**Journal of Actuarial, Finance and Risk Management (JAFRM)**
Vol. 1(2), November, 2022.

# Forecasting The Number of Aircraft Passengers Arriving Through Soekarno-Hatta Airport Using Arima Model

Windi Marnizal Putri[1*], Fauziah Nur Fahirah Suding[2]

[1,2] *Study Program of Actuarial Science, School of Business, President University, 17550, Indonesia*
[*]*windi.putri@student.president.ac.id*

***Abstract***— Soekarno-Hatta International Airport is well known as the busiest airport in Indonesia with the number of airplane passengers normally grow from year to year. In 2010, there were more than 43 million passengers, and had increased up to 62.4 million over the year 2011. Risk of overcapacity became an issue. Thus, in the following year, the airport was planned for an expansion. Predicting the frequency of passengers can be helpful for future planning and to improve airport facilities and policy. This research used Autoregressive Integrated Moving Average (ARIMA) to forecast the number of aircraft passengers. ARIMA (0,1,1) is the most suitable model used with MAPE 110%, the results is 2,405,205 passengers. Actual data and predictive data are not much different.

***Keywords***— ARIMA; Forecasting; MAPE; Soekarno-Hatta Airport

## I. INTRODUCTION

In Indonesia, there are many types of airlines and airports that fly aircraft, one of which is Soekarno-Hatta International Airport. It was established in 1985 to replace Kemayoran International Airport. Then in 1991, it was expanded because it replaced the overcapacity Halim Perdana Kusuma International Airport and became the main airport that served the greater area of Jakarta. [1]

Passengers traffic growth has been dramatically increase. In 2010, The airport reported that there were around 43,7 millions passengers and 338,711 aircraft movement. By 2011, air passengers reached 62.4 millions. [2] The overcapacity became large issue, consequently expansion of the airport was planned to overcome the crowded since it affected the service and time management. Predicting the frequency of people take flight could also be one solution since the results can be used for future planning in improving airport capacity and facilities. [3]

Many studies have been carried out in forecasting the number of passengers. Fahik and Jatipaningrum (2021) compare two methods, Holt-Winters Exponential Smoothing and Seasonal ARIMA in predicting total passengers in Soekarno Hatta International Airport. Holt-Winters Exponential Smoothing produced smaller MAPE value (4.407%) than Seasonal ARIMA (5.306%). Predicting result for March 2020, is estimated to 732,129 people. [4]

Another study from Hasanah (2019), she also analyzed number of airport passengers in Juanda International Airport. By comparing three methods, ARIMA, Regression Time Series and TBATS, resulting ARIMA to be best model. Kuswanto et all.,(2014), also estimated the frequency of passengers who travel through Juanda International Airport using ARIMA and Transfer Function. In this study, the writers also get ARIMA as the best method. [5]

From previous studies above, researchers were encouraging to use ARIMA to predict the frequency of aircraft passengers through Soekarno-Hatta International Airport in 2012-2016. Research result were expected to show that simple model can also produce good result. And as to be information for government and PT. Persero, to overcome the crowded such that it can improve facilities at the airport that will increase the value of Soekarno-Hatta airport.

## II. LITERATURE REVIEW

### A. *Introduction to Time Series Analysis*

A time series is a collection of observed data from a distribution with random variables that are recorded within a certain time but continuously, a time series is obtained. When observed continuously for some time interval. What distinguishes the time series is that the analysis can show how the variables change from time to time, or it can be called a variable which is an important factor because it will show a picture of the data that corresponds to the data points. [6]

### B. *Stochastic*

Process A stochastic processor also called a random process is a random variable that is indexed to several sets in mathematics. This process is the process of observing the times and determining the results where these results are called random variables. Time series of random variables is part of the stochastic process. Where to get the value of the book of Jesus using the formula, as follows:

1) *Mean*

$$\mu_t = E(X_t) \quad (1)$$

2) *Variance*

$$Var(X_t) = t\sigma_e^2 \quad (2)$$

3) *Autocovariance*

$$\gamma_k = Cov(X_t, X_s) \quad (3)$$

4) *Autocorrelation*

$$\rho_k = Corr(X_t, X_s) \quad (4)$$

*C. Stationary*

Stationary is an important factor that must be present in time series analysis. Stationary can be known in several ways such as from graphs, correlograms, or through Root. Stationary itself means it is a time series of power that has a tendency to move on the average (mean). Stationary data usually crosses the horizontal axis when pulled, or its autocorrelation will decrease over a long period of time. [7]

*D. Transformation of Time Series Analysis*

In the transformation of time series analysis, there are differences in the forecasting of the two methods in the time series. This difference is a linear operation of the conditional expectation. There are two forecasting methods considered, namely:

1) Non-stationary Time Series
This time series can be used with the formula:

$$e_t = X_t - X_{t-1} \quad (5)$$

2) Stationary Time Series
This Time Series can be used with the formula:

$$W_t = X_t - X_{t-1} \quad (6)$$

Or in its general form using the formula:

$$BX_t = X_{t-1} \quad (7)$$

*E. Formulas in Time Series Analysis*

As previously mentioned, Time series has several important formulas that must be included in time series analysis, namely Autocovariance, Autocorrelation (ACF) functions, Partial Autocorrelation Functions (PACF). These three types of correlation are usually used in model specifications

1) Autocovariance Function
The definition of the Autocovariance function is [8]:

$$\rho_{xx}(t_1, t_2) = \frac{K_{xx}(t_1,t_2)}{\sigma_{t1}.\sigma_{t2}} = \frac{E[(X_{t1}-\mu_{t1})]}{\sigma_{t1}.\sigma_{t2}} \quad (8)$$

And can also be denoted by:

$$\gamma_k = Cov(X_t, X_s) \text{ for } k = 0, \pm1, \pm2, \pm3, \ldots \quad (9)$$

The set of values $\gamma_{k},_{k} = 0,1,2,\ldots\ or\ \gamma_0 = \sigma_y^2$ is a function symbol of the Autocovarian Function, where [9]:

$$Covv(X_t, X_s) \quad = E(X_t - \mu_t)(X_t - \mu_s) = E(X_t, X_{t-s}) - \mu_t.\mu_s$$

2) Autocerrelation Function (ACF)

Autocorrelation function is usually used to find stationarity of analysis time series data. In general, the function of Autocorrelation is to define how similar the data is to the time version itself. The random process X(t) is also called $E[X^2(t)]t < \infty$ for each $t \in T$ if it is second order. The autocorrelation functions of X(t) and X(s) are denoted by $R_{xx}(t, s)$ and are also denoted by:

$$R_{xx}(t,s) = E[X(t)X(s)] = E[X(s)X(t)] = R_{xx}(s,t)$$
$$R_{xx}(t,t) = E[X^2(t)] \quad (9)$$

Or you can also use the formula:

$$\rho_k = Corr(X_t, X_s) \; for \; k = 0, \pm 1, \pm 2, \pm 3, \ldots \quad (10)$$

Where:

$$Corr(X_t, X_s) = \frac{Cov(X_t, X_s)}{\sqrt{Var(X_t)Var(X_s)}} = \frac{\gamma_k}{\gamma_0} \quad (11)$$

*3) Partial Autocorrelation Function*

Partial Autocorrelation Function is the correlation between two variables to calculate the value of a set of variables.

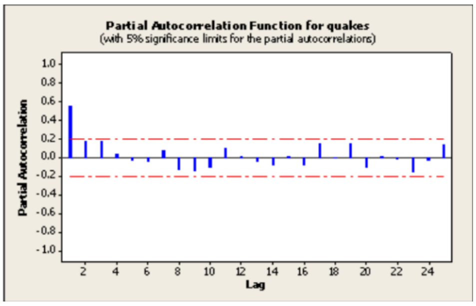


Figure. 1 Particial Autocorrelation Function

Partial Autocorrelation Function is denoted by [10] :

$$\frac{Cov(y, x3|x1, x2)}{\sqrt{Var(y|x1, x2)Var(x3|x1, x2)}} \quad (12)$$

$$y = \beta_0 + \beta_1 x^2 \; and \; y = \beta_0 + \beta_1 + \beta_2 x^2 \quad (13)$$

Or

$$\boldsymbol{\phi_{kk} = Corr(X_t, X_{t-k}|X_{t-k}, X_{t-2}, \ldots, X_{t-k+1}} \quad (16)$$

## III. Analysis and Results

### *A. Data Preparation*

The data used in this paper is data on the number of passengers arriving through Soekarno-Hatta International Airport from January 2012 to December 2016.

TABLE 1
THE NUMBER OF PASSENGERS ARRIVING BY SOEKARNO-HATTA INTERNATIONAL AIRPORT

| mm/yyyy | Data | mm/yyyy | Data |
|---|---|---|---|
| Jan-12 | 2314964 | Jul-14 | 2092201 |
| Feb-12 | 2127022 | Aug-14 | 2810608 |
| Mar-12 | 2362070 | Sept-14 | 2296555 |
| Apr-12 | 2341738 | Pkt-14 | 2382573 |
| May-12 | 2409964 | Nov-14 | 2306529 |
| Jun-12 | 1978000 | Des-14 | 2492669 |
| Jul-12 | 2384206 | Jan-15 | 2196439 |
| Aug-12 | 2472378 | Feb-15 | 1959646 |
| Sep-12 | 2421433 | Mar-15 | 2103921 |
| Okt-12 | 2405838 | Apr-15 | 2104874 |

| | | | |
|---|---|---|---|
| Nov-12 | 2346410 | May-15 | 2293224 |
| Des-12 | 2483418 | Jun-15 | 2171041 |
| Jan-13 | 2374483 | Jul-15 | 2550866 |
| Feb-13 | 2112179 | Aug-15 | 2339590 |
| Mar-13 | 2481657 | Sep-15 | 2067674 |
| Apr-13 | 2406460 | Okt-15 | 2240216 |
| May-13 | 2545296 | Nov-15 | 2250089 |
| Jun-13 | 2764786 | Des-15 | 2478472 |
| Jul-13 | 2394358 | Jan-16 | 2304570 |
| Aug-13 | 2772952 | Feb-16 | 2072698 |
| Sep-13 | 2449839 | Mar-16 | 2278964 |
| Okt-13 | 2546954 | Apr-16 | 2275596 |
| Nov-13 | 2450975 | May-16 | 2581790 |
| Des-13 | 2543140 | Jun-16 | 2126278 |
| Jan-14 | 2481027 | Jul-16 | 2856824 |
| Feb-14 | 1986114 | Aug-16 | 2466109 |
| Mar-14 | 2190455 | Sep-16 | 2318798 |
| Apr-14 | 2141703 | Okt-16 | 2331367 |
| May-14 | 2386400 | Nov-16 | 2293547 |
| Jun-14 | 2507504 | Des-16 | 2654623 |

Source : bps.go.id

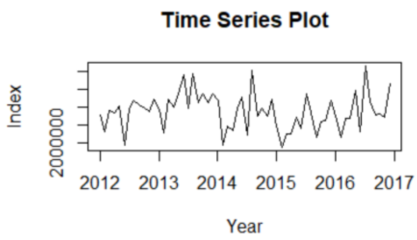


Figure. 2 Training Data

*B. Stationary*

Our data is said to be stationary if the p value is less than 0.05. We can check the stationarity of the data by using the ADF test in R. According to the results, the p-value is 0.1599 which is greater than 0.05, thus data is not stationary. Therefore, we have to differencing the data until it is < 0.05 or becomes stationary. After differencing, the p-value of the data finally became smaller than 0.05, which means that the data we have is stationary. The following is a plot of Time Series, Autocorrelation Function, and Partial Autocorrelation Function from data on the number of passengers arriving through Soekarno-Hatta International Airport in 2012-2016.

*C. AR(p) and MA(q)*

Orders of AR and MA can be determine using PACF for Order p and ACF for Order q.

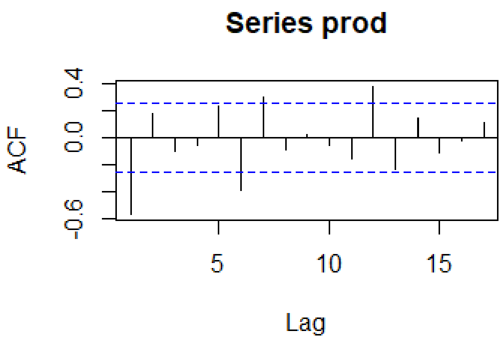


Figure. 3 ACF

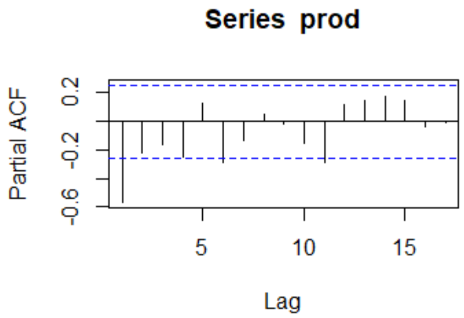


Figure. 4 Particial ACF (PACF)

It can be seen that the cut off for PACF and ACF are both at lag time 1. Thus, the AR order, namely p, is 1 and the MA order, q, is 1.

*D. Model Specification*

Here are the model specifications for the ARIMA model, so there will be 7 models with d equal to 1.

TABLE 2
ARIMA MODEL SPECIFICATION

| **Model ARIMA** | **P** | **d** | **q** |
|---|---|---|---|
| ARIMA(1,1,0) | 1 | 1 | 0 |
| ARIMA(0,1,0) | 0 | 1 | 0 |
| ARIMA(1,1,1) | 1 | 1 | 1 |
| ARIMA(0,1,1) | 0 | 1 | 1 |

*E. Parameter Estimation*

The following are the parameter estimates for all models. We can determine it after we know the ARIMA model and we can determine the estimated coefficients consisting of AR1, AR2, AR3, Log likelihood, MSE, RMSE, MAE, MAPE, and AIC which will be considered for forecasting later.

TABLE 3
ESTIMATION MODEL ARIMA DARI AR1,MAI,LOG LIKELIHOOD, MSE, RMSE, MAE, MAPE, AND AIC

| Model ARIMA | Coefficient Estimation Result | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AR1 | MA1 | Log Likeli-hood | MSE | RMSE | MAE | MAPE | AIC |
| ARIMA (1,1,0) | -0.5797 | | -810.29 | 66766845170 | 258392.8118 | 66766654436 | 89% | 1622.58 |
| ARIMA (0,1,0) | | | -822.11 | 65191765260 | 255326.7813 | 65191588517 | 86% | 1644.21 |
| ARIMA (1,1,1) | -0.0661 | -0.8063 | -805.94 | 84079744816 | 289965.0752 | 84079507306 | 108% | 1615.87 |
| ARIMA (0,1,1) | | -0.8421 | -806 | 87123131570 | 295166.2778 | 87122888062 | 110% | 1614 |

*F. Residual Analysis*

From the residual analysis, the p-value must be more than 0.05 to get the best model for our data. To determine the best model, we can use the Saphiro test and Ljung's test. The model that passes the test will be the best model.

TABLE 4
Residuals Analysis

| Model ARIMA | Shapiro Test | Ljung Test | AIC |
|---|---|---|---|
| ARIMA(1,1,0) | 0.3765 | 0.2736 | 1622.58 |
| ARIMA(0,1,0) | 0.4188 | 6.497e-06 | 1644.21 |
| ARIMA(1,1,1) | 0.8213 | 0.9603 | 1615.87 |
| ARIMA(0,1,1) | 0.7587 | 0.7101 | 1614 |

*G. Best Model Evauation*

After testing the model through the Shapiro and Ljung-Box tests, from the forecasting results we can see which model is close to the actual data, and in this case the model that is suitable for forecasting is Model 4, namely ARIMA(0,1,1).

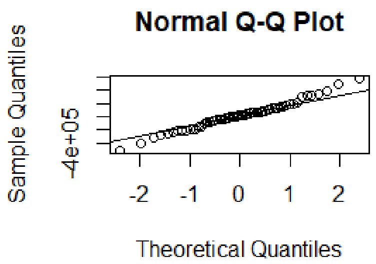


Figure. 5 Model 4 Residual Plot

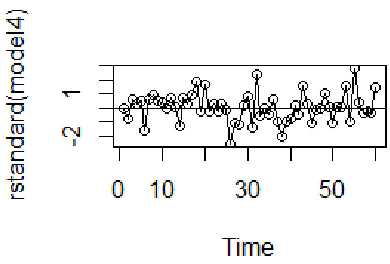


Figure 6 Model 4 Standardised Residual Plot

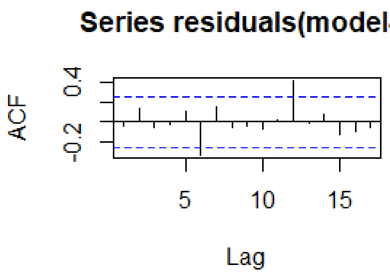


Figure 7 Model 4 ACF Residual

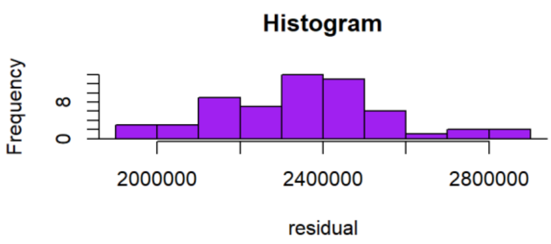


Figure 8 Model 4 Histogram

*H. Forecasting*

The following are the parameter estimates for all models. We can determine it after we know the ARIMA model and we can determine the estimated coefficients consisting of AR1, AR2, MA1, Log likelihood, and AIC which will be considered for forecasting later.

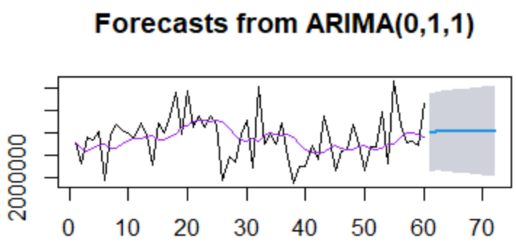


Figure 9 Model 4 Forecast

*I. Comparison Between Actual and Forecast Data*

The following is a comparison table for forecast data training with test data (12 final actual data).

TABLE 5
TABLE OF ACTUAL AND FORECAST DATA

| Year | Mount | Point Forecast | Lower Boundary | Upper Boundary | Actual Data |
|---|---|---|---|---|---|
| 2017 | Januari | 2405205 | 1872161 | 2938248 | 2611736 |
| 2017 | Februari | 2405205 | 1865560 | 2944849 | 2160303 |
| 2017 | Maret | 2405205 | 1859039 | 2951370 | 2506223 |
| 2017 | April | 2405205 | 1852595 | 2957814 | 2459888 |
| 2017 | Mei | 2405205 | 1846225 | 2964184 | 2663541 |
| 2017 | Juni | 2405205 | 1839927 | 2970482 | 2201103 |
| 2017 | Juli | 2405205 | 1833699 | 2976710 | 3158315 |
| 2017 | Agustus | 2405205 | 1827537 | 2982872 | 2627000 |
| 2017 | September | 2405205 | 1821441 | 2988968 | 2585256 |
| 2017 | Oktober | 2405205 | 1815408 | 2995001 | 2580263 |
| 2017 | November | 2405205 | 1809435 | 3000974 | 2614908 |
| 2017 | Desember | 2405205 | 1803522 | 3006887 | 2718013 |

## IV. CONCLUSION

The data used in this paper was taken from bps website. After processing and analyzing, the data on the number of passengers who came through Soekarno-Hatta Airport in 2012-2016 used the ARIMA model (0,1,1) which is the best data from the smallest number of AIC after the Shapiro Test and Ljung Boox. Results show that the point forecast during 2017 is 2405205 people, which are close to actual data. Using 90% confidence interval, we can conclude that the actual data values are within estimation interval.

## REFERENCES


[1] Wikipedia, " Bandar Udara Internasional Soekarno–Hatta," *en.wikipedia.org*, 2022 . Accessed on: August 28, 2022, Available: https://id.wikipedia.org/wiki/Bandar_Udara_Internasional_Soekarno%E2%80%93Hatta

[2] Badan Pusat Statistik, " Jumlah Penumpang Pesawat Udara yang Berangkat dan Datang Melalui Pelabuhan Udara Soekarno-Hatta 2018-2021," BPS, Jakarta. Accessed on: August 28, 2022, Available: https://jakarta.bps.go.id/indicator/17/309/1/jumlah-penumpang-pesawat-udara-yang-berangkat-dan-datang-melalui-pelabuhan-udara-soekarno-hatta.html

[3] H. Kuswanti, Suhartono, and A.M. Huda, "Forecasting The Frequency Of Domestic Air Passengers At Juanda Airport Using ARIMA and Transfer Function As A Basis For Future Development Of Airport Scenario," *Jurnal Tata Kota dan Daerah,* vol. 6, No. 1, 2014.

[4] D.S. Fahi,k and M.T Jatipaningrum, "Peramalan Jumlah Penumpang Penerbangan International Di Bandar Udara Soekarno-Hatta dengan Metode Holt-Winter Exponential Smoothing dan Seasonal ARIMA," *Jurnal Statistika dan Industri dan Komputasi*, vol. 6, No.1, 2021, pp.77-87

[5] S. Hasanah, "Peramalan Jumlah Penumpang di Bandara International Juanda Menggunakan Metode ARIMA Regresi Time Series," *TBATS. JUSTEK : Jurnal Sains dan Teknologi*, Vol 2, No.1 , 2019, pp. 29-37, ISSN 2620-5475, http://journal.ummat.ac.id/index.php/justek

[6] Tableau, " Time Series Analysis: Definition, Types, Techniques, and When It's Used," tableau.com, 2003. Accessed on: August 28, 2022, Available: https://www.tableau.com/learn/articles/time-series-analysis#:~:text=Time%20series%20analysis%20is%20a,data%20points%20intermittently%20or%20randomly.

[7] T. J. Long, "Uji Stasioneritas Data Time Series," *jagostat.com*. Accessed on: August 28, 2022, Available: https://jagostat.com/analisis-time-series/uji-stasioneritas-data-time-series#:~:text=%2DPerron%20Test).,Stasioneritas%20merupakan%20konsep%20penting%20dalam%20analisis%20time%20series.,rata%2Drata%20dan%20variansnya%20konstan.

[8] Wikipedia, "Autocovariance," *en.wikipedia.org*, 2021. ccessed on: August 28, 2022, Available: https://en.wikipedia.org/wiki/Autocovariance#:~:text=In%20probability%20theory%20and%20statistics,of%20the%20process%20in%20question.

[9] Science Direct, " Autocorrelation Function ," *ScienceDirect*, 2001. Accessed on: August 28, 2022, Available: https://www.sciencedirect.com/topics/mathematics/autocorrelation-function#:~:text=Introduction%20to%20Random%20Processes&text=Basically%20the%20autocorrelation%20function%20defines,for%20each%20t%20%E2%88%88%20T.

[10] PennState, " Partial Autocorrelation Function (PACF), " The Pennsylvania State University , 2022. Accessed on: August 28, 2022, Available: https://online.stat.psu.edu/stat510/lesson/2/2.2