Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia

Vol. 11, 2022, page 44 - 51

Comparison Study of SARIMA and ANFIS Model in Weather Forecasting of Temperature Data in Malaysia

Nur Atiqah Morshidi, Siti Mariam Norrulashikin*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: sitimariam@utm.my

Abstract

Weather forecast is one of the crucial types of forecasting as daily human activities heavily depends on it. Hence, there is a high significance in determining the best models that could predict weather more accurately. The aim of this study is to make a comparison study of the two different approaches of weather-forecasting method which are Seasonal Auto-regressive Integrated Moving Average (SARIMA) and Adaptive Network-based Fuzzy Interference System (ANFIS). The two types of weather forecast model are chosen and proposed in this study to predict the temperature data forecast by using two different approaches where SARIMA is a statistical-based forecasting model while ANFIS is a neural-network-based forecasting model. The data chosen for the case study is the average monthly temperature data in the duration of ten years, from January 2010 until September 2019, obtained based in Senai station, Malaysia. At the end result of the two models' analysis, the objective is to evaluate the forecasting performance through error measures which are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Final results were compared between SARIMA and ANFIS method and it was found that with the same ratio of training and testing data, ANFIS method has lower error measures in forecasting the temperature data which brings to the result that ANFIS is a better forecasting method for the temperature data based in Senai station.

Keywords: Weather forecasting; SARIMA; ANFIS

1. Introduction

Time series can be defined as a collection of observations, which are data points, being measured at equal time intervals. Evolving over time, the applications of time series analysis has expanded greatly across multiple areas especially in forecasting and analysis such as economic forecasting, weather forecasting, stock market analysis, sales forecasting and many more. Forecasting is a technique or a process of analysing past and current data by observing trends to make predictions. Due to its application and significance in many sectors such as in marine, agriculture and aviation industry, there has been many methods developed by scientists and mathematicians through studies and research that available for forecasting, depending on the situation and type of forecasting analysis to be made.

Weather forecasting is undoubtedly one of the important types of forecasting and has been a part of human lives since decades. Being able to predict weather for the next few hours or days allows human activities to be planned properly and hence, the activities can be carried out more smoothly, taking into consideration of the weather. In ancient times, weather forecasting was done mainly from the observation of weather pattern. Across time, human has become more and more advance in forecasting, accelerated with the evolving knowledge and advancement in computers and software. As compared to ancient times, in present days, there are many methods and models available, which have been developed and improved over time by scientists and mathematicians. This is further supported by the fact that weather forecasting has increased significantly in every aspect of life where a good forecasting is crucial for maximum work efficiency. In many businesses, having a good estimation on weather forecasting further drives the business on the right track as most of the plan will follow accordingly.

Human activities and natural climate changes as the earth grew older has crucially affected the weather in Malaysia. The impact thus can obviously be observed over the years by looking at the huge changes of temperature in which the trend shows an increase in the annual mean temperature.

Meteorologist uses temperature prediction to do forecasting on the state of atmospheric parameters [1] and in general, temperature itself are used to be measured at a higher accuracy as compared to other variables in weather forecasting [2]. Therefore, it is significant to have a good forecasting model which would be able to predict the temperature more accurately in order for better-planned activities can be executed, taking into consideration as well of the actions that could affect the temperature changes in the long run.

Although there are different models available for weather forecasting, this study will only highlight on two models which are Seasonal Auto-regressive Integrated Moving Average (SARIMA) and Adaptive Network-based Fuzzy Interference System (ANFIS). This comparison study is done to verify which model gives better performance for these set of data by using two different approaches where SARIMA is a statistical-based forecasting model while ANFIS is a neural-network-based forecasting model. Hence, this study is a comparison study of statistical-based forecasting model with network-based forecasting model which are SARIMA and ANFIS in determining the best model for temperature prediction in Malaysia. This quantitative study will be using the data recorded by several weather stations in Peninsular Malaysia for analysis.

This research aims to make a comparison between SARIMA and ANFIS models in order to find the best model that fits the temperature data in Malaysia and to propose the best forecasting model for temperature trend analysis in Malaysia. The performance indicator are the error metrics RMSE and MAE in which the smaller the RMSE and MAE value, the better the forecasting results will be.

2. Literature Review

2.1. SARIMA Model

2.1.1. SARIMA concept

ARIMA model, which was introduced by Box and Jenkins in the year 1978, is a class of mathematical models used for forecasting [3]. It helps in explaining a given time series based on its past values, which were its own lags and the lagged forecast errors so that the equation can be used to forecast future values. The nature of ARIMA itself is very flexible [4] and it requires only few assumptions, hence the extensive use of the ARIMA models. The ARIMA model consists of 3 components which are:

i. Auto-regressive (AR) component which is denoted by p;

$$X_t = a + \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t$$

where:

a = $constant,$
ϕi = $the\ parameter\ for\ the\ model$
εt = $the\ term\ for\ error$

ii. Moving average (MA) component which is denoted by q;

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

where:

θ_i = the parameter for the model
x_t = the data observed at t
ε = random error

iii. Moving average (MA) component which is denoted by q;

$$Y_t = \alpha + \sum_{i=1}^{p} \phi_i \bullet Y_{t-i} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}$$

where:

Y_t = a stationary stochastic process
α = a constant
ε_t = the white noise disturbance term
ϕ_i = the coefficient of auto-regression
θj = the moving average coefficient.

If the data is concerned with seasonality, then a SARIMA model is considered where the equation is given by:

$$\emptyset(B)\Phi_p(B^S)(1-B)^d(1-B^S)^D y_t = \delta + \theta(B)\Theta_Q(B^S)\varepsilon_t$$

where ε_t is the white noise, $\emptyset(B)$ is the ordinary AR and $\theta(B)$ is the MA component, $\Theta_Q(B^S)$ and $\Phi_p(B^S)$ are the seasonal AR and MA components respectively, $(1-B)^d$ and $(1-B^S)^D$ are the ordinary and seasonal differencing component of order *d* and *D*.

2.2. ANFIS Model

2.1.2. ANFIS concept

ANFIS model is proposed by Roger Jang in 1993, serving as a basis in the construction of a set of fuzzy 'if-then' rules with appropriate membership functions in generating pairs of input and output [5]. It is based on the Sugeno Model which can generally described (for first-order models) in the case of two inputs which are x and y , with one output, z is as follows:

$$\text{Rule 1: If x is } C_1 \text{ and y is } C_2\text{, then } f_1 = p_1x + q_1y + r_1$$
$$\text{Rule 2: If x is } D_1 \text{ and y is } D_2\text{, then } f_2 = p_2x + q_2y + r_2$$

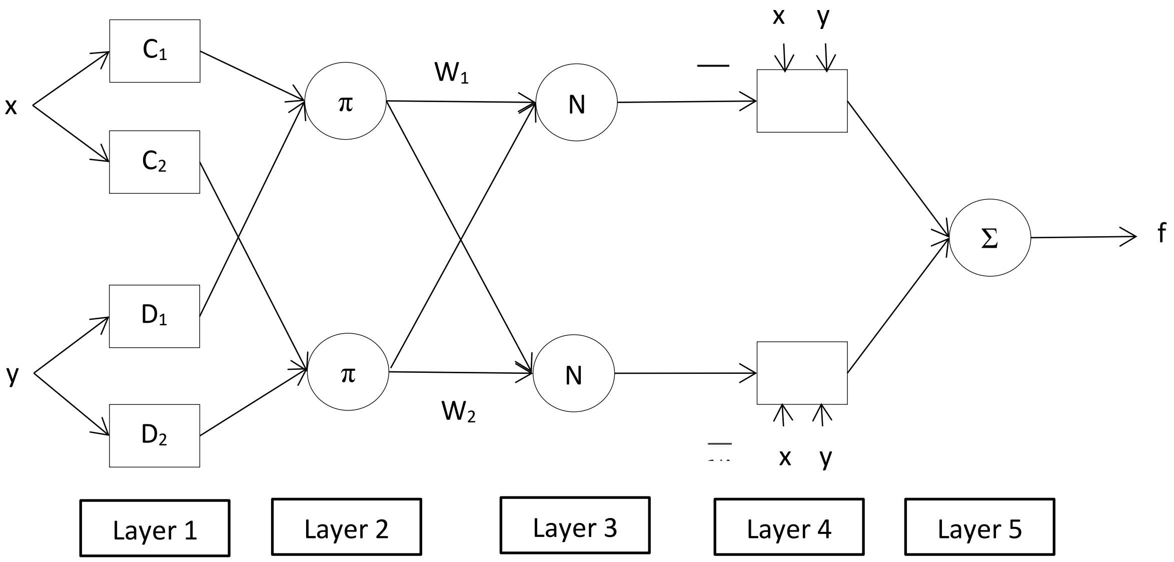

Figure 1 ANFIS architecture

3. Methodology

3.1. Introduction

The data in this study is the monthly temperature data from January 2010 to September 2019, obtained from Senai meteorological station. The temperature data is then classified into training data and testing data. Data in January 2010-December 2018 is chosen as the training data. This data will be in the same ratio for both SARIMA and ANFIS model. The values of the error metrics then will be compared to evaluate the accuracy of forecasting results by RMSE and MAE.

3.2. SARIMA Model

The research used R software for analysis purposes. The steps for forecasting using SARIMA Model are as follows:

1. *The graph for the data is plotted for visualization and identification of the characteristics of the data.*
2. *ACF is plotted to consider the components in a time series data.*
3. *Data is splitted into training and testing data with the ratio of 80:20.*
4. *Seasonal differencing and non-seasonal differencing is performed.*

5. Identification of seasonal and non-seasonal component to list down possible models.
6. Parameter estimation for each model is identified.
7. Diagnostic checking is performed.
8. Forecasting of data.
9. Model Validation.
10. Error measurements of the best model.

3.2. ANFIS Model

The research used MATLAB software for analysis purposes. The steps for forecasting using ANFIS Model are as follows:

1. The matrix for training and testing data is entered in MATLAB
2. The initialization stage: entering small number of input lag
3. Number of epochs and type of membership function are set
4. Prediction is computed
5. Error is obtained
6. Repeat for training and testing data

4. Results and discussion

4.1. SARIMA Model

The data obtained is plotted to consider the characteristics of time series data. Then, the ACF is plotted for training data where we obtained the plot as in Figure 2 and hence, seasonal differencing and non-seasonal differencing is performed to stabilize the mean of the time series data.

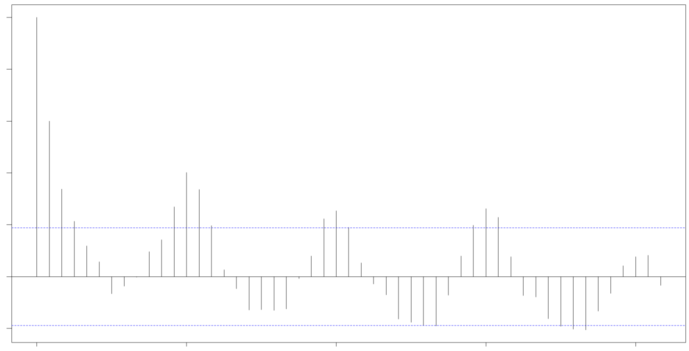

Figure 2 ACF plot of training data

For model identification, the values of *p, q*, *P* and *Q* of our SARIMA model will be obtained from the ACF and PACF plot of the differenced train data.

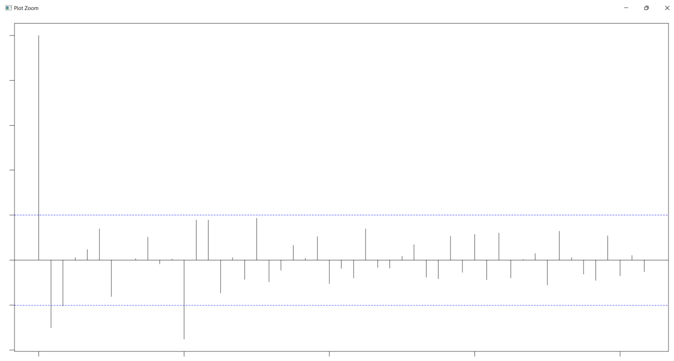

Figure 3 ACF plot of differenced train data

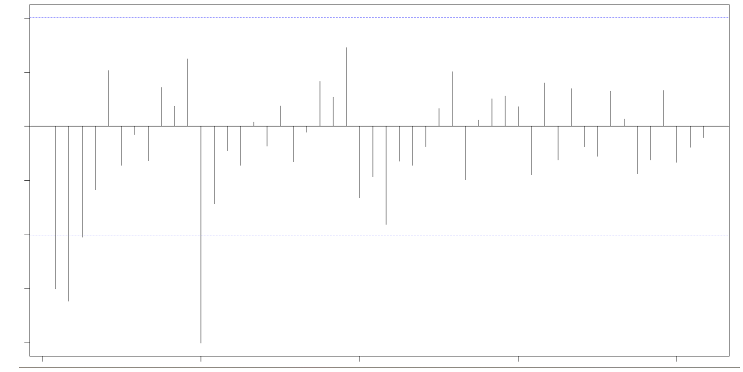

Figure 4 PACF plot of differenced train data

From Figure 3 and Figure 4 as shown above in our analysis , the possible models obtained for SARIMA model are SARIMA(1,1,1)×(1,1,1)$_{12}$, SARIMA(2,1,1)×(1,1,1)$_{12}$, SARIMA(2,1,2)×(1,1,1)$_{12}$ and SARIMA(3,1,1)×(1,1,1)$_{12}$.

Next, parameter estimation is observed for each possible models. The result is displayed in Table 1.

Table 1 Parameter estimation of possible models of SARIMA

Possible Models	Log Likelihood	AIC
SARIMA(1,1,1)×(1,1,1)$_{12}$	-90.05	190.1
SARIMA(2,1,1)×(1,1,1)$_{12}$	-85.73	183.46
SARIMA(2,1,2)×(1,1,1)$_{12}$	-83.41	180.83
SARIMA(3,1,1)×(1,1,1)$_{12}$.	-83.43	180.87

Based on the result of parameter estimation shown in the above table, it can be observed that the model with the lowest AIC is SARIMA(2,1,2)×(1,1,1)$_{12}$ indicating the model with the best fit for our data. Hence, in the next steps, we will only consider the model SARIMA(2,1,2)×(1,1,1)$_{12}$.

Next, diagnostic checking is performed by using arch-test and Ljung-Box test. The results shows there are no ARCH effects present and the serial correlation exist in the temperature data.

We proceed to the next step which is forecasting. The plot of forecast for the best model of SARIMA are shown as below.

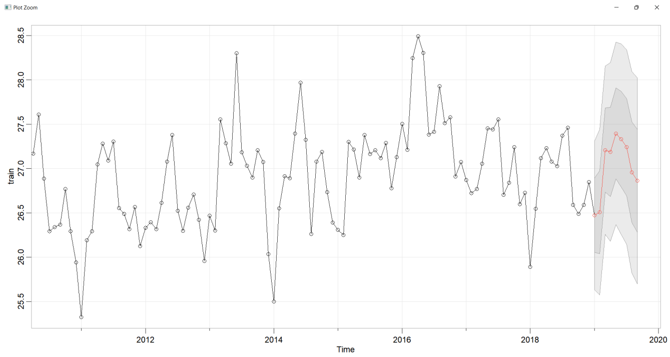

Figure 5 Forecast plot for SARIMA(2,1,2)×(1,1,1)$_{12}$

Lastly, model validation is done by finding the error values of the models.

Table 2 Error values of fitted data for SARIMA models

Selected Model	Training Sets Error Metrics		Testing Sets Error Metrics	
	RMSE	MAE	RMSE	MAE
SARIMA(2,1,2)×(1,1,1)$_{12}$	0.4828	0.3565	0.6507	0.5498

Based on the results in Table 2, it can be observed that for SARIMA(2,1,2)×(1,1,1)$_{12}$, the values of RMSE and MAE are 0.4828 and 0.3565 respectively for training and for the testing sets of data, RMSE = 0.6507 and MAE = 0.5492. The difference between the error values for training and testing sets are relatively small. Hence, this indicates that the selected model is a good fit for our data. Thus, the best SARIMA model that fits the temperature data based in Senai station is SARIMA(2,1,2)×$(1,1,1)_{12}$ based on its AIC value which is the lowest and by using its RMSE and MAE, we will compare SARIMA(2,1,2)×$(1,1,1)_{12}$ with the model from ANFIS.

4.2. ANFIS Model

4.2.1. Methodology

For ANFIS testing, the software used is MATLAB. The variation for testing is according to the number of membership functions where the membership functions are used in the mapping of non-fuzzy input values to fuzzy linguistic terms in the fuzzification and defuzzification steps of a fuzzy logic system. In this study, the membership functions chosen is the generalized-bell membership function (gbellmf) with the number of membership functions is 3 3 3 since we have 3 inputs and 1 output. Next, we vary the number of epochs, which the number of epochs chosen for this research is 10,20,30 and 50. The results of RMSE and MAE is then obtained through calculation, based on the forecast and fitted data obtained from MATLAB.

4.2.2. Results & Discussion

Checking plots are obtained from MATLAB to validate the ANFIS output with the training and testing data on whether it matches the training data well. This helps to decide on whether here is a need to improve the match by increasing the number of membership functions or by increasing the number of training epochs. Based on the checking plots obtained, it can be observed that the match between the ANFIS output with the training and testing data has improved as the number of training epochs increases.

Below is the checking plots for training and testing data at each epoch.

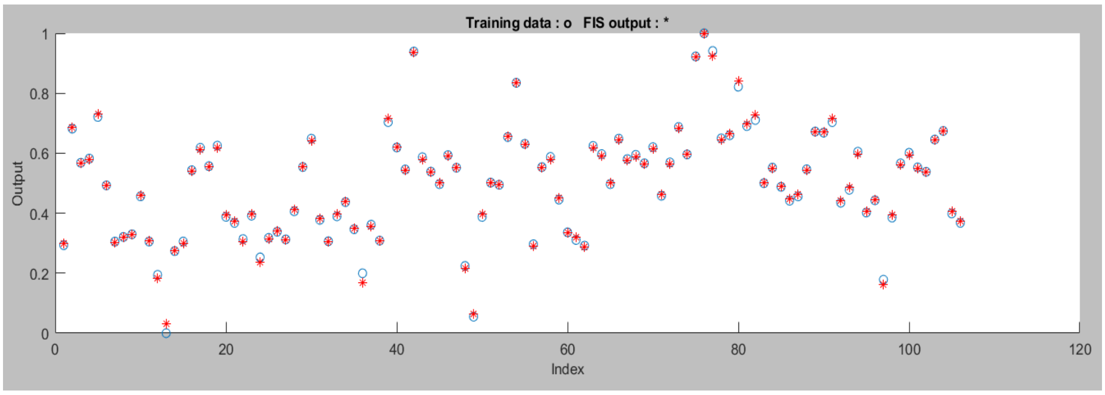

Figure 6 Checking plot for training data with epoch=10

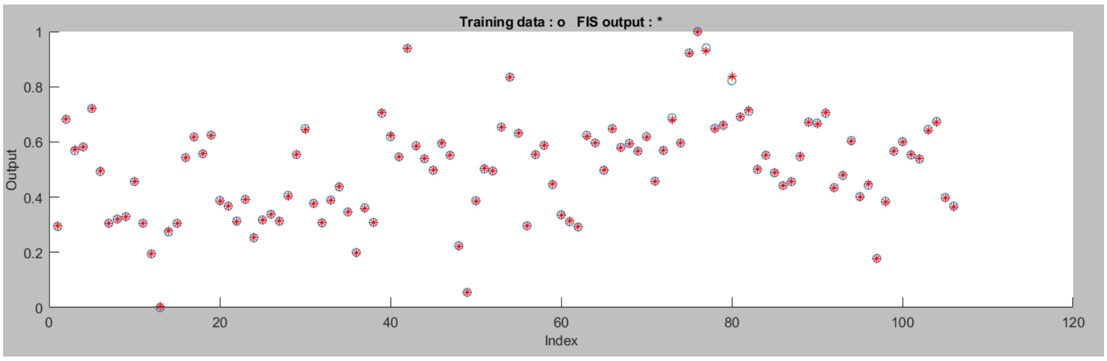

Figure 7 Checking plot for training data with epoch=20

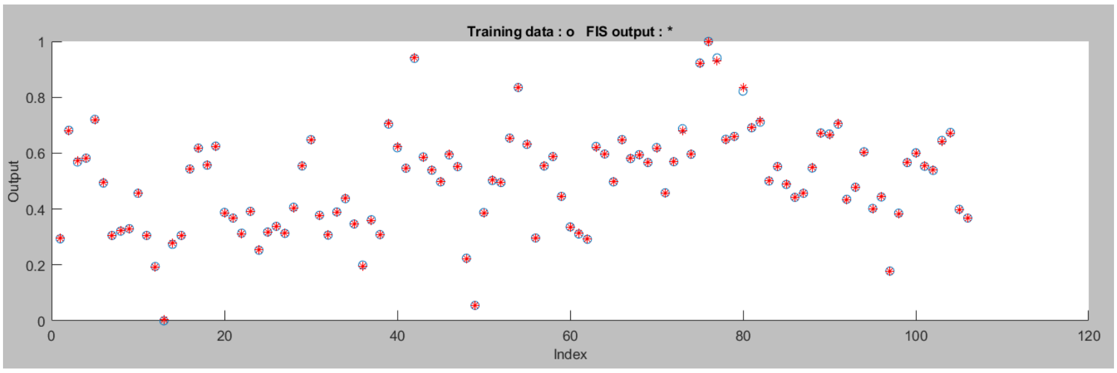

Figure 8 Checking plot for training data with epoch=30

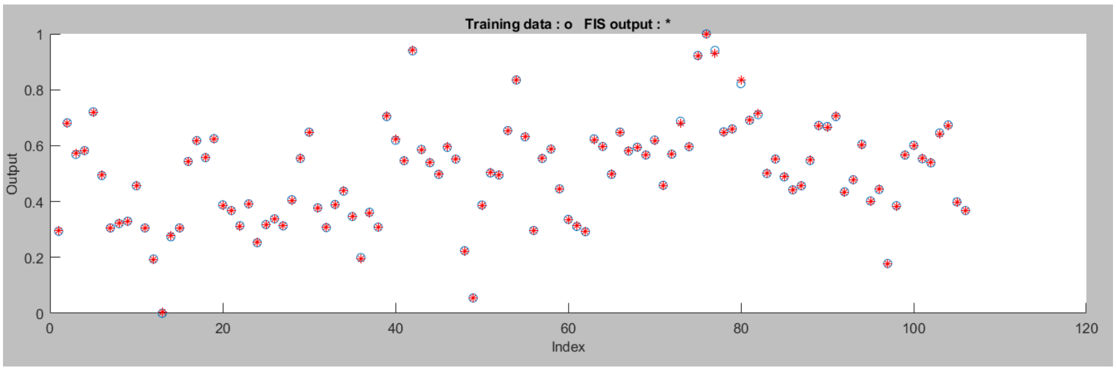

Figure 9 Checking plot for training data with epoch=50

Based on the checking plot shown above, it can be observed that the match between the ANFIS output with the training data has improved as the number of training epochs increases. The same step is repeated for testing data to observe on whether the ANFIS output for testing data matches the testing data well. The results show the same trend are seen where as we increase the number of epochs, the ANFIS output has a greater match with testing data.

The error values for the fitted and forecast models are simplified in the table below:

Table 3 Error values of fitted data for ANFIS models

Generalized Membership Function (gbellmf)		
Number of epoch	Training Sets Error Metrics	
	RMSE	MAE
10	0.02365	0.01680
20	0.008485	0.005458
30	0.007835	0.005191
50	0.007835	0.005191

Table 4 Error values of forecast data for ANFIS models

Generalized Membership Function (gbellmf)		
Number of epoch	Error Metrics	
	RMSE	MAE
10	0.03797	0.03178
20	0.01694	0.01307
30	0.01331	0.01046
50	0.01331	0.01046

Based on the table above for testing data, it can be observed that the value of RMSE and MAE starts becoming constant from epochs=30, which is also the same case for training data. As shown in the two tables above, it can be observed that, the number of epochs cannot be increased greater than 30 as this will lead to overfitting of data.

Conclusion

From the SARIMA method for the forecasting of temperature data, it is observed that SARIMA$(2,1,2)\times(1,1,1)_{12}$ is the best model that fits the best for the data as shown by its lowest AIC value. By using the ANFIS method of forecasting, based on the performance indicator result, the best model for the set of data is 3 3 3 with epochs = 30. In conclusion based on the outcomes of each method, by comparison with ARIMA$(2,1,2)\times(1,1,1)_{12}$, ANFIS method shows a lower error measurement of RMSE and MAE which is what we are trying to achieve. Therefore, for the dataset of average monthly temperature in Senai station ranging from January 2010 until September 2019, ANFIS is better in terms of forecasting as it has lower RMSE and MAE values than SARIMA$(2,1,2)\times(1,1,1)_{12}$.

Acknowledgement

The researcher would like to deeply thank all those who have given their support and ideas in completing the research especially the project supervisor, Dr. Siti Mariam Norrulashikin for her help and support along the way. Special appreciation to the friends and family members as well for their continuous support and believe in the researcher’s work since the start of the research.

References

[1] Tyagi, H., Suran, S., & Pattanaik, V. (2016). Weather - temperature pattern prediction and anomaly identification using artificial neural network. International Journal of Computer Applications, 140(3), 15–21. https://doi.org/10.5120/ijca2016909252.

[2] Tektaş, M. (2010). Weather forecasting using ANFIS and ARIMA models. Environmental Research, Engineering and Management, 51(1), 5-10.

[3] Fahimifard, S. M., Homayounif, M., Sabouhi, M., &amp; Moghaddamn, A. R. (2009). Comparison of ANFIS, Ann, Garch and Arima techniques to exchange rate forecasting. Journal of Applied Sciences, 9(20), 3641–3651. https://doi.org/10.3923/jas.2009.3641.3651.

[4] Ho, S. L., & Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis. Computers & industrial engineering, 35(1-2), 213-216.

[5] Rahman, M., Islam, A. H. M. S., Nadvi, S. Y. M., Rahman, R. M. (2013). Comparative study of ANFIS and Arima model for weather forecasting in Dhaka. 2013 International Conference on Informatics, Electronics and Vision (ICIEV). https://doi.org/10.1109/iciev.2013.6572587.