NSIEx +RS 2025
NEGERI SEMBILAN INTERNATIONAL EXPOSITION

HARMONY IN PROGRESS:

INTEGRATING SCIENCE, TECHNOLOGY, AND SOCIAL INNOVATION FOR SUSTAINABLE DEVELOPMENT GOALS

e-Book of Extended Abstract
Negeri Sembilan International (NSIEx) & Research Symposium
2025

RSST002 - FORECASTING DAILY LRT AMPANG LINE RIDERSHIP USING SARIMA MODEL

Suhaila Bahrom[1*], and Eszleen Sies[2]

[1]Department of Mathematics, Centre for Foundation Studies, International Islamic University Malaysia

[2]College of Computing & Informatics, Universiti Tenaga Nasional, Putrajaya Campus

Corresponding author: suhaila_b@iium.edu.my

ABSTRACT

Accurate forecasting of public transport ridership is crucial for effective urban transit planning and operations. In particular, short-term prediction of LRT ridership in Kuala Lumpur plays a pivotal role in optimizing scheduling and resource allocation. The main objective of this study is to develop a reliable statistical model for forecasting the daily ridership of the LRT Ampang line. This study employs the Seasonal Autoregressive Integrated Moving Average (SARIMA) model based on Box-Jenkins methodology to analyze daily ridership data obtained from the Malaysia Official Open Data Portal. The dataset covers the period from January 2023 to May 2025. Model identification and estimation involved stationarity checks, ACF and PACF analysis, and parameter selection using AIC and BIC. The final model selected, SARIMA$(1,1,2)(0,1,1)_7$, demonstrated strong forecasting performance with a MAPE of 9.78%. Residual analysis confirmed model adequacy through the ACF plot of residuals and the Ljung-Box test. This study concludes that SARIMA is a robust and interpretable statistical method for forecasting daily LRT ridership. It offers a structured statistical approach that contributes to improved decision-making in public transportation planning through short-term ridership forecasting.

Keywords – LRT Ridership, Box-Jenkins, SARIMA, time series analysis

INTRODUCTION

Forecasting is vital in modern urban transit planning, especially in fast-developing cities where transport demand is highly dynamic. The Light Rail Transit (LRT) system in Kuala Lumpur is among the most significant modes of public transportation, serving thousands of commuters daily. Accurate ridership forecasts are critical to ensure sufficient train frequency, avoid overcrowding, and allocate operational resources effectively. Despite the growing availability of transport data, traditional methods often fall short in capturing seasonality and autocorrelation in time series data, underscoring the need for models specifically designed to address these features.

Recent international evidence demonstrates that the Seasonal Autoregressive Integrated Moving Average (SARIMA) model remains a robust and interpretable tool for short-term ridership forecasting, particularly in contexts with clear seasonal fluctuations. For example, a study on metro passenger demand in Beijing showed that SARIMA effectively captured recurring patterns with high accuracy, making it suitable for operational planning in rail networks (Li et al., 2021). Similarly, research in China and India confirmed that SARIMA and its hybrid variants provided reliable forecasts of subway and bus passenger flows, reinforcing its relevance as a baseline model for transport applications (Patel et al., 2025; Zhang, 2024). More recent work also demonstrated that SARIMA remains valuable when integrated into ensemble frameworks for metro ridership forecasting, further validating its effectiveness in diverse settings (Sun et al., 2020).

In Malaysia, forecasting studies on public transport ridership remain comparatively underexplored, particularly at the daily level for rail services. Earlier work on the Ampang Line applied ARIMA models with monthly data, which provided baseline insights but did not capture short-term fluctuations in passenger demand (Shitan et al., 2014). More recently, Al Nafees et al. (2024) examined Rapid Bus KL ridership using machine learning techniques and reported that Random Forest substantially outperformed ARIMA, achieving R^2 of 0.9656 and MAPE of 3.4%. While this demonstrates the growing interest in advanced approaches, daily-level ridership forecasting for Malaysia's rail systems remains scarce, and empirical evidence on the applicability of SARIMA to this context is still relatively limited. The main objective of this study is to apply the SARIMA model to develop a time series forecasting model for predicting the daily ridership of the LRT Ampang line, contributing new evidence of SARIMA's capability to generate accurate and interpretable forecasts that can support short-term operational planning and provide a methodological baseline for future comparative studies in transport demand forecasting.

METHODOLOGY

SARIMA is a time series forecasting model that accounts for seasonal and non-seasonal variations. This study used open-access secondary data on LRT Ampang ridership from the Malaysia Open Data Portal, which contained no personal or sensitive information; therefore, ethical approval was not required. The dataset covered the period from January 2023 to May 2025 and contained daily ridership records, providing sufficient observations for forecasting seasonal and short-term patterns. The dataset was divided into training and testing subsets using a 90:10 split. The general form of the SARIMA $(p,d,q)(P,D,Q)_s$ can be expressed as Equation (1):

$$\phi_P\left(B^S\right)\varphi_p\left(B\right)\nabla_S^D\nabla^d y_t = \Theta_Q\left(B^S\right)\theta_q\left(B^q\right)a_t \tag{1}$$

where $p,d,$ and q represent the non-seasonal autoregressive, differencing, and moving average orders respectively, and P,D,Q represent their seasonal counterparts with seasonality period s. To stabilize variance, the Box-Cox transformation was applied, and the Augmented Dickey-Fuller (ADF) test was conducted to confirm stationarity in the mean (Osborne, 2010) . The autocorrelation (ACF) and partial autocorrelation (PACF) plots guided the selection of model parameters. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) was used to identify the optimal SARIMA configuration. Residual analysis and the Ljung-Box test Ljung & Box (1978) validated the model fit, with all procedures following the Box-Jenkins methodology (Box et al., 2015).

RESULTS

The time series plot in Figure 1 illustrates daily fluctuations in LRT Ampang line ridership from January 2023 to May 2025, revealing distinct seasonality and short-term variability in the data. The training dataset included data from 1st January 2023 to 3rd March 2025, while the testing dataset covered the remaining observations through the end of May 2025. Following the four stages of the Box-Jenkins methodology, model identification, parameter estimation, diagnostic checking, and forecasting, the best model selected was SARIMA(1,1,2)(0,1,1)$_7$. The model was chosen using the auto.arima function, which selected the model with the lowest AIC and BIC values. The model captured weekly seasonality and underlying ridership patterns. Figure 2 presents the diagnostic plots of residuals, where the ACF showed no significant autocorrelation. The Ljung-Box test results at lags 12, 24, and 36 produced *p*-values of 0.605, 0.658, and 0.936, respectively, greater than 0.05. This indicates that the null hypothesis of no autocorrelation could not be rejected, confirming that the residuals behave like white noise and supporting the adequacy of the selected model.

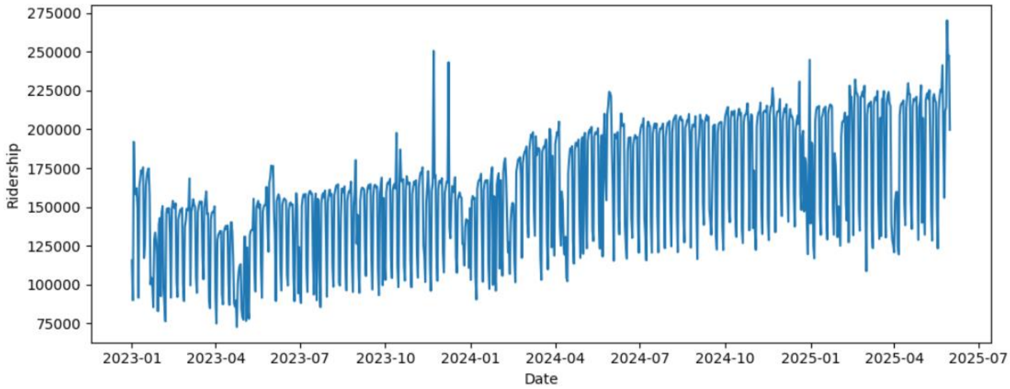

Figure 1: Daily LRT Ampang ridership trends from 1st January 2023 to 31st May 2025

Following the four stages of the Box-Jenkins methodology, model identification, parameter estimation, diagnostic checking, and forecasting, the best model selected was SARIMA(1,1,2)(0,1,1)$_7$, which captured weekly seasonality and underlying ridership patterns. Figure 2 presents the diagnostic plots of residuals. The residuals exhibited no significant autocorrelation and were approximately normally distributed, supporting the adequacy of the selected model.

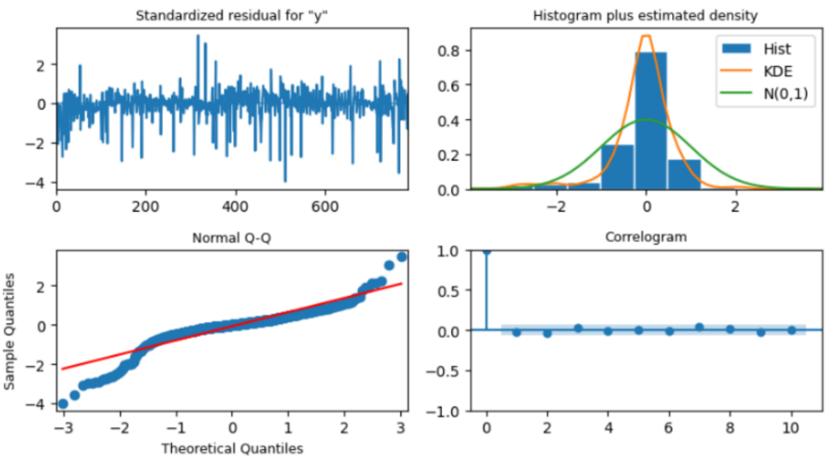

Figure 2: Diagnostic plots of the SARIMA(1,1,2)(0,1,1)$_7$ residuals

The Mean Absolute Percentage Error (MAPE) stood at 9.78%, which, according to Lewis (1982) indicates good predictive performance, as shown in Table 1. Figure 3 compares the actual ridership values with the SARIMA(1,1,2)(0,1,1)$_7$ forecasts for the testing period. The model followed the observed data trend, including the weekly fluctuations, with minimal forecast error.

Table 1: Interpretation of typical MAPE

MAPE(%)	10%	10%-20%	20%-50%	>50%
Evaluation	High accuracy prediction	Good prediction	Reasonable prediction	Inaccurate prediction

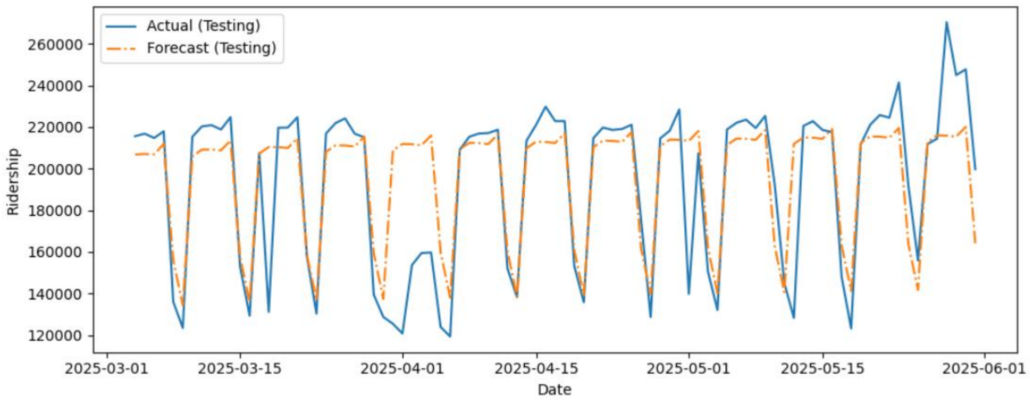

Figure 3: Actual vs forecasting plot of the SARIMA(1,1,2)(0,1,1)7

DISCUSSION AND CONCLUSION

Implementing the SARIMA model in this study demonstrates its effectiveness as a forecasting tool for public transportation systems. The model successfully captured the trend and seasonality in LRT Ampang's ridership data, making it a reliable planning and resource allocation approach. These forecasts can assist transport agencies in scheduling trains more efficiently during peak and off-peak hours, reducing overcrowding, and improving passenger flow management. Moreover, anticipating demand enables better resource allocation, such as optimizing staff deployment, maintenance scheduling, and energy usage.

SARIMA offers transparency and interpretability, essential for policy-making and operational decisions. The methodology used in this study can be extended to other urban transit systems with similar seasonal and trend-driven dynamics. Future work may consider hybrid models or incorporate external variables such as weather conditions, public holidays, and special events to improve forecasting accuracy further.

REFERENCES

Al Nafees, A., Hassan, M., Paul, A., Shraban, S. S., & Deb Mahin, H. (2024). Public Transport Ridership Forecasting Using Machine Learning: A Case Study of Rapid Bus KL in Malaysia. *2024 27th International Conference on Computer and*

Information Technology (ICCIT), 3366–3371. https://doi.org/10.1109/ICCIT64611.2024.11021879

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.

Lewis, C. D. (1982). *Industrial and Business Forecasting Methods*. Butterworth Scientific.

Li, W., Sui, L., Zhou, M., & Dong, H. (2021). Short-term passenger flow forecast for urban rail transit based on multi-source data. *EURASIP Journal on Wireless Communications and Networking*, *2021*(1), 9. https://doi.org/10.1186/s13638-020-01881-4

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, *15*(1).

Patel, M., Patel, S. B., Swain, D., & Mallagundla, R. (2025). Enhancing Accuracy in Hourly Passenger Flow Forecasting for Urban Transit Using TBATS Boosting. *Modelling*, *6*(2), 32. https://doi.org/10.3390/modelling6020032

Shitan, M., Karmokar, P. K., & Ng Yung Lerd. (2014). Time series modeling and forecasting of ampang line passenger ridership in Malaysia. *Pakistan Journal of Statistics*, *30*(3).

Sun, S., Yang, D., Guo, J., & Wang, S. (2020). AdaEnsemble learning approach for metro passenger flow forecasting. *ArXiv Preprint ArXiv:2002.07575.*

Zhang, M. (2024). Enhancing periodic mobility planning: ARIMA-driven prognostications for subway passenger volume dynamics. *Applied and Computational Engineering*, *101*(1), 224–232. https://doi.org/10.54254/2755-2721/101/20241038