Forecasting the National Health Insurance Fund Membership Enrolment in Tanzania Using the SARIMA Model

Alfred Tembo[1]
Bahati Ilembo[2]
Joseph Lwaho[3]

[1]alfredtembo83@gmail.com
[2]bmilembo@mzumbe.ac.tz
[3]jolwaho@mzumbe.ac.tz

[1,2,3]Mzumbe University, Morogoro, United Republic of Tanzania

ABSTRACT

This paper aimed at forecasting membership enrolment in the National Health Insurance Fund (NHIF) in Tanzania using quarterly time series data. This study used 88 time series data to fit the seasonal Autoregressive Integrated Moving Average model (SARIMA). ARIMA (3,1,1) (0,1,0)[4] model was built and used for forecasting. The results show that there will be an increasing membership enrolment overtime over the years and no signs of decreasing. Thus, the government, apart from continuing subsidizing the cost of accessing health insurance services, should also improve the National Health Insurance (NHI) coverage to accommodate the increased enrolment and discourage dropouts. In turn, this will help to achieve the Universal Health Coverage (UHC) ultimate aim of ensuring equitable access to essential and manageable healthcare services, regardless of individuals' financial situations, their location, and personality.

Key words: ARIMA, Box- Jenkins, National Health Insurance, SARIMA, Seasonal

I. INTRODUCTION

Health insurance (HI) schemes in developing countries generally do not reach the targeted populations to ensure access to essential quality health services (Osei Afriyie et al., 2022). In recent years, enrolment into Health Insurance Scheme (HIS) as a strategy to attain the Universal Health Coverage (UHC) has become a global discussion. Numerous studies including those of Boateng (2024); Yego et al., (2023); Ng'ang'a, (2021); Alesane and Anang, (2018) have documented efforts taken by individual countries to promote the use of HI as a means to achieve a (UHC). However, according to Kathrin et al. (2021), enrolment rates of many low to middle income countries, especially those located in the Sub-Saharan Africa (SSA) region are still low and rarely exceed 10%. Compared to the rural, urban regions experience a remarkably higher number of new entrants to the health insurance scheme (Sharma, 2023). On the other hand, the urban and peri- urban areas have also experiencing a significant number of renewals compared to other areas (Nsiah-Boateng, and Aikins, 2018). This trend has been related to the affordability of the premium, and contribution (Yego et al., 2023; Kusi et al. 2015). Further, there is a direct relationship between indigenous health equity and enrolment in the health insurance scheme, an important aspect to consider in addressing the disparities faced by indigenous populations in accessing healthcare. Either way, enrolment in HI is a viable means of improving the sustainability of the schemes thereby enabling access and utilization of healthcare services towards UHC.

Studies exist in determining and assessing factors that hinder enrolment to the NHIS (Boateng, 2024; Afriyie et al., 2023; Yego et al., 2023; Ng'ang'a, 2021; Ghimire et al., 2019; Alesane & Anang, 2018; Kotoh et al., 2018) and none was able to predict future enrolment. Few studies (Marinova & Todorova, 2023; Yego et al., 2023, 2021; Kathrin et al., 2021) attempted to predict health insurance uptake using machine learning techniques, nevertheless, none of them considered time series forecasting. This study brings a new dimension of predicting future membership enrolment using historical data on enrolment. Given the seasonality nature of the membership enrolment time series data, this paper used the strengths of time series analysis, particularly SARIMA model. The model is appropriate for predicting quarterly, monthly, and yearly membership enrolment into NHIF in Tanzania but also its ability to show good prediction performance as suggested by Wiah et al. (2022). This paper is organized as follows, section 2 presents the literature review, section 3 presents the methodology. Findings and discussion is on section 4 and the paper concludes in section 5 on conclusions and policy recommendations.

II. LITERATURE REVIEW

2.1 Conventional Theory of Health Insurance

According to Nyman (2001), conventional theory holds that people purchase insurance because they prefer the certainty of paying a small premium to the risk of getting sick and paying a large medical bill. Conventional theory also holds that any additional health care that consumers purchase because they have insurance is not worth the cost of producing it. Therefore, economists have promoted policies-copayments and managed care-to reduce consumption of this additional, seemingly low-value care. This presents a new theory of consumer demand for health insurance which holds that people purchase insurance to obtain additional income when they become ill. Thus, additional income generates purchases of additional high-value care, often allowing sick persons to obtain life-saving care that they could not otherwise afford. Regarding risk, the new theory relies on empirical studies showing that consumers actually prefer risk of large risk to incurring a smaller loss with certainty. It is further argued that, if consumers purchase insurance, it is not because they desire to avoid risk. Instead, the new theory suggests consumers simply pay a premium when healthy in exchange for a claim on additional income (effected when insurance pays for the medical care) if they become ill. Health insurance is substantially more valuable to the consumer under the new theory. The new theory moreover implies that copayments and managed care-central health policies of the last 30 years-were directed at solving problems that largely did not exist. Because these policies either reduced the amount of income transferred to ill persons or limited access to valuable health care, they may have done more harm than good. The new theory also provides a solid theoretical justification for insuring the uninsured and for implementing national health insurance. It is from this theoretical background that this paper is set to forecast membership enrolment in the NHIF Tanzania using Seasonal Autoregressive Moving Average methodology of time series analysis.

2.2 Empirical Literature Review

The approaches to study the status of membership enrolment in the health insurance schemes has been the use of binary regression models and in many cases, the analysis was driven by the uptake of health insurance than predicting membership enrolment (Yego et al., 2023; Ng'ang'a, 2021; Alesane & Anang, 2018). At least for Boateng (2024), Marinova and Todorova (2023), and Yego et al. (2021), the analysis on predicting membership enrolment is seen, though all the studies do not consider time series as the tool for predictions and rather use selected families of regressions such as logit and probit because of the nature of the dependent variable. Regressions may not be sufficient to predict membership enrolment given time series data which are seasonal as they do not provide robust analysis to carter for the seasonal nature of the enrolment (quarterly, monthly and yearly data). As a result, the seasonal autoregressive moving average (SARIMA) becomes appropriate to perform the forecast. In other health related forecasts, time series analysis has also gained popularity. Kathrin et al. (2021) forecasted distribution and pattern of a health claim system whose aim was to analyze the distribution and future pattern of insurance health claim system using time series approach. Akaike information criterion and Schwarz Bayesian criterion were used to select the adequate model through maximum likelihood estimation methods. ARIMA (0, 0, 0) (1, 0, 1) [12] model was chosen to forecast claim amounts. Existing studies in forecasting in health dominates the literature (Putri et al., 2023; Morgan et al., 2022; Jalalpour et al., 2015; Soyiri & Reidpath, 2012, 2013), predominantly on membership enrolment, claims payments and health expenditure but have been using Autoregressive Integrated Moving Average (ARIMA). This study employs the SARIMA which is capable of handling non stationary data but also good for seasonal data such as membership enrolment to a health insurance scheme. Subsequently, the study identifies an appropriate type of model based on the Box-Jenkins methodology for making the forecasts. The paper applies the static one step ahead forecasting method to the annual data over the period 1970-2015.

III. METHODOLOGY

3.1 Data Source and Model Justification

The study used secondary univariate quarterly time series data obtained from the NHIF in Tanzania past 22 years and contained 88 enrolee observations a minimum requirement of the generalization ability of time series analysis (Lwaho & Ilembo, 2023). SARIMA$(p, d, q)(P, D, Q)[4]$ model was used in the analyzing membership enrolment. To achieve this the function auto. arima () in R software was used to select the best model for forecasting membership enrolment based on the minimum value of Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) and Maximum log-likelihood. According to Wiah et al. (2022), the SARIMA model shows good prediction performance when used to forecast seasonal data compared to ARIMA models which are not capable of capturing seasonal patterns. Therefore, this study used the strengths of the SARIMA model, which is capable of capturing seasonality (Box and Jenkins, 1970).

3.2 The Conceptual Framework

The analysis in this paper is guided by the Box- Jenkins methodology of forecasting (Box & Jenkins, 1970). This model involves three major steps namely identification, estimation and forecasting. In such fields such as health, business and economics, usually time series contain a seasonal periodic component which repeats every "s" observations. This study considered quarterly observations where the $s = 4$, but the application to other values of s is straightforward. According to Box and Jenkins (1970) the generalized ARIMA model to deal with seasonality and defined general multiplicative seasonal model is given by:

$$\vartheta_p(B)\varphi_p(B^4)w_t = \theta_q(B)\xi_Q(B^4)a_t \quad \ldots \quad 3.1$$

Whereby $B =$ backward shift operator

$\vartheta_p, \varphi_p, \theta_q, \xi_Q$ = polynomials of order p, P, q, Q respectively and

$\{a_t\}$ = independent random variables with zero mean and variance σ_a^2

The shift operator $(B^4)w_t$ is such that $(B^4)w_t = w_{t-4}$. Therefore, equation (3.1) defines a stationary model provided that the roots of $\vartheta_p(B)\varphi_p(B^4) = 0$ lie outside the unit circle. In order to fit the model to a non- stationary series, Box and Jenkins (1970) again, suggest differencing the original series to remove both trend and seasonality following the procedure underneath:

$$w_t = \nabla^d \nabla_4^D x_t$$

Whereby $\nabla_4 x_t = x_t - x_{t-4}$ and that:

$$\nabla^{/}\nabla_4 x_t = \nabla_4 x_t - \nabla_4 x_{t-1} = x_t - x_{t-1} - x_{t-4} + x_{t-5}.$$

The values of the integers d and D do not usually need to exceed a unit. Details that describe the Box and Jenkins procedure to forecasting can also be found in the old literature by Naylor et al. (1972), Chatfield and Prothero (1973), and Thompson and Tiao (1971). However, this paper provides the detailing procedure described by Box and Jenkins in their popular work of 1970.

The procedure consists of fitting a mixed autoregressive integrated moving average (ARIMA) model to a given set of time series data and then taking conditional expectations. The main stages in setting up a Box-Jenkins forecasting model are as outlined below:

Model identification: In this step, the aim is to examine the data to reveal which member of the class of ARIMA processes appears to be the most appropriate. Estimation*:* In this step, the aim is to estimate the parameters of the chosen model by least squares. Diagnostic checking*:* In this step, the aim is to examine the residuals from the fitted model to see if it is adequate. Alternative model consideration*:* if the first model appears to be inadequate for some reason, then consideration of an alternative model other than ARIMA models may be tried until a satisfactory model is found. For non-seasonal data, first-order differencing is usually appropriate. Moreover, for seasonal data of period 4, the operator $\nabla\nabla_4$ is often used if the seasonal effect is additive, while the operator ∇_4^2 maybe used if the seasonal effect is multiplicative. In our case, the general SARIMA model is defined as $\vartheta_p(B)\varphi_p(B^4)w_t = \theta_q(B)\xi_Q(B^4)a_t$ has been used.

3.2.1 Stationarity Test

To test for stationarity, the Augmented Dickey fuller (ADF) test was used. The null hypothesis that there is a unit root will not be accepted at a given level of significance. The Augmented Dickey-Fuller test is a unit root based on stationarity (Dickey & Fuller, 1979). The unit-root based test is associated with the first lag of the time series variable. If the coefficient $(\gamma = 1)$ has a unit root, the time series behaves similarly to the random walk model which is non-stationary, and if the coefficient $|\gamma| < 1$ then, there is no unit root. Hence, we can test statistically whether the coefficient (γ) is equal to one or not. The Dickey-Fuller test adopts this procedure by carefully manipulating the equation, given as:

$$y_t = \alpha + \beta t + \emptyset\, y_{t-1} + e_t \quad \ldots \quad 3.2$$

Also, written as

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta t + \gamma\, y_{t-1} + e_t \quad \ldots \quad 3.3$$

In Dickey-Fuller test, we test the hypothesis

$H_0: \emptyset = 1$

$H_1: \emptyset \neq 1$

Correlograms

ACF and PACF are statistical measures that help to analyze the relationship between a time series and its lagged values. They are generally producing plots that are very important in finding the order of Autoregressive (AR) and Moving Average (MA) models.

Autocorrelation Function (ACF)

ACF measures the linear relationship between a time series and its lagged values. It assesses how much the current value of a time series depends on its past values. Autocorrelation is fundamental in time series analysis, helping identify patterns and dependencies within the data. The correlation between the current observation (y_t) and the previous observation (y_{t-k}) is given as:

$$\rho_k = \text{corr}(y_t, y_{t-k}) = \frac{\text{Cov}(y_t, y_{t-k})}{\sqrt{\text{Var}(y_t) \cdot \text{Var}(y_{t-k})}} = \frac{\gamma_k}{\gamma_0} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.4$$

Where, k = 1, 2, …..

Partial Autocorrelation Function (PACF)

PACF removes the influence of intermediate lags, providing a clearer picture of the direct relationship between a variable and its past values. Unlike Autocorrelation, partial Autocorrelation focuses on the direct correlation at each lag. The partial Autocorrelation function at lag k for time series is given as:

$$\emptyset_{11} = \text{Corr}(Y_{t+1}, Y_t) = \rho_1$$

$$\emptyset_{kk} = \text{Corr}(Y_{t+k} - \hat{Y}_{t+k}, Y_t - \hat{Y}_t), \text{k} \geq 2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.5$$

The suitable values of p and q will be selected by observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series data. The appropriate ARIMA models will be selected by observing the behaviour of ACF and PACF spikes based on the order identified (Hyndman & Athanasopoulos, 2018).

Seasonality Test

An approach to seasonality testing is to conduct seasonal decomposition using an additive model. The model decomposes time series into its trend, seasonal, cyclical, and regular components by assuming time series can be modeled by using those components. Visual examination can detect the seasonal component for regular patterns that repeat at a fixed interval. If the seasonal component exhibits regular patterns, this indicates the presence of seasonality in the time series.

3.3 Model Estimation

The parameters of the selected seasonal ARIMA (SARIMA) model with the specific values of (p, d, q) × (P, D, Q) $_s$ needs to be estimated. The maximum likelihood estimation (MLE) estimates the coefficients of the suggested models at the identification stage. Selection of the best model was based on AIC and BIC.

Akaike Information Criterion (AIC)

Kullback et al. (1951) developed a measure to capture the information that is lost when approximating reality. Kullback and Leibler measure is a criterion for a good model that minimizes the loss of information. Two decades later, Akaike established a relationship between the Kullback-Leibler measure and the maximum likelihood estimation (MLE) method that was used in many statistical analyses for model selection (Akaike, 1974). This criterion referred to as AIC, is generally considered the first model selection criterion that should be used in practice. The AIC is given as:

$$\text{AIC} = -\log(\hat{\theta}) + 2k \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.6$$

Where; θ is the set of model parameters, L ($\hat{\theta}$) is the likelihood of the candidate model given the data when evaluated at the maximum likelihood estimate of θ and k is the number of estimated parameters in the candidate model. Since AIC does not consider the effect of sample size, for small sample sizes, the second-order equation of the Akaike information criterion (AIC_C) is defined as:

$$\text{AICc} = -2\text{logL}(\hat{\theta}) + 2\text{k} + \frac{(2\text{k} + 1)}{(\text{n} - \text{k} - 1)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.7$$

where n denotes the total number of observations.

A small sample size is when $^{n}/_{k}$ less than 40, also that when the number of observations increases, the third term in AIC_C approaches zero and will therefore give the same result as AIC in the equation 3.6

Bayesian information criterion (BIC)

Bayesian information criterion is another model selection criterion based on information theory but set within a Bayesian context. The difference between the BIC and AIC is the greater penalty imposed for the number of parameters

$$\text{BIC} = -2\log L(\hat{\theta}) + k log\, n \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.8$$

where n denotes the total number of observations.

The BIC strongly penalizes the number of involved parameters. High values of AIC mean that the observed data does fit the model, while lower values indicate strong evidence that the observed data fit the models. Similarly, lower values of BIC indicate better fitting of the models.

3.3.3 Diagnostic Checking

Diagnostic check evaluating the acceptance of the fitted SARIMA model by examining the residuals, which are the difference between observed and predicted values. The aim is to ensure the residuals are random and do not contain any patterns or structures. The diagnostic checks involved the use of the Ljung-Box test statistic and the forecasting accuracy.

Ljung-Box Test

The Ljung-Box test helps to check whether the errors or residuals in our model have any pattern or correlation. The hypotheses under the Ljung-Box test is defined as:

H_0 = Residuals are independently distributed, correlation in the population from which the sample is taken is 0

H_1 = Residuals are not independently distributed; they exhibit serial correlation.

The Ljung-Box test statistic is given as;

$$Q = \mathrm{n(n+2)} \sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.9$$

Where n is the sample size, $\hat{\rho}_k^2$ is the sample autocorrelation at lag k, and h is the number of lags being tested. Under the null hypothesis, the test statistic Q asymptotically follows a $\mathcal{X}_{(h)}^2$, for the significant level α, the critical region is rejected if;

$$\mathrm{Q} > \mathcal{X}_{(1-\alpha),\mathrm{h}}^2$$

Where $\mathcal{X}_{(1-\alpha),h}^2$ is the $(1-\alpha)$ quantile of Chi-square distribution with h degrees of freedom.

After diagnostic checking, the fitted model will used in forecasting future values if the model is adequate. Otherwise, we need to repeat the selection and estimation method. Try with another potential candidate model (Ramasubramanian, 2007).

Forecasting and Forecasting Accuracy

Once the selected model has been verified, the model will be then used to predict membership enrolment in the next 24 months. After the forecasting, this study employs the Mean Percentage Error (MAE) and Mean Absolute Percentage Error (MAPE) to evaluate the forecasting accuracy of the selected model. The MAE and MAPE are given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.10$$

$$\text{and } MAPE = \frac{100}{n}\sum\left|\frac{(p_i - o_j}{o_j}\right| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.11$$

Where p_i is the predicted value for the i^{th} observations, o_i is the observed value for the j^{th} observation, n is the number of non-missing residuals.

IV. FINDINGS & DISCUSSIONS

4.1 Model Identification Process

Figure 1 shows the time series plots for the membership enrolment from 2002 Q1 to 2023 Q4. The plot **s**hows an unpredictable pattern in the long-term and the series serves to be seasonally increasing upward movement and nonstationary. However, the stationarity of the time series is confirmed by using the ADF test.

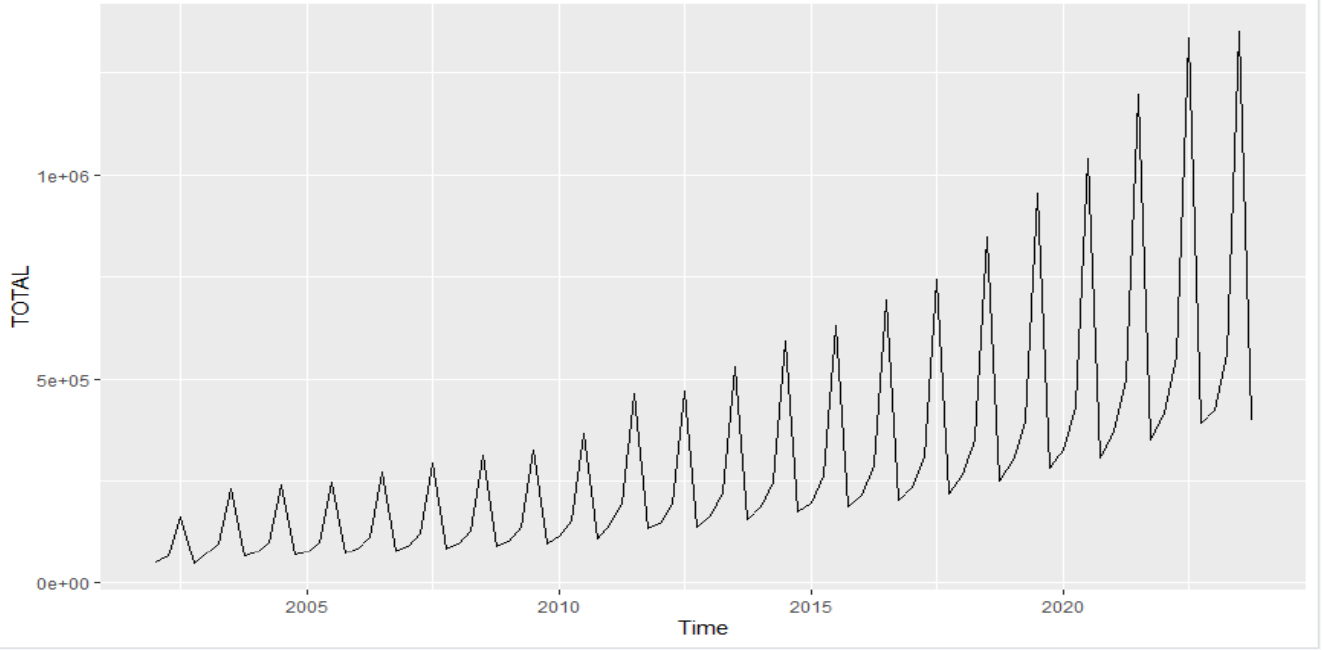

Figure 1
Time Series Plot of the Quaterly NHIF Membership Enrolment in Tanzania, 2002-2023

It was observed that there was an increasing upward but the peaks were not repeated with the same interval of time and not repeated with the same intensity rather than peaks increasing with increasing time. This shows that the series was not stationary. Since the original data show that they were nonstationary, the transformation was taken by natural logarithm and first differencing (d = 1) to make them stationary. Further checks were performed including the Dickey-Fuller test.

4.2 Augmented Dickey-Fuller (ADF) test

Table 1 shows that the results of the ADF test indicating that the null hypothesis is not rejected at 5% level of significance. Therefore, the series contains a unit root and it confirms that the data is not stationary. This result can be also inferred by examining correlograms in Figure 2 below.

Table 1
The Result of the ADF Test

Augmented Dickey-Fuller Test		
Dickey-Fuller = −0.963	Lag order = 4	P-value = 0.9389

Alternative hypothesis: Stationary

It was observed that there was an increasing membership enrolment upward but the peaks were not repeated with the same interval of time and not repeated with the same intensity rather than peaks increasing with increasing time. This shows that the membership enrolment was not stationary. Since the original data show that they were nonstationary, the transformation was taken by natural logarithm and first differencing (d = 1) to make them stationary.

4.2.1 Time Series Plot of Residuals

The adequacy of the model selected for forecasting is checked at this stage through the residuals concerning the specific variable. The residuals plot of ACF and PACF are given in Figure 2 below.

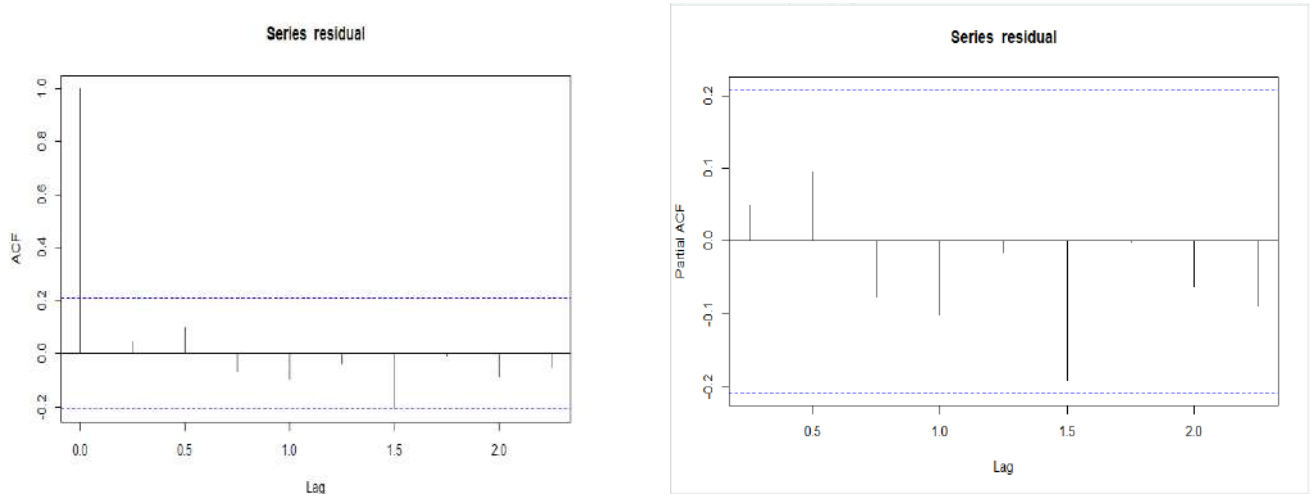

Figure 2
The ACF and PACF Residuals of the Fitted Model Total Seasonal ARIMA Model

The graphs of ACF and PACF in Figure 2 of the residual shows that most of the lags are within the limit which shows that the information was well captured through the model used and thus the model suggested is good for fitting the specific variable.

The seasonal ARIMA (3, 1, 1) (0, 1, 0) [4] was selected because of the lowest Akaike Information Criterion Corrected of 1896.91 and the largest log-likelihood of -943.46 among other models, and it was considered the best model for forecasting. The R-Software confirms the required model automatically. The model that was identified and selected was seasonal ARIMA (3, 1, 1) (0, 1, 0) [4]

4.2.2 Plots of the Residuals

Residuals of the variable are required to be normally distributed, the histogram of the residual of the total variable of the study showed the normality to prove that the model selected was the best fit. The Q-Q plot shows some data was out of the straight line so it shows that some data are not well fit. This is shown in Figure 3.

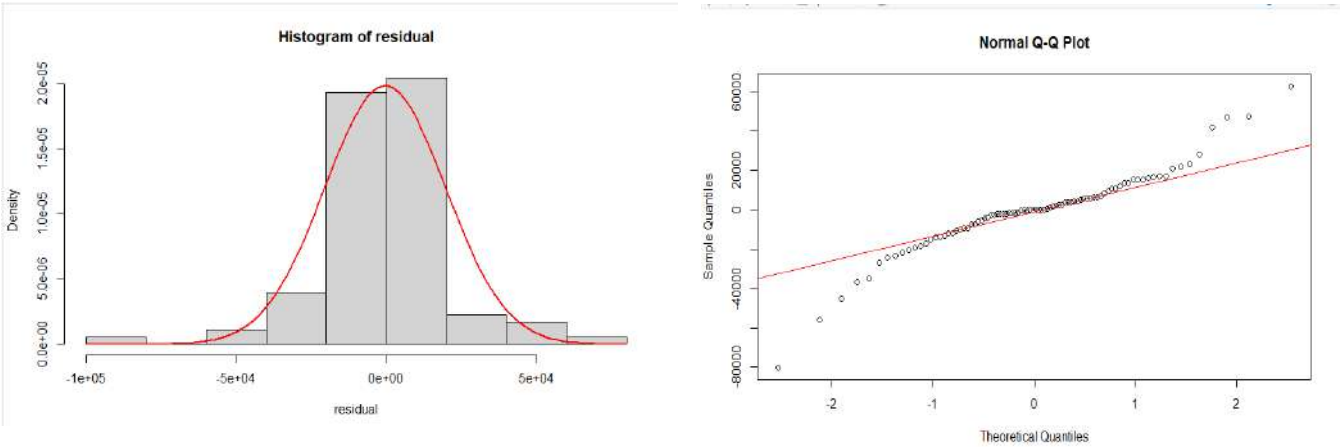

Figure 3
The histogram and Q-Q Plot of the Residuals

The time plot shows evidence of constant variance, the ACF plot shows that the errors are approximately uncorrelated, and the histogram shows that the errors are approximately normally distributed with mean zero. The adequacy test result is also confirmed by the formal test Ljung-Box in section 4.2.3 underneath.

4.2.3 Ljung-Box Test

The Ljung-Box test was performed on the residuals of the model and the results showed the p-value obtained was 0.2989 which is greater than 0.05 significance level (test value), indicating that residuals are uncorrelated and they are pure randomly.

4.3 Seasonality Test

Figure 4 shows the result of seasonal decomposition using an additive model conducted to detect seasonality. It observed that the seasonal component exhibits regular patterns or cycles, which indicate the presence of seasonality in time series. This is confirmed by the seasonality test which proves the presence of seasonality. This means that the SARIMA model could be an appropriate forecasting model to be used in predicting yearly membership enrolment to the National Health Insurance Fund.

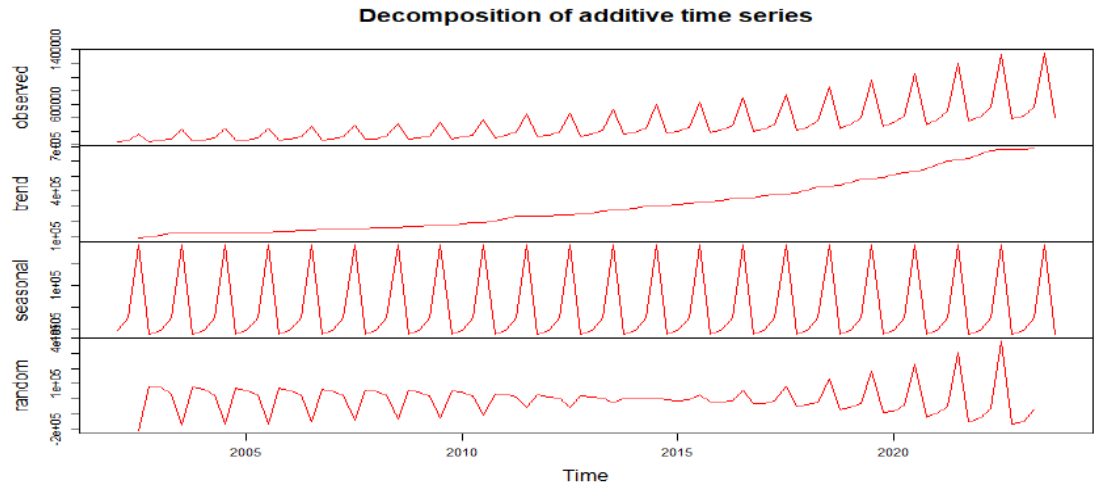

Figure 4
Time Series Decomposition

4.4 Model Selection and Estimation

Since the time series tends to be stationary, we have to identify the order of Seasonal ARIMA$(p, d, q)(P, D, Q)[4]$. To achieve this the function auto. arima () in R software was used to select the required model for forecasting membership enrolment with the minimum value of AIC and BIC and Maximum log-likelihood. The study arrived at a SARIMA model which is the Seasonal ARIMA (3,1,1) (0,1,0)[4] and was selected due to the lowest AIC of 1896.91, BIC of 1909.01 and the largest log-likelihood of -943.46 among the other models and it was considered as the best model for forecasting membership enrolment in the NHIF.

4.5 Validating the Model

The model validation is usually done to assess the precision of the model fit in estimating the observed values. The forecasted predictions for the validation set are plotted against the observed values as seen in Figure 5. The predicted values are fitted using the original data. It can be concluded that the model is best for the series.

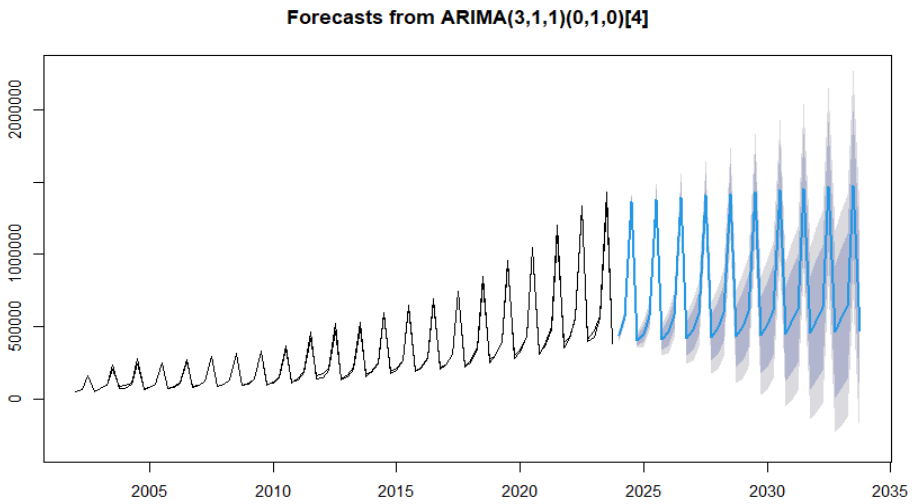

Figure 5
Model Validation of Membership Enrolment Using SARIMA (3,1,1) (0,1,0)[4]

4.6 Forecasting

Based on the SARIMA (3,1,1) (0,1,0)[4] model, the membership enrolment for the next 28 quarters is provided in Table 2, and Figures 6 and 7 show the trend of forecasted membership enrolment.

Table 2

Forecasted Values for the TOTAL Variable

Point	Forecast	Point	Forecast
2024 Q1	437937.6	2027 Q3	1407652.8
2024 Q2	568403.0	2027 Q4	423210.6
2024 Q3	1367749.9	2028 Q1	491066.5
2024 Q4	403209.4	2028 Q2	596782.6
2025 Q1	452009.3	2028 Q3	1420140.7
2025 Q2	575001.9	2028 Q4	430828.2
2025 Q3	1381497.4	2029 Q1	503319.4
2025 Q4	409295.3	2029 Q2	604665.4
2026 Q1	465482.1	2029 Q3	1432296.9
2026 Q2	581932.0	2029 Q4	438786.1
2026 Q3	1394786.7	2030 Q1	515279.7
2026 Q4	416002.7	2030 Q2	612817.6
2027 Q1	478474.1	2030 Q3	1444164.9
2027 Q2	589196.9	2030 Q4	447030.0

In Table 2, the forecasted enrollment numbers are increasing and others remain stable over time, so this indicates growing participation in the NHIF. The increase in registration membership for the coming years is shown evidently in Figures 6 and 7.

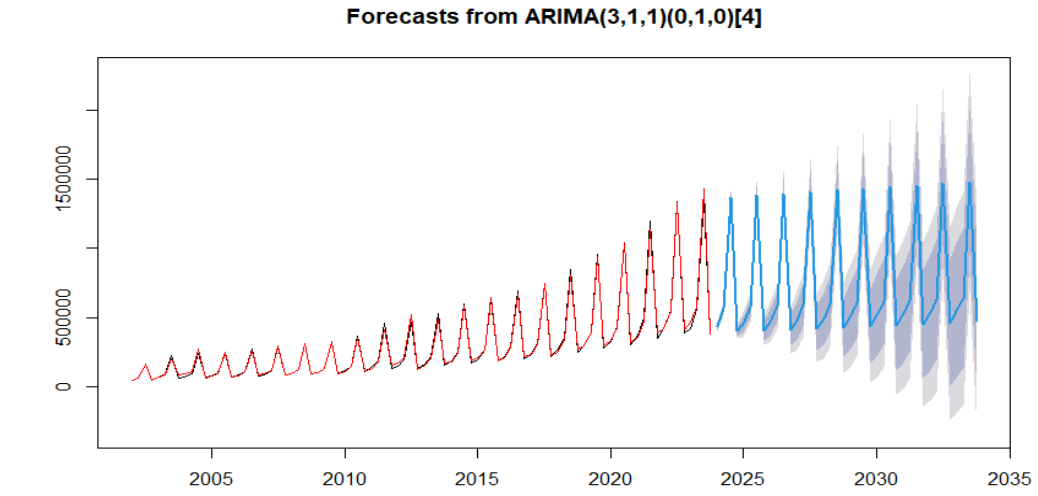

Figure 6

The Time Series Plot of the Forecasted Variable

Figure 7 presents the plot of total enrolment members against year (quarterly). The results show that there is a small upward increase in the enrolment members from 2024 to 2030, which is also a linear trend. Despite the fluctuations through increasing or decreasing from one year to another, this shows that through the line, there is an increasing trend of total enrolment of members over the years. Because the years before forecasting in 2022 (Q1, Q2, Q3, Q4) were (418516, 556034,1338918, 393610) and 2023 (Q1, Q2, Q3, Q4) were (423084, 562104,1353538, 397908) and remaining years were recorded respectively, so through the trend analysis of forecasted trend it shows there is an upward increase of enrolment for the years or quarterly. The trend line in the graph described below shows a positive slope most likely indicating an upward trend, signifying an increase in enrolment over time. This means the

data points show a general increase in total enrollment over time, The line slants upwards from left to right. This also signifies that as the x-axis value (time) increases, the y-axis value (enrolment) also increases.

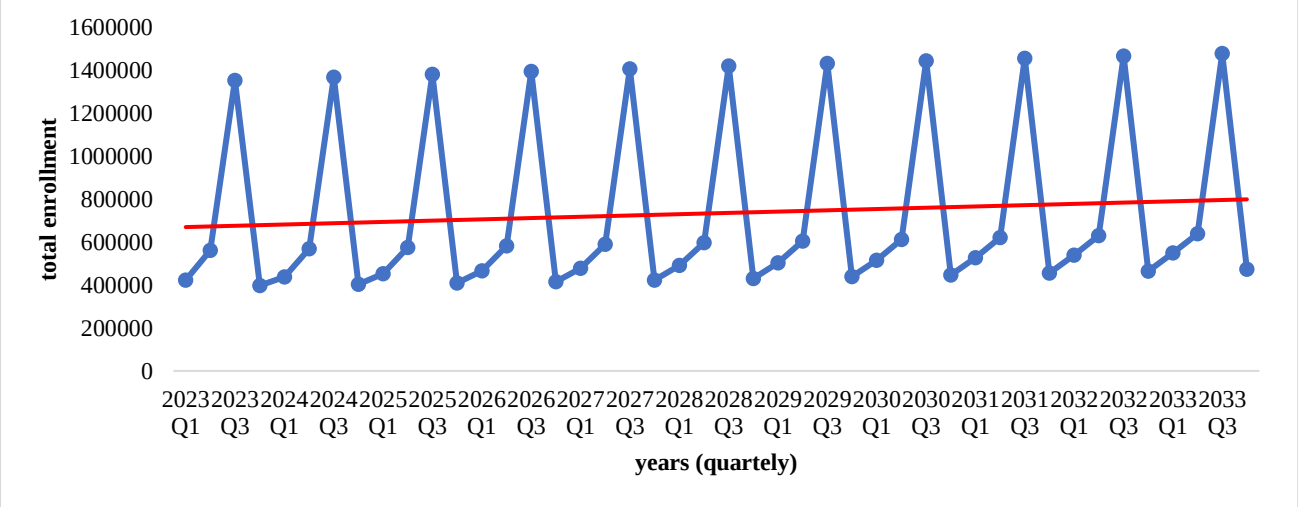

Figure 7
Trend Analysis of Forecasted NHIF Membership

V. CONCLUSIONS & RECOMMENDATIONS

5.1 Conclusions

The general objective of the study was to forecast membership enrolment to the NHIF in Tanzania using the Seasonal autoregressive integrated moving average (SARIMA) model. SARIMA (3,1,1) (0,1,0)[4] model was used and was built following Box and Jenkins methodology for handling seasonal data. The results show that there will be an increasing membership enrolment overtime over the years and no signs of decreasing. The SARIMA model used, helped in determining future membership enrolment to the fund in the country. However, due to the fluctuation in the data series, this research needs to be extended by applying other methodologies such as the Autoregressive Integrated Moving Average with exogenous variable (SARIMAX), Simple Exponential Smoothing (SES) or the Holt-Winters Exponential Smoothing (HWES). These methodologies may improve the results from this study and widening the scope on how the forecasting of the membership enrolment to the country's national health insurance funds can be best handled.

5.2 Policy Recommendations

Findings from this study will enable policy makers in Tanzania and government officials responsible for the health sector, especially those responsible for the National Health Insurance Fund to make a well- informed decision in matters pertaining to NHIF enrolment, access and management of the insurance fund in general. Also, the government, should improve the NHI coverage to accommodate the increased enrolment and discourage dropouts. An increased enrolment when coupled with assured accessibility to the health services, will help to achieve the Universal Health Coverage (UHC) aim of ensuring equitable access to essential and manageable healthcare services which is inclusive, regardless of individuals' financial situations, their setting, and personality.

REFERENCES

Afriyie, D. O., Masiye, F., Tediosi, F., & Fink, G. (2023). Confidence in the health system and health insurance enrollment among the informal sector population in Lusaka, Zambia. *Social Science & Medicine*, *321*, 115750.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Alesane, A., & Anang, B. T. (2018). Uptake of health insurance by the rural poor in Ghana: determinants and implications for policy. *Pan African Medical Journal*, *31*(1), 1-10.

Boateng, R. (2024). Micro Level Analysis on Health Insurance Enrolment among Selected Women in Ghana: Barriers and Predictors. *Journal of Health Statistics Reports. SRC/JHSR-118*, *3*(1), 2-5.

Box, G., & Jenkins, G. (1970). Time series analysis: Forecasting and control. San Francisco: Holden-Day

Chatfield, C., & Prothero, D. L. (1973). Box-Jenkins seasonal forecasting: Problems in a case study. *Journal of the Royal Statistical Society: Series A (General)*, *136*(3), 295-315.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427-431.

Ghimire, P., Sapkota, V. P., & Poudyal, A. K. (2019). Factors Associated with Enrolment of Households in Nepal's National Health Insurance Program. *International Journal of Health Policy and Management*, *8*(11), 636–645. https://doi.org/10.15171/ijhpm.2019.54

Hyndman, R. J., & Athanasopoulos, G. (2018*). Forecasting Principles and Practice*. Melbourne: OTexts.

Jalalpour, M., Gel, Y., & Levin, S. (2015). Forecasting demand for health services: Development of a publicly available toolbox. *Operations research for health care*, *5*, 1-9.

Kathrin, D., Günther, I., & Harttgen, K. (2021). Using machine learning to predict health insurance enrolment and take-up of health services.

Kornelio, S., Balan, R., & Deogratias, E. (2024). Forecasting students' enrolment in Tanzania government primary schools from 2021 to 2035 using ARIMA model. *International Journal of Curriculum & Instruction*, *16*(1). 162-174.

Kotoh, A. M., Aryeetey, G. C., & Van der Geest, S. (2018). Factors that influence enrolment and retention in National Health Insurance Scheme. *International Journal of Health Policy and Management*, *7*(5), 443

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79-86.

Kusi, A., Enemark, U., Hansen, K. S., & Asante, F. A. (2015). Refusal to enroll in Ghana's National Health Insurance Scheme: is affordability the problem? *International journal for equity in health*, *14*, 1-14.

Lwaho, J., & Ilembo, B. (2023). Unfolding the potential of the ARIMA model in forecasting maize production in Tanzania. *Business Analyst Journal*, *44*(2), 128-139.

Marinova, G., & Todorova, M. (2023, November). Regression Analysis for Predicting Health Insurance. In *2023 4th International Conference on Communications, Information, Electronic and Energy Systems (CIEES)* (pp. 1-4). IEEE.

Morgan, A. K., Adei, D., Agyemang-Duah, W., & Mensah, A. A. (2022). An integrative review on individual determinants of enrolment in National Health Insurance Scheme among older adults in Ghana. *BMC Primary Care*, *23*(1), 190.

Naylor, T. H., Seaks, T. G., & Wichern, D. W. (1972). Box-Jenkins methods: An alternative to econometric models. *International Statistical Review/Revue Internationale de Statistique*, 123-137.

Ng'ang'a, E. W. (2021). *Determinants of Health Insurance Uptake Among Low-Income Populations in Kibera-Nairobi, Kenya* (Doctoral dissertation, University of Nairobi).

Nsiah-Boateng, E., & Aikins, M. (2018). Trends and characteristics of enrolment in the National Health Insurance Scheme in Ghana: a quantitative analysis of longitudinal data. *Global health research and policy*, *3*, 1-10.

Nyman, J. A. (2001). *The theory of the demand for health insurance* (No. 311). Discussion Paper.

Osei Afriyie, D., Krasniq, B., Hooley, B., Tediosi, F., & Fink, G. (2022). Equity in health insurance schemes enrollment in low and middle-income countries: A systematic review and meta-analysis. *International Journal for Equity in Health*, *21*(1), 21.

Putri, N. K., Laksono, A. D., & Rohmah, N. (2023). Predictors of national health insurance membership among the poor with different education levels in Indonesia. *BMC Public Health*, *23*(1), 373.

Ramasubramanian, V. (2007). Time series analysis. *New Delhi*: I.S.R.I

Sharma, R. (2023). Inequality and disparities in health insurance enrolment in India. *Journal of Medicine Surgery and Public Health*, *1*, 100009.

Soyiri, I. N., & Reidpath, D. D. (2012). Evolving forecasting classifications and applications in health forecasting. *International journal of general medicine*, 381-389.

Soyiri, I. N., & Reidpath, D. D. (2013). An overview of health forecasting. *Environmental health and preventive medicine*, *18*, 1-9.

Wiah, E. N., Buabeng, A., & Agyarko, K. (2022). Statistical Model for the Forecast of Electricity Power Generation in Ghana. *Open Journal of Statistics*, *12*(3), 373-384.

Yego, N. K. K., Nkurunziza, J., & Kasozi, J. (2023). Predicting health insurance uptake in Kenya using Random Forest: An analysis of socio-economic and demographic factors. *Plos one*, *18*(11), e0294166.

Yego, N. K., Kasozi, J., & Nkurunziza, J. (2021). A Comparative Analysis of Machine Learning Models for the Prediction of Insurance Uptake in Kenya. *Data, 6*(11), 116. https://doi.org/10.3390/data6110116