文章编号：1003-1421(2026)02-0142-10
引用格式：范家乐，景云.基于潜在类别模型的高铁旅客画像建模方法研究[J].铁道运输与经济，2026，48(2)：142-151.
FAN Jiale，JING Yun. Research on High Speed Railway Passenger Portrait Modeling Method Based on Latent Class Model[J]. Railway Transport and Economy，2026，48(2)：142-151.

基于潜在类别模型的高铁旅客画像建模方法研究

范家乐[1]，景　云[2]

(1.北京交通大学 詹天佑学院，北京　100044；2.北京交通大学 交通运输学院，北京　100044)

摘　要：随着高速铁路的不断发展，高速铁路旅客出行需求呈现出异质性特点。为更好地服务旅客出行需求，有必要针对旅客群体异质性的特点开展高速铁路旅客画像建模方法研究。首先，利用潜在类别模型对旅客进行分类，选取模型拟合指标BIC确定分类数目，选取熵衡量模型分类准确性；其次，根据样本描述性统计分析不同类别旅客的个人属性，构建MNL模型研究不同类别旅客出行选择行为；最后，准确剖析不同类别旅客的特征，提取不同类别旅客的画像语义标签。实际案例表明，京沪高速铁路旅客可划分为“舒适型”和“经济型”2类，“舒适型”旅客注重出行体验，关注出行服务质量，“经济型”旅客注重出行费用，关注价格合理性。本次调查中，2类旅客在不同属性上区分度高，模型分类效果好，高铁旅客画像构建精准。

关键词：高速铁路；旅客画像；潜在类别模型；样本描述性统计；MNL模型
中图分类号：U293.1　　**文献标识码**：A　　**DOI**：10.16668/j.cnki.issn.1003-1421.20250401002

Research on High Speed Railway Passenger Portrait Modeling Method Based on Latent Class Model

FAN Jiale[1]，JING Yun[2]

(1. Jeme Tienyow Honors College, Beijing Jiaotong University, Beijing 100044, China; 2. School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China)

Abstract: With the continuous development of high speed railway, passenger travel demands exhibit significant heterogeneity characteristics. To better serve passenger travel demands, it is necessary to conduct research on passenger portrait modeling methods based on the characteristics of passenger group heterogeneity. Firstly, the latent class model was applied to classify passengers, with the BIC selected as the model fitting indicator to determine the number of classifications, and entropy was used to measure model classification accuracy. Then, based on sample descriptive statistics, personal attributes of different passenger categories were analyzed, and MNL model was constructed to explore travel choice behaviors of different passenger categories. Finally, the characteristics of different passenger categories were accurately analyzed, and semantic labels for passenger profiles were extracted. A practical case study shows that passengers on the Beijing-Shanghai High Speed Railway can be classified into two groups: “comfort-oriented” and “economy-oriented”. Comfort-oriented passengers prioritize travel experience and pay close attention to service quality, while

收稿日期：2025-04-01
通信作者：景云(1981—)，男，陕西宝鸡人，北京交通大学交通运输学院教授。
基金项目：国家自然科学基金项目(52372300)；中国国家铁路集团有限公司科技研究开发计划课题(N2024X035)

economy-oriented passengers focus on travel costs and price rationality. In this survey, the two passenger groups demonstrate high discriminability in different attributes, with excellent classification performance and precise passenger portrait construction.

Keywords: High Speed Railway; Passenger Portrait; Latent Class Model; Sample Descriptive Statistics; MNL Model

0　引言

随着我国高速铁路的不断发展，高速铁路面临的主要问题由基础设施建设转为运营管理。现今，高铁旅客不只满足于空间位移需求，在选择客运产品时，还会对客运产品的票价、旅行时间、出发时间段、席别、列车等级提出要求[1]。客运服务产品设计的目标是满足旅客出行需求，而旅客出行需求表现出群体异质性。通过高铁旅客画像建模，对以高铁为出行方式的旅客进行特征刻画，分析旅客的个人属性、出行属性等信息，抽象出合适的语义标签，描述旅客类别，形成一个关于高铁旅客的信息全貌[2]，为探究不同类别旅客的出行选择行为、应用收益管理理论奠定基础。

目前高铁旅客画像构建主要使用聚类分析法及其改进算法，研究思路类似于市场细分。不同领域的学者针对市场细分已有比较深入的研究。刘彦麟等[3]为研究铁路票价浮动前后收益的平衡点，基于铁路用户的行为数据和交易数据，挖掘旅客群体特征，细分旅客需求，构建用户画像。Mohiuddin等[4]为提高其研究区域的共享单车利用率及交通公平性，从不同研究角度使用聚类算法进行若干次市场细分，最终提取出各类别用户特征，为政策建议提供理论依据。尹溪琛[5]为构建机场旅客画像与标签体系，基于K-means算法对旅客特征数据集进行聚类分析，并结合轮廓系数法和主成分分析法优化旅客聚类结果，为构建机场旅客出行链提供保障。张安忠等[6]基于多源数据构建城市轨道交通乘客画像，使用K-means算法对乘客进行分类，通过对不同类别乘客的特征分析，结合不同类别乘客的出行需求，最终将乘客分为通勤乘客、旅游乘客和生活类乘客，并在乘客画像的应用方面做出探索。然而以上研究均使用聚类算法，聚类算法基于距离度量对样本集进行分析，对初始聚类中心及数据输入顺序敏感，导致分类结果可解释性差。因此，相关学者使用潜在类别模型进行市场细分研究。乔珂等[7]分析潜在类别模型相较于聚类算法的优点，并基于京沪高铁客票数据进行潜在类别分析，得出结果后分别对各类别旅客进行样本描述性统计，赋予各类别旅客语义标签。Wen等[8]使用潜在类别模型对航空旅客进行细分，对比研究“台北—香港”和“台北—东京”2条航线，根据模型评价指标判断分类结果较好，说明了潜在类别模型在小样本数据集上的适用性。朱海等[9]考虑潜在类别模型与混合Logit模型结合，研究铁路疏解衔接系统中的用户细分，根据混合Logit模型的参数求解结果，为各类别用户赋予语义标签。苏焕银等[10]基于潜在类别模型研究广珠城际铁路旅客细分，根据分类相对误差评价分类效果，为提高铁路运输企业收益提供建议。以上研究赋予用户类别标签的方法包括样本描述性统计和参数解释，样本描述性统计通常提取个人属性特征，参数解释通常提取出行属性特征，但2种方法往往孤立使用，研究结论不全面。

铁路领域分析旅客出行选择行为多使用Logit模型及其改进模型，常用的Logit模型包括多项式Logit模型(MNL)、巢式Logit模型(NL)和混合Logit模型(ML)。王文宪等[11-12]分别使用MNL和NL对铁路旅客客运产品选择行为进行研究。Newman等[13]分析各类Logit模型的特点，明晰各类Logit模型的使用条件和使用场景。陈林[14]基于揭示嗜好(Revealed Preference，RP)和表明嗜好(Stated Preference，SP)调查，应用MNL和NL对成渝通道交通选择行为进行分析，并使用弹性分析对Logit模型显著变量的作用进行解读。程谦等[15]使用ML研究旅客对短途高速铁路列车的选择行为。Han等[16]考虑用户心理潜变量的影响，使用结构方程模型(SEM)进行潜变量的量化，并将其纳入到Logit模型的效用函数中，以此得出最终的研究结论。杨林远等[17]为研究城际旅客出行选择行为，在分析城际旅客出行链构成的基础上，考虑引入若干潜变量，构建SEM-Logit模型，通过实例分析表明了模型真实有效。以上研究在构建Logit模型之前均未对旅客进行分类，不同

类别的旅客出行选择行为可能大不相同，从而影响了研究结论的准确性，有进一步改进的空间。

在上述研究的基础上，首先利用潜在类别模型对旅客进行分类，其次同时使用样本描述性统计和MNL模型提取旅客特征，其中，样本描述性统计提取旅客的个人属性特征，MNL模型提取旅客选择行为特征，且分别为不同类别旅客构建MNL模型，对多组参数进行分析，最后根据旅客画像的不同特征命名语义标签，并使用Apriori算法挖掘强关联规则，提出了基于潜在类别模型的高铁旅客画像建模方法。

1　高铁旅客分类的潜在类别模型

1.1　模型构建

潜在类别模型是一种对属性均为离散变量的数据集进行分类的统计学模型，包括不可直接测量的潜在变量和可以直接测量的外显变量。模型中的潜在变量是一个类别变量，即潜在类别，其不同水平是指参数估计后所得到的不同类别值；外显变量也是一个类别变量，具体取值可以由问卷或量表调查得到，其不同水平是指受试者在各题目上所属的不同选项。

潜在类别模型利用统计原理进行概率分析，模型将潜在类别概率和条件概率视为参数，用以构建似然函数。其中，潜在类别概率和条件概率满足相应概率之和为1的规范性约束。

1.2　模型参数估计

模型通常使用极大似然法进行参数估计。在估计的过程中，通常使用期望最大化(Expectation Maximization，EM)算法[18]。

根据EM算法的原理，推导出潜在类别模型的迭代步骤[19]。假设调查问卷包含n个题目，每个题目均有m个水平。这n个题目便是外显变量，命名为A_1，A_2，…，A_n，其对应的水平分别记为a_{1i}，a_{2j}，…，a_{nl}。潜在类别变量用c来表示，并假设其分为K类。最终得到的结果是后验概率，如式(1)所示。

$$P\left(c=k\middle|A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right) \tag{1}$$

式中：k取值为1，2，…，K；i，j和l取值均为1，2，…，m。

根据贝叶斯公式得出式(2)

$$P\left(c=k\middle|A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)=\frac{P\left(c=k,\ A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)}{P\left(A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)} \tag{2}$$

假设外显变量具有局部独立性，参数$P(c=k)$，$P\left(A_1=a_{1i}|c=k\right)$，$P\left(A_2=a_{2j}|c=k\right)$，…，$P\left(A_n=a_{nl}|c=k\right)$设定初始值后，得到式(3)

$$P\left(A_1=a_{1i}|c=k\right)=\left[\sum_{j=1}^{m}\cdots\sum_{l=1}^{m}f\left(A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)P\left(c=k\middle|A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)\right]/P(c=k) \tag{3}$$

式中：$f\left(A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)$为数据集中该样本的频率；$\sum_{j=1}^{m}\cdots\sum_{l=1}^{m}f\left(A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)P\left(c=k\middle|A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)$为多重求和，即$n-1$个加和号嵌套作用在$f\left(A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)P\left(c=k\middle|A_1=a_{1i},\ A_2=a_{2j},\ \cdots,\ A_n=a_{nl}\right)$上。

同理求得$P\left(A_2=a_{2j}|c=k\right)$，$P\left(A_3=a_{3j}|c=k\right)$，…，$P\left(A_n=a_{nl}|c=k\right)$。至此，完成一次迭代过程，不断迭代直到前后2次的估计值差异小于一个指定的数值，则模型收敛。

1.3　模型分类数目确定与模型评价

通常会从潜在类别数为1的模型(Null model)开始估计，逐步增加潜在类别数，求解每个模型的参数估计结果和模型适配指标，进行比较，从而选择最佳的潜在类别数。常用的模型适配指标为赤池信息准则(AIC)值和贝叶斯信息准则(BIC)值，两者数值越小，模型适配度越高。其中，BIC值所起作用更大[8]。

因为潜在类别模型是统计模型，需要进行假设检验。常用的检验方法为似然比卡方检验(以下简称“G^2检验”)。

通常使用熵衡量分类准确性，熵的定义如式(4)所示。

$$E_K=1-\frac{\sum_{i}^{N}\sum_{k}^{K}-p_{kli}\ln p_{kli}}{N\ln K} \tag{4}$$

式中：N为样本量；K为类别数量；p_{kli}为个体i属于类别k的后验概率；熵的取值在0~1之间，超过0.8

表示模型可以接受[20]。

2　高铁旅客画像标签体系设计

2.1　不同类别旅客样本描述性统计

样本描述性统计是指用于概括和描述数据集基本特征的一组统计方法，包括了多种测量方式，用来反映数据集的集中趋势、离散程度和分布形态等。集中趋势指标用来表示一组数据的典型值，常见的有均值、中位数和众数；离散程度指标反映了数值之间的差异程度或分散程度，主要包括极差、方差、标准差和四分位差等；分布形态指标揭示数据分布的形状，如偏度和峰度等。

针对不同类别旅客的多组数据集进行对比分析，可以非常方便地明晰各类别旅客的基本特征。

2.2　不同类别旅客选择行为分析

在MNL中，决策者n一共有K个方案可供选择，其中$K \geqslant 3$。

引入U_{in}表示方案i的全部效用，引入V_{in}表示方案i的效用中可以观测得到的、确定性的部分，引入ε_{in}表示相应的随机效用。于是总效用可表示为固定效用和随机效用之和，如式(5)所示。

$$U_{in} = V_{in} + \varepsilon_{in} \tag{5}$$

假设所有的ε之间均相互独立，并考虑随机变量ε_{in}服从位置参数$\mu = 0$，尺度参数$\beta = 1$的Gumbel分布。于是可得到最终结果如式(6)所示。

$$P_n(i) = \frac{e^{V_{in}}}{\sum_k e^{V_{kn}}} \tag{6}$$

对于式(6)中的固定效用V_{in}，为参数估计的方便，通常表示成相关变量的线性组合。

通常使用McFadden拟R^2指标，如式(7)所示，衡量最终模型较所有参数为0的模型的改进效果，当R^2值为0.2~0.4时可认为模型拟合效果较好[9]。

$$R^2 = 1 - \frac{LL}{LL(0)} \tag{7}$$

式中：LL为迭代收敛后的对数似然函数值；$LL(0)$为所有参数设置为0时的对数似然函数值。

3　案例分析

3.1　数据收集与整理

为构建高铁旅客画像，在北京南站进行旅客问卷调查。调查时间为2025年1月5—11日，调查地点为北京南站2层候车大厅。调查的乘车区段包括“北京—上海”“北京—南京”“北京—济南”“北京—青岛”，出行距离分别为1 318 km，1 023 km，406 km，630 km。剔除明显不合理的问卷后，本次调查共收集有效问卷376份。

调查分3部分内容。①旅客个人属性调查。包括旅客的性别、年龄、月收入、职业、受教育程度。②旅客出行属性调查。包括旅客的出行目的、可接受票价浮动范围、行前接驳时间、候车时间、候车环境满意程度、可接受旅行时间偏差。③旅客客运产品选择结果。包括旅客所选车次、目的地、席别。

根据旅客所选车次，结合12306平台公布的信息，可以得到该车次的票价、出发时间段、旅行时间，而这三者一定程度上决定了一件客运产品的特性。

调查采用揭示偏好调查(RP调查)和陈述偏好调查(SP调查)混合模式，既使用RP调查方法，调查旅客实际出行行为，又在调查旅客可接受票价浮动范围和可接受旅行时间偏差时，使用SP调查方法，调查旅客出行意愿。

根据旅客个人属性和旅客出行属性调查数据进行统计，样本比例分布情况如表1所示。由表1可知，本次调查所得样本在性别、年龄、月收入、职业、受教育程度、出行目的上的比例分布较为均衡。

调查旅客的可接受票价浮动范围、候车大厅候车环境满意程度、候车时间、行前接驳时间、可接受旅行时间偏差时，将其视为离散有序变量，设置为单项选择题供旅客选择。①可接受票价浮动范围设置选项“0%~10%”“10%~20%”“20%~30%”“30%~40%”“40%~50%”，分别对应离散值1，2，3，4，5。②候车大厅候车环境满意程度使用李克特5级量表形式，分别将“非常不满意”“不满意”“一般”“满意”“非常满意”设置为离散值1，2，3，4，5。③候车时间和行前接驳时间设置选项“0~20 min”“20~40 min”“40~60 min”“60~80 min”“80~100 min”“100~120 min”，分别对应离散值1，2，3，4，5，6。④可接受旅行时间偏差是在给定旅客一些车次的旅行时间以供参考的基础上，调查旅客最高可以接受的旅行时间。通过12306平台查询旅客实际旅行时间，得到旅客可接受旅行时

表1　样本比例分布情况

Tab.1　Distribution of sample proportions

变量	变量选项	频数	百分比/%
性别	男	191	50.8
	女	185	49.2
年龄/岁	25以下	102	27.1
	26~35	92	24.5
	36~45	119	31.6
	46~55	53	14.1
	56以上	10	2.7
月收入/元	3 500以下	93	24.7
	3 500~6 500	34	9
	6 500~9 500	56	14.9
	9 500~15 000	62	16.5
	15 000~25 000	85	22.6
	25 000以上	46	12.2
职业	公职人员	6	1.6
	企事业单位人员	168	44.7
	个体经营户	55	14.6
	学生	91	24.2
	其他	56	14.9
受教育程度	高中及以下	23	6.1
	大学	218	58
	研究生及以上	135	35.9
出行目的	公务(可报销)	97	25.8
	商务(不可报销)	53	14.1
	休闲旅游	70	18.6
	探亲	57	15.2
	其他	99	26.3
出行距离/km	406	57	15.2
	630	20	5.3
	1 023	100	26.6
	1 318	199	52.9
总计		376	100

间与实际旅行时间的差值，再计算该差值与旅客实际旅行时间之比，根据比值的最大值和最小值均匀划分为5级，对应离散值1，2，3，4，5。

考虑到总的数据量为376，旅客最终所选择的客运产品数量不能太多。因旅客出行距离的差异，将列车票价、旅行时间设置为单位公里票价、单位公里旅行时间，根据最大值和最小值均匀划分为5级。根据样本比例分布情况所得到的数据特点，经过若干次对比试验，考虑试验效果，最终根据变量“列车出发时间段”和“分级处理后的单位公里票价”为依据对客运产品进行编号。

根据问卷题目选项的设置，“列车出发时间段”分为6:00—9:00，9:00—12:00，12:00—15:00，15:00—18:00，18:00—21:00，以此编号1，2，3，4，5。根据旅客所选列车，在12306平台上查出相关信息(票价、旅行时间等)，计算票价与出行距离之比得到“单位公里票价”。根据376个样本的最大值和最小值，等值划分为5级，分别编号为1，2，3，4，5。

经对比试验，设置客运产品编号。①列车出发时间段为1，2，3且单位公里票价为1，2时，客运产品编号为1。②列车出发时间段为1，2，3且单位公里票价为3，4，5时，客运产品编号为2。③列车出发时间段为4，5且单位公里票价为1，2时，客运产品编号为3。④列车出发时间段为4，5且单位公里票价为3，4，5时，客运产品编号为4。

因此，客运产品1的特征为早、午间出发，价格较低；客运产品2的特征为早、午间出发，价格较高；客运产品3的特征为下午、晚间出发，价格较低；客运产品4的特征为下午、晚间出发，价格较高。

3.2　潜在类别模型运行结果

为求解合适的潜在类别模型，从潜在类别数目为1(Null model)开始，使用Mplus软件进行探索性分析，比较模型的适配指标。模型适配度结果如表2所示。Mplus运行结果中的AIC值和BIC值表明，2-class model和3-class model是可以接受的。2-class model的AIC值大，BIC值小，而3-class model的AIC值小，BIC值大。依据文献[8]，BIC值所起作用较大，所以选择2-class model作为最终模型。在2-class model中，类别1旅客数量166，占比为44.15%；类别2旅客数量210，占比为55.85%。

Mplus参数估计结果如表3所示。潜在类别模型的熵值为0.889，大于0.8，表明模型可以接受。

3.3　高铁旅客画像标签体系提取

首先从2个角度对旅客特征进行刻画。

(1) 针对2个潜在类别旅客若干属性(性别、年龄、受教育程度、职业、出行目的)进行对比分析。不同类别旅客特征对比分析如图1所示。

由图1可知，类别1旅客男性偏多，年龄多分

表2　模型适配度结果
Tab.2　Model fit results

Model	χ^2	G^2	AIC值	BIC值	df	Number of Free Parameters
Null model	9 999.065	2 047.450	6 577.148	6 675.388	17 230	25
2-class model	5 547.644	1 520.810	6 016.219	6 216.628	17 211	51
3-class model	4 553.150	1 418.421	5 938.423	6 241.001	17 188	77
4-class model	4 505.029	1 413.069	5 940.527	6 345.275	17 164	103

表3　Mplus参数估计结果
Tab.3　Mplus parameter estimates

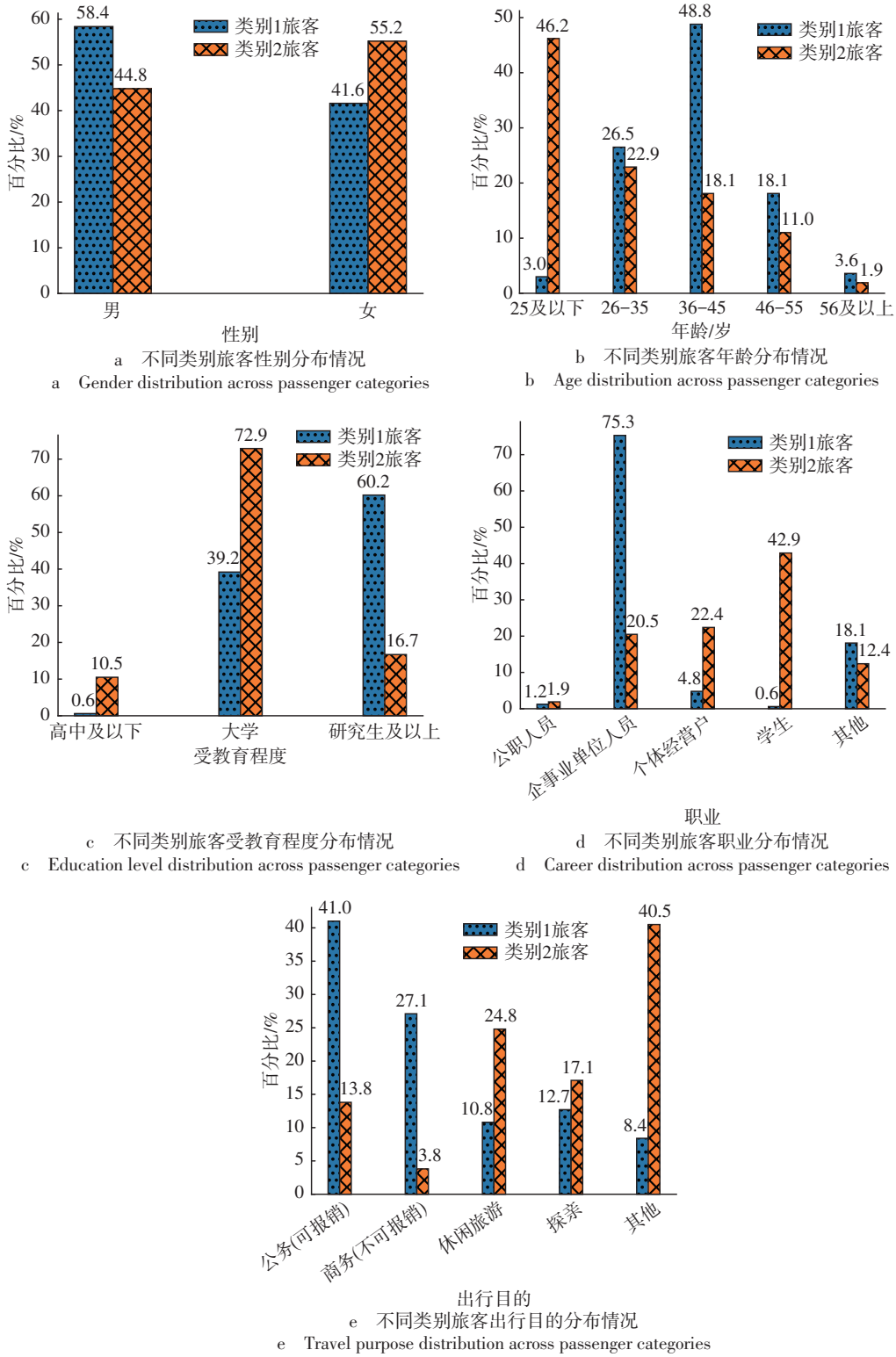
项目	编号	潜在类别	
		类别1	类别2
旅客月收入	1	0.008	0.434
	2	0.000	0.161
	3	0.046	0.229
	4	0.177	0.156
	5	0.509	0.005
	6	0.260	0.015
可接受票价浮动范围	1	0.061	0.639
	2	0.159	0.344
	3	0.465	0.016
	4	0.315	0.000
	5	0.000	0.000
旅客候车环境满意度	1	0.217	0.043
	2	0.267	0.280
	3	0.208	0.567
	4	0.308	0.110
	5	0.000	0.000
旅客候车时间	1	0.017	0.015
	2	0.525	0.154
	3	0.399	0.257
	4	0.045	0.329
	5	0.013	0.231
	6	0.000	0.014
行前接驳时间	1	0.067	0.014
	2	0.489	0.130
	3	0.376	0.251
	4	0.045	0.268
	5	0.007	0.236
	6	0.017	0.014
可接受旅行时间偏差	1	0.383	0.094
	2	0.474	0.421
	3	0.126	0.323
	4	0.017	0.133
	5	0.000	0.028
潜在类别概率		0.438	0.562

布在36~45岁，占比48.8%，受教育程度总体较高，职业多为企事业单位人员，占比75.3%，公务、商务出行偏多，两者比例之和高达68.1%；类别2旅客女性偏多，年龄在25岁及以下偏多，占比46.2%，受教育程度总体较低，职业多为学生，其次是个体经营户，两者比例之和达到65.3%，出行目的其他占比40.5%，多为学生回家(调查时认为“探亲”和学生“回家”是2种出行目的)，其次是休闲旅游，占比24.8%。2个类别的旅客在上述属性上区分度较高，分类效果较好。

（2）考虑将旅客月收入、可接受票价浮动范围、行前接驳时间[15]、可接受旅行时间偏差作为自变量，将候车时间、候车环境满意程度、年龄[12]作为协变量(协变量是指在模型中与主要自变量一同考虑的额外变量，本身不是建模的主要焦点，但可能对因变量有影响，或者可能会干扰模型对主要自变量效果的解释)，构建Logit模型。因为模型的自变量和协变量均为旅客个人属性变量，未涉及客运产品属性变量，使用NL模型和ML模型存在局限性，故使用MNL模型。

旅客选择高铁客运产品时，可能关注列车出发时间段，会对高峰期和平峰期产生预判。一些旅客，尤其年龄较大的旅客，不希望在高峰期时有较长的候车时间和拥挤的候车环境。因此，将这3个协变量纳入分析会使得模型参数估计更加准确。将编号为1的客运产品(Result1)作为参照组，应用Stata软件进行参数估计。Stata参数估计结果如表4所示。

类别1旅客的MNL模型McFadden拟R^2指标为0.381 1，大于0.2，拟合效果可以接受，因此可以对参数进行解读。随着旅客月收入水平的增加，类别1旅客更倾向于选择客运产品4。随着旅客可接受票价浮动范围的增加，类别1旅客同样偏好客运产品4。协变量方面，随着候车环境越来越舒适，



a　不同类别旅客性别分布情况
a　Gender distribution across passenger categories

b　不同类别旅客年龄分布情况
b　Age distribution across passenger categories

c　不同类别旅客受教育程度分布情况
c　Education level distribution across passenger categories

d　不同类别旅客职业分布情况
d　Career distribution across passenger categories

e　不同类别旅客出行目的分布情况
e　Travel purpose distribution across passenger categories

图1　不同类别旅客特征对比分析

Fig.1　Comparative analysis of passenger characteristics across categories

类别1旅客也会倾向于选择客运产品4。客运产品4下午、晚间出发，价格较高。下午、晚间通常为平峰期，候车大厅相对不拥挤，候车环境相对较好，而且通常票价高的客运产品旅行时间短。由此可知，类别1旅客追求出行品质，注重出行全过程的体验，对价格不敏感。

表4　Stata参数估计结果

Tab.4　Stata parameter estimates

变量	类别1旅客			类别2旅客		
	Result2	Result3	Result4	Result2	Result3	Result4
旅客月收入	1.913***	0.545**	2.277***	4.797***	0.417**	—
可接受票价浮动范围	—	—	1.301**	—	—	—
行前接驳时间	—	—	—	—	−0.235*	—
可接受旅行时间偏差	—	—	—	−2.007*	0.340**	—
候车时间	—	—	—	—	—	—
候车环境满意程度	3.229***	0.446*	6.309***	—	—	—
年龄	—	−0.419*	—	—	—	—
常数项	−22.475***	—	−45.673***	−22.564*	—	—
样本总数	166			210		
极大似然估计值(Log likelihood)	−129.802			−137.613		
似然比检验	LR chi(21)=159.87	Prob > chi2=0.000 0		LR chi(21)=69.15	Prob > chi2=0.000 0	
McFadden拟R^2指标	0.381 1			0.200 8		

注：“*”“**”“***”分别表示在0.1，0.05，0.01水平上显著；“—”表示参数不显著。

类别2旅客的MNL模型McFadden拟R^2指标为0.200 8，大于0.2，拟合效果同样可以接受，因此对参数进行解读。随着旅客月收入水平的增加，类别2旅客更倾向于选择客运产品2。随着旅客可接受旅行时间偏差的增加，类别2旅客更倾向于选择客运产品3。客运产品2早、午间出发，价格高；客运产品3下午、晚间出发，价格低。月收入增加后，类别2旅客会考虑选择价格高的产品，但不关注出行舒适性。可接受旅行时间偏差的增加表明类别2旅客愿意增加旅行时间以换取价格降低。由此可知，类别2旅客不过多关注出行体验，注重出行价格合理性。

综合上述分析，提取语义标签，构建高铁旅客画像。将类别1旅客命名为追求出行品质、注重出行体验的“舒适型”旅客，该类别旅客年龄和受教育程度均较高，职业多为企事业单位人员，侧重关注出行服务质量，对价格不敏感，铁路运输企业可以为该类别旅客提供优先候车服务，设计直达或停站次数少的客运产品；将类别2旅客命名为追求经济效益的“经济型”旅客，该类别旅客年龄和受教育程度均较低，职业多为学生和个体经营户，相较于出行品质、出行体验，更关注价格合理性，铁路运输企业可以为该类别旅客提供淡季折扣票，以吸引其出行。

3.4　关联规则分析

关联规则分析作为一种较为实用的数据挖掘技术，旨在从数据集中提取数据属性频繁出现的先验关联关系，其基本概念参见文献[21]，挖掘强关联规则通常使用Apriori算法。

为了方便铁路运输企业判断旅客类别，在jupyter notebook上使用Python运行Apriori算法，设置最小支持度为0.2，最小置信度为0.3[22]，共得到28个频繁项集，其中频繁1−项集为17个，频繁2−项集为10个，频繁3−项集为1个。根据频繁项集产生关联规则，进而得到强关联规则，共计24个，仅考虑“旅客类别”作为关联规则后项，则得到满足条件的强关联规则8个。强关联规则结果如表5所示。强关联规则分析的意义在于根据旅客所表现出的相关特征，可以帮助铁路运输企业非常方便地推理出旅客所属类别，进而针对不同类别旅客提供差异化服务。由表5的强关联规则可以看出，若旅客候车时间为20~40 min(相对较短)、可接受票价浮动范围为20%~30%(较高)、月收入在15 000~25 000元之间(较高)、可接受旅行时间偏差较低(期望更短的在途时间)，则旅客很有可能属于类别1。若旅客对候车大厅候车环境容易满足(对出行全过程体验关注较少)、可接受票价浮动范围为0% ~ 10%(低)、月收入低，则旅客很有可能属于类别2。

根据表5中的强关联规则，铁路运输企业能够实现仅依据旅客的少量信息就推断出旅客所属类别，并根据支持度、置信度和提升度来评价推断结果的有效性和准确性。

表5　强关联规则结果

Tab.5　Strong association rule results

强关联规则	支持度 (Support)	置信度 (Confidence)	提升度 (Lift)
“候车时间为20~40 min”⇒“旅客为第1类”	0.234	0.739	1.675
“候车大厅候车环境满意”⇒“旅客为第2类”	0.322	0.786	1.407
“可接受旅行时间偏差较低”⇒“旅客为第1类”	0.210	0.473	1.071
“可接受票价浮动范围为0%~10%”⇒“旅客为第2类”	0.356	0.924	1.655
“可接受票价浮动范围为20%~30%”⇒“旅客为第1类”	0.207	0.975	2.208
“月收入为15 000~25 000元”⇒“旅客为第1类”	0.226	1.000	2.265
“月收入为3 500元以下(学生群体)”⇒“旅客为第2类”	0.245	0.989	1.771
“可接受票价浮动范围为0%~10%”且“候车大厅候车环境满意”⇒“旅客为第2类”	0.213	0.988	1.768

4　结论

基于北京南站376份旅客调查问卷，构建潜在类别模型，参考拟合指标BIC将旅客分为2类，结果表明模型的熵值为0.889，分类结果可靠。通过样本描述性统计分析2类旅客的个人属性特征，类别1旅客年龄和受教育程度均较高，职业多为企事业单位人员；类别2旅客年龄和受教育程度均较低，职业多为学生和个体经营户。构建MNL模型研究2类旅客的出行选择行为特征，类别1旅客侧重关注出行服务质量，对价格不敏感，可命名为追求出行品质、注重出行体验的“舒适型”旅客；类别2旅客相较于出行品质和出行体验，更关注价格合理性，可命名为追求经济效益的“经济型”旅客。类别1和类别2旅客的模型McFadden拟R^2指标分别为0.381 1和0.200 8，2类旅客区分度高，模型拟合效果较好。通过高铁旅客画像建模，铁路运输企业可依托旅客群体异质性设计相应的高铁客运产品，对不同类别旅客提供差异化的出行服务，实行差异化的市场营销策略，提高自身经济收益。

考虑旅客心理潜变量的潜在类别集成模型和涉及客运产品属性的Logit模型将是进一步研究的主要内容。

参考文献：

[1] 孙国锋，景　云，马亚雯. 考虑旅客多维出行需求的动态列车开行方案优化[J]. 铁道学报，2022，44(11)：10-18.
SUN Guofeng，JING Yun，MA Yawen. Optimization of Dynamic Train Line Planning Considering Multi-Dimensional Travel Demand of Passengers[J]. Journal of the China Railway Society，2022，44(11)：10-18.

[2] 俞诚成，吕红霞. 基于大数据铁路旅客画像的客运营销策略研究[J]. 铁道学报，2020，42(8)：23-28.
YU Chengcheng，LYU Hongxia. Research on Passenger Transport Marketing Strategy Based on Big Data Railway Passenger Portrait[J]. Journal of the China Railway Society，2020，42(8)：23-28.

[3] 刘彦麟，周姗琪，颜　颖，等. 基于用户画像的铁路客运票价浮动平衡点研究[J]. 铁道运输与经济，2025，47(1)：210-214，224.
LIU Yanlin，ZHOU Shanqi，YAN Ying，et al. Research on Floating Balance Point of Railway Passenger Ticket Prices Based on User Portraits[J]. Railway Transport and Economy，2025，47(1)：210-214，224.

[4] MOHIUDDIN H，FITCH-POLSE D T，HANDY S L. Examining Market Segmentation to Increase Bike-Share Use and Enhance Equity：The Case of the Greater Sacramento Region[J]. Transport Policy，2024，145：279-290.

[5] 尹溪琛. 基于旅客画像和出行链模型的交通出行推荐方法及应用[D]. 北京：北京交通大学，2021.

[6] 张安忠，许心越，叶红霞. 城市轨道交通乘客画像构建方法及应用策略研究[J]. 铁道运输与经济，2024，46(1)：139-146，164.
ZHANG Anzhong，XU Xinyue，YE Hongxia. Research on Construction Method and Application of Urban Rail Transit Passenger Portrait[J]. Railway Transport and Economy，2024，46(1)：139-146，164.

[7] 乔　珂，赵　鹏，文佳星. 基于潜在类别模型的高铁旅客市场细分[J]. 交通运输系统工程与信息，2017，17(2)：28-34.
QIAO Ke，ZHAO Peng，WEN Jiaxing. Passenger Market Segmentation of High Speed Railway Based on Latent Class Model[J]. Journal of Transportation Systems Engineering and Information Technology，2017，17(2)：28-34.

[8] WEN C H，LAI S C. Latent Class Models of International Air

Carrier Choice[J]. Transportation Research Part E：Logistics and Transportation Review，2010，46(2)：211-221.

[9] 朱　海，罗　霞，陈　欣，等. 基于潜类混合Logit的铁路疏解衔接系统用户细分[J]. 华南理工大学学报(自然科学版)，2019，47(4)：67-75.
ZHU Hai，LUO Xia，CHEN Xin，et al. Customer Segmentation of Railway Egress Shuttle System Based on Latent Class Mixed Logit Model[J]. Journal of South China University of Technology (Natural Science Edition)，2019，47(4)：67-75.

[10] 苏焕银，陶文聪，彭舒婷，等. 基于潜在类别模型的城际高速铁路客运市场细分[J]. 铁道科学与工程学报，2021，18(6)：1410-1417.
SU Huanyin，TAO Wencong，PENG Shuting，et al. Passenger Market Segmentation of Inter-City High Speed Railway Based on Latent Class Model[J]. Journal of Railway Science and Engineering，2021，18(6)：1410-1417.

[11] 王文宪，倪少权，吕红霞，等. 基于非集计理论的高速铁路旅客乘车选择行为研究[J]. 交通运输系统工程与信息，2015，15(3)：13-18，43.
WANG Wenxian，NI Shaoquan，LYU Hongxia，et al. Boarding Choice for High Speed Railway Passengers Based on Disaggregate Theory[J]. Journal of Transportation Systems Engineering and Information Technology，2015，15(3)：13-18，43.

[12] 王文宪，陈钉钧，吕红霞，等. 基于Nested Logit模型的铁路旅客客运产品选择研究[J]. 交通运输系统工程与信息，2015，15(4)：140-146.
WANG Wenxian，CHEN Dingjun，LYU Hongxia，et al. Research of Transport Product Choice for Railway Passengers Based on Nested Logit Model[J]. Journal of Transportation Systems Engineering and Information Technology，2015，15(4)：140-146.

[13] NEWMAN J P，LURKIN V，GARROW L A. Computational Methods for Estimating Multinomial，Nested，and Cross-Nested Logit Models that Account for Semi-Aggregate Data[J]. Journal of Choice Modelling，2018，26：28-40.

[14] 陈　林. 基于Logit模型的成渝通道交通选择行为研究[D]. 成都：西南交通大学，2013.

[15] 程　谦，杨　光，胡启洲. 基于混合Logit模型的旅客对短途高速铁路列车选择行为[J]. 中国铁道科学，2021，42(2)：183-192.
CHENG Qian，YANG Guang，HU Qizhou. Passenger Choice Behaviors for Short-Haul HSR Trains Based on Mixed Logit Model[J]. China Railway Science，2021，42(2)：183-192.

[16] HAN Y，LI W Y，WEI S S，et al. Research on Passenger's Travel Mode Choice Behavior Waiting at Bus Station Based on SEM-Logit Integration Model[J]. Sustainability，2018，10(6)：1996.

[17] 杨林远，刘兰芬，刘佳伟. 基于出行链的城际旅客多模式出行选择研究[J]. 铁道运输与经济，2024，46(7)：61-68.
YANG Linyuan，LIU Lanfen，LIU Jiawei. Multi-Mode Travel Choice of Intercity Passengers Based on Travel Chain[J]. Railway Transport and Economy，2024，46(7)：61-68.

[18] 邱皓政. 潜在类别模型的原理与技术[M]. 北京：教育科学出版社，2008.

[19] 温忠麟，谢晋艳，王惠惠. 潜在类别模型的原理、步骤及程序[J]. 华东师范大学学报(教育科学版)，2023，41(1)：1-15.
WEN Zhonglin，XIE Jinyan，WANG Huihui. Principles，Procedures and Programs of Latent Class Models[J]. Journal of East China Normal University (Educational Sciences)，2023，41(1)：1-15.

[20] LUBKE G，MUTHÉN B O. Performance of Factor Mixture Models as a Function of Model Size，Covariate Effects，and Class-Specific Parameters[J]. Structural Equation Modeling：A Multidisciplinary Journal，2007，14(1)：26-47.

[21] 李海杰，苗　蕾，聂　磊，等. 基于关联规则和主成分分析的高铁旅客购票行为特征研究[J]. 铁道科学与工程学报，2023，20(6)：2013-2025.
LI Haijie，MIAO Lei，NIE Lei，et al. High Speed Railway Passenger Ticketing Behavior Characteristics Based on Association Rules and Principal Component Analysis[J]. Journal of Railway Science and Engineering，2023，20(6)：2013-2025.

[22] 陈权超. 基于大数据的高速铁路客流分析[D]. 成都：西南交通大学，2020.

(责任编辑　汪沁萱)