SARIMA Time Series Modeling for Forecasting Malaria Positivity Rate: A Case Study of Kisumu County, Kenya

Naomy Odero Onyuka[1,2], Lisette Schutte[3], Tobias F. Rinke de Wit[3,4], Fredrick Onyango[1]

[1]Department of Statistics and Actuarial Science, Maseno University, Kenya

[2]PharmAccess Foundation, Kisumu, Kenya

[3] PharmAccess Foundation, Amsterdam, the Netherlands

[4] Department Global Health, Amsterdam University Medical Center, the Netherlands

DOI: https://doi.org/10.51584/IJRIAS.2026.11080046

Received: 10 August 2026; Accepted: 15 August 2026; Published: 03 September 2026

ABSTRACT

Malaria transmission in Kisumu County, Kenya, follows a seasonal cycle that repeatedly strains health facility capacity, yet routine Kenya Health Information System (KHIS) surveillance data had not previously been applied to systematic short-term forecasting in the County. This study developed and validated a Seasonal Autoregressive Integrated Moving Average (SARIMA) model for forecasting monthly malaria positivity rate (MPR) in Kisumu County and extended the analysis to its seven constituent sub-counties: Kisumu Central, Kisumu East, Kisumu West, Nyando, Muhoroni, Nyakach, and Seme. Secondary KHIS data were extracted for 25 health facilities meeting minimum testing-volume and reporting-completeness criteria, covering January 2016 to December 2024. Malaria Positivity Rate was computed as confirmed malaria cases divided by tests conducted, multiplied by 100. Of 2,700 facility-month records, 232 (8.59%) were removed for logical inconsistency, yielding a complete 108-month county time series. The dataset was split into training (January 2016 to December 2023, n=96) and out-of-sample testing (January to December 2024, n=12) periods. Six candidate SARIMA models were compared using the corrected Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC). The optimal model was SARIMA(1,0,1)(0,1,1)[12] (AICc=541.74, BIC=550.95; ar1=0.531, ma1=0.111, sma1=-1.000), with residual adequacy confirmed by the Ljung-Box test ($p \approx 0.42$-0.49). County-level forecasts for 2024 achieved a Mean Absolute Error (MAE) of 2.27 percentage points, a Root Mean Squared Error (RMSE) of 3.01, and a Mean Absolute Percentage Error (MAPE) of 9.80%, within the 8-15% benchmark reported for comparable studies. The same model structure was optimal for all seven sub-counties, with sub-county MAPE ranging from 11.05% (Kisumu Central) to 41.54% (Seme), and forecast accuracy directly related to underlying data quality. These findings demonstrate that SARIMA modelling of routine surveillance data provides a practical, low-cost approach to short-term MPR forecasting at both county and sub-county levels, supporting spatially differentiated malaria preparedness planning without requiring additional data collection.

Keywords: SARIMA; Malaria Positivity Rate; Time Series Forecasting; Kisumu County; Sub-County Surveillance

INTRODUCTION

Malaria remains one of the world's most persistent public health challenges. In 2022 alone, over 249 million malaria cases were reported globally, with approximately 608,000 deaths, the majority occurring in Sub-Saharan Africa, continuing a burden trend little changed from the prior year [1,19]. Kenya continues to experience a significant malaria burden, and Kisumu County, located in western Kenya near the shores of Lake Victoria, consistently reports high malaria transmission rates. The county's warm temperatures, high humidity, lake-shore environment, and bimodal rainfall pattern create favorable conditions for mosquito breeding and sustained transmission [2].

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

Malaria is caused by Plasmodium parasites transmitted through the bites of infected female Anopheles mosquitoes, with Plasmodium falciparum predominating in Sub-Saharan Africa [5]. Kenya implements several interventions to reduce malaria transmission, including insecticide-treated nets, indoor residual spraying, and artemisinin-based combination therapies, coordinated through the national malaria control programme [3,6]. Despite these efforts, Kisumu County remains malaria endemic, with transmission occurring year-round and seasonal peaks placing considerable pressure on health facilities. A persistent challenge is the reactive nature of routine surveillance: case reporting depends on facility attendance after individuals fall ill, and preparedness measures are frequently implemented only after transmission has already begun to rise [4].

Time series forecasting offers a means of strengthening malaria preparedness by anticipating periods of increased transmission before they occur. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the ARIMA framework by incorporating seasonal components, making it well suited to diseases exhibiting recurring seasonal fluctuations [7]. SARIMA has been applied successfully to malaria surveillance data in other endemic settings [8,9,10], but evidence for its application to Malaria Positivity Rate (MPR), specifically within Kisumu County, has remained limited [4]. Kisumu County additionally comprises seven administratively and ecologically distinct sub-counties (Kisumu Central, Kisumu East, Kisumu West, Nyando, Muhoroni, Nyakach, and Seme), yet no prior study has produced sub-county specific SARIMA forecasts to support spatially differentiated preparedness planning.

This study therefore developed and validated a SARIMA model for forecasting monthly MPR in Kisumu County using KHIS routine surveillance data from 2016 to 2023, evaluated forecast performance against observed 2024 data, and extended the analysis to all seven sub-counties. The specific objectives were to: (i) analyze monthly MPR trends and seasonal patterns at county and sub-county level; (ii) assess stationarity and seasonality as prerequisites for SARIMA modelling; (iii) identify and estimate optimal SARIMA parameters at both levels; (iv) assess model adequacy using residual diagnostics; (v) generate 2024 forecasts and evaluate predictive accuracy using MAE, RMSE, and MAPE; and (vi) characterize sub-county level MPR variation and seasonal patterns.

METHODOLOGY

Research Design and Study Area

This study adopted a retrospective quantitative time series design, applying SARIMA modelling to secondary routine surveillance data following the Box-Jenkins methodology of model identification, parameter estimation, diagnostic checking, and forecasting [7,18]. The study was conducted in Kisumu County, western Kenya (approximately 0°06'S, 34°45'E), an area of approximately 2,085 km² with an estimated population of 1.15 million [2]. The county is malaria endemic year-round, with seasonal increases associated with its bimodal rainfall pattern, and comprises seven sub-counties: Kisumu Central, Kisumu East, Kisumu West, Nyando, Muhoroni, Nyakach, and Seme.

Data Source and Preparation

Monthly data on confirmed malaria cases and malaria tests conducted were extracted from the Kenya Health Information System (KHIS) for all registered facilities in Kisumu County, covering January 2016 to December 2024 (108 months). The initial extract covered over 100 facilities. Facilities were retained for analysis if they met two conditions: a mean monthly testing volume of at least 50 patients, to reduce instability in MPR estimates arising from small denominators, and reporting completeness exceeding 50 of 108 possible months. This yielded 25 facilities across all seven sub-counties, corresponding entirely to the sentinel facility network established under the Digital Innovations and Diagnostics for Infectious Diseases in Africa (DIDIDA) programme in Kisumu County, reflecting the strength of DIDIDA's existing surveillance infrastructure as a foundation for this analysis. Records with logical inconsistencies (tests conducted equal to zero, or confirmed cases exceeding tests conducted) were removed. Monthly confirmed cases and tests were summed across included facilities to construct county-level and sub-county level monthly time series.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

The outcome variable, monthly Malaria Positivity Rate (MPR), was computed as:

$$\text{MPR} = \frac{\text{Confirmed Malaria Cases}}{\text{Malaria Tests Conducted}} \times 100$$

MPR provides an indicator of transmission intensity that adjusts for variation in testing volume over time [1,4], and constitutes the Yt series modelled in this study.

SARIMA Model Specification

The SARIMA model extends ARIMA by incorporating seasonal autoregressive (P), seasonal differencing (D), and seasonal moving average (Q) components operating at a fixed seasonal period s, in addition to the non-seasonal autoregressive (p), differencing (d), and moving average (q) components [7]. For monthly data with annual seasonality, s=12. A SARIMA(p,d,q)(P,D,Q)12 model is written in backshift operator notation as:

$$\Phi_P(B^{12})\ \varphi_p(B)\ (1 - B^{12})^D\ (1 - B)^d\ Y_t = \Theta_Q(B^{12})\ \theta_q(B)\ \varepsilon_t$$

where B is the backshift operator (BYt = Yt−1) and B12 is the seasonal backshift operator (B12Yt = Yt−12). Model identification used the Augmented Dickey-Fuller (ADF) test for stationarity and inspection of autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Candidate models were fitted by maximum likelihood and compared using AICc and Bayesian Information Criterion (BIC), with the lowest-scoring model selected. Residual adequacy was assessed using the Ljung-Box test, under the null hypothesis that residuals are white noise. The same five-stage procedure (trend/seasonality analysis, stationarity testing, model identification, diagnostic checking, and forecast evaluation) was applied independently to the county series and to each of the seven sub-county series.

Forecast Generation and Accuracy Evaluation

The training period (January 2016 to December 2023, n=96) was used for model fitting; the testing period (January to December 2024, n=12) was withheld entirely from model development and used for out-of-sample evaluation. Forecast accuracy was assessed using three metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{100}{n}\ \Sigma\ |Yt - \hat{Y}t|\ /\ Yt$$

Prediction intervals were computed at the 80% and 95% confidence levels (α=0.20 and α=0.05 respectively), consistent with standard practice in time series forecasting [13]. All analyses were conducted in R (version 4.x) using the forecast, tseries, and ggplot2 packages.

Ethical Considerations

This study used only aggregate, facility-level secondary surveillance data extracted from KHIS. No patient-level or personally identifiable information was accessed, collected, or reported at any stage of the analysis. Authorization to access and use the KHIS data for this research was obtained from the Kisumu County Department of Health.

RESULTS

Data Preparation

The cleaned dataset contained 2,700 facility-month records from 25 health facilities, covering January 2016 to December 2024. A total of 232 records (8.59%) were removed for logical inconsistency, defined as records where the number of tests conducted was zero or the number of confirmed cases exceeded the number of tests conducted, both of which indicate data entry or reporting errors rather than valid surveillance values. This left

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

2,468 valid records and a complete 108-month county time series with no missing months, satisfying the continuity requirement for SARIMA modelling. All 25 included facilities submitted a monthly record in every one of the 108 study months; no facility permanently exited the dataset at any point in the study period. However, individual facility-months were excluded from the aggregate where records failed validity checks, meaning the specific facilities contributing valid data to a given month's aggregate varied modestly according to record-level data quality rather than facility turnover. Table Ia reports monthly participation, the testing volume range, and removal rates for each of the 25 individual facilities (coded to preserve confidentiality), and Table Ib presents the same information aggregated to sub-county level.

Table Ia. Facility-Level Monthly Participation, Testing Volume Range, and Record Removal Rates (n=25 facilities, 108 months each).

Facility	Sub-County	Months Reported	Records Removed	% Removed	Monthly Test Range (min–max)
X1	Kisumu Central	108/108	0	0.00%	948-4234
X2	Kisumu Central	108/108	0	0.00%	125-2802
X3	Kisumu Central	108/108	0	0.00%	181-2545
X4	Kisumu Central	108/108	0	0.00%	159-2211
X5	Kisumu Central	108/108	0	0.00%	53-856
X6	Kisumu Central	108/108	0	0.00%	239-647
X7	Kisumu East	108/108	7	6.48%	132-1068
X8	Kisumu East	108/108	3	2.78%	225-1763
X9	Kisumu East	108/108	2	1.85%	51-1285
X10	Kisumu East	108/108	1	0.93%	62-795
X11	Kisumu West	108/108	9	8.33%	56-1138
X12	Kisumu West	108/108	6	5.56%	57-1131
X13	Kisumu West	108/108	5	4.63%	104-1524
X14	Muhoroni	108/108	6	5.56%	56-639
X15	Muhoroni	108/108	4	3.70%	62-1083
X16	Nyakach	108/108	5	4.63%	65-2860
X17	Nyakach	108/108	0	0.00%	164-1309
X18	Nyando	108/108	9	8.33%	62-803
X19	Muhoroni	108/108	47	43.52%	57-897
X20	Kisumu West	108/108	22	20.37%	58-961

X21	Nyando	108/108	0	0.00%	381-2170
X22	Seme	108/108	4	3.70%	144-987
X23	Seme	108/108	42	38.89%	72-568
X24	Nyando	108/108	37	34.26%	55-920
X25	Seme	108/108	23	21.30%	62-1023

Table Ib. Sub-County Monthly Participation, Testing Volume Range, and Record Removal Rates.

Sub-County	**Facilities**	**Total Records**	**Records Removed**	**% Removed**	**Monthly Test Range (min-max)**
Kisumu Central	6	648	0	0.00%	53-4234
Kisumu East	4	432	13	3.01%	51-1763
Kisumu West	4	432	42	9.72%	56-1138
Muhoroni	3	324	57	17.59%	56-1083
Nyakach	2	216	5	2.31%	65-2860
Nyando	3	324	46	14.20%	55-2170
Seme	3	324	69	21.30%	62-1023
County (total)	25	2700	232	8.59%	51-4234

To assess whether inclusion of lower-quality facilities affected forecast accuracy, a sensitivity analysis was conducted restricting the county-level aggregate to the 20 facilities with individual removal rates below 10% (excluding Facility X19, X20, X23, X24, and X25). The resulting restricted-facility model retained the same optimal structure, SARIMA(1,0,1)(0,1,1)[12] (AICc=547.68, BIC=556.89), but achieved a test-period MAPE of 11.24%, higher than the 9.80% achieved using all 25 facilities. This indicates that excluding lower-quality facilities did not improve forecast accuracy, and by extension that the original facility inclusion criteria (Section II.B) provided a reasonable balance between data quality and the statistical power gained from aggregating a larger number of reporting facilities.

County-Level MPR Trends and Seasonality

County MPR ranged from 9.36% to 46.6% over the study period (Fig. 1), with a prominent spike around 2020 plausibly reflecting disruption to routine health service utilization during that period, due to COVID-19. The training series exhibited a recurring annual pattern consistent with Kisumu County's bimodal rainfall seasonality. The Augmented Dickey-Fuller test supported stationarity of the county training series ($p<0.01$), and the ACF showed significant autocorrelation at lag 1 and at seasonal lags (multiples of 12), motivating a seasonal model with s=12.

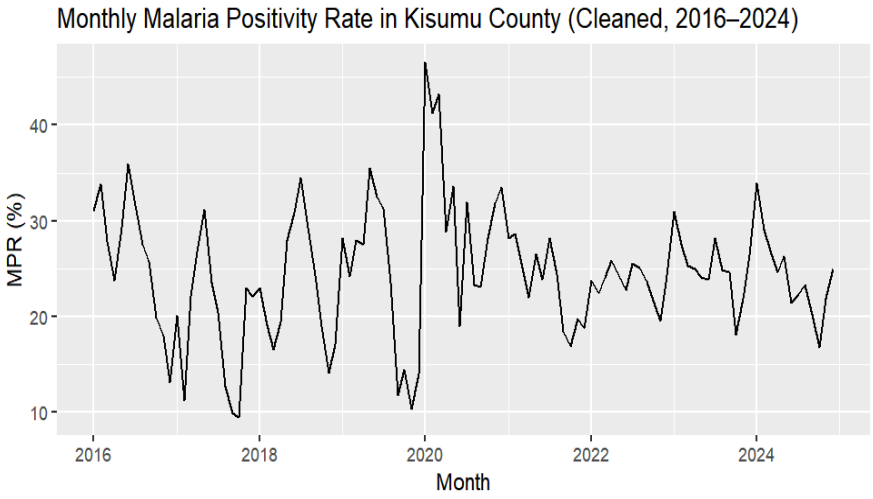

Fig. 1. Monthly malaria positivity rate, Kisumu County (cleaned series, January 2016 to December 2024).

County Model Selection and Diagnostics

Six candidate SARIMA models were compared using AICc and BIC (Table II). Non-seasonal and seasonal differencing requirements were assessed prior to fitting: the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests both indicated no requirement for non-seasonal differencing (d=0), while the Canova-Hansen test (joint statistic=1.555, p=0.484) and the OCSB test (statistic=−6.27, critical value=−1.80 at the 5% level) indicated no requirement for seasonal differencing (D=0). Candidate SARIMA models were subsequently evaluated using information criteria and out-of-sample forecast performance rather than the unit-root tests alone. SARIMA(1,0,1)(0,1,1)[12] achieved the lowest AICc (541.74) and BIC (550.95), with a margin exceeding 58 AICc units over the nearest alternative, and the estimated coefficients were ar1=0.531, ma1=0.111, and sma1=−1.000. The selected model therefore retained d=0 but incorporated one seasonal difference (D=1), reflecting the model-selection results rather than the seasonal unit-root tests alone. This is notable given that two of the five alternative candidates already evaluated in Table II were themselves D=0 specifications, each scoring substantially worse than the selected model (AICc higher by 66-70 units; Table III), reinforcing that the seasonally differenced specification was empirically superior despite the formal unit-root test result.

As a further robustness check on the boundary sma1 estimate, the model was re-estimated with sma1 fixed at -0.90, bounding it away from the boundary, while ar1 and ma1 remained freely estimated. This constrained specification achieved a marginally lower AICc (539.86 versus 541.74), lower BIC (546.85 versus 550.95), and marginally lower test-period MAPE (9.57% versus 9.80%) than the freely-estimated model. Given the small magnitude of these differences, the freely-estimated model (sma1=-1.000) was retained as the primary reported result, since the two specifications are not meaningfully distinguishable in either fit or accuracy; the constrained result nonetheless confirms that the boundary estimate does not reflect a fitting failure, as an alternative specification bounded away from the boundary performs equivalently well.

Table II. Candidate County SARIMA Model Comparison, Training Period 2016-2023 (n=96). Ranked by AICc

Candidate Model	**AICc**	**BIC**
SARIMA(1,0,1)(0,1,1)$_{12}$	541.74	550.95
ARIMA(3,0,0)	599.88	612.03
SARIMA(2,0,1)(1,0,1)$_{12}$	605.88	620.32
SARIMA(1,0,1)(1,0,1)$_{12}$	607.71	619.86

SARIMA(1,0,0)(1,0,0)$_{12}$	609.28	619.44
SARIMA(0,0,1)(1,0,1)$_{12}$	617.6	629.75

Table III. Sensitivity Comparison: Seasonally Differenced (D=1) vs. Non-Differenced (D=0) Model Alternatives, County Training Period 2016-2023, Evaluated on 2024 Test Data. AICc and BIC for SARIMA(1,0,1)(1,0,1)[12] and SARIMA(2,0,1)(1,0,1)[12] are reproduced from Table II, where these models were also evaluated as candidates; test-period MAPE has been added for all models here.

Model	**AICc**	**BIC**	**Test MAPE (%)**
SARIMA(1,0,1)(0,1,1)[12] — D=1,selected	541.74	550.95	9.80
SARIMA(1,0,1)(1,0,1)[12] — D=0	607.71	619.86	9.76
SARIMA(2,0,1)(1,0,1)[12] — D=0	605.88	620.32	8.94
SARIMA(1,0,1)(1,0,0)[12] — D=0	610.94	623.1	12.93
SARIMA(1,0,1)(0,0,1)[12] — D=0	611.53	623.68	13.12

The optimizer converged successfully (convergence code = 0), with residual variance (σ^2) = 25.64. Inspection of the characteristic autoregressive and moving-average polynomial roots confirmed the model was stationary and invertible (root moduli > 1); the seasonal moving-average root had modulus approximately 1.00, indicating the annual seasonal cycle is close to a unit-root process, consistent with the boundary sma1 estimate discussed above. Residual diagnostics confirmed model adequacy in most respects: the residual time plot fluctuated around zero with no visible trend, the residual ACF showed no significant remaining autocorrelation, and the Ljung-Box test, conducted at lag 24 with degrees of freedom equal to lag minus the number of estimated model parameters, was non-significant ($p \approx 0.42$-0.49), confirming the residuals approximated white noise. An ARCH-LM test for heteroskedasticity at lag 12 was likewise non-significant (LM statistic=9.21, p=0.68), indicating no evidence of time-varying residual variance. However, formal normality tests rejected the null hypothesis of normally distributed residuals (Jarque-Bera: statistic=42.88, $p<0.001$; Shapiro-Wilk: W=0.941, $p<0.001$), despite the residual histogram appearing approximately symmetric on visual inspection. This departure from normality does not affect the validity of the point forecasts, which depend on the residual independence confirmed by the Ljung-Box test, but it means the nominal coverage of the 80% and 95% prediction intervals, which assume Gaussian residuals, should be interpreted with some caution (Section V).

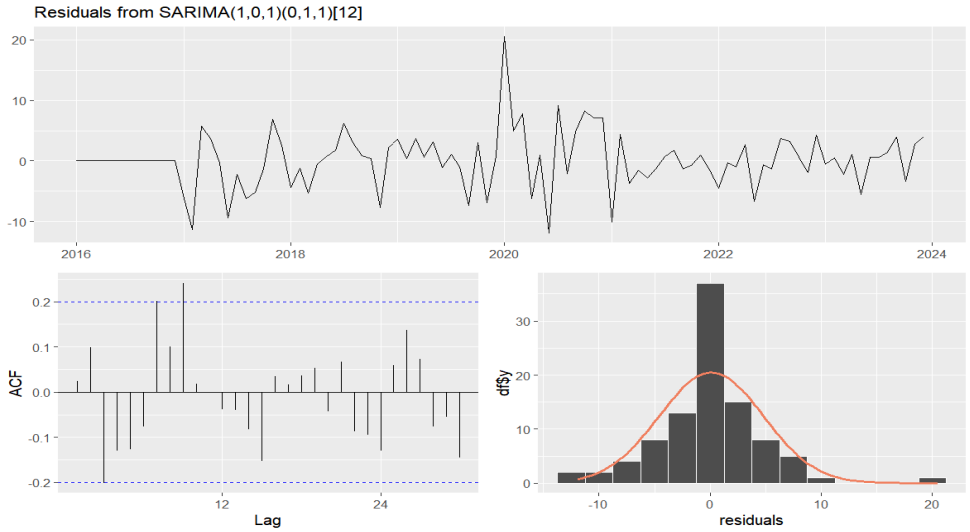

Fig. 2. County residual diagnostics for SARIMA(1,0,1)(0,1,1)[12]: residual time plot (top), residual ACF (bottom left), and histogram of residuals with normal density overlay (bottom right).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

County-Level Forecast Evaluation

Using the selected model, forecasts were generated for January to December 2024 and evaluated against observed values (Table IV). Forecasts tracked observed 2024 MPR closely overall; the largest deviations occurred in June and July, where the model overestimated observed MPR. Most observed values fell within both the 80% and 95% prediction intervals, indicating appropriate uncertainty calibration. To contextualise this performance, a seasonal-naive benchmark (repeating the observed value from the same calendar month one year earlier) was also evaluated on the same test period, achieving MAE=2.18, RMSE=2.70, and MAPE=9.30% (Table IV). The SARIMA model's point-forecast accuracy was therefore comparable to, rather than substantially better than, this simple benchmark. Its principal advantages over the naive approach are the provision of formal prediction intervals, explicit residual diagnostics confirming model adequacy, and a coherent modelling framework that extends to sub-county series with varying data characteristics (Section III.E), none of which the naive method provides. As an additional check on forecast robustness, a rolling-origin (walk-forward) validation was conducted, re-estimating the model and forecasting one month ahead sequentially across the 2024 test period, rather than fixing the model at the end of the training period and forecasting all twelve months at once. This produced a MAPE of 8.50%, comparable to, and marginally better than, the fixed-origin evaluation (9.80%), indicating that the model's forecast accuracy was not dependent on the specific point at which it was fitted and holds under a more rigorous, sequential evaluation scheme. Consistent with the fixed-origin result, the largest rolling-origin errors occurred in June, July, and August (15.0-16.0% absolute percentage error).

Table IV. County-Level Forecast Accuracy Metrics, January-December 2024, Including Seasonal-Naive Benchmark Comparison.

Metric	**Value**
MAE (percentage points)	2.27
RMSE	3.01
MAPE (%)	9.8
Published benchmark range (MAPE)	8-15%
Seasonal-naive MAE (percentage points)	2.18
Seasonal-naive RMSE	2.7
Seasonal-naive MAPE (%)	9.3

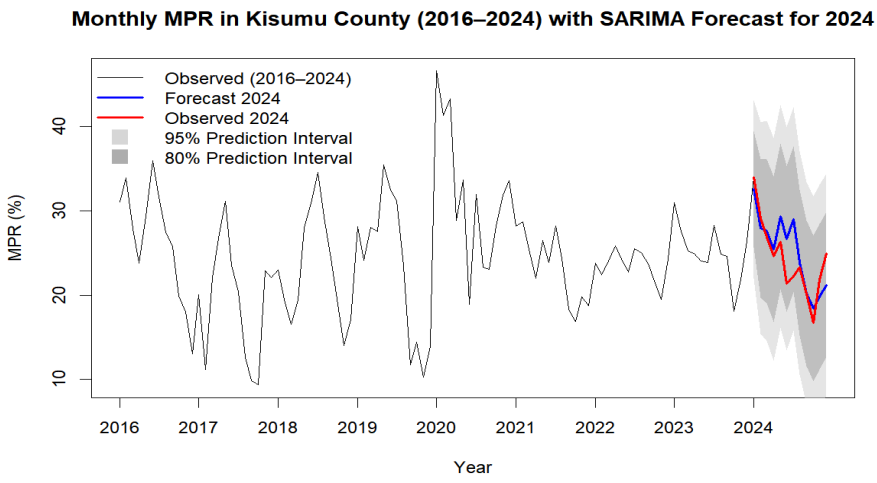

Fig. 3a. Monthly MPR in Kisumu County (2016-2024) with SARIMA(1,0,1)(0,1,1)[12] forecast for 2024, 80% and 95% prediction intervals.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

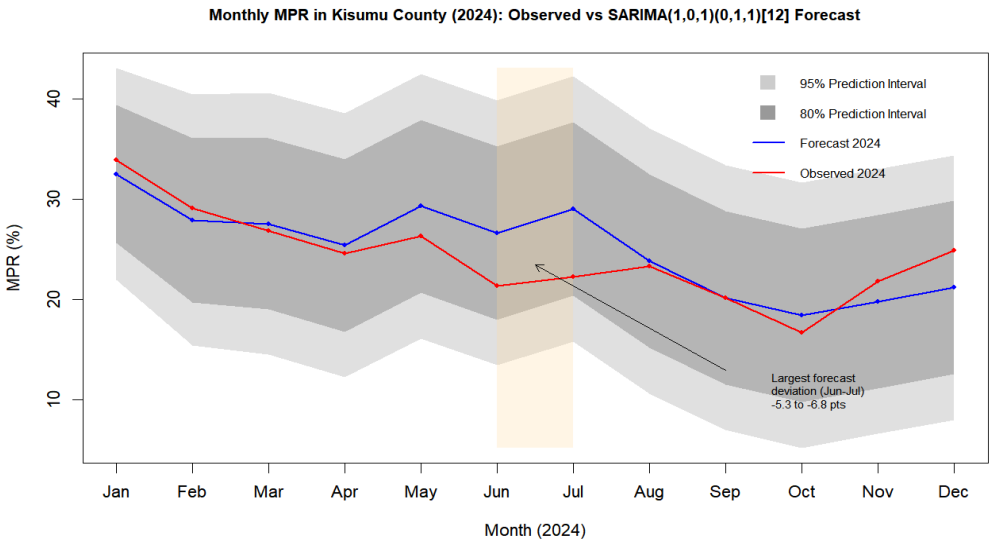

Fig. 3b. Monthly MPR in Kisumu County (2024): observed vs. SARIMA(1,0,1)(0,1,1)[12] forecast, with 80% and 95% prediction intervals. June-July shows the largest deviations (-5.32 and -6.77 percentage points).

Sub-County Level Analysis

The same SARIMA(1,0,1)(0,1,1)[12] structure was independently optimal for all seven sub-counties, with AICc margins over the nearest alternative ranging from approximately 58 to 81 units. Estimated ar1 coefficients ranged from 0.440 (Kisumu Central) to 0.754 (Nyakach), reflecting differences in short-term persistence across sub-counties. Five of seven sub-counties (Nyando, Kisumu West, Kisumu East, Kisumu Central, and Muhoroni) were stationary by the ADF test ($p<0.01$); Nyakach ($p=0.080$) and Seme ($p=0.078$) were borderline, with seasonal differencing ($D=1$) within SARIMA adequately addressing this. Residual diagnostics confirmed adequacy for six of seven sub-counties; Kisumu East showed a marginal Ljung-Box result at lag 24 ($p=0.039$), interpreted with additional caution.

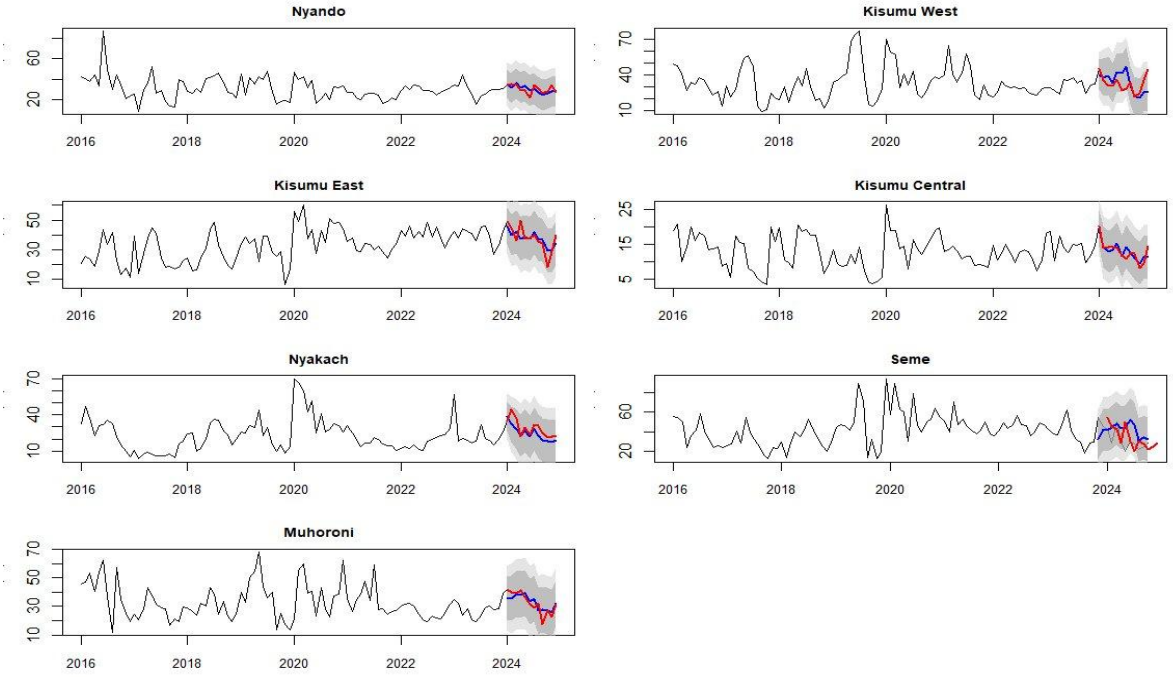

Fig. 4. Combined sub-county MPR time series (2016-2024) with SARIMA(1,0,1)(0,1,1)[12] forecasts for 2024. Blue = forecast; red = observed 2024.

Table V summarises coefficients, diagnostics, and forecast accuracy across all seven sub-counties and the county-level reference. Sub-county MAPE ranged from 11.05% (Kisumu Central) to 41.54% (Seme). Kisumu

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

Central achieved the best sub-county forecast accuracy, consistent with its urban profile and zero invalid records removed during cleaning. Seme's comparatively poor accuracy coincided with its highest data-removal rate (21.30%) and highest residual variance (σ^2=212.5) of any sub-county series, indicating that forecast reliability at sub-county level tracked underlying data quality rather than deficiency in the model structure itself.

Table V. Sub-County SARIMA Models: Coefficients, Residual Diagnostics, and Forecast Accuracy, with County-Level Reference. *Marginal Ljung-Box result; interpreted with caution.

Sub-County	AICc	BIC	ar1	ma1	sma1	LB p (lag 24)	MAPE (%)
Nyando	654.58	663.8	0.502	-0.155	-0.749	0.613	11.39
Kisumu West	645.38	654.6	0.451	0.278	-1	0.301	24.36
Kisumu East	635.56	644.78	0.62	-0.113	-0.719	0.039*	13.68
Kisumu Central	497.02	506.24	0.44	0.09	-1	0.45	11.05
Nyakach	641.09	650.3	0.754	-0.046	-1	0.303	16.48
Seme	702.4	711.5	0.65	-0.285	-0.999	0.178	41.54
Muhoroni	668.74	677.95	0.493	-0.103	-0.884	0.39	13.25
County	541.74	550.95	0.531	0.111	-1	0.42-0.49	9.8

DISCUSSION

The results presented in this study demonstrate that SARIMA modelling applied to monthly MPR derived from routine KHIS data produced reliable short-term forecasts at both county and sub-county levels in Kisumu County, Kenya. The county-level MAPE of 9.80% falls within the 8-15% benchmark range reported for comparable SARIMA and ARIMA malaria forecasting applications in Ghana, western Kenya, Zambia, Pakistan, Ethiopia, and Mumbai [8,9,10,14,15,16], and is consistent with the broader literature applying time series methods to routine epidemiological surveillance data [17].

The consistent selection of SARIMA(1,0,1)(0,1,1)[12] as the optimal model across all seven sub-counties as well as at the county level, despite differing ecological profiles, facility counts, MPR ranges, and stationarity characteristics, provides strong evidence for the robustness of the seasonal differencing and seasonal moving average structure in capturing annual malaria seasonality throughout Kisumu County. This consistency is plausibly attributable to a shared ecological driver: the county's bimodal rainfall pattern and Lake Victoria microclimate operate across all sub-counties, consistent with the established relationship between climate and malaria transmission risk [12], producing a common seasonal autocorrelation signature that the same model structure captures efficiently with only three estimated parameters [7,11].

Kisumu Central achieved the best sub-county forecast accuracy (MAPE=11.05%), consistent with its lower mean MPR, zero invalid records removed during cleaning, and stable, urban epidemiological profile. In contrast, Seme's comparatively poor accuracy (MAPE=41.54%) coincided with a substantially higher data removal rate and the highest residual variance of any sub-county model. This pattern reinforces that sub-county forecast reliability in this study was primarily a function of underlying surveillance data quality rather than model inadequacy, a distinction with direct implications for where investment in data quality improvement would yield the greatest returns.

These sub-county forecasts can support differentiated malaria preparedness planning across Kisumu County's administrative hierarchy, enabling earlier and more intensive commodity pre-positioning and facility readiness in sub-counties with higher predicted burden or greater historical volatility, while allowing comparatively lighter anticipatory response where MPR has proven stable and well-forecast.

CONCLUSION AND RECOMMENDATIONS

Routine KHIS surveillance data, subjected to appropriate inclusion criteria and validity filtering, provided a reliable basis for constructing complete monthly MPR time series suitable for SARIMA forecasting at both county and sub-county levels in Kisumu County. SARIMA(1,0,1)(0,1,1)[12] was the optimal model at every level of analysis examined, and county-level forecasts met the published accuracy benchmark. Sub-county forecasting was feasible for the majority of sub-counties, with accuracy directly related to underlying data quality rather than model structure.

Kisumu County health surveillance teams should consider integrating routine monthly MPR forecasting into malaria preparedness planning, implementable using existing KHIS data and open-source statistical software without additional data collection. Forecast outputs should be reviewed alongside contextual information, including rainfall patterns and health-service utilization trends, since a univariate model cannot anticipate discrete external shocks such as the 2020 disruption observed in this study, most likely caused by COVID-19. Targeted data-quality strengthening in Seme, Muhoroni, and Nyando, the sub-counties with the highest record-removal rates, would directly improve future sub-county forecast reliability. Future research should evaluate SARIMAX models incorporating rainfall and temperature covariates, compare SARIMA against alternative approaches such as Prophet or long short-term memory networks, and assess transferability of this framework to other Lake Victoria endemic counties including Homa Bay, Siaya, and Migori.

This study is limited by its reliance on routine surveillance data subject to residual reporting inconsistency, its univariate specification excluding climate and intervention covariates, and evaluation against a single calendar year of out-of-sample data, albeit assessed under both fixed-origin and rolling-origin validation schemes. Results for Seme and Kisumu West in particular should be interpreted with the documented data-quality limitations in view. A sensitivity analysis restricting the county model to the 20 facilities with individual removal rates below 10% did not improve forecast accuracy relative to the full 25-facility model (Section III.A), suggesting the original inclusion criteria were reasonable. The residual normality assumption underlying the reported 80% and 95% prediction intervals was formally rejected (Section III.C), so interval coverage should be interpreted with some caution alongside the point forecasts, which are less affected by this violation. Because MPR is a bounded, denominator-dependent proportion, future work should also evaluate logit-transformed or count-based (binomial) formulations using confirmed cases and testing volume directly, which may better respect the statistical properties of the outcome than the untransformed rate modelled here.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

The data supporting this study were derived from the Kenya Health Information System (KHIS) and are subject to county-level data-sharing authorization. Aggregated data is available from the county upon reasonable request.

Acknowledgements

The authors thank the Kisumu County Department of Health for facilitating access to routine surveillance data used in this study. This study was partly supported by the EU through the DIDIDA project (https://dididahealth.eu/).

REFERENCES

1. World Health Organization (WHO). World Malaria Report 2022. Geneva: WHO; 2022.
2. Kenya Malaria Indicator Survey (KMIS). Kenya Malaria Indicator Survey 2022. Nairobi: Ministry of Health, Kenya; 2022.
3. Ministry of Health (MOH). Kenya Malaria Strategy 2019-2023: Updated for 2023. Nairobi: Government of Kenya; 2022.
4. Oduor, C., Onyango, C., & Musyoka, P. (2023). Evaluating ARIMA models for malaria forecasting in Kenya's lake endemic region. African Health Sciences, 23(1), 99-108. https://doi.org/10.4314/ahs.v23i1.12
5. Centers for Disease Control and Prevention (CDC). Malaria biology and life cycle. Atlanta: CDC; 2023.
6. Ministry of Health (MOH). National Malaria Control Programme Annual Report 2022. Nairobi: Government of Kenya; 2022.
7. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control (5th ed.). Hoboken: Wiley.
8. Asare, C., & Asiedu, E. (2020). Comparative analysis of ARIMA and SARIMA models for malaria forecasting in Ghana. Malaria Journal, 19(1), 256. https://doi.org/10.1186/s12936-020-03411-x
9. Wanjala, C. L., & Mwangangi, J. M. (2022). Application of ARIMA models in malaria trend prediction in western Kenya. East African Medical Journal, 99(4), 278-285.
10. Mwale, F. D., Chirwa, G. C., & Siziya, S. (2017). Time series analysis of malaria in Zambia using ARIMA models. The Pan African Medical Journal, 28, 154. https://doi.org/10.11604/pamj.2017.28.154.13510
11. Teklehaimanot, H. D., Schwartz, J., Teklehaimanot, A., & Lipsitch, M. (2004). Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia. Malaria Journal, 3(1), 41. https://doi.org/10.1186/1475-2875-3-41
12. Paaijmans, K. P., Read, A. F., & Thomas, M. B. (2009). Understanding the link between malaria risk and climate. Proceedings of the National Academy of Sciences, 106(33), 13844-13849. https://doi.org/10.1073/pnas.0903423106
13. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). Melbourne: OTexts.
14. Khan, M. A., Ahmad, N., & Mehmood, M. (2020). Time series modeling and forecasting of malaria in Punjab, Pakistan. Infectious Diseases of Poverty, 9(1), 65. https://doi.org/10.1186/s40249-020-00679-4
15. Tulu, T., Raj, S., & Atnafu, A. (2020). Modeling and forecasting malaria cases using ARIMA models in Jimma Zone, Ethiopia. Journal of Parasitology Research, 2020, 1249298. https://doi.org/10.1155/2020/1249298
16. Ghosh, S., Patil, R. R., & Singh, N. (2019). Forecasting malaria incidence using time series models in Mumbai, India. Tropical Medicine and Health, 47(1), 50. https://doi.org/10.1186/s41182-019-0178-0
17. Zhang, X., Zhang, T., Young, A. A., & Li, X. (2013). Applications and comparisons of four time series models in epidemiological surveillance data. PLoS ONE, 8(2), e56725. https://doi.org/10.1371/journal.pone.0056725
18. Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control (1st ed.). San Francisco: Holden-Day.
19. World Health Organization (WHO). World Malaria Report 2021. Geneva: WHO; 2021.