Forecasting the Number of Goods Transported in Java Using ARIMA Models

Chelisca Natasha[1*], Fauziah Nur Fahirah Suding[2]

1,2 Study Program of Actuarial Science, School of Business, President University, 17550, Indonesia
**Corresponding author: chelisca.k@student.president.ac.id*

***Abstract*—** As the most populated island in Java, the island needs to have very efficient out-of-town land transportation. The train is really important transportation in Java. To avoid an unexpected surge in delivering the goods, proper forecasting is required. ARIMA (Autoregressive Integrated Moving Average) model is a method that can be used to predict the number of goods transported in the future. In this analysis, ARIMA (1,1,0) is the best model to use because it has the smallest MAPE among the other model which is 66.6%. The objective of this analysis is to predict the number of goods so the train company can anticipate surges in delivering the goods and may be useful in handling the number of goods in the future by making efficient policies.

***Keywords*—**ARIMA, Trains, Forecasting

I. INTRODUCTION

Freight trains or baggage trains are trains used to transport goods. Goods that can be transported by freight trains are very diverse, such as fertilizers, mining products, containers, oil or liquid commodities, and even livestock can be transported by freight trains [1]. History records that freight transport plays an important role in railways in Indonesia. The growth in the number of motorized vehicles and regional developments are very influential in decreasing the speed of delivery of goods and the high number of accidents and traffic, making the delivery of goods less efficient. Because of this, trains are much more efficient in distributing goods for both medium and long distances for out-of-town deliveries compared to other land vehicles [2]. Freight train is also one of the transportation that plays an important role in the mobility of the population. According to Badan Pusat Statistika Indonesia [4], the demand for freight train services tend to increase every year, this shows a very large public interest in the use of freight trains. However, there are unstable increases and decreases at certain times, so it is necessary to adjust the number of goods each time. This of course requires that it can reduce the risk.

Many previous studies used ARIMA to estimate freight demand. In [6], Su and Su forecast railway, BRT, and bus system demand in Instanbul. According to analysis, ARIMA showed moderate prediction for their data set, and for railway itself, it had 2.55% yearly prediction error. Zhao *et al.* [7] predicted railway freight volume of Ningxia in 2016. Based on test result, the construction of ARIMA (2,2,2) model gave a pretty good fitting precision, with forecasting volume of 5681.457 million tons in 2016, an increase of 0.89% over 2015. Furthermore, they stated the accuracy of ARIMA model will reduce if the predicting period was extended. Thus for future research, they suggested to improve the parameters so that the accuracy for long term forecasting can be improved as well.Since previous researches above succeed in freight prediction, writers motivated to study freight trains demand as well using ARIMA with the data used in the analysis were obtained from Badan Pusat Statistik Indonesia. The purpose of forecasting the number of goods by train in Java is to estimate the number of goods or the surge in goods in every situation. So that this forecasting can help PT Kereta Api Indonesia to prepare and take effective and efficient policies thus later it can anticipate unexpected things in the future.

II. LITERATURE REVIEW

A. Introduction to Time Series Analysis

Time Series data is a collection of data observations on an object $\{X_t\}$ that occurs sequentially in time t, where $t = 1,2,3,$ ...In analyzing the Time Series, the Autoregressive Integrated Moving Average (ARIMA) method can be used to observe accurate short-term data [5]. Based on the name, this method is a combination of regressive and moving average models, wherein making predictions, this method uses independent variable data based on past and present values.

B. Stochastic Process

A stochastic process is a sequence of random variables $\{Y_t = 0, \pm1, \pm2, \pm3, \dots\}$ which presents the observed time series model. The models are mean, variance, autocovariance, and autocorrelation.

1) The Mean is defined by

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

μt is the expected value at process time t. In general, t can be different at any time *t.*

2) Function Variance is

$$-[Var\ (Y_t) = t\sigma_e^2 \dots \quad (2)$$

3) The Autocovariance defined by

$$\gamma_{t,s} = Cov\ (Y_t, Y_s) \qquad \text{for } t, s = 0, \pm1, \pm2, \ldots \tag{3}$$

where,

$$Cov\ (Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t, Y_s) - \mu_t \mu_s \tag{4}$$

4) The *Autocorrelation* is

$$\rho_{t,s} = Corr\ (Y_t, Y_s) \qquad \text{for } t, s = 0, \pm1, \pm2, \ldots \tag{5}$$

where,

$$Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var\ (Y_t) Var(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} \tag{6}$$

C. Stationary

In the analysis of time series, stationary is an important concept in observing data. Stationary is some simple assumptions to make statistical conclusions about the stochastic process based on the obtained data. Time series can be said to be stationary if the mean and variance are constant.

1) The Strictly Stationary

Stochastic process $\{Y_t\}$ can be said to be strong if the combined distribution of $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ is the same as the combined distribution of $Y_{t_1-k}, Y_{t_2-k}, \ldots, Y_{t_n-k}$ for all time points $t_1, t_2, \ldots, t_n$ and for all times k.

For n =1, the univariate distribution of Y_t is equal to Y_{t-k} for all *t* and *k*. In other words, *Y* is an identical distribution. Then, $\mathrm{E(Y_t)} = \mathrm{E(Y_{t-k})}$ and $\mathrm{Var(Y_t)} = \mathrm{Var(Y_{t-k})}$ for all *t* and *k.* resulting in *mean* and variance from time to time.

For *n*= 2, Y_t and Y_s must equal $Y_{t-k}$and Y_{s-k}. So,

$$Cov(Y_t, Y_s) = Cov(Y_{t-k}, Y_{s-k}) \text{ for all } t, s, \text{ and } k \tag{7}$$

k = *s* and *k* = *t*, get

$$\begin{aligned} \gamma_{t,s} &= \mathrm{Cov(Y_{t-s}, Y_0)} \\ &= \mathrm{Cov\ (Y_0, Y_{s-t})} \\ &= \mathrm{Cov(Y_0, Y_{|t-s|})} \\ &= \gamma_{0,|t-s|} \end{aligned} \tag{8}$$

From The notation can be concluded that $\gamma_k = Cov(Y_t, Y_{t-k})$ and $\rho_k = Corr(Y_t, Y_{t-k})$

`

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{9}$$

2) Weakly Stationary

Stochastic process $\{Y_t\}$ can be said to be weak or second-order stationary if the mean function is constant over time and if $\gamma_{t,t-k} = \gamma_{0,k}$ for all time t and lag k.

D. Transformation of Time Series Analysis

To use the ARIMA model, stationary time series data is required because the model can only be used on a stationary time series, therefore data transformation is required. The transformation that will be used is to perform differencing, which is to calculate the change or difference in the value of the observation. There are 2 methods for doing Forecasting, namely

1) Non-Stationary Time Serie\\

$$\mathrm{e_t = X_t - X_{t-1}} \tag{10}$$

2) Stationary Time Series

$$\mathrm{W_t = X_t - X_{t-1}} \tag{11}$$

The general form of the backward Shift Operator,

$$BX_t = X_{t-1} \tag{12}$$

Like polynomials, other real *B* and *W* are manipulated in the same way.

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

1) Autocovariance

Function Autocovariance function is the covariance of the variable itself at a certain time. Autocovariance is defined as,

$$\gamma_k = Cov\ (X_t, X_s) \text{ for } k = 0, \pm1, \pm2, \pm3, \ldots \quad (13)$$

Where,

$$Cov(X_t, X_s) = E[(X_t - \mu_t)(X_t - \mu_s)] = E\left[X_t, X_{t,s}\right] - \mu_t\mu_t \quad (14)$$

2) Autocorrelation Function (ACF)

Autocorrelation is a correlation that is arranged based on the time sequence between the data before and the data after it. Autocorrelation is denoted by,

$$\rho_{t,s} = Corr(Y_t, Y_s) \text{ for } t, s = 0, \pm1, \pm2, \ldots \quad (15)$$

Where,

$$Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_S)}{\sqrt{Var\ (Y_t) Var(Y_S)}} = \frac{\gamma_k}{\gamma_0} \quad (16)$$

Following are some important properties of autocovariance and autocorrelation functions:

$$\gamma_{t,t} = Var(Y_t) \qquad ,\rho_{t,t} = 1 \quad (17)$$

$$\gamma_{t,s} = \gamma_{s,t} \qquad ,\rho_{t,s} = \rho_{s,t} \quad (18)$$

$$\left|\gamma_{t,s}\right| \leq \sqrt{\gamma_{t.t}\gamma_{s,s}} \qquad ,\left|\rho_{t,s}\right| \leq 1 \quad (19)$$

3) Partial Autocorrelation Function (PACF)

The partial Autocorrelation Function is a function used to measure the correlation between data at k that have elapsed and current observations [5]. Partial autocorrelation is formulated as follows

$$\phi_{kk} = Corr(X_t, X_{t-k} | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}) \quad (20)$$

III. ANALYSIS AND DISCUSSION

A. Data Preparation

The data used in this analysis is the number of goods via rail transportation in Java from January 2016 to December 2018.

TABLE 1
MONTHLY GOODS TRANSPORTED 2016-2018

Year	Month	Number of goods	Year	Month	Number of goods
2016	January	927	2017	July	1081
2016	February	734	2017	August	1176
2016	March	785	2017	September	1083
2016	April	967	2017	October	1197
2016	May	873	2017	November	1143
2016	June	945	2017	December	1110
2016	July	766	2018	January	1227
2016	August	1019	2018	February	1021
2016	September	936	2018	March	1205
2016	October	975	2018	April	1193
2016	November	973	2018	May	1338
2016	December	991	2018	June	846
2017	January	974	2018	July	1357
2017	February	861	2018	August	1323
2017	March	966	2018	September	1330
2017	April	967	2018	October	1391
2017	May	1101	2018	November	1284
2017	June	781	2018	December	1300

B. Stationary Check

The next step is to check whether the data is stationary or not. To check whether the data is stationary, we can use the Augmented Dickey-Fuller (ADF) Test on R. Stationary data can be seen from its p-value. If the p-value of the data is less than 0.05, then the data can be said to be stationary.

Based on the data obtained using the ADF test, the p-value of the data is 0.2476, which means the data is not stationary because it is greater than 0.05. Therefore, it is necessary to do differencing until the data is stationary. After doing the differencing, we can see that the p-value of the data becomes 0.01, which means that the data is stationary because the p-value is already below 0.05. The following is a time series plot, autocorrelation function, and partial autocorrelation function for the number of goods transported in java for 2016-2018.

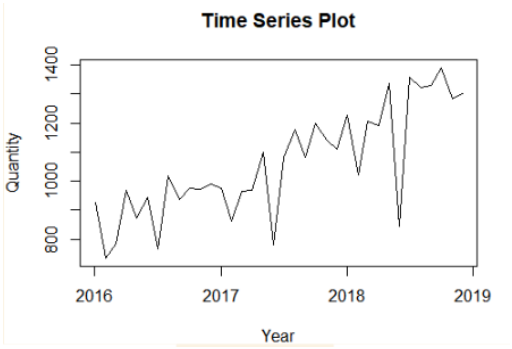

Figure 1 Time Series Plot

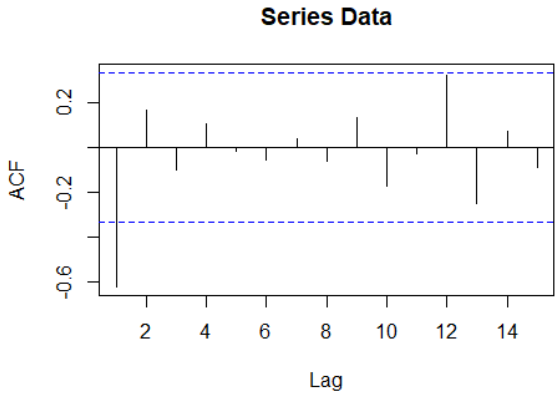

Figure 2 Autocorrelation Function

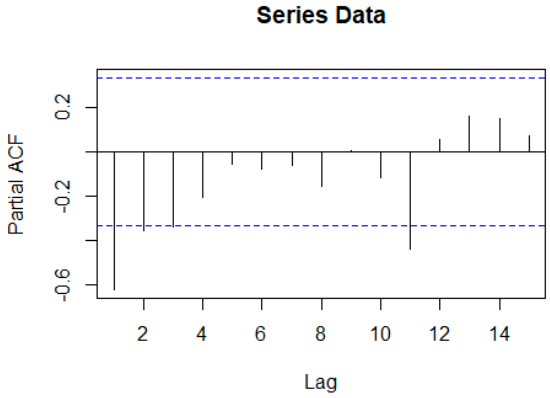

Figure 3 Partial Autocorrelation Function

Based on the results of differencing, autocorrelation functions, and our partial autocorrelation function gets a *p* of 3, *d* which has a value of 1, and *q* which has a value of 1.

C. Model Specification

Based on the ACF and PACF data, there are 4 models for the ARIMA model with the same d value, namely 1. Following are the specifications of the models,

TABLE 2
MODEL SPECIFICATION

ARIMA Model	p	d	q
ARIMA (0,1,0)	0	1	0
ARIMA (1,1,0)	1	1	0
ARIMA (2,1,0)	2	1	0
ARIMA (3,1 ,0)	3	1	0
ARIMA (0,1,1)	0	1	1
ARIMA (1,1,1)	1	1	1
ARIMA (2,1,1)	2	1	1
ARIMA (3,1,1)	3	1	1

D. Parameter Estimation

The following are parameter estimates that can be determined after knowing the ARIMA model. The following parameter estimation coefficients consist of AR1, AR2, AR3, MA1, and Log-likelihood which will later be considered in forecasting.

TABLE 3

PARAMETER ESTIMATION

Model	Coefficient Estimation Result				
	AR1	AR2	AR3	MA1	Log-Likelihood
ARIMA (0,1,0)					-230.75
ARIMA (1,1,0)	-0.6167				-222.38
ARIMA (2,1,0)	-0.8375	-0.3444			-220.16
ARIMA (3,1.0)	-0.9264	-0.5753	-0.2709	-0.5580	-218.85
ARIMA (0.1,1)				-0.699	-220.71
ARIMA (1,1.1)	-0.3497			-0.5580	-219.27
ARIMA (2, 1,1)	-0.5703	-0.1376		-0.4617	-219.11
ARIMA (3,1.1)	-0.756	-0.4358	-0.2171	-0.1868	-218.76

E. Residual Analysis

In this residual analysis, 2 types of tests were performed, namely, the Shapiro test, and the Ljung to determine which is the best model that can be used for forecasting. The model that is passed is only if the value of the p-value is greater than 0.05.

TABLE 4
RESIDUAL ANALYSIS

Model	Shapiro Test	Ljung Box Test	Description	AIC
ARIMA (0,1,0)	0.2476	0.0001097	Rejected	461.51
ARIMA (1,1,0)	0.2595	0.154	Accepted	446.76
ARIMA (2,1,0)	0.07362	0.3934	Accepted	444.33
ARIMA (3,1.0)	0.1712	0.4422	Accepted	443.70
ARIMA (0.1,1)	0.02209	0.03444	Rejected	443.42
ARIMA (1,1.1)	0.08938	0.4187	Accepted	442.55
ARIMA (2,1.1)	0.07735	0.4645	Accepted	444.23
ARIMA (3,1.1)	0.1671	0.4631	Accepted	445.51

Based on table 3.4, it can be seen that from eight models there are only six models passed the Shapiro test and Ljung Box test, namely the ARIMA model (1,1,0), ARIMA (2,1,0), ARIMA (3,1,0), ARIMA (1,1,1), ARIMA (2,1,1), and ARIMA (3,1,1) models. However, because the difference in AIC values is not too big between all the models that pass, it is necessary to check all forecasting points in order to get more efficient results.

TABLE 5
BEST MODEL EVALUATION

Actual Data	Forecasting Points					
	ARIMA (1,1,0)	ARIMA (2,1, 0)	ARIMA (3,1,0)	ARIMA (1,1,1)	ARIMA (2,1,1)	ARIMA (3,1,1)
1243	1290,133	1323,452	1330,204	1310,115	1322,689	1329.22
975	1296,218	1298,301	1322,006	1306,578	1309,817	1323,389
1169	1292,466	1311,287	1307,891	1301,815	1312,749	1311,588
1158	1294,779	1309.073	1317,501	1307.382	1331,141	1316,707
1203	1293,353	1306.355	1318,939	1307,534	1312,553	1319,246

After all the estimated points are compared, it can be seen that the model that is closest to the actual data is the ARIMA model (1,1,0) compared to the other models. With the Formula:

$$y_t = -0.6167\, Y_{t-1} + e_t \qquad (21)$$

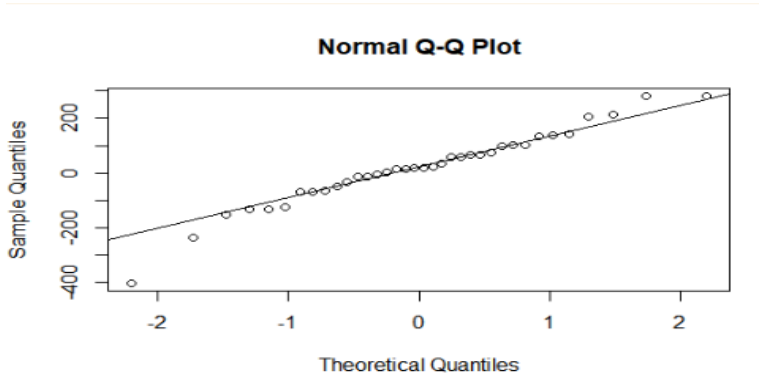

Figure 4 QQ Plot

F. Forecasting

The following is a graph forecasting the number of goods via rail transportation in Java with a confidence interval of 99. The forecasting graph below starts from January 2019 until May 2019 with the ARIMA model (1,1,0).

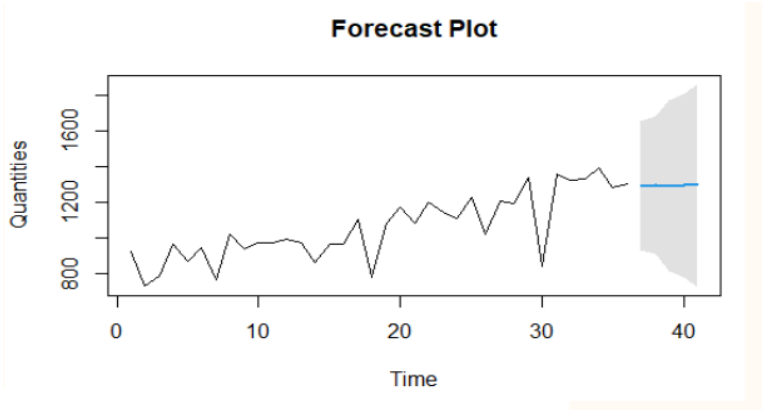

Figure 5 Forecasting Graph

Based on the *forecasting results*, the following are the point forecast, lower limit, and upper limit for forecasting for the next 5 months based on the data.

TABLE 6
FORECASTING RESULT

Year	Month	Estimated Point	Lower	Limit Upper Limit
2019	January	1290,133	929,2140	1651,053
2019	February	1296,218	909,6881	1682,747
2019	March	1292,466	817,7435	1767,188
2019	April	1294,779	783,0863	1806.473
2019	May	1293,353	726,8233	1859,882

G. Comparison

The following table is the forecasting points and the actual data from January 2019 to May 2019. Although the forecast point is not too close to the actual data, this forecasting can be said successful because the actual data is still within the estimation interval.

TABLE 7
ERROR ANALYSIS

Year	Month	Estimated Point ($\hat{y}$)	Actual Data (y)	$\lvert\hat{y} - y\rvert$	$(\hat{y} - y)^2$	$\frac{(\hat{y} - y)^2}{y}$
2019	January	1290.133	1243	47.13	2221.52	1.72
2019	February	1296.218	975	321.22	103181.00	79.60
2019	March	1292,466	1169	123.47	15243.85	11.79
2019	April	1294,779	1158	136.78	18708.49	14.4563
2019	May	90,353,381	1203	1203.66	1293,353	6.31

With that information from table 7, we can get MSE for 12,293.21, RMSE 110.87, MAE 12,233.3, and MAPE 66.6%.

IV. CONCLUSION

Based on the results of forecasting conducted on data on the number of goods via rail transportation on the island of Java in 2016-2018, the best model used to predict the number of goods in the next 5 months with MSE values 12,293.21, RMSE 110.87, MAE 12,233.3, and MAPE values 66.6% are ARIMA models (1,1,0). Based on the residual analysis, the ARIMA formula (1,1,0) is:

$$y_t = -0.6167\, Y_{t-1} + e_t \tag{22}$$

It is proved from the results of the forecasting point which is not far from the actual data. Although the forecasting results are not the same, the actual data are still within the estimation interval. It is important for PT. Kereta Api Indonesia to prepare for the number of goods that will be transported in the future so they can make efficient policies.

REFERENCES

[1] "Kereta Api Barang," June 19, 2019. Available: https://id.wikipedia.org/wiki/Kereta_api_barang.

[2] PT. Kereta Api Indonesia, "Tentang Kami", 2020. Available: https://cargo.kai.id/site/about.

[3] A. Rufaidah and M.A. Iffindi, "Analisis Peramalan ARIMA BOX-JENKINS pada pengiriman barang," *Nusantara Journal of Computers and Its Applications*, vol. 2, no. 1, 2017.

[4] Badan Pusat Statistik, "Jumlah Barang Melalui Transportasi Kereta Api Menurut Pulau 2022." Available: https://www.bps.go.id/indicator/17/73/1/jumlah-barang-melalui-transportasi-kereta-api-menurut-pulau.html.

[5] J. D. Cryer and K.S. Chan, *Time Series Analysis with Application in R*, 2nd ed., Springer, 2008.

[6] E. Su and O.A. Su, "Public Transport Demand Forecast Using Arima: The Case of Instanbul." Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=402064

[7] J. Zhao, J. Cai, and W. Zheng, "Research on Railway Freight Volume Prediction Based on Arima Model" in *18th COTA International Conference of Transportation Professionals*, Beijing, China, July 5-8, 2018. DOI: https://doi.org/10.1061/9780784481523.043