ICEBIT
The 5th International Conference On Economics , Business and Information Technology
E-ISSN 3025-4892
August 24, 2024 pp. 323-333

PREDICTION NUMBER OF TOURIST ARRIVALS & PASSENGERS AT THE AIRPORT USE THE TIME MODEL FORCESTING SERIES

Abdul Wahid
Institut Teknologi dan Sains Mandala
Bondowoso, East Java
abdulwahid@itsm.ac.id
085257076330

Angga Ade Permana
Institut Teknologi dan Sains Mandala
Jember, East Java
angga@itsm.ac.id
081252554205

Syarif Aminul Khoiri
Unversitas Ibrahimy
Situbondo, East Java
syarifak@ibrahimy.ac.id
081249058246

ABSTRACT

The progress of tourism in today's life is very common in every country in the world. Improving the quality of tourism is very important for every country, considering that tourism is one of the sources of state income. Therefore, one of the most important parameters for this is knowing the number of visitors or tourists at any time, and being able to utilize existing historical data to predict the number of tourists in the future. In this research, prediction/forecasting of the number of tourists and passengers at the airport will be carried out using the Seasonal Auto Regressive Integrated Moving Average (SARIMA), Long-short Term Memory (LSTM), and Prophet methods on two time series datasets with monthly frequency. Of the three forecasting models, the results of each were obtained and then compared, the SARIMA model was the model with the best performance with the smallest RMSE and MSE values.

Keywords : *Time Series, Univariate, Forecasting*.

1. INTRODUCTION

Progress in the world of tourism is very common in this era of super sophisticated technology. Every country will always try to improve the quality of its tourism for every tourist who visits their country. One parameter that is very important for this is knowing the number of visitors or tourists at any time [1]. Analysis of data on the number of visitors will be very useful for the authorities to improve and improve the quality of tourism. Data on the number of tourists collected from a certain time period can be used to predict future events based on events in the past, based on this data [2]. This will be very useful for the authorities in making decisions to improve the quality of tourism in their country.

These predictions can be applied to time series data on the number of tourists in a certain period, which can be daily, weekly, monthly or yearly. There are many time series data prediction methods or models that can be used to carry out forecasting (forcesting) in the future time range [1]. Like research [3] [4], which uses a combined method or what is called a hybrid method between the Autoregressive integrated moving average (ARIMA) and Artificial Neural Network (ANN) methods. There are 3 different datasets that are used to carry out experiments to demonstrate the hybrid model [3], namely: Wolf's sunspot data, Canadian lynx (Iynx cat) data, and pound sterling and US dollar exchange rates. The results of this experiment were then compared with the results of non-hybrid methods, namely ARIMA and ANN. The hybrid method or combination of ARIMA and ANN succeeded in getting
the smallest Mean Square Error (MSE), this indicates that the hybrid model has better performance than the ARIMA and ANN methods.

This is different from research [2] which predicted air pollution levels in several Indian cities based on historical data on air pollution levels using the Seasonal Autoregressive Integrated moving average (SARIMA) and Prophet methods or models. In this research, the two models have good performance in predicting air pollution levels, however the result of comparing the two methods is that the Prophet method has better performance because it managed to get smaller RMSE and MSE values than SARIMA.

Furthermore, the Long Short Term Memory (LSTM) model which is part of the Recurrent Neural Network (RNN) is used in research [5]. The LSTM method has often been applied to make predictions in various different areas. In this research it is used to predict the Chinese stock market. The dataset used for training data is 900,000 and test data is 311,361. The LSTM method designed in this research succeeded in increasing the stock prediction accuracy value from 14.2% to 27.2%.

From the various methods of predicting or forecasting time series data, it can be concluded that there are many areas that can be predicted by these methods. One of them is in the world of tourism as discussed in this research. The author will forecast/predict the number of tourists and passengers at the airport using the SARIMA, LSTM, and Prophet methods. There are two datasets that will be used, namely data on the number of Indonesian tourists in Taiwan and the number of arriving passengers at Singapore's Changi Airport.

All of these datasets are time series data with a monthly frequency. At the end of this research will be carried out Compare the results of the three forecasting models based on the RMSE and MSE values of each model, in order to find out the performance of the best and most appropriate model for the time series data.

2. TIME SERIES FORCESTING MODEL

A time series is a series of observational data sequentially based on time [6]. Time series models have been widely implemented in various fields, for example engineering, geophysics, economics, agriculture and medicine. Time series forecasting is the process of forecasting or predicting future events based on events in the past using historical data in the nature of a time series [7]. The following will explain several prediction methods (forecasting) that will be used in this research::

2.1. Arima & Sarima

ARIMA (Auto Regressive Integration Moving Average) or what is usually called Box-Jenkins is a method for predicting time series data. The ARIMA method is usually used to predict short term events [8] [2]. The ARIMA method can be defined as follows:

a. Autoregressive (AR)

The autoregressive method or abbreviated as AR is a method that says that events in the current period are influenced by event data in the previous period [8]. Autoregressive method with order p (AR(p)) or order (p,0,0) in the ARIMA model can be expressed as follows:

$$Yt = \alpha 0 + \theta 1 Yt-1 + \ldots + \theta p Yt-p + et \quad (1)$$

$Yt = nilai\ observasi$ ke (t)
$\alpha 0 = konstanta$
$\theta p = parameterautoregressivekep$
$et = nilai$ error ke (t)

b. Moving Avarage (MA)

Moving average (MA) is a model that states that all time data values are related to the prediction error values in the present and the past respectively. The moving average model with order q is abbreviated as MA(q) or ARIMA (0,0,q) [7]. The basic equation of this model can be stated as follows:

$$Yt = \theta 0 + \theta 1 et-1 + \ldots + \theta q\ et-q \quad (2)$$

$\theta 0 = konstanta$

$\theta q = moving\ average\ ke\ q$

$et-p = nilai$ error ke $t - k$

c. ARMA

The ARMA model or Auto Regressive Model Average is a combined model of AR and MA. So it can be said that current event data is influenced by previous period data and also data errors in the past. The ARMA model with orders p and q is synchronized ARMA (p,q) or ARIMA (p,0,q) [8]. The equation of this model is as follows:

$$Y_t = \alpha 0 + \theta 1\, e_t-1 + \ldots + \theta_q e_t-q + \alpha 1\, Y_t-1 + \ldots + \alpha_p Y_t-p \quad (3)$$

d. ARIMA

Auto Regressive Integrated Moving Average (ARIMA) is the same as the ARMA model but there is an additional differencing method to make time series data into stationary data, this is done if the time series data is not stationary [3]. The ARIMA equation with the order (p,d,q) is as follows:

$$Yt = (1 + \alpha 1)\, yt-1 + \ldots + (1 + \alpha p)\, yp-1 + et + \theta 1\, et-1 + \ldots + \theta q\ et-q \quad (4)$$

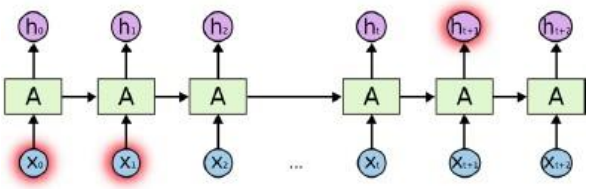
e. SARIMA

Seasonal Auto Regressive Integrated Moving Average is the same model as ARIMA, but there are additional seasonality parameters to it. This model is used when the observed time series data has a seasonal pattern (seasonality) that repeats itself over a certain time period [8]. The equation can be formed as follows:

$$ARIMA\ (p,d,q)(P,D,Q)s \quad (5)$$

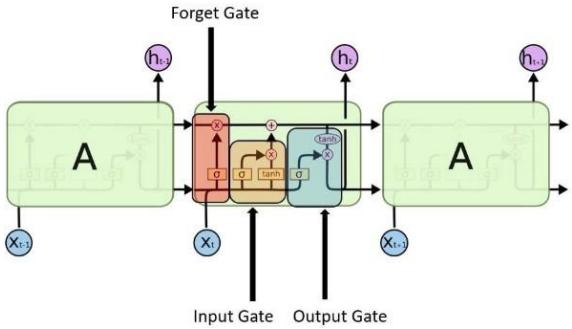
2.2. LTSM

Long Short-term Memory is a technique derived from the Recurrent Neural Network (RNN) model [3]. RNN is a model that has good performance for analyzing data on events that occur in sequential time, such as time series data [5]. Meanwhile, LSTM itself has its own advantages over simple RNN, as shown in Figure 1 below:



Gambar 1 Memory in RNN

In Figure 1 it is explained that the shortcomings of the RNN can be seen in the input, 1 which carries a very large range of information with t, 1+1, so that when the output Ht+1 requires input values that correspond to X1, X0 the RNN cannot learn to adjust the information because The old memory has been overwritten and replaced by data in the new memory over time. However, this problem can be resolved using the LSTM model which can manage the memory for each input by utilizing memory cells and gate units, namely Input Gate, Forget Gate, and Output Gate. As depicted in figure 2:



Gambar 2 LSTM Memory Cells and Gate Units

2.3. Prophet

Forecasting Model or Prophet prediction, is a prediction method developed by Facebook, which is available in R and Python. The prediction process is based on an additive model where non-linearity is relevant to the influence of seasonality such as yearly, weekly and daily, as well as the influence of the holiday season. This prophet model is very good for applying to time series data that has strong seasonality. The Prophet model can work well on data with missing values and can handle outliers well.

The following is the equation for the Prophet prediction model:

$$y(t) = g(t)+s(t) + h(t)+ \in t \quad (6)$$

Where the model parameters g(t), s(t), h(t), ÿt: are gradual linear curves for modeling non-linear changes in time lags in time series data, changes in time periods, effects of irregular time patterns such as holidays, error values result from changes that are not accommodated by the model sequentially. To apply the Prophet forcing model to time series data that has seasonality and want to predict based on seasonality, you can use the Fourier series which can create a flexible model. The effect of seasonality s(t) can be expressed in equation 7 [2]::

$$s(t) = \sum_{n=1}^{N} a_n \cos\left(\frac{2\Pi nt}{P}\right) + b_n \sin\left(\frac{2\Pi nt}{P}\right) \quad (7)$$

3. DATASETS

Dataset used The data set that will be observed to apply the forecasting model is the global data set obtained from www.kaggle.com , there are two time series data sets, namely;

a. Data on the number of visitors to Taiwan with monthly frequency from 2011 to 2018, there are 96 data records. In this dataset, researchers only selected the number of visitors from Indonesia, to further simplify the analysis process.

b. Data on the number of passenger arrivals at Singapore's Changi Airport, with monthly time frequency from 2009 to 2019, there are 132 data records.

Both datasets are univariate time series data that have a monthly time frequency.

4. DATA ANALYSIS

Before forecasting, data analysis is needed to gain insight from the time series data. The following will display a line plot diagram of the two datasets:

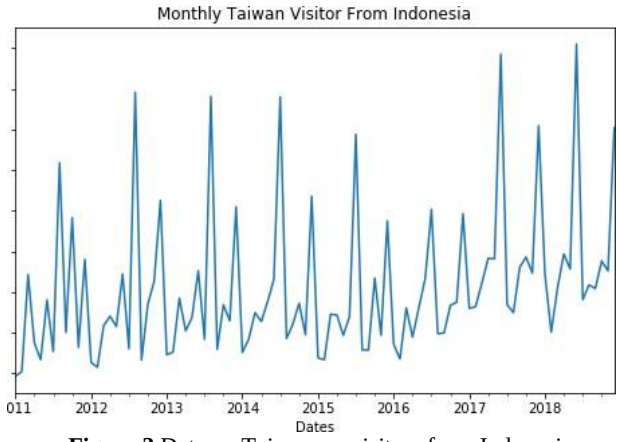


Figure 3 Data on Taiwanese visitors from Indonesia

From the line plot diagram, it can be seen that the number of visitors is quite fluctuating every year, and has almost a seasonal pattern every year. For greater clarity, the following is a line plot diagram with a sample period of only 3 years:

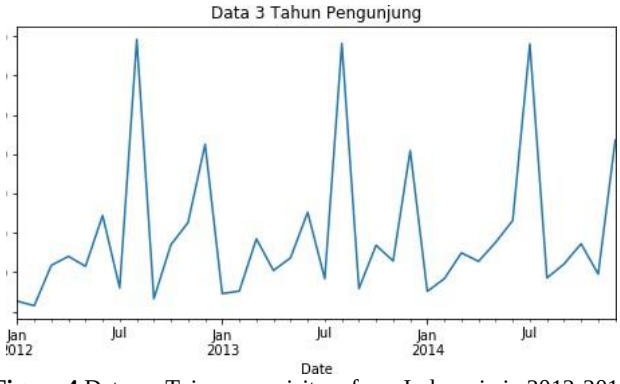


Figure 4 Data on Taiwanese visitors from Indonesia in 2012-2014

It is quite clear that fluctuations in the number of visitors experience the same seasonal pattern every year. Amount Visitors always increase in August and December, this is thought to be due to the end-of-semester school holidays and also the end-of-year holidays. However, the number of visitors will decline again in February and January. Every time series data must have at least these three elements; Trend, Seasonality, and Residual. The following is a line diagram of these three elements in the dataset of the number of visitors to Taiwan from Indonesia:

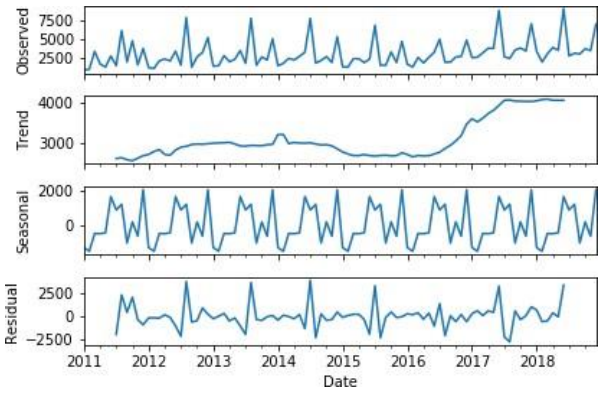


Figure 5 Trends, Seasonality and Residuals of Taiwanese Visitors from Indonesia

This data has a fairly stable trend in 2011-2016, and has an upward trend in 2016-2018. The seasonality of the data also looks quite significant.

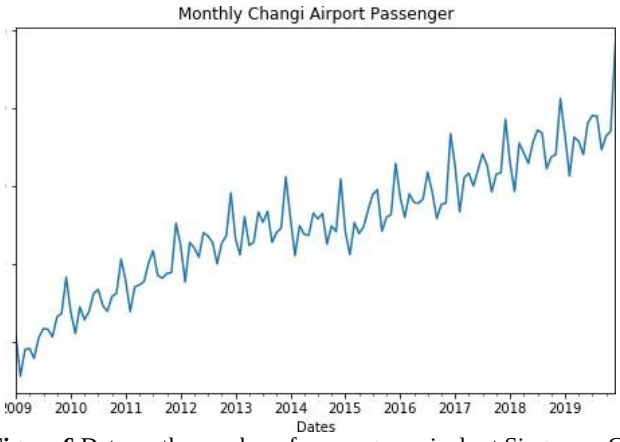


Figure 6 Data on the number of passenger arrivals at Singapore Changi Airport

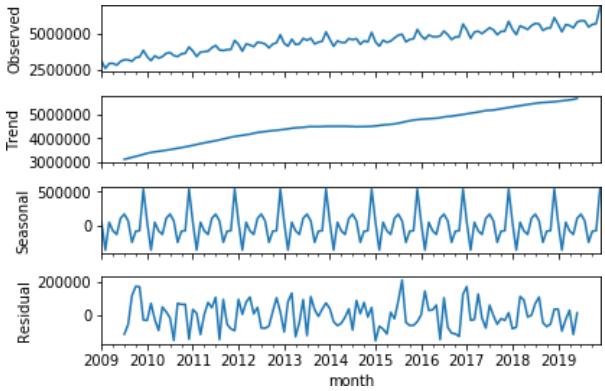


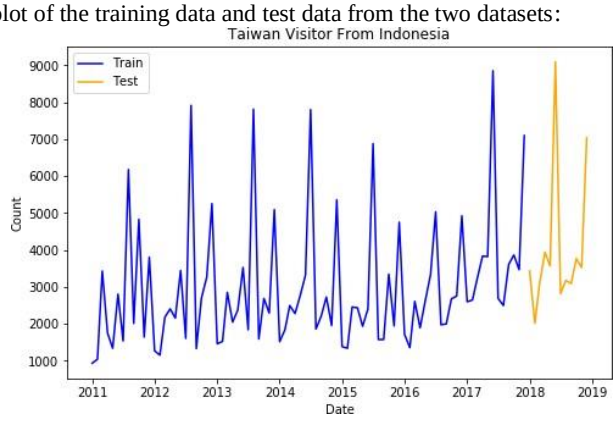
Figure 7 Trend, Seasonal and Residual Number of Passenger Arrivals at Singapore Changi Airport

Not much different from visitor data in Taiwan, in the dataset the number of passenger arrivals at Changi airport is also quite fluctuating. There is a seasonality pattern that occurs almost every year, the number of passenger arrivals increases in July and December, and decreases again in February. The trend is quite strong, here there is an upward trend

5. IMPLEMENTATION AND RESULTS OF TIME SERIES FORECASTING MODEL

stage a forcesting experiment will be carried out on the two datasets. The models that will be used are SARIMA, LSTM (RNN) and Prophet Forcesting Model. In this implementation, all datasets will be divided into two, namely training data and test data in order to validate the results of implementing the model on the training data.

For the dataset on the number of visitors to Taiwan from Indonesia, the training data is 84 (in the period 2011- 2017), and the test data is 12 (last 1 year, 2018). Meanwhile, for the dataset on the number of passenger arrivals at Changi Airport, the training data is 120 (in the 2009-2018 time period), and the test data is 12 (1 year last 2019). The following is a line plot of the training data and test data from the two datasets:



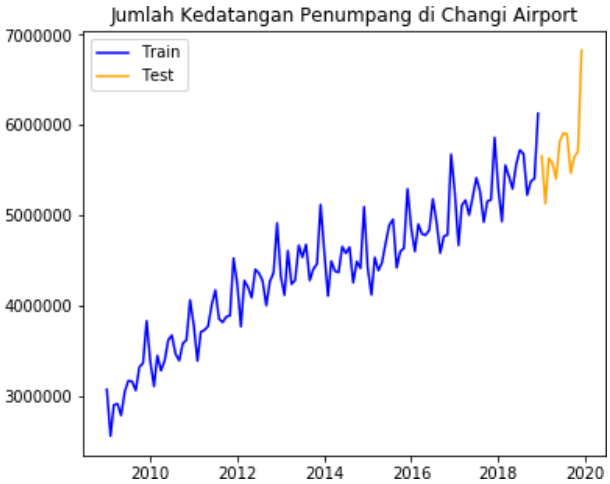


Figure 8 Training data and test data dataset of number of visitors to Taiwan from Indonesia

Figure 9 Training data and test data dataset for the number of passenger arrivals at Singapore Changi Airport

5.1 Implementation of SARIMA

In this Seasonal Auto Integrated Moving Average Model, the dataset must be stationary. If the data is not stationary, preprocessing can be carried out, namely using the differencing method. Therefore, to find out if the dataset is stationary or not, a stationary test can be carried out , one of which is the Dickey Fuller Statistical Test technique.

The following are the results of the Dickey Fuller Statistical Test ::

Tabel 1 ADF Test

Taiwan Visitor Dataset from Indonesia	
Dickey Fuller Statistical Test	0
P-value	1
Used Lags	12
Number of comments used	83
Critical Value (1%)	-3
Critical Value (5%)	-3
Critical Value (10%)	-3

Tabel 2 ADF Test

Dataset Number of Passenger Arrivals at Changi Airport Singapore	
Dickey Fuller Statistical Test	-0
P-value	1
Used Lags	12
Number of comments used	119
Critical Value (1%)	-3
Critical Value (5%)	-3
Critical Value (10%)	-3

Based on the ADF results, all ADF statistical results are greater than critical values, so it can be concluded that all datasets are non- stationary. The statistical test results still have to be matched or confirmed with visual techniques on the line diagram and experiments must be carried out on the Integrated (I/Differencing) input values when implementing the SARIMA model.

Next, an experiment will be carried out using the SARIMA method, with the ARIMA input sequence (p,d,q)(P,D,Q)s as explained in previous sub-chapter 11.1. The input values will be tested with different differencing values and then the RMSE and MSE results will be compared with the training data.

a. SARIMA model on a dataset of Taiwanese visitors from Indonesia.

Table 3 SARIMA Forecasting Result

(p,d,q)(P,D,Q)s	MSE	RMSE
(1,0,0)(1,1,1)12	319425,29	565,18
(1,1,0)(1,1,1)12	2926073,47	1710,58

From the experimental results on training data and then validation on test data, it was found that the ARIMA input value (1,0,0)(1,1,1)12 had the smallest MSE and RMSE values. The following is a line plot of forecasting results using the SARIMA model on training data:

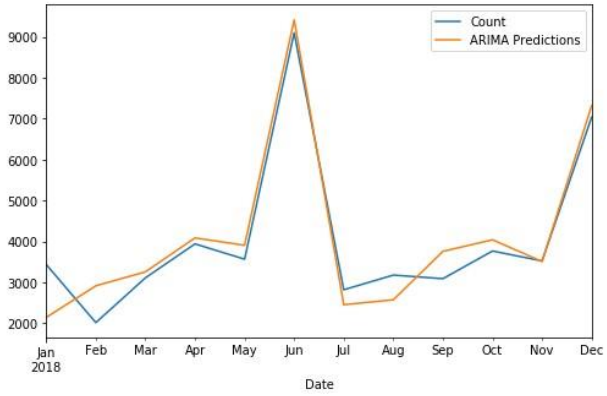


Figure 10 Comparison of test data with SARIMA prediction results (1,0,0)(1,1,1)12

b. SARIMA model on passenger arrival dataset at Changi Airpot.

Table 4 SARIMA Forecasting Result

(p,d,q)(P,D,Q)s	MSE	RMSE
(0,0,1)(2,1,2)12	29.657.018.261	172212,1316
(0,1,1)(2,1,2)12	23.452.430.508	153141,864

From the experimental results on training data and then validation on test data, it was found that the ARIMA input value (0,1,1)(2,1,2)12 had the smallest MSE and RMSE values. The following is a line plot of forecasting results using the SARIMA model on training data:

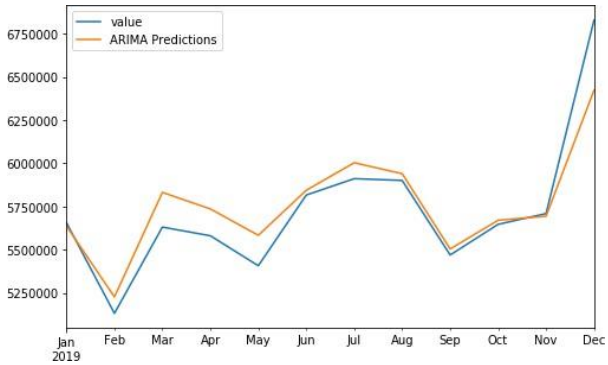


Figure 11 Comparison of test data with SARIMA prediction results (0,1,1)(2,1,2)12

5.2. LSTM RNN Implementation

a. Processing

Before testing the LSTM model, the dataset is first normalized to reduce the level of error values. The min-max scaling normalization technique will be used on both datasets. Min-max scaling is changing real data into interval range values [0,1]

b. Long-short Term Memory (LSTM) RNN

After carrying out preprocessing, a forecasting experiment will be carried out using the LSTM model. In LSTM here, it uses 1 hidden layer with 200 dimensions, and uses activation
function ReLu, and Epoch= 100.
Following are the MSE and RMSE results from LSMTM on the two respective datasets.

Table 5 LSTM Forecasting Result

LSTM: Taiwan Number of Visitors Dataset From Indonesia	
MSE	**RMSE**
1813932,153	1346,822985

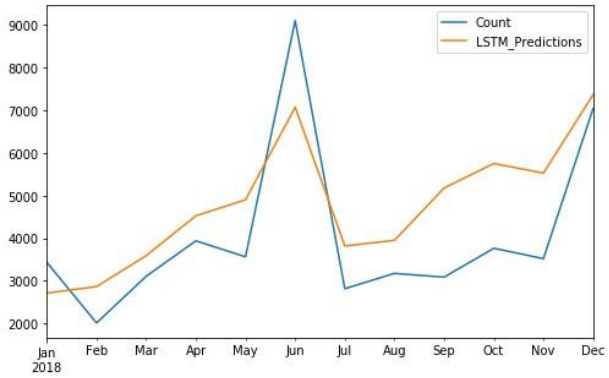


Figure 12 Comparison of test data with LSTM prediction results Taiwan Visitor Dataset

Table 6 LSTM Forecasting Results

LSTM: Arrival Count Dataset Passengers at Changi Airport	
MSE	**RMSE**
36.550.379.311	191.181

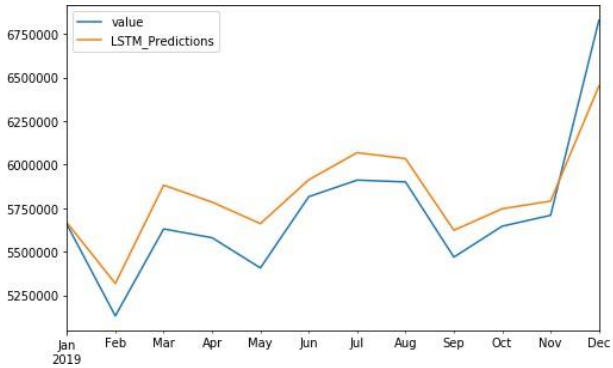


Figure 13 Comparison of test data with LSTM prediction results Changi Airport Arrival Dataset

5.3. Implementation of the Prophet Forecasting Model

The following are the MSE & RMSE results from forecasting experiments using the Prophet model, with input into the equations explained in sub-chapter II.3.

Table 7 Prophet Forecasting Result

Prophet: Number of Visitors Dataset Taiwanese from Indonesia	
MSE	**RMSE**
3003489,275	1733,057782

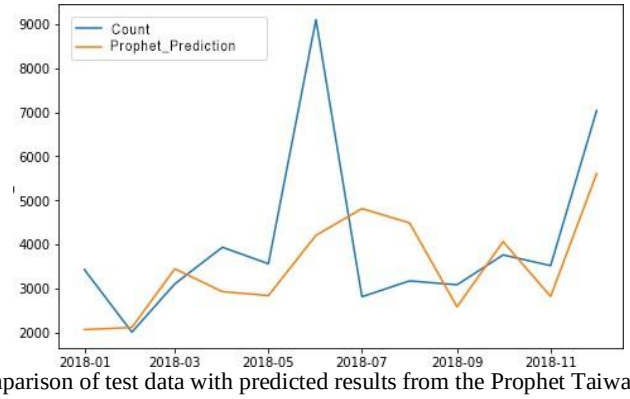


Figure 14 Comparison of test data with predicted results from the Prophet Taiwan Visitor Datase

Table 8 Prophet Forecasting Result

Prophet: Arrival Count Dataset Passengers at Changi Airport	
MSE	**RMSE**
24.558.451.552	156711,3638

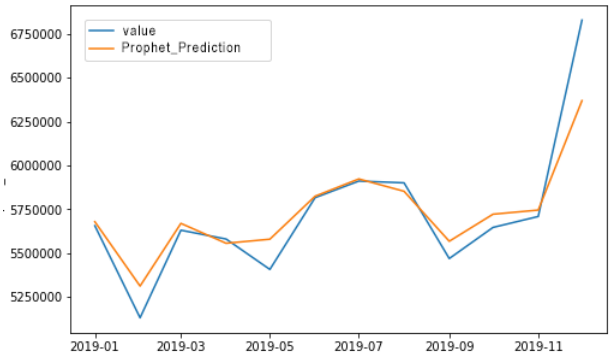


Figure 15 Comparison of test data with Prophet Changi Airport Arrival Dataset prediction results

5.4. Implementation Result

The following are comparison results of the results of all experiments on three forecasting models, namely SARIMA, LSTM, and Prophet:

Table 9 Comparison of Three Forecasting Models

Dataset on the Number of Taiwanese Visitors from Indonesia		
Model	**RMSE**	**MSE**
SARIMA	565	319.425
LSTM	1.347	1.813.932
Prophet	1.733	3.003.489

Table 10 Comparison of Real Data and Forecasting Results on Taiwan Visitor Test Data from Indonesia

Month	Real	ARIMA	LSTM	Prophet
01/18	3434	2142	2715	2073
02/18	2014	2914	2867	2120
03/18	3109	3255	3592	3450
04/18	3939	4084	4525	2933
05/18	3565	3904	4902	2843
06/18	9101	9415	7071	4209
07/18	2817	2452	3819	4815
08/18	3176	2571	3951	4491
09/18	3088	3756	5177	2588
10/18	3765	4039	5752	4067
11/18	3522	3508	5526	2824
12/18	7039	7316	7363	5608

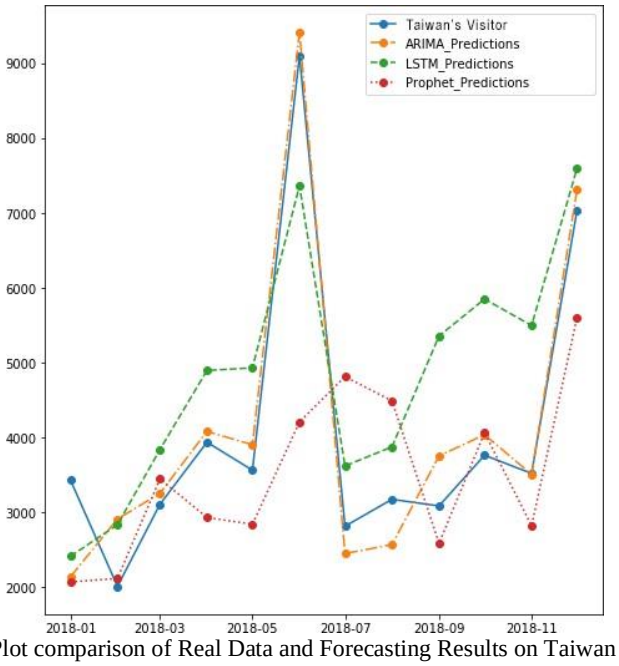


Figure 16 Line Plot comparison of Real Data and Forecasting Results on Taiwan Visitor Test Data from Indonesia

Table 11 Comparison of Three Forecasting Models

Dataset Number of Passenger Arrivals at Changi Airports		
Model	**RMSE**	**MSE**
SARIMA	153.142	23.452.430.508
LSTM	191.181	36.550.379.311
Prophet	156.711	24.558.451.552

Table 12 Comparison of Real Data and Forecasting Results on Changi Airport Arrival Number Test Data Singapore

Month	Real	ARIMA	LSTM	Prophet
01/19	5.656.076	5.636.642	5.648.514	5.679.077
02/19	5.131.807	5.226.755	5.366.953	5.312.444
03/19	5.630.780	5.831.258	5.866.676	5.669.597
04/19	5.580.503	5.735.820	5.852.081	5.556.350
05/19	5.407.308	5.583.248	5.727.086	5.579.520
06/19	5.816.089	5.843.826	5.967.431	5.824.505
07/19	5.910.782	6.002.966	6.108.785	5.922.897
08/19	5.900.629	5.939.459	6.079.748	5.852.532
09/19	5.469.342	5.504.248	5.687.570	5.567.841
10/19	5.646.643	5.670.161	5.719.881	5.721.932
11/19	5.708.993	5.693.691	5.765.265	5.745.476
12/19	6.827.843	6.422.891	6.426.182	6.368.971

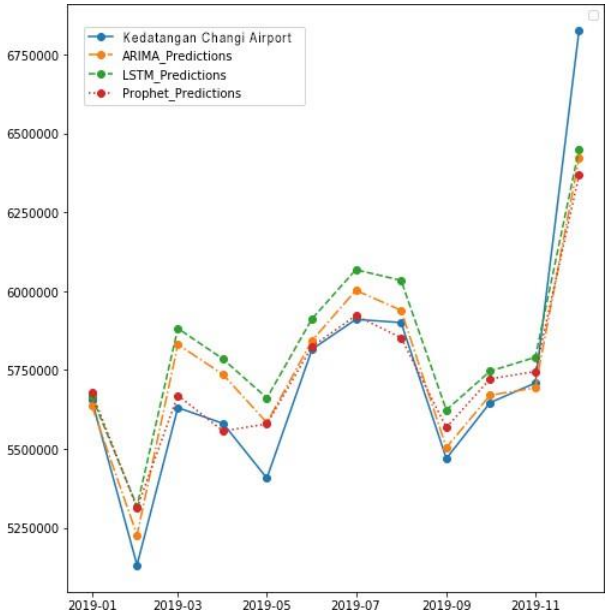


Figure 17 Line Plot comparison of Real Data and Forecasting Results on Changi Arrival Number Test Data Singapore Airport

6. CONCLUSION

Prediction or forcing on both time series datasets; the number of visitors to Taiwan from Indonesia, and the number of passenger arrivals at Changi Airport using three forecasting models, namely SARIMA, LSTM, and Prophet, it can be seen from the results of the comparison of RMSE and MSE values tested on test data, the SARIMA model has the best performance among the three this model, with the smallest RMSE and MSE values.

7. REFERENCES

[1] L. C. Jason, L. Gang, C. W. Doris and S. Shujie, "Forecasting Seasonal Tourism Demand Using a Multiseries Structural Time Series Method," Journal of Travel Research, pp. 1-12, 2017.

[2] R. S. K. Krishna, S. B. Korra, K. D. Santosh and A. Abhirup, "Time Series based Air Pollution Forecasting using SARIMA and Prophet Model," in Association for Computing Machinery, Singapore, 2019.

[3] G. P. Zheng, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing 50 (2003) 159 – 175, vol. 50, pp. 159-175, 2003.

[4] D. O. Faruk, "A hybrid neural network and ARIMA model for water quality time series prediction," Engineering Applications of Artificial Intelligence, no. 23, p. 586–594, 2010.

[5] C. Kai, Z. Yi and D. Fangyan, "A LSTM-based method for stock returns prediction: A case study of China stock market," in 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, USA, 2015.

[6] B. Kasun, B. Christoph and S. Slawek, "Forecasting Across Time Series Databases using Recurrent Neural Networks on Groups of Similar Series: A Clustering Approach," in Elsevier, 2019.

[7] C. Anila, H. M. Raviraj and G. Varghese, "Modelling and Forecasting Bus Passenger Demand using Time Series Method," in International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2018.

[8] L. Ziyu, B. Jun and L. Zhiyin, "Passenger Flow Forecasting Research for Airport Terminal Based on SARIMA Time Series Model," in MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing, 2012.