

基于 ARIMA 的网约公交出行需求预测研究

易坤炎^{1,2}, 李 祥^{1,2}, 熊 刚^{1,2}, 文健峰^{1,2}

(1. 长沙中车智驭新能源科技有限公司, 长沙 410083;

2. 中车时代电动汽车股份有限公司, 湖南 株洲 412007)

摘 要:本文针对网约公交出行需求预测问题,提出一种基于 ARIMA 模型的预测方法。通过整合乘客出行数据、天气数据和时间特征数据,构建了综合特征的时间序列预测模型。研究结果表明,考虑天气温度和时间特征的 ARIMA 模型较基础 ARIMA 模型具有更好的预测精度,能够为网约公交的运营调度提供科学依据。本研究为城市微循环交通系统的智能化提供了理论和实践参考。

关键词:ARIMA 模型; 网约公交; 出行需求预测

中图分类号:U491.1⁺2; O211.61 **文献标志码:**A **DOI:**10.15917/j.cnki.1006-3331.2025.04.006

Research on Travel Demand Forecasting of Ride-hailing Bus Based on ARIMA

YI Kunyan^{1,2}, LI Xiang^{1,2}, XIONG Gang^{1,2}, WEN Jianfeng^{1,2}

(1. Changsha CRRC Intelligent Control and New Energy Technology Co., Ltd., Changsha 410083, China;

2. CRRC Electric Vehicle Co., Ltd., Zhuzhou 412007, China)

Abstract: This paper aims at the problem of travel demand forecasting for ride-hailing buses and proposes a prediction method based on the ARIMA model. It constructs a time series prediction model with comprehensive features by integrating passenger travel data, weather data, and time feature data. The research results show that the ARIMA model, which considers weather temperature and time characteristics, has better prediction accuracy than the basic ARIMA traditional model and can provide a scientific basis for the operation and scheduling of ride-hailing buses. This study provides theoretical and practical references for the intelligence of urban micro circulation transportation systems.

Key words: ARIMA model; ride-hailing bus; travel demand forecast

随着城市交通需求的多样化和个性化发展,网约公交作为一种灵活的公共交通方式,在解决“最后一公里”问题中发挥着重要作用。株洲市荷塘区职教城沿北环路半径 3 km 范围内有 9 所学校,需保障共计约 10 万名师生的日常出行。因此,积极探索微循环交通出行方式,已成为亟待解决的现实问题。由于网约公交服务具有需求分散、时空特征复杂的特点,传统的运营模式难以满足精准调度需求。因此,开展网约公交出行需求预测研究,对于优化交通资源配置、提高运营效率具有重要意义。

网约公交高效运营的关键之一在于精准的出行需求预测。研究表明,该网约公交出行需求受到多重

因素影响,其中时间特性、外部环境因素的影响尤为显著^[1-2]。尽管已取得诸多研究成果,但也存在一些问题:传统自回归积分滑动平均模型(Autoregressive Integrated Moving Average, ARIMA)方法依赖研究者对自相关函数(Autocorrelation Function, ACF)和偏自相关函数(Partial Autocorrelation Function, PACF)图的经验解读^[3],或者仅采用单一信息准则衡量统计模型。前者主观性强且易误判,后者在大样本下可能导致过参数化,均影响模型的稳定性和泛化能力。因此,本文在 ARIMA 模型应用中实现了两方面创新:首先,在模型定阶过程中引入贝叶斯信息准则(Bayesian Information Criterion, BIC)与 ACF-PACF 联合分析的

收稿日期:2025-05-10。

第一作者:易坤炎(1983—),男,助理工程师,主要从事智能网联方面的管理工作。E-mail:yikunyan.cy@crrecg.cc。

混合优化机制,通过双重验证确保参数选择的鲁棒性;然后,构建温度-时间交叉特征,采用滑动窗口法计算连续3日温度变化率与周末状态的交互效应,有效捕捉特殊天气条件下的需求突变规律。

1 原理与方法

1.1 ARIMA 模型原理

ARIMA 模型是一种广泛应用于时间序列预测的统计模型^[4]。该模型通过3个参数(p, d, q)来描述时间序列的趋势和季节性特征,其中 p 为自回归项的阶数, d 为差分次数, q 为滑动平均项的阶数。ARIMA 模型的基本思想是将非平稳时间序列转化为平稳时间序列,然后通过建立合适的自回归和滑动平均模型进行拟合和预测。在应用 ARIMA 模型时,需先验证序列的平稳性,可通过检验统计量单位根检验。当检验统计量大于临界值(如5%显著性水平对应-2.86)时,判定序列非平稳,此时需要进行差分处理。差分是通过构造新序列消除趋势成分的核心方法,一阶差分即计算相邻观测值的差值,其公式为: $\nabla y_t = y_t - y_{t-1}$ 。若一阶差分仍不平稳,则需进行更高阶($d > 1$)的差分,二阶差分公式为: $\nabla^2 y_t = \nabla y_t - \nabla y_{t-1}$ 。具体而言,ARIMA 模型的数学表达式见式(1):

$$\varphi(B)(1-B)^d y_t = \theta(B) \varepsilon_t \quad (1)$$

式中: y_t 为时间序列数据; B 为滞后算子; $\varphi(B) = 1 - \varphi_1 B^1 - \varphi_2 B^2 - \dots - \varphi_n B^n$, 为自回归多项式; $\theta(B) = 1 + \theta_1 B^1 + \theta_2 B^2 + \dots + \theta_n B^n$, 为滑动平均多项式; ε_t 为白噪声序列。

1.2 评价指标

对于文中株洲网约公交出行需求预测的实例,为了客观判断模型性能,本文选择平均绝对误差(Mean Absolute Error, MAE)和均方根误差(Root Mean Square Error, RMSE)作为评价指标。这两个指标能够直观反映模型预测的准确性^[5],且适用于连续数值型数据的评估,具体计算公式见式(2)和式(3):

$$E_{MA} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

$$E_{RMS} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

式中: n 表示样本数量; $|y_i - \hat{y}_i|$ 表示绝对误差取绝对

值; $y_i - \hat{y}_i$ 表示第 i 个样本的真实值 y_i 与预测值 $\hat{y}_i$ 的绝对误差; E_{MA} 为预测值与实际值之间绝对差异的平均值,用于衡量预测误差的平均大小,不考虑误差的方向, E_{MA} 值越小,表明模型的预测准确性越高; E_{RMS} 为预测值与实际值之间差异平方的平均值的平方根, E_{RMS} 对较大的误差更为敏感,由于误差被平方后会进一步放大, E_{RMS} 的取值范围在 $[0, +\infty)$ 之间,数值越小,表示模型的预测误差越小,预测能力越强。

在实际应用中,MAE 和 RMSE 有各自的适用场景与优缺点。MAE 的优点是计算简单、直观,并且对异常值的鲁棒性较好,不易受到个别大误差的影响;而 RMSE 对大误差更敏感,能更有效地反映模型在极端情况下的表现,更适用于误差呈正态分布的数据。因此,在模型评估中,通常建议同时参考 MAE 和 RMSE,以全面评估模型的预测性能。

2 模型构建

2.1 数据准备

文中所用数据为株洲网约公交的数据集,涵盖了2024年10月15日至2025年3月31日期间的所有订单数据,主要字段包括:ID、乘客乘车起点ID、乘客乘车起点名称、乘客乘车终点ID、乘客乘车终点名称、订单时间。同时对订单数据按日期进行了统计,统计结果如图1所示。

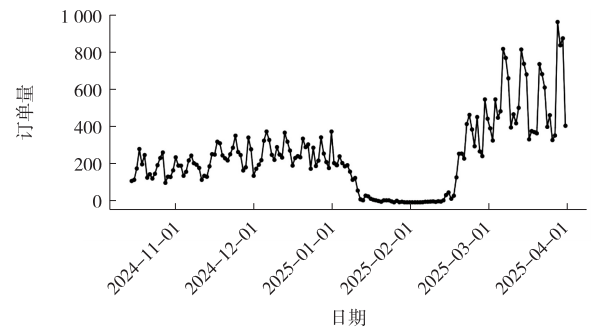


图1 订单统计图

从图1可知,订单量总体呈上升趋势。在非寒暑假期间(2024年10月15日至2025年1月7日,2025年2月17日至2025年3月31日)订单呈现明显的周期规律。而在寒暑假期间,订单量则相对较少。

基于订单时间提取日期特征,根据日期生成订单日期温度表与订单日期特征表。其中,订单日期特征

表包含以下字段:日期、星期几、是否周末、是否节假日;订单日期温度表包含以下字段:日期、日平均温度。

使用 Python 中的 pandas 库对数据进行清洗和处理:①将“订单时间”字段中的日期提取为标准格式(如 YYYY/MM/DD);②确保所有数值字段(如日平均温度)为数值类型。最终形成了包含以下字段的综合数据集:日期、订单出行需求量、日平均温度、星期几、是否周末、是否节假日。

2.2 模型构建

基于整合后的数据,构建了考虑天气和时间特征的 ARIMA 模型。具体步骤如下:

1) 平稳性检验。通过增广迪基-富勒检验(Augmented Dickey-Fuller Test, ADF)检验确定原始序列的平稳性^[6]。将数据输入 Python 程序,ADF 检验结果为 $p\text{-value}=0.8098$ 。由于该值大于 0.05,需要对数据进行差分以消除趋势。首先对原始序列进行一阶差分,一阶差分后的 ADF 检验结果为 $p\text{-value}=0.0914$ 。由于该值仍大于 0.05,需要进行二次差分。进行二次差分后的 ADF 检验结果为 $p\text{-value}\approx 7.5524\times 10^{-9}$,此值远小于 0.05,因此可以认为数据已经平稳。

2) 模型定阶。在完成数据清洗与平稳性检验后(已通过 ADF 检验确定差分次数 $d=2$),本文在 BIC 准则指导下,采用 ACF-PACF 联合分析方法确定 ARIMA 模型的最优参数组合 (p,d,q) ^[7-10]。首先构建参数搜索空间,其中 $p\in[0,5]$, $q\in[0,5]$,通过网格搜索计算不同组合的 BIC 值。当 $p=2, q=1$ 时, BIC 值最小(-423.56),且 ACF 和 PACF 的截尾特征符合统计规律,最终确定 ARIMA(2,2,1) 为最优模型。特征工程方面,除基础时间特征外,本研究创新性地构建了温度累积效应因子和节假日衰减系数。

①温度累积效应因子 W_t 的计算公式见式(4),该参数旨在捕捉近期温度对目标变量的影响。

$$W_t = 0.6T_t + 0.3T_{t-1} + 0.1T_{t-2} \quad (4)$$

式中: T_t 表示当前时刻的温度; T_{t-1} 和 T_{t-2} 分别表示前一时刻和前两时刻的温度。

赋予近期温度更高的权重(0.6、0.3、0.1)是因为近期温度对目标变量的影响更为显著。这种权重分配方式能够更好地反映温度变化的即时效应。权重选择依据是基于历史数据的分析和试验验证。通过对比不同权重组合下的模型预测效果,最终确定了上

述权重分配方案,以达到最优的预测性能。

②节假日衰减系数 H_t 的计算公式见式(5),该参数旨在量化假期影响的时序衰减特征。

$$H_t = 1 + 0.3e^{0.5d} \quad (5)$$

式中: d 为距节假日的天数。

公式中的 $e^{0.5d}$ 表示随着距节假日天数 d 的增加,衰减系数逐渐减小。这种指数衰减形式能较好地模拟假期影响随时间逐渐减弱的规律。系数 0.3 和指数项的组合确保了在节假日前后,衰减系数能够显著偏离 1,从而有效反映假期对目标变量的影响。随着距节假日天数的增加,衰减系数逐渐趋近于 1,表示假期影响逐渐消失。

3) 模型训练。使用选定的参数构建 ARIMA 模型并进行训练,其中 2024 年 10 月 15 日至 2025 年 3 月 21 日的数据作为训练集,2025 年 3 月 22 日至 2025 年 3 月 31 日的数据作为测试集。以训练集的前 30 天数据为初始窗口,逐步向后滚动扩展。每新增一天的数据,就重新拟合模型并预测次日的需求量。

3 基于模型的出行需求预测结果

3.1 预测结果

预测的订单量见表 1。

表 1 预测订单量与实际订单量对比

序号	日期	实际订单量	预测订单量	相对误差/%
1	2025/3/22	688	663	4
2	2025/3/23	617	570	8
3	2025/3/24	405	421	4
4	2025/3/25	468	511	9
5	2025/3/26	334	329	1
6	2025/3/27	359	375	4
7	2025/3/28	969	1017	5
8	2025/3/29	843	805	5
9	2025/3/30	881	864	2
10	2025/3/31	411	443	8

由表 1 可知,预测订单量与实际订单量的误差稳定在 10% 以内。预测误差的出现可能与模型对某些特殊天气条件或节假日的预测不够准确有关,这或许是因为模型未能充分考虑这些特殊因素的影响。

3.2 可视化分析

预测订单量与实际订单量的整体趋势如图 2 所示。

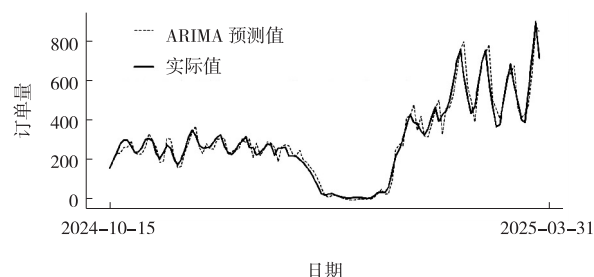


图2 ARIMA 模型预测效果对比

从图2可以看出,预测值与实际值的走势基本一致,表明 ARIMA 模型能较好地捕捉数据的变化趋势。

图3展示了 ARIMA 模型预测误差的分布情况。从图中可以看出,误差主要集中在某个范围内,表明模型的预测结果相对稳定。

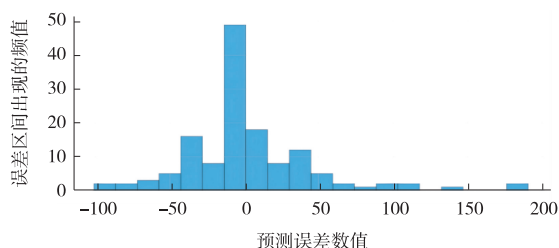


图3 误差分布图

3.3 模型评价指标

为验证模型创新点的有效性,设置另外两组模型与本文模型进行对比:第一组仅使用历史订单数据进行传统 ARIMA 预测;第二组在相同特征输入的基础上进行长短期记忆网络(Long Short-Term Memory, LSTM)模型预测,具体结果见表2。

表2 模型对比

模型	MAE	RMSE
传统 ARIMA	143	206
LSTM	80	97
本文模型	28	42

本文 ARIMA 模型的 MAE 较传统 ARIMA 降低了 67%,验证了多源特征融合的有效性;与 LSTM 相比,本文模型在训练数据量有限(<1 000 样本)时仍保持至少 10%的精度优势,说明 ARIMA 在小样本场景的适用性。传统 ARIMA 模型的参数选择(如差分次数、自回归阶数和滑动平均阶数)需要手动调整,对用户的技术要求较高。相比之下,LSTM 模型虽然在处理非线性关系和长期依赖方面具有优势,但其模型结

构复杂,训练过程耗时较长,且对数据量和计算资源要求较高。改进后的 ARIMA 模型在保持较高精度的同时,简化了模型结构,降低了计算复杂度。

4 结束语

本文通过基于温度和时间特征的 ARIMA 模型对网约公交出行订单进行时间序列预测,研究结果表明:2025 年 3 月 22 日至 3 月 31 日内出行订单量预测相对误差稳定控制在 10% 以内(平均绝对误差 MAE=28),证明该模型在网约公交出行订单预测场景中具有方法论层面的普适性。同时,在理论意义方面,本文通过引入差分次数自适应调整机制,完善了传统 ARIMA 模型在复杂交通场景中的应用范式;在实践层面,为株洲网约公交建立了出行订单预测系统,从而实现动态预警,提高公交运营效率。后续可在以下方面继续开展研究:①设计基于实时路况数据的 ARIMA 动态参数调整预测系统;②融合 LSTM 深度学习算法构建混合预测模型。

参考文献:

- [1] 田镡豪,王先玉,秦锦波,等. 基于 ARIMA 模型预测的中国茶叶出口额分析[J]. 农业技术与装备,2024(12):77-81.
- [2] 段依幸. 基于 ARIMA 的城市度电产值变化趋势研究[J]. 智能城市,2024,10(11):57-59.
- [3] 徐志昊. 多源因素影响下的出租车出行需求预测研究[D]. 青岛:青岛大学,2023.
- [4] 周坤,许云飞,祁浩伟. 基于改进 ARIMA 的新能源发电短期动态调度模型[J]. 电脑与信息技术,2024,32(1):56-61.
- [5] 毕辉. 基于订单数据挖掘的共享单车出行需求特性研究[D]. 南京:东南大学,2023.
- [6] 王春霞. 基于用户出行需求预测的共享汽车车辆调度优化研究[D]. 北京:北京交通大学,2022.
- [7] 余彦翘,李思杰,刘志钢. 基于 ARIMA 与 LSTM 的铁路车站客流预测方法比较[J]. 上海工程技术大学学报,2024,38(3):278-283.
- [8] 俞乐澜,邵梓轩,徐程,等. 基于季节性 ARIMA 模型的短时交通流预测方法研究[J]. 交通世界,2024(25):2-5.
- [9] 谭林清. 基于 AFC 数据的城市轨道交通车站短时客流量预测研究[D]. 重庆:重庆交通大学,2024.
- [10] 董敏,娄峰. 基于 SARIMA 和 ARIMA-GARCH 模型的共享单车用户量预测[J]. 现代商业,2023(17):46-49.