M – 13

Univariate and Multivariate Time Series Models to Forecast Train Passengers in Indonesia

Lusi Indah Safitri[1], Suhartono[2], and Dedy Dwi Prastyo[3]
[1,2,3]Department of Statistics, Institut Teknologi Sepuluh Nopember
Email: lusi14@mhs.statistika.its.ac.id

***Abstract*—** Time series model is one of quantitative methods that frequently used for forecasting a number of train passengers in certain route. In general, there are two types of time series models, i.e. univariate and multivariate time series. The objective of this paper is to apply ARIMA model as a univariate method and VARIMA as a multivariate method for forecasting a number of executive train passengers in Indonesia, particularly Surabaya-Jakarta route. The number of daily train passengers in three types of executive classes that departure from Surabaya Pasar Turi station, i.e. Argo Bromo Anggrek Pagi, Argo Bromo Anggrek Malam, and Sembrani, are used as case study. The data are consisted 761 observations and recorded from January 1st, 2014 till February 27th, 2016 and divided into two parts, i.e. January 1st, 2014 to January 30th, 2016 and 1-27 February 2016 as training and testing data, respectively. Root mean of squares error (RMSE) in testing data is used as criteria to select the best forecasting model. The results show that ARIMA yields more accurate forecast at two data, i.e. number of passengers at Argo Bromo Anggrek Pagi and Sembrani, whereas VARIMA gives better forecast at Argo Bromo Anggrek Malam. Hence, this result inlines with the first conclusion of M3 competition, i.e. statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones.

Keywords: *forecasting, train passengers, ARIMA, VAR, RMSE.*

I. INTRODUCTION

The train is known as the mode of transport that has multiple advantages, such as energy saving, land-saving, environmentally friendly, high safety levels, able to transport large amounts, as well as adaptive to technological development [1]. The number of train tickets is sold uncertainty (fluctuatively) everyday. In general, the demand of train tickets at the weekend usually increase compared to normal days. Moreover, the peak of train tickets demand usually occurs one to five days ahead of Idul Fitri due to an annual *mudik* tradition in Indonesia. The term *mudik* refers to the exodus of Indonesian workers from the cities back to their hometowns ahead of Idul Fitri. Not only the Muslim community of Indonesia will return to their places of origin, but also people adhering to other religions traditionally use this public holiday to visit their parents or make a short holiday. This demand peak usually continues after Idul Fitri due to they must going back after this holiday.

A problem often faced by the railway operator is the large supply train ticket quotas are not appropriate to the number of train passengers. The number of train passengers usually increase in the days ahead of the national holidays or certain religious holidays. Due to the railway operator usually only provides the number of tickets as a normal day, this can lead to frustration of the passengers train because many passengers did not get a ticket. Hence, an accurate prediction of the number of rail passengers in the future is important to minimize the number of train passengers who did not get a ticket.

There are some researches on forecasting the train passengers have been conducted such as Andalita [2] who studied about forecasting the number of train passengers in economy class using ARIMA (Autoregressive Integrated Moving Average) and ANFIS (Adaptive Neuro Fuzzy Inference System) methods. Furthermore, Hermawan [3] applied the model NN (Neural Network) for forecasting the number of railway passengers in Jabodetabek. Additionally, Rosyidah [4] employed ARIMA modeling for forecasting of passenger trains on DAOP IX Jember.

In this research, forecasting the number of train passengers in executive class with univariate and multivariate time series. The data will be used is the number of train passengers in three executive trains, i.e. Argo Bromo Anggrek Pagi, Argo Bromo Anggrek Malam and Sembrani. The univariate time series modeling used ARIMA while the multivariate time series modeling employed Vector Autoregressive Integrated Moving Average (VARIMA). Modeling of multivariate time series are not only able to predict the number of passengers in the future, but also could explain the relevance of other types of executives trains. Once the best model of univariate and multivariate time series was obtained, the models were compared based on the accuracy to predict the number of train passengers. In forecasting, multivariate methods are usually more complicated than the univariate method. However as said by Makridakis and Hibon [5] that statistical methods are more sophisticated or more complicated does not always provide more accurate estimates than a simple method. Through the comparison of the two models time series derived from both methods will be obtained the best model to predict the number of rail passengers in the executive class Pasar Turi station.

II. LITERATURE REVIEW

A. Autoregressive Integrated Moving Average (ARIMA)

ARIMA(*p, d, q*) model is a combination of the AR (Autoregressive) order *p* model and MA (Moving Average) order *q* model with differencing order *d*. ARIMA model can be used in the seasonal and non-seasonal data. The ARIMA (*p, d, q*) can be written as follows [6]:

$$\phi_p(B)(1-B)^d Y_t = \mu + \theta_q(B)a_t,$$

Where μ is a constant, $\phi_p(B) = (1-\phi_1 B - \cdots - \phi_p B^p)$ is polynomial backshift operator for AR and $\theta_p(B) = (1-\theta_1 B - \cdots - \theta_q B^q)$ polynomial backshift operator for MA. The procedures used to obtain the forecasting value of using ARIMA consists of four steps starting from the model identification, parameter estimation, diagnostic testing and selection of the best models, forecasting. The identification can be done using ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots. This step is valid if the time series is stationary that can be visually checked through time series plot. If the data is not stationary in mean, then do differencing whereas if the data is not stationary in variance, then the Box-Cox transformation can be used. The complete steps to do ARIMA modeling can be seen in [4].

B. Vector Autoregressive (VAR)

The VAR model with order one denoted as VAR(1) follows this equation [6]:

$$\boldsymbol{Y_t} = \phi_o + \boldsymbol{\Phi} Y_{t-1} + \boldsymbol{a}_t, \tag{2}$$

The model in (2) that consists of two series can be written in matrix form as follows:

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1,t} \\ a_{2,t} \end{bmatrix}, \tag{3}$$

In general, the VAR model of order p denotes as VAR (*p*) is formulates as [4]:

$$\boldsymbol{Y_t} = \phi_o + \boldsymbol{\Phi} Y_{t-1} + \cdots + \boldsymbol{\Phi}_p Y_{t-p} + \boldsymbol{a}_t. \tag{4}$$

The complete description of VAR model can be seen in [6].

C. Model Identification, Diagnostic Checking, and Model Selection

Identification of time series model can be done by creating ACF and PACF plots for the univariate models. The identification step for the multivariate models can be done by employing MPACF (Matrix of Partial Autoregression Function) and the AIC (Akaike Information Criterion) [6].

Checking the assumptions of model is conducted after the identification and parameter estimation steps. The purpose of this step is to determine whether the model fulfills the assumptions. The ARIMA model assumes that the residual is white noise and normally distributed whereas the VARIMA model assumes that the vector of residual is white noise and follows multivariate normal distribution [6].

Selection of the appropriate model was done based on smallest RMSE (Root Mean Squared Error) of out of sample data. The RMSE is calculated as follows [6].

$$\text{RMSE}_{out\ of\ sample} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(Y_{n+l} - \hat{Y}_n(l)\right)^2}.$$

with L is number of observation in out of sample data, Y_{n+l} is observation l in out sample, and $Y_n(l)$ is the value of l-step forecasting.

III. RESEARCH METHODOLOGY

A. Data and Variables

The data used in the study were the data of the number passengers in daily train executive class with Surabaya-Jakarta route in the period from January 1, 2014, until February 29, 2016. These data are secondary data which obtained from Pasar Turi Train Station. There are three types of train executive class data, i.e. the number of passengers in the train Argo Bromo Anggrek Pagi, the number of Sembrani train's passengers, and the number of Argo Bromo Anggrek Malam passengers.

The variables in this study are denoted by $Y_{m,t}$ with m is stating the type of trains and t is stating the time (days). Following is the details of the variables in the study:

$Y_{1,t}$: The daily number of Argo Bromo Anggrek Pagi train passengers

$Y_{2,t}$: The daily number of Argo Bromo Anggrek Malam train passengers

$Y_{3,t}$: The daily number of Sembrani train passengers.

B. Steps of Analysis

1. ARIMA modeling with the following steps:
   a) Stationary inspection of data by looking at the data, ACF and PACF plots. If the data is not stationary variance, we can transform this data. If the data is not stationary in mean, we can difference this data. Formally, Augmented Dickey-Fuller test is used to check the data in the stationary mean.
   b) Identification of the model by ACF and PACF plots to determine the order of AR and MA
   c) Parameter estimation using OLS method.
   d) Selection of the best model by AIC criterion.
   e) Diagnosis check.
   f) Forecasting for data out samples with the selected order.
2. VAR modeling with the following steps:
   a) Inspection stationary of data such as the ARIMA model.
   b) To identify the model by MPACF and minimum AIC value thus obtained VAR order.
   c) Diagnosis check.
   d) Forecasting for data out samples with the model selected.
3. Compare the best model between univariate and multivariate models that have been selected.

IV. ANALYSIS AND RESULTS

A. Descriptive Statistics

The first step in this study was dividing the data into in sample, i.e. January 1, 2014 until January 14, 2015, and out of sample data span from January 15 until February 29, 2016. The descriptive statistics of the data are presented in Table 1. The largest average of the number of train passengers is for train ABAP (Argo Bromo Anggrek Pagi), followed by train ABAM (Argo Bromo Anggrek Malam) and Sembrani. However, the variance for Sembrani is the largest. The correlation between number of train passengers across trains are shown in Table 2.

TABEL 1. DESCRIPTIVE STATISICS FOR THE NUMBER OF TRAIN PASSENGERS

Train	Average	Variance	Minimum	Maximum
ABAP	350	6304	129	635
ABAM	338	5072	109	540
Sembrani	317	6307	98	544

TABEL 2. CORRELATION OF NUMBER OF PASSENGERS OF BETWEEN TRAINS

Train	ABAP	ABAM	Sembrani
ABAP	1	-	-
ABAM	0.677	1	-
Sembrani	0.666	0.761	1

Table 2 shows that number of passengers between trains are correlated. The strongest correlation is between ABAM and Sembrani. All the correlation are significant with *p*-values are less than 0.05. These facts motivated the use of multivariate time series to model the data.

B. ARIMA Model

ARIMA modeling begins with the identification step given the series stationary. Following is time series plot of the three train's passenger data: ABAP (left), ABAM (middle), and Sembrani (right). Based on time series plot in Figure 1, the data seems to be not stationary. They have high fluctuations. In order to know the stationary of these data, the Box-Cox transformation was used to check it where the results were reported in Table 3.

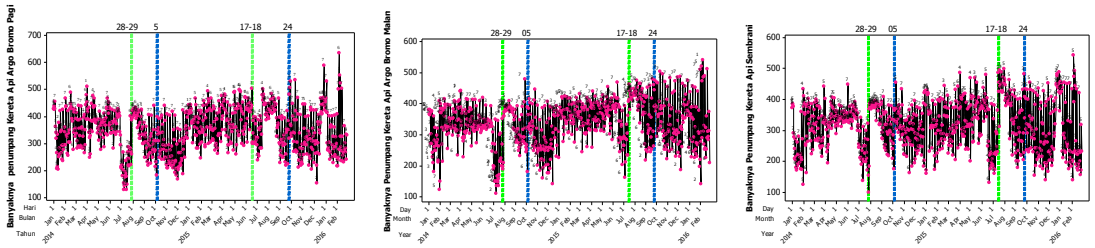

FIGURE 1. TIME SERIES PLOT OF NUMBER OF PASSENGERS TRAINS

TABEL 3. BOX-COX TRANSFORMATION

Variable	Rounded Value	Lower Limit	Upper Limit
ABAP	1	0.82	1.48
ABAM	1.36	1.06	1.71
Sembrani	1.33	1.06	1.65

In the ABAP, the rounded value for the parameter estimate in Box-Cox transformation is one. This indicates that only the ABAP variables are stationary in variance. The series for ABAM and Sembrani are required to be transformed. The next step is to check the stationary in the mean. The checking can be viewed via time series plot and ACF. ACF plot shows a form of a cut off pattern. This shows that the data has not been stationary in the mean. It is necessary to do differencing, i.e. differencing order one and seven. After differencing, the next step is to look at the ACF and PACF plot of the data that has been stationary. Through ACF and PACF plot can be determined the order of ARIMA models. The ACF and PACF plots had been stationary using differencing order one and seven.

The ARIMA models for the ABAP data is ARIMA (0, 1, 1) $(0, 1, 1)^7$ because the ACF plot occurs in the lag 1,6,7,8 are cut off. The modeling for ABAM and Sembrani series need an outlier detection approach because the normality assumption in ARIMA models was not fulfilled. The model for ABAM series is ARIMA ([6,7,8], 1,2)$(0,1,1)^7$ with few outliers whereas the model for Sembrani is ARIMA (7, 1, 1) $(0, 1, 1)^7$ with few outliers. All models have fulfilled the assumptions, i.e. white noise and normality distribution. Thus, the model for $Y_{1,t}$ is:

$$Y_{1,t} = Y_{t-1} + Y_{t-7} - 0.610a_{t-1} - 0.812a_{t-7} + a_{1,t}.$$

The model for $Y_{2,t}$ is:

$$Y_{2,t} = \frac{(1-0.29B-0.21B^2)(1-0.83B^7)}{(1+0.10B^6-0.3B^7+0.27B^8)} a_t + W_{AO} I_t^{(T)}$$

with $W_{AO} I_t^{(T)}$ is:

$$\begin{aligned}
&= -2554.1I_t^{(33)} - 2033.3I_t^{(458)} - 2144.5I_t^{(31)} + 2173.8I_t^{(634)} + 1350.5I_t^{(631)} - 950.3I_t^{(453)} - 1040.3I_t^{(365)} - 1438.7I_t^{(416)} + \\
&\quad - 1548.3I_t^{(39)} + 1116.1I_t^{(452)} - 1389.1I_t^{(486)} + 1084.6I_t^{(14)} + 1224.9I_t^{(107)} + 1103.6I_t^{(271)} + 1116.4I_t^{(361)} - 1031.4I_t^{(208)} + \\
&\quad - 943.9I_t^{(347)} + 741.7I_t^{(689)} - 1447.5I_t^{(609)} - 1319.0I_t^{(610)} + 1154.7I_t^{(358)} + 1046.4I_t^{(134)} - 687.0I_t^{(730)} + 1005.9I_t^{(291)} + \\
&\quad - 979.1I_t^{(121)} + 1025.2I_t^{(10)} + 1028.4I_t^{(705)} + 1016.3I_t^{(551)} + 1169.0I_t^{(206)} + 1006.7I_t^{(444)} + 1272.1I_t^{(187)} + 1168.1I_t^{(194)} + \\
&\quad 986.8I_t^{(190)} + 781.8I_t^{(176)} + 792.9I_t^{(594)} + 974.8I_t^{(718)} + 838.5I_t^{(161)} - 822.2I_t^{(632)} - 764.6I_t^{(707)} - 1029.9I_t^{(38)} + 881.5I_t^{(652)} + \\
&\quad 1234.4I_t^{(655)} + 1093.3I_t^{(648)} + 830.7I_t^{(280)} - 735.4I_t^{(431)} + 492.5I_t^{(725)} + 1006.2I_t^{(326)} + 634.1I_t^{(269)} - 707.4I_t^{(337)} + 612.4I_t^{(297)} + \\
&\quad 961.5I_t^{(199)} - 824.9I_t^{(471)} + 904.5I_t^{(681)} + 717.4I_t^{(16)} + 1384.1I_t^{(30)} - 595.1I_t^{(92)} + 696.7I_t^{(152)} + 724.2I_t^{(211)} - 842.2I_t^{(311)} + \\
&\quad 522.7I_t^{(455)} - 502.7I_t^{(533)} - 662.8I_t^{(483)} - 699.6I_t^{(378)} + 649.4I_t^{(185)} + 659.9I_t^{(414)}.
\end{aligned}$$

The model for $Y_{3,t}$ is :

$$Y_{3,t} = \frac{(1-0.44B)(1-0.88B)^7}{(1+0.09B^2-0.28B^7+0.12B^8)} a_t + W_{AO} I_t^{(T)} + W_{ls} I_t^{(T)},$$

with $W_{AO} I_t^{(T)}$ and $W_{ls} I_t^{(T)}$ is:

$$\begin{aligned}
&= 2334.9I_t^{(631)} - 1249.5I_t^{(31)} + 1786.0I_t^{(452)} - 1467.7I_t^{(486)} + 1243.9I_t^{(457)} + 1299.5I_t^{(358)} + 1441.3I_t^{(414)} - 1538.4I_t^{(39)} + \\
&\quad - 1131.3I_t^{(530)} + 1268.8I_t^{(485)} + 1296.6I_t^{(444)} - 1689.9I_t^{(366)} - 1392.7I_t^{(208)} + 1211.1I_t^{(306)} + 1396.1I_t^{(636)} + -1237.3I_t^{(458)} \\
&\quad + 1123I_t^{(498)} + 1073.9I_t^{(696)} + 1098.0I_t^{(87)} - 1075.6I_t^{(283)} + 1003.5I_t^{(368)} + 1024I_t^{(623)} - 1000.2I_t^{(416)} - 888.9I_t^{(564)} - 973.2I_t^{(53)} + \\
&\quad 981.5I_t^{(735)} + 896.8I_t^{(402)} - 1122.5I_t^{(736)} - 1630.6I_t^{(365)} + 1523.4I_t^{(30)} + 1372.3I_t^{(634)} - 1060.0I_t^{(737)} + 1559.4I_t^{(361)} + \\
&-654.8I_t^{(5)}.
\end{aligned}$$

C. VARIMA Model

VARIMA modeling begins with the identification step based on MACF and MPACF plots and AIC. The series for three train passenger were differenced order one and seven because the data are not stationary in the mean. These series were transformed using Box-Cox transformation because they are not stationary in variance. The minimum value of the AIC indicated that the proper model is VARIMA $(30,1,0)(0,1,0)^7$.

The results of parameter estimation of VARIMA $(30,1,0)(0,1,0)^7$ indicates that the model has 270 parameters. However, the *p*-value of each parameter some parameters were not significant. So, the model need restrict some parameters. This restriction process start from the parameter estimate with the highest *p*-value until all parameter estimates had *p*-value less than Type-I error ($\alpha = 0.05$). The results of parameter estimation VARIMA $(30,1,0)(0,1,0)^7$ with restriction showed that there are 78 parameters that were significant in the model. The VARIMA $(30,1,0)(0,1,0)^7$ model for the series of the number of train passengers is:

$$\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}-\begin{bmatrix}-0.7&0,01&0.1\\0&0.7&0\\0&0&-0.7\end{bmatrix}B-\begin{bmatrix}-0.4&0&0.1\\0.3&-0.6&0\\0&0&0.6\end{bmatrix}B^{2}+\begin{bmatrix}-0.3&0&0.1\\0.7&-0.4&0.4\\0&0&0.3\end{bmatrix}B^{3}-\begin{bmatrix}-0.2&0&0\\0.7&-0.3&0\\0.1&0&-0.3\end{bmatrix}B^{4}-\begin{bmatrix}-0.1&0&0.1\\0&-0.2&0.5\\0.1&0&-0.2\end{bmatrix}B^{5}+\right.$$
$$-\begin{bmatrix}-0.1&0&0\\0&-0.2&0\\0&0&0.1\end{bmatrix}B^{6}-\begin{bmatrix}0.7&0&0\\0&-0.7&0\\0&0&-0.7\end{bmatrix}B^{7}-\begin{bmatrix}-0.44&0&0.1\\0&-0.5&0\\0&0&-0.5\end{bmatrix}B^{8}-\begin{bmatrix}-0.24&0&0\\0&-0.4&0\\0&0&0.4\end{bmatrix}B^{9}-\begin{bmatrix}0.2&0&0\\0&-0.2&0\\0&0&-0.2\end{bmatrix}B^{10}-\begin{bmatrix}0.1&0&0\\0&-0.2&0\\0&0&-0.1\end{bmatrix}B^{11}+$$
$$-\begin{bmatrix}0&0&0\\0&-0.1&0\\0&0.1&0\end{bmatrix}B^{12}-\begin{bmatrix}0.4&0&0\\0&-0.4&0\\0&0&-0.5\end{bmatrix}B^{14}-\begin{bmatrix}-0.3&0&0\\0&-0.3&0\\0&0&-0.4\end{bmatrix}B^{15}-\begin{bmatrix}-0.1&0&0\\0&-0.2&0\\0&0&-0.3\end{bmatrix}B^{16}-\begin{bmatrix}-0.1&0&0\\0&-0.1&0\\0&0&-0.2\end{bmatrix}B^{17}-\begin{bmatrix}-0.1&0&0\\0&0&0\\0&0&0\end{bmatrix}B^{18}+$$
$$-\begin{bmatrix}0&0&0\\0&0&0\\0.1&0&0\end{bmatrix}B^{19}-\begin{bmatrix}0.1&0&0\\0&0&0\\0.1&0&0\end{bmatrix}B^{20}-\begin{bmatrix}-0.3&0&0\\0&0.3&0\\0&0&-0.3\end{bmatrix}B^{21}-\begin{bmatrix}-0.2&0&0\\0&-0.2&0\\0&0&-0.2\end{bmatrix}B^{22}-\begin{bmatrix}0&0&0.1\\0&0.1&0\\0&0&-0.2\end{bmatrix}B^{23}+\begin{bmatrix}-0.1&0&0\\0&0.1&0\\0&0&-0.1\end{bmatrix}B^{24}+\begin{bmatrix}-0.1&0&0\\0&0&0\\0&0&0\end{bmatrix}B^{25}+$$
$$\left.\begin{bmatrix}-0.2&0&0\\0&-0.2&0\\0&0&-0.2\end{bmatrix}B^{28}+-\begin{bmatrix}-0.1&0&0\\0&-0.1&0\\0&0&0\end{bmatrix}B^{29}\right)\times\begin{bmatrix}(1-B)(1-B^{7})Y_{1,t}\\(1-B)(1-B^{7})Y_{2,t}\\(1-B)(1-B^{7})Y_{3,t}\end{bmatrix}+\begin{bmatrix}a_{1,t}\\a_{2,t}\\a_{3,t}\end{bmatrix}.$$

Based on the model that had been established, the next step is testing the assumptions on residual. In multivariate time series modeling, to testing the assumption of white noise on the residual can be done by looking at the results of the portmanteau test. In this study, testing the white noise used AIC criterion calculated from the residual of VARIMA $(30,1,0)(0,1,0)^{7}$ model. Based on Table 4, the smallest AIC value is in AR (0) and MA (0). This suggests that the residuals of the model have fulfilled the white noise assumption.

TABEL 4. MINIMUM INFORMATION CRITERION OF RESIDUAL

Lag	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR(0)	27.26852	27.32797	27.32771	27.32819	27.32541	27.3277
AR(1)	27.28481	27.31673	27.32693	27.3307	27.33645	27.34313
AR(2)	27.29714	27.32953	27.33551	27.33822	27.34943	27.35946
AR(3)	27.3034	27.3347	27.33804	27.34664	27.35837	27.37177
AR(4)	27.30585	27.3421	27.35016	27.35857	27.37946	27.39254
AR(5)	27.31611	27.35402	27.36436	27.37477	27.39558	27.4099

The next assumption that must be fulfilled is multivariate normal distribution for the vector of residual. The null hypothesis of the test is that the residuals follows have multivariate normal distribution. The null hypothesis will be fail to be rejected if the *p*-value of the statistic test exceed the value of Type-I error. The evaluation of the multivariate normal assumption test can also be done visually through QQ plot of the residual. The assumption is fulfilled when residual plot tends to form a straight diagonal line as displayed by Figure 2. Moreover, if the proportion of values which generated through calculation are at least 50% greater than the value of statistic test, in this study is about 83%, the null hypothesis is fail to be rejected. Therefore, residuals of the VARIMA $(30,1,0)(0,1,0)^{7}$ model have multivariate normal distribution.

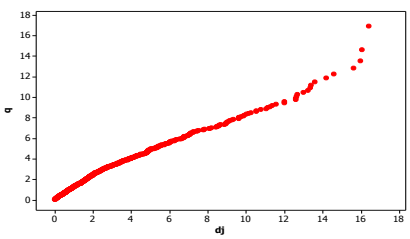

FIGURE 2. QQ PLOT FOR RESIDUALS

Once the best model for ARIMA and VARIMA were obtained, the out of sample forecast can be calculated for the series. The RMSE out of the samples were reported in Table 5. The models for ABAP

and Sembrani have relatively large RMSE for both univariate and multivariate models. This indicates that the two models are less good than the model for ABAM. The comparison of RMSE out samples for univariate and multivariate models indicated that multivariate model, which is more complex, were not always resulted in better prediction than the simpler model. Often in some studies suggest that more complicated method will yield better accuracy but the reality is not always so. Table 6 reported the forecasting value obtained from ARIMA model for each train.

TABEL 5. THE COMPARISON RMSE OUT OF SAMPLE FOR UNIVARIATE AND MULTIVARIATE

Variable	RMSE Out of Sample for ARIMA	RMSE Out of Sample for VARIMA
ABAP	111.5	133.8
ABAM	91.2	77.5
Sembrani	102.1	120.0

TABEL 6. THE RESULTS OF FORECASTING FOR EACH TRAIN

Dates	ARIMA			Dates	ARIMA		
	ABAP	ABAM	Sembrani		ABAP	ABAM	Sembrani
15-Jan-16	274	372	287	6-Feb-16	253	285	150
16-Jan-16	295	285	213	7-Feb-16	339	432	308
17-Jan-16	382	432	369	8-Feb-16	255	305	167
18-Jan-16	297	305	226	9-Feb-16	197	286	153
19-Jan-16	240	286	198	10-Feb-16	190	322	165
20-Jan-16	232	322	219	11-Feb-16	213	281	162
21-Jan-16	255	281	204	12-Feb-16	218	369	234
22-Jan-16	260	369	269	13-Feb-16	239	281	137
23-Jan-16	281	281	182	14-Feb-16	325	430	297
24-Jan-16	368	430	337	15-Feb-16	241	302	155
25-Jan-16	283	302	198	16-Feb-16	183	283	140
26-Jan-16	225	283	180	17-Feb-16	176	318	152
27-Jan-16	218	318	194	18-Feb-16	199	280	150
28-Jan-16	241	280	188	19-Feb-16	204	366	223
29-Jan-16	246	366	256	20-Feb-16	225	278	124
30-Jan-16	267	278	165	21-Feb-16	311	428	287
31-Jan-16	354	428	320	22-Feb-16	227	299	142
1-Feb-16	269	299	181	23-Feb-16	169	281	127
2-Feb-16	211	281	166	24-Feb-16	162	315	139
3-Feb-16	204	315	178	25-Feb-16	185	278	137
4-Feb-16	227	278	175	26-Feb-16	190	363	212
5-Feb-16	232	372	245	27-Feb-16	211	276	110

V. CONCLUSIONS AND RECOMMENDATIONS

The empirical results found that the best model for univariate approach to model number of passenger train are ARIMA $(0,1,1)(0,1,1)^7$ for Argo Bromo Anggrek Pagi, ARIMA$([6,7,8],1,2)(0,1,1)^7$ with outliers detection for Argo Bromo Anggrek Malam, and ARIMA $(7,1,1)(0,1,1)^7$ with outliers detection for Sembrani. The multivariate models appropriate to model these three series is VARIMA $(30,1,0)$ $(0,1,0)^7$. Based on the RMSE value for out of sample data, it can be concluded that the ARIMA model had better prediction for two series whereas the VARIMA model outperformed in one series, i.e. series for Argo Bromo Anggrek Malam.

The empirical results also showed that many outliers were found in the data and influenced the forecast accuracy of both univariate and multivariate models. Hence, more detail about outlier detection can be done for further research. Moreover, non linear time series models such as ANN (Artificial Neural Network) which is flexible to overcome data with outliers also could be considered as a future research for forecasting train passengers in Indonesia.

REFERENCES

[1] PT. Kereta Api Indonesia (2016). www.kereta-api.co.id. Situs Resmi PT. Kereta Api Indonesia (Persero).

[2] Andalita, I. (2015). Peramalan Jumlah Penumpang Keret Api Kelas Ekonomi Kertajaya Menggunakan ARIMA dan ANFIS. Tugas Akhir Statitika ITS. Surabaya.

[3] Hermawan, N. (2014). Aplikasi Model Neural Network dan Neuron Fuzzy Untuk Peramalan Banyaknya Penumpang Kereta Api Jabotadetabek. Tugas Akhir UNY. Yogyakarta.

[4] Rosyiidah, U. (2013). Pemodelan ARIMA Dalam Peramalan Penumpang Kereta Api Pada DAOP IX Jember. Jurnal. Jember.

[5] Makridakis, S., & Hibon, M. (2000). The M3-Competition: Results, Conclusions, and Implications. *International Journal of Forecasting*, 16, 451-476

[6] W. W. Wei. *Time Series Analysis (Univariate and Multivariate Methods).* United States of America:Pearson Education,Inc. (2006)