Forecasting amount of passanger of ships in Madura strait port using ARIMA Box-Jenkins method

M Prastuti¹, B S S Ulama¹ and P P Nurindraprasta¹

¹Department of Business Statistics, Institute of Technology Sepuluh Nopember, Surabaya, Indonesia

E-mail: mike_p@statistika.its.ac.id

Abstract. Madura Island and the city of Surabaya, used to be only connected by Ujung-Kamal Port. Later in 2009, Suramadu Bridge began operating. Some people would rather drive over the Suramadu bridge than using the ships so that the amount of ships passanger decreases. This has made PT. ASDP Indonesia Ferry, as the port manager, lose some of the port’s revenue that leads into loss. Along with the free access to pass the Suramadu bridge, it is feared that the amount of ships passanger will continue to decrease. Therefore PT. ASDP Indonesia Ferry requires effort to prevent the greater loss. One of the efforts is to reschedule the ship trips. The ship trips can be affected by the amount of ships passanger. Therefore this research aims to forecast the amount of ships passanger, includes pedestrian, motorbike rider, and car rider. The forecasting is done by the ARIMA Box-Jenkins method. The result shows that ARIMA ([1,12,13],1,0), ARIMA (0,1,[1,11,12]), and ARIMA (0,1,1) are the best model to forecast the amount of padestrian, motorbike rider, and car rider. The forecast results show that the amount of padestrian passenger is continue to decrease, while the amount of motorbike and car rider are constant.

1. Introduction
Madura strait port is connecting Madura Island and Java Island, precisely in Bangkalan district and the city of Surabaya. This port has been operating since 1949. This port was also the only access to cross the Madura strait until in mid-2009, the Suramadu bridge began operating. Suramadu bridge operation has an impact on the reduced number of ship passangers. This condition occurs because some people would rather drive over the Suramadu bridge than using the ships to cross the Madura strait.

The reduced number of ships passangers affects various parties, including traders and public transportation driver who operating around the port area. They experience a decrease in income which then makes some traders and public transportation drivers choose to switch jobs. Not only affected the traders and public transportation drivers, the decline in ship passanger also affected PT. ASDP Indonesia Ferry (Persero) Surabaya as the port manager. This condition caused PT. ASDP Indonesia Ferry (Persero) Surabaya to suffer losses each year.

In order to prevent greater loss, PT. ASDP Indonesia Ferry (Persero) Surabaya requires several precautions. Factors that can determine the number of ship trips include the number of ship passangers that includes pedestrian, motorbike rider, and car rider. Therefore, this research aims to forecast the number of ship passangers in 2019. The forecasting is done by ARIMA Box-Jenkins method.
2. Literature review

2.1. Time series analysis
Time series analysis is qualitative analysis method that can be use to forecast the values of data in the future. Purpose of time series analysis is to determine patterns in time series data and explore the data to forecast the future values. Generally, the purpose of time series analysis is to forecast the future conditions, to know the connections of data in a period with the data afterwards, and to control a process [1].

2.2. ARIMA Box-Jenkins
ARIMA Box-Jenkins is one of time series analysis. There are several steps in this method, which are as follows.

2.2.1. Identifications of data patterns.
The data used to forecast must be stationary data. Stationary data is the data that does not experience growth or decline, in other words the data has a horizontal pattern along the time axis. The stationarity of the data can be determined through a time series plot or autocorrelation function (ACF) plot. As for the data that is not stationary, differencing is performed using the following equation [1].

\[ Z'_t = Z_t - Z_{t-1} \]  

(1)

2.2.2. Identifications of ARIMA model.
The ARIMA model is determined by the pattern of autocorrelation function (ACF) plot and partial autocorrelation function (PACF) plot. By paying attention to the pattern of ACF and PACF plot, ARIMA models that can be determine are explain in the following table [2].

| Model          | ACF                  | PACF                  |
|----------------|----------------------|-----------------------|
| AR (p)         | Dies down            | Cut off after lag p   |
| MA (q)         | Cut off after lag q  | Dies down             |
| AR (p) or MA (q) | Cut off after lag q | Cut off after lag p   |
| ARMA (p,q)     | Dies down            | Dies down             |

1. AR Model represent the value of the data at t period of time is affected by as many as p data values in the previous time period. The equation of AR model is shown in the below.

\[ Z_t = \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \alpha_t \]  

(2)

2. MA Model represent the value of the data at t period of time is affected by as many as q residuals in the previous time period. The equation of AR model is shown in the below.

\[ Z_t = \alpha_t - \theta_1 \alpha_{t-1} - \ldots - \theta_q \alpha_{t-q} \]  

(3)

3. ARMA model is a combination of both the ar and ma models. The equation of ARMA model is shown below.

\[ Z_t = \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \alpha_t - \theta_1 \alpha_{t-1} - \ldots - \theta_q \alpha_{t-q} \]  

(4)

4. ARIMA model is the result of the expansion of ARMA model. ARIMA model also pay attention to as much as d times of differencing. The equation of ARIMA model is shown below.

\[ \phi(B)(1-B)^d Z_t = \theta_0 + \theta_q (B)\alpha_t \]  

(5)

5. Seasonal ARIMA model is the another form of arima model that can be used on seasonal data. The equation of seasonal ARIMA model is shown below.

\[ \Phi_p (B^s) (1 - B^s)^d Z_t = \Theta_q (B^s) \alpha_t \]  

(6)

2.2.3. Parameter estimation.
The next step of ARIMA Box-Jenkins method is to estimate the values of each parameter. There are several methods that can be used to estimate the parameter value such as least square (conditional least square) method, maximum likelihood method, unconditional least square method, and non linear estimation method. Conditional least square method is used in this study. This method estimating the parameter by minimizing in the sum square error of the of the model[3]. Then the results of estimation are being tested to decide that the values are significant or not.

Diagnostic check. Model with significant parameter need to go through diagnostic check. The purpose of doing diagnostic check is to make sure that the model is valid. In diagnostic check, residuals have to fulfill two assumption, include assumption of white noise and assumption of normal distribution. Assumption of white noise can be checked by Ljung-Box test using hypothesis that the residuals has white noise[2]. Assumption of normal distribution can be checked by Kolmogorov-Smirnov test using the hypothesis that distribution of residuals are normal[4].

2.2.4. Best model selection.
If there are two or more ARIMA model that have significant parameter and already fulfill the assumption of white noise and assumption of normal distribution, then we should select the best model. The best model selection is based on the value of RMSE and MAPE. Model with the smallest value of RMSE and MAPE is the best model. The equation of RMSE and MAPE is shown below.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Z_t - \hat{Z}_t)^2}
\]

\[
MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{Z_t - \hat{Z}_t}{Z_t} \right| \times 100\%
\]

2.2.5. Forecasting.
After finding the best model, the final step is to forecast the value of data in the future. The best ARIMA model is being used to forecast the value of data in the future. For an example, the equation of AR (1) is shown below.

\[
\hat{Z}_t(k) = \mu + \phi(k-1) - \mu
\]

3. Methodology
In this study, we use 3 variables that represent the different types of ship passengers. Those 3 variables are pedestrians, motorbike rider and car rider. We are using data that is start since January 2018 to December 2018. The data is given by PT. ASDP Indonesia Ferry (Persero) Surabaya. Analysis steps in this study include the following.

1. Describe and identify food pattern data.
2. Divide data into samples in and out of the sample.
3. Check the stationarity of the data.
4. Identify the arima model that might appear.
5. Estimate the parameter values.
6. Check the assumption of white noise and normal distribution.
7. Choose the best arima model.
8. Estimating the value of data in the future.
9. Interpret the results.
10. Draw conclusions.

4. Analysis

4.1. Characteristics of the amount of ship passengers
The amount of ship passengers at the Madura Strait port is calculated by the amount of tickets sold. For example, if there is four people in a car. The car driver should buy the car rider ticket, while the other people in the car should buy pedestrian ticket.

![Figure 1. Time series plot of the amount of pedestrian passanger.](image1)

![Figure 2. Time series plot of the amount of motorbike rider passanger.](image2)

![Figure 3. Time series plot of the amount of car rider passanger.](image3)

Figures above show that there is decreasing in the amount of ship passengers. The decrease in the amount of pedestrian passangers in 2018 reached 72.8% compared to 2011. While the decrease of motorbike riders and car riders reached 30.6% and 60.1% in 2018 compared to 2011.

4.2. Forecasting the amount of pedestrian passengers

The time series plot of passengers in Figure 1 shows a pattern of trends in the amount of pedestrian passengers. It is also suspected that there is a seasonal pattern. This pattern indicates that the data is not stationary and needs to be differencing. The differencing in the amount of pedestrian passengers is done twice. The first difference in is used to remove the effect of the trend pattern and the second difference in is used to remove the effect of seasonal pattern. Therefore, the ACF and PACF plot are made based on the results of differencing.
Based on ACF and PACF plot for differencing 1, there are thirty two models that might appear. While based on ACF and PACF plot for differencing 1 and 12, there are two models that might appear. the total number of possible models for pedestrians is 34 models.

The result of parameter estimation for all of the 34 models shows that there are only 9 model left with all the parameter in the model is significant. Then after diagnostic check, there is only 3 model that fulfill the assumption of white noise and the assumption of normal distribution. Those 3 model are ARIMA ([1,12,13],1,0), ARIMA (0,1,2)(0,1,0)\(^{12}\), and ARIMA (2,1,0)(0,1,0)\(^{12}\). Since there are 3 ARIMA model that have significant parameter and also fulfill the assumption of white noise and assumption of normal distribution, one of 3 model should be select to forecast the amount of pedestrian. The value of RMSE and MAPE for each model is show in the following table.

Table 2. Best pedestrian model selection.

| Model                  | RMSE   | MAPE   |
|------------------------|--------|--------|
| ARIMA ([1,12,13],1,0)  | 8.01452| 5.73%  |
| ARIMA (0,1,2)(0,1,0)\(^{12}\) | 10.67533| 9.93%  |
| ARIMA (2,1,0)(0,1,0)\(^{12}\) | 11.15422| 10.82% |

Based on the value of RMSE and MAPE in table above, ARIMA ([1,12,13],1,0) have the smallest RMSE and MAPE value. This means that ARIMA ([1,12,13],1,0) is the best model to forecast the
amount of pedestrian passengers. The equation of ARIMA ([1,12,13],1,0) model is shown below and the forecast result of ARIMA ([1,12,13],1,0) for pedestrian is shown in the Figure 8 below.

\[
(1 - \phi_1 B - \phi_2 B^{12} - \phi_3 B^{13}) (1 - B) Z_t = \alpha_t,
\]

\[
(1 - \phi_1 B - \phi_2 B^{12} - \phi_3 B^{13} - B + \phi_1 B^2 + \phi_2 B^{12} + \phi_3 B^{13}) Z_t = \alpha_t,
\]

\[
Z_t = 0.226Z_{t-1} - 0.774Z_{t-2} + 0.4512Z_{t-12} - 0.1087Z_{t-13} - 0.3425Z_{t-14} + \alpha_t,
\]

Figure 8. Forecast result of the amount of pedestrian passanger.

Figure 8 shows that forecast result for the amount of pedestrian. The red line shows that the forecast result are fluctuating. The red line also shows that forecast result still decreasing. The lower limit and the upper limit are shown by green line and blue line.

4.3. Forecasting the amount of motorbike rider passengers

The time series plot of passengers in Figure 2 shows a pattern of trends in the amount of motorbike rider passengers. It is also suspected that there is a seasonal pattern. This pattern indicates that the data is not stationary and needs to be differencing. The differencing in the amount of motorbike rider passengers is done twice. The first differencing is used to remove the effect of the trend pattern and the second differencing is used to remove the effect of seasonal pattern. Therefore, the ACF and PACF plot are made based on the results of differencing.

Figure 9. ACF plot of the amount of motorbike rider passanger after differencing 1.

Figure 10. PACF plot of the amount of motorbike rider passanger after differencing 1.
Based on ACF and PACF plot for differencing 1, there are fourteen models that might appear. While based on ACF and PACF plot for differencing 1 and 12, there are four models that might appear. The total number of possible models for motorbike rider is 18 models.

The result of parameter estimation for all of the 18 models shows that there are only 12 models left with all the parameter in the model is significant. Then after diagnostic check, there is only 4 model that fulfill the assumption of white noise and the assumption of normal distribution. Those 4 model are ARIMA (0,1,1)(1,1,12), ARIMA (0,1,1)(1,1,0)\textsuperscript{12}, ARIMA (1,1,0)(1,1,0)\textsuperscript{12}, and ARIMA (1,1,3)(1,1)(0,1,0)\textsuperscript{12}.

Since there are 4 ARIMA models that have significant parameter and also fulfill the assumption of white noise and assumption of normal distribution, one of 4 models should be select to forecast the amount of motorbike rider. The value of RMSE and MAPE for each model is show in the following table.

**Table 3. Best motorbike rider model selection.**

| Model                          | RMSE   | MAPE  |
|--------------------------------|--------|-------|
| ARIMA ([1,12,13],1,0)          | 5.13088| 4.71% |
| ARIMA (0,1,1)(1,1,0)\textsuperscript{12} | 14.05883| 14.77%|
| ARIMA (1,1,0)(1,1,0)\textsuperscript{12} | 9.11422| 9.14% |
| ARIMA ([1,13],1,0)(0,1,0)\textsuperscript{12} | 14.54731| 15.36%|

Based on the value of RMSE and MAPE in table above, ARIMA ([1,12,13],1,0) have the smallest RMSE and MAPE value. This means that ARIMA ([1,12,13],1,0) is the best model to forecast the amount of motorbike rider passangers. The equation of ARIMA ([1,12,13],1,0) model is shown below and the forecast result of ARIMA ([1,12,13],1,0) for motorbike rider is shown in the Figure 13.

\[
(1 - B)Z_t = \left[ 1 - \theta_1 B - \theta_1 B^{11} - \theta_2 B^{12} \right] a_t \\
Z_{t-1} = a_t - \theta_1 a_{t-1} - \theta_1 a_{t-11} - \theta_2 a_{t-11} \\
Z_t = Z_{t-1} - \theta_1 a_{t-1} - \theta_1 a_{t-11} - \theta_2 a_{t-11} + a_t \\
Z_t = Z_{t-1} - 0.7187 a_{t-1} - 0.6047 a_{t-11} - 0.6206 a_{t-12} + a_t
\]
Figure 13. Forecast result of the amount of motorbike rider passenger.

Figure 13 shows that forecast result for the amount of motorbike rider passenger. The red line shows that the forecast result are constant and don’t have any trend. The lower limit and the upper limit are shown by green and blue line.

4.4. Forecasting the amount of car rider passengers

The time series plot of passengers in Figure 3 shows a pattern of trends in the amount of car rider passengers. This pattern indicates that the data is not stationary and needs to be differencing. The differencing in the amount of passengers is done once. The differencing in is used to remove the effect of the trend pattern. Therefore, the ACF and PACF plot are made based on the results of differencing.

Figure 14. ACF plot of the amount of car rider passengers after differencing 1.

Based on ACF and PACF plot for differencing 1, there are two models that might appear. The result of parameter estimation for the 2 models shows that both model have significant parameter. Then after diagnostic check, all of the models shows the result that fulfill the assumption of white noise and the assumption of normal distribution. Those 2 model are ARIMA (1,1,0) and ARIMA (0,1,1). Since there are 2 ARIMA model that have significant parameter and also fulfill the assumption of white noise and assumption of normal distribution, one of 2 model should be select to forecast the amount of car rider. The value of RMSE and MAPE for each model is show in the following Table 4.
Table 4. Best car rider model selection.

| Model          | RMSE   | MAPE  |
|---------------|--------|-------|
| ARIMA (1,1,0) | 1.397,5| 17.78%|
| ARIMA (0,1,1) | 1.377,7| 17.48%|

Based on the value of RMSE and MAPE in table above, ARIMA (0,1,1) have the smallest RMSE and MAPE value. This means that ARIMA (0,1,1) is the best model to forecast the amount of car rider passangers. The equation of the ARIMA (0,1,1) is shown below and the forecast result of ARIMA (0,1,1) for car rider is shown in the Figure 16.

\[(1-B)Z_t = (1- \theta_i B)a_t\]

\[Z_t - Z_{t-1} = a_t - \theta_i a_{t-1}\]

\[Z_t = Z_{t-1} - \theta_i a_{t-1} + a_t\]

\[Z_t = Z_{t-1} - 0.24395 a_{t-1} + a_t\]

Figure 16 shows that forecast result for the amount of car rider passangers. The red line shows that the forecast result are fluctuating. The red line also shows that forecast result still decreasing. The lower limit and the upper limit are shown by green and blue line.

5. Conclusion
The results of the forecast made, indicate that in 2019 the amount of pedestrians will decrease compared to the previous year, where the lowest occurred in November. While the amount of motorbike rider and car rider do not experience an increase or decrease. Based on the forecast results obtained, PT. ASDP Indonesia Ferry (Persero) Surabaya can reduce the amount of ship trips by adding a lag time to the ship.

6. References
[1] Makridakis S Wheelwright S C and McGee V E 1999 Metode dan Aplikasi Peramalan (Jakarta: Binarupa Aksara)
[2] Wei W 2006 Time Series Analysis, Univariate and Multivariate Methods (Boston: Peaseson Addison Wesley)
[3] Cryer J D and Chan K S 2008 Time Series Analysis with Application in R (New York: Springer)
[4] Daniel W W 1989 Statistika Nonparametrik Terapan (Jakarta: Gramedia)