文章编号:1009-6744(2018)01-0223-08 中图分类号:U239.5 文献标志码:A
DOI:10.16097/j.cnki.1009-6744.2018.01.033

基于售检票数据的城市轨道交通乘客分类

邹庆茹,赵 鹏*,姚向明
(北京交通大学 交通运输学院,北京 100044)

摘 要: 既有基于交通调查的乘客分类存在样本有限及分类标准主观性强等不足,本文以乘客真实出行记录为基础,从"消费行为"视角构建客观的乘客分类指标及方法.为满足大规模数据集处理需求,采用SPSS Modeler软件对全样本乘客进行聚类.选取北京轨道交通连续1个月自动售检票(AFC)数据进行实证分析,结果显示:将乘客分为5类时,聚类效果最佳;通过连续5个工作日聚类结果对比,验证了分类结果具有良好的稳定性.结合乘客分类结果进一步对北京市轨道交通低峰折扣票价策略下不同类型乘客的出发时间转移弹性进行测定.该研究提高了乘客分类客观性,能够为交通政策制定及运营策略评价提供方法支持.

关键词: 城市交通;乘客分类;两步聚类算法;自动售检票数据;城市轨道交通

Passenger Classification for Urban Rail Transit by Mining Smart Card Data

ZOU Qing-ru, ZHAO Peng, YAO Xiang-ming
(School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Traditional passenger classification methods based on traffic survey have drawbacks on limited sample and subjective standard, this paper constructs a new method and indexes in perspective of "consumer behavior" by using automatic fare collection (AFC) data. In order to meet the computation requirements of large data set, the SPSS Modeler is used to cluster the passengers. In case study, one month's AFC data of Beijing rail transit is applied and results shows that it is the best to cluster passengers in five classes, and the stability is verified by comparison with the clustering results in five consecutive days. The departure time transferring elasticities of different passenger types under pre-peak discount pricing strategy of Beijing transit are also analyzed. This study improves the objectivity of passenger classification and provides method support for traffic policy formulation and operation strategy evaluation.

Keywords: urban traffic; passenger classification; two-step clustering algorithm; automatic fare collection data; urban rail transit

0 引 言

乘客分类是研究出行者群体相似性行为及规律的重要方法.既有分类集中在以交通调查为基础的主观属性分类方面[1-2],如按出行目的、年龄、职业等.该分类的前提假设是相似个体属性或出行属性的乘客具有相似出行行为,存在主观性强和假设不合理等不足[3].本研究以轨道交通自动售检票数据(AFC)为基础,从出行强度、时间维度、空间维度及卡类型4个方面构建客观的分类指标,利用无监督聚类算法对乘客分类,从而提高分类的客观性.结合乘客分类结果,进一步对轨道交通峰前折

收稿日期:2017-06-08 修回日期:2017-07-23 录用日期:2017-08-17
基金项目:国家自然科学基金/National Natural Science of China(51478036&71701011);中央高校基本科研业务费专项资金资助/ Fundamental Research Funds for the Central Universities(2017RC032&2016JBM099);中国国家留学基金委资助(201707090039).
作者简介:邹庆茹(1987-),女,黑龙江齐齐哈尔人,博士生. *通信作者:pzhao@bjtu.edu.cn

扣票价策略下乘客出发时间转移弹性进行应用研究,为交通政策影响及评价提供方法支持.

AFC数据作为城市交通大数据中的重要组成部分,具备海量、持续、全样本特征,得以从个体视角精细化研究交通行为.Bagchi较早地阐述了AFC数据在交通管理中应用潜能[4];Pelletier进一步对AFC数据的应用进行了详细综述,其将应用划分为战略规划(如线网规划)、战术管理(如计划编制)和运营状态评价3个层次[5].在乘客分类或市场细分方面,Kieu基于密度的DBSCAN(Density-based Spatial Clustering of Applications with Noise)算法将乘客分为通勤乘客、出行起讫点(OD)稳定型、出发时间稳定型及不规律乘客4类,划分指标为出行起讫点和出发时间[6];在Kieu研究基础上,Venugopal利用OPTICS算法(Ordering Points to Identify the Clustering Structure)对乘客进行聚类,结果显示OPTICS算法比DBSCAN算法具有更高的准确性[7],该研究侧重于聚类算法效率和精度的提升,在分类指标方面略显单一.除AFC数据外,基于交易记录的用户分类在相关领域已较为丰富,如Tsai等从购买频率、消费金额、最近购买时间(RFM)构建指标对零售业顾客进行分类[3];张文欣改进RFM指标对航空客运市场进行了细分研究[8].综上分析,AFC数据挖掘与应用吸引了广大学者的关注,但从个体视角研究乘客出行行为及规律尚存在很大不足.公交IC卡号能对乘客进行标识,从而通过AFC记录能够捕捉单一乘客长期范围内的行为规律,为乘客出行模式识别、规律挖掘、行为变化追踪等精细化行为研究带来新的契机.

1 乘客分类指标构建

在RFM指标体系基础上,结合轨道交通AFC数据可获取的信息建立更为丰富的乘客分类指标.指标建立时以其能否表征不同类型乘客出行特征为基本原则.部分学者尝试通过AFC记录的内在关联推断其潜在信息(如出行目的)[9],但为避免推断误差造成的影响,构建指标时仅考虑AFC记录包含的直接信息.下面将从出行强度、时间维度、空间维度及卡类型4个方面阐述具体指标.

(1) 出行强度.

出行强度刻画乘客对轨道交通的利用程度,强度越大表明乘客对轨道交通忠诚度越高、依赖性越强.具体包括:①日均出行次数,刻画乘客对轨道交通的日均利用程度;②周均出行天数,刻画乘客1周内对轨道交通的利用程度;③周均出行天数标准差,刻画乘客出行的时间稳定性,一般通勤乘客出行稳定性高,而生活类出行波动大.

(2) 时间维度.

出发时间能在一定程度上反映乘客类型,如通勤乘客首末次出发时间一般位于早晚高峰.考虑到周末乘客出行不规律性,以及末次出发时间波动较大等因素,仅选取工作日首次出发时间作为表征变量.便于变量处理,将出发时间转化为分钟数(零点为起点).具体包括:①工作日首次出发时间,采用连续时期内乘客首次出发时间中位数表示;②工作日首次出发时间标准差,刻画乘客出发时间稳定性;③日均活跃时长,指1日内末次出行终止时间与首次出行起始时间差值.

前期研究发现1日内单次出行乘客占据一定比例[10],导致无法判断乘客是否为首次出行.为此,以中午12:00为分界点,在统计首次出发时间时不考虑晚于该时间点的出行记录.从较长连续时期来看,乘客每日出行仅为1次的概率较低,因此从统计角度来看对结果精度影响较小.

(3) 空间维度.

乘客在出行OD、出行距离方面均呈现一定特征.一般规律性乘客出行空间稳定性强.具体包括:①出行OD覆盖度,指出行OD对数与总出行次数的比值,OD覆盖度越小,乘客出行空间稳定性越高;②平均出行距离,出行距离与出行耗时、出行费用近似成线性关系(假定按里程计价),可用于刻画乘客的活动区域范围,本文选用出行耗时作为出行距离的替代指标.

(4) 卡类型.

卡类型能在一定程度上表征个体属性(乘客身份).IC卡常分为储值卡、学生卡、纪念卡、员工卡、临时卡(单程卡),以及车站工作卡,考虑到临时卡及车站工作卡的特殊性(无法对应固定乘客),该类

票卡对应的出行记录不在范围之内.

表1给出乘客分类指标的类型、取值范围等汇总信息.

表1　乘客分类指标汇总

Table 1　Summary of the passenger classification indexes

变量维度	名 称	单 位	类 型	取值范围
出行强度	周均出行天数	天/周	数值型	(0,5]
	周出行天数标准差	/	数值型	[0,2.5]
	日均出行次数	次/日	数值型	[1,9]
时间维度	首次出发时间中位数	min	数值型	[0,720]
	首次出发时间标准差	/	数值型	[0,210]
	日均活跃时长	min	数值型	[0,1 080]
空间维度	OD覆盖度	/	数值型	(0,1]
	平均出行耗时	min	数值型	(0,240]
个体属性	卡类型	/	分类型	{1,6,7,18,19}

注：卡类型1为储值卡，6为纪念卡，7表示员工卡，18表示学生卡1，19表示学生卡2；“/”表示无量纲.

2　北京市轨道交通乘客分类

考虑到聚类样本量十分庞大，经过一系列必选和分析，本研究采用高效的两步聚类算法来对乘客进行无监督客观分类.以北京市轨道交通乘客为对象进行分析，选取2016年12月连续1个月AFC记录为基础数据，计算每一张IC卡所对应乘客的出行指标(表1).为更好地分析1天内客流成分，选取2016年12月5日(周一)全天出行乘客为对象，共计约295万人.图1给出每日IC卡使用数统计结果，1个月内活跃IC卡总数约为1 249万张，单日平均活跃IC卡数约为269万张.图2给出12月5日内不同出行次数的客流量统计结果.

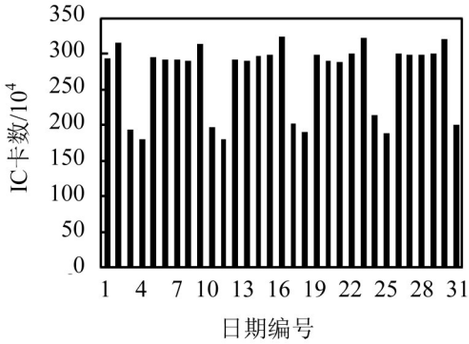


图1　IC卡使用数统计

Fig. 1　The number of used IC cards

2.1　指标筛选

指标筛选有助于缩减运算规模，提高聚类效率.在此，采用特征选择法对不同指标的重要程度进行分析，特征选择法即从众多输入变量中找出对分类结果有重要意义的变量.利用SPSS Modeler软件中的特征选择模型予以分析，结果如表2所示.一般特征值大于0.9的变量为重要变量.卡类型及平均出行耗时对分类结果影响较小，因此剔除这两个指标.

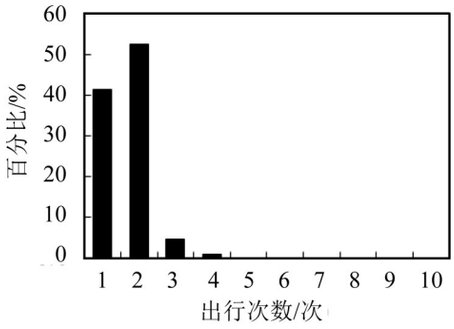


图2　不同出行次数下乘客比例

Fig. 2　Percentage of passengers with different trips

2.2　两步聚类过程

两步聚类算法是在BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies)算法基础上提出的改进算法.该算法特点包括：①对象间相似性采用似然距离测度，对分类变量和数值变量均适用；②采用CF树(Clustering Feature Tree)来提高聚类效率，能够解决大数据集的聚类问题；③能够根据Akaik判据(AIC)或贝叶斯判据(BIC)自动选择最优聚类数.

表2 出行指标特征值

Table 2 Characteristic value of travel indexes

编 号	出行指标	重要度	特征值
1	周均出行天数	重 要	1.000
2	周出行天数标准差	重 要	1.000
3	日均出行次数	重 要	1.000
4	首次出发时间中位数	重 要	1.000
5	首次出发时间标准差	重 要	1.000
6	日均活跃时长	重 要	1.000
7	OD覆盖率	重 要	1.000
8	平均出行耗时	不重要	0.679
9	IC卡类型	不重要	0.507

两步聚类包含2个阶段:①预聚类阶段,采用CF树生长的思想,在生成CF树的同时预先聚类密集区域的数据点,形成诸多子簇,该过程能够大幅提高聚类效率,如图3所示;②聚类阶段,以预聚类阶段得到的子簇为对象,利用凝聚法逐个合并子簇,直到得到期望的簇数量.

采用SPSS Modeler软件对1日内全样本乘客进行聚类.设定聚类数范围为2~15,对象间相似度采用对数似然距离,聚类准则采用BIC准则.结果显示将乘客分为5类时效果最佳.不同类型乘客占比如图4所示,表3给出不同类型乘客的聚类中心点.

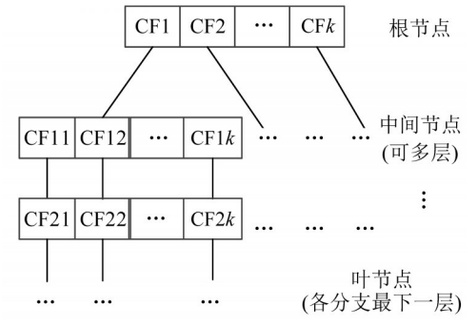


图3 CF树构建示意图

Fig. 3 The schematic for constructing CF tree

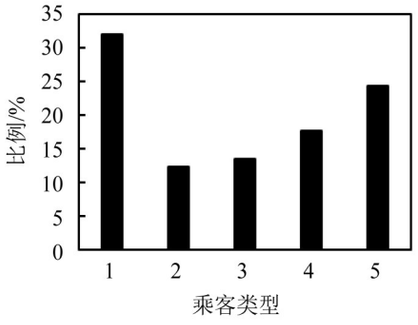


图4 不同类型乘客占比

Fig. 4 Percentage of different passenger types

表3 不同类型乘客聚类中心

Table 3 The center point for different clustering passengers

类 型	1	2	3	4	5
周均出行天数/天	4.64	4.49	3.21	2.29	0.96
周出行天数标准差	0.27	0.39	1.31	0.74	0.24
日均出行次数/次	1.88	2.25	1.76	1.72	1.41
首次出发时间中位数/min	484.98	474.90	474.61	554.62	364.47
首次出发时间标准差	13.56	49.99	23.07	62.23	5.85
日均活跃时长/min	737.16	777.39	700.68	550.47	171.59
OD覆盖度	0.17	0.33	0.30	0.62	0.80

为验证聚类结果稳定性,选取连续5个工作日为对象,对每日内乘客进行聚类,如表4所示.可以看出每日内客流成分趋于一致,验证了聚类结果具有较强的稳定性.

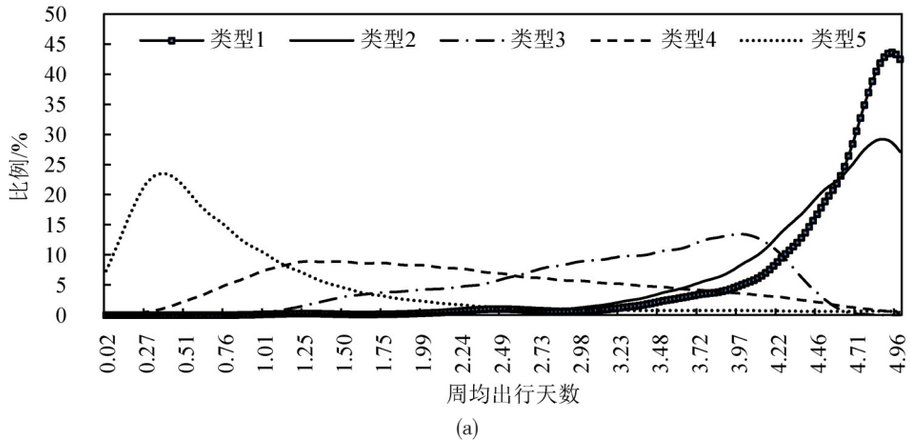
表4 不同工作日聚类结果对比

Table 4 Comparison of clustering results for different workdays (%)

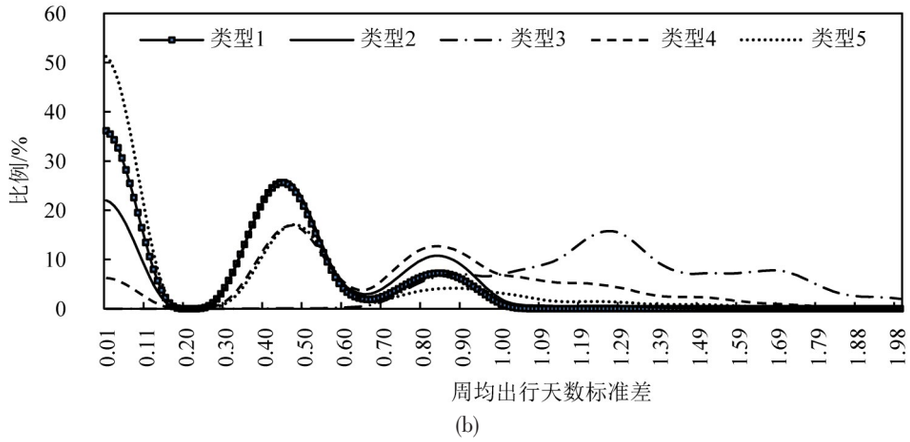
类 型	1	2	3	4	5
2016-12-05(星期一)	32.24	12.18	13.42	17.53	24.62
2016-12-06(星期二)	32.10	12.27	14.12	17.68	23.84
2016-12-07(星期三)	33.03	12.18	13.00	17.39	24.41
2016-12-08(星期四)	31.65	12.75	13.45	17.78	24.36
2016-12-09(星期五)	32.72	11.82	13.25	17.46	24.75
均 值	32.35	12.24	13.45	17.57	24.39

2.3 乘客出行特征分析

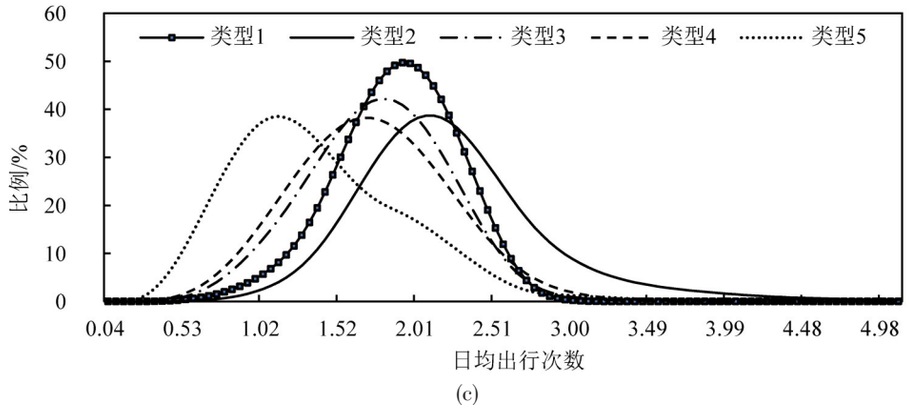
结合聚类结果对不同类型乘客的出行特征进行分析,具体出行指标分布如图5所示.


类型1
类型2
类型3
类型4
类型5
比例/%
0
5
10
15
20
25
30
35
40
45
50
0.02
0.27
0.51
0.76
1.01
1.25
1.50
1.75
1.99
2.24
2.49
2.73
2.98
3.23
3.48
3.72
3.97
4.22
4.46
4.71
4.96
周均出行天数

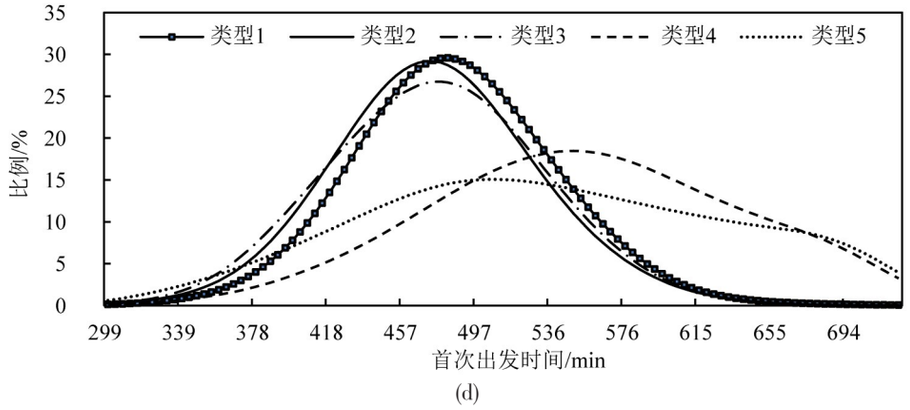
(a)


类型1
类型2
类型3
类型4
类型5
比例/%
0
10
20
30
40
50
60
0.01
0.11
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
1.09
1.19
1.29
1.39
1.49
1.59
1.69
1.79
1.88
1.98
周均出行天数标准差

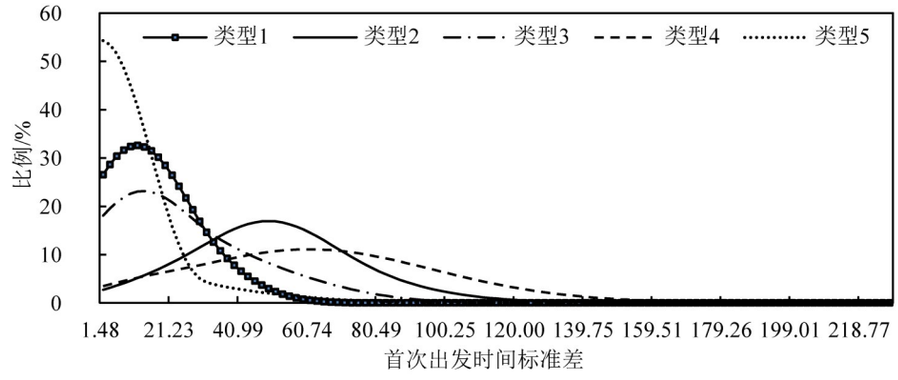
(b)


类型1
类型2
类型3
类型4
类型5
比例/%
0
10
20
30
40
50
60
0.04
0.53
1.02
1.52
2.01
2.51
3.00
3.49
3.99
4.48
4.98
日均出行次数

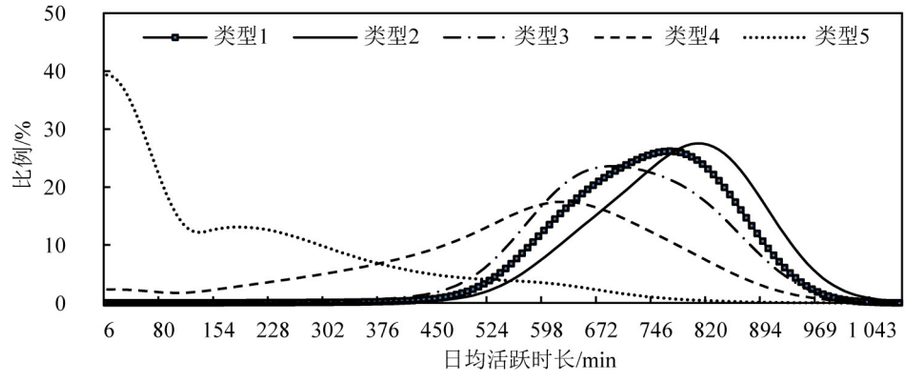
(c)


类型1
类型2
类型3
类型4
类型5
比例/%
0
5
10
15
20
25
30
35
299
339
378
418
457
497
536
576
615
655
694
首次出发时间/min

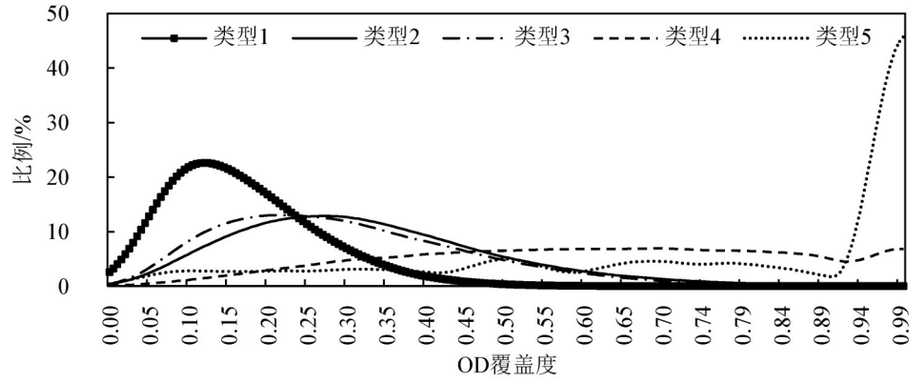
(d)



(e)



(f)



(g)

图5 不同类型乘客出行指标分布

Fig. 5 The distribution of travel indexes for different passenger groups

(1) 第1类乘客出行特征最为鲜明,表现为周均出行天数大(4.64天)、日均出行次数约为2次、首次出发时间和周均出行天数稳定性高(标准差小),且其占总体客流比例最大(约为32%),结合轨道交通客流主体为通勤客流,可认为该类乘客为标准通勤乘客.

(2) 第2类乘客与第1类乘客具有一定相似性,差异在于其日均出行次数更大(2.25次),周均出行天数标准差和首次出发时间标准差相对较大(时间稳定性较弱),且其OD覆盖度较大(出行空间稳定性较弱).种种特征表明该类乘客与业务型乘客相似,因此,将其定义为弹性通勤乘客(如业务性乘客).

(3) 第3类乘客与前2类的差异主要在:周均出行天数小,分布分散,出行时空稳定较弱,但出行OD较为稳定.因此,可将其定义为高频常乘客.

(4) 对比前3类乘客,第4类和第5类乘客具有明显差异,表现在出行频次低、时空不稳定;相比

而言，第5类乘客的活跃时间最短，出行OD极不稳定，出行频次很小.因此，可将其视为短期低频乘客，如旅游乘客、偶尔出行乘客；第4类乘客日均活跃时间和首次出发时间分布均较为分散，出行频次相比第5类乘客高，但其周均出行天数仍达到2.21天，将其定义为生活类乘客.

3　基于乘客分类的运营管理应用

乘客分类的目的在于从集计层面分析乘客的共性行为特征.在此，以北京地铁峰前五折票价优惠为应用场景，探求票价对乘客出发时间的影响.2016年末，北京地铁对八通线、昌平线、6号线共计24座车站在7:00前进站乘客实施五折票价优惠.考虑到八通线及昌平线在2015年底实行了峰前七折优惠，在此仅对6号线首次实施五折优惠的车站进行分析，包括北运河西、通州北关、物资学院路、草房、常营、黄渠、褡裢坡、青年路8座车站.首先，筛选出受影响的目标乘客，主要为居住在车站附近的居民，因此以乘客居住区是否为折扣票价车站为原则进行AFC记录抽取，居住区辨识算法参见文献[10].该过程可获取各站本源性交通需求，避免由其他车站到达该站而返回的客流影响.表5给出各站本源性乘客数及不同类型乘客的具体信息.

表5　目标车站不同类型乘客数统计

Table 5　The number of passengers in different types of analysis stations

车站编号	车　站	乘客数/人					
		类型1	类型2	类型3	类型4	类型5	总　计
1	北运河西	6 246	2 677	991	2 821	2 562	15 297
2	通州北关	2 399	808	312	1 024	832	5 375
3	物资学院路	10 038	3 755	1 450	4 485	3 597	23 325
4	草　房	9 280	4 582	1 886	4 435	3 827	24 010
5	常　营	7 674	3 901	1 482	3 468	3 760	20 285
6	黄　渠	7 848	2 879	1 070	3 084	3 145	18 026
7	褡裢坡	6 398	2 900	1 075	3 021	2 842	16 236
8	青年路	6 512	4 577	2 017	2 841	4 264	20 211

以折扣票价实施前后1个月为分析时段，对比乘客首次出发时间变化来判断乘客行为是否改变.折扣票价会导致部分乘客在出发时间上提前，但并非每次出行均提前.定义转移率来量化乘客出发时间转移弹性，计算公式为

$$p_i = \frac{n_i'}{N_i'} - \frac{n_i}{N_i} \tag{1}$$

式中：p_i为乘客i的出发时间转移率；n_i'为折扣票价实施后乘客首次出发时间早于7:00的次数，n_i为相应政策实施前的次数；N_i和N_i'为分析期内对应的总出行次数.

假设某类乘客集合为I，总人数为m，那么该类乘客的平均转移率$\bar{p}$为

$$\bar{p} = \sum_{i=1}^{I} p_i / m \tag{2}$$

图6给出各个目标车站不同类型乘客的出发时间转移率分析结果，可以看出：虽然各站客流量及客流结构存在差异，但不同类型乘客的转移率趋于一致，在一定程度上也说明了乘客分类的合理性.图7给出不同类型乘客的平均转移率，可以看出：①第1类和第2类乘客的转移率较低，表明其受折扣票价影响小，该类乘客出发时间约束较强，与实际情况保持一致；②第5类乘客出发频次小，且其出发时间分布较广，其受折扣票价的影响也较小；③第3类和第4类乘客的首次出发时间转移率较大，该类乘客出发时间弹性较大，因此受价格影响明显.

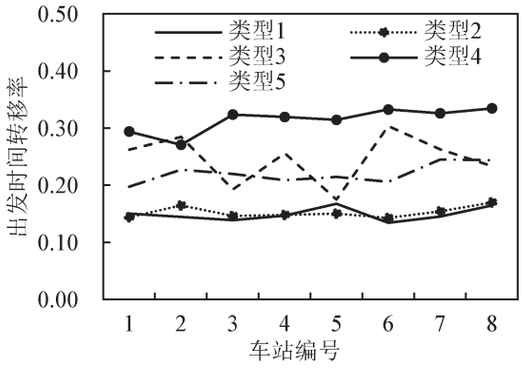


图6　各站乘客出发时间转移率

Fig. 6　Departure time transfer rate of each station

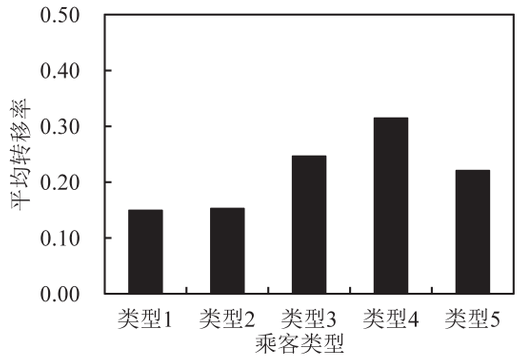


图7 乘客出发时间平均转移率

Fig. 7 Average departure transfer rate

在确定不同类型乘客出发时间转移弹性基础上,可进一步对折扣车站的客流转移效果进行测算.另外,还可对尚未实施折扣票价策略车站的客流结构进行分析,从而选取潜在转移效果明显的车站作为下一步策略实施的对象.

4 结 论

(1) 从乘客"消费行为"视角构建客观的乘客分类指标,并结合两步聚类算法对乘客进行分类;以北京市轨道交通连续1个月AFC数据进行实证分析,结果显示,将乘客分为5类时聚类效果最佳,分类结果具有良好的稳定性;并以北京地铁折扣票价策略实施为场景,分析不同类型乘客的首次出发时间转移弹性,为折扣票价策略的实施效果评估及推广应用提供了决策参考.

(2) AFC数据是交通系统中的一类重要数据源,能够为运营管理提供数据支撑,如何深入挖掘其潜在价值信息具有重要意义,后续将针对更长时间跨度内乘客的长期行为变化过程及规律进行深度挖掘.

参考文献:

[1] 吕红霞,王文宪,蒲松,等. 基于聚类分析的铁路出行旅客类别划分[J]. 交通运输系统工程与信息, 2016,16(1):129-134. [LV H X, WANG W X, PU S, et al. Classification of railway passengers based on cluster analysis[J]. Journal of Transportation Systems Engineering and Information Technology, 2016, 16(1): 129-134.]

[2] 史峰,邓连波,霍亮. 铁路旅客乘车选择行为及其效用[J]. 中国铁道科学, 2007, 28(6): 117-121. [SHI F, DENG L B, HUO L. Boarding choice behavior and its utility of railway passengers[J]. China Railway Science, 2007, 28(6): 117-121.]

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

[4] BAGCHI M, WHITE P R. The potential of public transport smart card data[J]. Transportation Policy, 2005, 12(5): 464-474.

[5] PELLETIER M, TRÉPANIER M, MORENCY C. Smart card data use in public transit: a literature review[J]. Transportation Research Part C: Emerging Technologies. 2011, 19(4): 557-568.

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

[7] VENUGOPAL S, DIVYA D. Transit passenger segmentation based on the travel patterns mined from smart card data using Optics algorithm[J]. International Journal of Advanced Information Science and Technology, 2016, 5(5): 49-56.

[8] 张文欣. 航空公司常旅客细分研究[D]. 南京: 南京航空航天大学, 2009. [ZHANG W X. Research on frequent flyer segmentation of airlines[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2009.]

[9] ZOU Q, YAO X, ZHAO P, et al. Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway[J]. Transportation, 2016: 1-26.

[10] 姚向明,赵鹏,韩宝明,等. 基于售检票数据挖掘的轨道交通乘客居住区辨识[J]. 交通运输系统工程与信息, 2016, 16(5): 233-240. [YAO X M, ZHAO P, HAN B M, et al. Home district identification for urban rail transit travelers by mining automatic fare collection data[J]. Journal of Transportation Systems Engineering and Information Technology, 2016, 16(5): 233-240.]