Neural Network Model Based on Data Preprocessing Technique for Foreign Tourists Prediction

Purwanto[1, a)], Sunardi[1, b)], and Fenty Tristanti Julfia[1, c)]

[1] *Universitas Dian Nuswantoro, Semarang, Central Java, Indonesia*

.

[a)]Corresponding author: purwanto@dsn.dinus.ac.id
[b)]sunardi@dsn.dinus.ac.id
[c)]fentytristanti@gmail.com

Abstract. There are various ways done by the government to increase regional income. One of the sectors to increase regional income is tourism sector. The uncertain arrival of foreign tourists makes it difficult for government to predict the number of foreign tourists. The prediction of foreign tourists is very important in assisting decision-making related to regional income. The prediction accuracy using a linear model has limitations in dealing with non-linear data. Thus, a reliable time series prediction model, especially in the field of tourism, is needed. This research proposes soft computing model that is neural network model based on data preprocessing technique for prediction of foreign tourists in Central Java Province, Indonesia. The data of this study are taken from the Department of Youth, Sports and Tourism of Central Java Province, to evaluate the proposed model. The results of this study indicate that neural network model has a better prediction performance in predicting foreign tourists. The results of this study prove that neural network model with NN (2-4-1) has better performance compared to linear regression (trend), moving average, single exponential smoothing, double exponential smoothing, triple exponential smoothing, and ARIMA models.

INTRODUCTION

The tourism industry plays an important role in increasing the country's foreign exchange earnings, which encourages every country to bring in international tourists (foreign tourists). According to the United Nations World Tourism Organization (UNWTO) [1], the world tourism organization, international tourist arrivals worldwide are expected to increase by 3.3% annually from 2010 to 2030. By 2030, the number of foreign tourists is expected to reach 1.8 billion. In 2013, foreign tourist arrivals grew 5.0% and reached 1.087 million. Asia and the Pacific recorded the highest growth of 6.0%. The UNWTO also reported that international tourism receipts in 2013 reached US $ 1,159 billion and US $ 1,078 billion in 2012.

In Indonesia, the number of foreign tourist arrivals reached 8.80 million in 2013, an increase of 9.42% when compared to the number of visits in 2012. The expenditures of foreign tourists per visit in 2013 averaged US $ 1,142.24 also increased by 0.74% compared to the average of foreign tourist expenditure per visit in 2012, which amounted to US $ 1,133.81 [2].

Many models of predictions with different concepts have been proposed by many researchers. Researchers have done their studies to develop better prediction models to improve prediction accuracy. The accuracy of the prediction model depends on the model and the complexity of the data. Therefore, it is very important to determine the best prediction model based on these data. According to Hanke and Wichern [3], there are two kinds of approaches in predictive models; they are casual prediction model and time series prediction model. The time series model determines future trends based on past values.

Statistical techniques such as linear regression, moving average, exponential smoothing, and autoregressive integrated moving average (ARIMA) are linear methods that can be used to predict time series. Statistical techniques have been used by many researchers and are conducted in various fields. Law Researcher [4] has used the moving

Human-Dedicated Sustainable Product and Process Design: Materials, Resources, and Energy
AIP Conf. Proc. 1977, 060006-1–060006-7; https://doi.org/10.1063/1.5043018
Published by AIP Publishing. 978-0-7354-1687-1/$30.00

average model for financial crisis. Similarly, exponential smoothing methods such as single exponential smoothing and double exponential smoothing have been applied for predictions in various fields [5].

In recent decades, researchers have done much research on international tourism predictions in the field of tourism. Most researches on tourism prediction use time series data. Gustavsson and Nordstrom [6] have implemented the ARMA model for predicting monthly tourist time series. The ARMA model has also been used to predict the time series of monthly tourist visits to Australia [7]. Similarly, the ARIMA model has been used for tourism predictions, such as [7, 8, 9]. Loganathan, *et al.* [10] have applied the SARIMA method to predict the demand of foreign tourism in Malaysia.

In the real world, many phenomena are non-linear (i.e., the relationship between past and current events is non-linear). The accuracy obtained using linear prediction models is not high in handling non-linear data [11]. Therefore, the linear time series model is not suitable for this case. For predictions in non-linear data, many researchers have implemented non-linear models. One of the prediction models that can handle non-linear data is Neural Network model. Niskaa *et al.* have applied neural networks to predict air pollution. They prove that Neural Network model gives better results [12].

This study proposes Neural Network based on data preprocessing technique to predict foreign tourists. To obtain better result, an experiment is performed by using various input based on data preprocessing technique to determine best configuration of Neural Network.

METHODOLOGY

In this research, the following steps are taken:

Data Collection

In this study, the data of foreign tourists in Central Java Province, Indonesia were collected from the Department of Youth, Sports and Tourism. The Foreign tourists' time series data were collected from January 1991 until December 2013. In other words, the data of foreign tourists consist of 276 months.

Data Preprocessing

The first step in this initial data processing is to check the missing value, and then change the univariate time series data into multivariate data. The univariate data are converted according to the pattern of data such as the pattern in the following table [13].

TABLE 1. Converting univariate data into multivariate data for monthly foreign tourists

Pattern	Input	Output/ Target
1	$y_1, y_2, y_3, y_4, \ldots, y_p$	y_{p+1}
2	$y_2, y_3, y_4, y_5, \ldots, y_{p+1}$	y_{p+2}
3	$y_3, y_4, y_5, y_6, \ldots, y_{p+2}$	y_{p+3}
...	...	....
m-p	$y_{m-p}, y_{m-p+1}, y_{m-p+2}, \ldots, y_{m-1}$	y_{m-p}

Furthermore, the pattern of data that has been formed is normalized data by using the formula:

$$x' = \frac{(x - \min)(\text{newmax} - \text{newmin})}{(\max - \min)} + new\min \tag{1}$$

Where x'= new data, x = old data, min = minimum data, max = maximum data, newmax = new maximum data and newmin = new minimum data. In this studi, we use newmax = 1, and newmin = 0.

Experiments using Statistical Techniques and Neural Network Models

The estimation models used in this study include statistical techniques such as linear regression, moving average, single exponential smoothing, double exponential smoothing, and ARIMA and Neural Network. Furthermore, the models and the techniques are evaluated using performance accuracy of prediction, namely MSE and RMSE. We compare the models and statistical techniques to obtain the best model.

Evaluation

To measure the performance of the estimation models, Root Mean Square Error (RMSE) and Mean Square Error (MSE) are used. The MSE and RMSE are calculated by the following formula [14]:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2}{n}} \qquad (2)$$

$$MSE = \frac{\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2}{n} \qquad (3)$$

Where Y_t = actual value and $\hat{Y}_t$ = value of prediction.

RESULTS AND DISCUSSION

The data set of foreign tourists that have been collected are then preprocessed by checking the missing value of data. Since there is no missing value in the collected data in this study, all of data can be used to construct the models. Furthermore, in converting univariate data into multivariate data, 12 period is used as input which includes x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}, x_{t-9}, x_{t-10}, x_{t-11}, and x_{t-12}. And the output is x_t. In this study, x_{t-1}, x_{t-2} as input and x_t for output of Neural Network model is also used.

Predicted models of foreign tourists

This study uses the model or method that include linear regression statistical techniques, moving average, exponential smoothing and ARIMA. This research also constructs models with k-NN and Neural Network. Model performance is measured by MSE and RMSE values.

The first model is linear regression. The equation of linear regression model obtained is as follows:

$$y_t = 0.36697 - 0.00053299 * t \qquad (4)$$

where y_t is value of prediction at t. The performance of the linear regression model obtained, that is RMSE and MSE, are 0.1644 and.02706, respectively.

For the Moving Average model, many experiments are performed using different values of parameter. The results of performance obtained are shown in Table 2.

Table 2 shows that Moving average model using parameter =2 gives minimal values of RMSE and MSE. This indicates that MA (2) is the best model for the moving average.

Experiments using the exponential smoothing model have been performed using various values of α. The results of experiments are shown in Table 3 as follows:

TABLE 2. Comparison of MSE and RMSE using Moving Average Model

MODELS	MSE	RMSE
MA (2)	0.016753	**0.129433**
MA (3)	0.017987	0.134116
MA (5)	0.019498	0.139635
MA (8)	0.020366	0.142709
MA (10)	0.019897	0.141057
MA (11)	0.019085	0.138148

TABLE 3. Comparison of MSE and RMSE using Single. Double and Triple Exponential Smoothing (ES) Model

MODELS			MSE	RMSE
Single ES		0.1	0.01933	0.13903
		0.3	0.01616	0.12713
		0.6	0.01508	**0.12279**
		0.8	0.01527	0.12357
		0.9	0.01565	0.12512
		1.0	0.01626	0.12752
DOUBLE ES			MSE	RMSE
	0.1	0.1	0.02130	0.14595
	0.1	0.2	0.02230	0.14932
	0.2	0.1	0.01913	**0.13830**
	0.2	0.2	0.02068	0.14381
TRIPLE ES			MSE	RMSE
(0.1, 0.1, 0.2)			0.08044	0.28361
(0.2, 0.2, 0.2)			0.02027	**0.14237**
(0.3, 0.3, 0.3)			0.02147	0.14653

Table 3 shows that Single Exponential Smoothing using α =0.6 gives minimal values of RMSE and MSE. Thus, Single Exponential Smoothing model using α =0.6 is the best model for Exponential Smoothing model.

Neural network models have been successful and applied for predictions in various fields. In this study, the neural network model will be used to predict the number of foreign tourists in Central Java, Indonesia. To find the best performance of the Neural network model, we will find the best configuration of this model, by experiment with iteration parameter = 10,000, learning rate = 0.1, momentum = 0.2, the several of number of neurons in the hidden layer, and the number of hidden layers is 1 and 2 layers. Table 4 is the experiment result of the prediction of foreign tourists using Neural Network model.

TABLE 4. Comparison of MSE and RMSE using Neural Network Model

MODELS	MSE	RMSE
NN(12-2-1)	0.01904	0.138
NN(12-3-1)	0.02190	0.148
NN(12-5-1)	**0.01613**	**0.127**
NN(12-7-1)	0.02103	0.145
NN(12-8-1)	0.02434	0.156
NN(12-5-5-1)	0.01769	0.133
NN(12-5-7-1)	0.01823	0.135
NN(12-5-9-1)	0.01638	0.128
NN(12-5-11-1)	0.01904	0.138
NN(12-5-13-1)	0.01988	0.141

Table 4 shows that the Neural Network model, i.e. NN (12-5-1) has RMSE. This shows that the NN (12-5-1) model is the best.

We also conduct experiments using other parameters of Neural Networks model, with inputs equal to 2, namely x_{t-1} and x_{t-2}, the number of iterations = 10,000, learning rate = 0.05, momentum = 0.5, varying the number of neurons in the hidden layer, the number of hidden layers equal to 1, and the activation function using hyperbolic tangent. Table 5 shows the performance results of the neural network models with the above parameters.

TABLE 5. Comparison of MSE and RMSE using Neural Network Model with different input

MODELS	MSE	RMSE
NN(2-3-1)	0.01261	0.11229
NN(2-4-1)	**0.01186**	**0.10890**
NN(2-5-1)	0.01438	0.11992
NN(2-6-1)	0.01446	0.12026
NN(2-7-1)	0.01427	0.11944
NN(2-8-1)	0.014622	0.12092

From Table 5, it appears that neural network configuration with 2 neurons inputs (NN(2-4-1)) gives better performance than Neural Network with 12 neuron inputs (NN(12-5-12)). The figure of Neural Network configuration with 2 neurons inputs (NN (2-4-1)) is shown in Figure 2.

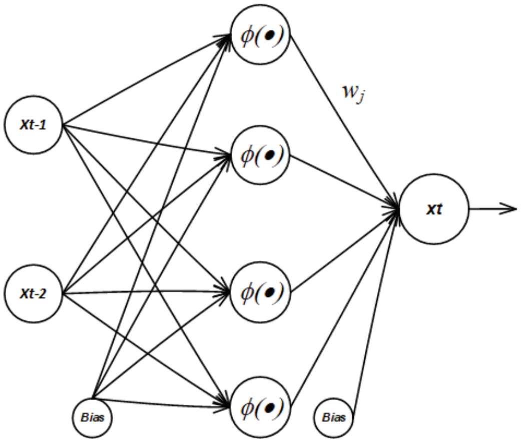

FIGURE 1. Neural network configuration with 2 neurons inputs

Models Comparison

Based on the performance of linear regression model, moving average, single exponential smoothing, double exponential smoothing, triple exponential smoothing, ARIMA and neural network model shown in Table 1 to Table 5, the performance results of those models can be compared. The comparison of model performance for the prediction of foreign tourists in Central Java Province is shown in Table 6.

TABLE 6. Comparison MSE and RMSE using Statistical techniques and Neural Network Model

MODELS	MSE	RMSE
Linear Regression	0.0271	0.1644
Moving Average	0.0168	0.1294
Single Exponential	0.0151	0.1228
DOUBLE ES	0.0191	0.1383
TRIPLE ES	0.0203	0.1424
ARIMA (1,1,1)	0.0139	0.1180
NN(2-4-1)	0.0119	**0.1089**

Figure 2 shows graphically the comparison of MSE and RMSE of models, i.e. Linear Regression (trend), moving average, Single Exponential Smoothing, Double Exponential Smoothing, Triple Exponential Smoothing, ARIMA, and Neural Network. From Figure 2 and Table 6, it is clear that Neural Network model gives best result compared to other models.

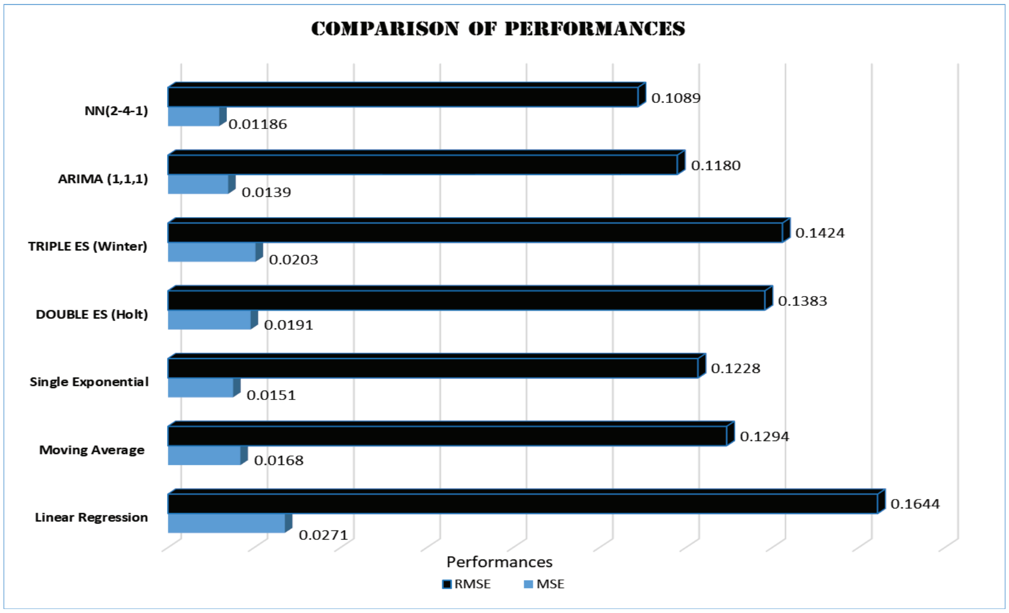

FIGURE 2. Comparison of MSE and RMSE using Statistical Techniques and Neural Network Model

Figure 3 shows the comparison of actual values and predicted values using Neural Network model. From the Figure 3, it can be seen that the prediction values using the Neural Network model is close to the actual values of the normalized data.

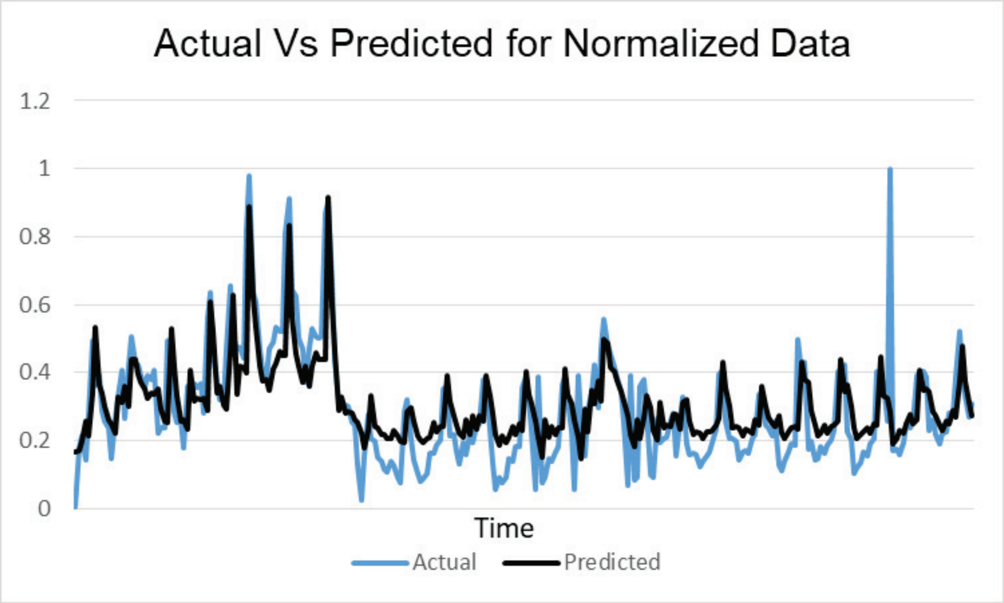

FIGURE 3. Comparison of Actual values and Predicted values for Normalization data

CONCLUSION

The prediction of foreign tourists in Central Java Province has been conducted through experiments on various models / methods with both statistical techniques, i.e. Moving Average, Single Exponential, Double Exponential, Triple exponential and ARIMA, as well as with soft computing method such as the Neural Network model. The result of model performance shows that Neural network model with input of 2 neurons, in hidden layer with 4 neuron, iteration = 10.000, learning rate = 0.05., Momentum = 0.5 gives best result compared to other models.

ACKNOWLEDGMENTS

This paper is the result of a study under research grant supported by the Ministry of Research, Technology and Higher Education, Republic of Indonesia in 2017.

REFERENCES

1. United Nations World Tourism Organization, *UNWTO Tourism Highlight, 2014 Edition*. (www.e-uwnto.org, 2014*)*
2. Ministry of Tourism and Creative Economy, *Performance Accountability Report of the Ministry of Tourism and Creative Economy*, (Ministry of Tourism and Creative Economy, Jakarta, 2014)
3. J. E. Hanke and D. W. Wichern, *Business Forecasting (9th Ed.)*. (Prentice-Hall, NJ, 2009).
4. R. Law, *Journal of Travel & Tourism Marketing* 10, 47–66 (2001)
5. F. A. Razak, M. Shitan, A. H. Hashim and I. Z. Abidin, *Jurnal Kejuruteraan* 21, 53-62 (2009)
6. P. Gustavsson and J. Nordstrom, *Tourism Economics* 7, 117–133 (2001)
7. C. Lim and M. McAleer, *Annals of Tourism Research* 28, 68–82 (2001)
8. A. Papatheodorou and H. Song, *Tourism Economics* 11, 11–23, (2005)
9. H. Hassani, E. S. Silvay, N. Antonakakis, G. Filis and R. Gupta, "Forecasting Accuracy Evaluation of Tourist Arrivals: Evidence from Parametric and Non-Parametric Techniques", in *Working Paper Series*, (University of Pretoria, South Africa, 2015)
10. N. Loganathan and I. Yahaya, *South Asian Journal of Tourism and Heritage* 3(2), pp.50-60 (2010)
11. G. P. Zhang, *Neurocomputing* 50, 159-175 (2003)
12. H. Niskaa, T. Hiltunena, A. Karppinenb, J. Ruuskanena, and M. Kolehmaine, *Engineering Applications of Artificial Intelligence*, 17, 159–167 (2004)
13. Purwanto, C. Eswaran and R. Logeswaran, "Improved Adaptive Neuro-Fuzzy Inference System for HIV/AIDS Time Series Prediction", *in Informatics Engineering and Information Science* 253, (Malaysia, 2011), pp. 1-13.
14. I. Rojas, O. Valenzuela, F. Rojas, A. Guillen, L. J. Herrera, H. Pomares, L. Marquez and M. Pasadas, *Neurocomputing* 71, 519 – 537 (2008)