European Journal of Mathematics and Statistics
Vol 5 | Issue 6 | December 2024
ISSN 2736-5484

RESEARCH ARTICLE

Time Series Modeling of National Hospital Insurance Fund Coverage in Kenya

Hellen Wawira Ndwiga*, Dennis Muriithi, and Daniel Mwangi

ABSTRACT

National Hospital Insurance Fund (NHIF) is a state-owned organization that was established in 1966 with the goal of providing Kenyans with social health insurance that is easily accessible, affordable, long-lasting, and of high quality. Only 24% of Kenyans have access to NHIF, this may affect the implementation and outcome of Universal Health Coverage (UHC). NHIF mandate can only be achieved if the whole population is under an insurance health cover. Understanding patterns, trends and forecasting of NHIF population Coverage using time series analysis would help in policy formulation and planning for proper implementation of UHC in Kenya. The main objective of the study was to model and forecast Kenya's NHIF population coverage using Seasonal Autoregressive Integrated Moving Average model. Time series research design was used as it involved data that was measured at regular intervals over a significant number of observations. This design followed the Box-Jenkins Seasonal Autoregressive Integrated Moving Average (SARIMA) model. Simulated time series data on NHIF enrollment for the period 1998–2023 was used for this study. R and R-studio was used in the statistical analysis of the data. The model which exhibited the least Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values was picked by fitting the SARIMA model. Finally, forecasting of the data after following the three Box-Jenkins methodologies, that is, model identification, estimation of parameters and diagnostic check is done. Having an AIC value of 2265.00 and BIC value of 2279.66 SARIMA (1,1,3) $(0,1,1)_4$ model fitted the data well. This model passed residual normality test and the forecasting evaluation statistics shows the errors as RMSE = −1263.392, MAPE = 5.872978 and MAE = 11197.31 The 3-year ahead forecasts showed that the enrollment had overall increasing trend. However, moving further into the future forecast the confidence intervals tend to widen. This indicated that the model's predictions became less certain with time. The SARIMA model proved to be a suitable approach for capturing the underlying patterns in the NHIF enrollment data, providing reasonable forecasts. The findings of the study would lead to robust sensitization by both national and county government and all other stakeholders on the importance of National Hospital Insurance Fund coverage which would lead to increased enrollment from 24% to almost 100% through the Social Health Authority (SHA). This in turn would lead to attainment of the Universal Health Coverage an objective of the third Sustainable Development Goal that stipulates healthy lives and promotes well-being for all at all ages.

Keywords: Coverage, modelling, NHIF, time series.

Submitted: October 16, 2024

Published: December 30, 2024

10.24018/ejmath.2024.5.6.382

Faculty of Physical Sciences, Engineering, and Technology, Tharaka University, Kenya.

****Corresponding Author:***
e-mail: hellenak4@gmail.com

1. Introduction

National Hospital Insurance Fund (NHIF) is one of Africa's first health insurance schemes. It was founded in 1966 to offer required health insurance to people working in formal employment, and in 1998, the NHIF reach was broadened in order to encompass those working under informal employment. Only 24% of Kenyans have access to NHIF, as reported in [1]. In urban areas compared to rural areas, a higher percentage of people have health insurance of some kind. The study also discovered that as wealth rises, so does the availability of health insurance. From 2.7 million in 2010 to 6.6 million in 2017 more Kenyans (primary members plus beneficiaries) were enrolled in the NHIF scheme [2]. The Universal Health Coverage (UHC) provides access to the whole range of basic health care, due to its affordability and advantages. NHIF coverage is crucial in this situation for the people to be able to obtain health services in Kenya. Time series is a group of data points that essentially consists of measurements taken in succession over a course of time. Data points obtained within a specific period of time could contain an internal structure (such as autocorrelation, trend, or seasonal fluctuation) that had to be considered. Time series analysis took this potential into account. Time series analysis was necessary to reveal hidden trends and seasonal patterns [3].

The SARIMA model is highly adaptable and can support various data trend and seasonality patterns [4]. Seasonality in a given time series is a predictable sequence of recurring patterns across time. The Box and Jenkins' Seasonal Autoregressive Integrated Moving Average (SARIMA) is one of the popular models that have been used in time series modelling to capture the seasonality aspect of time series data [5]. A system for projecting the Indian healthcare industry index was developed in India. Data spanning January 2010 through December 2021 was gathered. Values in the component parts provided some key observations into the pre- and peri-pandemic behavior of the industry. Al-Haque *et al*. [6] came up with a SARIMA model in response to the seasonal demand for medical services from traveling patients to predict the number of patients at the clinic with various needs. The RMSE of the SARIMA model was shown to be substantially less when compared to historical average forecasts.

1.1. Method

The study was carried out in Tharaka University. Simulated time series data on NHIF enrollment for the period 1998–2023 was used for this study. This statistic covered 100 quarters in a period of 25 years. This suggested that the study would focus on Kenya's NHIF coverage with 100 observations, which complied with the general guideline that there should be observations greater than fifty when using the Box-Jenkins method for time series forecasting. The study applied time series research design which is one of the four types of longitudinal research designs. Longitudinal research design uses continuous or repeated data over a prolonged period of time. Time series design was typically used in research projects that analyze variables using time series data. R and RStudio statistical program were used in data and statistical analysis. The data was analyzed using time series analysis because it required absolute forecast values and typically yielded superior results. The SARIMA process is a common mathematical forecasting tool.

1.2. Time Series Modelling

Time series approaches were applied in a variety of contexts, particularly in research using time-dependent data. Box and Jenkins introduced time series models for the first time in 1960; hence, the term box-Jenkins Model [4]. The original Box and Jenkins methodology had three steps: model selection, parameter estimation, and model checking. More recently, an initial stage of preparing data and the last stage of applying the model, forecasting, have been added.

1.3. Sarima Models Theory

a) Autoregressive (AR)

According to the AR model a realization at time t is a sequence of the p prior happenings with addition to the noise term. In terms of the lagged observations, autoregressive models reflect the provided time series x_t. In research, Autoregressive process of order p is expressed as an AR (p) process, as shown:

$$x_t = \mu + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots + \beta_p x_{t-p} + e_t$$

b) Moving Average (MA)

In time series analysis, one common method in modelling univariate time series is the MA model. A linear combination of the present value and several previous values of a random term is how the MA model generates the outcome variable. A moving average process of order q, represented as an MA (q) process, is defined as:

The general form of moving average or MA (q) process is considered as:

$$x_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} + \varepsilon_t$$

c) Autoregressive Moving Average (ARMA)

The AR and MA processes are both combined into the ARMA (p, q) processes in the ARMA Model [10]. A variety of stationary time series could be modeled using the ARMA technique. It follows from

TABLE I: Descriptive Statistics of NHIF Data

Statistics	Value
N	100
Mean	172690.4
Standard deviation	53970.57
Median	166410
Trimmed	170479.2
Mad	59591.62
Min	74799
Max	324000
Range	249201
Skew	0.35
Kurtosis	−0.45
Standard error	5292.25

this then a broad time series x_{t} can be expressed as one that is linear of past x_{t} values and errors, e_{t}. Result of combining the AR and MA models is the ARMA *(p, q)* process, expressed as;

$$x_t - \varnothing_1 x_{t-1} - \varnothing_2 x_{t-2} - \ldots - \varnothing_p x_{t-p} = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \theta_q e_{t-q}$$

d) Autoregressive Integrated Moving Average (ARIMA)

A given process is described by the ARIMA Model as a direct product of its own lagged observations of order p and lagged error terms of order q. These elements each have a clear definition as a model parameter. The ARIMA model is a model with a specified order, where p, d, and q are integers that exceed or equal to zero [7]. The parameter, p, describes the quantity of AR lags, d, the integration order ensures data stationarity, and q, the number of MA lags. The model works well with non-stationary data, when certain series must be differentiated in order to become stationary. According to Box and Jenkins, before the model is fitted, stationarity needs to be confirmed. If a time series' mean and variance remain constant across time, it is stationary.

e) Seasonal Autoregressive Moving Average (SARIMA)

Elements in a broad SARIMA model presented by Box and Jenkins include: The necessary number of differencing, auto regressive and moving average components. The model's three different categories of parameters are: the auto regressive parameter, the moving average parameter (q) and the number of differencing required (d). They are summed as ARIMA *(p, d, q)*.

The seasonal ARIMA (P, D, Q) parameters are recognized for the data. These are; seasonal auto regressive (P), seasonal moving average (Q) and seasonal differencing (D) components. These parameters were produced after differencing the series once at lag 1 and then once at lag 4. The SARIMA model extends the ARIMA model to fit seasonal time series data [8]. Seasonality in a given time series is a predictable sequence of recurring patterns across time.

The SARIMA (p, d, q) (P, D, Q)$_{\mathrm{m}}$ model equation is illustrated as;

$$\Phi_p(B)\varphi_p\left(B^m\right)(1-B)^d(1-B^m)^D X_t = \Theta_Q(B)\theta_q(B^m)\varepsilon_t$$

2. Results and Discussion

a) Descriptive Analysis

A descriptive analysis was done on simulated data on NHIF enrollment and the results are presented in Table I.

The minimum and maximum values of the data are 74,799 and 324,000 respectively with a mean and median of 172,690.4 and 166,410 respectively. The skew is 0.35, suggests a slight right-tailed distribution. The kurtosis is −0.45, suggesting a slightly flatter distribution than a normal distribution. With a standard deviation of 53970.57, the data points are quite spread out from the mean, indicating a significant amount of variability in the dataset. The standard error of 5292.25 indicates that the sample mean is a relatively precise estimate of the population mean.

b) Time Series Plot

In modelling the NHIF enrollment time series, the data set for the period starting 1998 through December 2023 was used to plot the time series graph. The analysis starts with the time series plots for the given data so that the general features of the time series can be seen visually (see Fig. 1).

There was an upward trend in the enrollment values. In the year 2007/2008 there was a drop in the enrollment values due to post election violence. The same scenario repeated itself in the year 2020/2021 due to the COVID-19 pandemic. It was possible to establish whether or not the series was stationary.

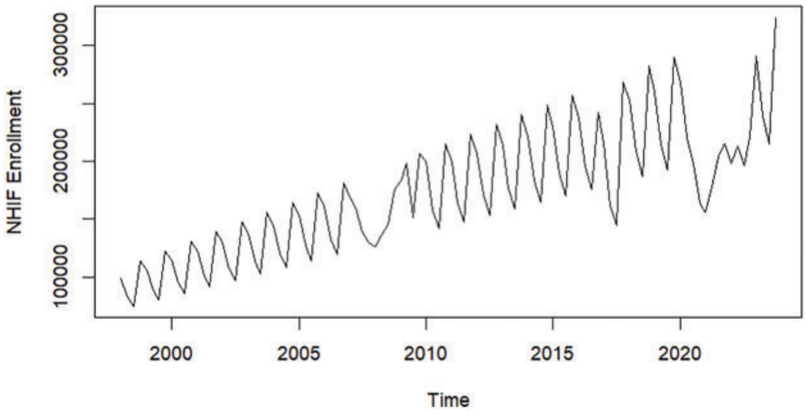

Fig. 1. Time series NHIF enrollment plot.

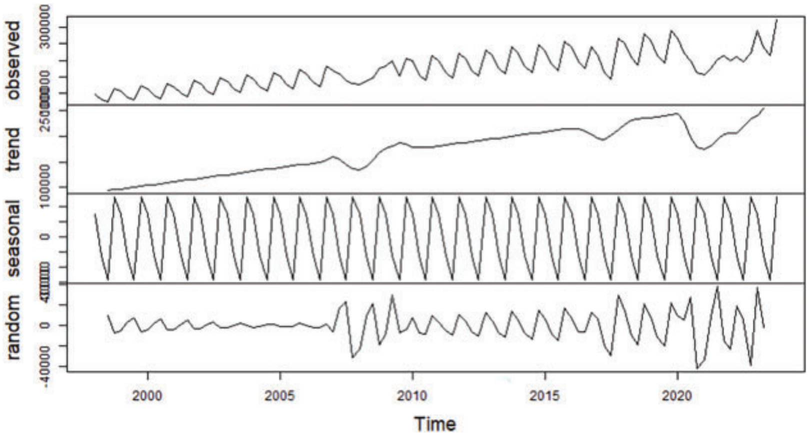

Fig. 2. Components of the quarterly enrollment time series.

TABLE II: Man Kendall Test Results

Tau	p-Value
0.628	2.22e-16

After concluding from a visual examination of the time series plots, tests are done to check if the series is stationary or not.

c) Decomposition of the NHIF Enrollment Time Series

The underlying patterns are divided into sub patterns in order to determine the component factors that affect each value in NHIF enrollment time series. The results are shown in Fig. 2.

i) Trend

There appears to be a linear upward trend component in the image. This indicates a consistent upward trend in the time series' overall value throughout the course of time. Analysis of the trend in the time series data was done using the Man Kendall test.

H_0 : *The correlation coefficient Tau* $= 0$ (*There is no correlation.*)

H_a : *The correlation coefficient Tau* $\neq 0$ (*There is a correlation.*)

Based on Table II, a Tau statistic of 0.628 and a very small p-value indicate that the two variables under study have a strong and statistically significant positive monotonic connection. That is to say, this relationship is unlikely to be the result of coincidence; as one variable rises, the other tends to rise as well. From Fig. 2, there is a drop in the enrollment in the year 2007/2008 due to the post-election violence and in 2020/2021 due to the COVID-19 pandemic.

ii) Seasonality

There is a definite periodic pattern in the seasonal component. The fact that it seems to be repeating every year indicates that the data may follow an annual cycle.

c) Stationarity Test

Having examined the ACF plot for periodic peaks at lags corresponding to the seasonal period of 4, the significant peaks at lag 4 suggest seasonality (Fig. 3).

There are significant spikes at lags 1 and 2, suggesting a strong correlation between a data point and its immediate past values.

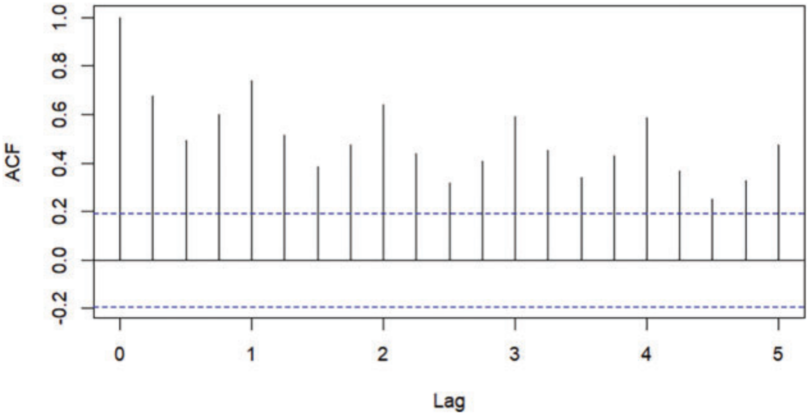

Fig. 3. ACF plot of NHIF enrollment data.

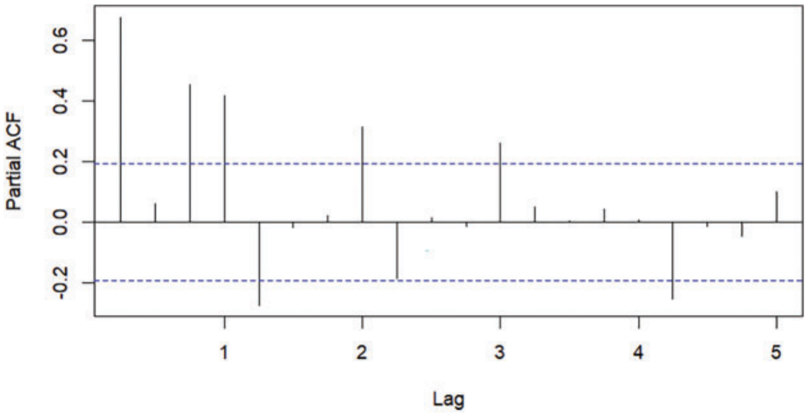

Fig. 4. PACF plot of NHIF enrollment.

TABLE III: ADF Test Results

ADF	P-Value	Lag order	Level of confidence
−3.9881	0.01246	4	95%

Significant spikes at certain lags indicate a direct relationship between a data point and its lagged values, independent of the effects of intermediate lags (Fig. 4).

To confirm the stationarity of the data the unit root test Augmented Dickey-Fuller test (ADF) was conducted. The stationary test, using Augmented Dickey-Fuller (ADF) test, a p-value below 0.05 was obtained (Table III). This suggested that the time series does not have a unit root and is stationary.

The ADF unit roots test showed stationarity of the series at level (ADF = −3.9881, p = 0.01246) for lag order 4. The significance level of 0.01246 in the ADF test indicated that the null hypothesis is rejected and the alternative hypothesis is accepted (time series is stationary).

d) Building SARIMA Model of NHIF Coverage

The SARIMA model for the NHIF enrollment time series data was developed by applying the Box-Jenkins Methodology. From the ADF test the data was found to be stationary. The SARIMA model requires the data to be stationary for it to be applied. The time series plot showed the data had the seasonality aspect from a visual point. The SARIMA model is used when the data has some seasonal aspects. This is in line with the study of Wanjuki *et al*. [7], where Evaluating the Predictive Ability of SARIMA Models, When Applied to Food and Beverages Price Index in Kenya was done taking into account the seasonal aspect of the data. The Box-Jenkins seasonal ARIMA model fitting procedure suggested several ARIMA models for the series. The fitted models were selected on the basis of some criteria as discussed in the following sections.

i) Model Identification

The ACF and PACF can be used to select the order of the best fit model, however it is not always accurate. In agreement with Nyoni and Nyoni [9], a grid search of possible orders was employed based on the lowest AIC value the best model is chosen (Table IV).

The model (1,1,3) $(0,1,1)_4$ (AIC = 2265, BIC = 2279.66). Low BIC and AIC values in the grid search gave preference to the model (1,1,3) $(0,1,1)_4$ these low values are in agreement with the study

TABLE IV: Model Identification Variables

Model	AIC value
SARIMA (0,1,0) (0,1,1) [4]	2288.79
SARIMA (0,1,0) (1,1,0) [4]	2296.092
SARIMA (0,1,0) (2,1,0) [4]	2297.941
SARIMA (0,1,1) (0,1,1) [4]	2282.383
SARIMA (1,1,3) (0,1,1) [4]	**2265**
SARIMA (0,1,2) (0,1,1) [4]	2266.396
SARIMA (0,1,3) (0,1,1) [4]	2267.999

TABLE V: Table of Coefficients

	Coefficients	Standard error
Ar. 1	−0.8254	0.1214
ma. 1	0.6272	0.1204
ma. 2	−0.7749	0.1014
ma. 3	−0.6634	0.0909
S. ma.1	−0.9558	0.1611

Note: log likelihood = −1126.04; AIC = 2264.09; AICc = 2265; BIC = 2279.66.

of Elgohari *et al*. [10], who predicted the quantity of hospital deaths using SARIMA (1,1,1) $(0,1,1)_{12}$ model which had the lowest values for RMSE and BIC as well as the lowest values for AIC.

ii) Model Estimation

This involves finding the values of the model's coefficients and the error between the model's predicted values and the actual observed values. The maximum Likelihood Estimate (MLE) was carried out to find the estimates of the coefficients of the model. The SARIMA (1,1,3) $(0,1,1)_4$ incorporates one non-seasonal AR process, three non-seasonal MA process and one seasonal MA process, with ($d = 1$), and ($D = 1$). The fitted model's final parameter estimates are illustrated in Table V.

The coefficients in Table V imply that the current time series is affected by its historical values (autoregressive terms), past period mistakes (moving average terms), and maybe seasonal trends (seasonal moving average term). The strength and direction of these correlations are shown by the particular values of the coefficients. An AR. 1 coefficient of −0.8254, indicates a correlation between a drop in the time series' current value and an increase in its value from one period prior. The first, second, and third moving average terms' coefficients (0.6272, −0.7749, −0.6634) show the connection between the time series' current value and its historical errors, or residuals. A seasonal MA term of order 1 (−0.9558). This is a key component of the seasonal pattern captured by the model.

The SARIMA model is fitted as shown.

- replacing the coefficients:

$$(1 + 0.8254\,\beta)\,y_t = \left(1 + 0.6272\,\beta - 0.7749\,\beta^2 - 0.6634\,\beta^3\right)\left(1 - 0.9558\,\beta^4\right)\varepsilon_t$$

- by expanding the model becomes

$$\begin{aligned}(1 + 0.8254\,\beta) = {} & \big(1 + 0.6272\,\beta - 0.7749\,\beta^2 - 0.6634\,\beta^3 - 0.9558\,\beta^4 - 0.5995\,\beta^5 + 0.7406\,\beta^6 \\ + {} & 0.6341\,\beta^7\big)\,\varepsilon_t\end{aligned}$$

iii) Model Diagnostics

Evaluating a fitted time series model is crucial. This was done by studying the residuals to see if any pattern remains unaccounted for hence verifying that the fitted model is adequate. To evaluate if the fitted model's residuals meet the assumptions of normality and autocorrelation the performance of two diagnostic tests was assessed. In diagnostic checking the ACF plots of the residuals in NHIF enrollments showed that the residuals lay within the 95% confidence limits. The results below indicated that, the residuals left over after fitting the models were white noise. Thus, there was no significant autocorrelation coefficients between residuals and given past residuals.

The findings in Fig. 5 are supported by the Ljung–Box Q-test showed insignificant Q-statistics (p-value = 0.3914).

The Ljung Box Test p-values exceed 5%, this suggests that there is some level of correlation but not strong enough (Table VI). This means the residuals do not deviate from white noise significantly.

A Q-Q plot is a graphical tool used to assess the normality of a dataset. It plots the quantiles of the sample data against the quantiles of a theoretical normal distribution. The Q-Q plot indicates that the data is likely normally distributed (Fig. 6).

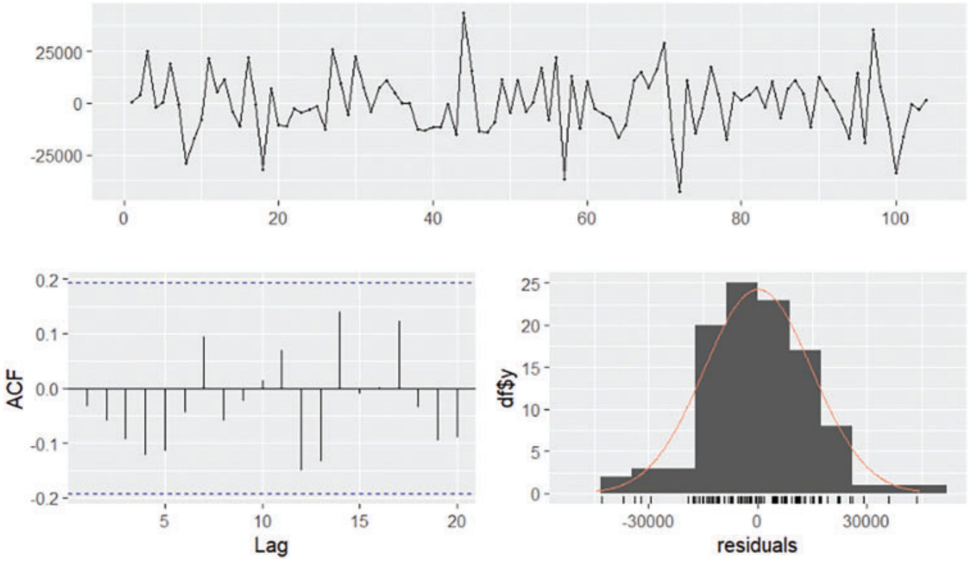

Fig. 5. Residuals from SARIMA (1,1,3) (0,1,1) [4].

TABLE VI: Box-Pierce Test Results

Q-statistic	Degrees of freedom	p-value
6.2906	6	0.3914

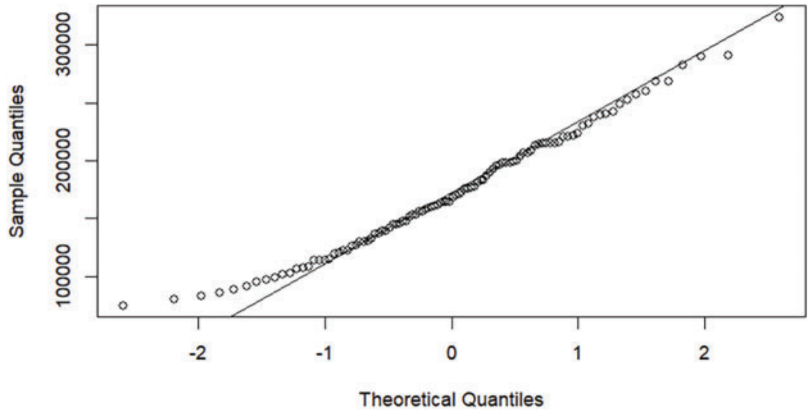

Fig. 6. Normal Q-Q plot.

TABLE VII: Standard Measures of Forecast Accuracy

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
−1263.392	19263.42	11197.31	−1.4926	5.872978	0.6163289	−0.05414976

The Quantile-Quantile (QQ) plot, illustrates that there may be a few small variations in the tails. Plotting the data set's quantiles against theoretical quantiles is shown by the dots. The data points seem to follow the diagonal line reasonably well, especially in the central part of the plot. The data seems to be roughly normally distributed based on the QQ plot.

To assess the models' ability to forecast, the standard measures of forecast accuracy were obtained as shown in Table VII. The values of these measures were obtained using the formulae mentioned in chapter three and the forecast errors.

The model's overall accuracy is relatively good, as indicated by the relatively low values of the RMSE, MAE, MAPE and MASE. This concurs with a study done [6] where the RMSE of the data was shown to be substantially less when compared to historical average forecasts.

e) Applying the SARIMA Model to Forecasting Future Enrolment Values

Since forecasting provides an understanding of future uncertainty, it aids in planning and decision-making. In order to formulate policies, forecasting models are necessary. The forecasts along with their corresponding lower and upper 95% confidence limit are presented in Table VIII.

The forecast column shows a general upward trend in NHIF enrollment over the forecasted period (2025–2027). By examining the forecasts across different quarters and years, there is a seasonal pattern

TABLE VIII: Forecasts

YEAR	QUARTERS	Forecast	Lo 95%	Hi 95%
2025	Q1	269735	217093	322377
	Q2	239832	187053	292611
	Q3	226304	173157	279452
	Q4	283791	230556	337026
2026	Q1	275212	221379	329044
	Q2	245954	191888	300020
	Q3	231894	177531	286257
	Q4	289820	235314	344325
2027	Q1	280878	225828	335928
	Q2	251920	196576	307264
	Q3	237613	181994	293232
	Q4	295743	239944	351542

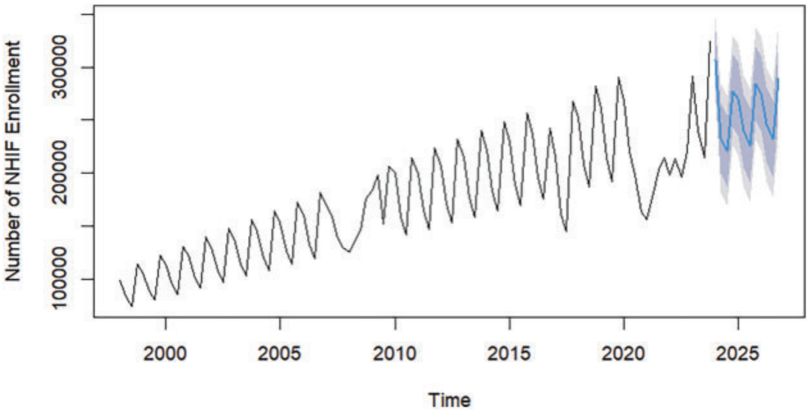

Fig. 7. Forecast from seasonal ARIMA (1,1,3) (0,1,1) [4].

where the forecast is higher in certain quarters than others. These intervals indicate the range of possible values for the future enrollment based on the model. This indicates that the model's predictions become less certain with time. While the model provides forecasts based on historical data and trends, external factors such as economic conditions, policy changes, or public health crises could influence actual enrollment.

The blue line in Fig. 7 provides the model's best estimate of the future enrollment values. The shaded area indicates the range of values within which the actual enrollment is expected to fall with a certain level of confidence. The shaded area represents the uncertainty associated with the forecasts. The upper and lower bounds of the shaded area define the prediction interval.

3. Conclusion

The SARIMA (1,1,3) $(0,1,1)_4$ captured the connection linking NHIF enrollment and economic indicators, suggesting that decline in the economy can cause temporary decreases in enrollment due to hardships in finances. The model was suitable for the NHIF population coverage in Kenya. From the historical records accurate three-year forecast were made. NHIF enrollment has exhibited a general upward trend of about 2% every year over the study period, indicating increasing awareness and acceptance of the program.

To address these challenges and capitalize on the opportunities, NHIF and SHA should work collaboratively to develop a comprehensive transition plan that addresses the key areas of integration, communication, and service continuity. By effectively managing the transition, NHIF and SHA can ensure a smooth transition to the new healthcare system and improve the overall health and well-being of Kenyans.

Recommendation to future researchers to investigate alternative modeling approaches such ANN (Artificial Neural Networks). This could provide valuable insights and enhance the precision of future forecasts. Future research could explore the influence of specific factors on SHA enrollment, such as economic conditions, policy changes, and demographic trends.

Acknowledgment

I would like to thank Tharaka University for the support. I am also grateful to Prof. Dennis Muriithi and Dr. Daniel Mwangi for their invaluable guidance and expertise.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

[1] KNBS, ICF. *Kenya Demographic and Health Survey 2022: Key Indicators Report*. Nairobi, Kenya: KNBS and ICF; 2023.

[2] Barasa E, Rogo K, Mwaura N, Chuma J. Kenya national hospital insurance fund reforms: implications and lessons for universal health coverage. *Health Syst Reform*. 2018 Oct 2;4(4):346–61.

[3] Kumar B, Yadav N. A novel hybrid model combining βSARMA and LSTM for time series forecasting. *Applied Soft Computing*. 2023 Feb 1;134:110019.

[4] Box GE, Jenkins GM, Reinsel GC. *Time Series Analysis, Forecasting and Control, Available from*. 3rd. Englewood Cliffs (NJ): Prentice Hall; 1994 [cited 2021 Dec 20]. Available from: Kenya and Rockville, Maryland, USA: KNBS and ICF.

[5] Sen J. A forecasting framework for the Indian healthcare sector index. *Int J Bus Forecast Mark Intell*. 2022;7(4):311–50.

[6] Al-Haque S, Ceyhan ME, Chan SH, Nightingale DJ. Responding to traveling patients' seasonal demand for health care services. *Mil Med*. 2015;180(1):111–7.

[7] Wanjuki TM, Wagala A, Muriithi DK. Evaluating the predictive ability of seasonal autoregressive integrated moving average (SARIMA) models using food and beverages price index in Kenya. *Eur J Math Stat*. 2022;3(2):28–38.

[8] Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. Melbourne: OTexts; 2018.

[9] Nyoni SP, Nyoni T. Hypertension trends at Silobela District Hospital (SDH): an application of the Box-Jenkins Model. *Int J Multidiscip Res*. 2020;6(1):109–20.

[10] Elgohari H, Abdulmajeed M, Elrefaey A. Application SARIMA models on time series to forecast the number of deaths in hospital. *Int J Appl*. 2019;7(4):9–18.