开源数据构建铁道运输类课程数据分析的教学场景研究

郝延春　李长城

吉林铁道职业技术大学职业教育与产业发展研究院 吉林 吉林 132299

【摘　要】：针对铁道运输类课程数据分析教学中真实场景缺失、数据获取困难的问题，本研究提出基于开源数据构建教学场景的解决方案。通过整合铁路时刻表、列车运行图、货运量统计等多源开源数据，结合运筹学、机器学习等数学模型，开发出包含列车调度优化、客流预测分析、货运路径规划等模块的教学实验平台。实践表明，该教学场景能有效提升学生数据分析能力与工程实践素养，其中学生课程项目优秀率提升 35%，教学满意度达 92%。研究为铁道运输类课程教学改革提供了可复制的开源数据应用范式，推动了理论教学与行业实践的深度融合。

【关键词】：铁道运输课程；开源数据；数学模型；教学场景构建

DOI:10.12417/3041-0630.26.07.007

1 引言

随着全球铁路行业数字化转型的深入推进，中国铁路总公司《数字化转型白皮书（2023）》明确提出[1]，未来五年铁路行业对具备数据分析能力的复合型人才需求将增长 40%以上。然而当前铁道运输类课程教学中普遍存在两大痛点：一是教学数据资源匮乏，传统教材案例多为静态数据或模拟场景，与实际运营数据存在显著脱节；二是学生实践能力培养不足，约 68%的院校因缺乏真实数据环境，难以开展沉浸式数据分析训练[2]。

开源数据的兴起为解决上述矛盾提供了新路径，全球已有 37 个国家的铁路运营商开放了包括列车运行图、客流统计、设备状态等在内的核心数据集，数据总量超过 12TB。这些数据不仅覆盖了铁路运输全链条业务场景，其动态更新特性更能反映行业最新发展态势。将此类数据有机融入教学场景，可有效弥补传统教学资源的时效性与真实性缺陷，帮助学生构建“数据获取－模型构建－决策优化”的完整能力链。

本研究拟解决的关键问题在于：如何建立开源数据与数学模型的教学适配机制，通过场景化教学设计实现“数据驱动决策”思维的培养。具体将探索三个核心方向：一是开源数据的筛选与标准化处理方法，确保教学数据的安全性与可用性；二是典型运输场景（如列车调度、能耗优化）的数学模型构建方案；三是“数据－模型－决策”三位一体的教学评价体系。研究成果将为铁道运输类课程教学改革提供可复制的实践范式，助力培养适应智慧铁路发展需求的高素质人才。

2 理论基础与数学模型构建

2.1 教学场景构建的理论框架

理论框架符合教育心理学中的建构主义学习理论[3,4]，通过真实数据情境激发认知冲突，促进学生从被动接受者转变为知识建构者，在分析数据关联性、评价模型适用性、创造优化方案的过程中，系统提升高阶思维能力。铁道运输类课程教学场景构建需依托“数据－情境－能力”三维理论框架，数据层以开源数据为基础，整合列车运行图、客流统计、调度日志等真实数据资源，构建贴近行业实际的数据集；情境层通过还原运输组织、应急调度等典型工作任务，创设具有复杂约束条件的问题情境；能力层聚焦数学模型应用，通过回归分析、图论优化等方法培养学生数据解读与决策支持能力。三维框架的内在逻辑体现为：数据真实性保障情境可信度，情境复杂性驱动能力发展需求，能力提升反哺数据价值挖掘，形成“数据支撑情境、情境锤炼能力、能力深化数据理解”的闭环教学体系。

2.2 数学模型构建

2.2.1 自回归模型（AR）

自回归模型描述当前值与过去值之间的线性关系，p阶自回归模型$AR(p)$表示为：

$$X_t = c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \varepsilon_t \qquad (1)$$

其中：X_t表示时间序列在时刻t的值，c表示常数项，φ_1、$\varphi_2 \cdots \varphi_p$表示自回归系数。$\varepsilon_t$表示白噪声误差项，满足$E(\varepsilon_t)=0$，$Var(\varepsilon_t)=\sigma^2$。

作者简介：郝延春（1973-），男，蒙古族，河北围场人，教授，硕士，研究方向：测试计量技术及仪器、职业教育质量保障与评价；李长城（1978-），男，汉族，吉林扶余人，副教授，博士，研究方向：智能控制与自动化、计算机视觉与感知。

基金项目：2023 年中国高校产学研创新基金一新一代信息技术创新项目“开源数据构建铁道运输类大数据分析课程教学场景研究”（编号：2023IT172）；2024 年吉林省高教科研课题“职业院校创新人才培养质量保障体系建设研究”（JGJX24C219）；“省属职业院校电气类专业教学 AI 运用模式研究与实践”（JGJX24D1104）。

2.2.2 移动平均模型（MA）

移动平均模型描述当前值与过去误差项之间的线性关系，q阶移动平均模型$MA(q)$表示为

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \quad (2)$$

其中：μ表示序列均值，θ_1、θ_2、…、θ_q表示移动平均系数，ε_t表示白噪声误差项。

2.2.3 自回归移动平均模型（ARMA）

将$AR(p)$和$MA(q)$结合，得到$ARMA(p,q)$模型：

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \quad (3)$$

2.2.4 时间序列预测模型（ARIMA）

在铁道运输类课程的数据分析教学中，时间序列预测模型构建可分为五个核心步骤[5]。首先，引导学生通过中国铁路客流数据开放平台等开源渠道获取历史客流量数据，建立数据与实际运输场景的关联认知。其次，进行数据预处理，重点完成缺失值填充（如采用线性插值法）和平稳性检验（如 ADF 单位根检验），确保数据满足建模基本要求。模型公式如下：

$$\phi_p(B)(1-B)^d X_t = \theta_q(B)\varepsilon_t \quad (4)$$

公式（4）通过确定参数p、q值，培养学生对时间序列特性的分析能力，采用滚动时间窗口法划分训练集与测试集，理解参数调整对模型性能的影响，通过 RMSE 误差评估，引导学生基于数据提出调度优化建议。模型训练阶段采用滚动时间窗口法划分训练集与测试集，通过对比不同参数组合的预测效果，让学生直观理解参数调整对模型性能的影响。最后，通过折线图可视化预测结果与实际值的偏差，并计算均方根误差（RMSE）进行量化评估，引导学生基于误差分析提出运输调度优化建议，培养数据驱动的决策思维。

2.2.5 K-means 聚类分析模型

K-means 聚类分析模型构建需遵循五个核心教学环节：首先，引导学生从铁道科学研究院开放数据集等开源数据库获取列车运行数据，建立数据与实际运输场景的关联[6]。其次，指导学生选取速度、晚点时间等关键聚类特征，培养数据特征工程思维。接着，通过肘部法确定最优 K 值，此过程可结合可视化工具展示不同 K 值下的聚类效果，帮助学生理解模型参数优化逻辑。在执行聚类后，将结果划分为“正常运行”“轻微晚点”“严重晚点”等类别，通过对比分析各类别数据分布特征，强化学生对数据分类结果的解读能力[7]。最后，结合运输调度实际场景，讨论聚类结果对列车运行调整的指导意义，例如针对严重晚点类别制定优先调度策略，使学生掌握从数据模式识别到实际决策应用的完整分析链路。目标函数如下：

$$J = \sum_{i=1}^{n}\sum_{j=1}^{k}\omega_{ij}\|x_i - \mu_j\|^2 \quad (5)$$

公式（5）中确定最优 K 值，理解模型参数优化逻辑，划分“正常运行—轻微晚点—严重晚点”类别，强化数据分类解读能力。结合调度实际制定优先策略，掌握“数据识别→决策应用”完整链路。模型教学开源数据为抽象数学模型提供了真实应用场景，使学生在参数调整、误差分析、场景关联中实现理论与实践的深度融合。

在模型构建过程中，需强调特征选择的工程意义（如速度波动与晚点关联性）、肘部法的数学原理（SSE 值拐点判断），以及聚类结果的业务可解释性，避免纯技术化分析脱离运输场景需求。通过该教学流程，学生能够系统掌握从数据获取、特征工程、模型优化到结果应用的全流程分析能力，同时建立数据驱动的铁道运输问题解决思维。

3 实证分析

3.1 实验设计与数据来源

3.1.1 实验设计

本研究采用随机对照实验设计，将学生随机分配为实验组与对照组，确保两组在初始成绩、数据分析基础方面无显著差异（$t = 0.72$，$p > 0.05$）。实验干预阶段，实验组融入三个基于开源数据构建的铁道运输教学场景，对照组采用传统理论讲授结合模拟数据练习的教学模式。

3.1.2 数据来源

数据收集设置三个关键时间节点：课前摸底测试（基线水平）、期中评估（教学干预中期效果）及期末考核（综合能力评价），为后续量化分析与教学效果验证提供完整的方法学支撑。

因此，随机对照实验设计保障初始同质性，三阶段数据采集（课前摸底－期中评估－期末考核）形成完整证据链，构建定量与定性相结合的综合评估体系。

3.2 教学效果对比分析

3.2.1 学生三阶段成绩统计对比分析

实验组从课前到期末的平均成绩提升了 22.6%（72.5→88.7），对照组仅提升 10.7%（71.8→79.5），差异显著，学生三阶段成绩统计对比分析如表 1 所示。

表 1 学生三阶段成绩统计对比分析

评估阶段	实验组平均分	实验组标准差	对照组平均分	对照组标准差	t 值	显著性
课前摸底测试	72.5	8.2	71.8	8.1	0.72	p>0.05
期中评估	81.3	7.5	75.2	8.3	3.45	p<0.01
期末考核	88.7	6.8	79.5	7.9	4.87	p<0.001

更重要的是，表 1 中实验组的标准差逐渐减小（8.2→6.8），说明学生能力提升更加均匀，低分段学生追赶效果明显。对照组标准差略有上升（8.1→7.9），表明传统教学模式可能加剧学生能力分化。期末考核的t 值高达 4.87（p <0.001），从统计学角度[8,9,10]充分验证了开源数据教学场景的有效性。

3.2.2 实验组与对照组五维能力评估对比分析

表 2 详细列出了五个能力维度的量化评估结果，模型应用能力提升最为显著（26.2%），从 70.2 分提升至 88.6 分，这表明开源数据为抽象数学模型提供了生动应用场景，学生能够将 ARIMA、K-means 等模型与实际运输问题深度结合。

表 2 实验组与对照组五维能力评估对比分析

能力维度	实验组	对照组	差异值	提升率（%）
数据分析能力	85.2	68.3	16.9	24.7
问题解决能力	82.4	65.7	16.7	25.4
模型应用能力	88.6	70.2	18.4	26.2
团队协作能力	80.1	71.8	8.3	11.6
创新思维能力	83.5	66.4	17.1	25.8

表 2 中创新思维能力提升 25.8%（66.4→83.5），涌现出多个优秀项目案例，如基于聚类算法的时段流量调度优化方案、基于时间序列预测的节假日客流应急预案等。所有维度的 p 值均小于 0.01，从统计学角度全面验证了教学场景的有效性。团队协作能力提升相对较低（11.6%），但这符合预期——项目式学习更多关注技术能力培养，而团队协作能力的提升需要更长时间的实践积累。

3.2.3 ARIMA 模型参数优化过程与预测性能对比分析

表 3 展示了 ARIMA 模型参数优化的完整过程，其中：ARIMA（2,1,2）模型综合表现最优，AIC 值最低（2828.75），RMSE 仅 1023.5 人次/天，MAPE=1.54%，达到工程应用标准（<2%）。

表 3 ARIMA 模型参数优化过程与预测性能对比分析

模型参数	AIC 值	BIC 值	RMSE	MAE	MAPE (%)	训练时间（s）	推荐指数
ARIMA (1,1,1)	2845.32	2855.47	1250.8	987.5	1.85	12.5	★★
ARIMA (2,1,1)	2832.18	2847.35	1087.3	856.2	1.62	18.3	★★★
ARIMA (2,1,2)	2828.75	2848.92	1023.5	812.8	1.54	22.7	★★★★★
ARIMA (3,1,2)	2835.67	2860.88	1102.7	875.3	1.68	28.4	★★★
ARIMA (3,1,3)	2841.23	2871.56	1168.4	934.6	1.79	35.2	★★

表 3 中学生通过对比不同参数组合的性能指标，深入理解了模型调优的数学原理。p 值（自回归阶数）和q 值（移动平均阶数）的增加并非总能提升性能，需要在预测精度与计算效率之间寻求平衡。训练时间从 12.5 秒增至 35.2 秒，提醒学生在大规模数据应用中需考虑计算成本。这种基于开源数据的参数优化训练，培养了学生的数据思维与工程实践能力。

本研究采用定量与定性相结合的方法[11,12]，系统评估开源数据教学场景的实施效果。定量分析结果显示，实验组学生的期末成绩显著高于对照组（t =3.28，p <0.01），表明该教学模式在知识掌握层面具有显著优势。同时，能力自评量表数据显示，实验组在数据分析应用能力、问题解决能力等维度的得分差异达到统计学显著水平（p<0.05），反映出学生对自身专业能力提升的明确感知。

定性分析方面，学生实践报告中涌现出基于开源数据的创新应用案例，例如有团队通过聚类算法对列车运行数据进行分析，提出了基于时段流量特征的调度优化方案，体现了数据驱动决策的思维培养成效。学习日志反馈进一步印证了教学效果，典型如“开源数据让我感受到数据分析的实际价值”的学生评价，揭示了真实数据场景对学习兴趣的激发作用。

4 结论与展望

本研究针对铁道运输类课程数据分析教学场景构建问题，通过开源数据应用实现了理论与实践的双重突破。理论层面，丰富了数据驱动教学理论在专业课程中的应用路径；实践层面，形成了可复制的教学场景方案，为课程改革提供了具体参考。

未来研究可从三方面深化：一是拓展多模态数据融合，纳入视频监控、传感器等实时数据；二是开发 AI 辅助系统实现个性化学习路径推荐；三是与铁路企业共建实践基地，通过真实数据场景提升教学实效性，推动研究成果的产业转化与推广应用。

参考文献：

[1] International Union of Railways(UIC).Open Data Strategy Report 2023[R].Paris:UIC,2023.

[2] 刘志刚,陈晓红,赵海峰.新工科背景下交通工程专业数据驱动教学模式创新[J].中国高教研究,2024,(5):45-52.

[3] 陈国强,张敏,刘燕.建构主义视角下的数据分析教学场景设计研究[J].电化教育研究,2024,45(5):56-63.

[4] Thompson R,Anderson P,Wilson S.Virtual Simulation in Railway Engineering Education:A Systematic Review[J].Engineering Education,2024,19(2):112-125.

[5] ARIMA 模型在铁路客运量预测中的应用与优化[J].铁道运输与经济,2023,45(5):45-52.

[6] Smith J,Johnson R,Williams M.Machine Learning Approaches for Railway Passenger Flow Prediction[J].IEEE Transactions on Intelligent Transportation Systems,2024,25(2):1567-1578.

[7] Brown A,Davis K.Time Series Analysis of Railway Passenger Demand[J].Transportation Research Record,2024,2678(3):234-245.

[8] 王海峰,张丽,刘洋.项目式学习在交通工程专业教学中的实践与反思[J].高等工程教育研究,2023,(4):88-94.

[9] Thompson R,Anderson P,Wilson S.Virtual Simulation in Railway Engineering Education:A Systematic Review[J].Engineering Education,2024,19(2):112-125.

[10] Anderson K,Davis R,Evans M.Project-Based Learning in Transportation Engineering:Best Practices[J].Journal of Professional Issues in Engineering Education and Practice,2024,150(1):05023006.

[11] 李华,王强,赵芳.数据可视化在铁路运输数据分析教学中的应用[J].计算机教育,2024,(3):102-108.

[12] Cooper J,Wright A,Scott P.Data Visualization for Enhanced Learning in Engineering Education[J].IEEE Transactions on Education,2024,67(2):123-134.