Forecasting Surabaya – Jakarta Train Passengers with SARIMA model

S W Astuti[1] and Jamaludin[2]

[1]Program of Study Railway Building and Track Engineering, Indonesian Railway Academy Madiun, Jalan Tirtaraya Kota Madiun, Jawa Timur, Indonesia.
[2]Program of Study Railway Transport Management, Indonesian Railway Academy Madiun, Jalan Tirtaraya Kota Madiun, Jawa Timur, Indonesia.

Abstract. PT. Kereta Api Indonesia (Persero) DAOP 8 Surabaya is a State Owned Enterprise which operates rail transport services covering passenger and freight transport especially in East Java Province. This study aims to determine a forecasting model that able to provide prediction number of train passengers in relation Surabaya- Jakarta so as to help PT. Kereta Api Indonesia (Persero) DAOP 8 Surabaya anticipate the increase of train passengers by using Seasonal ARIMA (SARIMA) models. The data used in this research is secondary data from PT. Kereta Api Indonesia (Persero) DAOP 8 Surabaya on January 2012 until July 2017. Data collected then processed and analyzed using SARIMA method. Various SARIMA models were assessed and the best one was selected. Then, the models was evaluated based on normality of the residuals distribution, correspondence between the fitted and real amounts, and calculation of Sum Squared Residual. The results shown that the best time series model based on minimum Sum Squared Residual (SSR) is SARIMA Model $(0,1,1)(1,10)_{12}$. Based on analysis using SARIMA model $(0,1,1)(1,10)_{12}$ obtained the forecasting of Surabaya – Jakarta train passengers in January 2018 until July 2018 ranged from 119,495 – 161,685. Higher volume of passengers flows in July 2018 resulted by school and college holidays. Seeing the increasing trend of passengers during peak season especially in July, it is expected that PT KAI Daop 8 Surabaya will improve performance so that the increasing number of railway passengers can be maintained.

1. Introduction

Surabaya city has high accessibility, where there are airport, ports, stations and terminals in one region. In addition, the city of Surabaya is a center of industrial, commercial and government activity in the province of East Java, thus creating a high mobility in the areas. The city of Surabaya has in common with the city of Jakarta as an urban area, hence forming the profile of the people's journey to and from the workplace tends to be higher. It is no wonder the needs of long-distance travel such as intercity or interprovincial by the people of Surabaya high especially let the peak season such as holidays and weekends. Daerah Operational 8 Surabaya or abbreviated to DAOP 8 Surabaya is one of the Indonesian railway operations area, under the environment of PT Kereta Api Indonesia (Persero).

Forecasting of future demand for transport service represent important element of success for a transport company [1]. There are two types of quantitative forecasting models used are time series models and causal econometric models. Causal models are based on the statistical analysis of data for other related (explanatory) variables, and the use of these variables to forecast the variable of interest [2]. Time series data is a sequence of observations of the defined variable at a uniform interval over a period of time in successive order. Most common series are in annual, quarterly, monthly, weekly and daily frequencies [3]. One of model for forecasting univariate time series data is ARIMA, first introduced by Box and Jenkins (1976). This model has been originated from the Autoregressive model (AR), the Moving Average model (MA) and the combination of the AR and MA, the ARMA models. In the case where seasonal components are included then the model is called as the Seasonal ARIMA (SARIMA) model [4]. We usually symbolize the ARIMA model as ARIMA (p, d, q), where p and q are non-negative integers that correspond to the order of the autoregressive, integrated and moving average parts of the model, respectively. We can build seasonal ARIMA (P, D, Q) model as SARIMA (p, d, q)(P, D, Q)s. the parameters P, D and Q are the relevant seasonal autoregressive parameter, seasonal integrated parameter and seasonal moving average parameter [5].

SARIMA modeling was applied for deriving a model to access rail passenger demand for a case of Serbian railways using a time series of monthly passenger flows from January 2004 to June 2014. The best fitting model was found on the base of Box-Jenkins model and normalized Bayesian Information Criterion (BIC) [1]. Widhianti [6] analyzed the number of train passengers on PT KAI DAOP VI Yogyakarta using passenger flows on seasonal Islamic holyday from 2005 to 2013. The ARIMA model $(0,1,1)(0,1,1)_{12}$ fitted for used because an estimation of the parameters significant against a model and met all the assumption in the non autocorrelation, residual analysis, and normality. Akaike Inforation Criterion (AIC) used to find the best fitted SARIMA model. ARIMA Box – Jenkins method also used to forecaster the number of train passengers on Java and Sumatera islands. The best model with the smallest root means square error (RMSE) is ARIMA model (1,1,[12]) [7].

In this paper, we present an application of Seasonal ARIMA models to find a forecasting model for the case of Surabaya – Jakarta train passengers so as to help PT. Kereta Api Indonesia (Persero) DAOP 8 Surabaya anticipate the increase or the decrease of train passengers by using SARIMA (Seasonal ARIMA) models. SARIMA models aimed at time-series forecasting when they became stationary by differencing[1] Data collected then processed and analyzed using SARIMA method with Eviews 9.5 Student / Lite Version Program help. Different SARIMA models were applied to find the best fitting model and met all the assumption in in the non autocorrelation, residual analysis, homoscedasticity and normality [8]. The selected model has tested again with past data to see if the model accurately describes the state of the data. Models that meet assumptions, compared with Sum Squared Residual. In the selection of the best method (the most appropriate method) used to forecast a data can be considered with minimize errors (residuals) that have an error size value the smallest model. The model is said to be ideal if it meets all assumptions and has smallest sum squared of residuals [5]. The last step of the time series process is the prediction or forecasting of the model that is considered the best, and predictably the value of several future periods.

2. Research methods

The data used in this research is secondary data from PT. Kereta Api Indonesia (Persero) DAOP 8 Surabaya on January 2012 until July 2017. Box and Jenkins created a family of models known as Autoregressive Integrated Moving Average (ARIMA) models. ARIMA models aimed at time-series forecasting when they became stationary by differencing [1]. This is due to the fact that ARIMA models give more emphasis on the recent past rather than distant past. A time-series can have seasonal and non-seasonal characteristics. A series has seasonal characteristics when these are repeated overs time periods. For example for data collected monthly, a full seasonal distinction can be calculated as follows:

$$X_t' = X_t - X_{t-12} = (1 - B^{12})X_t \tag{1}$$

SARIMA notation can be extended to handle seasonal aspects, a common notation used is:

$$SARIMA(p,d,q)(P,D,Q)_s \tag{2}$$

Where p, d, q is the non-seasonal part of the model, P, D, Q the seasonal part of the model and s sum of periods per season [9].

In general, the steps using SARIMA method are model identification, model estimation, diagnostic checking and forecasting/ validation of the developed model [1]. The first step is to make the data plot to know whether the data contain a trend, seasonal, outlier, a variance is not constant. If data is not stationary then it needs differencing process [5]. After stationary series next step is to look at the ACF plot (autocorrelation function) and PACF (partial autocorrelation function) of the colegram. From ACF and PACF plots can be identified several possible models suitable for modeling [5]. After establishing several possible matching models and estimate the parameters, the next step is performed significance tests on the coefficients. When the model coefficient is significant then the model is feasible to be used for forecasting. From several significant models, the assumption test was performed on the residual, such as autocorrelation test, homoscedasticity test and normality test [8]. Models that meet assumptions, compared with Sum Squared Residual (SSR) [5]. The model is said to be ideal if it meets all assumptions and has the smallest SSR values. The last step of the time series process is the prediction or forecasting of the model that is considered the best, and predictably the value of several future periods.

3. Results

The number of passengers Surabaya - Jakarta from January 2012 to July 2017 both economy class, business and executives for Surabaya - Jakarta relations from 2012 to 2015 tend to be stable. The lowest number of passengers in March 2015 was 37,926 and the highest in May 2014 was 73,852 with an average of 55,424 passengers per month. However, from July 2016 to July 2017 the number of passengers tends to increase quite high with the lowest number of passengers in April 2016 of 55,382 passengers, the highest number in July 2017 of 131,059 passengers with average passengers experiencing a high increase of 85,638 passenger. (Figure 1)

The plot of passenger data can be seen in the following figure:

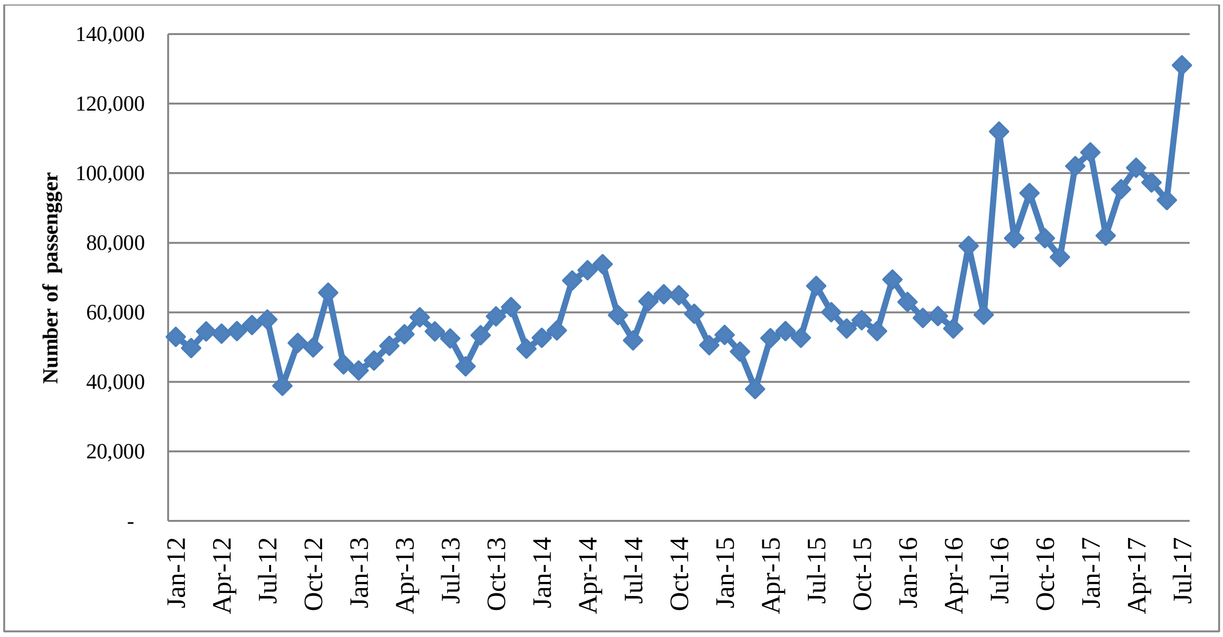


Figure 1. Plot of passengers.

Based on the image above there is a somewhat irregular fluctuation that identifies the possibility of seasonal factors in it. From the figure, there is also a tendency to increase the number of high passengers in the period of 2017. Because there are still trend elements, its need transformation to stabilize the variance [1]. The transformation to be taken is the transformation of natural logarithm (ln) [10]. (Figure 2)

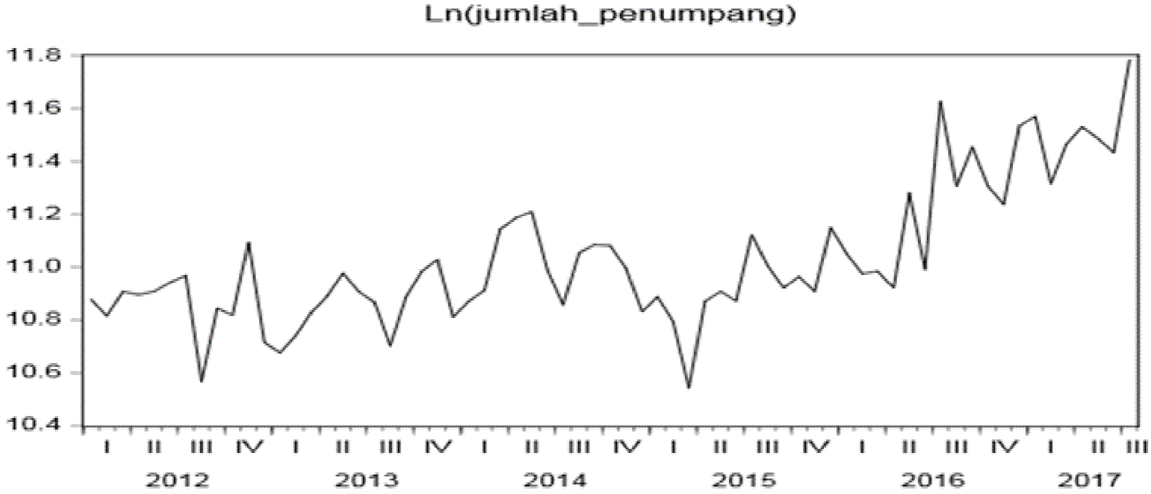


Figure 2. Number of passenger the result of transformation Ln.

From the picture above, it can be seen there are still elements of trends in the data. In order to eliminate the trends, it needs to be done the process of differencing. Seasonal difference terms are included in term, S=12 and D=1 and we selected SARIMA (p,1,q)(P,1,Q)$_{12}$ as the basic structure for the predictive SARIMA models. (Figure 3)

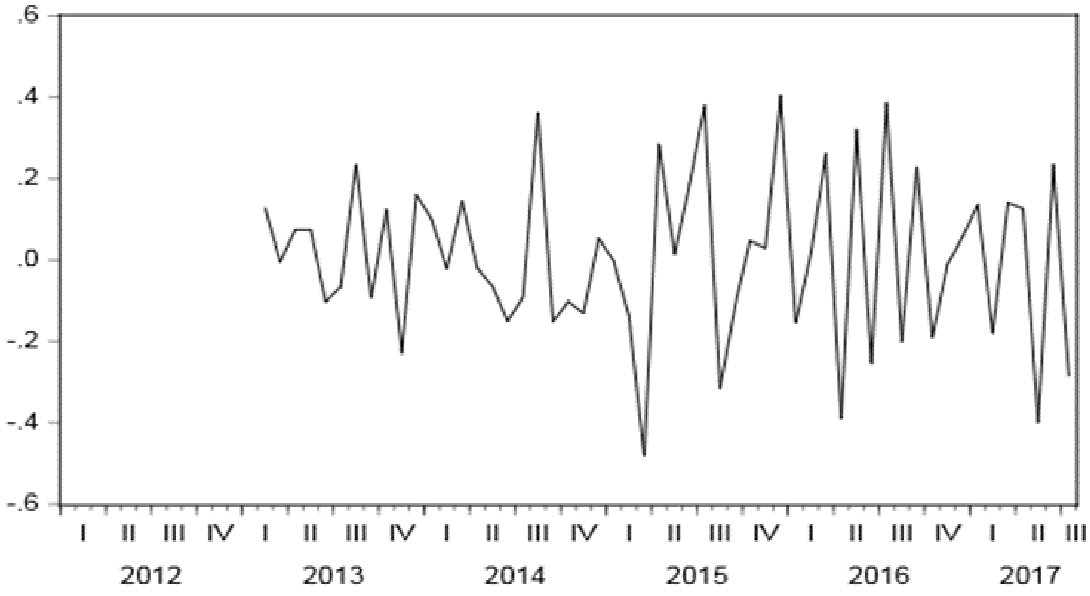
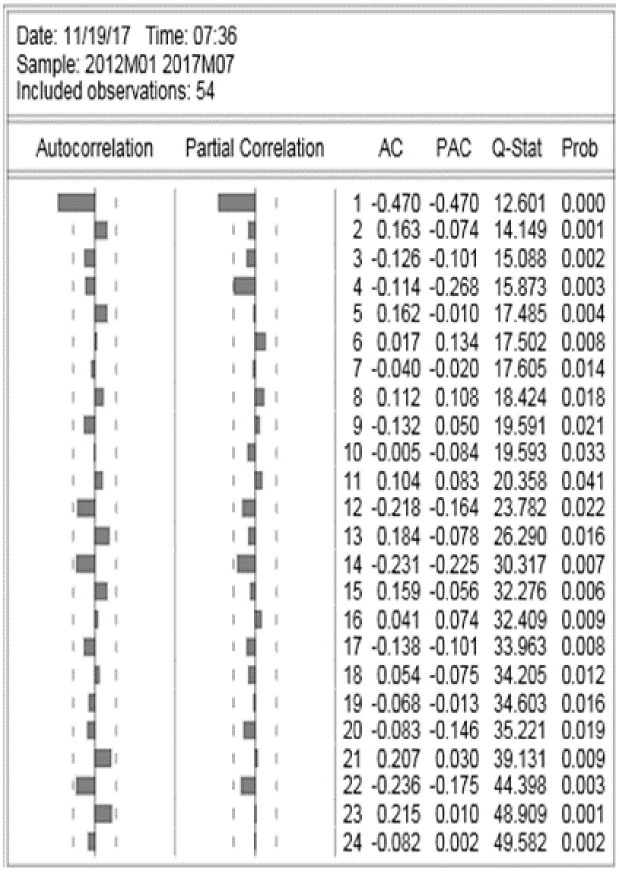


Figure 3. Number of train passenger result of differencing.

From the picture above shows that the data has been stationary both in the mean and variations. To further ensure the assumption of stationary residual average will be tested stationary test using Unit Root Test [9]. The stationary test was performed with the help of the Eviews program and application of the Augmented DickeyeFuller test (ADF) at a significance level of 5% [11]. Data is said to stationary if probability value ≤ 0.05. Dickey-Fuller test results show that the probability <0.05, so it can be concluded that the data is stationary in the average.

3.1. Identification of SARIMA model

In passing it is known that the time series data for the number of train passengers are influenced by seasonal factors. If it is considered based on table 1 above that in July there was a peak passenger spike and it repeats starting in 2015. While at the beginning of the year there was a decrease in passengers, therefore simply can be said that for the movement of seasonal passengers with seasonal season 12 months. (Figure 4)



```
Date: 11/19/17   Time: 07:36
Sample: 2012M01 2017M07
Included observations: 54

Autocorrelation   Partial Correlation      AC     PAC   Q-Stat  Prob

                                      1 -0.470 -0.470  12.601  0.000
                                      2  0.163 -0.074  14.149  0.001
                                      3 -0.126 -0.101  15.088  0.002
                                      4 -0.114 -0.268  15.873  0.003
                                      5  0.162 -0.010  17.485  0.004
                                      6  0.017  0.134  17.502  0.008
                                      7 -0.040 -0.020  17.605  0.014
                                      8  0.112  0.108  18.424  0.018
                                      9 -0.132  0.050  19.591  0.021
                                     10 -0.005 -0.084  19.593  0.033
                                     11  0.104  0.083  20.358  0.041
                                     12 -0.218 -0.164  23.782  0.022
                                     13  0.184 -0.078  26.290  0.016
                                     14 -0.231 -0.225  30.317  0.007
                                     15  0.159 -0.056  32.276  0.006
                                     16  0.041  0.074  32.409  0.009
                                     17 -0.138 -0.101  33.963  0.008
                                     18  0.054 -0.075  34.205  0.012
                                     19 -0.068 -0.013  34.603  0.016
                                     20 -0.083 -0.146  35.221  0.019
                                     21  0.207  0.030  39.131  0.009
                                     22 -0.236 -0.175  44.398  0.003
                                     23  0.215  0.010  48.909  0.001
                                     24 -0.082  0.002  49.582  0.002
```

Figure 4. ACF and PACF plots.

From the ACF and PACF correlogram the result of the first ln transform and the first differencing shows that there is a cut off of the first lag so it is assumed the data will be generated by MA (1) and AR (1)[5]. Because there are predicted seasonal factors, the temporary model that will be used is AR (1) and MA (1), with possible SARIMA model that is (1,1,1) $(1,1,1)_{12}$. But this does not rule out there other SARIMA models are formed. SARIMA models are available which may be as follows:

SARIMA (1,1,1) (0,1,1) $_{12,}$ SARIMA (0,1,1) (1,1,0) $_{12,}$ SARIMA (1,1,1) (1,1,1) $_{12,}$ SARIMA (0,1,1) (1,1,1) $_{12.}$ SARIMA (0,1,1) (0,1,1) $_{12}$ and SARIMA (1,1,0) (1,1,0) $_{12}$.

3.2. Estimation of SARIMA model

From the above models it is found that there are two models that are suitable for predicting model that is SARIMA (0,1,1) $(1,1,0)_{12}$ without constants and SARIMA (0,1,1) $(0,1,1)_{12}$ without constant. According to residual assumption test based on α=0.05, all of probability value (P-value) > 0.05 so that the residual data does not contain autocorrelation. From the correlogram square residuals also found p> 0.05 that the residual data does not contain homoscedasticity or it can be said that the residual variable is constant. According to the probability Jarque-Bera (the probability Jarque-Bera> 0.05), so it can be concluded that the residual follows the normal distribution or the assumption of residual normality is met. Based on the results of the above analysis, all residual assumptions are all met so that it can be said that the SARIMA model (0,1,1) $(1,10)_{12}$ and SARIMA model (0,1,1) $(0,1,1)_{12}$ can used to predict the number of passengers.

3.3. Selection of the best model

The best model will be selected from the value of Sum Squared Residual. (Table 1)

Table 1. Sum Squared Residual model SARIMA.

Model	Sum Squared Residual
$\text{SARIMA}(0,1,1)(1,1,0)_{12}$	1.637
$\text{SARIMA}(0,1,1)(0,1,1)_{12}$	1.832

The SARIMA model (0,1,1) $(1,1,0)_{12}$ has the smallest Sum Squared Residual so this model will be used to forecast the number of passengers.

3.4. Forecasting

The comparison between the number of actual passengers with the results of forecasting using the SARIMA model (0,1,1) $(1,1,0)_{12}$ are as follows: (Figure 5)

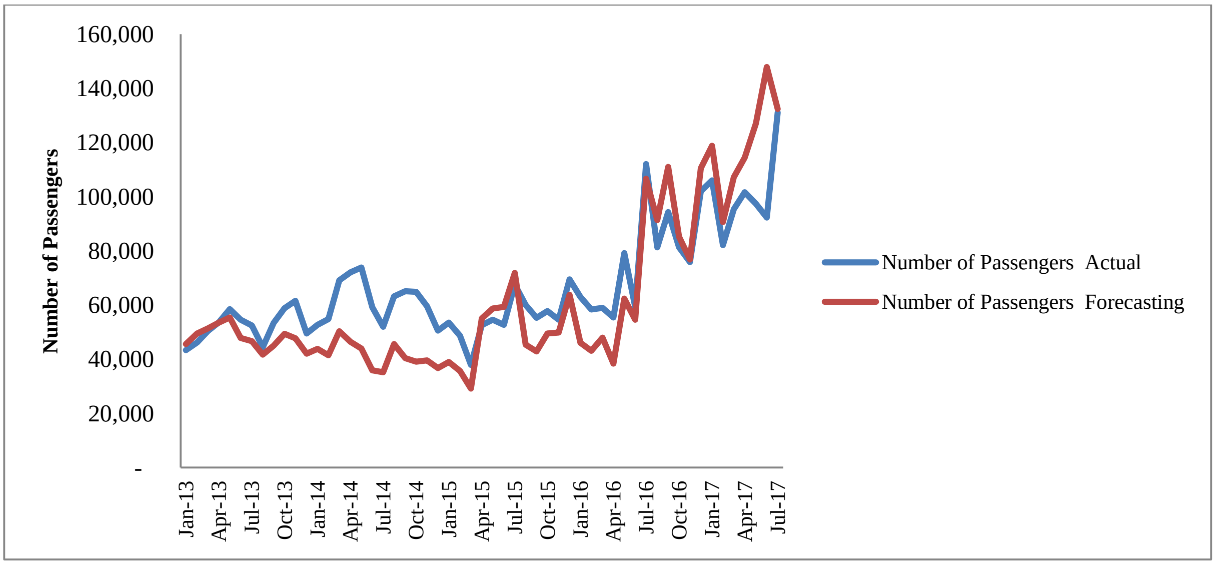


Figure 5. Actual vs forecasting.

From the figure above it is seen that result of the forecasting is lower than the actual from January 2013 to July 2016, and it increase after July 2016. But in general the pattern of the forecasting model is almost matched with the actual except from May 2017 to June 2017. The reason because of the number of passengers from July 2016 to July 2017 tends to increase quite high. Based on the above result also obtained mean absolute percentage error (MAPE) value of 18% where for the value of MAPE 18% is still quite good[10]. So the model can still be used for forecasting the number of passengers over the next period.

Table 2. Result of forecasting the number of passengers SARIMA $(0,1,1)(1,10)_{12}$.

Month	**Number of Passengers based on forecasting**

August 2017	95,459
September 2017	105,245
October 2017	95,411
November 2017	89,101
December 2017	118,559
January 2018	119,495
February 2018	96,372
March 2108	108,689
April2018	112,335
May 2018	118,972
June 2018	106,181
July 2018	161,685

Descriptively from the picture above shows that the number of train passenger relation Surabaya - Jakarta in October to November tend to decrease. However, in December - January will experience a fairly high spike. This is because in December - January there are Christmas and New Year holidays and Semester holidays for both school and college. The number of passengers decreased again in February, because this month there is no special moment and the number of days in the month of February is only 28 days and this is quite affecting the number of passengers. The number of passengers went back up to July. In July, there was a high increase because this month was peak season which in the month coincided with the school and college holidays that encourage the movement of higher by using the train mode.

4. Discussion

The result of a forecast is not a definite value in the coming period, considering other factors that also affect the volume of train passengers. So in the next research needs to be developed, for other forecasting modeling which takes into account factors other than the time series data of the passenger number.

5. Conclusion

Based on the above discussion can be summarized that the best time series model is based on model goodness value and fulfilment of assumptions to be used and with the smallest value of Sum Squared Residual is SARIMA Model (0,1,1) (1,10)$_{12}$. Seeing the increasing trend of passengers during peak season especially in July, it is expected that PT KAI Daop 8 Surabaya will improve performance so that the increasing number of railway passengers can be maintained.

Acknowledgements

The authors would like to thank the Indonesian Railway Academy Madiun for the support of this research and the PT. KAI (Persero) DAOP 8 Surabaya for providing the data.

References

[1] Milenković M, Švadlenka L, Melichar V, Bojović N and Avramović Z 2016 “SARIMA modelling approach for railway passenger flow forecasting,” *Transport* **4** 1–8.

[2] Lim C and McAleer M 2002 “Time series forecasts of international travel demand for Australia,” *Tour. Manag.* **23** (4) 389–396.

[3] Shrestha M B and Bhatta G R 2018 “Selecting appropriate methodological framework for time series data analysis,” *J. Financ. Data Sci.* **4** 71-89.

[4] Suhartono 2016 “Time Series Forecasting by using Seasonal Autoregressive Integrated Moving Average : Subset , Multiplicative or Additive Model,” **1** 20-27.

[5] Chang 2012 “Seasonal Autoregressive Integrated Moving Average Model for Precipitation Time Series,” *J. Math. Stat.* **8** (4) 500–505.

[6] Widhianti N and Wutsqa D U 2013 “Peramalan Banyak Penumpang Kereta Daerah Operasi Vi

Yogyakarta Menggunakan Model Time Series Dengan Variasi Kalender Islam Regarima," 978–979.

[7] Arianto B W 2017 "Peramalan Jumlah Penumpang Kereta Api Di Pulau Jawa Dan Sumatera Menggunakan Arima Box-Jenkins," 64.

[8] Gikungu S W 2015 "Forecasting Inflation Rate in Kenya Using SARIMA Model," *Am. J. Theor. Appl. Stat.* **4** (1) 15-19.

[9] Tadesse K B and Dinka M O 2017 "Application of SARIMA model to forecasting monthly flows in Waterval River, South Africa," *J. Water L. Dev.* **35 (**1) 23-25.

[10] Kumar S V and Vanajakshi L 2015 "Short-term traffic flow prediction using seasonal ARIMA model with limited input data," *Eur. Transp. Res. Rev.* **7** (3) 1–9.

[11] Hikichi S E, Salgado E G and Beijo L A 2017 "Forecasting number of ISO 14001 certifications in the Americas using ARIMA models," *J. Clean. Prod.* **147** 242–253.