Strathmore UNIVERSITY

Predicting malaria incidence in Kenya using the ARIMA and SARIMA models

BY:

Ali Amira Abdulkadir

100409

Strathmore Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

[July 2020]

ABSTRACT

Malaria is considered a public health challenge across the world. Approximately 40 percent of the population of the world is at risk of malaria. In this study we will employ time series analysis models to predict malaria incidence and it will also use climate variables such as temperature and rainfall as exogenous inputs. Future malaria incidences will be projected based on determined trend patterns. The Auto-regressive integrated moving average (ARIMA) and Seasonal Auto-regressive integrated (SARIMA) models were used in this study to predict and forecast monthly malaria incidence (the spread of malaria in Kenya). Considering the results, ARIMA (0, 1, 0) appeared fit for forecasting monthly malaria incidence in Kenya further on the SARIMA model was used to compare which model had the best results and to remove seasonality from the data the best fit for SARIMA model was (0,1,0) (0,1,0)[12].The models that usually give slightly better results are the ones that have the lowest AICc values. In addition to that a regression analysis was carried out to determine the effects of rainfall and temperature on malaria incidence in Kenya. The variables have different orders; we estimate a VAR regression because VAR regression enables us to dynamically measure variables with combination I different order. The results obtained from the regression analysis indicate that temperature has no significant impact on the number of malaria cases however rainfall has a significant impact.

LIST OF EQUATIONS

LIST OF TABLES

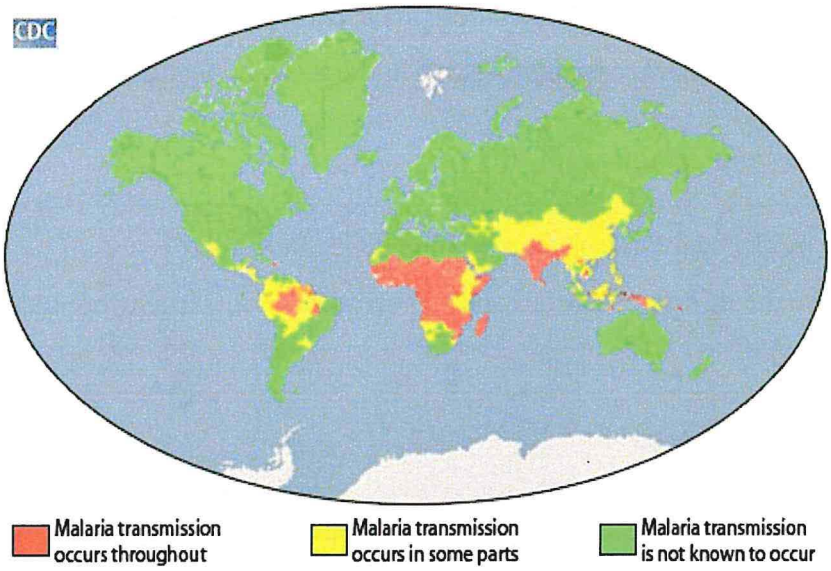

Figure 1. An approximation of the parts of the world where malaria transmissions occur.

Source https://www.cdc.gov/malaria/about/distribution.html

Studies have shown that approximately 70 percent of the population is at risk for malaria and it accounts for almost 16 per cent of outpatient consultations in Kenya. The transmission of malaria in Kenya is determined by temperature and rainfall patterns. It is important to understand the impact of climatic variables and understand the factors triggering transmission of malaria is important to prevent malaria outbreaks.

To prevent malaria outbreaks and re-introduction of malaria transmission factors triggering transmission should regularly be monitored (Ranjbar, et al., 2016). Therefore, understanding the impact of climatic variables on malaria prevalence is important for effective policy and management decisions locally and globally.

The President's Malaria Initiative (PMI) was launched in 2005. It was to play an important role in the prevention and treatment of malaria (Ye & Duah, 2019). The long-term strategy included treatment interventions and preventive measures. The goal by 2040-2050 is to eradicate malaria worldwide and free malaria zones in Africa

1.3. Research objectives and questions

The general objective of this research is to forecast malaria case admissions in Kenyan Health Facilities. The specific objectives are:

1. Develop malaria forecasting models in Kenya

early warning signals of changing malaria trends in the public health surveillance (Lewnard, Parikh, & Pitzer, 2016).

The aim of this study is finding an accurate model in predicting future malaria incidences in Kenya using previous malaria historical data. These will be helpful for malaria control and in elimination efforts in Kenya to understand the risk and severity of malaria epidemic and to make decisions on effective control actions. However, an ineffective or poor model can lead to incomplete plan and bad management of malaria (Malinga, 2015).

District, Ghana. The possible decline was due to the effective intervention strategies put in place by the government.

Obtaining advance information about the location and timing of malaria epidemics facilitates and enables efficient allocation of the available resources for prevention and emergency response (Wimberly, et al., 2014). In conformity with (Cox, 2007) it is possible to deliver epidemic control interventions within days of the malaria case being identified under existing modes of surveillance. Best and effective responses need to be screened first and selectively applied. Different countries adapt to different practices and this is obtained through interactive research.

2.3. Empirical Framework

In this section we look at the different approaches used in predicting malaria incidences in malaria prone areas across different countries around the world. This section discusses the empirical framework using the ARIMA models, Poisson Autoregressive models, GAMBOOST model.

Appropriate mathematical and statistical methodologies have been developed to measure the malaria burden in many malaria endemic countries such as those in SSA, with malaria forecasting being a critical part of many malaria control programmes and research organizations worldwide (Malinga, 2015). Disease forecasting has no single approach, various different methods and approaches have been adopted to forecast health conditions (Permanasari, Hidayah, & Bustoni, 2013). The selection of an appropriate method is important to achieve a better prediction.

2.3.1: Poisson Autoregressive model

In Nigeria, (Abdulkadir & Mutah, 2018) used the Poisson Autoregressive model (PAR) to model the trend and forecasts of malaria incidences. Poisson autoregressive PAR (p) methodology is used in modeling, forecasting the time series count data and the future based on the time series model. The author used it in this research because it is considered a natural starting point to extend to dependent count data.

We first define The PAR model:

others that the Government should ensure the provision of insecticides, insecticide-treated nets and anti-malaria drugs in the rural areas in Nigeria. The author found the ARIMA model to be useful in forecasting malaria epidemic hence should be used in future studies on the outbreak of malaria epidemic in Nigeria and other African countries.

2.3.3: SARIMA model

(Ebhuoma, Gebreslasie, & Magubane, 2018) employed the SARIMA model in predicting and modelling malaria cases in KwaZulu-Nata (low transmission region in South Africa. Malaria monthly cases were predicted using R statistical software. The SARIMA forecast model can serve as a useful tool for public health workers. The author indicates that the information obtained can provide relevant authority to act proactively and can also be applied as a malaria early-warning system.

(Permanasari, Hidayah, & Bustoni, 2013) analyses and uses the SARIMA method in developing a forecasting model that provided prediction of malaria incidence. Forecasting methods are very useful when it comes to predicting number of disease occurrences. It can also serve as a tool for providing relevant information to locals and visitors prior to high malaria transmission months.

2.3.4: GAMBOOST model

(Sewe, et al., 2017) developed and compared statistical models using malaria data to forecast malaria admissions. The data was obtained from a district hospital in Western Kenya. Two models were tested to forecast malaria admissions. In predicting monthly admissions at the district hospital, the GAMBOOST model proved to have the best accuracy

The findings of the study indicate that using boosting regressions in GAM models is beneficial in early warning systems hence providing enough time in intensifying malaria control interventions. Malaria response plans are therefore put in place

2.3.5 Comparing different approaches

(Hussien, Eissa, & Awadalla, 2017) used four methods of time series analysis to forecast and predict malaria incidences in Sudan with patterns of historical data alone. The models used were: Autoregressive integrated moving average (ARIMA); moving average (MA); transformation model and exponential model. Forecasting methods is useful in predicting

be used to provide information to support policy makers and public health efforts so that intervention resources can be provided and channelled in a sustainable and effective way.

Conforming to (Malinga, 2015) the ability to forecast future malaria incidence is a major milestone as it facilitates the implementation, planning , prevention of malaria incidences trough optimal distribution of the available resources.

2.5. Conclusion

This chapter captured the different studies that have been done and different approaches with regards to predicting the malaria incidences in malaria prone areas across different parts of the world. It allows for the review work of different authors in both theoretical and empirical literature and the relevance of forecasting malaria trends. In line with (Bansal, Azad, & Liò, 2013) it is important to monitor malaria trends to see if the malaria control campaigns are effective and making improvements.

ARIMA models are regressions that are designed to account for serial autocorrelation in time series. (Zinszer, et al., 2015). With the ARIMA models the random error term ε_t is assumed to be a white noise (Malinga, 2015) .This implies that it is identically and independently distributed and has a mean zero and common variance σ^2. $\varepsilon t \sim iid(0, \sigma^2)$.

The parameters of the ARIMA model are defined as follows:

- **p**: is the number of lag observations included in the model (represents the partial autocorrelation between data).
- **d**: The degree of differencing (value representing the differential of the trend)
- **q**: The order of moving average (represents the autocorrelation between the data).

The autoregressive (AR) process and the moving average process (MA) are combined. For the AR process Yt is expressed as a function of its past values. For the MA process Yt is expressed as a function of the error term ε. The AR does autocorrelation between current and past observations, The MA describes the autocorrelation structure of the error (residuals).

Equation III

$$Yt = \theta Y_{t-1} + \theta Y_{t-2} + \dots\dots\dots + \theta_p Y_{t-p} - \varepsilon t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots \dots \phi_q \varepsilon_{t-q}$$

Where

- θs are the coefficient of the AR process
- ϕs are the coefficient of the MA process, p and q are the number of past values of Y_t.

or a process yt is ARIMA (p, d, q) if:

Equation IV (non-seasonal case)

$$\Phi_p (L) \ (1 - L)^d \ y_t = \theta_q (L) \varepsilon_t$$

where

- Φ_p (L) is the AR polynomial of order p
- θ_q(L) is the MA polynomial of order q

residuals are studied therefore determine whether the autocorrelations and partial autocorrelations are significant hence the model is considered adequate. One simple test of the chosen model is to see if the residuals estimated from the model obtained are white noise. If the model is a white noise, we accept the fit. For the SARIMA model to test the adequacy of the selected model we use the Box-Ljung test.

Step 4: Forecasting - At this stage, the estimated parameters are used to obtain the forecasted values. Forecasting can only be done once differencing of data has taken place. The forecasted values will be compared to the actual values of rainfall. The SARIMA model that is selected in step 3 will be used to forecast malaria cases.

3.5 Regression Analysis

This forms the second part of methodology. The aim of this regression analysis is to be able to analyze how malaria incidence is affected by rainfall and temperature. We will then evaluate whether incorporating climate variable will improve the models' fit and forecasting ability. R statistical package and Stata 12 will be used to carry out the analyses and develop the ARIMA models.

Estimation steps that will be used in Stata are the first step is to specify the model. We then time set the application, if this is not done then Stata will not perform the time series analysis. The third step is to then perform the stationarity tests. The fourth step is to then determine the optimal lag length of the model. The fifth step is to estimate the VAR model. The sixth step is to perform some diagnostic tests which entails autocorrelation, normality, and stability.

The Figure above displays the time series plot (Yt) of the reported monthly Malaria incidence in Kenya. The plot shows 108 monthly observations from January 2008 up to December 2016.

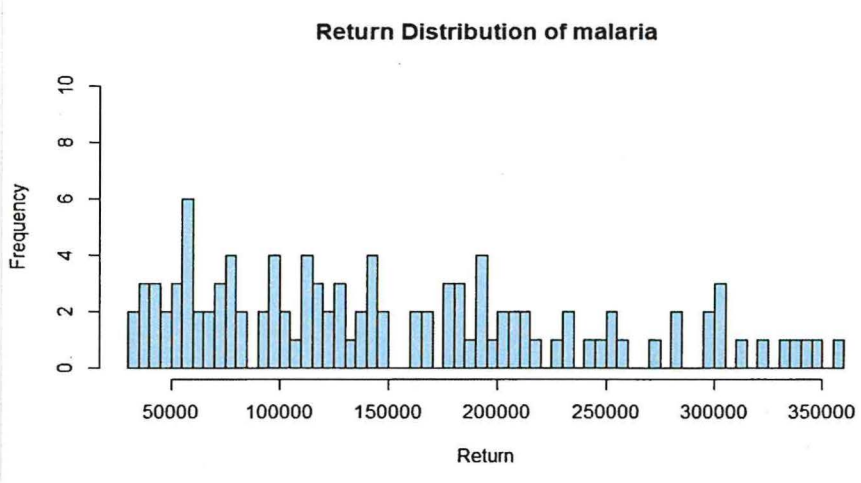

Figure 4: Seasonal indexes of malaria incidence

4.2 ARIMA model

4.2.1 Model Identification

We carry out the Augmented Dickey Fuller test to test for stationarity. The null and alternative hypothesis for the Augmented dickey fuller test is:

$$H_0 : No\ stationarity\ (presence\ of\ unit\ root)$$

$$H_a : Stationarity\ exists$$

Data	Malaria
Dickey-Fuller	-5.0326
Lag order	4
P-value	0.01
Alternative hypothesis	Stationary

Table 2: Augmented Dickey Fuller test

4.2.3 Diagnostic checking

Data	Malaria
X-squared	35.895
df	20
p-value	0.01582

Table 4: Box pierce Test

P-value is less than 0.05 hence it is statistically significant hence a non-stationary signal will have a low p-value.

4.2.4 Forecasting

The final step for Stochastic ARIMA modelling is to forecast. The data was forecasted with respect to 5 years (2017-2022)

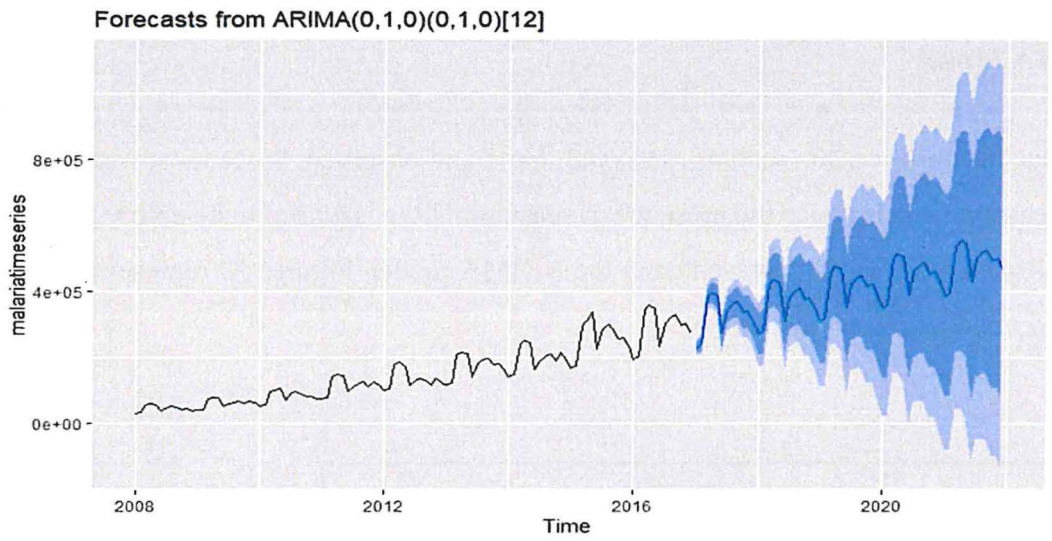

Figure 7: The graph for the five-year malaria forecasts using ARIMA model

As seen from the graph above it can be observed that in Kenya the malaria cases will experience a steady increase.

Figure 9: Monthly malaria incidence in Kenya

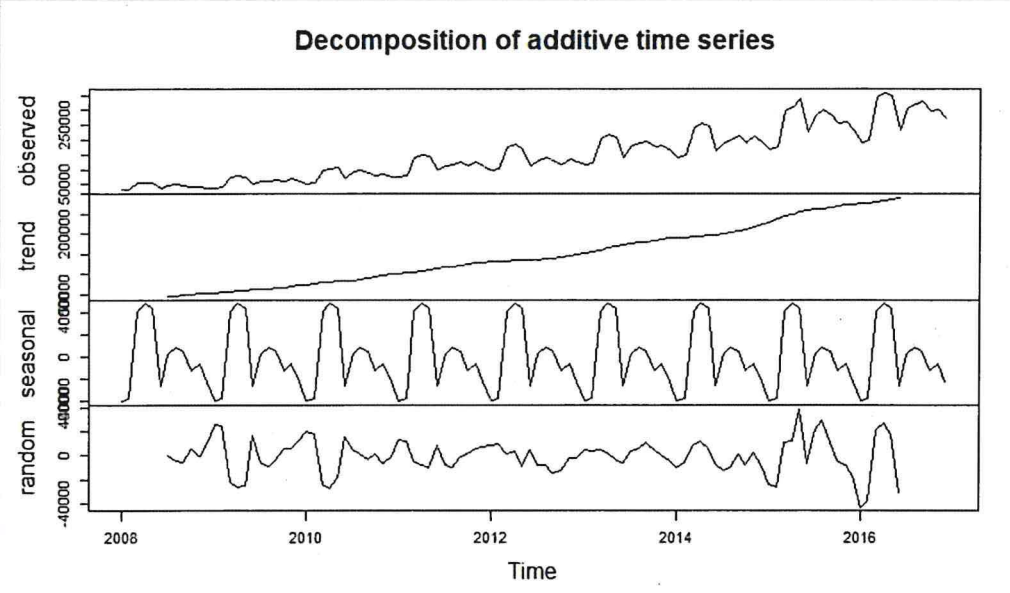

Figure 10: Decomposed malaria dataset

The data was decomposed into random, seasonal, trend and observed.

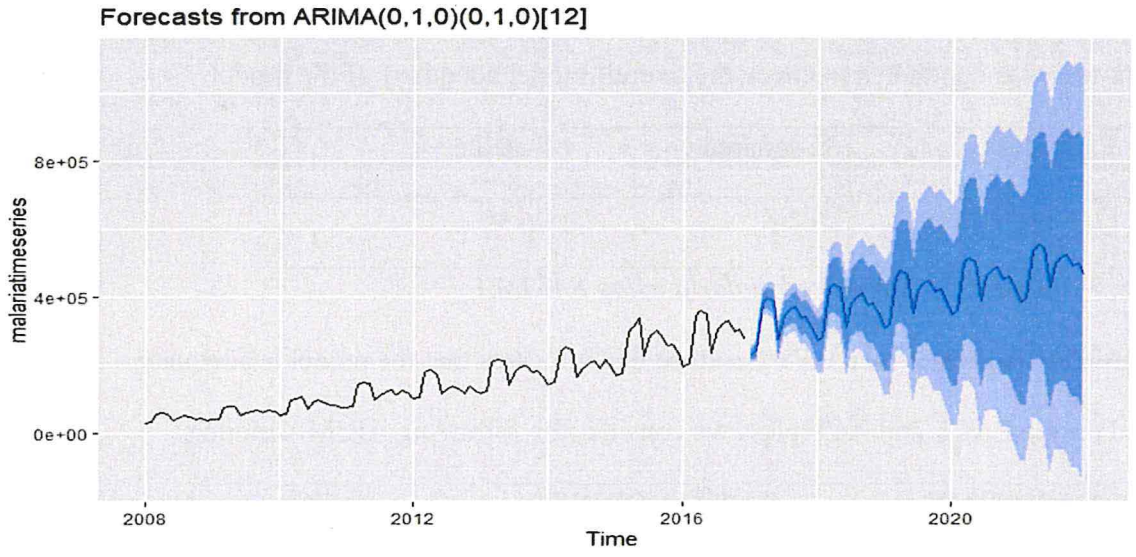

Figure 11: Sarima malaria forecasts

Therefore, from the analysis SARIMA is a better model in forecasting climate variables than ARIMA model. The models that usually give slightly better results are the ones that have the lowest AICc values.

4.5 Regression Analysis

We run a regression analysis using Temperature, Rainfall and Malaria as our variables. The first step followed is to check for stationarity. Since our variables have different orders, we estimate a VAR regression because VAR regression it enables us to dynamically measure variables with combination I different orders. We also need to know the optimal lag selection for us to estimate the VAR.

The null and alternative hypothesis for the Augmented dickey fuller test is:

$$H_0 : No\ stationarity\ (presence\ of\ unit\ root)$$

$$H_a : Stationarity\ exists$$

VARIABLE	COEFFICIENT	P-VALUE
Malaria Cases		
Rainfall L1	-282.5701	0.003
Rainfall L2	-419.0856	0.000
Temperature L1	5896.464	0.136
Temperature L2	8768.851	0.050
Rainfall		
Malaria Cases L1	0.00016	0.058
Malaria Cases L2	0.000064	0.446
Rainfall L1	0.2831559	0.002
Rainfall L2	-0.2460905	0.005
Temperature L1	16.6747	0.000

CHAPTER 5 CONCLUSION AND RECOMMENDATIONS

The overall objective of the study was to forecast malaria case admissions in Kenya and ccompare the applicability and assess the accuracy of the different forecasting methods in predicting malaria case admissions in the Kenyan health facilities. From the analysis SARIMA is a better model in forecasting climate variables than ARIMA model because it has a lower AIC value.

The results obtained from the regression analysis indicate that temperature has no significant impact on the number of malaria cases however rainfall has a significant impact. As seen from the results obtained in chapter 4 it can be observed that in Kenya the malaria cases will experience a steady increase. The forecasts obtained enables the health authorities in monitoring malaria outbreaks, prioritizing allocation of resources and taking necessary actions to prevent the spread especially in areas that are prone to Malaria in Kenya

It is recommended that in some instances, indoor residual spraying should be applied to minimize environmental contamination. Indoor residual spraying is suggested for application prior to the rainy season to prevent and control epidemic outbreaks. The use of Insecticide Treated Nets is also recommended especially in the rainy seasons when malaria cases increase.

This research study investigates the relationship between malaria incidence and climate variables (temperature and rainfall). Future studies should consider using non-climatic variables such as level of immunity in human hosts, migration, and urbanization. Future researchers should incorporate the impact of migration and urbanization on the spread of malaria incidence.

Darkoh, E. L., Larbi, J., & Lawer, E. A. (2017). A Weather-Based Prediction Model of Malaria Prevalence in Amenfi West District, Ghana. *Malaria Research and Treatment, 2017*(2017), 7820454-7820454. Retrieved 5 21, 2020, from https://hindawi.com/journals/mrt/2017/7820454/abs

Ebhuoma, O., Gebreslasie, M., & Magubane, L. (2018). A Seasonal Autoregressive Integrated Moving Average (SARIMA) forecasting model to predict monthly malaria cases in KwaZulu-Natal, South Africa. *South African Medical Journal, 108*(7), 573-578. Retrieved 5 25, 2020, from http://samj.org.za/index.php/samj/article/view/12327/8516

Hussien, H. H., Eissa, F. H., & Awadalla, K. E. (2017). Statistical Methods for Predicting Malaria Incidences Using Data from Sudan. *Malaria Research and Treatment, 2017*, 4205957-4205957. Retrieved 5 17, 2020, from https://hindawi.com/journals/mrt/2017/4205957

Lekia, N., & Wonu, N. (2018). *Application of time series analysis on the forecasting of the outbreak of malaria epidemic in Nigeria*. Retrieved 5 17, 2020, from http://mathsjournal.com/pdf/2018/vol3issue6/partb/3-6-13-156.pdf

Lewnard, J., Parikh, S., & Pitzer, V. (2016). *Time series analysis of malaria in Afghanistan : using ARIMA models to predict future trends in incidence.*

Malaria - Bill & Melinda Gates Foundation. (n.d.). Retrieved 5 14, 2020, from Gatesfoundation.org: http://www.gatesfoundation.org/What-We-Do/Global-Health/Malaria

Malinga, J. K. (2015). *Forecasting malaria case admissions in three Kenyan health facilities*. Retrieved 5 17, 2020, from http://erepository.uonbi.ac.ke:8080/xmlui/handle/11295/90009

Malinga, J. K. (2015). *Forecasting Malaria Case Admissions in three Kenyan Health Facilities*

.

Merkord, C. L., Liu, Y., Mihretie, A., Gebrehiwot, T., Awoke, W., Bayabil, E., . . . Wimberly, M. C. (2017). Integrating malaria surveillance with climate data for outbreak detection and forecasting: the EPIDEMIA system. *Malaria Journal, 16*(1), 89-89. Retrieved 5 17, 2020, from https://ncbi.nlm.nih.gov/pubmed/28231803

Zinszer, K., Kigozi, R., Charland, K., Dorsey, G., Brewer, T. F., Brownstein, J. S., . . . Buckeridge, D. L. (2015). *Forecasting malaria in a highly endemic country using environmental and clinical predictors* .

Zinszer, K., Kigozi, R., Charland, K., Dorsey, G., Kamya, M. R., & Buckeridge, D. L. (2014). Predicting Malaria in a Highly Endemic Country using Environmental and Clinical Data Sources. *Online Journal of Public Health Informatics, 6*(1). Retrieved 5 24, 2020, from https://ncbi.nlm.nih.gov/pmc/articles/pmc4050902

```
## Forecast
F <- forecast(malariamodel, 60)
F
library(ggplot2)
autoplot(F)
accuracy(F)

library("ggplot2")

## SARIMA MODEL
library("forecast")
library("tseries")
library("MASS")

## Malaria
malariasarima <- ts(monthlymalariadata,start = c(2008,1), frequency = 12)
plot(malariasarima)
title(" Kenya Malaria")
components <- decompose(malariasarima)
plot(components)

install.packages("statsmodel")
```

```
install.packages("forecast")
library(forecast)

autoplot(F) + xlab("Year") + ylab("Malaria Forecast")

ggtsdisplay(diff(malariatimeseries,12))

fit <-arima(malariatimeseries, order = c(0,0,1),
                                        seasonal = c(0,1,1))
fit
ggtsdisplay(residuals(fit))

res <- residuals(fit)
res
autoplot(res) + xlab
Box.test(res, lag = 36, fitdf = 12, type = "Ljung")

autoplot(forecast(fit, h= 12))
auto.arima(malariatimeseries)
```