Aldi SAPUTRA[1], Rendra GUSTRIANSYAH[2], Ahmad SANMORINO[3*], Zaid Romegar MAIR[4], Dewi SARTIKA[5], Shinta PUSPASARI[6]

Department of Informatics Engineering, Faculty of Computer and Science, Universitas Indo Global Mandiri (1)(2)(4)(5)(6), Department of Information System, Faculty of Computer and Science, Universitas Indo Global Mandiri (3)
ORCID: 1. https://orcid.org/0009-0000-8340-0859; 2. https://orcid.org/0000-0001-7600-1147; 3. https://orcid.org/0000-0002-4949-4377; 4. https://orcid.org/0009-0008-2654-5881; 5. https://orcid.org/0009-0001-9747-3183; 6. https://orcid.org/0000-0001-5906-3362

doi:10.15199/48.2024.10.07

Prediction Passenger Numbers in Light Rail Transit using Seasonal Autoregressive Integrated Moving Average (SARIMA)

Abstract. *Light Rail Transit (LRT) plays a role in supporting the mobility of the people of a city. However, the increase in LRT use presents challenges, requiring effective solutions to anticipate changes in the number of passengers. This research aims to design and implement a prediction model using the Seasonal Autoregressive Integrated Method Moving Average to anticipate and predict the number of LRT passengers. The prediction results using the parameter model (0,1,1)(0,1,0) obtained a MAPE value of 16.69%, thus, the accuracy level obtained was 83.31%.*

Streszczenie. *Tranzyt koleją lekką (LRT) odgrywa rolę we wspieraniu mobilności mieszkańców miasta. Jednakże wzrost wykorzystania LRT stwarza wyzwania wymagające skutecznych rozwiązań umożliwiających przewidywanie zmian w liczbie pasażerów. Celem badania jest zaprojektowanie i wdrożenie modelu predykcyjnego wykorzystującego sezonową, zintegrowaną metodę autoregresyjną, średnią ruchomą do przewidywania i przewidywania liczby pasażerów LRT. Wyniki predykcji z wykorzystaniem modelu parametrycznego (0,1,1)(0,1,0) uzyskały wartość MAPE na poziomie 16,69%, a zatem uzyskany poziom dokładności wyniósł 83,31%. (**Prognozowanie liczby pasażerów w transporcie szynowym przy użyciu sezonowej autoregresyjnej zintegrowanej średniej ruchomej (SARIMA)**)*

Keywords: Time series, Light Rail Transit, Prediction, Seasonal Autoregressive Integrated Moving Average
Słowa kluczowe: Szeregi czasowe, transport koleją miejską, prognoza, sezonowa zintegrowana średnia krocząca z autoregresją

Introduction

Mass transportation has become very important to support the mobility of urban communities in the modern era because urban infrastructure is increasingly developing and globalization is increasing. Reliable, efficient, and sustainable transportation systems are becoming increasingly important because urbanization continues to increase with population density in city centers. Choosing the right transportation method can impact residents' quality of life, economic productivity, and the environmental impact of urban growth. One of the rapid transportation initiatives in Palembang is Light Rail Transit (LRT). The LRT has 13 stations on its rapid transit lines, encouraging the growth of high-density, transit-oriented cities. The Palembang LRT has seen increased usage since its introduction. During holidays there is often a build-up or queue of people wanting to travel using the LRT. LRT management has to find solutions to overcome this. The problem that arises is how LRT management can recognize changes in passenger numbers and respond accordingly.

This research aims to predict the number of LRT passengers using the Seasonal Autoregressive Integrated Moving Average (SARIMA) method. Evaluation of the success of this research is by looking at the level of accuracy of the SARIMA method in predicting the number of Palembang LRT passengers.

Several previous studies regarding mass transportation predictions have been carried out by Alfikrizal et al. (2020) in their research entitled Monte Carlo Simulation in Predicting the Number of Mass Transport Passengers for Rapid Transit Buses in Padang City [1]. The results of the research show that the predicted number of passengers for 2018 based on 2017 data is 2,182,242 with an average accuracy of 82.43 percent, and the predicted number of passengers for 2019 based on 2018 data is 2,689,626 with an average accuracy amounting to 83.80 percent. Then, research by Utomo and Fanani (2020) entitled forecasting the Number of Train Passengers in Indonesia Using the Seasonal Autoregressive Integrated Moving Average (SARIMA) Method predicted the number of train passengers [2]. The research results show that the prediction for 2020 is 36,941,500 passengers, with an MSE error calculation of 0.046875, and supported by a MAPE value of 6.26 percent. The SARIMA method was also used in Inka Durrah et al. (2018) research entitled Forecasting the Number of Airplane Passengers at Sultan Iskandar Muda Airport using the SARIMA Method [3]. The results of this research show that the highest number of passengers in January was at Sultan Iskandar Muda Airport, and is predicted to increase in December 2017, namely 101,484 people. Meanwhile, the lowest number of passengers is predicted to occur in March 2017, namely 87,899 people.

Method

The steps to achieve the objectives in this research are illustrated in Figure 1.

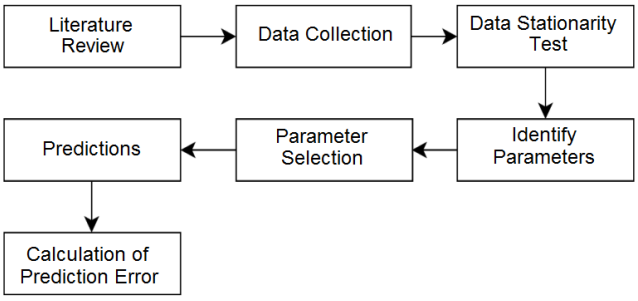

Fig.1. Research method

The literature study carried out by researchers was by reading journals or papers related to time series analysis, the use of the SARIMA method, and predictions of the number of public transportation passengers. Data collection began with secondary data obtained from the South Sumatra Light Railway Management Center. This dataset includes the number of passengers per month from August 2018 to December 2021, which will be training data. Data for the period January 2022 to December 2023 will be used as testing data, allowing testing of model performance in a period that has never been seen before. Table 1 shows the number of LRT passengers for 2018-2023 which has been collected from the South Sumatra Light Railway Management Center.

Table 1. LRT passengers for 2018-2023

Month	Year					
	2018	2019	2020	2021	2022	2023
Jan		146,954	313,502	102,902	193,990	316,521
Feb		109,053	243,181	91,641	146,524	261,581
Mar		153,979	160,473	117,323	195,847	279,621
Apr		194,345	14,759	112,413	166,525	361,765
May		150,941	12,589	147,884	352,840	336,260
Jun		322,628	21,924	159,596	260,725	355,364
Jul		277,801	31,470	102,580	286,978	370,468
Aug	284,618	202,526	48,483	89,996	254,043	345,592
Sep	176,229	211,105	39,673	125,082	268,444	351,418
Oct	144,653	255,546	46,806	178,417	310,179	344,388
Nov	155,699	244,722	52,462	173,679	294,964	342,228
Dec	186,035	361,559	68,171	197,619	356,676	417,513

Testing data stationarity involves evaluating the variance and mean of the data. Because time series plots are not always sufficient to identify data stationarity, other methods such as the Augmented Dickey-Fuller (ADF) test are used to test the average, while the Box-Cox Test is used to test the variance [4, 5]. By combining these methods, stationarity analysis can be performed more comprehensively, allowing for a more accurate assessment of the stationarity properties of the data. At the parameter identification stage, autocorrelation feature (ACF) [6], and partial autocorrelation feature (PACF) analysis [7] is used to determine the best parameters in the SARIMA model (p,d,q)(P,D,Q)s. ACF Plots help identify direct correlations with specific lags, while PACF Plots indicate correlations across lags. The non-seasonal components (p,d,q) can be observed from lag 1, while the seasonal components (P,D,Q) can be identified from peaks located at multiples of the seasonal length.

The parameter selection stage involves selecting the most suitable parameters by carrying out t-tests and statistical tests such as Ljung-Box [8]. Testing is used to evaluate the individual significance of each parameter in the model, providing information about whether each parameter contributes significantly to the model [9]. Meanwhile, the Ljung-Box statistical test was carried out to identify whether the residuals from the model parameters were white noise. Testing of residuals is carried out by examining the p-value. P-value is a statistical metric that reflects how strongly data evidence supports or rejects the null hypothesis. If the p-value exceeds 0.05 (a commonly chosen significance level), this indicates that there is insufficient evidence to reject the null hypothesis, which states that the residual is white noise.

After successfully identifying the best parameters through historical data analysis, the next step in applying the SARIMA method is to integrate these parameters into the predicting process. It is important to understand the role of each SARIMA parameter, such as orders that include an autoregressive component (p), a differences component (d), and a moving average component (q). By understanding these characteristics, we can identify optimal parameters to improve prediction accuracy. After the process of selecting parameters and model formation, the next step involves calculating the MAPE value (mean absolute percentage error) based on the actual data using the parameter model that has been selected for forecasting. The MAPE indicates the percentage of the average error between the prediction value and the actual value. Assessment of the resulting mape value aims to determine the extent of the accuracy of the parameter model. The comparison process between predictive results obtained from training data with actual data from January 2022 to December 2023 will provide an overview of how well the model can produce predictions that are close to the actual value. If the MAPE value is close to zero, it shows that the model has a high level of accuracy. Conversely, a higher mape value indicates a greater level of error in predicting.

Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA is a time series method that combines seasonal (seasonal), autoregressive (AR), average moving (MA), and integration. General notation that forms the SARIMA model (p, d, q)(P, D, Q)$_S$ is explained as follows: *p* is the sequence of auto-regression components, *d* is the amount of differentiation to achieve stationary, *q* is the sequence of moving levels components, *P* is the sequence of components Seasonal Auto-regression, *D* is a sequence of seasonal differentiation, *Q* is the sequence of seasonal movable levelling components, and S is the number of periods per season: *p* - Autoregression component sequence, *d* - The amount of differentiation applied to achieve stationary, *q* - The sequence of moving leveling components, *P* - Sequence of seasonal autoregression components, *D* - Seasonal differentiation order, *Q* - Sequence of seasonal leveling components, *S* - Number of periods per season. The general form of SARIMA (p,d,q)(P,D,Q)$_S$ is shown in Equation 1.

$$(1)\quad \Phi_p B^S \phi_p(B)(1-B)^d (1-B^S)^D z_t = \theta_q(B)\Theta_q(B^S)a_t$$

where: $\phi_\rho(B)$ – AR non-seasonal, $\Phi_\rho B^S$ - AR seasonal, $(1-B)^d$ - differencing non seasonal, $(1-B^S)^D$ - differencing seasonal, $\theta_q(B)$ - MA non seasonal, $\Theta Q(B^S)$ - MA seasonal.

Evaluation Metrics

Mean Absolute Percentage Error (MAPE) is an evaluation metric that measures the average percentage error between the actual value and the predicted value. The initial step in calculating MAPE is to calculate the absolute error for each period by comparing the data used with the observed values for that period [10]. Next, the absolute error results are divided by the observation value in that period. Finally, MAPE is calculated by taking the average of these absolute percentage errors. Equation 2 presents the general form of the Mean Absolute Percentage Error (MAPE).

$$(2)\quad MAPE = \left(\frac{1}{n}\right) * \sum \left| \frac{y_i - y}{y_i} \right| * 100\%$$

where: n - amount of data used in the calculation, y_i - actual value of the ith data, y - predicted value for the ith data. There is a range of values that are used as indicators to measure the ability of a predicting model. The range is shown in Table 2.

Table 2. The MAPE value range

Range	Explanation
<10%	The predicting model is very good
10-20%	Good predicting model
20-50%	Feasible predicting model
>50%	Bad predicting model

Result and Discussion

Table 3 displays the predicted, actual, and MAPE values obtained from January 2022 to December 2023.

In Table 3, the MAPE calculation results for each month and year are shown in detail. MAPE is an important evaluation metric in evaluating the accuracy of a prediction model. The MAPE calculation results for each month show the relative error level of the model predictions to the actual values. From Table 3, the MAPE value is generated by calculating the average of each MAPE calculation, resulting

in a MAPE value of 16.69%. This means that on average, model predictions have an error of 16.69% from the actual value. From this MAPE value, the model accuracy level can be estimated at 83.31%, calculated by subtracting the MAPE value from 100%, or 100%−16.69% = 83.31%. This shows that the prediction model has a fairly good level of accuracy in predicting the value of the observed variables.

Table 3. MAPE metrics evaluation

Month	Year	Predicted	Actual	MAPE
Jan	2022	234,137	193,990	20.69%
Feb	2022	222,877	146,524	52.11%
Mar	2022	248,558	195,847	26.91%
Apr	2022	243,648	166,525	46.31%
May	2022	279,119	352,840	20.89%
.....	.....	.....	.....	.....
Nov	2023	436,149	342,228	27.44%
Dec	2023	460,089	417,513	10.19%
			Average	16.69%

This analysis depicts prediction error rates for various periods, providing insight into the quality and reliability of the prediction model. For example, May 2022 has an error of 20.89%, indicating significant inaccuracy. Meanwhile, June 2022 has an error of 11.54%, showing relatively accurate predictions. October 2022 shows an error of only 0.17%, indicating a very good prediction. However, January 2023 has an error of 15.43%, indicating a fairly large deviation. Finally, March 2023 has an error of 35.82%, indicating predictions are far from actual values. The average prediction error of all data is 16.69%, which is an indicator of the overall quality of the prediction model.

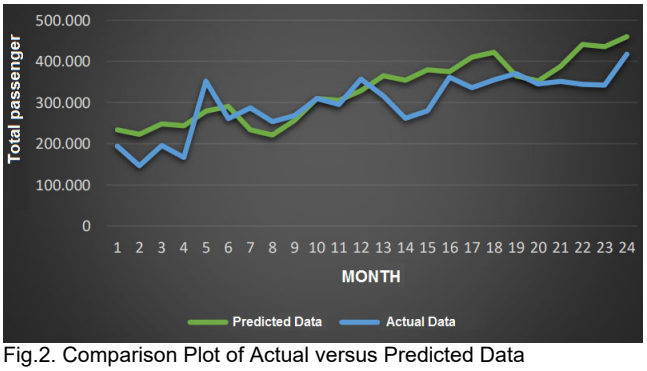

Fig.2. Comparison Plot of Actual versus Predicted Data

Figure 2 shows the comparison of actual and predicted values which is very important for model evaluation. The trend comparison results show consistency between the predicted and actual lines from Month to Month, indicating the model's ability to follow and reproduce the trend well. The minimal difference between the two lines gives an idea of the model's accuracy, where smaller deviations indicate better predictions. In other words, the closer the predicted line is to the actual line, the more accurate the model is in predicting the actual value.

Conclusions

Based on the discussion that has been carried out, it is concluded that this research has succeeded in predicting the number of Palembang LRT passengers using the Seasonal Autoregressive Integrated Moving Average method. With a MAPE value of 16.69%, the prediction results can be considered good because they are at the MAPE range value level of 10%-20%. Thus, the prediction accuracy level was obtained at 83.31%.

Acknowledgments

We would like to acknowledge Universitas Indo Global Mandiri for supporting this study.

***Authors**: Aldi SAPUTRA, Department of Informatics Engineering, Faculty of Computer and Science, Universitas Indo Global Mandiri, Indonesia, E-mail: 2020110051@students.uigm.ac.id;*
Rendra GUSTRIANSYAH, Department of Informatics Engineering, Faculty of Computer and Science, Universitas Indo Global Mandiri, Indonesia, E-mail: rendra@uigm.ac.id;
Ahmad SANMORINO, Department of Information System, Faculty of Computer and Science, Universitas Indo Global Mandiri, Indonesia, E-mail: sanmorino@uigm.ac.id;
Zaid Romegar MAIR, Department of Informatics Engineering, Faculty of Computer and Science, Universitas Indo Global Mandiri, Indonesia, E-mail: zaidromegar@uigm.ac.id;
Dewi SARTIKA, Department of Informatics Engineering, Faculty of Computer and Science, Universitas Indo Global Mandiri, Indonesia, E-mail: dewi.sartika@uigm.ac.id;
Shinta PUSPASARI, Department of Informatics Engineering, Faculty of Computer and Science, Universitas Indo Global Mandiri, Indonesia, E-mail: shinta@uigm.ac.id.

*Corresponding Author: sanmorino@uigm.ac.id

REFERENCES

[1] Alfikrizal, K., Defit, S. & Yunus, Y. Simulasi Monte Carlo dalam Prediksi Jumlah Penumpang Angkutan Massal Bus Rapid Transit Kota Padang. *J. Inform. Ekon. Bisnis* 3, 78–83 (2020).
[2] Utomo, P. & Fanani, A. Peramalan Jumlah Penumpang Kereta Api di Indonesia Menggunakan Metode Seasonal Autoregressive Integrated Moving Average (SARIMA). *J. Mhs. Mat. Algebr.* 1, 169–178 (2020).
[3] Durrah, F. I., Yulia, Y., Parhusip, T. P. & Rusyana, A. Peramalan Jumlah Penumpang Pesawat Di Bandara Sultan Iskandar Muda Dengan Metode SARIMA (Seasonal Autoregressive Integrated Moving Average). *J. Data Anal.* 1, 1–11 (2018).
[4] Gianfreda, A., Maranzano, P., Parisio, L. & Pelagatti, M. Testing for integration and cointegration when time series are observed with noise. *Econ. Model.* 125, 106352 (2023).
[5] Ichihara, K., Yamashita, T., Kataoka, H. & Sato, S. Critical appraisal of two Box-Cox formulae for their utility in determining reference intervals by realistic simulation and extensive real-world data analyses. *Comput. Methods Programs Biomed.* 242, 107820 (2023).
[6] Podulka, P. *et al.* Roughness evaluation of turned composite surfaces by analysis of the shape of autocorrelation function. *Meas. J. Int. Meas. Confed.* 222, (2023).
[7] Balawi, M. & Tenekeci, G. Time series traffic collision analysis of London hotspots: Patterns, predictions and prevention strategies. *Heliyon* 10, e25710 (2024).
[8] Schoukens, J., Westwick, D., Ljung, L. & Dobrowiecki, T. Nonlinear system identification with dominating output noise - A case study on the silverbox. *IFAC-PapersOnLine* 54, 679–684 (2021).
[9] Sanmorino, A., Gustriansyah, R., Terttiaavini & Isabella. The toolkit of success rate calculation of broiler harvest. *Telkomnika* 15, 1947-1954 (2017).
[10] Gustriansyah, R., Ermatita, E. & Rini, D.P. An approach for sales forecasting. *Exp. Sys. With App.* 207, 118043 (2022).