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

Vol. 10, 2022, page 142 - 152

Forecasting Amount of Rainfall in Peninsular Malaysia Using Holt-Winter Method and Seasonal Autoregressive Integrated Moving Average (SARIMA)

Lee Chi Poh, Siti Mariam Norrulashikin*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: sitimariam@utm.my

Abstract
Rainfall is an important component in many developing countries especially to those agriculture producing country this is because water is vital for the maintenance of physiological and chemical processes that takes place within plant structures. As result, forecasting the amount of rainfall critical for the government as well as the rescue team. This study aims to model monthly amount of rainfall in Peninsular Malaysia using Autoregressive Integrated Moving Average (ARIMA) and Holt-Winter models. The models' performance is measured using Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). In this study, we are using the amount of rainfall data from January 2010 until September 2019 from six different stations which is Senai, Pulau Langkawi, Kuala Terengganu, Gong Kedak, Cameron Highlands and Alor Setar. The data from January 2010 until December 2018 were used as the training set data to fir the forecasting model, while the 9 months data from January 2019 until September 2019 used to assess the forecasting accuracy of the models that have been implemented. The forecasting model with the lowest prediction error is chosen. This studies shows that ARIMA in overall performance is better than Holt-Winter method. Hence, ARIMA is considered the best model to forecast the amount of rainfall.

Keywords: Rainfall; ARIMA; Holt-Winter method

1. Introduction

Time series analysis is the most widely used field of data science and machine learning, it decomposes the past historical data to depict the trend, seasonality, and noise to derive the future trends from it. It's a type of predictive analysis that forecasts the value of a variable in future occurrences based on history. The predicted values can be influenced by certain external factors which are known as independent variables like in the case of sale of a product is influenced by the discount percentage on its prices or the temperature on a particular day is influenced by the humidity or wind speed etc.

Much like other statistical analysis, by using time series people can also perform analysis and this type of analysis is so called the time series analysis. Time series analysis covers methods for evaluating time series data in order to derive relevant statistics and other properties of the data. The main objective of time series analysis is to comprehend or model the behaviour of observed series and the changing pattern over time. Knowing well about the underlying naturalistic process helps in evaluating the consequences of either planned or unplanned intervention. Furthermore, time series analysis may assist in forecasting the future values of the series, and managing the quality of the process in statistical quality control.

2. Literature Review

This chapter will discuss pros and cons of rainfall brings to human beings and a brief explanation on rainfall. Moreover, this chapter will also discuss the forecasting model that was mentioned in Chapter 1 which is Holt-Winter method and ARIMA model. Furthermore, we will also review some article which related to the application of Holt-Winter method and ARIMA model.

2.1. Rainfall

Rainfall is the most recognized triggering factor of disaster such as flood, landslide, cause damage to buildings and infrastructure and loss crops and livestock, especially in tropical regions with hot and humid climatic conditions. Based on the Climate Prediction Center Merged Analysis of Precipitation (CMAP) data [18], the June–September rainfall in the Asian-Pacific summer monsoon (APSM) region accounts for approximately 30% of the total tropical rainfall, albeit this region occupies only 10% of the tropics between 30S and 30N. The rainfall in this region plays a critical role in maintaining the global energy/water cycle and driving the monsoon climate variability and has farreaching impacts on global circulation [18]

2.2. ARIMA Model

Autoregressive integrated moving average (ARIMA) is one of the most important and widely used in time series model [19]. The popularity of the ARIMA model is due to its statistical properties as well as the well-known Box–Jenkins methodology in the model building process. In addition, various exponential smoothing models can be implemented by ARIMA models. In ARIMA model, it is assumed that the future value and the past observation values of the time series satisfy the linear relationship. In fact, most of the time series data contain nonlinear relationship, which limits the scope of the application of ARIMA model.

2.3. Holt-Winter Model

The main intention of the time series analysis of the rainfall data is to find out trend, and seasonality in the data and use this information for forecasting the region-wise rainfall for the future [4]. Holt-Winter method is one of the traditional methods used to forecasting a leveled time series data. While the past observations moving averages are weighted equally, Exponential Smoothing allocates rapidly reducing weights as the observations become huge. Holt-Winter's method can be classified into Additive and Multiplicative effect assumption. Manideep et al. (2018) made a prediction using different methods of Holt-Winters algorithm with big data approach and it is found out that multiplicative method works better than additive method in most of the cases.

3. Methodology

3.1. Introduction

In this study, we will apply either the Multiplicative Holt Winters (MHW) or Additive Holt Winter (AHW) methods to predict the rainfall for a given dataset. The decision in choosing between additive method and multiplicative method depends on the time series characteristics as well because different methods will be suitable for different data and each method has its drawback as well. When the seasonal component is directly proportional to trend level the additive method should be preferable, whereas if size of the seasonal component is directly proportional to trend level multiplicative method should be preferable.

3.2. ARIMA Model

The ARIMA approach, has been one of the most widely applied linear frameworks in time-series forecasting over the past three decades [19]. The popularity of the ARIMA model is due to its statistical properties and it is well-known as Box–Jenkins methodology [6]. The popularity of ARIMA arose from its analytical ability, which is valuable in the process of model building, and the quality of its forecasts [2]. In addition, various exponential smoothing models can be implemented by ARIMA models [11]. Because the ARIMA model implies a linear correlation structure between the time series values, no nonlinear features can be captured.

AR: p = order of the autoregressive part, I: d = degree of differencing involved and

MA: p = order of the moving average part.

The autoregressive model can be represented as follows:

$$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t$$

where ε_t is the white noise, c is the arbitrary constant, $\emptyset_p$ represent the regression coefficient and y_t is the actual value at time t.

A moving average model of order q can be written as:

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where ε_t is the white noise, μ is the mean of the model, and θ_q is the coefficient of the moving average to the estimated.

The combination of differencing with autoregression and moving average model will form the non-seasonal ARIMA model as below:

$$y'_t = c + \emptyset_1 y'_{t-1} + \cdots + \emptyset_p y'_{t-p} + \cdots + \emptyset_q y_{t-q} + \varepsilon_t$$

where y'_t is the differenced series.

The differencing process builds ARIMA (p, d, q) with p as the order of the autoregressive part, d as degree of first differencing involved and q as order of the moving average part. The general equation of the ARIMA model can be written in backshift notation, B as

$$\emptyset_p(B)(1-B)^d y_t = \theta_q(B)\varepsilon_t$$

with

$$\emptyset_p(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

where B is the backshift and defined as $By_t = y_{t-1}$.

3.3. Holt-Winter

Holt-Winter's method is advanced the double exponential smoothing to model time series with seasonality. The method is also known as the Triple exponential smoothing in respects of the term of the discoverers. Holts-Winter's method is developed by the Holt's Linear method with adding a third parameter to deal with seasonality. Therefore, this method consents for leveled time series while the level, trend, and seasonality are different. There are two main differences in the Holt-Winter model: trend and seasonality and they mostly depend on the type of seasonality. To grip seasonality, a third parameter is added in this model.

Holt-Winter's model can continue to provide forecast with the same accuracy over time. The Holt-Winter's method has an additive and a multiplicative form. The difference between those forms is in the nature of the seasonal component

3.3.1 Additive Holt-Winter

The additive model is preferred when the variation of seasonal component is almost stable through the series while the multiplicative is used when seasonal variations changes proportionally to the level of the series. Holt winter's method have three parts which is *Lt* is the level at time *t*, *Tt* is the trend at time *t*, *St* is the seasonal component at time *t*. The additive model is represented by the model as in :

Level :
$$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1})$$

Trend :
$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

Seasonal :
$$S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s}$$

Forecast :
$$F_{t+m} = L_t + b_t m + S_{t-s+m}$$

3.3.2 Multiplicative Holt-Winter

The multiplicative method is chosen when the variations in seasonal changed proportional to the level of the series while, the multiplicative model is represented as in

Level :
$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$$

Trend :
$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

Seasonal :
$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s}$$

Forecast :
$$F_{t+m} = (L_t + b_t m)S_{t-s+m}$$

From the equation, L_t is the level; b_t is the trend; S_t is the seasonal; Y_t is the value from the given data, while *t* is the time period for the component of L_t, b_t, S_t, and Y_t. *Ft* is the forecast value of a period ahead. Additionally α : level, β: trend and γ: seasonal are smoothing coefficients. *m* is the forecast period and *s* is the seasonal duration.

3.4 Accuracy Checking

The error measurement component is important to measure the accuracy and the suitability of the forecasting method in order to demonstrate the effectiveness of the forecasting result. In this work, Mean Absolute Percentage Error (MAPE) and root mean square error (RMSE) will be used to measure the accuracy of the Holt- Winter's method. The equations are shown below:

$$\text{MAPE} = \frac{1}{n}\sum\nolimits_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| * 100$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

Given n is the value of total observations, y_t is the actual value and $\hat{y}_t$ is the forecast value of period t.

4. Results and discussion

4.1 Data Collection Analysis

In this research, we collected 6 data set from Peninsular Malaysia which is Senai, Pulau Langkawi, Kuala Terengganu, Gong Kedak, Cameron Highlands and Alor Star and labeled as Station A, Station B, Station C, Station D, Station E and Station F respectively.

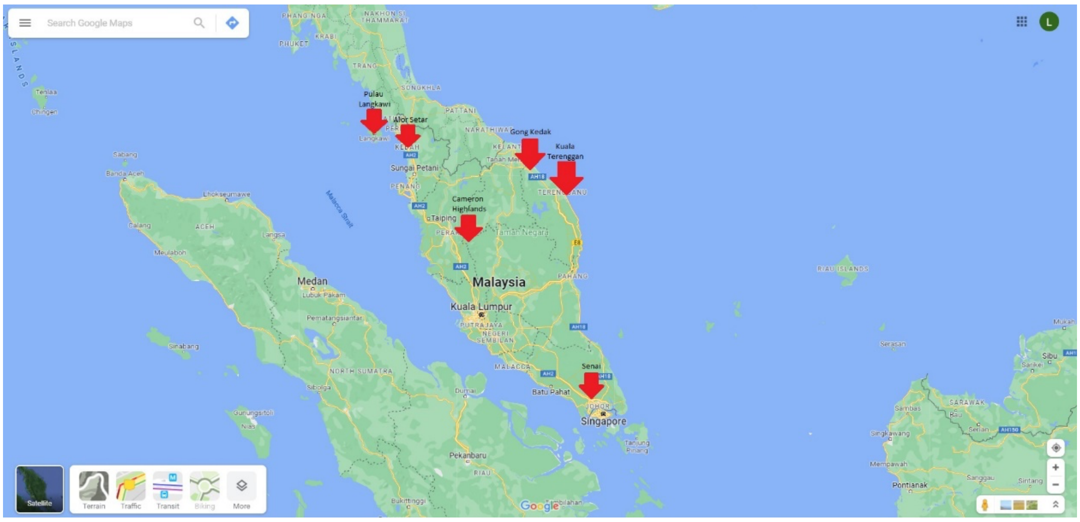
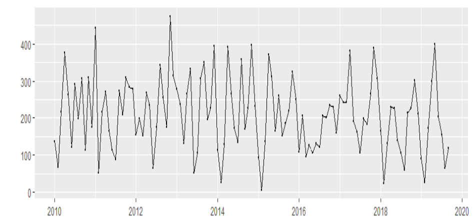
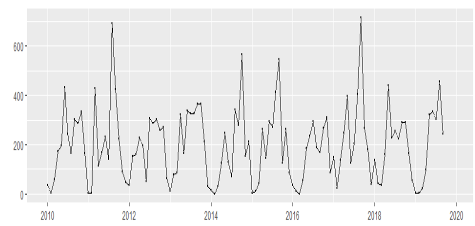
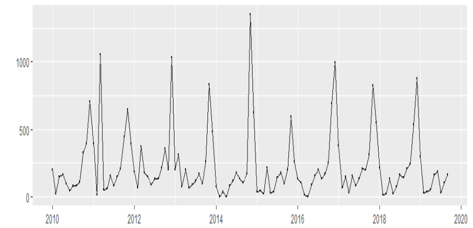
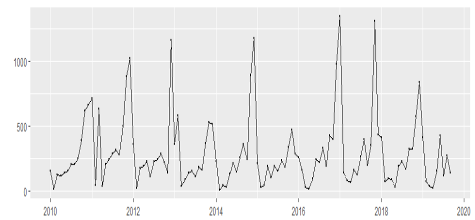
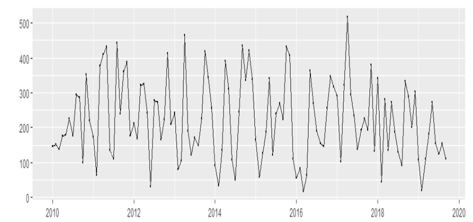
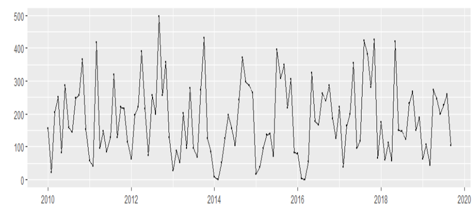


Figure 1: Selected Stations

Table 1: Time Series Plot Across All Station

 Senai (Station A)	 Pulau Langkawi (Station B)
 Kuala Terengganu (Station C)	 Gong Kedak (Station D)
 Cameron Highland (Station E)	 Alor Setar (Starion F)

4.2 Holt-Winter's Method

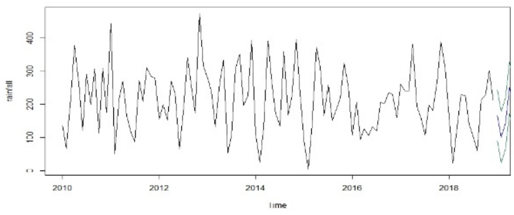
In this research, the measurement of the original monthly amount of rainfall taken from January 2010 until September 2019 by using mathematical software R-Studio to conduct our analysis.

From the time series plot in Table 1, we observed that it has several fluctuations over time. There are several peaks of the amount of rainfall that occurred. The amount of rainfall has presented inconsistent trend which is changing over time. We decided to use Multiplicative Holt-Winter's (MHW) method to solve the data. Since the data fluctuated and behave seasonality, Additive Holt-Winter (AHW) is not suitable to forecast the data. The data set are divided into 2 sets which is the training data, use to mdoel the in-sample data, while the testing data is to compare the original data with the out sample data which is forecasted from the training data.

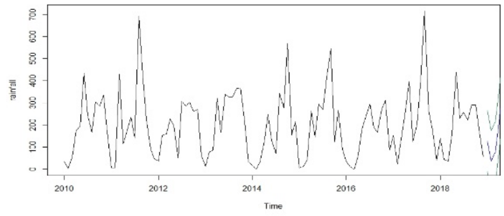
Table 2: Parameter estimation of AHW

Stations		A	B	C	D	E	F
Smoothing Parameters	α, Alpha	0.0238	0.0293	0.0573	0.1077	0.1270	0.0049
	β, Beta	0.0748	0.0892	0.0807	0.0368	0.0446	0.4867
	γ, Gamma	0.4176	0.3251	0.5426	0.2368	0.3882	0.3555

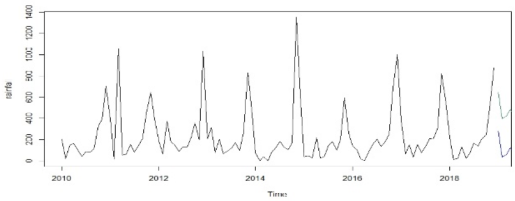
Table 3: AHW plot with the Forecasted Value



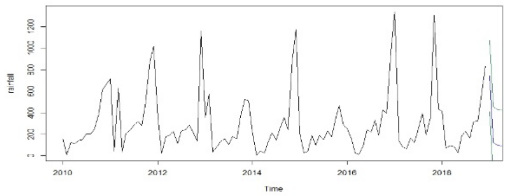
Senai (Station A)



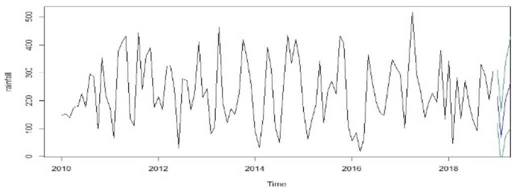
Pulau Langkawi (Station B)



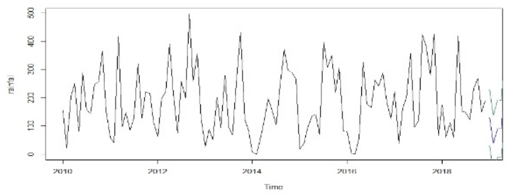
Kuala Terengganu (Station C)



Gong Kedak (Station D)



Cameron Highland (Station E)



Alor Setar (Starion F)

The black colour line was the original data and the blue is the forecasted value with the green line as the upper boundary and lower boundary.

4.3 Box-Jenkins Method

Data analysis of Box-Jenkins method can be classified into four main steps. They are model identification, parameter estimation process, diagnostic checking and forecasting process. This research are conducted by using R-Studio software for our analysis process.

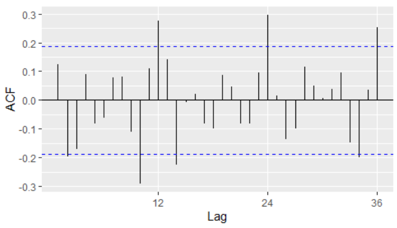
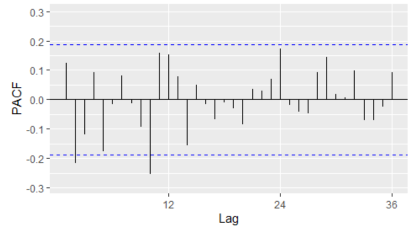
Table 4: ADF test for training data across all station

Augmented Dickey-Fuller Test	
Station	p-value
A	0.4739
B	0.2552
C	0.3633

D	0.5439
E	0.2182
F	0.5355

Table 4 shows the ADF test for the training data set across all station with lag = 12.

Table 5: ACF and PACF plot of Station A

Station	ACF	PACF
A		

From the ACF and PACF plot of station A from Table 5, we could conclude that the data reveals a seasonality and the ACF and PACF shows some seasonality across the lags. This is becasue ACF graph show a significant spike at lag 12, 24 and 36. This indicates that the data consist of seasonality. Hence, the seasonal difference should performance at lag = 12, because our data behave monthly.

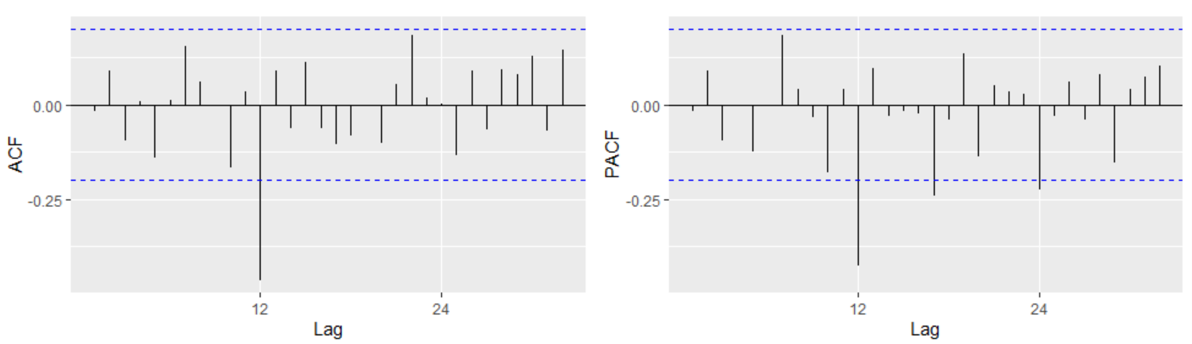


Figure 2: ACF and PACF after seasonal differencing of Station A

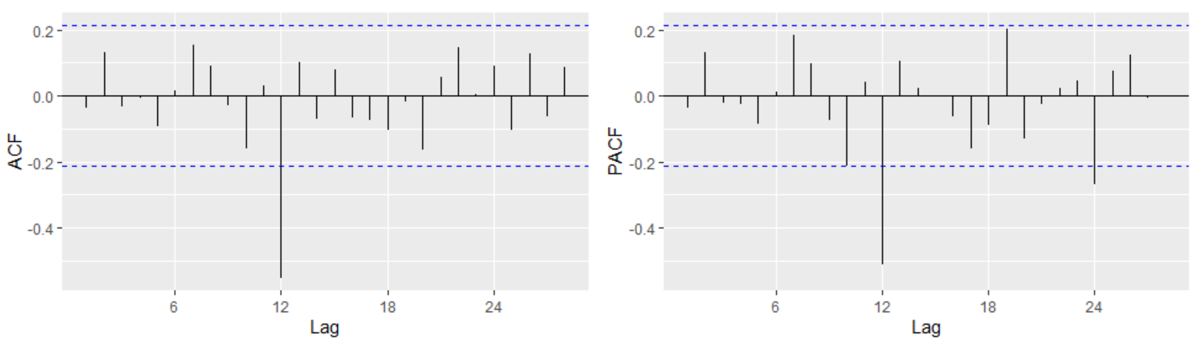


Figure 3: ACF and PACF after second seasonal differencing of Station A

After the second seasonal differencing, we check again the diagnostic with ADF test. ADF shows that the p-value = 0.036 which is smaller than 0.05. This indicates that H_0 is rejected at 5% significant level and the data is stationary.

Hence, we can form our SARIMA model with the information given in Figure 3 which is $SARIMA(0,0,0)(2,2,1)_{12}$, $SARIMA(0,0,0)(2,2,0)_{12}$, $SARIMA(0,0,0)(1,2,1)_{12}$, $SARIMA(0,0,0)(1,2,0)_{12}$, $SARIMA(0,0,0)(0,2,1)_{12}$, and $SARIMA(0,0,0)(0,2,0)_{12}$.

Table 6: AIC of all the possible SARIMA model

ARIMA	AIC
(0,0,0), (2,2,1)	1059.021
(0,0,0), (2,2,0)	1073.733
(0,0,0), (1,2,1)	1060.354
(0,0,0), (1,2,0)	1089.355
(0,0,0), (0,2,1)	1074.817
(0,0,0), (0,2,0)	1130.679

From Table 6, we can said that $SARIMA(0,0,0)(2,2,1)_{12}$ is the most suitable model to forecast since it has the smallest value of AIC.

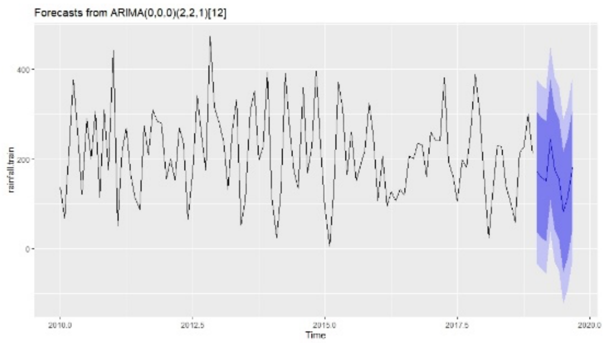


Figure 4: Fitted plot of Station A

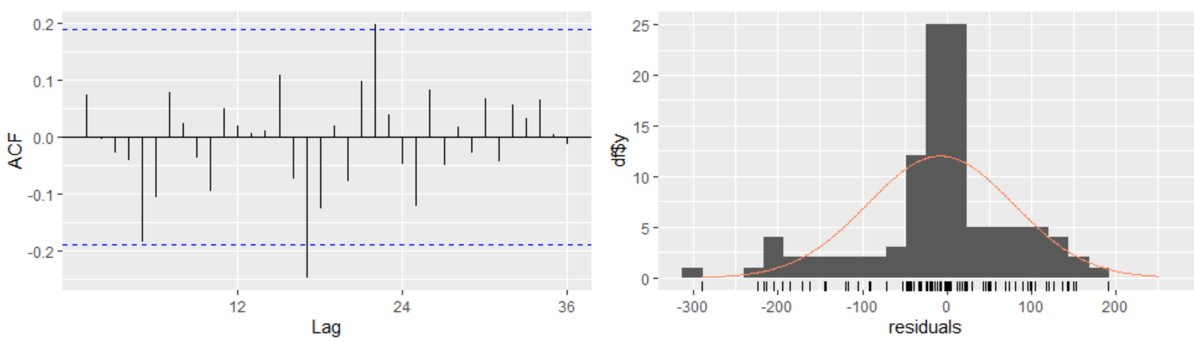


Figure 5: Residual plot for $\boldsymbol{SARIMA(0,0,0)(2,2,1)_{12}}$

The residual plot in Figure 4.5 behaves stationary with a constant variance and the ACF plot does not have significant spike cut off the standard error limit. But the residual histogram consist of 2 mode which might bring out the bi-modal distribution. We can conclude that the model is adequate as the residuals nearly the properties of white noise.

4.4 Error Estimation

Table 7: Error estimation of AHW and SARIMA model

Station	Model	MAPE	RMSE
A	$SARIMA(0,0,0)(2,2,1)_{12}$	98.2402	100.5456
	Additive Holt-Winter	73.1299	89.8384

Conclusion

The performances of the forecasting for these deployed models were evaluated by using the error metrics namely Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). In this study, the best prediction model of SARIMA was displayed in Table 7.

In conclusion, the SARIMA model is more appropriate model to forecast the amount of rainfall in station senai than AHW model.

References

[1] Abdel-Aal, R. E. (2008) ‘Univariate modeling and forecasting of monthly energy demand time series using abductive and neural networks’, Computers and Industrial Engineering, 54(4), pp. 903–917.
[2] Assis, K., Amran, A. and Remali, Y. (2010) ‘Forecasting Cocoa Bean Prices Using Univariate’, Journal of Arts Science & Commerce, 1(1), pp. 71–80.
[3] Bandyopadhyay, G. (2016) ‘Gold Price Forecasting Using ARIMA Model’, Journal of Advanced Management Science, 4(2), pp. 117–121.
[4] Dhamodharavadhani, S. and Rathipriya, R. (2019) Region-Wise Rainfall Prediction Using MapReduce-Based Exponential Smoothing Techniques, Advances in Intelligent Systems and Computing. Springer Singapore.
[5] Elmunim, N. A., Abdullah, M., Hasbi, A. M. and Zaharim, A. (2015) ‘Forecasting ionospheric delay during quiet and disturbed days using the Holt-Winter method’, International Conference on Space Science and Communication, IconSpace, 2015-Septe, pp. 132–135.
[6] G.E.P. Box, G. J. (1970) *Time Series Analysis: Forecasting and Control*.
[7] Hendrix, C. S. and Salehyan, I. (2012) ‘Climate change, rainfall, and social conflict in Africa’, *Journal of Peace Research*, 49(1), pp. 35–50.
[8] Jadhav, V., Chinnappa Reddy, B. V. and Gaddi, G. M. (2017) ‘Application of ARIMA model for forecasting agricultural prices’, *Journal of Agricultural Science and Technology*, 19(5), pp. 981–992.
[9] KARA, E. and DİKEN, A. (2020) ‘CLIMATIC CHANGE: THE EFFECT OF RAINFALL ON ECONOMIC GROWTH’, *Süleyman Demirel Üniversitesi Vizyoner Dergisi*, 11, pp. 665–679.
[10] Majed Saleh Al-Hafid, D. and Hussein Al-maamary, G. (2012) ‘Short term electrical load forecasting using holt-winters method’, *AL-Rafdain Engineering Journal (AREJ)*, 20(6), pp. 15–22.
[11] McKenzie, E. (1984) ‘General exponential smoothing and the equivalent arma process’, *Journal of Forecasting*, 3(3), pp. 333–344.

[12] Perone, G. (2021) 'Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy', *European Journal of Health Economics*. Springer Berlin Heidelberg, (0123456789).

[13] Rahman, A., Al-adawiyyah, D. A. S., ana, M., Rivika, S. V., Adnan, A. and Yendra, R. (2021) 'Holt-Winter Forecasting Method for Inflow and Outflow of Bank Indonesia in Riau', *International Journal of Economics and Management Studies*, 8(7), pp. 71–76.

[14] Razali, S. N. A. M., Rusiman, M. S., Zawawi, N. I. and Arbin, N. (2018) 'Forecasting of Water Consumptions Expenditure Using Holt-Winter's and ARIMA', *Journal of Physics: Conference Series*, 995(1).

[15] Tsay, R. S. (2000) 'Time series and forecasting: Brief history and future research', *Journal of the American Statistical Association*, 95(450), pp. 638–643.

[16] Vilas Jadhav and B. V. C. Reddy and G. M. Gaddi (2017) 'Application of ARIMA Model for Forecasting Agricultural Prices', *Journal of Agricultural Science and Technology*, 19, pp. 981–992.

[17] Wolters, G. K. J. and Hassler, U. (2012) *Introduction to Modern Time Series Analysis*. 2nd edn. New York Dordrecht London: Springer Heidelberg.

[18] Xie, P. and Arkin, P. A. (1997) 'Global Precipitation: A 17-Year Monthly Analysis Based on Gauge Observations, Satellite Estimates, and Numerical Model Outputs', *Bulletin of the American Meteorological Society*, 78(11), pp. 2539–2558.

[19] Zhang, P. G. (2003) 'Time series forecasting using a hybrid ARIMA and neural network model', *Neurocomputing*, 50, pp. 159–175.