Journal of Biostatistics and Biometric Applications
Volume 4 | Issue 1
ISSN: 2455-765X

Research article Open Access

Comparison of Statistical Techniques for Forecasting Malaria Cases in Ghana

Twumasi-Ankrah S[*1], Pels WA[1], Nyantakyi KA[2] and Addo DK[1]

[1]Department of Mathematics, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana
[2]Department of Management Science, School of Business, Ghana Institute of Management and Public Administration, Accra, Ghana

[*]**Corresponding author:** Twumasi-Ankrah S, Department of Mathematics, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana, Tel: +233244974531, E-mail: stankrah2017@gmail.com

Citation: Twumasi-Ankrah S, Pels WA, Nyantakyi K, Addo DK (2019) Comparison of Statistical Techniques for Forecasting Malaria Cases in Ghana. J Biostat Biometric App 4(1): 102

Received Date: November 09, 2019 **Accepted Date**: June 17, 2019 **Published Date:** June 19, 2019

Abstract

Background and Aim: The purpose of this study was to determine an appropriate statistical technique for forecasting the monthly Malaria cases in Ghana.

Methods and Materials: Monthly data spanning from January 2008 to December 2017 were obtained from the District Health Information Management System (DHIMS) 2, Ghana Health Service. The four competing forecasting techniques that were applied to the Malaria cases data were the Seasonal Autoregressive Integrated Moving Average (SARIMA), Artificial Neural Network (ANN), Exponential smoothing (ETS) and a Combination technique. The four competing forecasting techniques were compared using their respective forecast accuracy measures in order to choose the appropriate technique for forecasting Malaria cases in Ghana.

Results: It was observed that the SARIMA technique was the appropriate statistical technique. The "best" model for forecasting the monthly malaria cases in Ghana was SARIMA $(2, 1, 0)\ (2, 0, 0)_{12}$ which passed all the required diagnostic tests.

Conclusion: A two-year monthly forecast from the "best" model revealed that, in 2018, we should expect a decrease in Malaria cases in the last quarter but should expect an increase in Malaria cases during the first half of 2019.

Keywords: Malaria Cases; SARIMA; Artificial Neural Network; Forecast Accuracy; Exponential Smoothing

Introduction

Malaria is one of the key micro parasitic diseases causing human mortality in the world [1,2]. According to [3], there were an estimated 216 million cases of malaria in 91 countries in 2016, this accounted for about 80% of the global malaria burden; and an increase of 5 million cases over 2015 [3]. About 90% of the world's 300 - 500 million malaria cases and 1.5 - 2.7 million deaths due to malaria annually are from the sub-Saharan Africa [4,5]. The rate of malaria as indicated by [5] still records 40% of all outpatient attendance, with the most vulnerable people being under 5 children and pregnant women [6,7]. There were an estimated 445,000 deaths from malaria globally in 2016 as compared to 446, 000 estimated deaths in 2015 [3]. The Sub-Saharan Africa, accounted for 91% of all malaria deaths in 2016, followed by the South - East Asia Region (6%) [3].

In 1995, the estimated total cost of malaria to Africa was US$ 1.8 billion and US$ 2 billion in 1997 [8]. In 2016, 74% of an estimated US$ 2.7 billion invested in malaria control and elimination efforts globally was spent on Sub-Sahara African countries [3]. Thus, malaria is not only a public health problem but also a developmental problem.

In Ghana, malaria is the number one cause of morbidity accounting for 40 - 60% of out-patient [6]. According to [18], the burden of malaria in Ghana in 2002 was estimated at US$ 2.63 per capita or US$ 13.51 per households [18]. The 2015–2020 Ghana Malaria Strategic Plan which was recently adopted aims at reducing malaria burden by 75.0% [9,10]. According to [11] there are few readily available mathematical models to support decisions for planning and evaluation of these strategies remains a challenge [11]. Thus, forecast of malaria cases is vital for its control and intervention. Because of its devastating effects, some researchers in Ghana have tried to come up with appropriate models to forecast malaria cases, which will help policy makers. The gaps in some of the research articles on forecasting malaria cases in Ghana are discussed.

Takyi Appiah S, *et al*., indicated that the ARIMA (2, 1, 1) model was the "best" fitted model for forecasting the number of malaria cases in the Ejisu-Juaben Municipality for a period of two years from 2014 and 2016. The authors employed only the graphical method to examine whether the data was stationary or not, which is not evidence enough [12].

Alhassan EA, *et al.*, identified an ARIMA (1, 0, 1) model for forecasting malaria cases in Kasena Nankana municipality in Ghana. They indicated that malaria cases in the Navrongo municipality were growing at a constant quadratic rate. However, their "best" model did not capture the trend nature of the data, hence assumed a non-stationary data as stationary [13].

Furthermore, Bosson-Amedenu S used the Box-Jenkins (ARIMA) methodology on a monthly malaria incidence data of children below the age of five in Edum Banso. The choice of the order of the differencing was not justified since only the graphical approach was used to check for stationarity [14].

Again, Anokye R, *ε.*, used the quadratic model for forecasting the half year incidence of Malaria whiles ARIMA (1, 1, 2) was used for forecasting monthly malaria incidence for the years 2018 and 2019 in Kumasi Metropolis [15]. A major drawback on their work is that, in the ARIMA modelling, the authors treated their data as a stationary data whereas the data was non-stationary and thus require differencing (that is, their data exhibited a quadratic trend and was also a seasonal data).

From the above literatures, it is obvious that several of these studies only considered the graphical method to decide whether a series was stationary and not conducting a formal test of hypothesis like the unit root test (e.g. the Augmented Dickey Fuller Test). Also, none of these studies considered different forecasting techniques in order to select the most appropriate one; and again, the above studies did not consider forecasting malaria cases for the entire country. Hence this study seeks to develop and validate an appropriate forecasting technique that can be used to forecast malaria transmission in Ghana. This will bring out a solid model by comparing several competing models using different forecasting techniques. This study employs three different forecasting techniques: Autoregressive Integrated Moving Average (ARIMA) model, Artificial Neural Network (ANN) and Exponential smoothing approach (ETS) to model and forecast malaria cases in Ghana.

Method

Data Source

Data used for the study are monthly malaria cases spanning from January 2008 to December 2017 and were obtained from the Ghana Health Service database called the District Health Information Management System (DHIMS)2. For the purpose of cross validation of forecasting techniques, we used the data period from 2008 to 2015 to fit the model while 2016 and 2017 period was used for out-sample forecast accuracy measure.

The Unit Root Test: The unit root test considered in this study is the Augmented Dickey-Fuller (ADF) test which has the null hypothesis that the time series has a unit root (that is, it is not stationary). Thus, the alternative hypothesis is that the time series does not have a unit root (that is, it is stationary). This test is made up of approximating the regression model below:

$$\Delta Y_t = \alpha + \alpha_1 t + \gamma Y_{t-1} + \sum_{i=1}^{m} \beta \Delta Y_{t-1} + \varepsilon_t \qquad (2.1)$$

Where ε_t is a pure white noise error term and the ADF test follows an asymptotic distribution.

Autoregressive Integrated Moving Average (ARIMA) Model: A non-stationary ARMA (p + d, q) process is written in a lag operator form as:

$$\phi'(L) y_t = \theta(L)\varepsilon_t \qquad (2.2)$$

where the roots of $\phi'(L)y_t$ is the lag operator form of autoregressive model, while $\theta(L)\varepsilon_t$ is the lag operator form of the moving average. In similar instances, equation (2) can be written as a stationary process ω_t so that $\phi'(L)\omega_t = \theta(L)\varepsilon_t$, where $\omega_t = \nabla^d y_t$. It can be said that y_t is an ARIMA (p, d, q) and ω_t is an ARMA (p, q). Conditionally, the tail off of the ACF and the PACF implies a mixed ARMA model.

Exponential Smoothing Technique (ETS): The various parts or components of time series include the trend (τ), cycle (c), seasonal (s), and irregular or error (ε) parts. All these could be added together in a diverse number of ways. A purely additive model can be expressed as

$$y = \tau + s + \epsilon \qquad (2.3)$$

Where the three parts are combined to form the series that is seen. The trend component, an amalgamation of a level term (τ_l) and a growth term (τ_g) at all times, is the beginning component in exponential smoothing. If the error part is overlooked, there are precisely fifteen exponential smoothing methods as shown in the table below.

To be clear, according to Hyndman RJ, Athanasopoulos G, the simple exponential smoothing method is defined by cell (N, N), Holt's linear method by cell (A, N), the damped trend method by cell (Ad, N), Holt-Winters' additive method by cell (A, A), and Holt-Winters' multiplicative method is given by cell (A, M) [16].

Trend Component	Seasonal Component		
	N (None)	A (Additive)	M (Multiplicative)
N (None)	N, N	N, A	N, M
A (Additive)	A, N	A, A	A, M
Ad (Additive Damped)	Ad, N	Ad, A	Ad, M
M (Multiplicative)	M, N	M, A	M, M
Md(Multiplicative Damped)	Md, N	Md, A	Md, M

Table 1: Variations in the Combination of the Trend and Seasonal Components

Model Selection Criteria

In this study, the Corrected Akaike information criterion (AICc) was used, and it is given by:

$$AIC_c = -2\ln\ \mathrm{L}(\hat{\theta}_k) + \frac{2kn}{n-k-2} \qquad (2.4)$$

where $L(\hat{\theta}_k)$ is the likelihood of the fitted model, $\mathrm{K} = \mathrm{p}+1$ representing size of model, p = total of parameters, and n is the total number of observation.

Diagnostic Checking

When the 'best' model is selected, it is important to access the model for likely inconsistencies. Having fitted the model, residuals are found to access the likely inconsistencies, as well as help to modify the selected model. It is necessary that the residuals satisfy the model assumptions. If this is not the case, then the model which has been fitted will not be statistically correct.

Forecast Accuracy Measures

Root Mean Square Error (RMSE): RMSE is a measure of spread of the forecast errors about the actual data points. This implies that the RMSE informs how far or near the forecasted values of an estimated model are from the real data points. The formula is given as:

$$RMS\ E_{forecast} = \sum_{i=1}^{N} \left(\frac{\hat{\mathrm{Y}}_t - Y_t}{N}\right)^{\frac{1}{2}} \qquad (2.5)$$

where $\hat{\mathrm{Y}}_t$ is the forecasted values, Y_t the actual data points, an N is the sample size.

Mean Absolute Percentage Error (MAPE): MAPE is a measure of the size of error of a forecast in percentage. It is used to measure the accuracy of a forecast using the formula below:

$$MAPE_{forecast} = \left(\frac{1}{N}\sum\frac{|\ Y_t - \hat{\mathrm{Y}}_t\ |}{|\ Y_t\ |}\right) \times\ 100\% \qquad (2.6)$$

Principle of the Analysis

For replication purposes, the principles were followed:

1) The series is plotted to observe the various features
2) Augmented Dickey-Fuller test is used to test for stationarity in the series
3) For each of the forecasting techniques, competing models are fitted to the malaria cases series; and the "best" model is selected by the minimum information criterion.
4) The "best" model from each of the forecasting techniques in step (3) is compared using their predictive performance or lowest accuracy measure to select the "best" method for forecasting malaria cases in Ghana.
5) The "best" forecasting technique in step (4) is used to forecast the malaria cases in Ghana.

Results

Firstly, we employed three different univariate time series techniques in this section to model and forecast malaria incidence in Ghana. The time series techniques are ARIMA, ETS and ANN. We compared these techniques by using their predictive performance or lowest accuracy measure to select the "best" method for forecasting malaria cases in Ghana. The R software, specifically forecast 8.3 package Hyndman R, *et al.*, is used to run the time series models. For each forecasting technique, appropriate competing models are constructed and their AICc's recorded. The model with the least AICc is chosen as the 'best' model for forecasting the malaria time series data [17].

Data Visualization

Time Series Plot: Generally, in Figure 1, there is a strong upward trend in malaria cases from 2008 to 2012, whereas from 2013 to 2017, the trend changes to a decreasing one. This is an indication that the malaria cases series is not stationary.

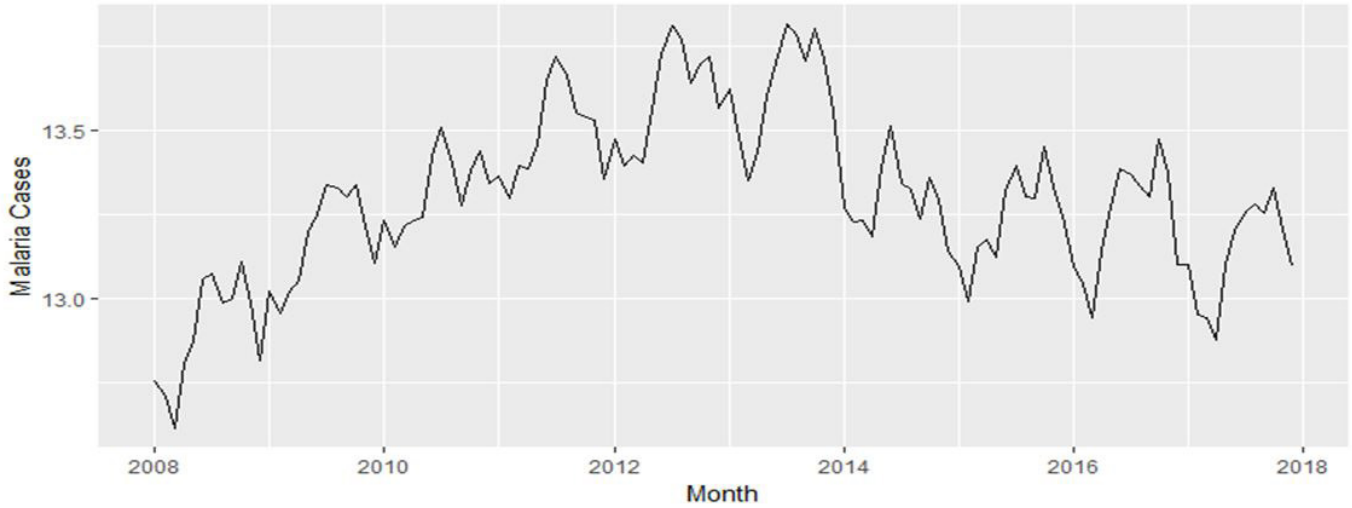

Figure 1: Time series plot of Monthly Malaria Cases (Jan. 2008 – Dec. 2017)

Seasonal Plot: In Figure 2, there is an observed sharp decrease in malaria cases at the start of each year (i.e., January and February) but relatively higher as compared to the cases at the end of the previous year. An upward trend in malaria cases tends to occur from March every year through to the middle (July) of each year where the annual peak mostly occurs. Finally, a downward trend is observed typically in the last quarter of every year as the number of malaria cases at the start of that year reduces by the end of the same quarter. Remarkably, malaria cases tend to be evidently lower during the dry season in Ghana than the wet season.

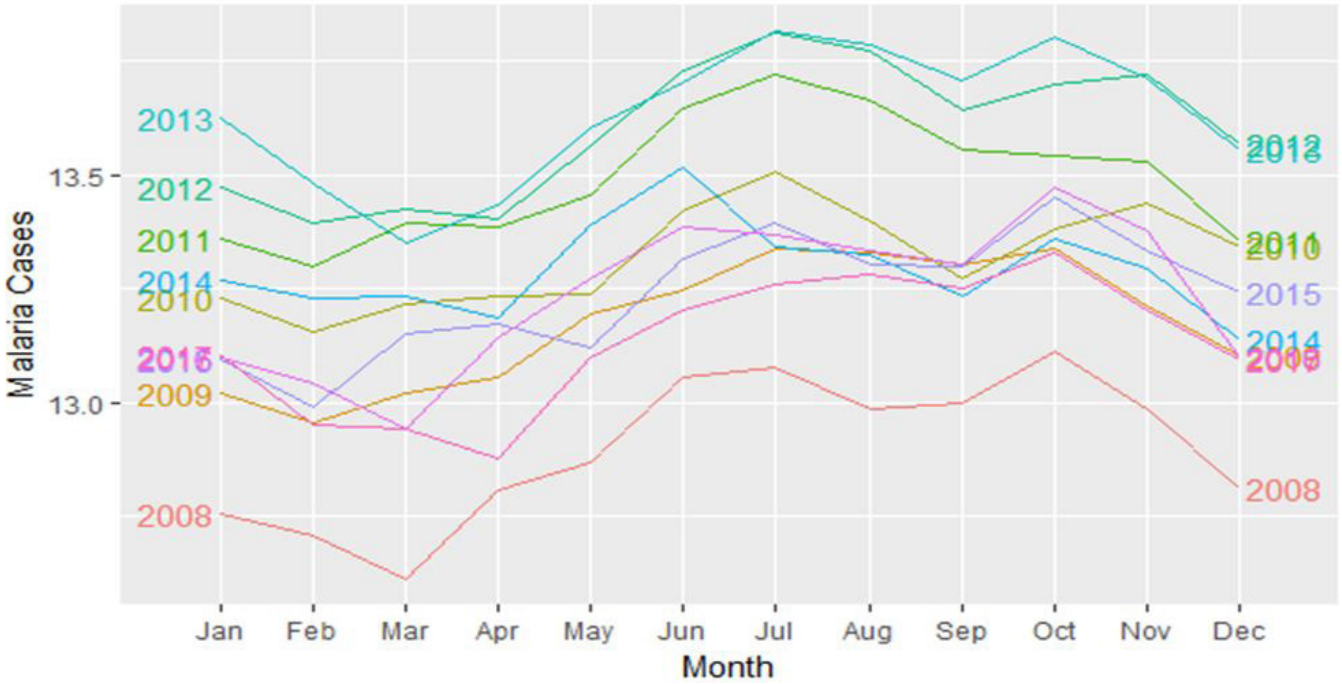

Figure 2: Seasonal plot of Monthly Malaria Cases (Jan. 2008 – Dec. 2017)

ARIMA Model

Test of Stationarity: As observed in Figures 1 and 2, the trend and seasonality imply that our time series data is non-stationary. This is confirmed by the Augmented Dickey-Fuller test in Table 2 which resulted in a p-value (0.3315) greater than 5% significance. Thus, we do not have enough evidence to reject the null hypothesis that the data was non-stationary. Nevertheless, a first difference of the Malaria data made it stationary, as confirmed by the same test.

Variable	Order of Differencing	P-value	Conclusion
Malaria Cases	I (0) (original data)	0.3315	The data is not stationary
Differenced Malaria Cases	I (1) (differed data)	0.01	The data is stationary

Table 2: ADF Unit root Test on Malaria Cases Data

Model Selection: In model building, the standard practice is to fit more than one model to the dataset in order to choose the "best" model. With the "differencing" information acquired at the test of stationarity in Table 2, a set of competing models and their respective information criterion are given in Table 3. The model with the minimum AICc value among the competing models is chosen as the "best" model, therefore SARIMA (0, 1, 2) $(0, 1, 1)_{12}$ is the "best" ARIMA model.

Model	AICc
(1, 1, 0) (1, 1, 0)	-178.28
(0, 1, 1) (1, 1, 0)	-181.52
(0, 1, 2) (1, 1, 0)	-182.12
(0, 1, 1) (0, 1, 1)	-187.88
(0, 1, 2) (0, 1, 1)*	-188.15
(1, 1, 1) (1, 1, 0)	-180.79
(1, 1, 1) (0, 1, 1)	-187.33
(1, 1, 1) (0, 1, 0)	-162.85
(1, 1, 2) (0, 1, 1)	-186.06
(1, 1, 2) (0, 1, 1)	-180.66

*The chosen model

Table 3: Competing models and their AICc values

Exponential Smoothing Technique

The appropriate exponential smoothing technique for the malaria series is the Holt-Winters' seasonal method. This is mainly because of the observed trend and seasonality in Figure 1 and 2. For the analysis, five competing exponential smoothing models were fitted to the Malaria data, all exhibiting either multiplicative error or additive error. Table 4 summarizes each of the competing models and their respective information criterion.

Model	AICc	BIC
M, N, M	88.7279	123.7425
M, Ad, M *	-19.0453	21.5477
M, Md, M	-18.6612	21.9317
A, A, N	43.5431	56.3655
A, Ad, A	-18.9906	21.6024

*The chosen model

Table 4: Competing Models with respective AICc and BIC

Conclusively, the 'best' model is ETS (M, Ad, M), because it has the least AICc value. This implies that the model that best fits the malaria data using the exponential smoothing technique is one that has a multiplicative error, a damped additive trend, and a multiplicative seasonality.

Modeling with Artificial Neural Network

Several competing artificial neural networks were constructed after setting seed; and NNAR (1, 1, 2) [12] model is considered as the "best". In Figure 3, the forecast values over a two year period, that is January, 2016, to December, 2017 from NNAR (1, 1, 2) [12] is plotted. It can be observed that the forecast does not bear much resemblance to the original malaria cases series.

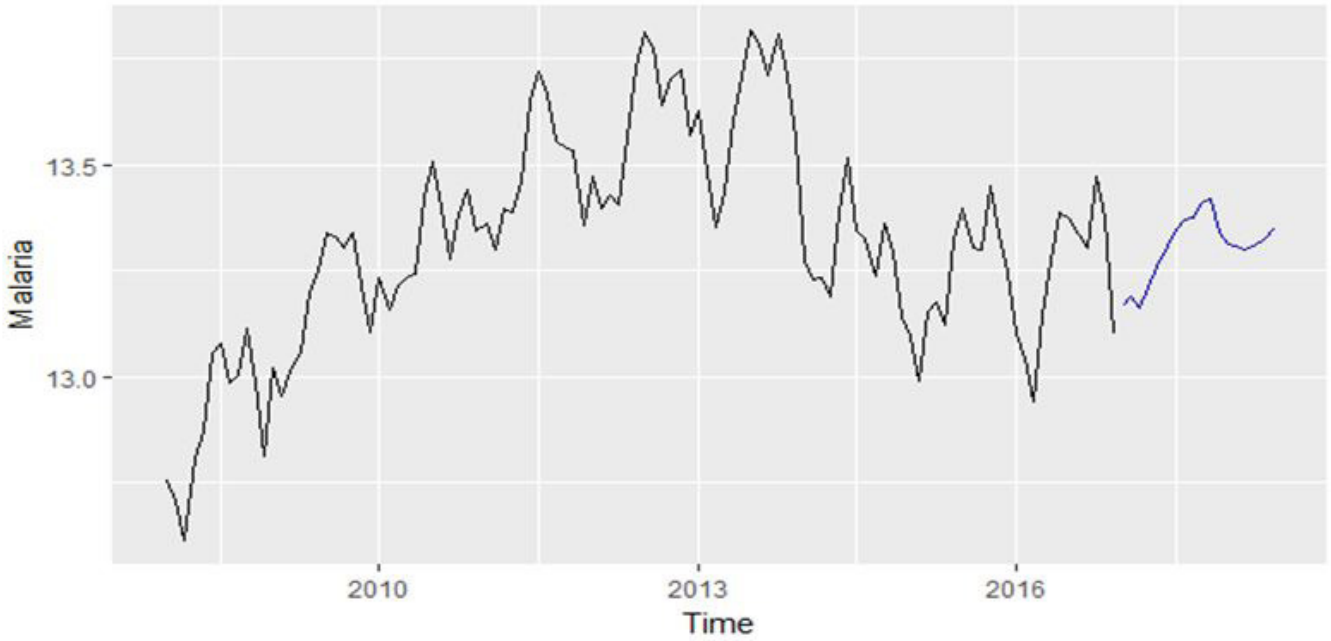

Figure 3: Forecast plot of malaria cases using NNAR (1, 1, 2) [12].

Comparison of Forecasting Techniques: ARIMA, ETS and NNAR

To select the appropriate forecasting technique for malaria cases in Ghana, the respective "best" forecasting models were separately used to forecast the next two years. The forecast values were then compared to the actual data (which is the test data), and the accuracy of each of the 'best' forecast technique was computed using the Root-Mean-Square Error (RSME) and the Mean Absolute Percentage Error (MAPE). From Table 5, the forecast values of the three original methods (ARIMA, ETS, NNAR) are combined by simple arithmetic mean to form the Combination technique. The RMSE and MAPE of the combined method and the other three methods are noted. It is obvious that the SARIMA model results have the least RMSE and MAPE values among the other three forecasting techniques. Thus, it is concluded that the SARIMA (0, 1, 2) $(0, 1, 1)_{12}$ model is the 'best' model for forecasting the malaria time series data in Ghana.

MODEL	RMSE	MAPE
SARIMA(0, 1, 2) $(0, 1, 1)_{12}$*	0.07727	0.4569
ETS (M, Ad, M)	0.1231	0.8141
NNAR (1, 1, 2)	0.1818	1.2421
Combination (ARIMA, ETS, NNAR)	0.1207	0.7958

*The chosen model

Table 5: Forecast Accuracy Measures of Competing Forecast Techniques

Estimates and Diagnostic Checking of the "Best" Model - SARIMA (0, 1, 2) $(0, 1, 1)_{12}$

From Table 6, SARIMA (0, 1, 2) $(0, 1, 1)_{12}$ was selected as the 'best' forecasting technique because it had the least forecast accuracy measure. From the diagnostic checking, the chosen forecast technique, SARIMA (0, 1, 2) $(0, 1, 1)_{12}$, fitted the malaria cases series well.

Variable	Estimates		Standard Error	P – Value
Malaria Cases	MA(1)	−0.2471	0.1042	0.0198
	MA(2)	−0.1575	0.0987	0.1140
	SMA(1)	−0.6235	0.1212	0.0000

Table 6: Parameter Estimates of the chosen ARIMA Model

Diagnostic Checking

In Figure 4, the diagnostic checks on the residuals of the chosen forecasting method [SARIMA (0, 1, 2) (0, 1, 1)12] is presented. This is done to see if it does not violate any of the assumptions. From Figure 4, we observed the following:

a) There is no apparent trend in the plot of the standardized residuals over the years.
b) As well, a plot of the ACF of the residuals confirms that none of their lags are statistically significant implying that the residuals are not correlated.

c) The Normal Q-Q plot shows that most of the errors are normally distributed except for a few at either ends that move away from the normal line. Again, potential cases of outliers are observed.

d) Finally, from the plot of the p-values for the Ljung-Box statistic, it is observed that all the p-values are above the 5% significance level. This indicates that the null hypothesis cannot be rejected, as well; the plotted lags are not significantly different from zero.

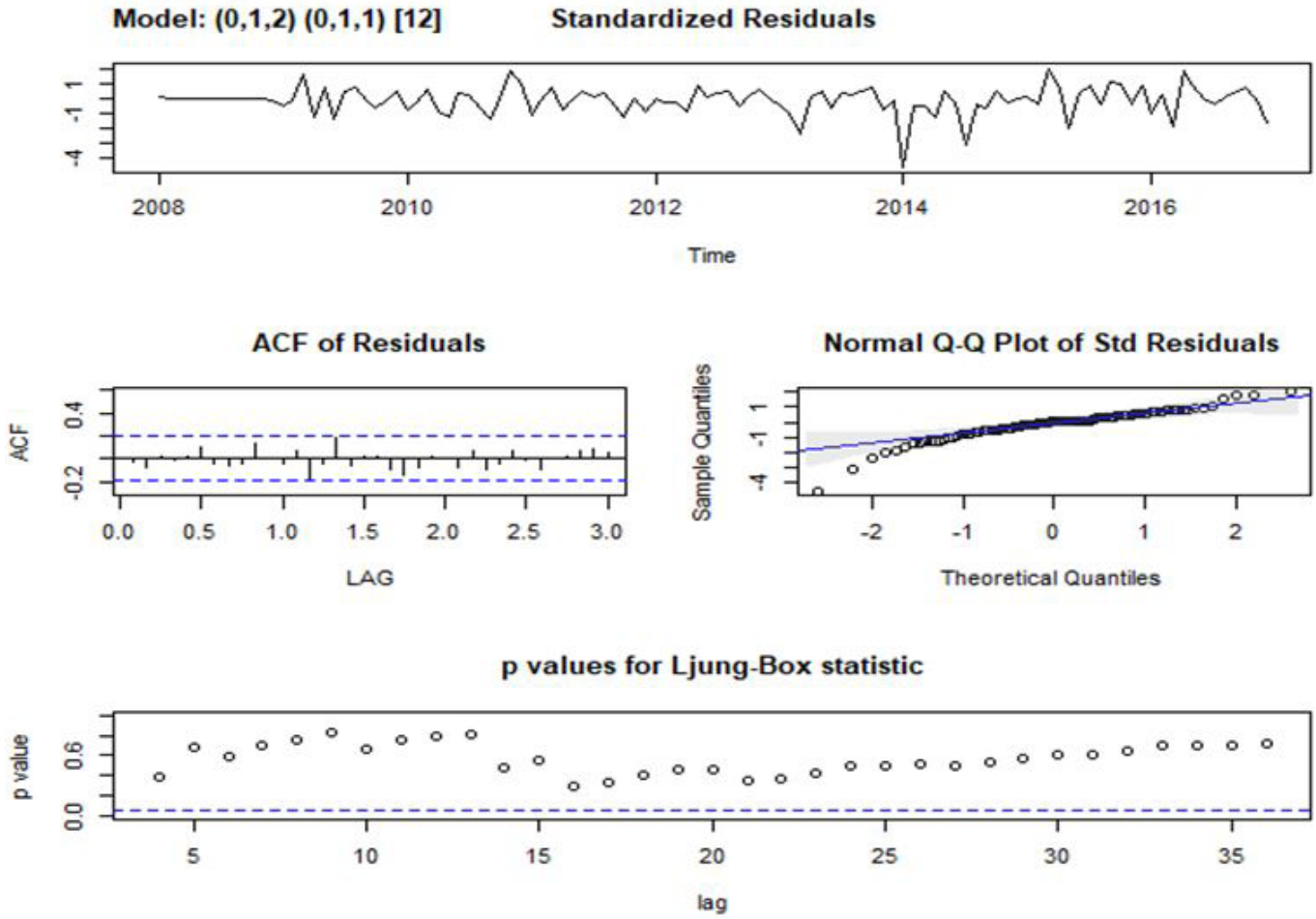

Figure 4: Diagnostic checking on the residuals of SARIMA (0,1,2)(0,1,1)[12]

Forecasting Malaria Cases Using SARIMA (0, 1, 2) (0, 1, 1)[12]

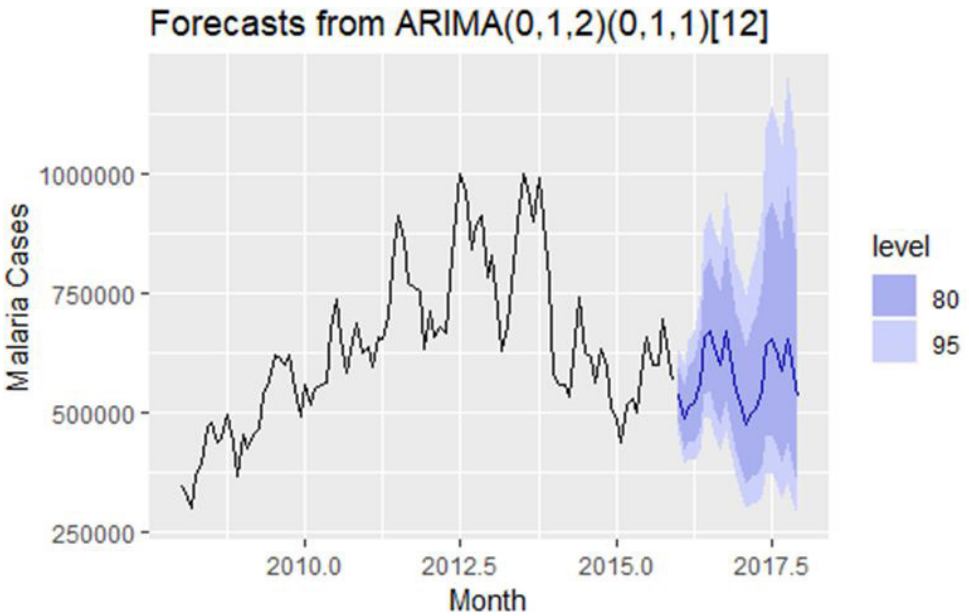

Figure 5: Forecast plot for Malaria Cases using S ARIMA (0,1,2)(0,1,1)[12]

In Figure 5, monthly forecast of Malaria cases in Ghana for two years that is 2018 to June 2019 is presented. In 2018, we should expect a decrease in Malaria cases in the last quarter but should expect an increase in Malaria cases during the first half of 2019.

Figure 6 gives the actual data from 2008 to 2017 and forecast values from 2016 to 2017 of malaria cases; it is obvious that the suitable model captures the data well especially for 2016.

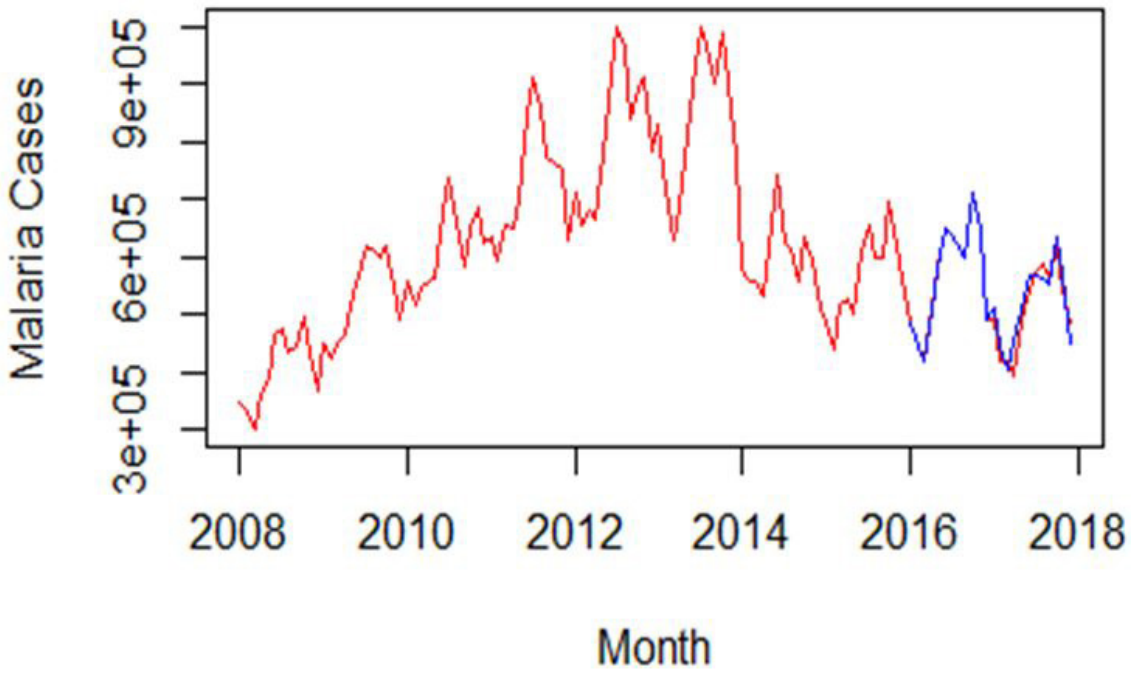

Figure 6: Actual plot (Red color) from 2008 – 2017 and Forecast plot (Blue color) from 2016 – 2017 for Malaria Cases using SARIMA (0,1,2)(0,1,1)[12]

Discussion

The main aim of the study is to identify appropriate statistical technique for forecasting the monthly Malaria cases in Ghana. Thus, strict statistical procedures were followed in order to achieve a suitable model for forecasting. Firstly, the monthly malaria cases data were assessed to know the pattern or characteristics of the data, and this is clearly shown in Figure 1 and 2.

For building suitable statistical models, four competing forecasting techniques were compared in order to choose the appropriate technique. Four competing forecasting techniques (that is SARIMA, ETS, NNAR and a combination of the three) were applied to malaria cases data spanning from January 2008 to December 2017. And for each forecasting technique, the principle of model selection was applied in order to select an appropriate model. That is, suitable model is selected under each forecasting technique and the suitable models from the four techniques are finally compared to choose the overall suitable model (Table 3 and 4). The SARIMA technique is the most suitable statistical model for forecasting malaria incidence in Ghana (Table 5). The suitable model for forecasting (i.e. SARIMA (2, 1, 0) (2, 0, 0)) passed all the needed diagnostic tests [18].

Conclusion

Forecast of malaria cases is vital for its control and intervention. This can diminish the huge impact of morbidity and mortality caused by the malady. National control and aversion methodologies will be greatly upgraded through better capacity to forecast future trends in the disease incidence. This research sought to identify appropriate statistical technique for forecasting the monthly Malaria cases in Ghana. Therefore, four competing forecasting techniques were compared in order to choose the appropriate technique. Four competing forecasting techniques (that is SARIMA, ETS, NNAR and a combination of the three) were applied to malaria cases data spanning from January 2008 to December 2017. Our findings reveal that the SARIMA technique is the appropriate statistical model for forecasting malaria incidence in Ghana. The "best" model for forecasting is SARIMA (2, 1, 0) (2, 0, 0) which passed all the needed diagnostic tests. A two year monthly forecast from the "best" model revealed that, in 2018, we should expect a decrease in Malaria cases in the last quarter but should expect an increase in Malaria cases during the first half of 2019.

References

1. Acharya P, Garg M, Kumar P, Munjal A, Raja KD (2017) Host–Parasite Interactions in Human Malaria: Clinical Implications of Basic Research. Front Microbiol 8: 889.

2. Abuaku BK, Koram KA, Binka FN (2004) Antimalarial drug use among caregivers in Ghana. Afr Health Sci 4: 171-7.

3. World Health Organization (2017) World Malaria Report 2014. World Health Org Geneva 23: 238.

4. Suh KN, Kain KC, Keystone JS (2004) Malaria. Can Med Assoc J 170: 1693-1702.

5. Ameme DK, Afari EA, Nyarko KM, Malm KL, Sackey S, et al. (2014) Direct observation of outpatient management of malaria in a rural Ghanaian district. Pan Afr Med J 19: 367.

6. Mba CJ, Aboh IK (2006) Prevalence and management of malaria in Ghana: a case study of Volta Region. Afr Popul Stud 22.

7. Ghana Health Service (2013) Ghana malaria programme review final report. Ghana Health Serv.

8. World Health Organization (1997) World malaria situation in 1994, WHO weekly Epidemiological Record. World Health Org 36: 269-74.

9. Ghana Health Service (2013) Ghana malaria programme review final report. Ghana Health Serv.

10. National Malaria Control Program (2013) National Malaria Control Program An epidemiological profile of malaria and its control in Ghana. Natl Malar Control Program.

11. Awine T, Malm K, Bart-Plange C, Silal SP (2017) Towards malaria control and elimination in Ghana: challenges and decision making tools to guide planning. Glob Health Action 10: 1381471.

12. Takyi Appiah S, Otoo H, Nabubie IB (2015) Times Series Analysis of Malaria Cases In Ejisu-Juaben Municipality. Int J Sci Technol Res 4: 220-6.

13. Alhassan EA, Isaac AM, Emmanuel A (2017) Time Series Analysis of Malaria Cases in Kasena Nankana Municipality. Int J Statis Appli 7: 43-56.

14. Bosson-Amedenu S (2017) Nonseasonal ARIMA Modeling and Forecasting of Malaria Cases in Children under Five in Edum Banso Sub-district of Ghana. Asian Res J Math 4: 1-11.

15. Anokye R, Acheampong E, Owusu I, Obeng E (2018) Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. Cogent Social Sci 4: 1461544.

16. Hyndman RJ, Athanasopoulos G (2018) Forecasting: Principles and Practices (2nd Edn).

17. Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, et al. (2018) forecast: Forecasting functions for time series and linear models. R package version 8.3.

18. Asante AF, Asenso-Okyere K (2003) Economic Burden of Malaria in Ghana. A Technical Report submitted to World Health Organization. World Health Org Africa.