OPEN ACCESS
J DATA SCI APPL, VOL. 3, NO. 1, PP. 031-037, JANUARY 2020

E-ISSN 2614-7408
DOI: 10.34818/JDSA.2020.3.45

JOURNAL OF DATA SCIENCE AND ITS APPLICATIONS

Forecasting Number of Passengers of TransJakarta using SARIMAX Method

Maftukhatul Qomariyah Virati[1], Diory Paulus Pamanik[2], Setia Pramana[3*]

[1]BPS' Statistics Indonesia
[2]Jakarta Smart City, Jakarta Indonesia
[3]Politeknik Statistika STIS, Jakarta Indonesia

[*]setia.pramana@stis.ac.id

Received on 16-09-2019, revised on 22-11-2019, accepted on 11-05-2020

Abstract

TransJakarta is one of the most common public transportation modes used by public in Jakarta. Every day, there are more than 300.000 people who uses TransJakarta. Jakarta Smart City collaborating with PT. Transjakarta to integrate mass transportation system in Jakarta. PT. Transjakarta has data on the number of TransJakarta passengers every day that can be used to predict the number of users in the coming week. This information can be used to optimize Transjakarta's bus performance. For example, when to do maintance on Transjakarta's buses, cleaning buses, etc. The results of the study showed that there is a pattern in the number of Transjakarta passengers where the number of users will increase on weekdays and decrease on weekends. To predict that, SARIMA which able to overcome seasonal effects the data had can be used. Then it was also found that the number of users decreased significantly in the Indonesian holiday season, such as on Eid al-Fitr and Indonesian Independence Day. To overcome that pattern, the help of x-factor, where x-factor is a dummy variable of holiday in Indonesia, can be used. Therefore, this study uses the SARIMAX model to predict the number of TransJakarta users in the next 7 days. The model with the best results is the SARIMA(0,0,0)(2,1,0) with x-factor and with an analysis error of MSE = 162402173, MAPE = 2.6122 and MASE = 0.211698.

Keywords: sarimax, time-series, transjakarta, x-factor

I. INTRODUCTION

TransJakarta is a bus rapid transit (BRT) system in Jakarta, Indonesia, which was the first BRT system in Asia. It is now the main public transportation in Jakarta with 155 routes. The TransJakarta system is a shuttle bus that will stop at each designated bus stop. One of advantage of using a TransJakarta bus is that it can reduce congestion in Jakarta. TransJakarta buses have special routes that can only be passed by TransJakarta . The price for the use of TransJakarta is also very cheap, which is 3500 rupiah regardless of the distance traveled. In 2018 about 189.8 million passengers used the service. Currently, it serves more than 700 thousands passengers daily, and it is aimed to serve one million passengers daily.

Jakarta Smart City is part of the DKI Jakarta Provincial Government to implement smart cities in Jakarta. Smart City is a city that uses information and communication technology to help that city to build their competitive advantages [1] [2]. There are six categories targeted by Jakarta Smart City, smart living, smart mobility, smart governance, smart environment, smart economy, and smart people. Smart mobility means the public can have access to diverse modes of transportation, prioritizing environmentally friendly transport and not motorized vehicles, and integrated with information and communication technology.

PT Transjakarta has data on the number of Transjakarta's passengers every day. Before boarding a TransJakarta bus, one must have an e-money card. This card will be tap-in at the entrance gate available at each TransJakarta bus stop. If passengers wants to get out of the bus stop, then the passengers must tap out at the exit gate. These tap-in and tap-out transactions are recorded by PT. Transjakarta.

Jakarta Smart City in collaboration with PT. Transjakarta to help realize Smart City. By using the help of Application Programming Interface (API). Data on the number of users who tap-in and tap-out every day can be monitored by Jakarta Smart City.

PT Transjakarta has a limited number of TransJakarta's busses. But, TransJakarta's passengers always want the best service deliver. There's days when Transjakarta's passengers at the peak, but there's a low season when the bus is almost empty. By understanding the pattern of TransJakarta's passengers tap-in, prediction can be made. With this prediction, we can optimize the service of TransJakarta's buses. The maintenance of Transjakarta's buses can be done, without compromising the passengers.

Several studies regarding the prediction of bus passengers have been carried out [3] [4]. Seeing the data obtained from PT Transjakarta is daily data from tap-in & tap-out by Transjakarta's passengers, predictions can be made using the time-series method. There is a pattern in the tap-in data, which causes predictions cannot be made using only the ARIMA model, but must use the Seasonal ARIMA (SARIMA) model that can overcome the seasonal pattern. [5]. However, simple time series approach is sometimes less accurate because besides the time effect, there are other variables that affect the data movement. This can be overcome by adding x-factors in the form of additional variables. [6] Therefore in this study the SARIMAX model is implemented.

II. RESEARCH METHODS

A. TransJakarta Dataset

Data of this study were obtained from PT. TransJakarta through Jakarta Smart City. Data in the form of the number of tap-ins and tap-outs that occur every day from 5 June until 19 August 2017. There are some data missing in the dataset. Generally, there’s around 350.000 taps-in in weekday, and 250.000 taps-in in weekend. The challenge of this study is the absence of a clear id on the electronic card. Therefore, if an electronic card is used to tap-in twice in a day, morning and evening, it will be considered taps-in twice.

B. Time Series Approach

ARIMA

The simple time series approach is Autoregressive Integrated Moving Average (ARIMA) models. Univariate (single vector) ARIMA is a forecasting technique that projects the future values of a series based entirely on its own inertia [7]. The model is written like this

$$ARIMA(p, d, q),$$

where:

p = order of the auto regressive part,
d = degree of first differening involved,
q = order of the moving average part.

The formula is written like this

$$(1-\ \phi_1\ B-\ldots-\phi p\ B)(1-B)\ yt\ =\ (1-\ \theta_1 B-\cdots-\ \theta qB)\ et_.$$

where

$\phi =$ autoregressive parameter,
$\theta = moving\ average\ parameter,$
$et_{.=}$ the error term at time t,
B = Backward shift operator,

When the data has a seasonal weekly effect so it is suitable to use Seasonal Autoregressive Integrated Moving Average (SARIMA). Trend is the movement of data up or down. In this case the data had fallen during the Lebaran but back up after Eid. Seasonal is a repetitive movement. Reassuring that data does have a weekly seasonal effect. The random is data that can not be explained by Trend and Seasonal effects,

SARIMA

Seasonal Autoregressive Integrated Moving Average (SARIMA), is a method for time series forecasting with univariate data containing trends and seasonality [7].The model is written as follow:

$$ARIMA(\ p\ ,d\ ,q\ )(\ P,D,Q)s,$$

where

P = order of the seasonal auto regressive part,
D = degree of seasonal differening involved,
Q = order of the seasonal moving average part,
s = The number of time steps for a single seasonal period.

The formula is defined as follows:

$$(1-\ \phi 1\ B-\ldots-\phi p\ B)\ (1-\ \Phi\ 1\ Bs-\ldots-\Phi\ P\ Bs)(1-B)\ (1-Bs)\ yt\ =\ (1-\ \theta 1B-\cdots-\ \theta qB)\ (1-\ \ \Theta\ 1Bs-\cdots-\ \ \Theta\ QBs)\ et\ .$$

where

$\Phi =$ seasonal autoregressive parameter,
Θ = seasonal moving average parameter,
Bs = Backward Seasonal shift operator.

The X Factors

X-factors that affect the number of TransJakarta passenger are holiday, and Eid holidays, where the number of TransJakarta users drops dramatically during that period. This can be overcome by adding x-factors in the form of additional variables. Some research [8] takes into account calendar variation effect by the inclusion of x-factor.

The dataset for this study is from 5 June until 19 August 2017. At that date, 25 – 26 June 2017 is Eid al-Fitr which is a big holiday in Indonesia. There is also some National holiday for example 17 August 2017 which is Independence National Day.

In this study the x-factor variable is a dummy data. Selection of data dummy is done by tentative by seeing the actual data movement. When it is a holiday, the dummy is -1, when it is a lebaran the dummy is -4. This is because the decrease of passenger in Lebaran is 4 times greater than holiday

SARIMAX

Time series modeling is sometimes less accurate because there are other variables that affect data movement. This can be overcome by adding x-factors in the form of additional variables. In this study the additional variable is a dummy data to overcame Lebaran and holiday like National Independent Day [9]. Selection of data dummy is done by tentative by seeing the actual data movement. When it is a holiday the dummy is -1, when it is a lebaran the dummy is -4. This is because the decrease of passenger in Lebaran is 4 times greater than holiday. In this case, we know that we need to use SARIMA and X-factor to handle the holiday effect. Hence, the SARIMAX is chosen to forecast the number of TransJakarta Passenger.

Interpolation

Some data on the TransJakarta dataset is missing. It is known that this data was lost due to PT. TransJakarta upgrades and repairs the system. However, the time series analysis can not be performed if there is missing data. It can be overcome by performing imputation of the data based on interpolation. Interpolation is the filling of lost data with a particular method.

There is a lot of method to interpolate time series data [10]. The interpolation in this study is implemented with function na.interp of R package Forecast. For seasonal series, a robust Seasonal and Trend decomposition using Loess decomposition is first computed. Then a linear interpolation is applied to the seasonally adjusted data, and the seasonal component is added back [11].

Heteroscedascity Test

The variance of the errors should be consistent for all observations, known as homoscedasticity, is one of the assumptions about residual/error in ordinary least square regression. When this assumption is violated, it is called heteroscedasticity. Heteroscedascity test is needed because SARIMAX cannot handle the high violatility. The heteroscedascity test in this study is implemented with function McLeod.Li.test from R package TSA [5].

Forecast Accuracy

It is important to evaluate forecast accuracy using several forecasts Key Performance Indicators. in this stydu the following measurements are used. **Mean Absolute Percentage Error (MAPE)** is one of the most commonly used to measure forecast accuracy. It Measures by total of absolute error divided by the true value. Actually, it is the average of the percentage errors. **Mean Absolute Error (MAE)** is the mean of the absolute error. **Mean Squared Error (MSE)** is defined as the square root of the average squared error. Mean Absolute Scaled Error (MASE) measures from average scaled error [12] .The best model is selected based on the smallest error.

All the analysis is conducted using the R software [13] [14].

III. RESULTS AND DISCUSSIONS

There are around 350.000 passengers everyday at weekday and 250.000 passengers at weekend during 5 June until 19 August 2017. The tap-in and tap-out pattern is shown in Figure 1. However, there are around 50.000 passengers in difference because the tap out system has only just begun to be implemented in September 2016. We observe a weekly effect as the number of passengers in every weekend is dropping. It shows that people mainly use the TransJakarta to go to work and/or school. The number of passengers also drop at the holiday period. For example, at Indonesian Independence Day, 17 August 2017, the number of passenger is dropping, even though it is fall on Thursday. In Ramadhan period, the effect is much stronger. We observed low number of passenger at 4 days before and after Eid Fitr.

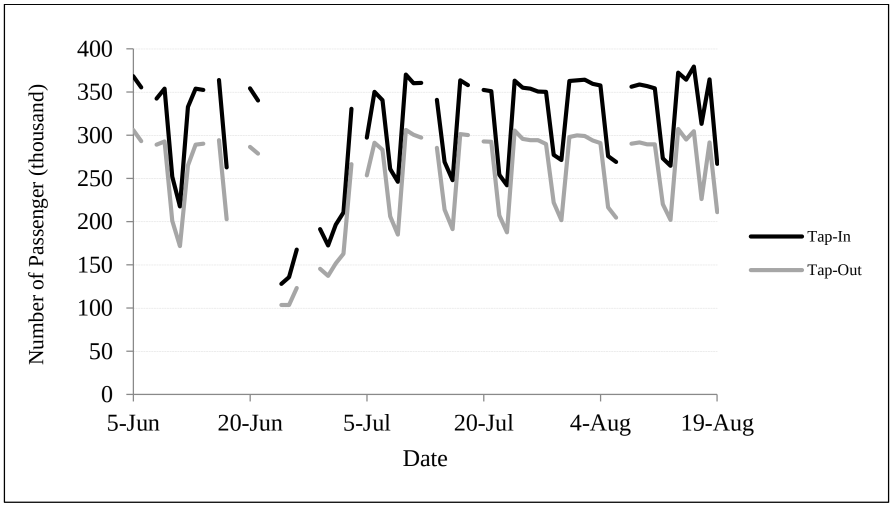

Fig 1. Number of Tap-in & Tap-out of TransJakarta from 5 June – 19 August 2017

The data unfortunately have a lot of missing data. This is due to the fact that TransJakarta is changing the system. A linear interpolation for non-seasonal series and periodic structural decomposition using seasonal series is being implemented to impute the missing data and the result is shown in Figure 2. The result shows that the interpolation resulted in nice data imputation as it follows the pattern of the series.

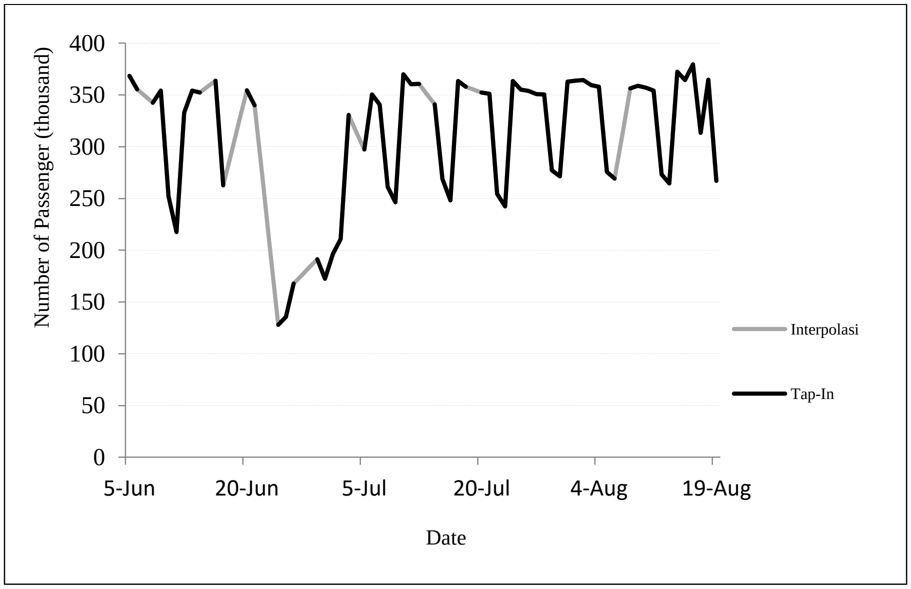

Fig 2. Number of Tap-in & Tap-out of TransJakarta after interpolation from 5 June – 19 August 2017

Using the inputted data, several SARIMAX models are being derived and compared based on the forecast accuracy measurements. The results shown in Table 1 revealed that SARIMA(0,0,0)(2,1,0)[7] gives the smallest MSE, MAE, MAPE and MASE. Hence we can use SARIMA(0,0,0)(2,1,0)[7] as the selected model for this dataset.

TABLE I
Error Measurement

MODEL	MSE	MAE	MAPE	MASE
SARIMA(2,1,1)(0,1,1)[7]	706196269	23905.14	7.14%	0.626538
SARIMA(1,0,0)(2,1,0)[7]	249027679	13061.67	3.90%	0.342338
SARIMA(0,0,0)(2,1,0)[7]	**162402173**	**8077.20**	**2.41%**	**0.211698**

Comparison between forecast values based on model SARIMA(0,0,0)(2,1,0)[7] and the actual data can be seen from Figure 3 and Table 2. We observed that the forecast is fitting nicely with the series pattern.

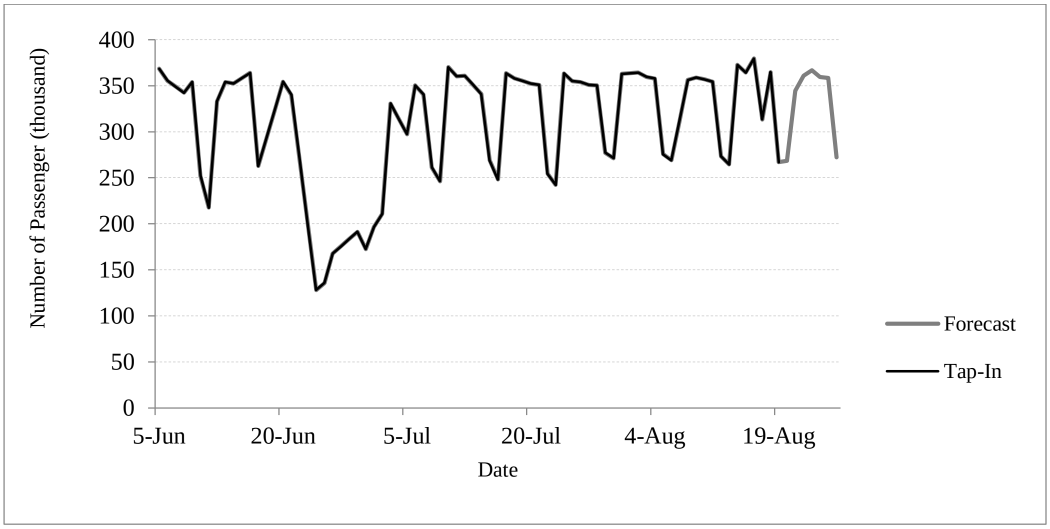

Fig 3. Result of the Forecast

TABLE II
Forecast and Actual Data

Date	Forecast	Actual
20-Aug	268342	248147
21-Aug	344558	371031
22-Aug	360876	361953
23-Aug	366657	364641
24-Aug	359449	359444
25-Aug	358513	361963
26-Aug	272078	275403

Figure 3 shows that the general pattern of TransJakarta's passengers are low in number during weekend and remained high during weekday. However, in some situation such as holiday, the number of passengers is very low. In this case we observed very low numbers during Ramadan and Eid Fitr Period.

After seeing the number of Transjakarta's passengers in the next 7 days, it can be compared between the number of real Transjakarta's passengers and the forecasting number of passengers using existing models. The comparison of this data can be seen in Table 2. From the results of the assessment can be calculated error values, such as MSE, MAE, MAPE, and MASE which can be seen in Table 1. Figure 4, shows the graph comparison of actual data of Transjakarta’s passengers and the forecast of SARIMA (0,0,0) (2,1,0) [7]

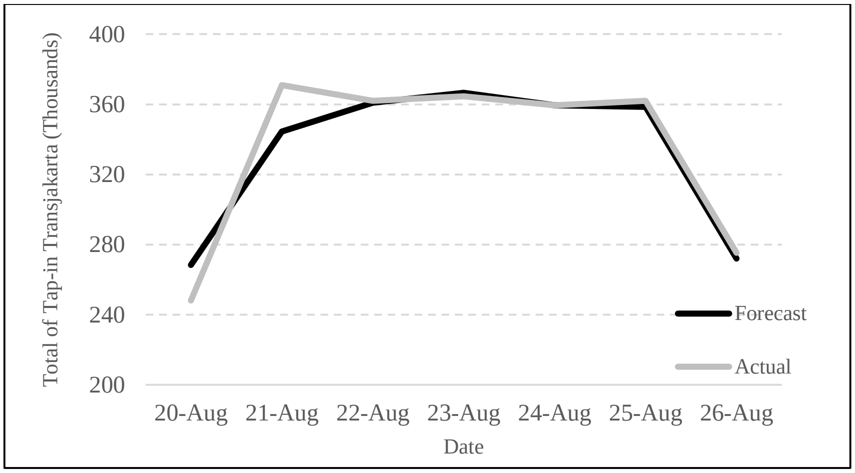

Figure 4. Forecast vs Actual data of SARIMA(0,0,0)(2,1,0)[7]

IV. CONCLUSION

The prediction results of the number of Transjakarta's passengers using the SARIMA model (0.0.0) (2.1.0) [7] with x-factor proved to be quite good with a fairly small error, namely with of MSE = 162402173, MAPE = 2.6122 and MASE = 0.211698.. The results of these predictions can be used to add insight to PT Transjakarta. In addition, this prediction can also be used by Jakarta Smart City to bring smart mobility closer to Jakarta.

REFERENCES

[1] Y. T. &. V. K., "Knowledge-based urban development: The local economic development path of Brisbane, Australia.," *Local Economy,* vol. 23(3), pp. 195-207, 2008.

[2] A. D. B. C. &. N. P. Caragliu, "Smart cities in Europe.," *Journal of Urban Technology,* vol. 18(2), p. 65–82, 2011.

[3] R. &. S. D. J. &. C. S. Xue, "Short-Term Bus Passenger Demand Prediction Based on Time Series Model and Interactive Multiple Model Approach.," *Discrete Dynamics in Nature and Society. ,* no. 10.1155/2015/682390. , 2015.

[4] F. K. M. C. E. T. M. &. O. L. Toqué, "Short & long term forecasting of multimodal transport passenger flows with machine learning methods.," *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC),* pp. 560-566, 2017.

[5] J. Cryer and K.-s. Chan, Time Series Analysis: With Applications in R., Springer, 2010.

[6] C. K. a. T. Kruangpradit, "Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) Model for Thailand Export," *International Institute of Forecasters,* 2017.

[7] R. J. Hyndman and G. Athanasopoulus, Forecasting : principles and practice, OTexts, 2013.

[8] M. H. Lee and N. A. Hamzah, "Calendar variation model based on ARIMAX for forecasting sales data with Ramadhan effect," *Proceedings of Regional Conference on Statistical Sciences 2010,* pp. 349-361, 2010.

[9] F. Seyyed, A. Abraham and M. & Al-Hajji, "Seasonality in stock returns and volatility: The Ramadan effect.," *Research in International Business and Finance, 19,* pp. 374-383, 2005.

[10] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer' and J. Stork, "Comparison of different Methods for Univariate Time Series Imputation in R.," 2015.

[11] R. J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software,* 2008.

[12] R. J.Hyndman and a. A. B.Koehle, "Another look at measures of forecast accuracy," *International Journal of Forecasting 22,* pp. 679-688, 2006.

[13] R. C. Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.," 2017.

[14] S. Pramana, R. Yordani, R. Kurniawan and B. Yuniarto, Dasar-dasar Statistika dengan Software R Konsep dan Aplikasi. 2nd Edition, Jakarta: InMedia, 2017.