# Developing Stochastic Model For Forecasting Malaria Cases In Addis Zemen, South Gondar, Ethiopia: A Time Series Analysis

Bantie Getnet, Salie Ayalew

Department of Statistics, University of Gondar, Ethiopia
Email addresses: bantiegetnet@gmail.com (Bantie Getnet), salie.ayalew55@gmail.com (Salie Ayalew)

**Abstract**: Malaria is a serious public health problem in developing countries like Ethiopia. Early prediction of malaria cases is very important for its control and intervention. This study aimed for developing stochastic model for forecasting malaria cases in Addis Zemen, South Gondar, Ethiopia. Data of monthly malaria cases from January 2007 to June 2016 were obtained from Addis Zemen health center, south Gondar, Ethiopia. The autoregressive integrated moving average (ARIMA) model, is typically applied to forecast the malaria cases; it can take into account changing trends, seasonal variation, and random disturbances in time series. Generalized Autoregressive conditional heteroscedasticity (GARCH) models are the prevalent tools used to deal with time series heteroscedasticity. In this study, based on the data of the malaria cases, the researcher establish the seasonal ARIMA $(1, 1, 1)\ (2, 1, 1)_{12}$ model and GARCH (1, 1) model, which can be used to forecast the malaria cases successfully in Addis-Zemen. Although both two families of models could reasonably forecast the malaria cases, the GARCH model demonstrated better goodness-of-fit than the SARIMA model. The seasonal trend of malaria cases is predicted to have lower monthly malaria cases in January and higher malaria cases in October. To the best of the researcher's knowledge, this is the first study to establish the ARIMA model and GARCH model for forecasting the monthly malaria cases in Addis-Zemen.

[Bantie Getnet, Salie Ayalew. **Developing Stochastic Model For Forecasting Malaria Cases In Addis Zemen, South Gondar, Ethiopia: A Time Series Analysis.** *Nat Sci* 2017;15(4):64-77]. ISSN 1545-0740 (print); ISSN 2375-7167 (online). http://www.sciencepub.net/nature. 10. doi:10.7537/marsnsj150417.10.

**Key Words:** malaria cases, time series, ARIMA models, ARCH-GARCH models

## 1. Introduction

Malaria in humans is caused by five species of parasites belonging to the genus Plasmodium. Four of these – P. falciparum, P. vivax, P. malariae and P. ovale – are human malaria species that are spread from one person to another via the bite of female mosquitoes of the genus Anopheles. There are about 400 different species of Anopheles mosquitoes, but only 30 of these are vectors of major importance. In recent years, human cases of malaria due to P. knowlesi have been recorded – this species causes malaria among monkeys in certain forested areas of South-East Asia. Current information suggests that P. knowlesi malaria is not spread from person to person, but rather occurs in people when an Anopheles mosquito infected by a monkey then bites and infects humans (zoonotic transmission). P. falciparum and P. vivax malaria pose the greatest public health challenge. P. falciparum is most prevalent on the African continent, and is responsible for most deaths from malaria. P. vivax has a wider geographical distribution than P. falciparum because it can develop in the Anopheles mosquito vector at lower temperatures, and can survive at higher altitudes and in cooler climates. It also has a dormant liver stage (known as a hypnozoite) that can activate months after an initial infection, causing a relapse of symptoms. The dormant stage enables P. vivax to survive for long periods when Anopheles mosquitoes are not present (e.g. during winter months). Although P. vivax can occur throughout Africa, the risk of infection with this species is quite low there because of the absence in many African populations of the Duffy gene, which produces a protein necessary for P. vivax to invade red blood cells. In many areas outside Africa, infections due to P. vivax are more common than those due to P. falciparum, and cause substantial morbidity.

Malaria remains a major public health problem in many countries of the world. Despite the progress in reducing malaria cases and deaths, it is estimated that 214 million cases of malaria occurred worldwide in 2015, leading to 438 000 malaria deaths (WHO, 2015).

Malaria is transmitted by mosquitoes carrying malaria parasites. Malaria's distribution depends on the availability and productivity of mosquito breeding habitat. The availability of the breeding habitat is related to stagnant water that remains after rainfall while productivity of the breeding habitat is a function of the ambient temperature (Githeko A, 2008). Rainfall rises the abundance of the breeding habitat while higher temperature increases the malaria risk by shortening the malaria parasites development-cycle (Hay et al, 2000). The average life span of a mosquito carrying malaria parasites is about 21 days. It takes 19

days for the malaria parasite to mature inside the mosquito at 22 degrees Celsius and 8 days to mature at 30 degrees Celsius. Apart from the African highlands and the farthest southern and northern African regions, the annual mean temperature on the African continent is above 25 degrees Celsius (Githeko A, 2008). Therefore, the increase in mean temperature under climate changes (IPCC. Climate Change 2007) may result in a faster parasite development and a potentially higher incidence of malaria.

In Ethiopia, malaria is one of the most public health problems, with more than three-quarters of the landmass (altitude <2000 m) of the country is either malarious or potentially.

malarious, and an estimated 68% (>50 million people) of the total population resides in areas at risk of malaria infections (Adhanom et al, 2006). Annually, half a million microscopically confirmed cases of malaria are reported to the Federal Ministry of Health (FMOH) from basic health services. However, the actual number of malaria cases in the country is estimated to be more than 5 million each year. According to the 2007/2008 report of the FMOH, malaria was the leading cause of outpatient visit accounting for 12% of cases and the second cause of (10%) admit next only to admit for delivery (MOH, 2007/2008).

P. falciparum and P. vivax are the dominant malaria parasites distributed all over Ethiopia and account for about 60% and 40% of malaria cases, respectively(MOH, 2007/2008).

Based on the findings or reports from Amhara regional health bureau report, 2011/2012 Amhara region is one of the malarious regions of Ethiopia. The prevalence among sex was male greater than women and the dominant species of malaria in the region are P.falciparum and P.vivax. Due to these problems the researcher motivated to investigate the malaria cases in Addis Zemen, South Gondar, Ethiopia using the stochastic time series models for forecasting these malaria cases, and to give indications for what factors should be done more activities to solve the problem.

**1.2. Statement Of The Problem**

In Ethiopia, malaria is one of the most public health problems, with more than three-quarters of the landmass of the country and an estimated 68% of the total population is considered at risk of malaria infections (Adhanom et al, 2006). Ethiopia is implementing a range of malaria control interventions that aim to improving access and equity to preventive as well as curative health services, which include prompt and effective malaria treatment, selective vector control using insecticide treated nets and indoor residual spraying. Effective and timely prevention and control of malaria epidemics is also part of the main strategies. The need for developing comprehensive and high impact communication strategies for malaria control is imperative (ACIPH, 2009).

The occurrence of malaria epidemics has been more frequent and wide-spread in recent years. Although rainfall-associated breeding of the major vector Anopheles arabiensis is the main cause of seasonal malaria epidemics in Ethiopia, abnormal climatic changes have often given rise to major epidemics in the past. These epidemics have usually inflicted high incidence of mortality upon the non-immune population. Most of the epidemic-affected areas are highlands or highland fringe areas where the population lacked immunity to malaria and thus all age groups are frequently affected. The somewhat large-scale periodic epidemics have been associated with increase in temperature, abnormally high rainfall as well as unusually prolonged dry seasons or in other words malaria cases are depending on the environmental, seasonal, climatic and others different socioeconomic factors. Reducing malaria incidence at any level will need identifications of seasonal, environmental and climatic variation of malaria incidence. There is at present a need for a strengthened epidemic management at all levels due to increasing problem in early detection, prevention and control. Therefore, to prevent and control the malaria diseases for the future it is better to forecast the coming malaria cases, and this study will address the future malaria cases by forecasting and identify the peaks on and off period using a stochastic model.

**1.3. Objective Of The Study**

**1.3.1. General Objective Of The Study**

The main objective of this study is developing stochastic model for forecasting malaria cases in Addis Zemen, South Gondar, Ethiopia.

**1.3.2. Specific Objective Of The Study**

- ✓ Compare the models to find the best fit model using time series model selection techniques and forecasting future malaria cases.
- ✓ To identify the months in which the malaria cases mostly occur.
- ✓ To compare the forecasting power of the models from the malaria cases data.
- ✓ To see the influence of past malaria cases to the present time.

**1.4. Significance Of The Study**

This study will have the following significances:

- ➢ The study will provide information for the Government/concerned bodies about the malaria cases.
- ➢ The study will provide background information to those who want to conduct further detailed studies in development of stochastic time series models.

- The findings will also help for peoples who are living in Adiss Zemen by providing information about malaria.

## 2. Data And Methodology

A retrospective study was conducted at Addis Zemen health center from January 2007 to June 2016 for malaria cases. Addis Zemen is found in South Gondar administration zone in the Amhara region of northwestern Ethiopia and is around 637 km far from the capital city of Ethiopia. Addis Zemen which has a total population of 20,412 is the capital town of Libo Kemkem wereda (district), which has average populations of 198,374. It has an average altitude of less than 2,000 m above sea level Addis Zemen has a latitude and longitude of 12°07′N 37°47′E. The health center serves not only Libo Kemkem district but also the nearby districts like Fogera which has estimated populations of 226,595. This district is malarious and the majority of the population depends on subsistence farming. Malaria is the most prevalent seasonal disease in the area, accounted as second of all the reported diseases in the health center and October to December is the peak malaria transmission season in the area. Both P.vivax and P.falciparum exist in the area with P.falciparum prevailing all year.

A time series is a series or sequence of data points measured typically at successive times. These data points are commonly equally spaced in time (Chatfield, 2004).

The first task in any time series analysis is to check the data is a time series data or not, to do this task there are many methods to check the data are random or not. Among these tests, turning point test, phase length test, rank test and difference sign test are the most common tests. Here the researcher applies only the turning point test and this test is described as the test against systematic oscillation.

After checking the nature of the data i.e. whether the data is time series or not the next step is checking the stationarity of the series. In the event that the series exhibits nonstationarity, appropriate transformations, will be applied to make the series stationary. Alternatively, other approaches such as smoothing, in the form of for example exponential smoothing may also be used to transform the data.

### 2.1. Time Series Models

Family of ARIMA and family of ARCH-GARCH models were used in this paper.

After describing various time series models, the next step is model identification here, once the model is tentatively established, the parameters and the corresponding standard errors can be estimated using statistical techniques, such as Maximum Likelihood (ML), least square and Yule-Walker estimation method, the other step is model checking which, includes the analysis of the residuals as well as model comparisons. If the model fits well, the standardized residuals should behave as an independent and identically distributed sequence with mean zero and variance one (Cryer and Chan, 2008). A standardized residuals plot or a Q-Q plot can help in identifying the normality (Stoffer and Dhumway, 2010) and the formal test Box-Pierce-Ljung test were used to check the model and finally forecasting were applied using different forecasting techniques like exponential smoothing(simple exponential, double exponential and triple exponential).

## 3. Results

This chapter presents results and discussions of developing stochastic model for forecasting malaria cases using the data which is obtained from Addis-Zemen health center. Data management and analysis was done in R-Gui Software (version 3.3.1) and R-Studio.

As explained on section 3.2.2.1., before analyzing the data the first task is to check whether the data is a time series or not. Thus Table4.1.1 below shows that the series is not random which rejects the null hypothesis that the series is random; this indicates that the data is a time series data, implies that it is possible to apply time series analysis for these malaria cases data.

Table 4.1.1 Turning point test for randomness of malaria cases data

| Turning point test | | |
|---|---|---|
| Statistic | Number of observations | p-value |
| -3.508 | 114 | 0.0004514 |

### 3.1. Descriptive Data Analysis

The monthly reported malaria cases that were used for this research are covering the period from January 2007 to June 2016, which consists of 114 monthly malaria cases registered in Addis Zemen health center, most of the malaria cases were clinical diagnosed and most of the cases occur on males above age group 15 and plasmodium falciparem is the most dominant as compared to plasmodium vivax and mix.

There was a relatively upward and downward trend (figure 4.1) within the given period and shows roughly seasonal fluctuations with the highest peak observed in October 2011 and the lowest peak observed in February 2016.

This figure 4.1 shows somewhat upward and downward trend even if there is no clear showing and clearly there is seasonal variation. In order to observe clearly it is better to decompose into three different parts i.e. into trend, seasonal and random part as shown in figure 4.2.

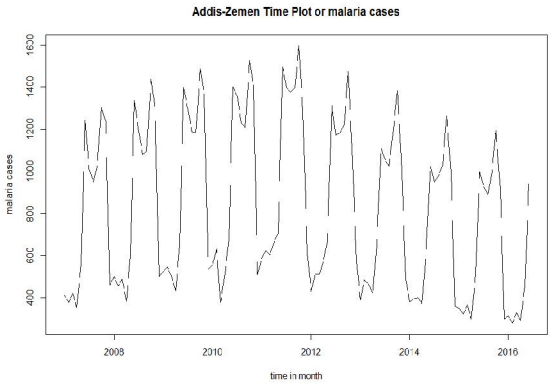



Figure 4.1 original malaria cases data plot.

The plot (figure 4.2) shows the original time series (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated random component (bottom). Here clearly observe that the estimated trend component shows an increasing trend from about 2007 to about 2011, followed by a steady decrease from 2012 to 2016.

### 3.2. Exploratory Data Analysis

This section is focused to fitting the ARIMA and GARCH family of models to Addis Zemen health center malaria cases data. The original data set consist of 114 monthly malaria cases and spanning from January 2007 to June 2016.

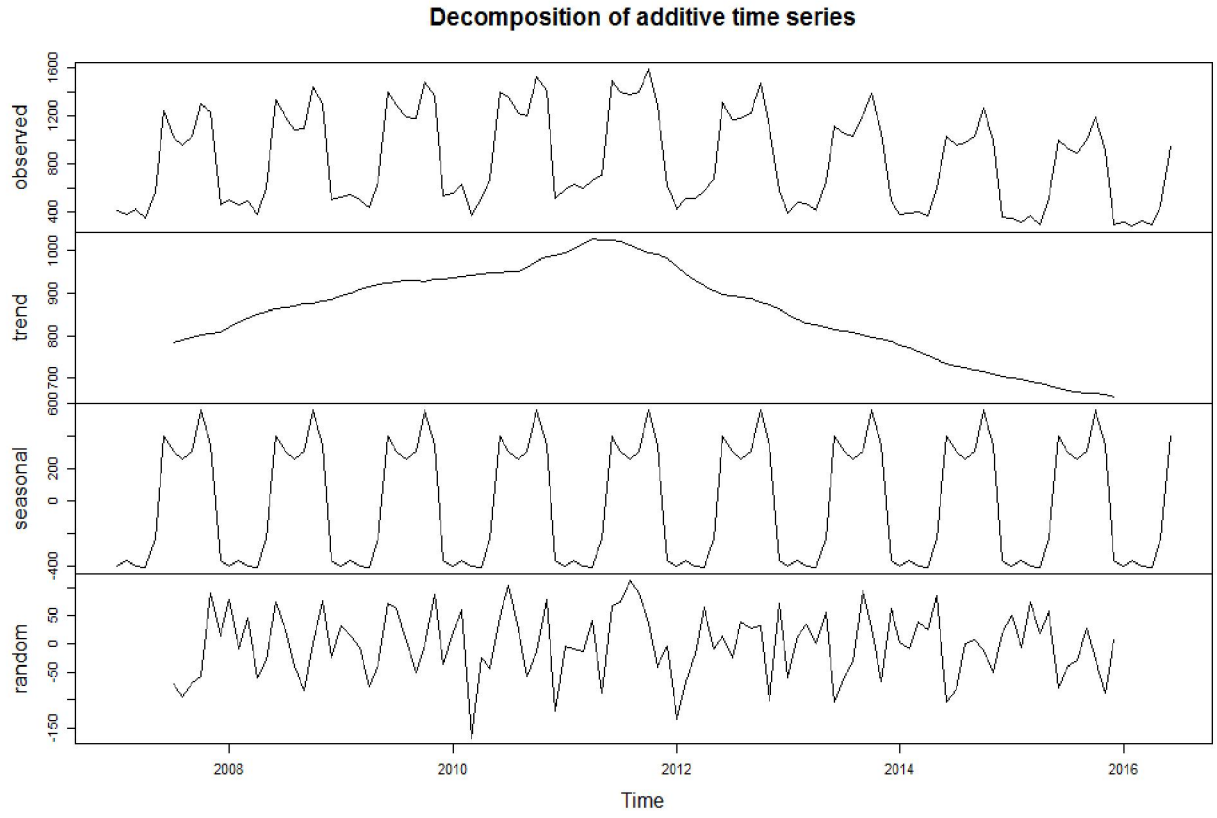



Figure 4.2: decomposition of time series data into trend, seasonal and random.

### 3.3. ARIMA Models

ARIMA model needs stationarity; therefore, the first step here is conduct stationarity tests to check if data is stationary, otherwise, it is difficult to forecast or predict the malaria cases. Here the ADF test ($p>0.05$) show the original time series is not stationary. In order to obtain a stationary time series, the researcher uses three steps to achieve. Firstly, first-order non-seasonal difference ($d = 1$) is computed, after that, ACF and PACF graphs indicate a high seasonal behavior with a circle of 12 (so $s = 12$), secondly, to remove monthly seasonality, first-order seasonal difference ($D = 1$) with a circle of 12 is computed, finally, to do ADF test, the result (as shown in Tbale4.2.1) is statistically significant ($p<0.01$), which confirms that the transformed time series is stationary.

Table 4.2.1: Augmented Dickey Fuller Test for stationarity

| Variables | Augmented Dickey Fuller Test | p-value |
|---|---|---|
| Original malaria cases | -3.9473 | 0.1429 |
| First differenced malaria cases | -15.353 | 0.01 |

### 3.3.1. Model Selection

After the series has been checked it's stationarity, the next step is fitting appropriate model. To fit the appropriate model as the researcher discussed on (3.5.2.) model selection is the most important technique that used to employ computationally simple techniques to narrow down the range of parsimonious models and this will lead to selecting the appropriate model that adequately describes the data.

First, the researcher should construct a time plot of the data and inspect the graph for any anomalies and to determine how many AR or MA terms are needed to correct any autocorrelation that remains in the differenced series (as shown figure 4.3.1). Thus, the numbers of AR and/or MA terms that are needed to fit a model are tentatively identified by looking the ACF and PACF plots of the series.

If the PACF of the series shows a cut off at lag k, it means that the series is not enough differenced and then by adding enough autoregressive terms can remove any autocorrelation left from a stationarized series. The lag at which the PACF cut off tells us how many AR terms are needed. For the same case for PACF, if the ACF cut off at lag k, this indicates that exactly k MA terms that are needed to remove the remaining autocorrelation from the series. Hence, by visual inspection of the Figure 4.3. below, the number of significant correlation lags from the ACF plot are almost 3 and the number of significant correlation lags from the PACF plot looks to be 2. Thus, the first tentatively model was ARIMA (2, 1, 3) where p=2 & q=3 are the order of autoregressive and moving average model respectively, and d=1 is the order of integration (non seasonal differencing) and for the seasonal effect the first tentative model was seasonal ARIMA or SARIMA $(2,1,3)(2,1,2)^{12}$ where P=2 and Q =2 are the order of seasonal autoregressive and moving average terms, and D=1 the order of seasonal differencing.

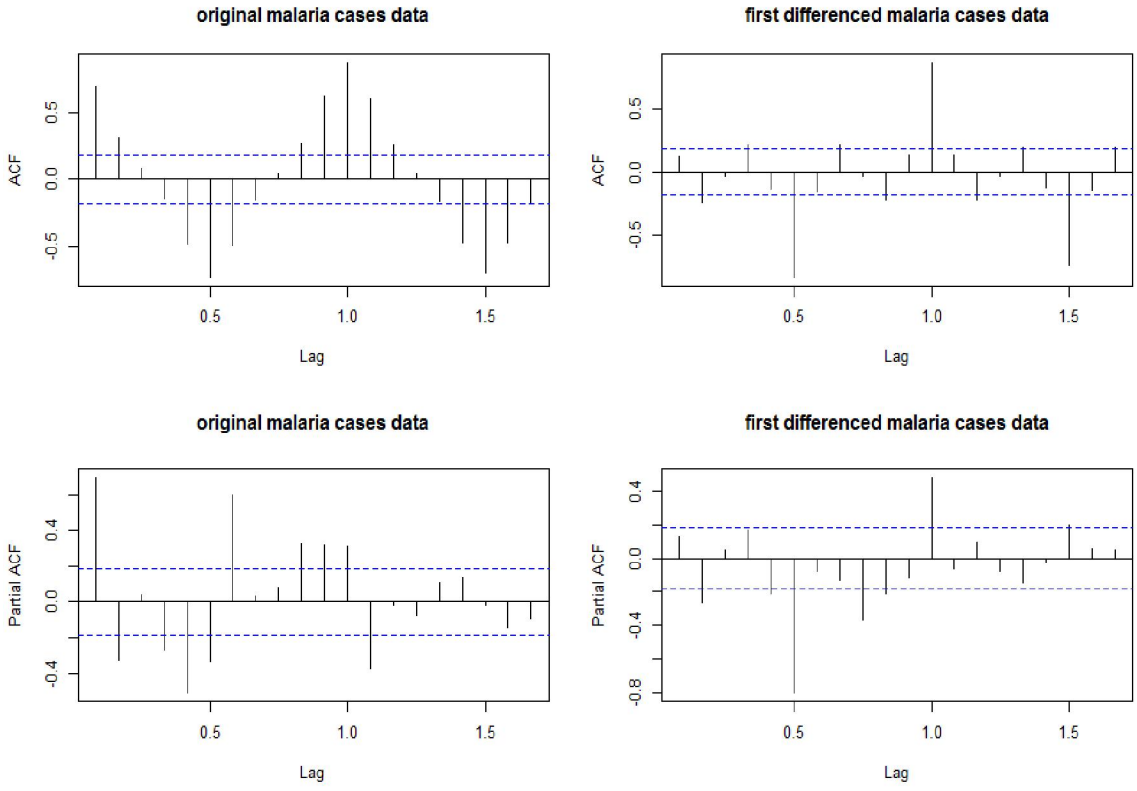



Figure 4.3. Addis-Zemen malaria cases data ACF and PACF plot

Based on the distribution characteristics, the researcher conducted seventeen possible models, SARIMA(1,1,1)(1,1,1), SARIMA(1,1,2)(1,1,1), SARIMA(1,1,2)(2,1,1), SARIMA(1,1,2)(2,1,2),

SARIMA(1, 1, 1)(2,1,2), SARIMA(1,1,1)(2,1,1), SARIMA(1,1,1)(1,1,2), SARIMA(2,1,1)(1,1,2), SARIMA(2,1,1)(1,1,1), SARIMA(2,1,1)(2,1,1), SARIMA(2,1,1)(2,1,2), SARIMA(1,1,3)(1,1,1), SARIMA(1,1,3)(2,1,2), SARIMA(2,1,2)(2,1,2), SARIMA(2,1,3)(2,1,1), SARIMA(2,1,3)(1,1,2) and SARIMA(2,1,3)(2,1,2) as shown in Table 4.3.1. Of all the models tested, the chosen model is normally the one with the least value of, AIC and satisfying also the parsimony principle which favors the least parameter possible in the model. Thus, SARIMA (1, 1, 1)$(2,1,1)_{12}$ was the one to satisfy the previous conditions with least value of AIC.

Table4.3.1. Comparison of tested ARIMA models

| Model | AIC |
|---|---|
| SARIMA(1, 1,1)(1,1,1) | 1114.58 |
| SARIMA(1,1,2)(1,1,1) | 1118.53 |
| SARIMA(1,1,2)(2,1,1) | 1114.08 |
| SARIMA(1, 1,2)(2,1,2) | 1122.20 |
| SARIMA(1,1,1)(2,1,2) | 1113.99 |
| **Sarima(1,1,1)(2,1,1)** | **1112.03** |
| SARIMA(1,1,1)(1,1,2) | 1112.76 |
| SARIMA(2,1,1)(1,1,2) | 1135.37 |
| SARIMA(2,1,1)(1,1,1) | 1114.65 |
| SARIMA(2,1,1)(2,1,1) | 1112.28 |
| SARIMA(2,1,1)(2,1,2) | 1140.58 |
| SARIMA(1,1,3)(1,1,1) | 1116.25 |
| SARIMA(1,1,3)(2,1,2) | 1120.19 |
| SARIMA(2,1,2)(2,1,2) | 1120.42 |
| SARIMA(2,1,3)(2,1,1) | 1124.42 |
| SARIMA(2,1,3)(1,1,2) | 1118.78 |
| SARIMA(2,1,3)(2,1,2) | 1128.54 |

From this the general model can be written as: p=1, q=1, d=1 and P=2, D=1, Q=1, s=12

$$(1-\Phi_1 B)(1-B_1B^{12}-B_2B^{24})(1-B)(1-B)X_t = C + (1-\Psi_1 B)(1-\theta_1 B^{12})\varepsilon_t \quad (4.1)$$

### 3.3.2. Model Estimation

This is the process of estimating the model parameters after selecting an appropriate model. The parameter estimates should be significant, with each providing a substantial contribution to the model for the most accurate forecasts. As the researcher discussed on 3.5.3, there are a number of ways to estimate autoregressive and moving averages parameters in ARMA models such as Maximum Likelihood and Least square estimates.

From the derived models, using the method of maximum likelihood the estimated parameters of each model is summarized in the following tables:

Table 4.3.2 Maximum Likelihood Estimates for parameters

| Coefficients | Estimates | p-value |
|---|---|---|
| ar1 | -0.4941 | <0.01 |
| ma1 | -0.9999 | <0.01 |
| sar1 | 0.9410 | <0.01 |
| sar2 | -0.1941 | <0.01 |
| sma1 | -0.9780 | <0.01 |
| Intercept | 772.32 | <0.01 |

As described on the above equation 4.1 the corresponding parameter values including the intercept term are C = 772.32, $\Phi_1$= -0.4941, $B_1$ = 0.9410, $B_2$ = -0.1941, $\Psi_1$ =-0.9999, $\theta_1$ = -0.978, and the final fitted model was

$$(1+0.49B)(1-0.94B^{12}+0.19B^{24})(1-B)(1-B)X_t = 772.32+(1+0.9999B)(1+0.98B^{12})\varepsilon_t \quad (4.2)$$

### 3.3.3. Diagnostic Checking Of The Seasonal ARIMA (1, 1, 1)$(2, 1, 1)^{12}$ Model

Model diagnostics is concerned with testing the goodness of fit of a model and suggesting appropriate recommendations if found to be poor and the goal of any statistical model development is to obtain which best describes the best model of the data; which means, having identified the final preliminary model the next step and most important in statistical analysis is to diagnose the fit of the model. The researcher then conduct the residual analysis by observing the ACF, PACF and conducting the Box-Pierce-Ljung test to goodness of fit to check if the residuals conform to the normal distribution.

The residual plot, ACF and PACF do not have any significant lag, indicating SARIMA$(1,1,1)(2,1,1)_{12}$ is a good model to represent the series.

In addition, Box-Pierce test also provides a different way to double check the model.

Basically, Box-Pierce is a test of autocorrelation in which it verifies whether the autocorrelations of a time series are different from 0. In other words, if the result do not rejects the hypothesis, this means the data is not independent and uncorrelated; which indicates, there still remains serial correlation in the series and the model needs modification.

Table 4.3.3. Box-Pierce Test for SARIMA (1, 1, 1)$(2,1,1)_{12}$ model

| Box-Pierce Test | | |
|---|---|---|
| $X^2$ statistic | Degree of freedom | p-value |
| 19.332 | 20 | 0.5003 |

Here, the output showed that p-value greater than 0.05, so the researcher cannot reject the hypothesis that the autocorrelation is 0 or the model fits the data well. Therefore, the selected model is an appropriate one for forecasting of malaria cases.

Generally, the time plot of the residuals look randomly distributed around zero and exhibits no clear pattern. The pattern of residuals together with the Box-Pierce statistics give overwhelming evidence that the residuals are independent implying the model fits the data well.

After identified and estimated a model that fits the data, the next step is to use the model to forecast future values of the series, which ideally is the principal goal of time series and the objective of this paper.

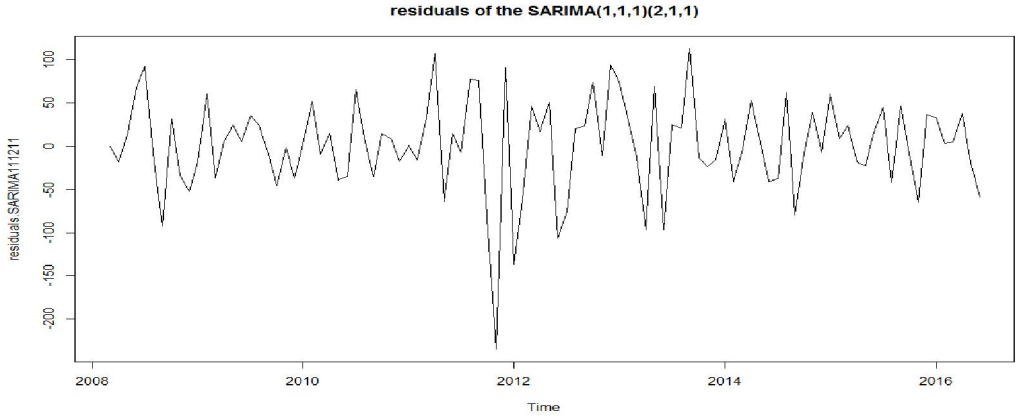
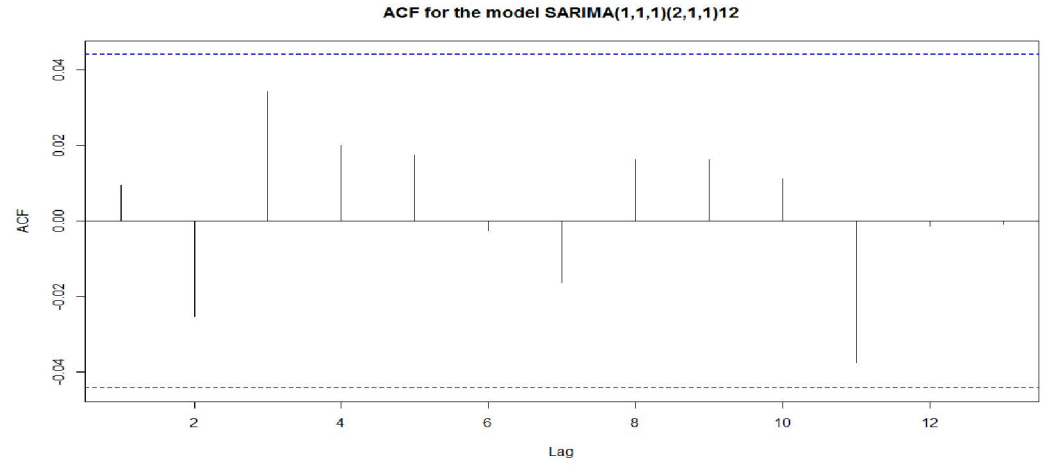
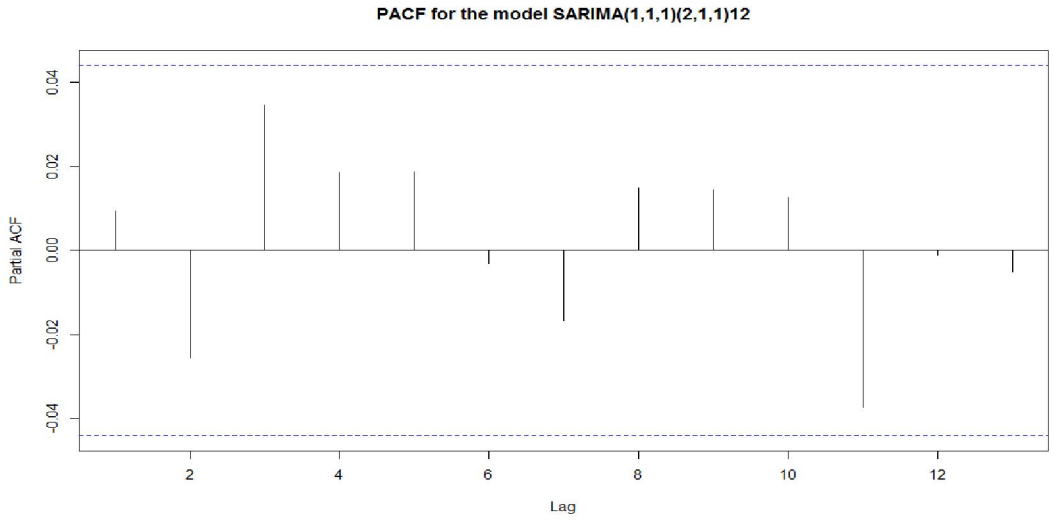



Figure 4.4. plot of residual, ACF and PACF for SARIMA $(1,1,1)(2,1,1)_{12}$ model



### 3.3.4. Forecasting With SARIMA (1,1,1)(2,1,1)12 Model

The main objective of this paper is to develop a stochastic model for forecasting future malaria cases before they are realized, and this section focuses on it. The forecasting techniques discussed in section (3.5.5.1) are employed as procedures for obtaining the forecast. The triple exponential (Holt-Winters) forecasting procedure that was discussed in section (3.5.5.1.3) is used since it has the capacity to cope with both trend and seasonality.

SARIMA $(1, 1, 1)(2, 1, 1)_{12}$ model given above was used to generate the forecast for given in the table below (4.3.4) for the years 2016(from July to December), 2017 and 2018.

The lower and upper 95% confidence limits are used to assess how good the forecasts are. This implies that the forecasts are expected to lie within the confidence limits with 95% confidence. As expected, the further into the future a forecast is the less precise it is, hence the wider the confidence limits indicating that the model has low forecasting power although it fits the data well.

### 3.4. Arch-Garch Modeling

Although PACF and ACF of residuals has no significant lags, and the time series plot of residuals shows some cluster of volatility. It is important to remember that ARIMA is a method to linearly model the data and the forecast width remains constant because the model does not reflect recent changes or incorporate new information. In other words, it provides best linear forecast for the series, and thus plays little role in forecasting model nonlinearly. In order to model volatility, ARCH/GARCH method comes into play. Now consider applying the ARCH-GARCH modeling to malaria cases data. But before applying the ARCH-GARCH modeling a formal test for heteroscedasticity was carried out in order to establish the presence of ARCH effect in the data. The lagrange multiplier and the Ljung Box Q-test (given in section 3.5) were used to check the validity of the ARCH effects in the data.

Table 4.3.4 thirty month's forecast of malaria cases obtained from the SARIMA $(1, 1, 1)(2, 1, 1)_{12}$ model.

| Time | point Forecast | Lower 95% confidence limit | Higher 95% confidence limit | Interval |
|---|---|---|---|---|
| Jul-2016 | 860 | 744.12 | 996.11 | 251.99 |
| Aug-2016 | 824 | 707.9 | 963.21 | 255.31 |
| Sep-2016 | 931 | 811.55 | 1072.61 | 261.06 |
| Oct-2016 | 1139 | 1015.66 | 1282.98 | 267.32 |
| Nov-2016 | 864 | 738.62 | 1012.69 | 274.07 |
| Dec-2016 | 258 | 129.12 | 410.46 | 281.34 |
| Jan-2017 | 270 | 126.52 | 425.63 | 299.11 |
| Feb-2017 | 238 | 89.64 | 396.99 | 307.35 |
| Mar-2017 | 281 | 119.23 | 445.33 | 326.1 |
| Apr-2017 | 238 | 71.21 | 406.52 | 335.31 |
| May-2017 | 400 | 229.22 | 574.21 | 344.99 |
| Jun-2017 | 888 | 710.24 | 1065.35 | 355.11 |
| Jul-2017 | 804 | 571.45 | 1041.02 | 469.57 |
| Aug-2017 | 769 | 531.49 | 1009.67 | 478.18 |
| Sep-2017 | 877 | 633.49 | 1120.72 | 487.23 |
| Oct-2017 | 1082 | 835.99 | 1332.69 | 496.7 |
| Nov-2017 | 809 | 557.38 | 1063.96 | 506.58 |
| Dec-2017 | 205 | -53.63 | 463.25 | 516.88 |
| Jan-2018 | 213 | -47.69 | 479.89 | 527.58 |
| Feb-2018 | 183 | -86.01 | 452.68 | 538.69 |
| Mar-2018 | 225 | -47.79 | 502.39 | 550.18 |
| Apr-2018 | 184 | -97.15 | 464.91 | 562.06 |
| May-2018 | 344 | 59.57 | 633.9 | 574.33 |
| Jun-2018 | 831 | 539.34 | 1126.29 | 586.95 |
| Jul-2018 | 749 | 411.05 | 1091.46 | 680.41 |
| Aug-2018 | 716 | 369.49 | 1061.7 | 692.21 |
| Sep-2018 | 821 | 469.94 | 1174.31 | 704.37 |
| Oct-2018 | 1027 | 670.91 | 1387.81 | 716.9 |
| Nov-2018 | 754 | 390.81 | 1120.58 | 729.77 |
| Dec-2018 | 149 | -221.67 | 521.32 | 742.99 |

Table 4.4.1. Lagrange Multiplier test for ARCH effect

| ARCH LM-test | | |
|---|---|---|
| $X^2$ statistic | Degree of freedom | p-value |
| 31.828 | 12 | 0.00171 |

The null hypothesis of homoscedasticity, the opposite of heteroscedasticity was tested and Table (4.4.1) gives the results for the Lagrange Multiplier (LM) test for heteroscedasticity. The p-value shows that an evidence for the presence of heteroscedasticity in the data, hence ARCH-GARCH modeling was deemed appropriate. According to Engle, (1982) any autocorrelations in the series have to be removed before an ARCH-GARCH model is constructed. This was done by regressing the squares of the series $X_t$ on its past squared values$X_t^2$, $X_{t-1}^2$,... with the number of lags determined by the form of the ACF and the PACF. The ACF and PACF suggested an AR (2) and MA (1) process respectively (as shown in figure 4.5 and 4.6 below), thus an AR (2) and MA (1) models were used in all autocorrelations being removed. Hence consider fitting the ARCH-GARCH models to the data.

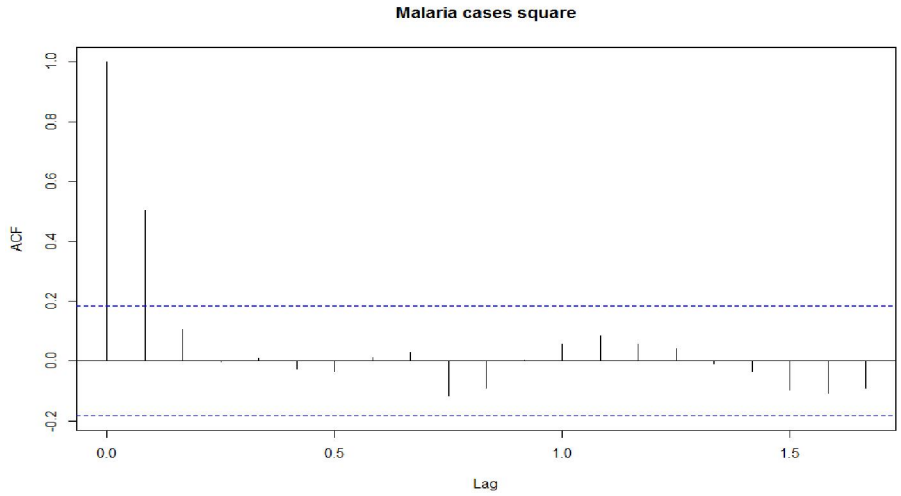



Figure 4.5. ACF squared plots for Malaria cases data

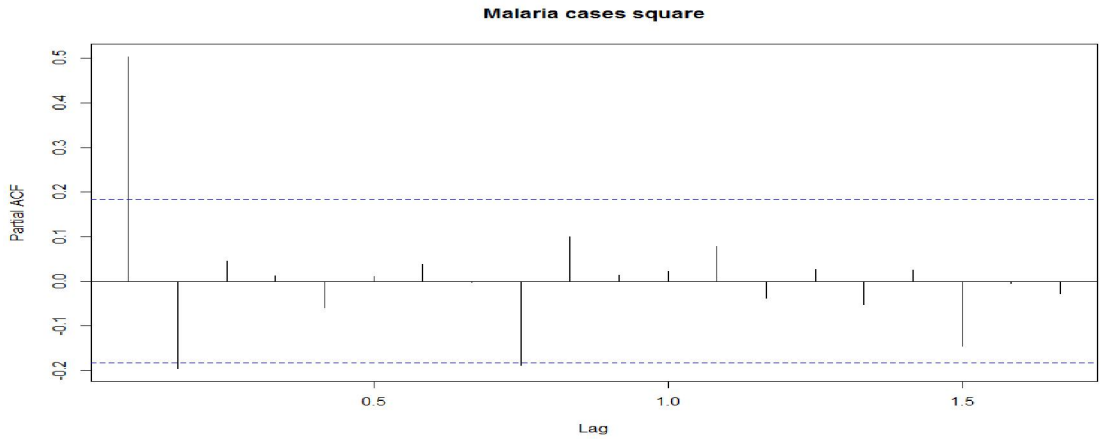



Figure 4.6. PACF squared plots for Malaria cases data

### 3.4.1. ARCH-GARCH Model Selection

Like model selection techniques in ARIMA modeling AIC and BIC are used and additionally $R^2$ and MSE were used to perform to determine the best ARCH-GARCH models.

As the researcher discussed on section (3.6.2.1) the GARCH model with AR errors is given by

$$X_t = \sigma_t \varepsilon_t \quad (4.3)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \quad (4.4)$$

The order of the parameters are determined by studying the ACF and the PACF in the same way as was done in the ARIMA modeling. Figure (4.5 above) shows the order of parameters and Table 4.4.2 gives the suggested models with their respective fit statistics.

Therefore in table (4.4.2), the smaller the AIC, BIC and the SIC the better the model that is GARCH (1,1) was judged to be the most appropriate according to the criteria above.

Table 4.4.2 comparison of tested GARCH models

| Model | AIC | BIC | SIC |
|---|---|---|---|
| GARCH(1,0) | 1.225 | 1.234 | 1.225 |
| **Garch(1,1)** | **1.125** | **1.137** | **1.125** |
| GARCH(2,0) | 1.189 | 1.200 | 1.189 |
| GARCH(2,1) | 1.127 | 1.141 | 1.127 |

**3.4.2. Estimating Parameters Of The GARCH(1,1)**

From the derived models, using the method of maximum likelihood the estimated parameters of GARCH (1,1) model is summarized in the Table4.4.3:

Table 4.4.3 maximum likelihood estimates of GARCH (1,1) model

| Coefficients | Estimates | Standard Errors |
|---|---|---|
| $\alpha_0$ | 0.0108 | 0.0028 |
| $\alpha_1$ | 0.1531 | 0.0264 |
| $\beta_1$ | 0.8060 | 0.0334 |

Table 4.4.3 suggests that the final model can be written as:

$X_t = \sigma_t \varepsilon_t$ (4.5)

and

$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2$ (4.6)

$\sigma_t^2 = 0.0108 + 0.1531 X_{t-1}^2 + 0.8060 \sigma_{t-1}^2$ (4.7)

Having estimated our parameters, the next step is to check how well the model fits the data and this can be explained in section 4.5.3 below.

**3.4.3. Diagnostic Checking Of The GARCH (1,1) Model**

One of the assumptions of GARCH models is that, for a good model, the residuals must follow a white noise process. If the model fits the data well, the residuals are expected to be random, independent and identically distributed following the normal distribution. The time plot of the residuals given in figure (4.7) is used to check whether the residuals are random.

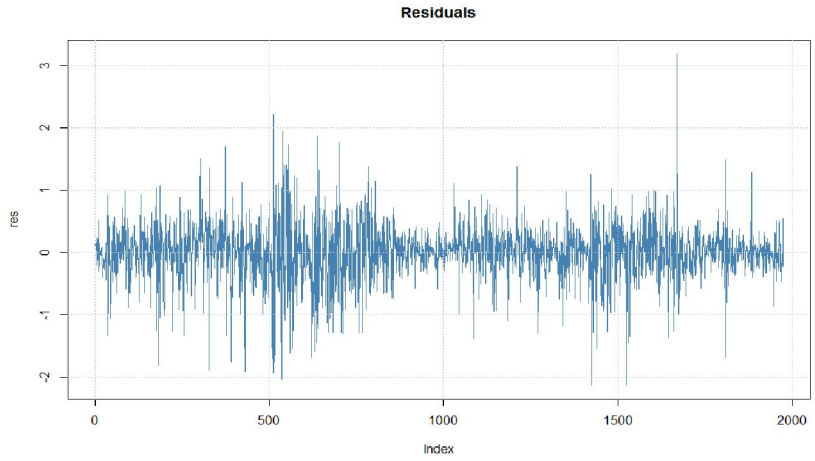



Figure 4.7 plots of residuals from GARCH (1,1) model

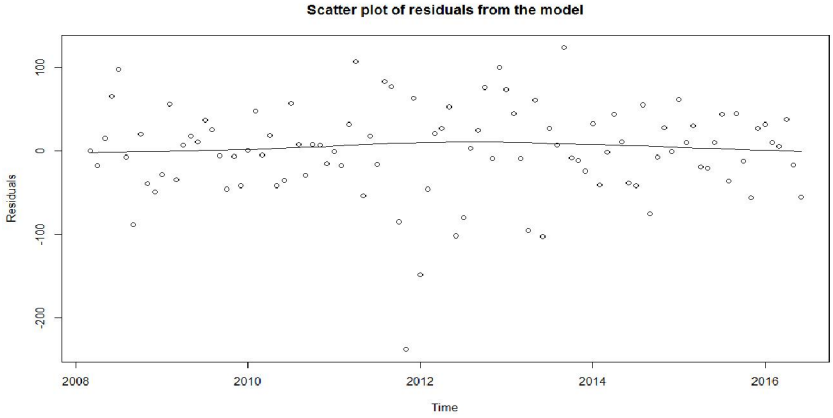



Figure 4.8. Scatter plot of the residuals

The normality check is also done by analyzing the histogram of residuals and normal probability plot. Figure (4.8) gives the histogram of the residuals from the GARCH (1,1) model.

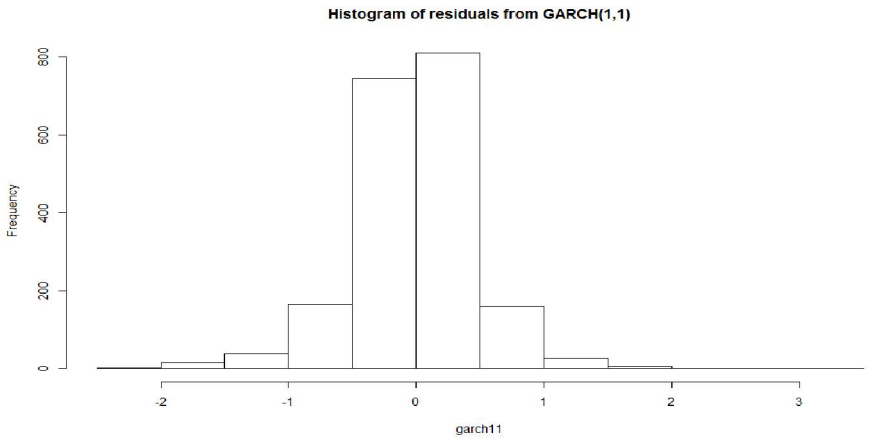



Figure 4.9. Histogram of the residuals from GARCH (1,1)

The histogram shows almost a symmetric bell shaped distribution which is indicative of the residuals following a normal distribution.

### 3.4.4. Forecasting With GARCH (1,1) Model

After successfully identifying and checking a potential model that describes well the historical data of malaria cases and a model that fits the data well is bound to give good forecasts. Using this selected model GARCH (1,1), forecasting was made for the next 30 months (July 2016–December 2018). The results are shown in table4.4.4 below.

Table 4.4.4 thirty month's forecast of malaria cases obtained from the GARCH (1,1) model.

| Time | Point Forecast | Lower 95% confidence limit | Higher 95% confidence limit | Interval |
|---|---|---|---|---|
| Jul-2016 | 861 | 705.07 | 986.1 | 281.03 |
| Aug-2016 | 825 | 666.74 | 952.45 | 285.71 |
| Sep-2016 | 931 | 770.47 | 1061.34 | 290.87 |
| Oct-2016 | 1138 | 974.53 | 1271.01 | 296.48 |
| Nov-2016 | 864 | 700.52 | 1000.09 | 299.57 |
| Dec-2016 | 258 | 95.22 | 397.35 | 302.13 |
| Jan-2017 | 270 | 106.1 | 412.25 | 306.15 |
| Feb-2017 | 237 | 73.71 | 383.33 | 309.62 |
| Mar-2017 | 281 | 117.88 | 431.45 | 313.57 |
| Apr-2017 | 238 | 74.56 | 392.49 | 317.93 |
| May-2017 | 401 | 241.01 | 559.75 | 318.74 |
| Jun-2017 | 887 | 722.530 | 1050.5 | 327.97 |
| Jul-2017 | 806 | 587.76 | 1022.44 | 434.68 |
| Aug-2017 | 770 | 547.85 | 990.38 | 442.53 |
| Sep-2017 | 876 | 650.03 | 1100.81 | 450.78 |
| Oct-2017 | 1083 | 852.58 | 1312 | 459.42 |
| Nov-2017 | 809 | 574.09 | 1042.55 | 468.46 |
| Dec-2017 | 203 | -36.64 | 441.23 | 477.87 |
| Jan-2018 | 214 | -30.15 | 457.52 | 487.67 |
| Feb-2018 | 182 | -67.88 | 429.95 | 497.83 |
| Mar-2018 | 226 | -29 | 479.37 | 508.37 |
| Apr-2018 | 183 | -77.59 | 441.67 | 519.26 |
| May-2018 | 345 | 79.64 | 610.15 | 530.51 |
| Jun-2018 | 832 | 559.98 | 1102.08 | 542.1 |
| Jul-2018 | 751 | 434.98 | 1064.25 | 629.27 |
| Aug-2018 | 715 | 393.58 | 1033.68 | 640.1 |
| Sep-2018 | 820 | 494.31 | 1145.57 | 651.26 |
| Oct-2018 | 1029 | 695.42 | 1358.19 | 662.77 |
| Nov-2018 | 755 | 415.54 | 1090.13 | 674.59 |
| Dec-2018 | 149 | -196.56 | 490.19 | 686.75 |

**4. Discussions**

Malaria is one of the most common infectious diseases in the world and one of the greatest global public health problems. In Ethiopia, it is one of the most important public health problems, with more than three-quarters of the landmass of the country and an estimated 68% of the total population is considered at risk of malaria infections (Adhanom et al. 2006). Based on the results from figure4.1 the number of malaria cases has an upward and downward trend, which indicates there is a challenge for malaria diseases control and prevention. Therefore, it is highly cost effective to detect a malaria epidemic in its early stages in order to optimize disease control and intervention in Addis-Zemen. However, up to now, there are no related articles for forecasting of monthly malaria cases in Addis-Zemen, South Gondar, Ethiopia. For early detection, prevent and control; forecasting the coming malaria cases are very important and this study aims to develop an appropriate model for forecasting malaria cases in Addis-Zemen.

According to U. Helfenstein, 1991 ARIMA models are useful in modeling the temporal dependence structure of a time series and a useful tool in epidemiological surveillance as they are particularly useful for diseases which show a seasonal pattern. In this study, the monthly malaria cases data from January 2007 to June 2016 was collected in Addis-Zemen, South Gondar, Ethiopia.

Family of ARIMA models are good models for modeling time series data, and it needs the data to be stationary, whereas family of ARCH-GARCH models deal with non-stationary of the time series data. The goal of this paper was to develop a stochastic model for forecasting malaria cases using the data which was obtained from Addis Zemen, South Gondar, Ethiopia. For the development of the model different stages have been discussed, like model formulation, identification, estimation and diagnostic checking for both the family of ARIMA and ARCH-GARCH models. The plots of autocorrelation function (ACF)(for identifying significant MA terms) and partial autocorrelation function(PACF) (for identifying significant AR terms) were applied for identifying the model. By using such techniques possible suggested family of ARIMA models, $SARIMA(1,1,1)(1,1,1)_{12}$, $SARIMA(1,1,2)(1,1,1)_{12}$, $SARIMA(1,1,2)(2,1,1)_{12}$, $SARIMA(1,1,2)(2,1,2)_{12}$, $SARIMA(1,1,1)(2,1,2)_{12}$, $SARIMA(1,1,1)(2,1,1)_{12}$, $SARIMA(1,1,1)(1,1,2)_{12}$, $SARIMA(2,1,1)(1,1,2)_{12}$, $SARIMA(2,1,1)(1,1,1)_{12}$, $SARIMA(2,1,1)(2,1,1)_{12}$, $SARIMA(2,1,1)(2,1,2)_{12}$, $SARIMA(1,1,3)(1,1,1)_{12}$, $SARIMA(1,1,3)(2,1,2)_{12}$, $SARIMA(2,1,2)(2,1,2)_{12}$, $SARIMA(2,1,3)(2,1,1)_{12}$, $SARIMA(2,1,3)(1,1,2)_{12}$ and $SARIMA(2,1,3)(2,1,2)_{12}$ were developed and criteria based such as SBC, SIC and AIC were applied to find out the best fitting model.

As explained on section 3.5.3 and 3.6.3.2 the parameters were estimated using the least square and maximum likelihood methods under the normality assumption. Plots of residuals from the estimated models and significant test via the *p*-values are used to validate the goodness of fit of the model. After the parameters estimated diagnostic checking was done for both the family of models and forecasting methods was outlined for both families of models. As the malaria cases data pattern showed both trend and seasonal variation Holt-Winter forecasting methods were applied to forecast the future 30 months of malaria cases from July 2016 to December 2018.

Every analysis such as plots and tables of the results analyzed using R-software version 3.3.1 and R-studio. The analysis showed that malaria cases data changing mean and unstable variance with upward and downward trend and seasonal variation. This prompted us to fit ARIMA models with both trend and seasonal terms in order to capture these variations, hence the best fitting ARIMA model is SARIMA $(1, 1, 1)(2,1,1)_{12}$. The Ljung–Box Q test given in Table (4.4.1) showed that a significant p-value. This is an indication of ARCH effect in the malaria cases series. Clear evidence to reject the null hypothesis of no ARCH effect was established from the fitted models for the malaria cases series. Hence, it indicates that GARCH modeling is necessary from the malaria cases series and ARCH-GARCH models were established to be plausible as they accommodate the time-varying variance nature of the data, Hence the best fitting ARCH-GARCH model is GARCH (1, 1).

Based on the analysis of the ARIMA (Table 4.3.1) and ARCH-GARCH (Table 4.4.2) family of models the criteria (AIC and BIC) has shown that the ARCH-GARCH modelling is superior than the seasonal ARIMA modeling because of GARCH model has smaller AIC and BIC values which explains the variation in the data better than the seasonal ARIMA model.

The two families of models were used to compute 30 months (from July 2016 to December 2018) forecasts for the malaria cases series. The forecasts from the seasonal ARIMA and GARCH models are given in Tables (4.3.4) and (4.4.4) respectively together with their respective 95% confidence intervals (CI) for each forecast value. The narrower the confidence interval the better the forecasts, (Granger and Newbold, 1986 and Granger,1989). The CI's from the ARIMA model are narrower than the CI's from the GARCH model in the early months of the forecasting period and became wider the further into the future a forecast is. This

probably indicates that the ARIMA model is better for short term forecasting than the GARCH model.

Several studies have used ARIMA model to fit and predict changing trends in infectious disease like malaria. IJSTR; 2015 used ARIMA for Times Series Analysis Of Malaria Cases In Ejisu- Juaben Municipality, Ashanti Region of Ghana. Varun et al, 2014 used ARIMA models for predicting monthly malaria slide positive using climatic factors including; mean monthly rainfall, mean maximum temperature and relative humidity, as risk factors in Delhi, India. Ezekie et al, (2014) used SARIMA models to model and forecast malaria mortality rate in Nigeria. Lin et al., (2009) have used ARIMA models for time series analysis to investigate the relationship between the falciparum malaria in the endemic provinces and the imported malaria in the non-endemic provinces of China. Asamoah et al. (2008) in their work used family of ARIMA models for total OPD reported cases, for Admission reported cases, for female OPD reported cases and for OPD pregnant cases in malaria reported cases. Wangdi et al, 2010 carried out an ARIMA model to develop prediction and forecasting models for malaria incidence in seven of the twenty malaria endemic districts in Bhutan. Generally this study agrees with all the above studies in cases of ARIMA models and differs in ARCH- GARCH family of models.

## 5. Conclusions

The ability to forecast future malaria cases will facilitate timely planning and implementation of control, prevention and case management interventions through optimal distribution of the available resources and this paper aims for developing of stochastic time series model for forecasting malaria cases in Addis Zemen. The best fitting model was selected based on how well the model captures the stochastic variation in the data. Based on minimum Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) values, it was observed that the best fit ARIMA model was SARIMA$(1,1,1)(2,1,1)_{12}$ and the best fit GARCH model was GARCH(1,1). Using both the selected models thirty months (from July 2016 to December 2018) malaria cases of Addis-Zemen were forecasted. Based on the forecasted values it is observed that the forecasted malaria cases are close to the actual malaria cases. From the results SARIMA$(1,1,1)(2,1,1)_{12}$ and GARCH(1,1) the further into the future a forecast is the wider the confidence limits indicating that the model has low forecasting power although it fits the data well but while observing the two models GARCH(1,1) has narrower confidence limits as compared to SARIMA$(1,1,1)(2,1,1)_{12}$ and hence GARCH(1,1) superior to forecast malaria cases in Addis-Zemen.

Finally, the general recommendation goes directly to Addis-Zemen health center and the concerned bodies for more control, prevention and intervention, because this study really shows that the forecasted malaria cases is likely to continue close to the actual value over time, which leads obviously to the serious number of malaria cases if nothing is done accordingly. Also, the researcher suggests that further studies can be considered as extensions and improvements to the GARCH model. These are the integrated GARCH (IGARCH), the exponential GARCH (EGARCH) and the stochastic volatility models and also a wider coverage of the study area is suggested for better results.

## Reference


1. Addis Continental Institute of Public Health (ACIPH)Report Submitted to Academy for Educational Development (AED) and NetMark.2009.
2. Akaike, H. 1974. *A new look at the statistical model identification. IEEE transactions on automatic control*. 19 (6):719-723.
3. Adhanom et al. 2006. *Epidemiology and Ecology of Health and Disease in Ethiopia*. Addis Ababa: Shama Books. PP. 556–576.
4. Amhara regional health bureau, PHEM annual report. 2011/2012. Bahir-Dar, Ethiopia.
5. Ansley, C.F and Newbold, P. 1986. "Stochastic processes and their applications." *Science direct*, 11(2): 201-206.
6. Asamoah et al. 2008. *Time Series Analysis of Malaria cases Ejisu-Juaben Municipality*.
7. Bollerslev et al. 1993. *ARCH models*. Discussion paper University of California, San Diego.
8. Bollerslev, T. 1986. "Generalized autoregressive conditional heterescedasticity." *Journal of econometrics*, Vol. 51: 307-327.
9. Box, G.E.P and Jenkins, GM. 1976. *Time series analysis: forecasting and control*. Holden-Day, Boca Raton.
10. Box, G.E.P and Pierce, D.A. 1970."Distribution of the autocorrelation in ARIMA time series models." *Journal of American statistical association*, Vol. 65: 1509-1526.
11. Brockwell et al. 2002. *Introduction to time Series and forecasting*.2nded. Springer Verlag.
12. Briet et al. 2008. "Temporal correlation between malaria and rainfall in Sri Lanka."*Malaria J*.
13. Chatfield, C. 2004. *The analysis of time series: an introduction*. 3rd ed. Chapman and Hall text in statistical science, London.
14. Cryer, J. D. and K.S. Chan.2008. *Time Series Analysis with Application in R*. 2nd ed. Springer, New York.
15. Deressa et al. 2003. "Self-treatment of malaria in rural communities Butajira, southern Ethiopia."*PMC free article*, BullWorld Health Organ, 81:261–268.
16. Diggle et al. 2002. *Analysis of LongitudinalData*. 2nd ed. Oxford England: Oxford University press.

17. Engle, R.F. 1982."Autoregressive Conditional Heteroscedasticity with estimates of variance of the United Kingdom inflation."*Econometrical*, 4(50): 987-1006.
18. Engle, R.F. 1982. "The use of ARCH/GARCH models in applied econometrics." *Journal of Economic Perspective*, 15(4): 157-168.
19. Ezekie et al. 2014."Modelling and Forecasting Malaria Mortality Rate using SARIMA Models (A Case Study of AbohMbaise General Hospital, Imo State Nigeria)."*Science Journal ofApplied Mathematics and Statistics*,2:31-41.
20. Federal Ministry of Health. 1999. *Malaria and Other Vector-borne Diseases Control Unit*. Addis Ababa, Ethiopia.
21. Githeko, A. 2008."Malaria, Climate Change and Possible Impacts on Populations in Africa."Edited by Carael, M and Glynn JR. *International Studies in Population HIV Resurgent Infections and Population Change in Africa*. Springer; New York. pp. 67–77.
22. Gourieroux et al. 1997. *Time series and dynamic models*. United Kingdom: Cambridge university press.
23. Granger, C.W.J. 1989. *Forecasting in Business and Economics*. 2nd ed. USA: Academic press Inc.
24. Granger, C.W.J and Newbold, P. 1986. *Forecasting Economic time series*. 2nd ed. Economic theory, Econometrics and Mathematical Economics. USA: Academic Press.
25. Harvey, A.C. 1991. *Forecasting, structural time series models and kalman filter*. UK: Cambridge university press.
26. Hay et al. 2000. "Earth observation, geographic information systems and plasmodium falciparum malaria in Sub-Saharan Africa." *Adv. Parasitol*. 47: 173–215.
27. *(IJSTR) International Journal of Scientific & Technology Research*, volume 4, issue 06, june 2015.
28. IPCC, Climate Change. 2007. "The Physical Science Basis Contribution of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change." Cambridge University Press; Cambridge, UK and New York, USA. pp. 1–996.
29. Lesaffre, E and Spiessens, B. 2001. "On the effect of the number of quadrature points in a logistic random-effects model." *Applied Statistics*. 50:325–335. doi: 10.1111/1467-9876.00237.
30. Lin et al. 2009. "Spatial and temporal distribution of falciparum malaria in China." *Malaria journal*, 8(30).
31. Lopez, J.A. 1999. *Evaluating the predictive accuracy of volatility models, Economic research*. Department of Federal Reserve Bank of San Francisco.
32. Malaria Early Warning System (MEWS). http://iridl.ldeo.columbia.edu/maproom/.Health/.Regional/.Africa/.Malaria/MEWS/ (accessed1 August 2008).
33. MARA/ARMA. 1998. "Towards an Atlas of Malaria Risk in Africa".; South Africa: Albany Print Ltd. pp. 1–31.
34. MOH. 2007/2008. *Health and Health Related Indicators*. Planning and Programming Department, Federal Democratic Republic of Ethiopia Ministry of Health, Addis Ababa.
35. Sachs, J and Malaney, Pia. 2002. *The Economic and Social Burden of Malaria*. Macmillan Pub. Ltd.415: 680 -685.
36. Stoffer, D.S. and R.H. Dhumway.2010. *Time Series Analysis and its Application*. 3rd ed. Springer, New York. pp. 596.
37. Talke, I.S. 2003. *Modelling volatility in time series data*. MSc thesis, University of Kwa-Zulu Natal.
38. Test of normality. http://en.wikipedia.org/wiki/Q − Q − plot.
39. Thomson et al. 2005. "Use of Rainfall and Sea Surface Temperature Monitoring for Malaria Early Warning in Botswana." *American Journal Tropical Medicine and Hygiene* 73(1): 214-221.
40. Tsay, R.S. 2002. *Analysis of financial time series*. 2nd Ed. New York.
41. Tulu, NA.1993. "The Ecology of Health and Disease in Ethiopia."*Malaria J*. Edited byKloos H, Zein AZ. USA: Westview Press Inc. pp. 341–352.
42. Wangdi et al. 2010. *Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan*. *Malaria J*, 9(251).
43. WHO. 2006. *Systems for the early detection of malaria epidemics in Africa: an analysis of current practices and future priorities, country experience*. Geneva, Switzerland.
44. World Health Organization. 2014. *Malaria Report*.
45. World Health Organization. 2015. *Malaria Report*.
46. U, Helfenstein. 1991. "The use of transfer function models, intervention analysis and related time series methods in epidemiology." *International Journal of Epidemiology*, 20(3): 808–815.
47. Varun et al. 2014. *Forecasting Malaria Cases Using Climatic Factors in Delhi, India: A Time Series Analysis*.
48. Ye Y et al. 2007. *Effect of meteorological factors on clinical malaria risk among children: an assessment using village-based meteorological stations and community-based parasitological survey*.
49. Zhou, G et al.2004. "Association between climate variability and malaria epidemics in the East African highlands". *Proc Natl Acad Sci*, USA. 101:2375–2380. doi: 10.1073/pnas.0308714100.

3/15/2017