文章编号：1003-1421(2026)02-0152-12
引用格式：谭思伦，单杏花，卫铮铮，等.铁路旅客购票策略性行为识别与风险OD预警技术研究[J].铁道运输与经济，2026，48(2)：152-163.
TAN Silun，SHAN Xinghua，WEI Zhengzheng，et al. Strategic Ticket-Purchasing Behavior Identification of Railway Passengers and Risk OD Early-Warning Techniques[J]. Railway Transport and Economy，2026，48(2)：152-163.

铁路旅客购票策略性行为识别与风险OD预警技术研究

谭思伦，单杏花，卫铮铮，王洪业，郭根材

(中国铁道科学研究院集团有限公司 电子计算技术研究所，北京 100081)

摘 要：针对铁路客运市场化经营中“买短乘长/买长乘短”旅客策略性购票乘车行为，聚焦其引发的OD客流统计偏移、超载停运风险及运力虚靡问题，构建了“行程还原-状态标定-智能预警”的风险OD预警方法体系。首先，通过融合闸机数据还原旅客策略性行为的行程轨迹，解决OD客流统计偏差问题；其次，引入超停数据及候补数据，量化超停风险态和沉没损失态2类风险OD；最后，挖掘实时售票、列车产品、空间属性和时间属性4个维度特征，运用SMOTE-XGBOOST算法应对风险OD样本不平衡问题，搭建预售期内的实时预警模型。选取某铁路局集团公司的真实数据验证分析，结果表明该方法有效重构了旅客策略性行为，合理界定OD的风险状态，并在预售期内实现风险OD的实时动态预警。研究有助于推动铁路客运运营决策从“经验驱动”向“数据驱动”转变。

关键词：铁路客运；旅客策略性行为；买短乘长；买长乘短；OD；数据挖掘；风险预警
中图分类号：U293.1^{+}3 **文献标识码**：A **DOI**：10.16668/j.cnki.issn.1003-1421.20250807001

Strategic Ticket-Purchasing Behavior Identification of Railway Passengers and Risk OD Early-Warning Techniques

TAN Silun，SHAN Xinghua，WEI Zhengzheng，WANG Hongye，GUO Gencai

(Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: In view of passengers' strategic ticket-purchasing and travel behavior, specifically the phenomena of “short-ticket overriding” and “long-ticket underriding” in the context of market-oriented railway passenger operations, this study focused on the resulting OD passenger flow miscount, overload-induced suspension risk, and capacity underutilization. A risk early-warning system for OD passenger flow, integrating “itinerary restoration, state calibration, and intelligent prediction,” was established. Firstly, gate machine data was fused to restore strategic behavior-induced passenger itineraries and resolve OD statistical biases. Secondly, overstop train data and waitlist data were introduced to quantify two risk states: overstop-risk state and sunk-loss state. Finally, features from four dimensions (real-time ticketing, train products, spatial attributes, and temporal attributes) were extracted, and the SMOTE-XGBoost algorithm was employed to address sample imbalance for precise risk-OD prediction. A real-time early-warning model for the pre-sale

收稿日期：2025-08-07
基金项目：中国国家铁路集团有限公司科技研究开发计划课题(N2024X037)

period was built. By using real operational data from a regional railway group company, the results show that the proposed approach can effectively reconstruct strategic passenger behavior, reasonably delineate the risk states of OD, and realize real-time dynamic early-warning of risk OD during the pre-sale period. This research facilitates operational decision-making and advances railway passenger transport from "experience-driven" to "data-driven" management.
Keywords: Railway Passenger Transport; Strategic Passenger Behavior; Short-Ticket Overriding; Long-Ticket Underriding; OD; Data Mining; Risk Early-Warning

0　引言

随着人民生活水平的提高，旅客出行需求越来越多样化，从“走得了”向“走得快”“走得好”逐步转变[1]。然而，在春运、十一等节假日高峰期，铁路旅客运输需求旺盛，具有极度集中、极不均衡、单向流动显著特点，热门线路和时段往往“一票难求”，部分旅客在理想的车票无法购买时，通过“买短乘长/买长乘短”的市场化购票乘车策略性行为实现出行目的。对旅客个人，“买短乘长”及上车补票需要承担拒载风险，且没有固定席位舒适度大打折扣；而“买长乘短”则意味着额外的车票支出。对铁路部门，随着第三方软件购票的诱导和所谓购票攻略的信息传递，旅客微观层面的个体行为逐步演化成宏观层面的群体行为，对运输组织造成显著干扰：“买短乘长”易引发车厢拥挤，秩序混乱，旅客在A站购短途票至B站，但实际乘车至更远的C站。当大量旅客采取此策略时，列车在B—C区段实际载客量可能超出列车定员，触发复兴号等动车组的超员报警系统，威胁行车安全；“买长乘短”则导致运力资源浪费[2]，旅客购买长途票额外占用的运力资源，可能导致其他旅客的需求未满足，降低了铁路运输系统的整体服务效能。

1　问题提出

造成该现象的本质是运输系统在市场化运营中供需矛盾下的多层博弈。旅客个性化需求和铁路刚性资源供给存在天然冲突，高峰期运输资源有限，列车开行方案需设置中间站以带动沿途客流，同一列车需服务多个起讫点(Origin-Destination，简称OD)，而售票组织策略动态调整长途与短途的票额，进一步导致部分旅客需求难以即时满足。铁路旅客的需求选择行为及风险OD的形成过程如图1所示，旅客初始需求[3]未被满足时，常规情况下，需求转移路径包括调整日期、席别、车次，选择中转出行、航空等其他交通方式，或变更目的地、取消出行。特殊情况下，购买同一车次其他OD车票的策略性行为(后文均指代“买短乘长/买长乘短”)成为综合时间、价格、风险后的博弈选择。

精准把握旅客真实需求是破解问题的关键。旅客通过12306平台选择的出行OD，作为需求最直接的显性载体，既是出发站至到达站旅客运输需求

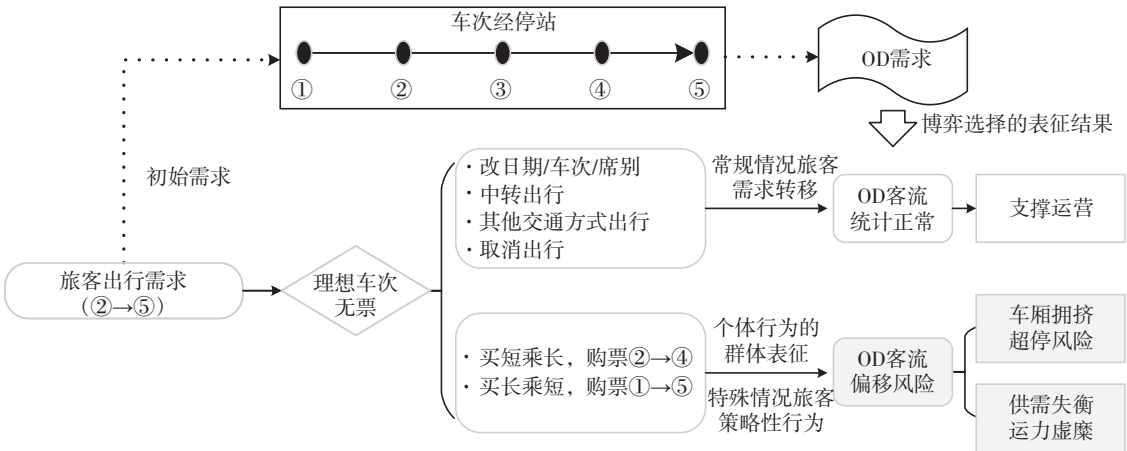

图1　铁路旅客的需求选择行为及风险OD的形成过程

Fig.1　Demand choice behavior of railway passengers and risk-OD formation mechanism

的量化表达，更是衔接旅客出行诉求与客运产品设计的核心要素。真实OD需求的获取通常依赖客流调查，但成本高昂；中国铁路客票发售和预订系统(以下简称“客票系统”)的销售数据记录了OD客流，但其表征的是旅客供需博弈后的选择结果[4]，无法还原“初始需求”，不过仍能支撑客流预测、票额预分、票价优化等业务实践工作[5]。尽管如此，如图1所示，旅客策略性行为造成OD客流的群体性偏移，若基于此类风险OD的偏移数据指导运营，不仅会加剧供需矛盾，还可能因偏移量累积引发超载停运(以下简称“超停”)风险与运力虚靡。

既有研究对此类问题的关注仍存在局限。现有研究中，仅孔德越等[6]建立概率模型预估旅客“买短乘长”的行为，部分研究针对特定售票过程进行预警和控制，武晋飞等[7]建立销售预警模型监控异常销售趋势，汪健雄等[8]针对冬奥服务列车在系统中设置超员控制策略。客运基础理论对需求和供给均有丰富的研究，需求侧中，一方面，销售数据作为OD需求的显性载体，被广泛应用于客流预测模型[9-10]；另一方面，学者们建立偏好选择模型，挖掘能力受限下的隐性旅客需求[11-12]，但是，现有研究均未量化策略性行为引发的数据偏移；供给侧中，学者们结合客流特征从票额预分[13-14]和开行方案调整[15-16]等角度研究合理的资源配置方案，优化目标通常聚焦收益最大化，忽视了策略性行为引发的安全风险和资源虚靡。借鉴城轨领域，AFC数据[17]及数据挖掘[18]等信息化技术逐步应用于解决地铁客运组织问题，打破数据壁垒是解决问题的必要途径。

综上，既有文献缺少针对“买短乘长/买长乘短”特殊场景的方法研究，现场应用中，铁路运营决策主要采用人工经验盯控和事后统计分析，缺乏系统性的事前预警机制。研究在大数据背景下引入数据融合与智能算法理论，构建实时预警模型，为铁路部门精准识别风险OD，实现了风险管控的前置化和智能化。

2　铁路旅客策略性行为识别与风险OD预警方法

2.1　整体框架

依托客运大数据平台，综合考虑各系统内部的业务规则，挖掘数据潜在价值和梳理数据关联，形成逻辑递进、数据互通的整体框架，基于数据驱动的旅客策略性行为识别和风险OD预警整体框架如图2所示，共包括3层，具体如下。

（1）旅客微观层。挖掘历史数据规律，分析旅客个体行为，引入客票系统的销售订单数据和闸机AFC核验数据，关联形成旅客从购票到出行的数据流程链路，标定“买短乘长/买长乘短”策略性行为，修正还原旅客真实的行程轨迹。

（2）OD宏观层。在还原所有旅客行程轨迹的基础上，聚合统计修正的OD客流，对比销售数据获取的OD客流，计算偏差形成客流统计漏报、客流统计虚高2种偏差情况；进一步，引入超停数据和能力虚糜数据，量化客流统计偏差的程度，识别所有OD为正常状态和风险状态。OD客流统计和状态标定，均以天为单位统计。

（3）实时预警层。考虑前2层依赖闸机记录、超停结果等“事后数据”的局限性，构建预售期内的实时预警机制实现“事前监测”。从需求和供给2方面引入特征构成数据集，运用SMOTE-XGBOOST算法训练智能模型。随预售期推进，实时补充新的售票数据迭代模型，输出3种OD状态预警结果，为优化票额策略和调整运力资源指明方向。

2.2　闸机数据修正旅客行程轨迹

一般地，旅客在一次完整的“购票-出行”的环节中，将与闸机发生至少3次信息交互，分别是实名制核验闸机、进站闸机和出站闸机。旅客i出行数据集$D_i(T, A_1, A_2, A_3)$，其中T代表购票信息，一般包括乘车日期、车次、出发车站、到达车站等，$A_j(j=1, 2, 3)$代表闸机信息，一般包括刷卡时刻、刷卡车站等，$j=1, 2, 3$分别对应上述3类闸机。常规情况下，A_1和A_2所对应的刷卡车站应该相同，且不同于A_3。

识别旅客策略性行为步骤如下。假设购票信息T中的出发车站和到达车站分别为O_T和D_T，对应闸机信息A_j中识别的刷卡车站为S_{A_j}，对旅客i的第k次铁路出行数据集进行异常行程轨迹判定，如式(1)—(3)所示。

$$abn_{i,k}^{\mathrm{IN}}:\ (O_T \neq S_{A_1}) \wedge (O_T \neq S_{A_2}) \tag{1}$$

$$abn_{i,k}^{\mathrm{OUT}}:\ D_T \neq S_{A_3} \tag{2}$$

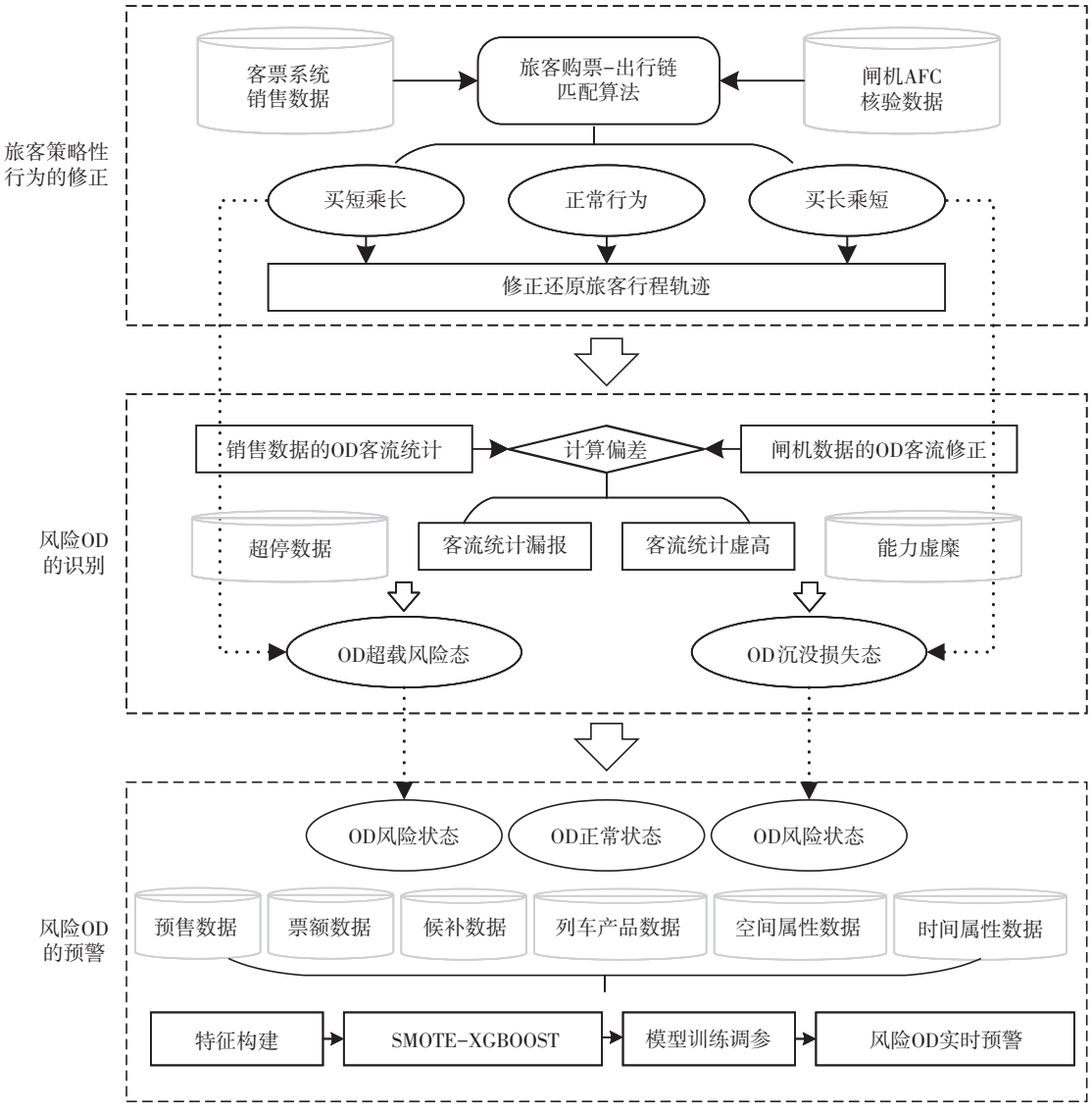

图2　基于数据驱动的旅客策略性行为识别和风险OD预警整体框架

Fig.2　Data-driven framework for strategic passenger behavior identification and risk-OD early-warning

$$abn_{i,k}: abn_{i,k}^{\mathrm{IN}} \vee abn_{i,k}^{\mathrm{OUT}} \tag{3}$$

式(1)解释了进站不匹配：条件$O_T \neq S_{A_1}$，$O_T \neq S_{A_2}$分别代表票面出发车站与实名制核验、进站闸机记录车站不一致；同样地，如式(2)所示，条件$D_T \neq S_{A_3}$代表票面到达车站与出站闸机记录车站不一致；如式(3)所示，总结旅客策略性行为和行程轨迹的关系，旅客策略性行为和行程轨迹的关系如表1所示。

刷卡闸机在实际应用过程中，部分出行数据集中存在空数据记录，需要对如下场景进行特殊处理。

（1）对于部分年龄较小的儿童旅客、年龄较大的老人旅客，考虑同行人关系，优先识别出行群体的其他旅客，替换识别结果。

表1　旅客策略性行为和行程轨迹的关系

Tab.1　Relationship between strategic passenger behavior and itinerary

策略性行为	闸机记录特征
买短乘长(晚出)	闸机出站序号>票面出站序号
买长乘短(晚进)	闸机进站序号<票面进站序号
买长乘短(早出)	闸机出站序号<票面出站序号

（2）对于部分中转换乘的旅客，其行程可能未被相应闸机记录，因此结合多段行程，从刷卡时间反推换乘车站，补充缺失的行程轨迹。

（3）对于进出站人流堆积或者遭遇突发事件的场景，客运人员会打开人工核验通道加快旅客的进站速度，该场景需要对具体的车次和上车/下车站进行标定，避免误判。假设车次t在车站z的出发客

流 $up_{t,z}^{\text{from}}$，到达客流 $up_{t,z}^{\text{to}}$，对应出行数据集中进站闸机数据为空的人数 $num_{t,z}^{\text{from}}$，出站闸机数据为空的人数 $num_{t,z}^{\text{to}}$，进行式(4)和式(5)判断。

$$(up_{t,z}^{\text{from}} > up_0) \wedge \left(\frac{num_{t,z}^{\text{from}}}{up_{t,z}^{\text{from}}} > \zeta\right) \tag{4}$$

$$(up_{t,z}^{\text{to}} > up_0) \wedge \left(\frac{num_{t,z}^{\text{to}}}{up_{t,z}^{\text{to}}} > \zeta\right) \tag{5}$$

式中：up_0 为最低客流值，一般取100人次；ζ 为比例阈值参数，一般取0.7，即当前车次在对应车站的进/出站客流大于100人次，且未刷卡人数的占比达到70%，属于人工核验。此参数综合车站进站闸机调研及数据统计分析确定。

2.3　风险OD的识别

2.3.1　OD客流统计偏差

个体层面的策略性行为可通过闸机数据直接识别，但OD层面的识别需通过量化客流偏差值界定边界。聚合所有旅客的修正轨迹得到闸机修正后的OD客流，对比销售数据OD客流，依次拆解闸机统计和销售统计客流值，并计算客流统计偏差，如式(6)—(8)所示。

$$N_{od}^{\text{gate}} = N_{od}^{\text{actual}} + \dot{N}_{od}^{\text{short-long}} + \dot{N}_{od}^{\text{long-short}} \tag{6}$$

$$N_{od}^{\text{sale}} = N_{od}^{\text{actual}} + N_{od}^{\text{short-long}} + N_{od}^{\text{long-short}} \tag{7}$$

$$\Delta_{od} = N_{od}^{\text{gate}} - N_{od}^{\text{sale}} = (\dot{N}_{od}^{\text{short-long}} + \dot{N}_{od}^{\text{long-short}}) - (N_{od}^{\text{short-long}} + N_{od}^{\text{long-short}}) \tag{8}$$

式中：N_{od}^{gate} 为闸机统计OD客流；N_{od}^{sale} 为销售统计OD客流；N_{od}^{actual} 为正常购票且正常乘车的旅客数量；$\dot{N}_{od}^{\text{short-long}}$ 为买短乘长从其他OD进入该OD的旅客数量；$\dot{N}_{od}^{\text{long-short}}$ 为买长乘短从其他OD进入该OD的旅客数量；$N_{od}^{\text{short-long}}$ 为买短乘长离开该OD的旅客数量；$N_{od}^{\text{long-short}}$ 为买长乘短离开该OD的旅客数量。

为验证不同策略性行为的贡献度，对2024年春运返程高峰日全国铁路OD数据进行了系统分析。为聚焦具有实际疏解价值的OD，采用的筛选条件包括：闸机日均统计OD客流和销售统计OD客流均>1 000人次、Δ_{od}>200人次，分别统计造成客流统计偏差的主导因素，旅客策略性行为对OD客流统计偏差的贡献度分析如表2所示。

因此，客流偏差 Δ_{od} 是多种策略性行为综合作用的结果。当 $\Delta_{od} > 0$ 时，策略性行为流入量>流出量，通常，以买短乘长流入占主导影响，体现了OD的能力不足。当 $\Delta_{od} < 0$ 时，策略性行为流出量>流入量，通常，以买长乘短流出占主导影响，体现了OD的能力过剩。

表2　旅客策略性行为对OD客流统计偏差的贡献度分析

Tab.2　Contribution of strategic passenger behavior to OD passenger flow statistical bias

偏差类型	筛选后OD样本数	主导因素分析	贡献度统计结果
闸机统计>票面统计	404	买短乘长流入值	贡献度超过75%的样本，236个
闸机统计<票面统计	423	买长乘短流出值	贡献度超过75%的样本，241个

2.3.2　超载风险态的标定

定义OD的超载风险态量化客流统计漏报程度，引入调度中心的超停数据(含超停日期、车次、车站、超停时长等)，结合列车的经停站信息，定位相关联的OD集合。

如式(9)所示，定义超载贡献系数，量化特定OD对某次超停事件的影响程度。该值越大，表明该OD的买短乘长客流对超停事件的“责任”越大，是事后追溯和事前防控的关键量化依据。

$$\sigma_{od} = \frac{\tau_k}{\tau_{\max}} \times \frac{N_{od}^{\text{short-long}}}{\upsilon_k} \tag{9}$$

式中：τ_k 为超停事件 k 的超停时长，若无关联到超停数据则为0；$\tau_{\max}$ 为历史最大超停时长，左项代表超停事件的严重程度，通常超停人数越多影响时长越长；$N_{od}^{\text{short-long}}$ 为在超停事件中 od 的买短乘长人数；υ_k 为超停人数，两者比值表征OD的责任占比。

通过分析历史上超停关联OD的 σ_{od} 概率分布，设定超载风险态的界定阈值 σ_{th} 为0.083。

2.3.3　沉没损失态的标定

定义OD的沉没损失态量化客流统计虚高程度。引入候补数据，统计各日期各OD未满足的旅客候补需求量，统计规则来自文献[12]，表征了OD的需求受限程度(含日期、OD、候补未兑现量)。

如式(10)所示，定义虚糜指数，综合衡量特定OD受买长乘短行为影响的运力浪费程度。该值越大，表明该OD掩盖的潜在短途需求越旺盛，票额错配越严重，是优化票额分配的决策指标。

$$\rho_{od} = \log_{10} hb_{od'} \frac{N_{od}^{\text{long-short}}}{N_{od}^{\text{sale}}} \tag{10}$$

式中：od' 为由于 od 买长乘短导致车上空座的闲置OD；$hb_{od'}$ 为闲置OD的候补未兑现需求。

考虑候补需求量差异过大且候补需求的边际效

应是递减的，为避免极端值主导模型，取对数表征候补需求受限程度，与右侧项$\frac{N_{od}^{\text{long-short}}}{N_{od}^{\text{sale}}}$客流虚高比例(即买长乘短导致的空座率)共同构成一个稳健的、符合业务逻辑的评估指标。

分析历史客流统计虚高OD的ρ_{od}概率分布，设定沉没损失态的界定阈值ρ_{th}为0.327。

2.3.4　风险OD的识别

结合客流偏差与风险量化结果，识别OD状态划分为3类。

（1）超载风险态：客流统计漏报且超载贡献系数达到阈值，即$\Delta OD_{od}>0 \wedge \sigma_{od}>\sigma_{th}$。

（2）沉没损失态：客流统计虚高且虚糜指数达到阈值，即$\Delta OD_{od}<0 \wedge \rho_{od}>\rho_{th}$。

（3）正常态：不满足上述情况。

2.4　基于SMOTE-XGBOOST的风险OD预警方法

2.4.1　特征构建

分析归纳预售期内旅客需求和票额供给博弈环节的影响因素，用于风险OD预警模型的特征构建，*pre*代表使用预售期内的数据份数(如使用提前14 d、提前7 d获取的数据，则*pre* = 2)合计特征数量为15 + 4*pre*个，风险OD预警模型的特征值如表3所示。

表3　风险OD预警模型的特征值

Tab.3　Feature values for risk-OD early-warning model

序号	特征量	特征量含义	特征数量
A1	已售客流	目标日期OD在提前天数内的售票订单统计值	*pre*
A2	余票数量	目标日期OD在提前天数内的剩余车票	*pre*
A3	票额能力	目标日期OD在提前天数内的共用票额量	*pre*
A4	候补未实现值	目标日期OD在提前天数内的未兑现候补订单量	*pre*
B1	车次数量	目标日期OD的全部/始发/终到/始发终到的列车数量	4
B2	旅行时间	目标日期OD的各列车旅行时间平均值/差异值	2
B3	热门车次数量	目标日期OD内热门时段/车次的列车数量	2
C1	车站/城市等级	OD对应出发站、到达站的车站等级及所在城市等级	4
C2	管理属性	OD为直通(跨铁路局集团公司)或管内(本铁路局集团公司)状态	1
C3	距离	OD之间的距离	1
D1	假期属性	目标日期在节假日的前期、中期和后期	1

（1）A类，实时售票。已售客流和候补未实现值直接体现旅客出行需求的强度，余票数量和票额能力则表征票额供给的规模与调整空间，4项特征均随预售期推进动态更新。

（2）B类，列车产品。刻画铁路向旅客提供的运力服务属性。车次数量反映列车产品丰富度；旅行时间体现产品的时间竞争力；结合历史购票数据标定各OD的高峰出行时段和标杆车，统计热门车次数量。

（3）C类，空间属性。聚焦OD的固有空间特征。车站和城市等级决定了基础客流规律和结构，管理属性影响票额调配效率，距离则关联旅客的选择偏好，3项特征共同构成OD客流空间层面的基础约束。

（4）D类，时间属性。捕捉周末及节假日周期对需求的影响，可适配不同阶段的OD客流波动规律，如春运前期的返乡OD高峰、后期的返程OD高峰。

2.4.2　XGBOOST算法

XGBOOST算法是基于梯度提升决策树算法的机器学习方法，其基本思想是集成学习，通过将多个分类与回归树进行组合，每棵决策树预测其之前决策树的预测结果和真实值的残差，迭代修正上一步误差，最终将所有决策树的预测值累加输出，得到准确率更高的模型[19]。

假设训练数据集为$\boldsymbol{D}=\{(\boldsymbol{x}_i,\ y_i)\}_{i=1}^{N}$，$\boldsymbol{x}_i$和$y_i$分别为样本$i$的特征向量和真实值，共有$N$个样本。XGBOOST算法的建模流程如下。

（1）初始化决策树$f_1(x)$，定义树的个数为M，那么每一棵决策树可表示为$f_m(x)$，$m=1,2,\cdots,M$，模型的最终输出为

$$\hat{y}_i=\sum_{m=1}^{M}f_m(x_i) \tag{11}$$

（2）确认最佳节点值。对每个样本计算负梯度(残差)，将上一棵决策树对样本i的残差值，更新为样本新的真实值，并将数据$\boldsymbol{D}_m=\{(\boldsymbol{x}_i,\ r_{m,i})\}_{i=1}^{N}$作为下一棵树的训练集，得到一棵新的决策树，其对应的叶子节点区域$\boldsymbol{R}_{m,j}$，$j=1,2,\cdots,J$，其中，J为决策树的叶子节点个数。此时，对$j=1,2,\cdots,J$个叶子节点，最佳节点值为

$$Obj_{m,j}=\sum_{i=1}^{N}l(y_i,\ \hat{y}_i^{(M)})+\sum_{m=1}^{M}\Omega(f_m) \tag{12}$$

式(12)中第1项为样本误差，第2项为正则项。样本总数共有N个，决策树共有y_i和$\hat{y}_i^{(M)}$为样本i的真实值和预测值，函数$l(\sim)$是样本误差函数。

（3）生成最终输出学习器。

$$\hat{f}(x)=f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}Obj_{m,j}J(x\in \boldsymbol{R}_{m,j}) \tag{13}$$

2.4.3　SMOTE算法的引入

超停数据本身较少，导致异常样本占比低，存在类别不平衡问题。SMOTE(合成少数类过采样技术)通过在少数类样本间插值生成新样本，平衡数据分布，提升XGBOOST的预测效果[20]。对少数类的训练集$\boldsymbol{D}^{y=C}k$，运用SMOTE操作。

（1）对每个样本$x\in \boldsymbol{D}^{y=C}$，以欧式距离为标准计算特征距离，在特征空间找到$k$近邻样本$xn$，$k$通常取5。

（2）对每一个近邻样本xn，依据式(14)分别与原样本生成新样本。

$$xnew=x+rand(0,\ 1)\times|x-xn| \tag{14}$$

式中：xn为k近邻样本；$xnew$为新生成样本。

（3）扩充少数类样本至合理比例，构建SMOTE平衡后的训练集$\boldsymbol{D}'$。模型搭建过程中训练集更新，而测试集不变。

2.4.4　模型训练流程和增量更新机制

（1）数据预处理：获取历史上识别后的风险OD，采集实时售票、列车产品等动态数据和空间属性、时间属性等静态数据，对动态特征进行离散值编码和数值归一化，经SMOTE平衡样本，划分训练集和测试集。

（2）模型训练：运用SMOTE-XGBOOST集成方法进行模型训练，考虑到不同风险状态对运输组织的影响权重不同，损失函数选取加权多分类交叉熵，强化模型对少数类样本的学习能力。优化学习率、最大树深度、迭代次数等超参数，采用混淆矩阵、ACC、召回率、宏平均$F1$值等指标对测试集的结果开展综合评价，迭代训练直到满足精度要求，保存模型。

（3）动态预测：预售期内按提前天数迭代，更新训练集中的实时售票数据，扩充特征数量，调用训练好的模型对预售期的风险OD进行预测，输出正常态、超载风险态、沉没损失态3类结果，自动触发预警信号。

3　验证分析

3.1　验证数据介绍

验证分析数据选取某铁路局集团公司的售票数据、闸机数据、超停数据、候补数据等，实验数据一览表如表4所示，覆盖18 833个直通OD、6 924个管内OD，共计25 757个OD。

表4　实验数据一览表

Tab.4　Overview of experimental data

数据类型	数据来源	核心字段	用途
售票数据	客票系统	订单号、乘车日期、OD、车次、席别	修正旅客行程轨迹、模型特征
闸机数据	AFC系统	刷卡日期、OD、车次、刷卡车站、闸机类型	修正旅客行程轨迹
超停数据	调度中心	超停日期、车次、车站、超停时长	标定OD状态
候补数据	客票系统	乘车日期、OD、候补兑现状态、候补需求数	标定OD状态、模型特征
票额能力	客票系统	乘车日期、OD、车次、总票额、剩余票额	模型特征
列车属性	客票系统	车次、停站、到发时刻、编组定员	模型特征

超停事件基本只发生于法定节假日和春运等高峰运营期，因此，只筛选节假日数据开展研究，定义节假日泛指法定节假日及其放假前后2天影响客流的日期及春运40天的日期，模型应用的时间范围定义如表5所示，包括元旦、春运、清明、五一、端午、中秋、十一，即研究所提方法的应用目标，实验最终选取2021年、2023年、2024年所有法定节假日及2025年春运，共计291 d，其中2022年的客流量受新冠疫情影响较大，不加入模型训练。

表5　模型应用的时间范围定义

Tab.5　Temporal scope definition for model application

节假日	春运	元旦	清明	五一	端午	中秋	十一
取数方式	固定40天	法定节假日及前后2天					
日期数量/d	40	7	7	9	7	7	11或12

技术实现基于以下平台和工具。数据处理环节，采用Spark分布式计算平台，使用Scala 2.12环境，处理海量客票数据，实现多源数据的高效融合；模型训练环节，使用Python 3.8环境，构建SMOTE-XGBOOST集成学习模型。

3.2　风险OD的识别效果分析

针对旅客策略性行为，从整体分析时间规律，2023年十一运输周期共12 d内，旅客策略性行为的占比超10%出现8 d，最高一天是2023年10月5日，达到14.3%，对应发生5起超停事件；2024年春运运输周期共40 d内，占比超10%出现15 d，最高一天是2024年2月17日，达到17.5%，对应发生33起超停事件，属于超停事件最多的一天。这表明策略性行为与超停风险存在显著时间关联性，且假期返程日易出现超停事件。

以2024年2月17日(春运返程高峰)为例，统计旅客策略性行为占比示例如表6所示。正常行为占比82.47%，策略性行为占比17.53%。其中“买短乘长”(含晚出、晚进且晚出)占比7.17%，“买长乘短”(含早出、晚进、晚进且早出)占比9.36%，反映出高峰期部分区段运力不足与票额分配失衡的双重问题。

表6　旅客策略性行为占比示例

Tab.6　Proportion distribution of strategic passenger behaviorss

旅客行程轨迹状态	细分状态	占比/%
正常	正常	82.47
买短乘长	晚进且晚出	0.48
	晚出	6.69
买长乘短	早出	2.78
	晚进	6.81
	晚进且早出	0.77

进一步，关联超停数据和候补数据，识别OD风险状态。通常客流较低的OD占绝大多数，且只承担了少部分的客流量，若加入模型会存在干扰数据质量的问题，筛选客流大于2 000人次的OD，获得107个直通OD、165个管内OD，共计272个OD，作为重点监测对象。

(1) 案例分析1：超载风险态OD识别。选取G6124次列车超停事件进行分析。列车停站方案为“永州—祁阳—祁东—衡阳东—长沙南—岳阳东”，在祁东站超停5 min，定员1 112人，实际载客约1 600人。根据式(7)计算各列车OD的超载贡献系数，超停事件示例中各列车OD的超载贡献系数如图3所示，其中“永州—长沙南”OD的责任占比达88.1%，统计数据中的最大超停事件影响时长为15 min，计算得到超载贡献系数为0.294，高于阈值0.083，被识别为超停风险态。进一步追溯发现，该OD的“买短乘长”客流主要来源于“永州—祁阳”“永州—祁东”等短途OD，图3中其贡献系数为负，表明需求被低估，验证了超载风险态识别的准确性。

(2) 案例分析2：沉没损失态OD识别。选取5个典型长途OD分析沉没损失态识别效果，沉没损失态部分OD的虚靡指数示例如表7所示，“深圳北—北京西”OD的虚靡指数达1.666，远高于阈值0.327，其客流虚高比例35.1%，实际出行集中于“广州南—北京西”OD，说明大部分旅客购买“深圳北—北京西”OD车票，实现“广州南—北京西”OD的出行目的；同时，候补受限程度达4.748，系数较大，说明深圳北至广州南区段存在严重的资源浪费，需向短途OD转移资源。

“福州南—深圳北”OD的客流虚高比例最高(37.0%)，说明票额配置与真实需求错配突出，长途

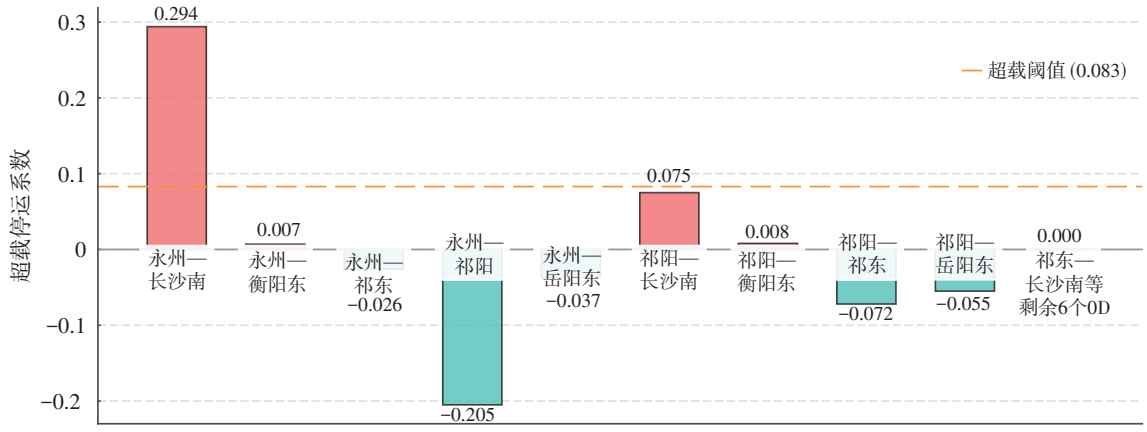

图3　超停事件示例中各列车OD的超载贡献系数

Fig.3　Overload contribution factors per train-OD in an overstop incident

票额供给过剩，而中途关键节点(莆田)的票额不足；此外，“南宁东—广州南”“厦门—深圳北”“柳州—广州南”等OD的虚糜指数均超过阈值，其实际出行分别集中于贵港、普宁、来宾北等中小城市至目的地的短途OD。这一现象表明，部分中小城市因直达票额不足，旅客被迫购买途经的长途车票，若继续使用虚高的客流指导运营方案，则会加剧此类中小城市的供需矛盾。

统计2024年2月17日的风险OD占比，春运返程高峰日的风险OD占比如表8所示，272个重点OD中，超载风险态32个，对应客流偏差量40 263人；沉没损失态29个，对应客流偏差量40 900人，受风险状态影响的客流超1/3，凸显了调控的必要性。

表7　沉没损失态部分OD的虚糜指数示例

Tab.7　Example values of underutilization index for selected sunk-loss state OD pairs

沉没损失态OD	客流虚高比例/%	主要的实际出行OD	候补受限程度	虚糜指数
南宁东—广州南	25.1	贵港—广州南	2.057	0.517
厦门—深圳北	30.2	普宁—深圳北	1.732	0.523
柳州—广州南	20.1	来宾北—广州南	1.699	0.341
福州南—深圳北	37.0	莆田—深圳北	2.530	0.937
深圳北—北京西	35.1	广州南—北京西	4.748	1.666

表8　春运返程高峰日的风险OD占比

Tab.8　Proportion of risk-state OD during peak return-trip days of Spring Festival travel season

风险OD状态	数量/个	客流统计偏差量/人	影响客流占比/%
超载风险态	32	40 263	17.1
沉没损失态	29	40 900	16.7
正常态	210	—	—

3.3　风险OD的预警效果分析

划分实验数据，搭建预售期内的预警效果测试条件：2021年、2023年、2024年法定节假日数据作为训练集(6 024个样本)，2025年春运数据作为测试集(1 476个样本)，用历史数据训练模型，预测未来风险状态，覆盖某铁路局集团公司管辖的所有相关线路。原始样本存在严重不平衡，正常态占比89.6%，采用SMOTE算法扩充少数类样本，其中k邻近取5，使超载风险态、沉没损失态样本各达到2 000个，平衡后训练集规模为9 395个样本，原始数据集和SMOTE处理后的实验数据划分如表9所示。

测试的预售场景是提前7 d预测，假设预测目标日期是2025年1月25日，则使用2025年1月18日及之前的实时数据，选取提前14 d，10 d，7 d的统计数据，即k=3，特征总数为27个，选择加权多分类交叉熵作为损失函数，设置权重占比为正常态∶沉没损失态∶超载风险态=1∶1.25∶1.5。

调参过程中采用网格搜索法确定XGBOOST最优超参数，XGBOOST超参数取值如表10所示，学习率控制每棵树的权重缩减系数，取值过大会导致模型过拟合，过小则收敛速度过慢；最大树深度限制单棵决策树的深度，用于平衡模型复杂度与过拟合风险；迭代次数指定集成树的数量，数量不足会导致欠拟合，过多则增加计算成本；分裂阈值指控制节点分裂的最小损失减少量，非负参数，值越大模型越保守；样本采样率是每棵树训练时的样本抽样比例，用于防止过拟合；特征采样率为每棵树训练时的特征抽样比例，降低特征间的共线性影响；L1正则化系数，对叶子节点权重施加L1惩罚，值越大正则化强度越高，有效抑制过拟合。

模型评价指标包括：①准确率(ACC)，整体分类正确率；②召回率(Recall)，异常状态的识别覆盖率；③宏$F1$值，平衡精确率与召回率的综合指标。

测试集风险OD预警效果混淆矩阵如图4所示，展示了对46个超载风险态和106个沉没损失态的预测对比图，可看出仅有3个超载风险态未被识别，说明模型能够精准挖掘特征，监测存在超载风险的OD；同样对于沉没损失态的召回率也达到93.4%，面对运力资源虚糜的OD也能及时提醒；部分正常态的样本存在误判，但比例不高，符合宁可误判不可漏判的标准。

表9　原始数据集和SMOTE处理后的实验数据划分

Tab.9　Partitioning of raw dataset and SMOTE-processed experimental data

状态	训练集				测试集			
	超载风险态	沉没损失态	正常态	总和	超载风险态	沉没损失态	正常态	总和
原始数据集	177	452	5 395	6 024	46	106	1 324	1 476
SMOTE处理的数据	2 000	2 000	5 395	9 395	—	—	—	—

表10　XGBOOST超参数取值

Tab.10　Hyperparameter settings for XGBOOST model

序号	参数	含义	搜索范围	最优值
1	learning_rate	学习率	[0.005，0.01，0.03，0.1]	0.01
2	max_depth	树的最大深度	[4，8，12，16]	12
3	n_estimators	树的数量	[50，100，200]	100
4	gamma	分裂阈值	[0，0.05，0.1]	0
5	subsample	样本采样率	[0.8，0.9，1.0]	0.8
6	colsample_bytree	特征采样率	[0.8，1.0]	0.8
7	reg_alpha	L1正则化系数	[0，0.1]	0.1

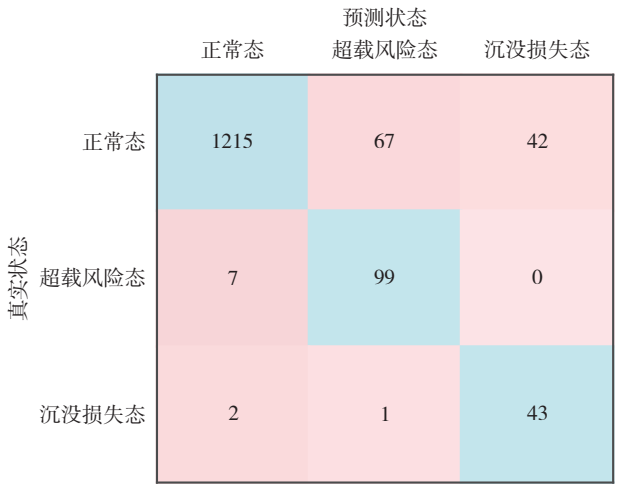

图4　测试集风险OD预警效果混淆矩阵

Fig.4　Confusion matrix for risk OD early-warning performance on test set

选取分类问题中的经典预测方法开展对比实验，逻辑回归为线性模型，适用于简单场景；支持向量机通过核函数处理非线性问题；MLP神经网络则依靠多层非线性变换拟合复杂模式；随机森林和XGBOOST属于集成算法，能融合多棵树的优势；而SMOTE-XGBOOST在XGBOOST基础上结合SMOTE算法，以应对样本不平衡情况。

对比结果显示，线性模型整体识别效果最差，随机森林、XGBOOST为基础的集成类算法整体性能显著优于支持向量机和MLP神经网络。随机森林、MLP神经网络、XGBOOST的宏$F1$值接近，但其在准确率和召回率上各有所长，而SMOTE-XGBOOST通过平衡样本数量，精准捕捉少数类异常特征，整体提升了各项评价指标，在保持和XGBOOST接近的准确率同时，从0.72到0.92大幅提升了召回率。这印证了方法在风险OD预警场景的适配性，既解决了正常态样本占比过高导致的模型偏向性，又通过集成学习挖掘了特征关联，实现“少漏报异常”的同时“少误报正常”。风险OD预警的多方法评价指标对比如表11所示。

表11　风险OD预警的多方法评价指标对比

Tab.11　Evaluation indicators for OD early-warning effects of multiple methods

算法	ACC准确率	召回率	宏$F1$值
逻辑回归	0.734	0.713	0.499
支持向量机	0.809	0.862	0.608
随机森林	0.892	0.746	0.669
MLP神经网络	0.916	0.641	0.667
XGBOOST	0.921	0.720	0.709
SMOTE-XGBOOST	0.919	0.929	0.779

综上所示，实验验证了所提方法的显著效能。识别精准性方面，分析了旅客策略性行为和风险OD状态的特点和分布，列举案例阐述了超载风险态和沉没损失态的量化识别，支撑精细化决策。预警优越性方面，SMOTE-XGBOOST模型在2025年春运测试中超载风险态的召回率达92.9%，对比传统方法可看出本研究方法合理解决了样本不平衡问题，达到了精准的预警识别效果，为高峰期铁路安全高效运营提供可靠技术支撑。

4　结论

面向“买短乘长/买长乘短”旅客策略性行为，剖析其本质为运输系统供需矛盾下旅客和铁路部门的博弈选择，揭示此类现象引发的OD客流统计偏移、超停风险及运力虚靡问题，构建了风险OD预警方法体系，量化旅客策略性行为影响程度并实时监测，为运营管理提供可靠的数据支撑，推动铁路客运决策从“经验驱动”向“数据驱动”转变。主要研究结论如下。

（1）将“买短乘长/买长乘短”问题从定性分析提升到定量建模层面，提出超载贡献系数和虚靡指数2个量化指标，将复杂的运营风险转化为可预警的数学模型。

（2）突破传统单一数据源局限，构建“售票-闸机-超停-候补”多源数据融合框架。

（3）紧密结合铁路售票、运输组织规律设计特征工程，考虑需求和供给中的多个相关指标。

（4）针对超停数据样本不平衡问题，系统性创新应用SMOTE-XGBOOST算法，面向实时预警的

业务场景，优化具体的权重占比、超参数设置。

研究验证表明，该方法可有效识别旅客策略性行为并实现预售期内的风险OD精准预警，进而推动预警OD应用于需求预测、票额分配和开行方案调整等业务，助力供需匹配与资源优化，减少旅客“买短乘长/买长乘短”的出现，提升购票乘车体验。

参考文献：

[1] 孙国锋，景　云，李和壁，等. 基于旅客多维出行需求的列车开行方案与票价联合优化方法[J]. 中国铁道科学，2024，45(6)：224-235.
SUN Guofeng，JING Yun，LI Hebi，et al. Joint Optimization Method of Train Line Planning and Ticket Pricing Based on Multi-Dimensional Travel Demand of Passengers[J]. China Railway Science，2024，45(6)：224-235.

[2] 刘帆洨，彭其渊，梁宏斌，等. 铁路客运票额预售控制决策模型研究[J]. 铁道学报，2018，40(1)：17-23.
LIU Fanxiao，PENG Qiyuan，LIANG Hongbin，et al. Control and Decision Model for Railway Passenger Ticket Pre-Sale[J]. Journal of the China Railway Society，2018，40(1)：17-23.

[3] 骆泳吉，刘　军，马敏书. 客票数据“受限”条件下旅客选择偏好估计[J]. 交通运输系统工程与信息，2015，15(2)：189-194.
LUO Yongji，LIU Jun，MA Minshu. Estimating Passenger Choice Preference of Railway Based on Censored Sales Data[J]. Journal of Transportation Systems Engineering and Information Technology，2015，15(2)：189-194.

[4] WANG X C，WANG H，ZHANG X N. Stochastic Seat Allocation Models for Passenger Rail Transportation under Customer Choice[J]. Transportation Research Part E：Logistics and Transportation Review，2016，96：95-112.

[5] JIANG X S，CHEN X Q，ZHANG L，et al. Dynamic Demand Forecasting and Ticket Assignment for High Speed Rail Revenue Management in China[J]. Transportation Research Record：Journal of the Transportation Research Board，2015，2475(1)：37-45.

[6] 孔德越，程　默，袁磊磊，等. 基于高峰期热门目的地识别的旅客买短乘长行为预估方法[J]. 铁路计算机应用，2024，33(8)：26-29.
KONG Deyue，CHENG Mo，YUAN Leilei，et al. Estimating Passenger's Act of Buying Short Distance Ticket for Long-Distance Travel Based on Identifying Popular Destinations during Peak Period[J]. Railway Computer Application，2024，33(8)：26-29.

[7] 武晋飞，单杏花，宋　超，等. 基于车次健康监测的铁路客票销售预警模型研究[J]. 铁道运输与经济，2022，44(6)：40-47.
WU Jinfei，SHAN Xinghua，SONG Chao，et al. Alarming Model of Railway Ticket Sales Based on Analysis of Train Health Monitoring[J]. Railway Transport and Economy，2022，44(6)：40-47.

[8] 汪健雄，阎志远，张瀛丹，等. 基于多模态信息融合的智能高速铁路旅客运输保障技术研究[J]. 铁道运输与经济，2023，45(2)：47-53.
WANG Jianxiong，YAN Zhiyuan，ZHANG Yingdan，et al. Intelligent Passenger Transportation Support Technology for High Speed Railway Based on Multi-Modal Information Fusion[J]. Railway Transport and Economy，2023，45(2)：47-53.

[9] 杨越迪，潘保霏，刘　军，等. 多分辨率多粒度时空特征提取的轨道交通短时OD客流预测方法[J]. 铁道学报，2024，46(11)：1-11.
YANG Yuedi，PAN Baofei，LIU Jun，et al. A Short-Term OD Passenger Flow Prediction Method for Rail Transit Based on Multi-Resolution and Multi-Granularity Spatio-Temporal Feature Extraction[J]. Journal of the China Railway Society，2024，46(11)：1-11.

[10] ZHANG Y，SUN K Y，WEN D，et al. Deep Learning for Metro Short-Term Origin-Destination Passenger Flow Forecasting Considering Section Capacity Utilization Ratio[J]. IEEE Transactions on Intelligent Transportation Systems，2023，24(8)：7943-7960.

[11] 郭　鹏，周　杰. 航线网络需求非限化估计中的广义随机偏好选择模型[J/OL]. 系统工程理论与实践，2025：1-20.(2025-05-15). https：//kns. cnki. net/KCMS/detail/detail. aspx?filename=XTLL20250515007&dbname=CJFD&dbcode=CJFQ.

[12] 谭思伦，卫铮铮，张　晗，等. 基于候补购票的铁路旅客出行需求挖掘方法研究[J]. 铁路计算机应用，2024，33(8)：19-25.
TAN Silun，WEI Zhengzheng，ZHANG Han，et al. Railway Passenger Travel Demand Mining Method Based on Candidate Ticket Purchase[J]. Railway Computer Application，2024，

33(8)：19-25.

[13] 单杏花，周亮瑾，吕晓艳，等. 铁路旅客列车票额智能预分研究[J]. 中国铁道科学，2011，32(6)：125-128.
SHAN Xinghua，ZHOU Liangjin，LYU Xiaoyan，et al. Research on Intelligent Pre-Assignment of Ticket Allotment for Railway Passenger Train[J]. China Railway Science，2011，32(6)：125-128.

[14] ZHU F，LIU S X，WANG R，et al. Assign-to-Seat：Dynamic Capacity Control for Selling High Speed Train Tickets[J]. Manufacturing & Service Operations Management，2023，25(3)：921-938.

[15] 赵　翔，武晋飞，单杏花. “一日一图”模式下高速列车开行方案与票额分配协同优化方法[J]. 铁道学报，2023，45(8)：10-17.
ZHAO Xiang，WU Jinfei，SHAN Xinghua. Collaborative Optimization of High Speed Train Line Planning and Seat Allocation under “Daily Train Timetabling”[J]. Journal of the China Railway Society，2023，45(8)：10-17.

[16] ZHAO S，WU R F，SHI F. A Line Planning Approach for High Speed Railway Network with Time-Varying Demand[J]. Computers & Industrial Engineering，2021，160：107547.

[17] 曾　璐，刘　军，秦　勇，等. 基于AFC数据的突发事件下城市轨道交通乘客路径决策研究[J]. 铁道学报，2019，41(8)：9-18.
ZENG Lu，LIU Jun，QIN Yong，et al. Route Choice Behavior of Passengers in Urban Rail Transit under Emergency Based on AFC Data[J]. Journal of the China Railway Society，2019，41(8)：9-18.

[18] ZHANG S X，ZHANG J L，YANG L X，et al. Physics Guided Deep Learning-Based Model for Short-Term Origin-Destination Demand Prediction in Urban Rail Transit Systems under Pandemic[J]. Engineering，2024，41：276-296.

[19] VLACHOGIANNIS D M，MOURA S，MACFARLANE J. Intersense：An XGBoost Model for Traffic Regulator Identification at Intersections through Crowdsourced GPS Data[J]. Transportation Research Part C：Emerging Technologies，2023，151：104112.

[20] SHANG W L，DING L，YANG X Y，et al. Complicated Imbalanced and Overlapped Data Oversampling Approach *via* Hypersphere Coverage and Adaptive Differential Evolution for Anomaly Detection of Industrial Internet of Things[J]. IEEE Internet of Things Journal，2025，12(10)：14002-14015.

(责任编辑　魏　然)