STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
8th article, Pages: 377 – 388
DOI: https://doi.org/10.32493/sm.v7i3.54879

FORECASTING FERRY PASSENGER TRAFFIC IN NEW YORK CITY USING THE SEASONAL ARIMA (SARIMA) MODEL

Linda Apriliana[1,*], Rana Aribah [2], and Gumgum Darmawan[3]

[1] Universitas Padjadjaran, Bandung
[2] Universitas Padjadjaran, Bandung
[3] Universitas Padjadjaran, Bandung

[*]Correspondance author: rana25003@mail.unpad.ac.id

ABSTRACT

This study addresses the seasonal and long-term fluctuating passenger volume patterns typical of water transportation systems such as NYC Ferry, necessitating practical forecasting methods to support operational decision-making and public transportation planning. The research aims to develop a forecasting model for NYC Ferry passenger counts using the Seasonal Autoregressive Integrated Moving Average (SARIMA) methodology. The analysis utilizes monthly historical passenger data from January 2020 to December 2024 for training data. Key analytical steps include testing data stationarity, splitting the dataset into training and testing subsets, modeling via RStudio, forecasting, and evaluating model accuracy using Mean Absolute Percentage Error (MAPE) compared against actual observations. Results indicate that the SARIMA (1,0,0) (0,1,1)12 model outperforms other methods, yielding the lowest MAPE of 5.04%, compared to Multiplicative Winters (8.57%), SARFIMA (17.62%), and Holt-Winters (32.93%). The SARIMA model effectively captures both seasonal and monthly trends, producing accurate passenger volume predictions. These findings demonstrate SARIMA's efficacy in monthly NYC Ferry ridership forecasting, contributing to time series literature, particularly within public transportation forecasting. Furthermore, the results offer practical insights for policymakers to strategize service capacity and enhance data-driven management of waterborne transit systems more efficiently.

Keywords: SARIMA, SARFIMA, Time Series Forecasting, Water Transportation, Time Series.

ARTICLE INFO

Submission received: 20 November 2025 | Accepted: 30 December 2025
Revised: 15 Desember 2025 | Published: 31 December 2025

Available on: https://doi.org/10.32493/sm.v7i3.54879

INTRODUCTION

Water transportation is one of the alternative modes of transport widely used in various urban areas, especially in coastal regions, densely populated areas, and major cities with river routes (25). Ships, boats, and ferries serve as important means to support community mobility in these regions. The increase in population and the pace of urbanization also heighten the demand for efficient water transportation (1). Therefore, data related to the use of this mode of transport becomes crucial to be statistically analyzed in order to support appropriate policymaking (12).

Information on passenger numbers plays an important role in assessing service effectiveness and supporting data-driven decision-making processes. Passenger statistics can illustrate patterns of community movement both on a daily and seasonal basis (23). Through careful analysis, service providers can adjust schedules or fleet capacity according to needs. These data also enhance transparency in the process of public transportation planning. Without adequate data, the resulting policies risk being inefficient and scientifically difficult to justify (4).

In addition, ferry passenger data serves to identify usage trends and detect potential problems early, such as periods of high congestion or declines in passenger numbers (24). The results of the analysis can form the basis for developing more adaptive and sustainable services. Understanding travel trends allows operators to adjust service strategies to be more responsive to user needs (30). Improvements in service innovation can also have a positive impact on passenger satisfaction and support the sustainability of water transportation systems (7).

In an international context, forecasting ferry passenger numbers in New York City (NYC) plays an important role for the city government in planning fleet capacity and port infrastructure efficiently (10). Through time series analysis and predictive models, passenger surges during certain periods, such as holiday seasons, can be predicted earlier (27). This step helps prevent fleet shortages or congestion at port facilities. Overall, accurate forecasting supports the smooth operation of water transportation services in major cities (20).

However, the water transportation sector faces various external uncertainties, such as seasonal influences, extreme weather conditions, regulatory changes, and economic dynamics (5). Fluctuations in passenger numbers triggered by these factors often create non-stationary data patterns, making time series modelling more challenging (17).

The ARIMA (Autoregressive Integrated Moving Average) method is widely used in time series analysis, but this approach has limitations in capturing complex seasonal patterns—such as spikes in passenger numbers during holiday seasons (21). To address this, models capable of accommodating seasonal components more flexibly are required, such as SARIMA (Seasonal ARIMA) and SARFIMA (Seasonal Autoregressive Fractionally Integrated Moving Average). These two methods are considered more capable of producing accurate forecasts for data with seasonal characteristics (26).

MATERIAL AND METHODS

Research design

There are several types of models for forecasting data with seasonal patterns, including SARIMA, SARFIMA, Holt–Winter, and Multiplicative Winter (29). The design of this study aims to forecast the number of ferry passengers in New York City using the SARIMA model. This method was selected to evaluate the capability of the models in handling monthly time series data that fluctuate and contain seasonal components. The SARIMA model is used to model data with seasonal patterns. In addition, the model is compared with the SARFIMA model, which is able to capture long-memory dependence in the data, as well as the Multiplicative Winter and Holt–Winter models.

Population and sample

The population of this study consists of all historical monthly data on the number of ferry passengers in New York City available through NYC Ferry Ridership. For analytical purposes, data from January 2020 to December 2024 are used as training data, while data from January to July 2025 are used as testing data to evaluate the model's performance in forecasting the number of passengers.

Instrument/Procedure

The main instrument used in this study is the SARIMA model, which is applied to model the passenger count data. The research procedure includes collecting secondary data from official sources, checking data completeness, examining data patterns, checking stationarity, normalizing the data, and dividing the data into training and testing sets. Furthermore, the SARIMA model is built using RStudio, followed by model training and forecasting. The accuracy of each model is evaluated using MAPE.

Data Analysis

Data analysis was carried out by comparing the forecasting accuracy of the SARIMA model. The MAPE value was used as the primary metric to assess how well the model predicts the number of passengers. In addition, comparisons were also made with the SARFIMA, Multiplicative Winters, and Holt–Winters models as references.

SARIMA (Seasonal Autoregressive Integrated Moving Average)

The SARIMA model is an extension of the ARIMA model designed to handle time series data with seasonal patterns (14). This model consists of two main components, namely the non-seasonal and seasonal components, which together describe the dynamics of the data over time (6). Data that exhibit seasonal patterns can be modeled using the Seasonal ARIMA (SARIMA), expressed as SARIMA (p, d, q) (P, D, Q)s. Here, p represents the autoregressive order, d is the differencing applied to achieve stationarity, and q denotes the moving average order (16). Meanwhile, P, D, and Q respectively represent the seasonal autoregressive order, seasonal differencing, and seasonal moving average, while s indicates the number of periods in one seasonal cycle. This model extends ARIMA by adding seasonal components (11). In general, the SARIMA model can be expressed in the following notation:

$$ARIMA(p,d,q)(P,D,Q)^s \tag{1}$$

The ARIMA model equation can be represented as a SARIMA model as follows:

$$\phi_{\mathrm{p}}(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D\dot{Z}_t = \theta_q(B)\Theta_Q(B^s)\alpha_t \tag{2}$$

Where α_t is a Gaussian white noise process for an s-period time series. The autoregressive and moving average components are represented by $\phi_p(B)$ and $\theta_q(B)$ of orders p and q. The seasonal autoregressive and moving average components are represented by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$, where P and Q are the seasonal orders. $(1-B)^d$ represents the non-seasonal differencing component, while $(1-B^s)^D$ represents the differencing for the seasonal component. B is the backshift operator (11).

SARFIMA (Seasonal Autoregressive Fractionally Integrated Moving Average)

The SARFIMA (Seasonal Autoregressive Fractionally Integrated Moving Average) model is an extension of the ARFIMA model, introduced by Porter–Hudak (1990). This model is designed to capture seasonal patterns and long-memory behavior in time series data, which cannot be fully explained by a standard SARIMA model (22). The SARFIMA model is denoted as follows:

$$SARFIMA(p,d,q)(P,D,Q)_s \tag{3}$$

with the general equation:

$$\Theta(L^s)\theta(\mathrm{L})(1-\mathrm{L})^{\mathrm{d}}(1-L^s)^D X_t = \Phi(L^s)\phi(L)\varepsilon_t \tag{4}$$

where L is the lag operator; s is the length of the seasonal period; d is the order of non-seasonal fractional differencing; D is the order of seasonal fractional differencing; $\Theta(L^s)\theta(\mathrm{L})$ is the non-seasonal and seasonal MA polynomials; $\Phi(L^s)\phi(L)$ is the non-seasonal and seasonal AR polynomials; ε_t is the random error (white noise)

The stationarity conditions of the model are:

1. Stationary if $d + D < 0.5$,
2. Long memory if $0 < d + D < 0.5$,
3. Non-stationary if $0.5 \leq d + D < 1$.

With parameters d and D that may take fractional values, SARFIMA is able to represent time series with long-term dependence and periodic seasonal fluctuations simultaneously (2).

Accuracy Measurement

In this study, forecasting accuracy is used as the main criterion for selecting the most appropriate forecasting method. The evaluation of accuracy is conducted using the Mean Absolute Percentage Error (MAPE), which is formulated as follows:

$$\mathrm{MAPE} = \left(\frac{1}{n}\right) \times \sum_{t=1}^{n} \left|\frac{Z_t - \hat{Z}_t}{Z_t}\right| \times 100 \tag{5}$$

Where Z_t is the actual value, $\hat{Z}_t$ is the forecast value at time t, and n is the number of observation periods (15). This measure is used to evaluate and compare the performance of forecasting models during the testing period (8).

RESULTS AND DISCUSSION

Statistical Eksperiments

This study focuses on the monthly volume of NYC Ferry passengers spanning from January 2020 to July 2025. The primary objective is to determine the most appropriate forecasting model for predicting passenger demand. Statistical analyses were performed using RStudio. The dataset was examined by aggregating passenger counts on a monthly basis to capture temporal patterns and variations in ferry usage.

Results

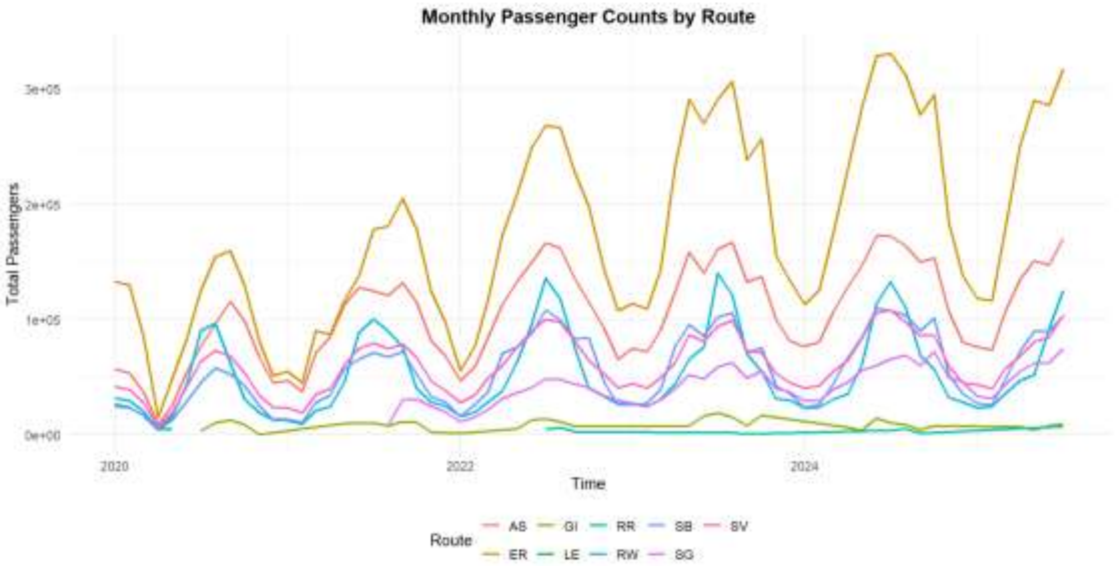

Figure 1. Monthly Passenger Counts by Route

The graph in Figure 1 illustrates the monthly passenger trends for each route during the 2020–2025 period. Each route represents the location where passengers board the ferry. Overall, all routes exhibit a clear seasonal pattern. The ER route shows the highest passenger volume throughout the observation period, indicating that it is the route with the highest demand. In contrast, the GI and RR routes record the lowest and least variable passenger counts. The similar patterns observed across routes suggest the presence of shared seasonal factors affecting all routes.

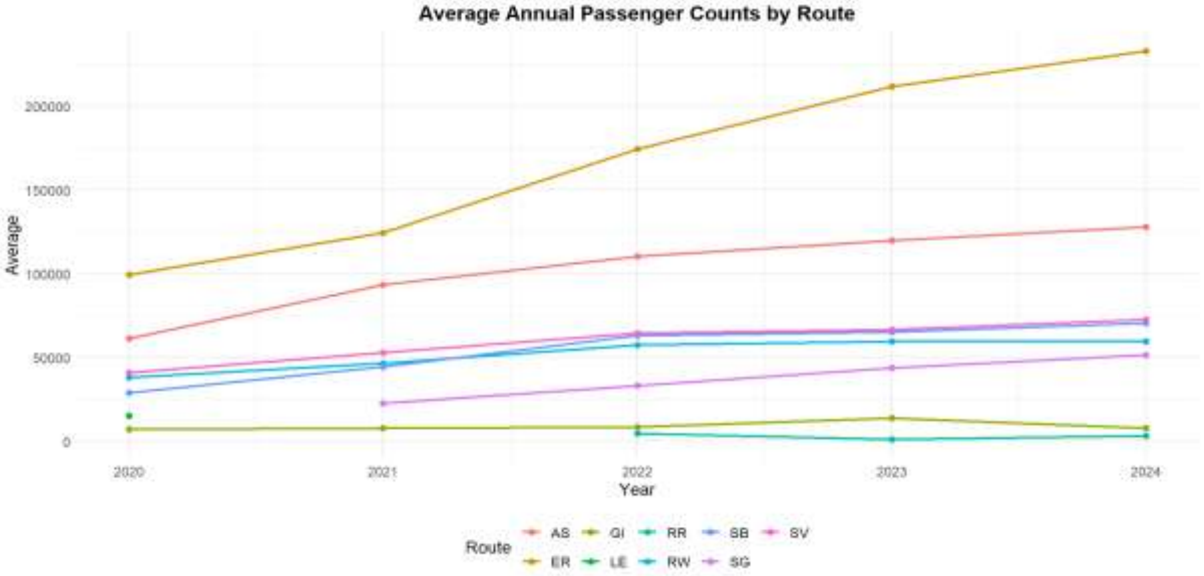

Figure 2. Average Annual Passenger Counts by Route

The graph in Figure 2 presents the average annual number of passengers by route for the period 2020 to 2024. Overall, all routes show an increasing trend in average passenger numbers from year to year. The ER route consistently records the highest annual average, indicating that it has the greatest level of passenger activity.

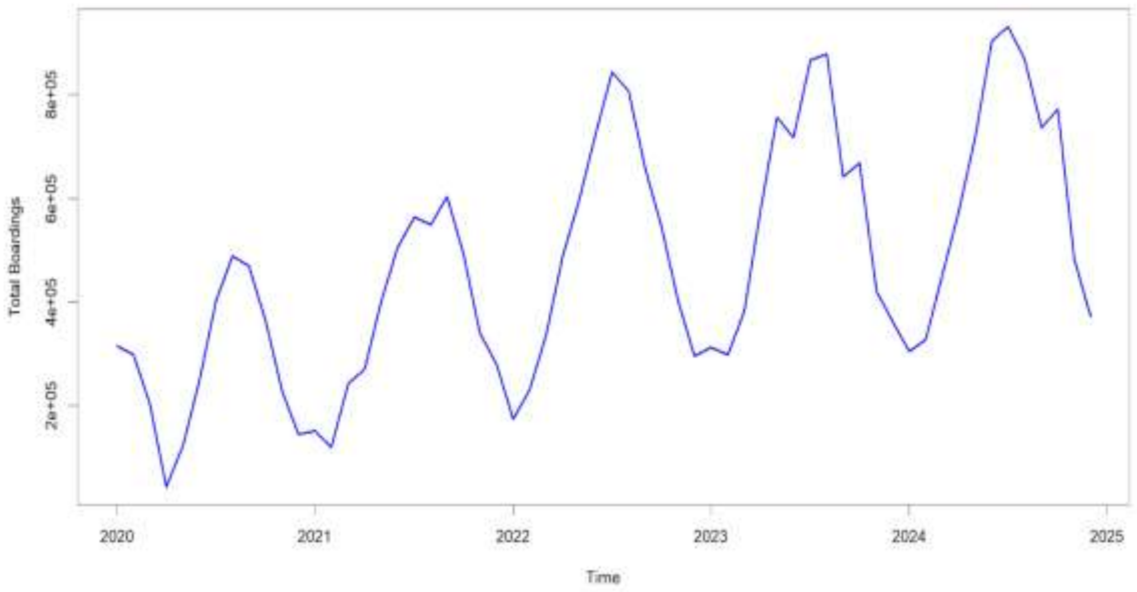

Figure 3. Monthly Passenger Counts by Route

Figure 3 presents the initial plot of NYC Ferry passenger counts from 2020 to 2025. The data show a clear upward movement over time, indicated by the progressively increasing trend line each year. The plot also reveals the presence of both trend and seasonal patterns.

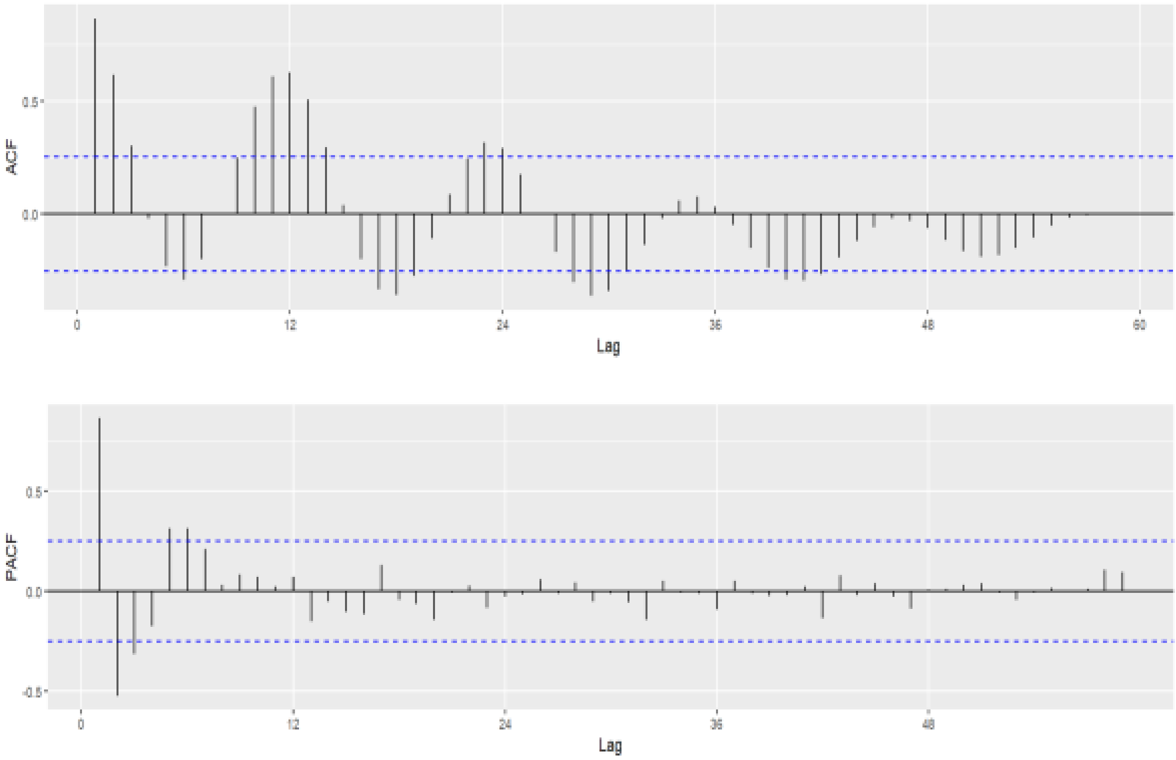

Figure 4. Autocorrelation (ACF) and Partial Autocorrelation (PACF) of the Original Series

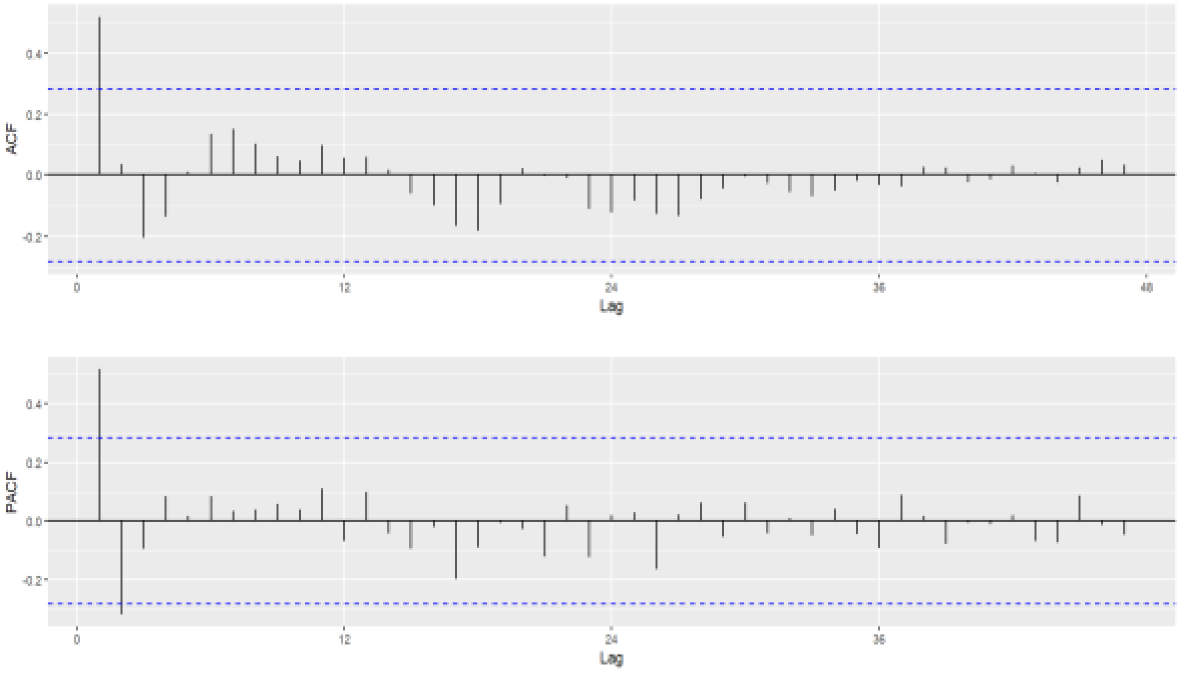

Figure 5. Autocorrelation (ACF) and Partial Autocorrelation (PACF) after Differencing (D = 1)

The Autocorrelation (ACF) and Partial Autocorrelation (PACF) of the original series are shown in Figure 4. Both the initial data plot in Figure 3 and the ACF pattern in Figure 4 indicate a clear seasonal pattern recurring every 12 months (s = 12) (9). In Figure 4, the ACF of the original series exhibits a slow decay, suggesting that the data are non-stationary and therefore require seasonal differencing. After applying one seasonal differencing, the series becomes stationary, as shown in Figure 5, where the ACF decays rapidly toward zero (11). These results support the consideration of a SARIMA $(p,d,q)(P,1,Q)_{12}$ model for further analysis. From the ACF plot after seasonal differencing (Figure 5), the first lag is significant, while the PACF shows a gradual decline with significant spikes at the first and second lags.

Based on Figure 5, the model components can be identified as follows: p (AR) = 2, based on the PACF; q (MA) = 1, based on the ACF.

The Augmented Dickey–Fuller (ADF) test indicates that the series is stationary ($p < 0.05$), so non-seasonal differencing is not required. For the seasonal components, P = 1 (from the seasonal PACF), Q = 1 (from the seasonal ACF), and D = 1 because one seasonal differencing was applied.

Based on the patterns observed in the ACF and PACF plots, several candidate models were considered to identify the best-fitting specification. The evaluated SARIMA models include SARIMA(0,0,1)(0,1,1)$_{12}$, SARIMA(0,0,1)(1,1,0)$_{12}$, SARIMA(1,0,0)(0,1,1)$_{12}$, SARIMA(1,0,0)(1,1,0)$_{12}$, SARIMA(1,0,1)(0,1,1)$_{12}$, SARIMA(1,0,1)(1,1,0)$_{12}$, SARIMA(1,0,1)(1,1,1)$_{12}$, SARIMA(2,0,1)(0,1,1)$_{12}$, SARIMA(2,0,1)(1,1,0)$_{12}$, SARIMA(2,0,0) (0,1,1)$_{12}$, SARIMA(2,0,0) (1,1,0)$_{12}$.

Table 1. Mean Absolute Percentage Error (MAPE).

Models	MAPE (%)
SARIMA (1,0,0) $(0,1,1)_{12}$	5.04
Multiplicative Winters	8.57
SARIMA (1,1,1) $(1,0,0)_{12}$	10.16
SARFIMA (1,0.438,1) $(1,0,0)_{12}$	17.62
Holt-winters Additive	32.93

Table 1 presents the MAPE values calculated using the testing period (January–July 2025). The best-performing model is SARIMA (1,0,0) $(0,1,1)_{12}$, which yields a MAPE of 5.04%.

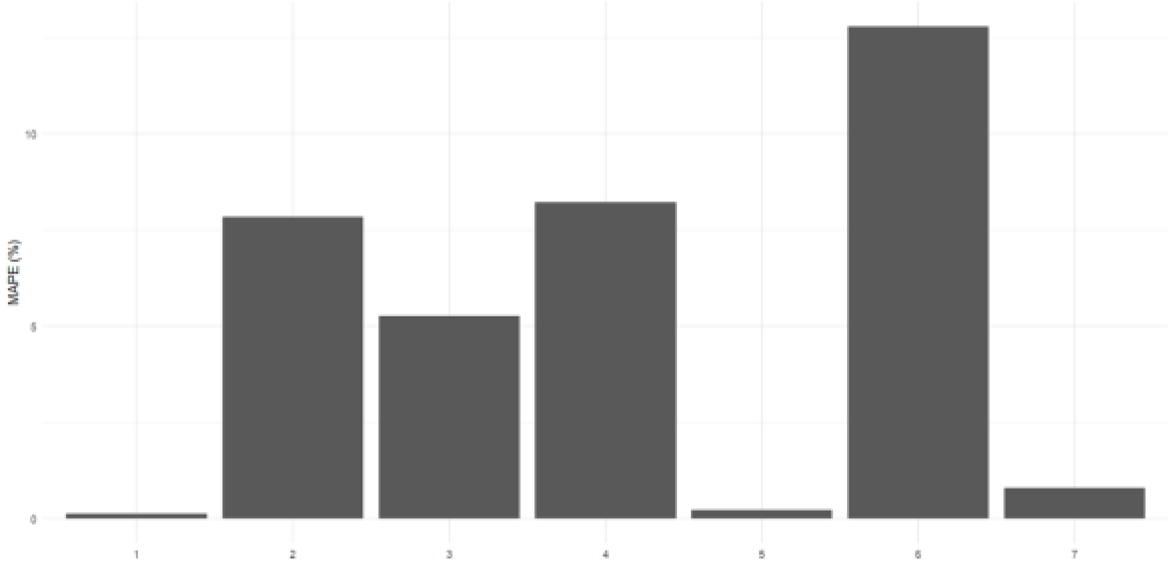

Figure 6. Monthly MAPE for the Forecast

Figure 6 presents the Mean Absolute Percentage Error (MAPE) for each month during the testing period. It can be observed that three months (1, 5, and 7) recorded MAPE values below 5%.

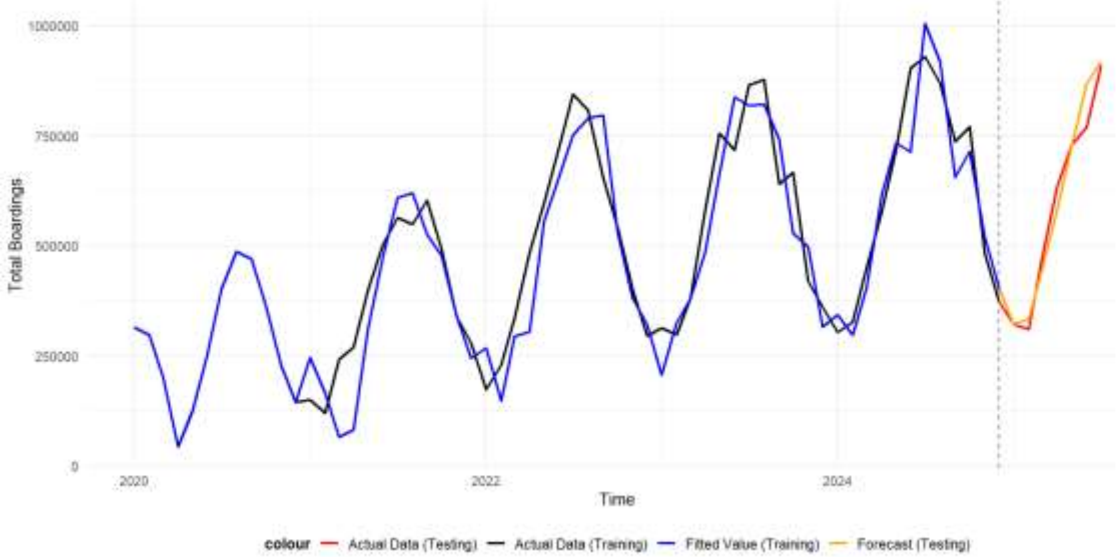

Figure 7. Actual vs. Predicted Values

Figure 7 presents the comparison between the actual data and the predicted values for both the training and testing periods. The graph shows that the model is able to follow the seasonal pattern and trend in the training data. The blue line (fitted values) and the black line (actual training data) demonstrate that the model successfully captures the seasonal fluctuations and overall trend in passenger counts. The orange line represents the forecasted values, which show an upward movement consistent with the historical trend. The actual values in the testing period (red line) also rise and remain close to the forecasted values, indicating that the model delivers reasonably accurate predictions. Overall, the model successfully represents both the seasonal pattern and the trend in the data across the training and testing periods.

CONCLUSION

The SARIMA model proved effective in forecasting passenger arrivals for the subsequent seven months, demonstrating strong capability in capturing both the seasonal pattern and historical trend. The model achieved a MAPE of 5.04%, which is considered highly accurate (18). Given that the evaluation was conducted only for a seven-month forecast horizon, it is recommended to regularly monitor and update the model to maintain predictive accuracy over longer periods.

AUTHOR CONTRIBUTIONS

For research articles with several authors, a short paragraph specifying their individual contributions must be provided.

Conceptualization, L.A.; methodology, L.A.; formal analysis, R.A.; investigation, L.A. and R.A.; resources, L.A.; data curation, R.A.; writing—original, L.A. and R.A.; writing—review and editing, L.A., R.A., and G.D.; visualization, R.A.; supervision, G.D. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support provided during the completion of this research. We would like to express our gratitude to the Master’s Program in Applied Statistics, Universitas Padjadjaran, for providing academic and technical support throughout the research process. We also extend our appreciation to the NYC Open Data for facilitating access to publicly available NYC Ferry Rideship data. Additionally, we are grateful for the constructive comments and insights shared by colleagues, which helped improve the clarity and quality of the manuscript.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. Abidin Z. Studi revitalisasi angkutan sungai sebagai moda transportasi perkotaan di Kota Banjarmasin. Jurnal Transportasi. 2017;17(2):123–35.
2. Adewole AI, Amurawaye FF. Modeling temperature forecast in Ogun State, Nigeria with SARFIMA and SARIMA. J Sci Inf Technol (JOSIT). 2024;18(1):1–12.
3. AM New York. NYC Ferry ridership reaches record levels as spring begins. 2025 Apr 15 [cited 2025 Oct 11]. Available from: https://www.amny.com/news/nyc-ferry-record-ridership-april-2025
4. Badan Pusat Statistik. Statistik Transportasi Laut Indonesia 2024. Jakarta: BPS; 2024.
5. Behdani B, Emmerich S. Resilience of maritime passenger transport under environmental and economic uncertainties. Mar Policy Manag. 2020;47(6):751–66.
6. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: Forecasting and control. 5th ed. Hoboken, NJ: John Wiley & Sons; 2015.
7. Chen Y, Zhang H, Liu W. Forecasting urban ferry ridership using time-series models: Evidence from New York City. Transp Res Procedia. 2023;71:12–21.
8. Carvalho-Silva M, Monteiro MTT, Sá-Soares F, Dória-Nóbrega S. Assessment of forecasting models for patients arrival at Emergency Department. Oper Res Health Care. 2018;18:112–8.
9. Chang X, Gao M, Wang Y, Hou X. Seasonal autoregressive integrated moving average model for precipitation time series. J Math Stat. 2012;8(4):500–5.
10. Ghosh D, Song W. Forecasting passenger demand for urban ferry services in metropolitan cities. J Transp Geogr. 2020;82:102580.
11. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 2nd ed. Melbourne, Australia: OTexts; 2018 [cited 2025 Oct 12]. Available from: OTexts.com/fpp2
12. Kementerian Perhubungan RI. Laporan Statistik Transportasi Laut dan Sungai Tahun 2022. Jakarta: Kemenhub; 2022.
13. Lindsey G, Wang F. Public ferry systems as sustainable urban mobility solutions: The case of NYC Ferry. Sustain Cities Soc. 2020;60:102217.
14. Majka M. Seasonal Time Series Analysis: Why SARIMA Outshines ARIMA. ResearchGate; 2024.
15. Makridakis S, Wheelwright SC, Hyndman RJ. Forecasting: Methods and

applications. 3rd ed. New York, NY: John Wiley & Sons; 1998.

16. Merabet F, Zeghdoudi H. On modelling seasonal ARIMA series: Comparison, application and forecast. WSEAS Trans Appl Theor Mech. 2020;15:1–10.
17. Moghimi B, Chen Z, Liu J. Non-stationary time series model for station-based ridership. J Big Data Anal Transp. 2022;4(3).
18. Montaño JJ, Palmer A, Sesé A, Cajal B. Using the R-MAPE index as a resistant measure of forecast accuracy. Psicothema. 2013;25(4):500–6.
19. NYC Open Data. NYC Ferry Ridershi. [cited 2025 Aug 29]. Available from: https://data.cityofnewyork.us/Transportation/NYC-Ferry-Ridership/t5n6-gx8c/about_data
20. OECD. Water transport and local economic development. Paris: OECD Publishing; 2021.
21. Pramesti DA, Nurhayati D. Peramalan jumlah penumpang kapal laut menggunakan model SARIMA pada Pelabuhan Tanjung Perak Surabaya. J Sains Seni ITS. 2021;10(2):A122–7.
22. Porter-Hudak S. Long-term memory modelling: A simplified spectral approach. Dissertation, University of Wisconsin–Madison; 1982.
23. Prawardani DSS. Analisis tren pergerakan penumpang di Pelabuhan Jayapura 2015–2024. J Teknik Transportasi, Universitas Muhammadiyah Palembang, Palembang; 2025.
24. Putri RD, Hidayat F. Analisis data penumpang kapal ferry untuk optimalisasi jadwal keberangkatan menggunakan metode time series. J Transportasi Logistik. 2020;7(2):95–104.
25. Rahma NA. Kajian transportasi sungai untuk menghidupkan kawasan tepian Sungai Kahayan Kota Palangkaraya. J Tataloka. 2014;16(1):1–12.
26. Saba T, Shahzad MN, Iqbal S, Rehman A, Abunadi I. A new hybrid SARFIMA-ANN model for tourism forecasting. Comput Mater Contin. 2021.
27. Sohifatul Khoiriyah N, Silfiani M, Novelinda R. Peramalan jumlah penumpang kapal di Pelabuhan Balikpapan dengan SARIMA. ResearchGate; 2023.
28. Susilo D, Santosa B. Analisis Penggunaan Data Penumpang untuk Optimalisasi Jadwal Kapal Penyeberangan di Indonesia. Jurnal Transportasi. 2021;21(2):88–97.
29. Xian X, Wang L, Wu X, Tang X, Zhai X, Yu R, et al. Comparison of SARIMA model, Holt-Winters model and ETS model in predicting the incidence of foodborne disease. BMC Infect Dis. 2023;23:803.
30. Zhou X, Li T. Data-driven adaptive service planning in public transportation: A case study on passenger behavior analysis. Sustain Cities Soc. 2022;80:103794.