Forecasting Using SARIMA and Bayesian Structural Time Series Method for Range Seasonal Time

M Rizal[1], **S U Zuliana**[1,*], **M W Musthofa**[1]

[1] UIN Sunan Kalijaga, Yogyakarta, Indonesia

*Corresponding author's e-mail : sri.zulianai@uin-suka.ac.id

ABSTRACT: SARIMA and Bayesian Structural Time Series are time series methods that can be used for data that contains seasonality. Data on the number of train passengers in the Java region has a seasonal pattern. This research aims to determine the steps of the SARIMA model and Bayesian Structural Time Series, applying the SARIMA model and Structural Bayesians Time Series, get the forecasting results of the SARIMA model and Bayesian Structural Time Series with MAPE measurements. The research method used is a quantitative method applied to data on the number of PT KAI train passengers in the Java region for 2006-2019. The results of this research show that the best model for forecasting the number of PT KAI train passengers in the Java region in 2006-2019 is SARIMA (2,1,0)(0,1,2)[12] with a MAPE value of 4.77% compared to the Bayesian method structural time series [12] namely 5.25%.

1. Introduction

Time series data forecasting and modeling has been widely used in various sectors, including transportation. Various methods have been developed to handle time series data analysis, most of them assuming that the residuals are normally distributed and stationary [1] . One of them is the Seasonal Autoregressive Integrated Moving Average (SARIMA), a method introduced by George Box and Gwilyn Jenkins in 1970 [2]. SARIMA is an adoption of the ARIMA method which is specifically for data that is influenced by seasonal factors.

However, besides conventional statistical methods such as SARIMA, there are also approaches that use Bayesian methods. Bayesian Structural Time Series models are one of several methods that can be used in forecasting data that contains seasonal patterns. The Bayesian Structural Time Series (BSTS) model, developed by Steven L. Scott and Hal Varian in 2014 [3], has the advantage of producing time series data that is more structured and often in a probabilistic form.

Therefore, the target of this research is to predict the number of commuter train passengers in the Java region for the next N periods by comparing two forecasting methods, namely Seasonal Autoregressive Integrated Moving Average (SARIMA) and Bayesian Structural Time Series (BSTS).

2. Theorical background

2.1. Forecasting

Forecasting is predicting something that has not yet happened [4]. Forecasting aims to predict the future value of a time series $x_{n+m}, m = 1,2,..,$ where x_{n+m} is the value that will be predicted from the time

series for the next m periods, starting from n. Based on data collected to date $x_{1n} = \{x_1, x_2, \dots, x_n\}$. Where x_{1n} is historical data collected from period 1 to n sequentially [5].

2.2. Time Series Analysis

Time series data is a series of values for a variable certain sequence each time. There are several things that must be considered when analyzing time series data, including data stationarity, autocorrelation function (ACF), and partial autocorrelation function (PACF) [6].

2.3. Seasonal Autoregressive Integrated Moving Average (SARIMA)

This Seasonal Autoregressive Integrated Moving Average model is denoted by SARIMA (p,d,q)(P,D,Q)s [7]. In this model there are 2 parts which are denoted by lowercase and capital letters, (p, d, q) is the notation for the part of the model that does not contain seasonal patterns. Meanwhile (P, D, Q) is a notation for the part of the model that contains seasonal patterns. The power s is the notation for the number of periods per season that will be calculated. The general equation of the Seasonal Autoregressive Integrated Moving Average model is as follows [8]:

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)\alpha_t \quad (1)$$

2.4. Bayesian Methods

The Bayesian model was developed from the Bayes method by Thomas Bayes in 1763. The method used for the Bayesian model is called the Bayesian method. In classical Bayes' theorem, probability theory can be written with the equation [9]

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \quad 2)$$

Where θ indicates the parameter to be estimated.

2.5. Bayesian Time Series Models

Bayesian Time Series Model Analysis is a linear dynamic model made from machine learning techniques used to view time series patterns. Bayesian Time Series models are also known as forecasting methods that contain probabilities that depend on joint probability distributions. For example, stating data in the t-th period in a time series observation, the structural time series model can be defined as follows:$y_{1:T} = y_1, \dots y_T p(y_{1:T})y_t$

$$y_t = x_t\beta + z_t\theta_t + \varepsilon_t \quad \text{and} \quad \theta_t = T_t\theta_{t-1} + \mu_t \quad (3)$$

where y_t is observation data in the t-th time period. x_t dimensional vector $1 \times K$ with independent variable elements and z_t is a $1 \times p$ dimension vector contains known input components of time series data (trend and seasonality) and is known to have a constant value. Meanwhile, β is the regression coefficient of the independent variable. θ_t represents $p \times 1$ dimensional vector containing p state parameter equations. Assumed ε_t independent is identical to the $N(0, h^{-1})$ distribution where h is the precision parameter which is defined as $h = \frac{1}{\sigma^2}$.

2.6. Measures of Forecasting Accuracy

The forecasting models obtained are then validated for their accuracy using several indicators, such as:

1. Mean Absolute Percentage Error (MAPE)
   The level of accuracy using MAPE is calculated by using the absolute error rate in each period divided by the actual absolute value, then the error percentage is averaged by dividing it by a number of observations and multiplying by 100% [10].

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (4)$$

2. Root Mean Square Error (RMSE)

Root Mean Square Error is a forecasting measure that is calculated by rooting the MSE obtained from an experience, where the MSE itself is obtained from the average of the average difference between the actual value and the actual value itself and then squared.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2 \quad (5)$$

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E((\hat{\theta} - \theta)^2)} \quad (6)$$

3. Methodology

3.1. Data Processing Methods

The data used in this research is data on the number of passengers PT.KAI trains for the Java region for the 2006-2019 period, which is secondary data. The data is published by the Badan Pusat Statistik (BPS). In this research, the data management and analysis methods for SARIMA and Bayesian Structural Time Series are as follows:

1. Divide the data into two parts, namely training and testing data.
2. Descriptive analysis of data on the number of PT.KAI train passengers in the Java region.
3. SARIMA method
   a. Carry out data stationarity checks.
   b. Identifying SARIMA models through ACF and PACF plots.
   c. Carrying out parameter estimation and checking the significance of parameters from the tentative model.
   d. Carry out model diagnostics to determine the suitability of the model.
   e. Overfitting to compare with the initial tentative model. Then a significant parameter test was carried out.
   f. Choose the best model from the SARIMA models that have been formed by looking at the smallest AIC and BIC values.
   g. Forecast as much as test data to calculate the MAPE value.
4. Bayesian Structural Time Series Method
   a. Selection or determination of prior distribution.
   b. Using the best prior as a reference and combining it with likelihood data.
   c. The prior and likelihood distributions that have been previously searched can be combined and will produce a posterior distribution that will be used in calculating parameter estimates.
   d. Parameter estimation calculations were carried out using MCMC simulation which uses the full posterior configurational distribution of each parameter.
   e. Identify seasonal periods from training data that will be used in forecasting the number of train passengers.
   f. Forecast as much as test data to calculate the MAPE value.
5. Comparing the forecasting results of the SARIMA and Bayesian Structural Time Series methods by looking at the smallest MAPE value.

4. Results and Discussion

4.1. Descriptive Analysis

Table 1. Descriptive Statistics

Amount of data	Min.	Standard deviation	1st Qu	Median	Mean	3rd Qu	Max.
168	10759	7809.612	15763	17465	21899	28802	38303

Based on data from PT KAI Indonesia, the growth of train passengers in the Java region, as in Figure 1, always increases every year. In the last five years the movement has always shown quite an increase. The highest number of passengers occurred in July 2019, namely 38,303 passengers. Meanwhile, the lowest number of passengers occurred in February 2007 with a total of 10,759 passengers.

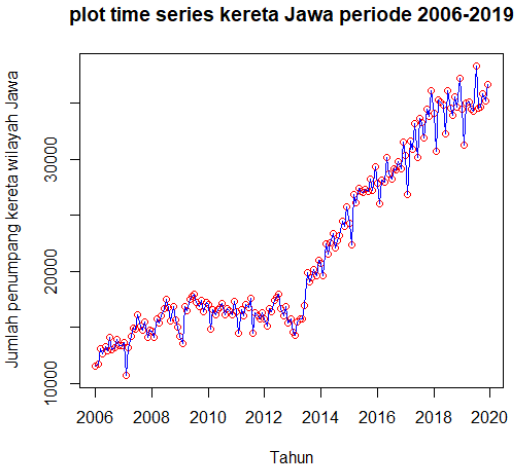

Figure 1. Time series data on the number of PT.KAI train passengers in the Java region 2006-2019

4.2. Seasonal ARIMA

a. Stationary in variance

On inspection Data stationarity in variance using Box-Cox obtained a rounded value (λ) of 1.059591. The lambda value is greater than or equal to 1 with a confidence interval of 95% so there is no need for data transformation. The lambda value shows that the data on the number of PT.KAI train passengers in the Java region.

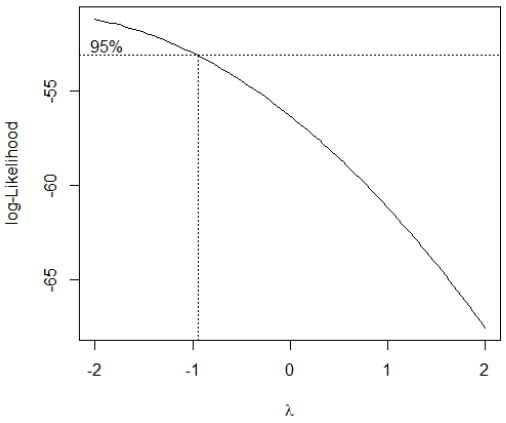

Figure 2. Cox data box

b. Stationary in the mean

Stationary in the average indicates a condition where the data is inaccurate at a constant average value. Stationarity in the average can be identified by looking at the autocorrelation function (ACF) plot, the autocorrelation value of data that is stationary will decrease to zero after a time difference (lag) in the fifth to sixth bracket depending on the pattern in the existing data. The ACF plot visualization can be seen in Figure 3.

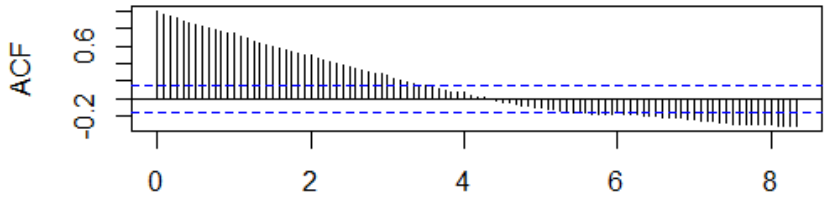

Figure 3. Non stationary ACF plot

On picture 3 It can be seen that in the ACF plot the lag decreases significantly towards zero so it can be ascertained that the data is not stationary in the mean. We can also check stationarity using the Augmented Dickey Fuller (ADF) test. The ADF test showed that the p-value was 0.9105. These results indicate that the data is not stationary in the mean because it produces a p-value that is greater than the significance level value. This indicates that there is a nonstationary nature in the average, so differencing is necessary.

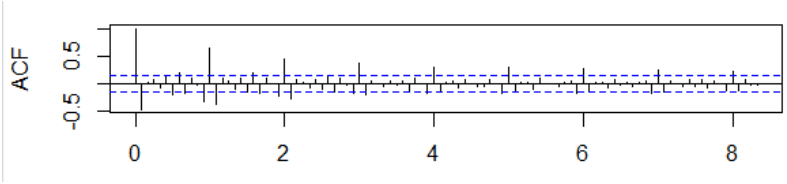

Figure 4. ACF plot after differencing

The stationarity of the data can be seen after differencing where the ADF value obtained is 0.01 or smaller than the significant level, so it can be concluded that the data on the number of PT.KAI train passengers in the Java region is in a stationary state on average. Data that is stationary in terms of mean and variance will be used to identify and estimate the parameters of the tentative SARIMA model.

c. Model Identification and Parameter Estimation

After testing the stationarity of the data for both the average and variance, a temporary model was obtained from the results of the data test. The model used is (p, d, q)(P, D, Q)s, where p is the assumption for non-seasonal AR values, while P is the assumption for seasonal AR values, d and D are seasonal and non-seasonal differencing processes, while q is the order of the non-seasonal MA and Q is the order of the seasonal MA.

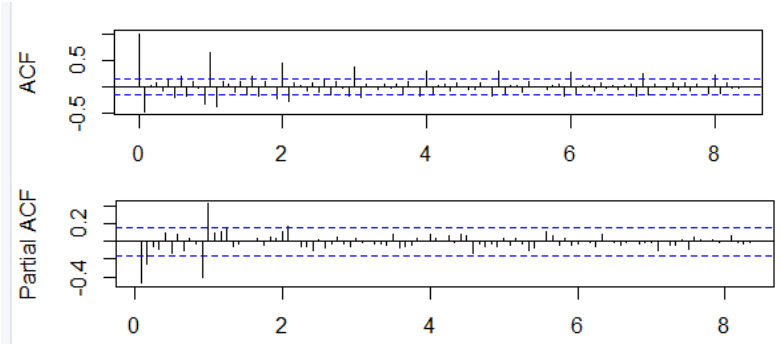

Figure 5. Non-seasonal differencing ACF and PACF plots

Because data on the number of train passengers PT. KAI in the Java region is seasonal, so a process is carried out *differencing* on *lag* to 12 because the data taken is monthly lag 12 means in the 12th month. The following is a plot of ACF and PACF after the process is carried out *differencing* on lag the 12th.

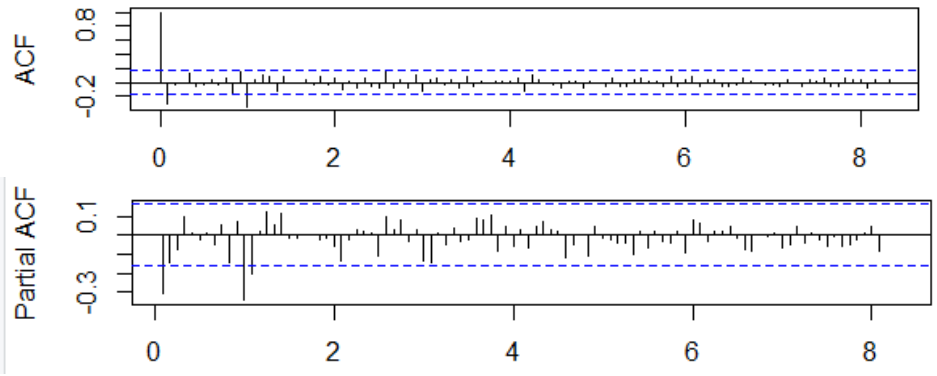

Figure 6.ACF and PACF plots *differencing* seasonal

From the results of data tests that have been carried out by observing the ACF and PACF plots for non-seasonal and seasonal, a tentative model was obtained. Possible initial tentative models based on ACF and PACF plots are SARIMA (2,1,0)(0,1,2)[12], SARIMA (1,1,0)(0,1,1)[12], SARIMA (1,1,0)(0,1,2)[12].

The parameter estimation results of the candidate models obtained at the model identification stage are presented in Table 2.

Table 2. Determination of the best model based on AIC and BIC

Model	Parameter	Coefficient Parameter	*P*-Value	AIC	BIC
SARIMA (2,1,0)(0,1,2)[12]	AR (1)	-0.4296	0.0000	2364.69	2379.5
	AR (2)	-0.2119	0.0119		
	SMA (1)	-0.5244	0.0000		
	SMA (2)	-0. 0693	0.4410		
SARIMA (1,1,0)(0,1,1)[12]	AR (1)	-0.3534	0.0000	2367.74	2376.62
	SMA (1)	-0.5116	0.0000		
SARIMA (1,1,0)(1,1,2)[12]	AR (1)	-0.3545	0.0000	2369	2380.85
	SMA(1)	-0.4933	0.0000		
	SMA (1)	-0.0787	0.3926		

Based on Table 2, the model that has significant parameter estimator values at the 95% significance level is the SARIMA model (2,1,0)(0,1,2)[12]. This model has the smallest AIC and BIC values. The selection of the smallest AIC and BIC is expected to produce the smallest possible error [11]. Significant models will undergo diagnostics.

d. Model Diagnostics

Diagnostic tests are used to see whether the residuals from the model are white noise or not. White noise shows that the data is random and stationary, which is a requirement for forecasting. Diagnostic checks can be carried out using the Ljung-Box Test. The p-value obtained from the diagnostic test is 0.991> α (0.05), then the decision fails to reject H0. So by using a 95% confidence level it can be concluded that there is no autocorrelation between the residuals. Based on the results of the diagnostic tests that have been carried out, it can be said that the white noise assumption is met in the best model, namely the SARIMA model (2,1,0)(0,1,2)[12], so that the model is suitable for use for forecasting.

e. SARIMA Method Forecasting Results

The following are the forecasting results for the 2019 using the SARIMA model equation (2,1,0)(0,1,2)[12].

Table 3. SARIMA forecasting (2,1,0)(0,1,2)[12]

No	Month	Actual Data	Forecasting Results
1	January	34435	35468.94
2	February	31282	32676.91
3	March	35068	36656.41
4	April	35106	36321.40
5	May	34514	37077.57
6	June	34261	35183.73
7	July	38303	37617.82
8	August	34542	36870.38
9	September	34615	36320.64
10	October	35814	37901.78
11	November	35228	37064.69
12	December	36710	39331.13
	MAPE		**4.77%**

4.3. Bayesian Structural Time Series

Jawa region train passenger data has seasonal patterns and trends caused by the effect of holiday. Calculations will be carried out using *Bayesian* with seasonal components S={12}, the S=12 index shows a seasonal pattern that repeats itself every 12 months. This model uses simulation *Markov Chain Monte Carlo* with iteration n = 1500 to find the estimated values of the model parameters. Next, this model applies a calculation pattern *dynamic* in his forecasting [12].

4.4. Forecasting Using Bayesian Structural Time Series

Known simple equations in the model *Bayesian Structural Time Series* with component input *trend* and seasonal, namely as follows:

$$y_t = \mu_t + \gamma_{1t} + \upsilon_t \sim N(0, \sigma^2{}_{obs})$$
$$\mu_t = \mu_{t-1} + \partial_{t-1} + \omega_{1t}; \sim N(0, \sigma^2{}_{level})$$
$$\partial_t = \partial_{t-1} + \omega_{2t}; \omega_{2t} \sim N(0, \sigma^2{}_{slope})$$
$$\gamma_{1t} = -\sum_{S-1}^{S-1} \gamma_{s,t-1} + \omega_{3t} \sim N(0, \sigma^2{}_{seas})$$
$$\gamma_{st} = \gamma_{s-1,t-1}$$

Below is a plot of the probability distribution of the four parameters in the equations above (σ^2_{obs},σ^2_{levels},σ^2_{slope} and σ^2_{seas}):

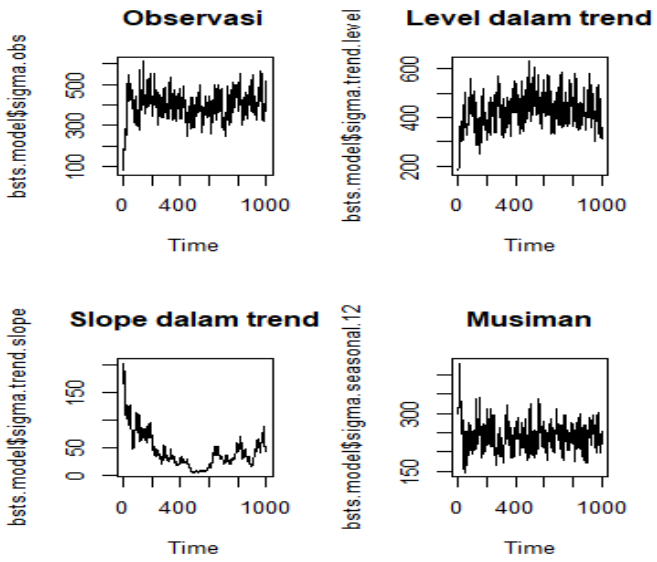

Figure 7. Plots *time series* variance parameters

Based on the plot time series Figure 7 shows that the parameters for the observation equation have a combination of components trend and seasonal. On the component plot trend It can be seen that there is an increase in the parameter value in each iteration and there is a slight decrease trend at the end of the iteration. As for parameters slope it is clearly visible trend down. In the seasonal component parameter plot, it appears that there is a periodic repetition in the parameter values even though they form trend down.

After knowing the model parameter values. Furthermore, the forecast for the number of PT.KAI train passengers in the Java region for 2006-2019 for the 2019 forecasting results is presented in the following table:

Table 4. Model forecasting *Bayesian Structural Time Series*[12]

No	Month	Actual Data	Forecasting Results
1	January	34435	35392.13
2	February	31282	32548.12
3	March	35068	36799.13
4	April	35106	36569.16
5	May	34514	37615.59
6	June	34261	35474.00
7	July	38303	38442.00
8	August	34542	37395.34
9	September	34615	36621.96
10	October	35814	38272.30
11	November	35228	37186.11
12	December	36710	39516.81
	MAPE		**5.25%**

4.5. *Comparison of SARIMA and BSTS Forecasting Performance*

Table 5. Comparison of SARIMA and BSTS Forecasting Performance.

No	Month	Actual Data	Forecasting Results	
			SARIMA(2,1,0)(0,1,2)[12]	BSTS [12]
1	January	34435	35468.94	35392.13
2	February	31282	32676.91	32548.12
3	March	35068	36656.41	36799.13
4	April	35106	36321.40	36569.16
5	May	34514	37077.57	37615.59
6	June	34261	35183.73	35474.00
7	July	38303	37617.82	38442.00
8	August	34542	36870.38	37395.34
9	September	34615	36320.64	36621.96
10	October	35814	37901.78	38272.30
11	November	35228	37064.69	37186.11
12	December	36710	39331.13	39516.81

Based on Table 5 which is the result of calculating the MAPE value from both methods, the MAPE value for forecasting the number of PT train passengers is obtained. KAI for the Java region in 2006-2019 used the SARIMA method (2,1,0)(0,1,2)[12] which was 4.77%. Meanwhile, for the Bayesian Structural Time Series method, the MAPE forecasting value was 5.25%. Based on the mean percentage error (MAPE) criteria, a model has good performance if the MAPE value is below 10%. So it can be said that the forecasting results from these two methods are very good because they have MAPE values below or less than 10%.

The MAPE value obtained from the SARIMA method (2,1,0)(0,1,2)[12] is 4.77% which is smaller than the Bayesian Structural Time Series method of 5.25%. So it can be concluded that the SARIMA (2,1,0)(0,1,2)[12] method is better to use for forecasting in this case.

5. Conclusion

1. Comparison of SARIMA methods and *Bayesian* structural time series Based on the forecasting results for 12 months, it was found that the two methods each had a forecasting error or MAPE value below 10%, which means that both methods were very good at forecasting for the next 12 months. However, the SARIMA method is still better than the Bayesian structural time series method. The accuracy of the MAPE forecasting measurement of the SARIMA model (2,1,0)(0,1,2)[12] is 4.77% and Bayesian structural time series [12] is 5.25%.
2. The forecasting results of the SARIMA method (2,1,0)(0,1,2)[12] on the number of PT.KAI train passengers in the Java region in 2006-2019 show that in 2019 the highest number of passengers occurred in December which reached 39331passengers and the lowest number of passengers occurred in February of 32676 passenger.
3. The forecasting results of the Bayesian Structural Time Series method [12] on the number of PT.KAI train passengers in the Java region in 2006-2019 show that in 2019 the highest number of passengers occurred in December, reaching 39516 passengers and the lowest number of passengers occurred in February of 32548 passenger.
4. The results of this forecasting have a positive impact on the railway industry in optimizing schedules, train capacity, efficient ticket pricing including route expansion and additional services.

References

[1] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and its Applications*, 4th ed. Springer, 2016.

[2] G. M. Box, G. E. & Jenkins, "Time series analysis: Forecasting and control," *san Fr. Calif Holden-Day.*, 1976.
[3] H. R. Scott, S. L. & Varian, "Predicting the present with bayesian structural time series," *Int. J. Math. Model. Numer. Optim.*, vol. 5, pp. 4–23, 2014.
[4] P. Subagyo, *Forecasting:Konsep dan Aplikasi*. BPPE–UGM, Yogyakarta, 1986.
[5] D. S. Shumway, R. H., Stoffer, D. S., & Stoffer, "Time series analysis and its applications (4th ed.)," *Springer*, vol. 3, 2016.
[6] W. W. S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed. New York: Departemen of Statistics The Fox School of Business and Management Temple University, 2006.
[7] P. Utomo, "Peramalan jumlah penumpang kereta api di Indonesia menggunakan metode seasonal autoregressive integrated moving average," PhD thesis, UIN Sunan Ampel Surabaya, 2020.
[8] M. Montgomery, D. C., Jennings, C. L., & Kulahci, *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.
[9] J. Pole, A., West, M., & Harrison, *Applied Bayesian forecasting and time series analysis*. CRC press, 1994.
[10] M. A. Maricar, "Analisa perbandingan nilai akurasi moving average dan exponential smoothing untuk sistem peramalan pendapatan pada perusahaan xyz," *J. Sist. dan Inform.*, vol. 13(2), pp. 36–45, 2013.
[11] R. S. Tsay, *Analysis of Financial Time Series*, 3rd ed. Chicago: Wiley, 2010.
[12] V. N. Putri, "Peramalan Curah Hujan Harian Kota Bandung Menggunakan Metode Bayesian Time Series Model," Universitas Padjajaran, 2020.