学，2017.
CHEN Xiuyuan. Analyzing classification of bus passengers based on their trip characteristics [D]. Chengdu: Southwest Jiaotong University, 2017. (in Chinese)

[22] 李海波，陈学武，陈峥嵘. 基于公交IC卡和AVL数据的客流OD推导方法[J]. 交通信息与安全，2015，33(6)：33-39，95.
LI Haibo, CHEN Xuewu, CHEN Zhengrong. A method for estimating origin-destination matrix of public transit based on smart card and AVL data [J]. Journal of Transport Information and Safety, 2015, 33(6): 33-39, 95. (in Chinese)

[23] 曾志南. 基于智能交通卡数据的轨道出行乘客特征研究[C]// 共享与品质：中国城市规划年会，杭州：中国城市规划学会，2018.
ZENG Zhinan. Research on passenger characteristics of rail travel based on intelligent transportation card data [C]// Sharing and Quality: Proceeding of Annual National Planning Conference. Hangzhou: Urban Planning Society of China, 2018. (in Chinese)

[24] KIEU L M, BHASKAR A, CHUNG E. Passenger segmentation using smart card data [J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(3): 1537-1548.

【中文编辑：方圆；英文责编：淡紫】