Malaria forecasting in Awka South (2024-2027)

Divine-Favour Ekemezie

December 2023

Despite ongoing efforts to tackle malaria, it remains a major public health challenge in Awka-South. This disease continues to cause high morbidity and mortality, particularly among vulnerable populations. Effective planning and resource allocation to combat malaria require accurate predictions of future prevalence rates.

Time series forecasting is a method used to predict future values based on previously observed data points collected over time. In this type of analysis, the data is sequential, typically consisting of time-ordered variables such as monthly malaria prevalence rates. The data was collected at regular intervals and structured in a sequential manner suitable for ARIMA modeling, sourced from the Chukwuemeka Odumegwu Ojukwu Hospital (Amaku Teaching Hospital), Awka.

The time series analysis to be done in this work will be evaluated using the AutoRegressive Integrated Moving Average (ARIMA) model.
The ARIMA model is a widely used time series forecasting method that combines three components: AutoRegression (AR), Integration (I), and Moving Average (MA).

ARIMA(p, d, q) Model

The ARIMA model is typically defined by three parameters:

- p: The number of lag observations in the autoregressive (AR) part.
- d: The number of differences needed to make the series stationary (I).
- q: The number of lagged forecast errors in the moving average (MA) part.

The general formula for the ARIMA(p, d, q) model can be written as:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_{1\epsilon_{t-1}} + \theta_{2\epsilon_{t-2}} + \cdots + \theta_{q\epsilon_{t-q}} + \epsilon_t \tag{1}$$

Where: Y_t is the current value at time t.
c is a constant.
$\phi_1, \cdots, \phi_p$ are the coefficients of the autoregressive terms.
ϵ_t is the error term (white noise).
$\theta_1, \cdots, \theta_p$ are the coefficients of the moving average terms.
$\epsilon_{t-1}, \epsilon_{t-2}, \cdots \epsilon_{t-q}$ are the past forecast errors.

Components of ARIMA

AutoRegressive (AR) part:

- This is the regression of the variable on its past values.
- The AR(p) component is defined as:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t \tag{2}$$

Integrated (I) part:

- This involves differencing the data to make it stationary (i.e., removing trends and seasonality).
- for instance, if d=1, the differenced series is

$$Y_t' = Y_t - Y_{t-1} \tag{3}$$

Moving Average (MA) part

- This incorporates the dependency between an observation and a residual (forecast error) from a moving average model.
- The MA(q) component is defined as:

$$\theta_{1\epsilon_{t-1}} + \theta_{2\epsilon_{t-2}} + \cdots + \theta_{q\epsilon_{t-q}} + \epsilon_t \tag{4}$$

Merging it all we will have for an ARIMA (1,1,1) model (i.e when p=1, d=1, q=1):

$$Y_t - Y_{t-1} = c + \phi_1 (Y_{t-1} - Y_{t-2}) + \theta_{1\epsilon_{t-1}} + \epsilon_t \tag{5}$$

DATA PRESENTATION AND ANALYSIS

We will present extensively the data and analysis to be conducted using ARIMA modelling. The data on malaria cases in Awka South from 2016-2023 is stated below as collected from Amaku Teaching Hospital, Awka.

Table 1: Malaria cases in Awka South from 2016-2023

Months	Jan	Feb	Mar	April	May	June	July	Aug	Sep	Oct	Nov	Dec
2016	42	21	10	28	16	25	18	16	24	10	18	28
2017	29	13	26	12	48	28	21	18	32	16	36	24
2018	30	46	35	16	20	38	21	44	28	21	25	38
2019	17	31	11	43	8	16	22	28	12	27	38	20
2020	28	16	37	5	20	19	24	26	23	15	22	29
2021	32	12	26	14	36	13	23	8	17	30	13	9
2022	19	34	20	42	26	14	9	33	17	14	42	38
2023	37	24	31	15	28	35	9	15	10	30	33	18

The plot of the yearly average cases of malaria in Awka South is shown below

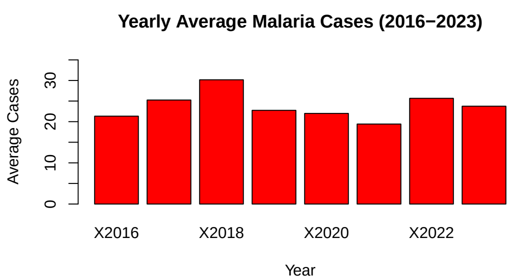

Figure 1: Yearly average malaria cases graph

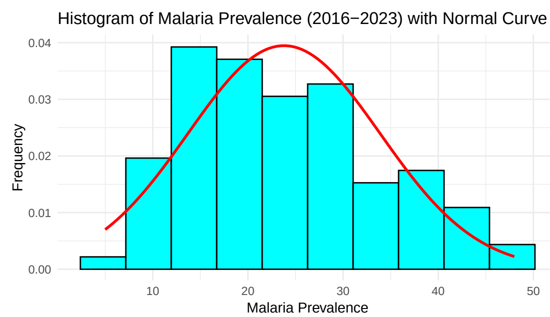

Figure 2: Histogram with normal curve

From the plot above, it can be deducted thatdeducedria prevalence over the years seems relatively stable, with occasional fluctuations. However, the data suggests that there are specific months or years (such as 2017 and 2018) where malaria prevalence spikes, as seen in the range and variability metrics.

Table 2: Descriptive Statistics of Reported Cases of Malaria in Awka South

Variable	N	Mean	SE. Mean	St. Dev.
Malaria	96	23.79	1.03	10.11
Variable	Skewness	kurtosis	Minimum	Maximum
Malaria	0.34	-0.73	5	48

From table 2, the minimum number of malaria cases recorded was 5, while the maximum was 48, indicating that monthly malaria cases in the Awka South ranged from 5 to 48. The average monthly cases were approximately 23.79, suggesting that the expected number of malaria cases per month is around this value. The data distribution is slightly positively skewed, suggesting that while the distribution is fairly symmetrical, it slightly extends toward higher values, meaning there are more small malaria case counts and fewer large ones. The negative kurtosis of -0.73 implies a more uniform spread of values with fewer extreme occurrences, which suggests that extreme highs or lows in malaria cases are relatively rare

Analysis

In order to model the data using ARIMA, we will need to determine if the data is stationary or non-stationary. To determine whether a time series is stationary, we begin by examining the graph of the Autocorrelation Function (ADF) Test and Partial Autocorrelation Function (PACF).

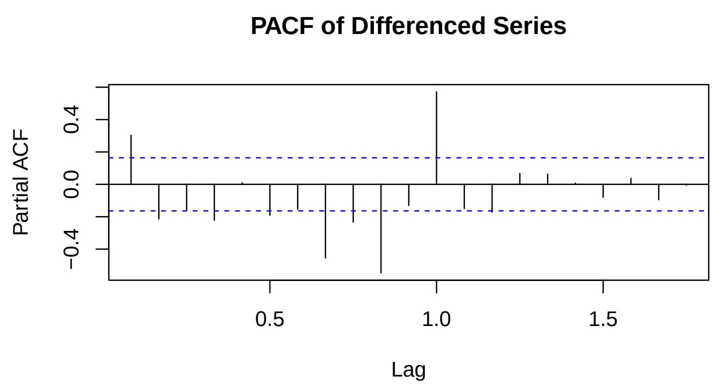

Figure 3: PACF of differenced series

Figure 2 represents the Partial autocorrelation function (PACF). Some of the lags show significant partial autocorrelation, which means that the differenced series still has relationships with some of its past values at those lags. The decay or significant spikes in PACF suggest the possible order of an autoregressive moving average (ARIMA) model that could fit the data.

Table 3: Normality Test Results

Test	Result	P-Value
Komogrov-Smirnov	0.0638	0.8104
Anderson Darling	0.3465	0.4750
Shapiro-Wilk	0.9916	0.7952

To test data, we test the null hypothesis
H_o ; there is no significant difference between the data and generated normal data.
H_1 ;there is a significant difference between the data and generated normal data.

All of the various tests of normatests in table 3 are greater than 0.05. Therefore, we will accept the null hypothesis and conclude that there is no difference between the data and the normal data generated indicating t, indicating that the data is normal.

Further Analysis

Here, we will establish the measure of accuracy using mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean squared error (RMSE). We will be comparing these three models to select the best model as the trend of malaria cases in Awka South.

Table 4: Measure of Accuracy

Model	MAPE	MAE	RMSE
Linear	24.34	6.1160	7.5826
Quadratic	20.50	4.9404	5.5350
Exponential	23.24	6.0439	7.6254

Given these results, the quadratic model, having the lowest values across the board, would be the most suitable model for forecasting malaria prevalence based on the data, as it provides more accurate predictions compared to the linear and exponential models.

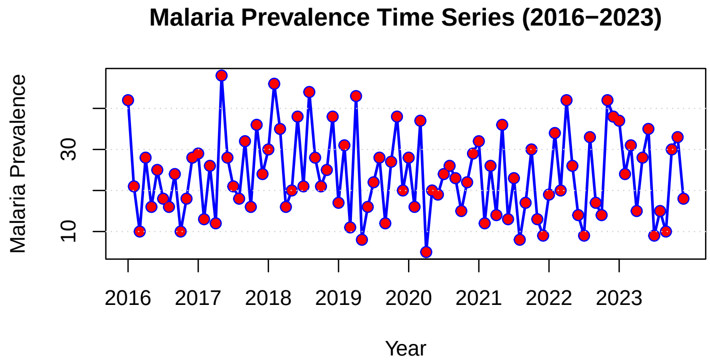

Figure 4: Time Series Plot of Malaria Cases

This plot shows a fluctuating pattern of malaria cases over the years, with no clear upward or downward trend. The prevalence appears to cycle periodically, with noticeable peaks in years like 2018 and 2022, where cases rise close to 30. In contrast, years such as 2019 and 2020 show lower malaria rates, with cases hovering around 10. Despite these fluctuations, the overall prevalence remains relatively stable within a range of 10 to 30 cases, indicating the absence of any extreme surges. The cyclical behavior suggests that malaria cases may be influenced by seasonal or external factors that cause the periodic rise and fall. This pattern could be valuable for forecasting future trends and timing public health interventions to coincide with anticipated peaks.

Now, we will continue to select the best ARIMA model for the prevalence data on malaria cases in Awka South. To achieve this objective, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) methods will be used to compare ARIMA (1,0,1), ARIMA (1,0,3), and IMA (1,0,0) to determine which one fits the data best.

Table 5: Model fitting of different ARIMA models

Models	AIC	BIC
ARIMA(1,0,1)	720.8593	731.1166
ARIMA(1,0,3)	723.8995	739.2856
ARIMA(1,0,0)	719.8276	727.5206

The model with the lowest AIC and BIC is ARIMA(1,0,0) which has an AIC value of 719.8276 and BIC value of 727.5206 . Since AIC measures the trade-off between model complexity and goodness of fit, the lower the AIC, the better the model fits the data while avoiding overfitting. Although ARIMA(1,0,1) is a close second, it is slightly more complex than ARIMA(1,0,0), yet the improvement in fit is not enough to justify the added complexity. On the other hand, ARIMA(1,0,3) has the highest AIC of 723.8995, indicating that it is the least favorable model among the three in terms of fitting the data while maintaining parsimony.

Model Fitting and Forecasting

With the help of the ACF and PACF plots, tentative models were fit to the data. ARIMA (1,0,0) was noted to fit the data well. The model was then used to forecast monthly cases of malaria for the next two years in Awka South as shown in the graph and table below.

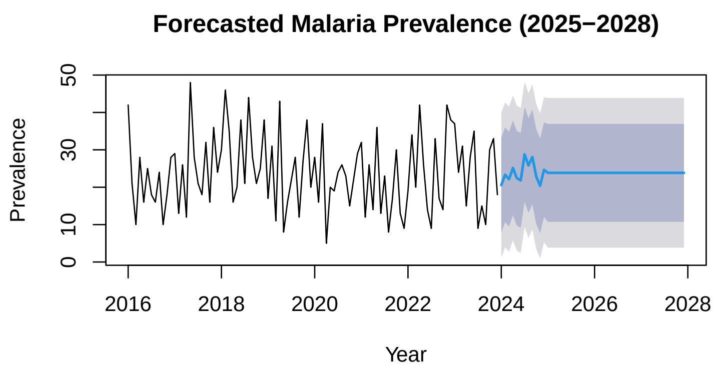

Figure 5: Forecasted Malaria Prevalence for 2025-2028

The plot of the forecasted malaria prevalence from 2024 to 2027 suggests that malaria cases will continue to fluctuate within a range similar to previous years, with prevalence generally staying between 10 and 30 cases. The forecast predicts some periods with higher prevalence approaching 30 cases, while other periods are expected to have lower levels around 10 cases, indicating the persistence of moderate variability in malaria prevalence over the forecast period.

Summary of ARIMA(1,0,0)

- AR(1): Uses one lagged value of the time series for prediction
- I(0): No differencing is needed; the time series is stationary.
- MA(0): No moving average terms are included; residuals are not adjusted using past errors.

This shows that p=1, d= 0 and q=0

The table below shows the forecast of malaria in Awka South using ARIMA (1,0,0), the model that best fits the data.

Table 6: Forecasted Malaria Cases in Awka South for 2024-2027

Period	**Forecast**	**Lo 80**	**Hi 80**	**Lo 95**	**Hi 95**
Jan-2024	24.56057	11.79326	37.32787	5.03465	44.08648
Feb-2024	23.66964	10.78515	36.55414	3.96451	43.37478
Mar-2024	23.79063	10.90398	36.67728	4.0822	43.49906
Apr-2024	23.7742	10.88751	36.66089	4.06571	43.48269
May-2024	23.77643	10.88974	36.66312	4.06794	43.48492
Jun-2024	23.77613	10.88944	36.66282	4.06764	43.48462
Jul-2024	23.77617	10.88948	36.66286	4.06768	43.48466
Aug-2024	23.77616	10.88948	36.66285	4.06768	43.48465
Sep-2024	23.77616	10.88948	36.66285	4.06768	43.48465
Oct-2024	23.77616	10.88948	36.66285	4.06768	43.48465
Nov-2024	23.77616	10.88948	36.66285	4.06768	43.48465
Dec-2024	23.77616	10.88948	36.66285	4.06768	43.48465
Jan-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Feb-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Mar-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Apr-2025	23.77616	10.88948	36.66285	4.06768	43.48465
May-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Jun-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Jul-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Aug-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Sep-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Oct-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Nov-2025	23.77616	10.88948	36.66285	4.06768	43.48465
Dec-2025	23.77616	10.88948	36.66285	4.06768	43.48465

Table 7: Forecasted Malaria Cases in Awka South for 2024-2027

Period	**Forecast**	**Lo 80**	**Hi 80**	**Lo 95**	**Hi 95**
Jan-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Feb-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Mar-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Apr-2026	23.77616	10.88948	36.66285	4.06768	43.48465
May-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Jun-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Jul-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Aug-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Sep-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Oct-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Nov-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Dec-2026	23.77616	10.88948	36.66285	4.06768	43.48465
Jan-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Feb-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Mar-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Apr-2027	23.77616	10.88948	36.66285	4.06768	43.48465
May-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Jun-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Jul-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Aug-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Sep-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Oct-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Nov-2027	23.77616	10.88948	36.66285	4.06768	43.48465
Dec-2027	23.77616	10.88948	36.66285	4.06768	43.48465

The forecasted results for malaria prevalence from 2024 to 2027 indicate a stable point estimate of approximately 23.77, beginning in July 2024. This consistency suggests that the malaria prevalence rate is projected to remain relatively unchanged, with minimal fluctuations expected throughout the forecast period. The confidence intervals (both 80% and 95%) remain wide, reflecting a degree of uncertainty in the forecasts, though the overall pattern shows the prevalence will likely stabilize around 23.77 with a lower bound of about 10.89 (80% CI) and an upper bound of 36.66 (80% CI).

Summary of Findings

The analysis of the forecasted data, which spans from January 2024 to December 2027, reveals a stable trend with minimal variation in the predicted values. The forecast consistently hovers around 23.776 each month, suggesting that the variable being analyzed, likely malaria prevalence, is expected to remain steady over the next few years. The forecast shows no significant increase or decrease, indicating no strong seasonal or cyclical variations. However, the confidence intervals, particularly the 95% range, reveal some level of uncertainty, with lower bounds as low as 4.07 and upper bounds reaching up to 43.48. This implies that while the general trend is stable, external factors could cause deviations from the forecasted values. The relatively wide confidence intervals highlight the potential for variability, even though the central forecast remains constant. In conclusion, the analysis suggests that the current conditions influencing the variable are expected to maintain a steady trend, but there remains a need for continued monitoring due to the uncertainty reflected in the confidence intervals. For malaria prevalence, this would suggest that while rates may remain consistent, external factors could still lead to fluctuations, necessitating ongoing vigilance and health interventions.

Conclusion

This research was undertaken with the primary motive of developing an adequate model for forecasting future trends. It was found that ARIMA (1,0,0) fits the data best after comparison using the AIC and BIC criteria. The trend analysis indicated that malaria cases in Awka South are growing at a constant quadratic rate. The forecasted malaria prevalence from 2024 to 2027 indicates that cases will fluctuate within a range consistent with previous years, with the number of cases generally staying between 10 and 30 per month.

Recommendation

Given that the forecast predicts malaria prevalence will fluctuate within a range similar to prior years, this works recommends that consistent monitoring of malaria cases remains crucial. Ongoing data collection will enable tracking of trends and allow for a prompt response to any notable changes. Public education campaigns should be strengthened to increase awareness about malaria prevention, particularly during peak seasons when case numbers typically rise.