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

Vol. 11, 2022, page 112 - 120

Investigating the Perfomance of SARIMA and ANN on Rainfall Forecasting

Nurul Aqilah Saiful Bahril, Siti Rohani Mohd Nor*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: sitirohani@utm.my

Abstract
Malaysia was identified as the top ten countries with the greatest rainfall in the globe. It is a Southeast Asian country located at the northern equator's point, situated in the Gulf of Thailand, with a path that stretches across Borneo Island and the Malay Peninsula. Geographically, the climate favors an equatorial weather pattern, with wet, hot, and humid weather throughout the year. Hence, rainfall forecasting is important to plan agriculture and domestic events as well to prevent catastrophic events especially during the monsoon season. Consequently, the accuracy of rainfall forecasting is demanding. As technology becomes more embedded in our daily lives by the minute with the growth of Artificial Intelligence (AI), rainfall forecasting by Artificial Neural Network (ANN) model has confiscated the interest of many more than the traditional forecasting method, Seasonal Autoregressive Integrated Moving Average (SARIMA). Therefore, this study aims to analyze the performance of ANN and SARIMA to an observed state in Peninsula Malaysia namely, Cameron Highland from January 2010 to September 2019. The accuracy of predicting performances for both models is evaluated by using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE), with the best forecasting model being chosen based on the lowest error. In projecting the observed state in Peninsula Malaysia, the ANN model outperformed the SARIMA model.

Keywords: SARIMA; ANN; Rainfall forecasting Malaysia

1 Introduction

Malaysia's annual rainfall is 80 percent which is between 2000mm to 2500mm with its average daily temperature between 25°C and 35°C. Since the Peninsula of Malaysia is geographically located in the southern land of Asia, its characteristics of rainfall are influenced by two monsoon seasons, the South West (SWM) and the North East (NEM) which often causes natural disasters and force many event to put on a halt.

Forecasting time series is a challenging problem with no simple solution. Despite being able to forecast seasonal and stationary data, SARIMA model has limitations to some time series analysis especially the weather forecasting. This has proven by a number of researchers and this model does not satisfy the linearity assumption in many time series events [9]. The study of ANN nonetheless, has expanded and is used in broad areas including the Rainfall prediction. Rainfall forecasting involves a complex nonlinear data pattern. In a study of rainfall in Bangkok, Thailand using the ANN model was able to learn from continuous input data which contained both rain and dry periods [1]. Hence, Neural Network model is to investigate its efficacy in weather forecasting, namely rainfall in Cameron Highland form early 2010 to September 2019, and compared it to SARIMA.

2 Research Methodology

2.1 The Dataset

The rainfall data set for this paper is from a secondary source and has a span of ten years from early 2010 to September 2019 which aggregates 120 months. In this study, the last nine months form January 2019 to September 2019 will be used to forecast the data set by using the best fitted model.

2.2 SARIMA mmodel

In modeling the data, a SARIMA model is formed by including additional seasonal terms in the ARIMA models which can be written as, the lowercase (*p, d, q*) which is the non-seasonal part of the model and the uppercase (*P, D, Q*) which is the seasonal part of the model. In detail, *p* is the degree of regular autoregressive parts, *d* is the degree of regular differencing, *q* is the degree of regular moving average processes. Thus, the SARIMA model can be shown as,

$$\phi(\mathrm{B})\Phi_P(B^s)\,(1-B)^d\;\;(1-B^s)^D y_t = \delta + \theta(\mathrm{B})\,\Theta_Q(B^s)\,\varepsilon_t \tag{1}$$

where $\varepsilon_t$is Gaussian white noise, ϕ(B) is the ordinary autoregressive and θ(B) is the moving average components while $\Theta_Q(B^s)$ and $\Phi_P\,(B^s)$ are seasonal autoregressive and moving average components, respectively, $(1-B)^d$ and $(1-B^s)^D$are the ordinary and seasonal difference components of order *d* and *D*. Unfortunately, the estimation procedures for the SARIMA model are available only for stationary series.

2.3 ANN model

A neural network resembles the human brain i.e. the systems of neurons (Neural Network Definition , December 2020) and can be organized as layers. There are three layers in a simple neural network: an input layer as predictors, an output layer as the forecast , and an intermediate layer in between as the hidden neurons. The ANN is a mathematical model that has the ability to identify the nonlinear relationship between input and output parameters. In forecasting, the Multilayer Perceptrons (MLP) architecture is the most widely used in the ANN design. The parameters for ANN modeling are basically network topology, neurones characteristics, training and learning rules [2].

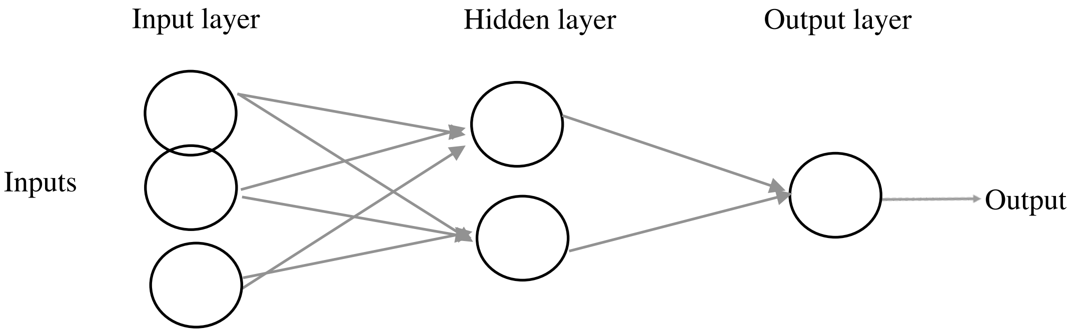

Figure 1 Neural Network with Three Inputs and One Output

Designing the number and structure of the network inputs are important in data science rainfall as it is a complex stochastic process which involves a number of unknown effects. In building the ANN for Cameron Highland rainfall forecasting, the rainfall data is separated into two sets: in-sample data and out-of-sample data. The in-sample data is used for forecasting as training data. To begin, the in-sample data are transformed to time series, and a number of ANN models with MSEs are produced.

The best model is an ANN model with the lowest MSE. The preferred ANN model is then fitted to the neural network. The fitted in-sample data is furthermore compared to the time series in-sample data using the measurement errors MAE, MAPE, and RMSE to see if it is appropriate for predicting. Following that, the projected data from the in-sample is compared to the out-of-sample data. Finally, the forecasted data generated from the in-sample is then compared to the out-of-sample data using the measurement errors MAE, MAPE, and RMSE.

3 Results and Discussions

3.1 Data Descriptive

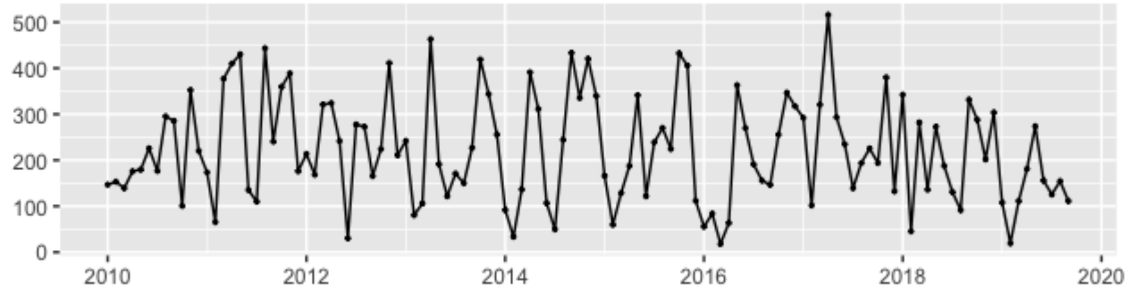

Figure 2 Time Series of Rainfall in Cameron Highland (2010-2019)

The time series plots shown above are not consistent for the time span from January 2010 to September 2019 and reveal the series to be nearly horizontal and show no recognisable patterns in the long run, indicating a stationary plot. Consequently, the plots portrays a seasonal cycle trends every 12 months. In addition to the Cameron Highland's rainfall distribution from 2010 to 2019, the place expected lower rainfall at the beginning of the year and higher rainfall in the mid of the year

Table 1 Data Descriptive of Rainfall(mm) of Cameron Highland

	Mean	Median	Min	Max	Variance	Skewness	Kurtosis
Cameron Highland	223.25	210.6	18	516	115.03	-0.70	223.25

Cameron Highland was thought to have a mild climate, with daytime temperatures not exceeding 25°C. Hence, this touristy spot has expected dewy seasons and heavy rainfall throughout the year. In proportion to the data descriptive of Cameron Highland above, it has a minimum of 18mm to 516mm of rainfall daily. Cameron Highland has a mean and median monthly rainfall of 223.25mm and 210.6mm, respectively, having October being the wettest and July being the driest, with a monthly variation of 115.03mm.

3.2 Analysis of SARIMA

Historical data for rainfall in Cameron Highland are splitted in two sets: in-sample (training set) and out-sample data (test set). The in-sample data are plotted into a time series to observe the pattern of the data. The ACF and PACF approach are used to determine the ideal SARIMA parameters and stationarity of the time series data.

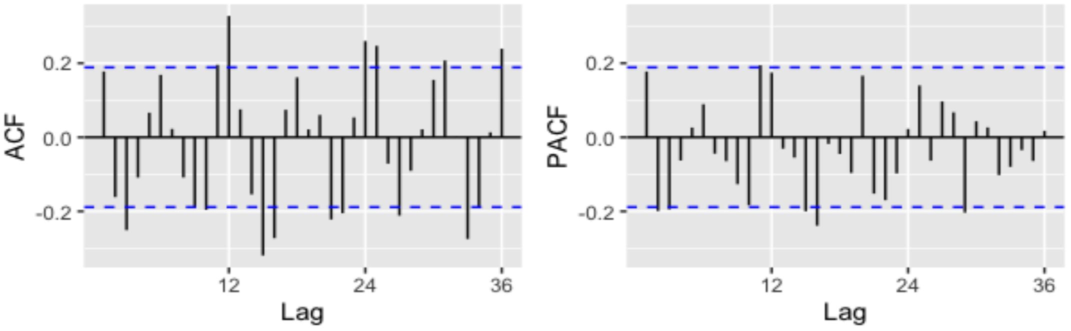

Figure 3 Time Series In-Sample ACF and PACF

The PACF charts are quite convincing that the data is stationary. The countless spikes in ACF plot of the time series, on the other hand, contradicts the previous statement as it does not drop to zero relatively quickly. Hence, the stationary of this time series is tested using the Augmented Dickey Fuller (ADF) test. The null and alternate hypothesis of this series are:

Ho: The time series is non-stationary (has unit root).
H1: The time series is stationary (has no unit root).

Table 2 Augmented Dickey-Fuller test result

Dickey-Fuller	Lag order	p-value
-2.8642	12	0.2182

According to the Augmented Dickey-Fuller test, the data fails to reject the null hypothesis since the *p*-value, 0.2182 which is greater than the critical value at 5%. Thus, the time series of Rainfall data in Cameron Highlands is a non-stationary data (has unit root) and establishing differencing is necessary for this time series.

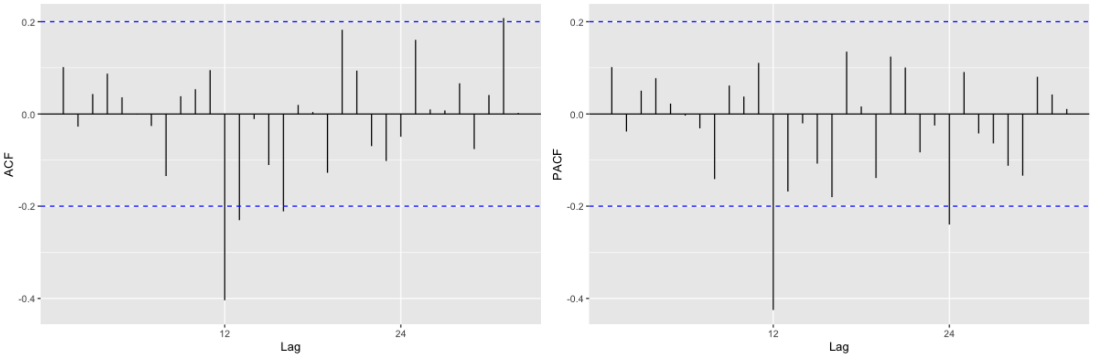

Figure 4 Time Series ACF and PACF after First Differencing

Table 3 Augmented Dickey-Fuller test result after First Differencing

Dickey-Fuller	Lag order	p-value
-4.3867	12	0.01

The Augmented Dickey-Fuller test expresses a *p*-value, 0.01 which is lower than the critical value at 5% and fails to reject the null hypothesis. Thus, the time series of Rainfall data in Cameron Highlands after the second differencing is a stationary data and are ready for forecasting. Thus, the six suggested tentative models that are suitable for rainfall forecasting. All of the provided tentative models are analyzed by comparing the modified Akaike information criterion (AIC) and Bayesian Information Criterion (BIC). The ideal fitted SARIMA model among those suggestions is SARIMA(0,0,0)x(2,1,1)**12** as it has the lowest AIC and BIC value

Table 4 Suggested SARIMA model and their corresponding AIC value

SARIMA model	**AIC value**	**BIC value**
SARIMA(0,0,0)x(0,1,0)**12**	1153.783	1158.645
SARIMA(0,0,0)x(0,1,1)**12**	1102.528	1109.821
SARIMA(0,0,0)x(1,1,0)**12**	1119.551	1126.843
SARIMA(0,0,0)x(1,1,1)**12**	1093.375	1103.098
SARIMA(0,0,0)x(2,1,0)**12**	1104.885	1114.608
SARIMA(0,0,0)x(2,1,1)**12**	1092.687	1104.841

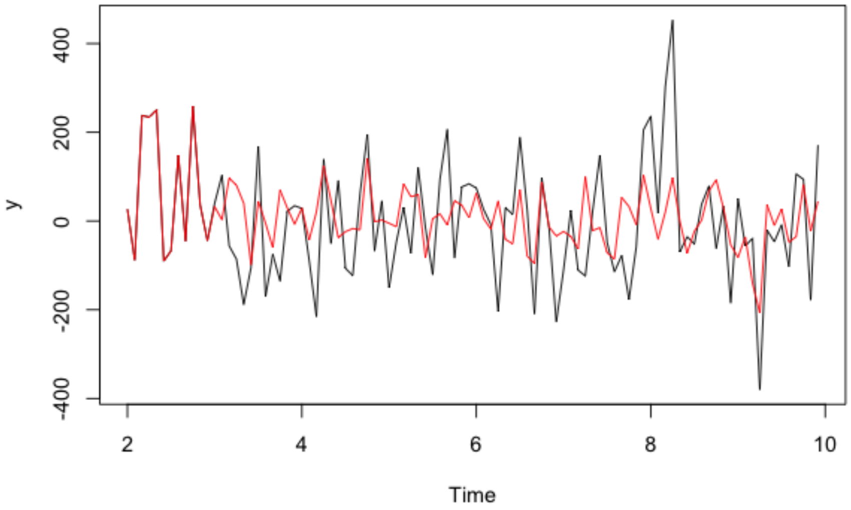

Figure 5 Time Series plot of Actual data after Second Differencing and Fitted data using SARIMA(0,0,0)x(2,1,1)**12**

Actual data after Second Differencing and Fitted data using SARIMA(0,0,0)x(2,1,1)**12** are a fit as the plots are relatively similar.

Table 5 Evaluation criteria of In-Sample data using SARIMA(0,0,0)x(2,1,1)**12**

RMSE	MAPE	MAE
0.4666	0.0043	0.3425

The errors displayed are minor. This demonstrates that the SARIMA(0,0,0)x(2,1,1)$_{12}$ model will anticipate future data with modest errors, implying that the projected data is closed to the actual data.

Following establishing model identification and parameter estimates, diagnostic checking is used to assess the adequacy of the model SARIMA(0,0,0)x(2,1,1)$_{12}$ in fitting the time series from year 2010 to year 2019. Hence, Ljung-Box tests are applied to examine the independence of the residuals. Hypothesis of Ljung-Box test is defined as:

H0: The residuals are independently distributed.
H1: The residuals are not independently distributed.

Table 6 Result of Ljung-Box Test for Residuals of SARIMA(0,0,0)x(2,1,1)$_{12}$

X-squared	df	*p*-value
0.61311	1	0.4336

From the Ljung-Box test, the *p*-value is 0.4336 which is larger than the critical value, therefore, there is not enough evidence to reject the null hypothesis at a significance level of 5%. Consequently the residuals are independently distributed and does not require extra modelling to improve forecasting performance accuracy.

3.3 Analysis of Neural Network Model

In this paper gathered data for all states were splitted into a training data set and test data set from January 2010 until December 2018 and January until December 2019 respectively. Subsequently, the time series training data set for each state is then fitted into the neural network and is tested for residuals. Below is the ANN model for Cameron Highland with its MSE.

Table 7 Suggested ANN model with its MSE

Model	**MSE**
ANN(15,6,1)$_{12}$	0.5892
ANN(15,7,1)$_{12}$	0.3793
ANN(15,8,1)$_{12}$	0.3536
ANN(15,9,1)$_{12}$	0.3376
ANN(15,10,1)$_{12}$	0.3851
ANN(15,11,1)$_{12}$	0.3732

The ANN(15,9,1)$_{12}$ proves to be the best model for as it comes with the lowest MSE compared to the other ANN models.

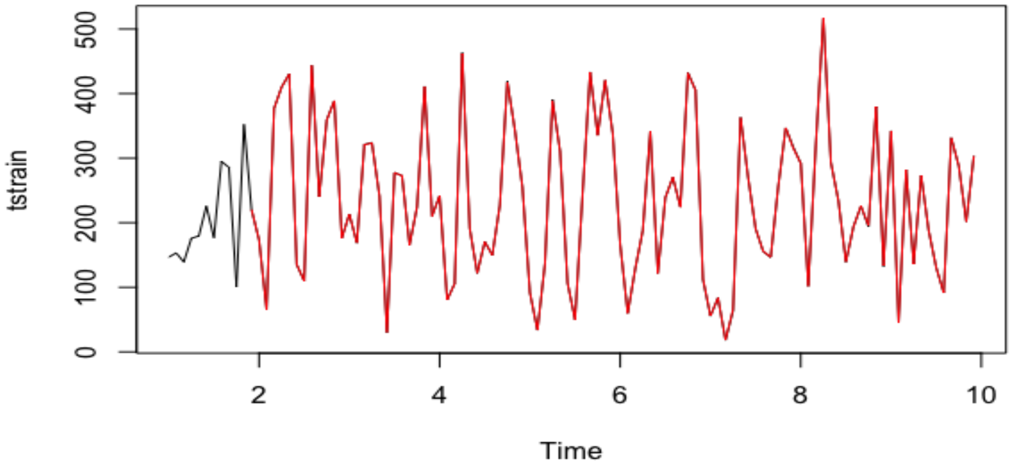

Figure 6 Time Series plot of Actual data and Fitted data using ANN(15,9,1)**12**

Actual data in the black line and Fitted data using ANN(15,9,1)**12** in the red line are a fit as the plots are relatively similar. Next, the residuals for the fitted NNAR are analyzed. Ljung-Box tests are applied to examine the independence of the residuals. Hypothesis of Ljung-Box test is defined as:

H0: The residuals are independently distributed.
H1: The residuals are not independently distributed.

Table 8 L-jung Box Test for ANN(15,9,1)**12**

States	Training data set	
	R-squared	*p*-value
Cameron Highland	0.86935	0.3511

The *p*-value is larger than 0.05 significant level. This shows no significant correlation among the residuals as the residuals lie among 95% confidence intervals. Hence, the residuals for the training data for Cameron Highland are independently distributed.

Next, the test data set is utilized for forecasting the distribution of rainfall from January 2019 to September 2019. Below are the forecast output for ANN(15,9,1)**12.**

3.4 Forecasting

The accuracy measurements, MAE, MAPE, and RMSE, are used to assess the modeling and forecasting capabilities of SARIMA(0,0,0)x(2,1,1)**12** and ANN(15,9,1)**12.** The goal of utilizing these assessment criteria is to corroborate the results reached about the optimal model for out-sample from January to September 2019.

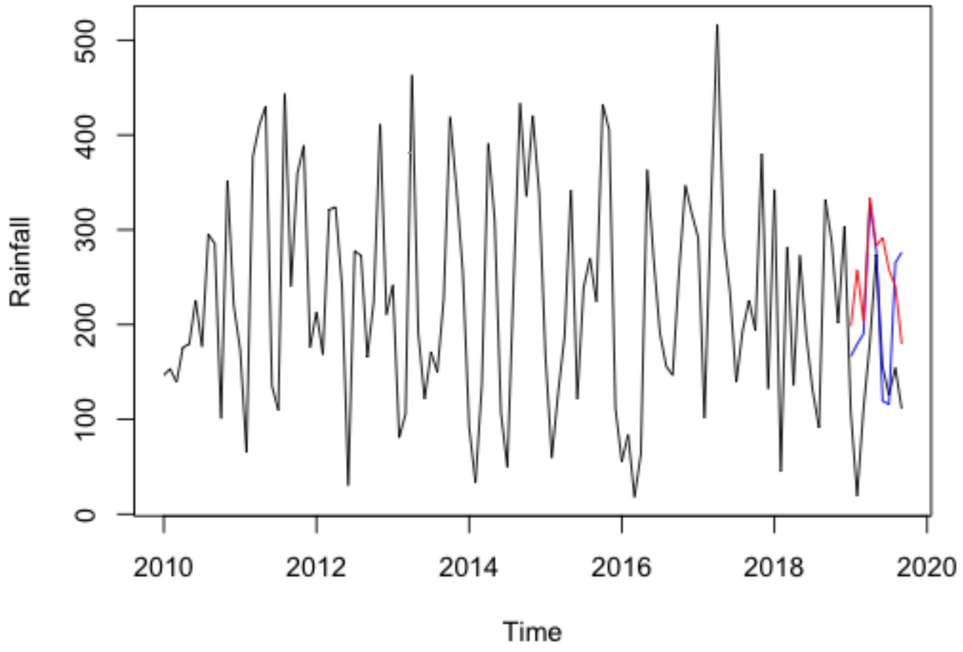

Figure 7 Comparison of Predicted and Actual data of Rainfall in Cameron Highland for SARIMA(0,0,0)x(2,1,1)**12** and ANN(15,9,1)**12**

There is a variability between the original Cameron Highland's data, the forecasted SARIMA(0,0,0)x(2,1,1)**12** data and the forecasted ANN(15,9,1)**12**. Nonetheless, the forecasted ANN(15,9,1)**12** model is represented by the blue line graph is clearly closer to the real data compared to the predicted outcome of SARIMA(0,0,0)x(2,1,1)**12** represented by the red line graph. Hence, the ANN(15,9,1)**12** is a superior model for rainfall forecasting in Cameron Highland than SARIMA(0,0,0)x(2,1,1)**12**

Table 9 Evaluation Criteria of In-Sample of SARIMA(0,0,0)x(2,0,0)**12** and ANN(15,9,1) **12**

Model	**RMSE**	**MAPE**	**MAE**
SARIMA(0,0,0)x(2,1,1)**12**	0.4666	0.0043	0.3425
ANN (15, 9, 1)**12**	0.5905	0.0022	0.3367

Table 10 Evaluation Criteria of Forecasting of SARIMA(0,0,0)x(2,0,0)**12** and ANN(15,9,1) **12**

Model	**RMSE**	**MAPE**	**MAE**
SARIMA(0,0,0)x(2,1,1)**12**	0.4648	0.0073	0.4092
ANN (*15*, *9*, *1*)**12**	0.3821	0.0052	0.3137

From these comparisons, ANN(15,9,1)**12** has lower accuracy measurements MAE, MAPE and RMSE as compared to SARIMA(0,0,0)x(2,1,1)**12** in both in-sample and out-sample testing. Hence, this can be concluded that ANN(15,9,1)**12** is a better model to forecast the rainfall in Cameron Highland as compared to SARIMA(0,0,0)x(2,1,1)**12.**

4 Conclusions

In conclusion, ANN(15,9,1)$_{12}$ is selected as the best model in fitting the data and forecasting the rainfall data of Cameron Highland. Mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) of this model achieved lower value for both in-sample and forecasting data in comparison to the SARIMA(0,0,0)x(2,1,1)$_{12}$ model by .

5 Acknowledgement

The authors would like to acknowledge the funder by the Ministry of Higher Education, Malaysia under FRGS, Registration Proposal No: FRGS/1/2020/STG06/UTM/02/6 and R.J130000.7854.5F370 and UTM Fundamental Research with vot no Q.J130000.2554.21H65.

References

[1] Hung, N. Q., Babel, M. S., Weesakul, S., & Tripathi, N. K. (2009). An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrology and Earth System Sciences, 13(8), 1413–1425. Retrieved from https://doi.org/10.5194/hess-13-1413-2009.

[2] Mekanink F. et al. (n.d.). Rainfall modelling using Artificial Neural Network for a mountains region in West Iran. Retrived from https://www.mssanz.org.au/modsim2011/I5/mekanik.pdf.

[3] Neural Network Definition. (2020, December 23). Investopedia. Retrieved from https://www.investopedia.com/terms/n/neuralnetwork.asp.

[4] Neural network models | Forecasting: Principles and Practice (2nd ed). (2016). Forecasting: Principles and Practice (2nd Ed). https://otexts.com/fpp2/nnetar.html.

[5] Nwokike, C. C., Offorha, B. C., Obubu, M., Ugoala, C. B., & Ukomah, H. I. (2020). Comparing SANN and SARIMA for forecasting frequency of monthly rainfall in Umuahia. Scientific African, 10, e00621. Retrieved from https://doi.org/10.1016/j.sciaf.2020.e00621.

[6] Sato, R. C. (2013). Gerenciamento de doenças utilizando séries temporais com o modelo ARIMA. *Einstein (São Paulo)*, *11*(1), 128–131.https://doi.org/10.1590/s1679-45082013000100024.

[7] Somvanshi V. K. et al. (2006). Modelling and prediction of rainfall using artificial neural network and ARIMA techniques (Vol.10, No.2, pp.141-151).

[8] Zhang, G., Eddy Patuwo, B., & Y. Hu, M. (1998). Forecasting with artificial neural networks: International Journal of Forecasting, 14(1), 35–62. https://doi.org/10.1016/s0169-2070(97)00044-7.

[9] Zhang, X. (2013, May 1). *Comparative Study of Four Time Series Methods in Forecasting Typhoid Fever Incidence in China*. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0063116