Number of Foreign Tourist Arrival Forecasting Using Percentile Error Bootstrap Based on VARIMA Model

Hidayatul Khusna[1], Muhammad Ahsan[1], and Dedy Dwi Prastyo [1]

Abstract*— Forecasting number of foreign tourist arrivals is important to improve the policies in the tourism sector. Better accuracy of forecast would help the government and investor to make operational, tactical, and strategic decisions. Data used in this research are monthly number of foreign tourist arrivals taken from Indonesia Central Bureau of Statistics. Multivariate forecasting at Soekarno-Hatta, Juanda, and Adi Sumarmo arrival gates was conducted using VARIMA ([12],1,0) (0,1,0)^{12} model. However, the longer step ahead to forecast, the larger variance error of corresponding models. As a result, the prediction interval become wider. This research computed the prediction interval using percentile error bootstrap based on VARIMA models that produced more precise forecast.

***Keywords*— Foreign tourist arrival, prediction interval, percentile error bootstrap, VARIMA.**

I. Introduction

As a developing country, one of the backbones of Indonesian economy is the tourism sector. The fascination arises from the variety of cultural attractions, the beauty of nature as well as the various dishes containing the value of high taste in culinary tourism. This attractiveness makes Indonesia known as one of the popular tourist destination. Thus, forecasting number of foreign tourist arrivals is important for improving the strategies are made in the tourism sector.

The tourism sector is expected to drive the economy as the sector most prepared in terms of facilities and infrastructure compared to other sector. This expectation is poured into Indonesia's foreign targets amounted to 240 trillion rupiah and the number of foreign tourists amounted to 20 million tourists in 2019 [1]. Therefore, the Ministry of Culture and Tourism set a target of tourist arrivals increased in every year as many as 7.7 million visitors.

Vector autoregressive integrated moving average (VARIMA) is one of the popular multivariate forecasting models. In this research, the VARIMA model was applied to forecast simultaneously the number of foreign tourist arrivals in three arrival gates, i.e. Soekarno-Hatta, Juanda, and Adi Sumarmo air ports. However, the longer step ahead to forecast, the larger variance error of VARIMA models [2]. As a result, the prediction intervals become wider. Bootstrap is one of nonparametric methods that used to built prediction intervals on the data that consist of outlier or non-linear pattern [3]. This research is aimed to forecast number of foreign tourist arrivals in those three gates using VARIMA models with percentile error bootstrap to built prediction interval.

II. Forecasting Method

A. VARIMA Model

ARIMA model is one of time series model that most popular and recently used. ARIMA (p,d,q) is built from AR (p) combine with MA (q) at non-stationary data then differencing by order d. ARIMA (p,d,q) can be written as mathematical formula:

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)(1-B)^d y_t=(1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q)a_t \quad (1)$$

where p is AR model order and q is MA model order. ARIMA for seasonal data pattern with seasonal period s, or ARIMA (p,d,q) (P,D,Q)^s can be written as follow:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D y_t=\theta_q(B)\Theta_Q(B^s)a_t, \quad (2)$$

where P is seasonal AR model order, Q is seasonal MA model order, $\Phi_P(B^s)=1-\Phi_1 B^S-\Phi_2 B^{2S}-\cdots-\Phi_P B^{Ps}$, and $\Theta_Q(B^S)=1-\Theta_1 B^S-\Theta_2 B^{2S}-\cdots-\Theta_Q B^{QS}$.

Conditional Least Square is one of the methods that used to estimate the parameter of ARIMA model [5]. Basically, this method is design to minimize mean squared error (MSE) of forecast. ARMA (p,q) can be represented to MA (∞) as follow:

$$Y_t=a_t+\psi_1 a_{t-1}+\psi_2 a_{t-2}+\psi_3 a_{t-3}+\cdots \quad (3)$$

Point forecast for h step ahead that minimize MSE forecast can be written as:

$$\hat{Y}_t(h)=\psi_h a_t+\psi_{h+1}a_{t-1}+\psi_{h+2}a_{t-2}+\cdots=\sum_{j=0}^{\infty}\psi_{h+1}a_{t-j} \quad (4)$$

By minimize MSE forecast, point forecast error can be written as:

$$e_t(h)=a_{t+h}+\psi_1 a_{t+h-1}+\psi_2 a_{t+h-2}+\cdots+\psi_{h-1}a_{t+1}=\sum_{j=0}^{h-1}\psi_j a_{t+h-j} \quad (5)$$

Mean and variance of point forecast error can be written as:

$$E[e_t(h)]=E[a_{t+h}+\psi_1 a_{t+h-1}+\psi_2 a_{t+h-2}+\cdots+\psi_{h-1}a_{t+1}=0 \quad (6)$$

$$Var[e_t(h)]=\sigma_a^2(1+\psi_1{}^2+\psi_2{}^2+\cdots+\psi_{h-1}^2)=\sigma_a^2\sum_{j=0}^{h-1}\psi_j{}^2 \quad (7)$$

Prediction intervals for ARIMA models can be written as:

$$\hat{Y}_t(h)+z_{\alpha/2}\sqrt{\mathrm{var}[e_t(h)]}<Y_t(h)<\hat{Y}_t(h)+z_{1-\alpha/2}\sqrt{\mathrm{var}[e_t(h)]} \quad (8)$$

where $\hat{Y}_t(h)$ is point forecast for h step ahead, $e_t(h)$ is point forecast error for h step ahead, and α is probability that prediction intervals cannot cover the actual data.

For a set of n time series variables $y_t=(y_{1t},y_{2t},...,y_{nt})'$, a VAR model of order p or VAR (p) can be written as follow [2]:

$$y_t=A_1 y_{t-1}+A_2 y_{t-2}+...+A_p y_{t-p}+a_t \quad (9)$$

[1]Hidayatul Khusna, Muhammad Ahsan, Dedy Dwi Prastyo are with Department of Statistics, Sepuluh Nopember Institute of Technology, Surabaya, 60111, Indonesia. E-mail: khusna.hidayatul15@mhs.statistika.its.ac.id; ahsan15@mhs.statistika.its.ac.id; dedy-dp@statistika.its.ac.id

where A_i are $n \times n$ coefficient matrices and $a_t = (a_{1t}, a_{2t}, ..., a_{nt})'$ is an unobservable identical independence and normal distributed with zero mean error term. Consider a two-variable VAR (1) with $k = 2$ can be written as:

$$y_{1t} = \phi_{11} y_{1t-1} + \phi_{12} y_{2t-1} + a_{1t} \quad (10)$$

$$y_{2t} = \phi_{21} y_{1t-1} + \phi_{22} y_{2t-1} + a_{2t} \quad (11)$$

with $a_{it} \sim i.i.d(0, \sigma_{ai}^2)$ and $\mathrm{cov}(a_i, a_j) = 0; i \neq j$

Following are representation of these equations in matrix form:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} \quad (12)$$

II *Percentile Error Bootstrap*

Percentile error bootstrap applied on state space model by Hyndman et al. [3]. Basicly, percentile error bootstrap can be applied by generate N error sample for each step of point forecast by using ordinary resampling. Following are detail step to built prediction interval by using percentile error bootstrap.

a. Calculate residuals
b. Generate N error sample for each step i by using ordinary resampling, where $i = 1, 2, \ldots, h$ step ahead and N is number of resampling
c. For each step of point forecast (i), order N error sample from smallest to biggest value
d. Calculate percentile error $\alpha/2$ and $1 - \alpha/2$ for each step of point forecast (i). Thus, the lower and upper prediction intervals can be written as follow:

$$\hat{Y}_t(i) + e_{[\alpha/2]} < Y_t(i) < \hat{Y}_t(i) + e_{[1-\alpha/2]}$$

Repeat step b until d to get prediction interval for each step of point forecast, $i = 2, 3, \ldots, h$.

III. METHODOLOGY

Data that used in this research are monthly number of foreign tourist arrivals taken from Indonesia Central Bureau of Statistics. Data from January 1996 until December 2013 are taken as training data. Then, data from January 2014 until March 2016 are taken as testing purpose [4]. Following are the steps of data analysis:

a. Explore the information of number of foreign tourist arrival data by using descriptive statistics analysis.
b. Identifying and selecting the best VARIMA model.
c. Forecast number of foreign tourist arrival by using VARIMA model.
d. Built prediction interval by using the best VARIMA model.
e. Built prediction interval by using percentile error bootstrap approach based on VARIMA model.
f. Compare both prediction interval with out-sample data, to check the accuracy of prediction intervals.
g. Forecast number of foreign tourist arrival for April 2016 until December 2017 by using the best prediction interval.

IV. RESULT AND DISCUSSION

This research is aimed to forecast number of foreign tourist arrival at three arrival gate, i.e. Juanda, Soekarno Hatta, and Adi Sumarmo. Time Series plots of those three series can be seen at Figure 1. It reveals that the numbers of foreign tourist arrivals data are not stationary both in mean and variance.

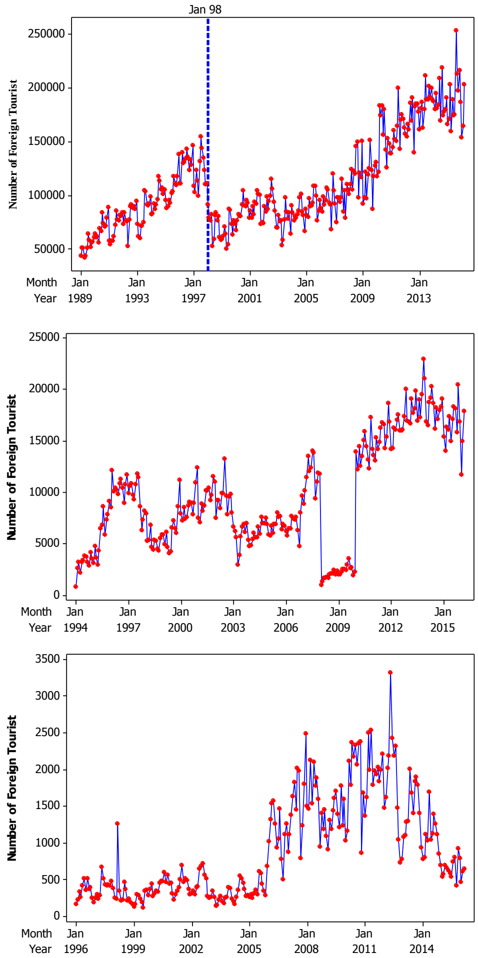


Figure 1. Time series plot of number of foreign tourist arrival at Soekarno Hatta, Juanda, and Adi Sumarmo airport

The first step of VARIMA modeling is identification the stationer of data. The multivariate cross-correlation function (MCCF) of actual data indicates non stationary pattern as displayed at Figure 2**.** Applying regular and seasonal differencing at lag 1 and 12, respectively, the MPCCF of data are shown at Figure 3. Thus the tentative VARIMA models that could be proposed is VARIMA ([12],1,0) (0,1,0)[12].

Variable/Lag	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
z1	+++	+++	+++	+++	+++	+++	+++	+++	.++	..+	...	...	---	---	---	---	---	---	---
z2	+++	+++	+++	+++	+++	+++	+++	+++	..+	...	...	--.	---	---	---	---	---	---	---
z3	+++	+++	+++	+++	+++	+++	.++	..+	...	...	--.	---	---	---	---	---	---	---	---

Variable/Lag	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
z1	---	---	---	---	---	---	---	---	..-	..-	...	++.	+..	+++	+++	+++	+++	+++	+++
z2	---	---	---	---	---	---	---	---	..-	...	+..	++.	++.	+++	+++	+++	+++	+++	+++
z3	---	---	---	---	---	---	..-	...	...	+..	++.	++.	+++	+++	+++	+++	+++	+++	+++

Variable/Lag	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
z1	+++	+++	+++	+++	+++	+++	..+	...	...	...	--.	--.	---	---	---	---	---	---	---
z2	+++	+++	+++	+++	+++	+++	.++	...	...	...	--.	--.	---	---	---	---	---	---	---
z3	+++	+++	+++	+++	+++	..+	..+	...	-..	--.	---	---	---	---	---	---	---	---	---

Figure 2. MCCF of actual data

Variable/Lag	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
z1	-..	...	...	...	...	...	..-	...	...	...	...	--.	.+.	--.	+..	...	...	...
z2	...	...	...	...	...	...	...	...	...	..-	-..	.-.	...	...	...	...	...	...
z3	..-	..-	...	...	+..	...	...	-..	...	...	-..	..-	...	...	...	...	...	...

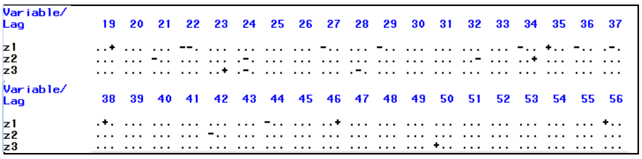


Figure 3. MPCCF after differencing lag 1 and lag 12

The result of parameters estimates significance test are shown at Table 1. The highlighted *p*-value means parameters are not significance because that *p*-values are greater than level of significance 0.05. Thus, VARIMA model for forecasting number of foreign tourist arrival at Soekarno Hatta, Juanda, and Adi Sumarmo can be written as follow:

$$\begin{bmatrix} Z_{1t}^* \\ Z_{2t}^* \\ Z_{3t}^* \end{bmatrix} = \begin{bmatrix} -0.273 & 1.218 & 0 \\ -0.019 & -0.248 & 0 \\ 0 & 0.019 & -0.360 \end{bmatrix} \begin{bmatrix} Z_{1,t-12}^* \\ Z_{2,t-12}^* \\ Z_{3,t-12}^* \end{bmatrix}$$

for $i = 1,2,3$. It is important to note that:

$$Z_{i,t}^* = (1-B)(1-B^{12})Z_{i,t}$$
$$Z_{i,t}^* = (1-B-B^{12}+B^{13})Z_{i,t}$$
$$Z_{i,t}^* = Z_{i,t} - Z_{i,t-1} - Z_{i,t-12} + Z_{i,t-13}$$

In addition, VARIMA model for number of foreign tourist arrival at each arrival gate can be written as follow:

a. Soekarno Hatta arrival gate

$$\begin{aligned}\widehat{Z_{1,t}} = Z_{1,t-1} + Z_{1,t-12} - Z_{1,t-13} \\ - 0.273(Z_{1,t-12} - Z_{1,t-13} \\ - Z_{1,t-24} + Z_{1,t-25}) \\ + 1.218(Z_{2,t-12} - Z_{2,t-13} \\ - Z_{2,t-24} + Z_{2,t-25})\end{aligned}$$

b. Juanda arrival gate

$$\begin{aligned}\widehat{Z_{2,t}} = Z_{2,t-1} + Z_{2,t-12} - Z_{2,t-13} \\ - 0.019(Z_{1,t-12} - Z_{1,t-13} \\ - Z_{1,t-24} + Z_{1,t-25}) \\ + 1.248(Z_{2,t-12} - Z_{2,t-13} \\ - Z_{2,t-24} + Z_{2,t-25})\end{aligned}$$

c. Adi Sumarmo arrival gate

$$\begin{aligned}\widehat{Z_{3,t}} = Z_{3,t-1} + Z_{3,t-12} - Z_{3,t-13} \\ + 0.019(Z_{2,t-12} - Z_{2,t-13} \\ - Z_{2,t-24} + Z_{2,t-25}) \\ - 0.36(Z_{3,t-12} - Z_{3,t-13} - Z_{3,t-24} \\ + Z_{3,t-25})\end{aligned}$$

The proposed VARIMA model was used to forecast 27 steps ahead as forecasting result of number of foreign tourist arrival at Soekarno Hatta, Juanda, and Adi Sumarmo for January 2014 until March 2016. These forecasting results can be compared with testing data as displayed at Figure 4. That also show prediction interval of VARIMA model. Both point forecasting result and testing data are inside prediction interval. The longer step ahead to forecast, the larger variance error of VARIMA model.

TABLE 1.
ESTIMATION AND PARAMETER SIGNIFICANCE TESTOF VARIMA MODEL

Location	Parameter	Estimate	SE	T	p-value
Soekarno Hatta (Z_1)	ϕ_{11}^{12}	-0.273	0.073	-3.75	0.0002
	ϕ_{12}^{12}	1.218	0.561	2.17	0.0311
	ϕ_{13}^{12}	6.721	3.604	1.86	0.0638
Juanda (Z_2)	ϕ_{21}^{12}	-0.019	0.009	-2.21	0.0284
	ϕ_{22}^{12}	-0.248	0.069	-3.59	0.0004
	ϕ_{23}^{12}	0.535	0.444	1.20	0.2305
Adi Soemarmo (Z_3)	ϕ_{31}^{12}	-0.001	0.001	-0.74	0.4628
	ϕ_{32}^{12}	0.019	0.011	1.76	0.0806
	ϕ_{33}^{12}	-0.360	0.072	-5.01	0.0001

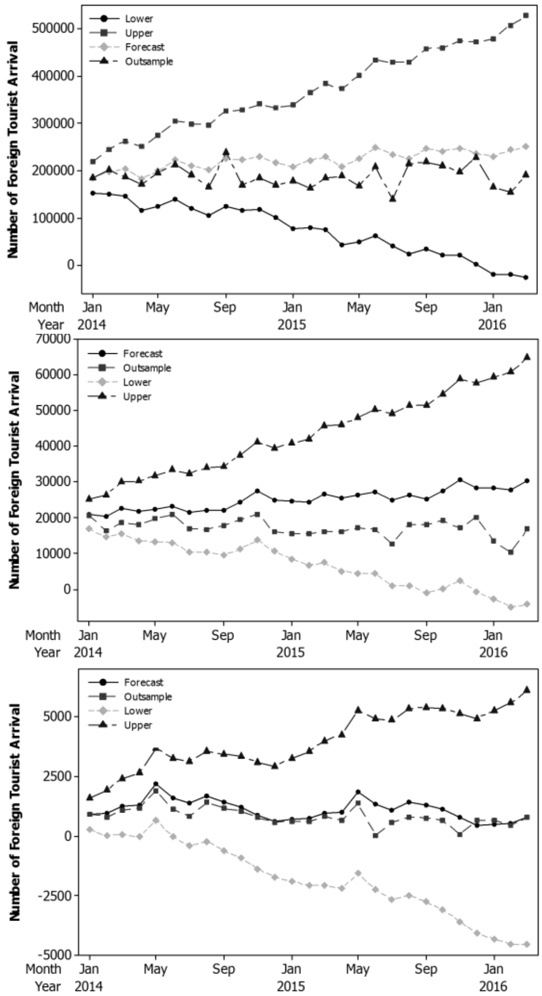


Figure 4. Comparisons of forecasting result, testing data, and prediction interval of VARIMA model at a) Soekarno-Hatta, b) Juanda, c) Adi Sumarmo.

Residual of VARIMA ([12],1,0) $(0,1,0)^{12}$ model then was used to build prediction interval using percentile error bootstrap approach. Comparisons of testing data, forecasting result, and prediction interval using percentile error bootstrap based on VARIMA model can be seen at Figure 5. Not all of the testing data are inside of prediction interval. Thus, the prediction intervals have not covered testing data perfectly. Moreover, it is important to note that each part of Figure 4 and Figure 5 have the same scale. As a result, prediction intervals at Figure 4 and Figure 5 can are comparable. Although prediction intervals using percentile error bootstrap based on VARIMA models have not covered all testing data perfectly, it has more stable result than ones produced by standard VARIMA model.

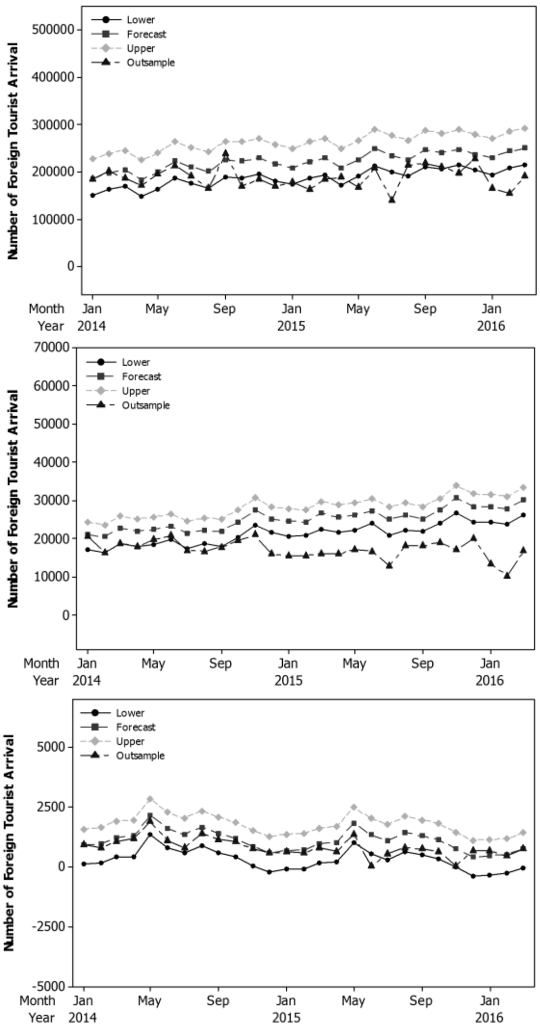


Figure 5. Comparisons of forecasting result, testing data, and prediction interval of percentile error bootstrap based on VARIMA model at a) Soekarno-Hatta, b) Juanda, c) Adi Sumarmo.

Point forecast accuracy of VARIMA ([12],1,0) $(0,1,0)^{12}$ model was evaluated using both training and testing symmetric MAPE (sMAPE). This sMAPE accuracy can be known at Table 1. Juanda has the highest sMAPE accuracy because there are significance outliers. In addition, testing sMAPE accuracy is smaller than testing sMAPE accuracy except for Soekarno-Hatta.

TABLE 1.
SMAPE ACCURACY OF VARIMA MODEL

sMAPE	Soekarto-Hatta	Juanda	Adi Sumarmo
Training	0.1200	1.7360	0.4152
Testing	0.1735	0.3722	0.3712

Not only point forecast but also prediction intervals need to be evaluated using certain measurement accuracy. Both prediction intervals of pure VARIMA model and prediction intervals of percentile error bootstrap were evaluated using percentage error as shown at Table 2. The percentage of number of testing data that outside the prediction intervals called percentage error. Actually, prediction intervals of pure VARIMA model have smaller percentage error than prediction intervals of percentile error bootstrap based on VARIMA model. It might be caused by some outliers at monthly number of foreign tourist arrivals data.

TABLE 2.
PERCENTAGE ERROR ACCURACY OF PREDICTION INTERVAL

Percentage Error	Soekarno-Hatta	Juanda	Adi Sumarmo
Pure VARIMA	0.0000	0.0000	0.0000
Percentile Error Bootstrap Based on VARIMA	0.4815	0.8148	0.0370

The proposed VARIMA model also used to forecast 21 steps ahead as forecasting result of number of foreign tourist arrival at Soekarno Hatta, Juanda, and Adi Sumarmo for April 2016 until December 2017 based on full data. These forecasting results are shown at Table 3**.** In addition, Figure 6 shows the point forecast and prediction intervals of monthly number of foreign tourist arrivals for three arrival gates completely.

TABLE 3.
FORECASTING RESULT OF NUMBER OF FOREIGN TOURIST ARRIVAL

Month	Year	Soekarno-Hatta	Juanda	Adi Sumarmo
April	2016	164177	18390	618
May	2016	185208	18947	797
June	2016	179675	17737	581
July	2016	179476	17775	599
August	2016	250229	18613	774
September	2016	211771	18171	773
October	2017	193687	16975	450
November	2017	220495	20342	722
December	2017	191391	18745	573
January	2017	156931	14067	298
February	2017	174167	16011	457
March	2017	211807	18643	526
April	2017	173674	18842	502
May	2017	196383	19822	603
June	2017	189814	18069	455
July	2017	187065	18709	424
August	2017	258437	19769	611
September	2017	218851	19355	627
October	2017	203411	17919	300
November	2017	225571	21414	629
December	2017	199113	19255	514

Residuals of proposed VARIMA model then used to build prediction intervals of percentile error bootstrap for April 2016 until December 2017. Point forecast and prediction intervals of monthly number of foreign tourist arrivals for Soekarno Hatta, Juanda, and Adi Sumarmo are presented at Figure 7.

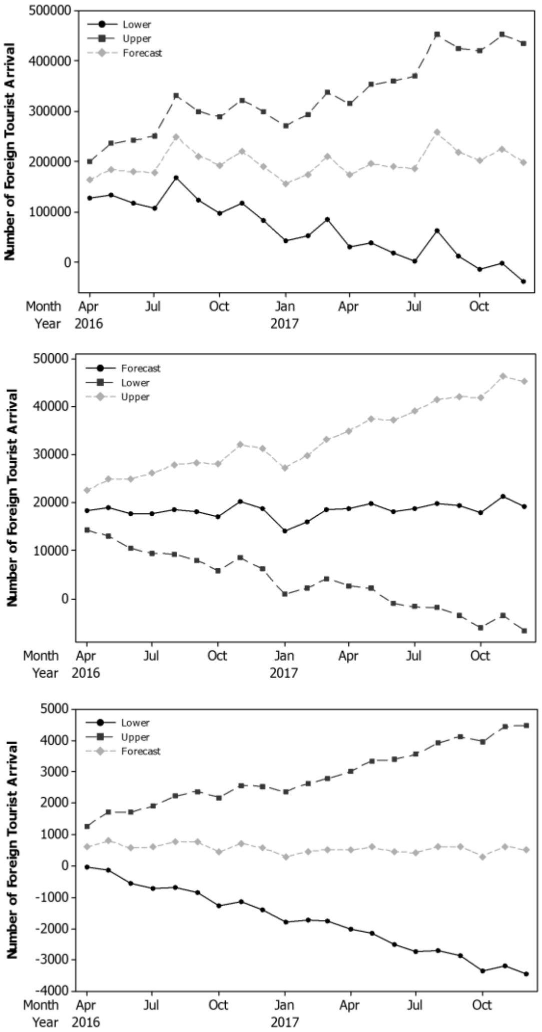
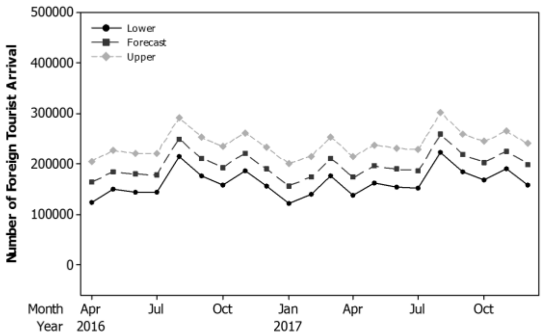


Figure 6. Forecasting result and prediction interval of VARIMA model at a) Soekarno-Hatta, b) Juanda, c) Adi Sumarmo.

It is important to know that both part of prediction intervals of Figure 6 and Figure 7 use equal scale. Longer step of forecasting, the wider prediction intervals produced by VARIMA model. Furthermore, prediction intervals using percentile error bootstrap based on VARIMA models are more stable while used to forecast for longer horizon.



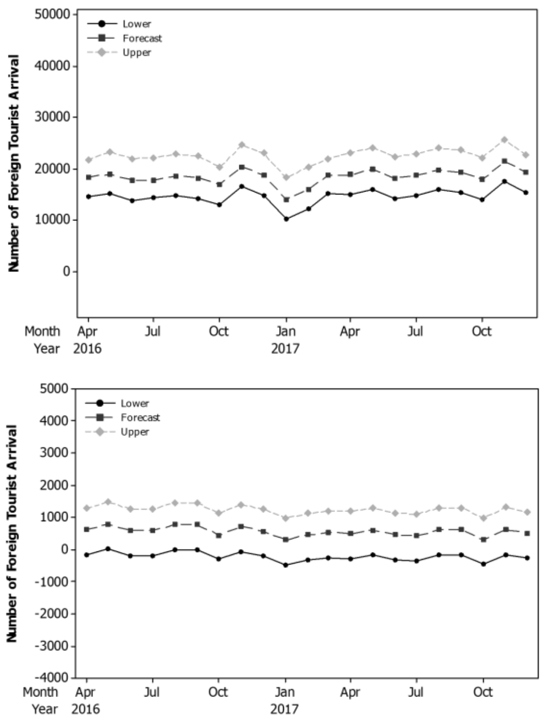


Figure. 7. Forecasting result and prediction interval of percentile error bootstrap based on VARIMA model at a) Soekarno-Hatta, b) Juanda, c) Adi Sumarmo.

V. CONCLUSION

The best model to forecast number of foreign tourist arrivals at Soekarno Hatta, Juanda, and Adi Sumarmo airport is VARIMA ([12],1,0)(0,1,0)^{12}. Longer horizon of forecasting, the variance error of VARIMA models get larger, therefore resulted in wider prediction interval. To overcome this drawback, the prediction interval using percentile error bootstrap based on VARIMA model was employed in this research. The empirical result of this approach outperformed the previous method.

VARIMA model is the extention of ARIMA model in multivariate problem. Both ARIMA and VARIMA model are established on tight residual assumption. The high percentage error of prediction interval of percentile error bootstrap might be caused from significance outlier at monthly number of foreign tourist arrivals. Hence, VARIMA modeling with exogenous variable or outlier condition might be useful for future research.

REFERENCES

[1] Republika, Monday 26 Januari 2015: Arif Yahya, Menteri Pariwisata RI: "Kita akan Siapkan Wisata Khas"

[2] W. W. S. Wei, "Time Series Analysis Univariate and Multivariate Methods Second Edition", New York: Pearson Addison Wesley, 2006.

[3] R. J. Hyndman, B. A. Koehler, D. R. Snyder and S. Grose, "A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods." *International Journal of Forecasting*, 18, 439-454, 2002.

[4] Badan Pusat Statistik [Online] Available at www.bps.go.id