Forecasting The Number of Train Passenger in Sumatra Using ARIMA Models

Fika Lestauli Sigalingging[1*], Maria Yus Irsan[2], Junianto Sesa[3]
[1,2]*Study Program of Actuarial Science, School of Business, President University, 17550, Indonesia*
[3]*Study Program of Mathematics, Faculty of Mathematics and Natural Sciences,Universitas Papua, Manokwari, Papua Barat 98314, Indonesia*
[*]*Corresponding author: Fika.sigalinggingu@student.president.ac.id*

Abstract— Indonesia is a country that has several means of transportation, one of which is a train. People tend to use the train because it is cheaper than some other means of public transportation. Train passenger data has the same pattern every year, which is always increasing and there is a surge in passengers in June and December. Therefore, it is very important to know the projection of train passengers for the purpose of planning and managing facilities and infrastructure as well as train fares. This work forecast the train passenger by using ARIMA model. ARIMA (2,0,0) is the most suitable model for use with a MSE is 1058.39, RMSE is 32.53, MAE is 1042.05, and MAPE is 31.4%. The forecasting process shows that passengers will have an upward trend pattern for several months starting in January 2019. The estimated train passenger data will be useful for planning train fares and improving facilities and infrastructure in the future.

Keywords— ARIMA; Forecasting; Train; Passenger.

I. INTRODUCTION

Indonesia is a developing country that has various types of transportation such as trains. A train is a means of transportation in the form of a vehicle with motion power, either running alone or in combination with other vehicles, which will or are moving on the rails.[1] Train is a means of mass transportation which generally consists of locomotives (vehicles with self-propelled power) and a series of trains or carriages (coupled with other vehicles).

Many researchers used ARIMA model to forecast the number of train passenger. In [2], Dewi and Darsyah forecast the number of train passengers using the Moving Average and Holt Winter methods. From the test results, the best forecasting value is using the Holt Winter method. With the combination α= 0.4; β= 0.25; γ= 0.15 produces MAPE and MAD values of 4 and 1382, respectively. In [3], Panjaitan *et al*. forecast the number of train passengers using the ARIMA, Intervention, and ARFIMA methods. Based on the analysis of the three methods, the best method of analyzing the number of local economy class train passengers in DAOP IV Semarang is the ARFIMA method with the model is ARFIMA (0; 0,367546; [1,13]). Utomo forecast the number of train passengers in Indonesia using the seasonal autoregressive integrated moving average method [5]. The SARIMA method and produces the best model, namely (1; 1; 2) (0; 1; 1)12, from this model it is obtained that the prediction of the total number of train passengers in Indonesia in 2020 is 492,230,700 passengers with an MSE value of 0, 046875 and the MAPE value of 6.26%.

In this work, number of train passenger in Sumatera will be forecasted using ARIMA model with data train passenger PT KAI from 2016 until 2018 [6]. With the estimated number of train visitors, it is hoped that it can help the government in developing transportation facilities, especially trains so that visitors can enjoy the trip more. In addition, it is hoped that visitors will increase again after the end of the Covid-19 pandemic.

II. LITERATURE REVIEW

In this section, the basic of concept ARIMA model is discussed. All subsection of this part refers to [6].

A. Time Series

A time series is a collection of observational data $\{x_t\}$ from the shared distribution of random variables, recorded at a certain time t. In discrete, a time series is one in which the set T_0 of the time at which the observations are made is a discrete set. Continuously, time series is obtained when observations are recorded continuously over several time intervals.

The purpose of time series analysis is generally twofold: to understand or model the stochastic mechanisms that give rise to the observed series and to predict or forecast the future value of the series based on the history of that series or other related factors.

B. Stochastic Processes

Random variable sequence {Yt: t = 0, ±1, ±2, ±3,…} is referred to as process stochastic and serves as a model for the observed time series. It is known that the complete probabilistic structure of such a process is determined by the distribution set of all finite collections of Y [6]. A stochastic process is a model for an observed time series of random variables $\{x_t\}$.

1) Mean

$$\mu_t = E(Y_t) \text{ for } t = 0, \pm1, \pm2, \ldots \quad (1)$$

2) Variance

$$var(Y_t) = t\sigma_e^2 \text{ for } t = 0, \pm1, \pm2, \ldots \quad (2)$$

3) Autocovariance

$$\gamma_k = cov(Y_t, Y_{t-k}) \text{ for k= } 0, \pm1, \pm2, \ldots \quad (3)$$

4) Autocorrelation

$$\rho_k = corr(Y_t, Y_{t-k}) = \frac{\gamma_k}{\gamma_0} \quad (4)$$

C. Stationarity

The basic idea of stationarity is that the probability laws governing the behavior of processes do not change over time. In a sense, the process is in statistical equilibrium. Specifically, a process $\{Y_t\}$ is said to be completely stationary if the shared division $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ equals the shared division $Y_{t_1-k}, Y_{t_2-k}, \ldots, Y_{t_n-k}$ for all time point options $t_1, t_2, \ldots, t_n$ and all k lagging time options. A stochastic process {Yt} is said to be weak (or second-order) stationary if

1. The average function is constant over time
2. $\gamma_{t,t-k} = \gamma_{0,k}$ for all time *t* and lag *k*

D. Autocovariance Function, Autocorrelation Function (ACF), and Partial Autocorrelation Function (PACF)

1) Autocovariance Function

$$\gamma_k = Cov(x_t, x_s) for\ k = 0, \pm1, \pm2, \pm3, \ldots \quad (5)$$

where,

$$Cov(x_t, x_s) = E(X - \mu_t)(X_t - \mu_s) = E[x_t, x_{t-s}] - \mu_t\mu_s \quad (6)$$

2) Autocorrelation Function (ACF)

$$\rho_k = Corr(x_t, x_s) \text{ for } t, s = 0, \pm1, \pm2, \pm3, \ldots \quad (7)$$

where,

$$Corr(x_t, x_s) = \frac{Cov(x_t, x_s)}{\sqrt{Var(x_t)\,Var(x_s)}} \quad (8)$$

3) Partial Autocorrelation Function (PACF)

$$\Phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \Phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \Phi_{k-1,j}\rho_j} \quad (9)$$

III. Analysis And Discussion

A. Data Preparation

The data that I used is the number of train passengers in Sumatra in 2016-2018 [6].

TABLE 1
NUMBER OF TRAIN PASSENGERS IN SUMATRA

Year	Month	Passenger
2016	January	472
2016	February	453
2016	March	461
2016	April	434
2016	May	527
2016	June	429
2016	July	615
2016	August	463
2016	September	497
2016	October	498
2016	November	512
2016	December	620
2017	January	590
2017	February	505
2017	March	558
2017	April	568
2017	May	588
2017	June	542
2017	July	641
2017	August	536
2017	September	577
2017	October	572
2017	November	563
2017	December	667
2018	January	610
2018	February	557
2018	March	603
2018	April	619
2018	May	605
2018	June	760
2018	July	711
2018	August	630
2018	September	626
2018	October	634
2018	November	661
2018	December	768

It can be seen in table 3.1 there is a trend that the data increase over time. in the number of passengers visits every June and December from 2016-2018, this shows that this data is the Seasonal Autoregressive Integrated Moving Average (ARIMA) Model.

B. Stationary Check

Our data is said to be stationary if and only if the p value is less than 0.05. We can check the stationarity of the data by using the Augmented Dickey-Fuller test in R. Based on the results of the Augmented Dickey-Fuller test, the p-value of the data is smaller than 0.05, which is 0.01, which means that our data is stationary. Therefore, we

do not need to do differencing. The following is a plot of Time Series, Autocorrelation Function, and Partial Autocorrelation Function from Railway Passenger data in Sumatra from 2016 to 2018.

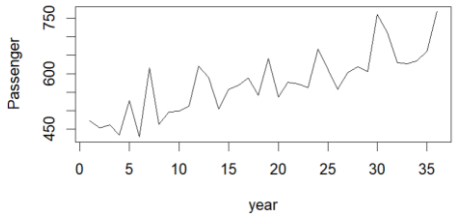

Figure 1. Plot time series of train passanger in Sumatera from 2016 to 2018

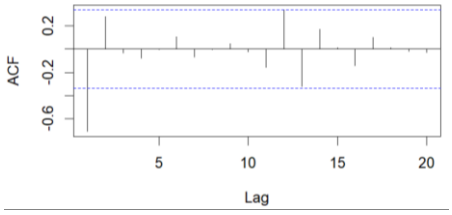

Figure 2. ACF plot

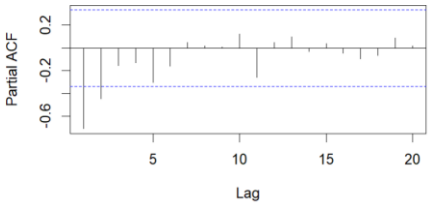

Figure 3. PACP plot

C. Model Spesification

Based on ACF and PACF plot at Figure 3, it is obtained that q=1 and p=2. Since there is no differencing process on the data, hence d=0. Thus, there are 6 model specifications for the ARIMA model as shown in Table 2.

TABLE 2
SPECIFICATIONS OF TRAIN PASSENGER DATA MODEL

Model	p	d	q
ARIMA(2,0,1)	2	0	1
ARIMA(1,0,1)	1	0	1
ARIMA(0,0,1)	0	0	1
ARIMA(2,0,0)	2	0	0
ARIMA(1,0,0)	1	0	0
ARIMA(0,0,0)	0	0	0

D. Parameter Estimation

We can determine the estimated coefficients consisting of AR1, AR2, MA1, Log-likelihood, and AIC which will be considered for further forecasting.

TABLE 3
PARAMETER ESTIMATION OF TRAIN PASSENGER DATA

Model	Cofficient of Estimation Result								
	ϕ_1(AR1)	ϕ_2(AR2)	θ_1(MA1)	Loglikehood	MSE	RMSE	MAE	MAPE	AIC
ARIMA(2,0,1)	0,8575	0,1167	-0,6284	-198,77	1020,21	31,94	1003,73	31,8%	405,54
ARIMA(1,0,1)	0,9771		-0,6747	-198,91	964,37	31,05	948,70	30,3%	403,82
ARIMA(0,0,1)			0,3716	-206,47	2690,71	51,87	2662,41	50,1%	416,95
ARIMA(2,0,0)	0,3937	0,4172		-200,2	1058,39	32,53	1042,05	31,4%	406,4
ARIMA(1,0,0)	0,9771		-0,6747	-198.91	1193,55	34,55	1176,34	30,5%	403,82

ARIMA(0,0,0)	-210,09	3596,24	59,97	42748,04	59,9%	422,17

E. Parameter Estimation

Here, we can determine the best model from our data by using Shaphiro test and Ljung-box test. A model is declared the best if it has a p value for each test of more than 0.05.

TABLE 4
RESIDUAL ANALYSIS OF TRAIN PASSENGER DATA

Model	Shapiro test	Ljung Test	AIC
ARIMA(2,0,1)	0,00618	0,6651	405,54
ARIMA(1,0,1)	0,01057	0,3913	403,82
ARIMA(0,0,1)	0,7764	0,3805	416,95
ARIMA(2,0,0)	0,3805	0,2101	406,4
ARIMA(1,0,0)	0,01057	0,3913	403,82
ARIMA(0,0,0)	0,5124	0,00125	422,17

Based on table 3.4, we can see that models 1, 2, 5 and 6 are declared not fulfilling because the p-value is not above 0.5. While models 3 and 4 are declared to be satisfactory because the p-vale is above 0.5. After testing the six models, it can be determined that model 4 is the best model for forecasting because it passes the Shapiro test, Ljung-Box test, and also has the smallest AIC value that meets the requirements. Model 4 is a model that will have a forecast value that is closest to the actual data. The mathematical equation of ARIMA (2,0,0) model can be expressed by

$$y_t = 0{,}3937Y_{t-1} + 0{,}4172Y_{t-2} + e_t \tag{10}$$

F. Forecasting

Forecasting the number of Railway Passengers in Sumatra for 5 months starting from January 2019 - May 2019 using the ARIMA(2,0,0) model.

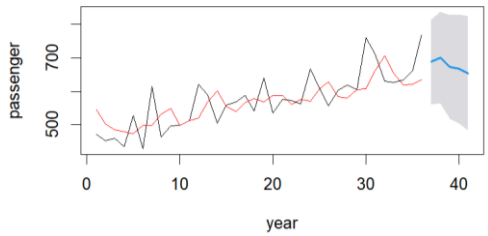

Figure 4. Forecasting of train passenger for next 5 month

TABLE 5
FORECASTING OF TRAIN PASSENGER DATA

Year	Month	Forecasting Point	Lower Limit	Upper Limit
2019	January	687.9363	560,8574	815,0151
2019	February	701,0549	564,4827	837,6271
2019	March	672,8182	518,0962	827,5401
2019	April	667,1746	504,7282	829,6210
2019	May	653,1729	483,2586	823,0873

Comparison between forecasting data and actual data is given by Table 6. Even though the results of the forecast data and actual data are not close together, all the actual data are within the forecast interval.

TABLE 6
COMPARISON OF FORECASTING POINT AND ACTUAL DATA

Year	Month	Forecasting Point	Actual Data
2019	January	687.9363	687
2019	February	701,0549	617
2019	March	672,8182	683
2019	April	667,1746	703
2019	May	653,1729	588

IV. CONCLUSION

Forecasting with the Arima model is good because the values between forecasts and actual data are not much different for the case of data with seasonal increases. There are several steps to create an ARIMA model in R. The first step is data preparation, where we prepare all the data to be analyzed. The next step is to check the stationarity of our data using the ADF test.

The data is said to be stationary if the p value is less than 0.05. If our data is not stationary, then we have to differencing the data until it shows a p-value less than 0.05. How much we do differencing is symbolized by d. After the p-value is below 0.05, the data can be said to be stationary.

The next step, the Model Specification uses the Auto Correlation Function (ACF) to determine q, and the Partial Auto Correlation Function (PACF) to determine p. The next stage is Parameter Estimation, at this stage we can use Moment Estimator Method, Least Square Estimator, and Maximum Probability Estimator.

The next step is Residue Analysis. In this step, I use Shapiro's test and Ljung's test to determine whether our ARIMA model is good or not. After we get the best ARIMA model, the last step is Forecasting. In this step, we can get the prediction number of passenger for the future of our data.

Based on the analysis results, the seasonal ARIMA model (1,0.0) is the best model to predict the number of train passengers in the next 5 months because it has the smallest AIC value, meets the L-Jung Box test and the Normality test.

REFERENCES

[1] Wikipedia, "Kereta," accessed on: April 6, 2022, Available:https://id.wikipedia.org/wiki/Kereta
[2] L.F. Dewi and M.Y. Darsyah, "Peramalan jumlah penumpang kereta api menggunakan metode moving average dan holt winter," in *Prosiding Seminar Nasional Mahasiswa Unimus*, vol. 1, 2018.
[3] H. Panjaitan, A. Prahutama, and S. Sudarno, "Peramalan jumlah penumpang kereta api menggunakan metode arima, intervensi dan arfima (Studi Kasus : Penumpang Kereta Api Kelas Lokal Ekonomi DAOP IV Semarang)," *Journalo of Gaussian*, vol. 7, no. 1, pp. 96-109, 2018, DOI: https://doi.org/10.14710/j.gauss.v7i1.26639
[4] P. Utomo, "Peramalan jumlah penumpang kereta api di Indonesia menggunakan metode seasonal autoregressive integrated moving average (SARIMA)," *Journal of Algebra Mathematics*, vol. 1, no. 2, 2020.
[5] Badan Pusat Statistik, "Jumlah Penumpang Kereta Api," Available: https://www.bps.go.id/indicator/17/72/7/jumlah-penumpang-kereta-api.html
[6] J.D. Cryer and K.S. Chan, *Time Series Analysis With Application in R Second Edition*, 2nd ed. Lowa City: Springer, 2008.