International Journal of Data Science and Analysis
2020; 6(1): 48-57
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20200601.16
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Comparative Analysis of Sarima and Setar Models in Predicting Pneumonia Cases in Kenya

Fredrick Agwata Nyamato*, Anthony Wanjoya, Thomas Mageto

Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya

Email address:
fredricknyamato@gmail.com (F. A. Nyamato), awanjoya@gmail.com (A. Wanjoya), ttmageto@gmail.com (T. Mageto)
*Corresponding author

To cite this article:
Fredrick Agwata Nyamato, Anthony Wanjoya, Thomas Mageto. Comparative Analysis of Sarima and Setar Models in Predicting Pneumonia Cases in Kenya. *International Journal of Data Science and Analysis*. Vol. 6, No. 1, 2020, pp. 48-57. doi: 10.11648/j.ijdsa.20200601.16

Received: February 24, 2020; **Accepted**: March 6, 2020; **Published**: March 18, 2020

Abstract: Kenya is a country located in Eastern part of Africa with approximate population of 46.5 million, with majority of the population constituting youths under the age of 35 years. The country has experienced increased morbidity rate arising from Pneumonia disease like other countries all over the world. As per recent studies 2 million children lose lives from pneumonia disease [1]. This study applies two models, one is linear model Seasonal autoregressive model (SARIMA) and another is a non-linear model called self-Excited Threshold Autoregressive (SETAR) in projection of cases in Kenya. Data for usage for purpose of this study was obtained Ministry of Health of Kenya of a period of 20 years from January 1999 to December 2018. The data collected is seasonal the number of case from period to period depending on climatic condition. Although both models performs well in pneumonia projection, non-linear SETAR models outperforms linear SARIMA. By carrying out a comparative analysis by use of Diebold-Mariano test, which revealed that there were no significant difference in the forecasting performance of the two models. The best model identified between the two models i.e. SETAR which best fit the data, can be applied in predicting pneumonia cases beyond the period under consideration. Other studies can be carried to come up with a model for every specific region in the country, to assist in resources allocation to specific parts of the country.

Keywords: Seasonal Autoregressive Integrated Moving Average, Self-excited Threshold Autoregressive, Stationarity and Linearity

1. Introduction

Pneumonia is characterized by an acute infection of the lungs, which results in coughing, fever, chills, muscle aches, and difficulty in breathing to its victims. It's also a leading killer in children globally [1]. Between 11 million and 20 million children with Pneumonia will require hospitalization and more than 2 million will die. Pneumonia accounts for approximately 1.9 million deaths globally in children under five each year. In the year 2011, there were an estimated 120 million episodes of childhood pneumonia globally of which 14 million progressed to severe disease with 1.3 million deaths [2]. Most of these deaths (81%) occurred in children under 2 years of age. The incidence and the severity of childhood Pneumonia were higher in Africa and South East Asia, which account for 30% and 39% respectively of the global burden of severe causes. First infections of Pneumonia cannot be traced to a specific period in history, mention of the disease found in early Greek Pneumonia infection has remained a serious medical concern throughout the global community despite a new breakthrough in its treatment and management. Millions of people continue to be hospitalized and losing lives due to their infections across the world. The World Health Organization (WHO) [3] estimated that there were more than 150 million cases of pneumonia each year and killing 1.6 million which accounts for 19% of all deaths worldwide. Developing countries have recorded the highest number of deaths caused by Pneumonia, Kenya ranking among the top 15 highest affected countries. The primary causative agent of Pneumonia is known by its scientific name Streptococcus pneumonia, through understanding this causative agent and reviewing strategies which have been deployed to manage the disease on a global scale will be important in reducing deaths caused it. Through this effort, the world will achieve easy access and efficiency in its treatment and eventually reduce its detrimental effects. Due

to the high incidence of Pneumonia death revealed on existing literature especially under five which can have a consequence on population growth and productivity in the future, there is the need for stringent measures to curb the trend. The availability of precise estimates and projections is crucial in supporting decisions and policymakers in planning and in developing programs to facilitate and improve health care programs designed to curb the negative impact of the disease. Projecting the future prevalence and its impact requires a sound methodology for projecting the number of future Pneumonia infections and determining the impact of those infections on the future pattern of adult and child deaths. Institutions need to devote billions of dollars to the health sector in order to resolve this problem. Therefore, knowing the pattern of this disease could aid world health bodies to plan and develop policies that could be used to reverse the growing trend of this killer disease. Hence this study compares two-time series models in predicting pneumonia cases in Kenya.

2. Methodology

2.1. Study Area and Data Source

Kenya is a state located in Eastern part of Africa, divided into 47 counties, with a population of 47million people as per census done in 2019. The data for this study was obtained from the Ministry of Health of Kenya.

2.2. Identification of Pneumonia Trend

The trend of a series reflects the long term growth of the time series over time. A time-series variable may exhibit different types of trend; the linear, linear constant growth, quadratic and quadratic constant growth among others. This study will evaluate the above different types of trend models for the disease under consideration

2.3. Test of Stationarity

The unit root test will be applied to test if the data under investigation is weakly stationary. A unit root test is performed to determine whether a stochastic or a deterministic trend is present in the series. When the roots of the characteristic equation lie outside the unit circle, then the series is considered stationary. We study employs the ADF test [4] to determine whether the disease involved contained a unit root (non-stationary) or has stationary covariance. The test statistic for the ADF test is given by:

$$F_r = \frac{\hat{\delta}}{SE(\hat{\delta})} \tag{1}$$

Where SE (δ) is the standard error of the least square estimate of $\hat{\delta}$ The null hypothesis is rejected if the test statistic is greater than the critical value.

2.4. SARIMA Model

There are five types of traditional time series models most commonly used in epidemic time series forecasting; Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), and Seasonal Autoregressive Integrated Moving Average (SARIMA) models. AR models express the current value of the time series linearly in terms of its previous values and the current residual, whereas MA models express the current value of the time series linearly in terms of its current and previous residual series. The ARMA model is a combination of two models Autoregressive (AR) and Moving Average (MA) models forming ARMA ((p, q) model, where p and q are the orders of the AR and MA processes respectively [5]. In ARMA the current value of the time series is expressed linearly in terms of its previous values and in terms of current and previous residual series. The ARMA model is given has:

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots \Phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots \theta_Q \varepsilon_{t-Q} \tag{2}$$

Where p and q are parameters of the autoregressive and moving average components respectively, i=1, 2. p and j=1, 2. q.

The ARMA (p, q) process is stationary if the roots of the polynomial in the AR component are less than one in absolute terms. On the other hand, the process is invertible on the condition that the absolute values of the roots of the polynomial in the MA component are less than one.

To incorporate the integrated component to carter for time-series data that are non- stationary in nature we come up with a model called Autoregressive Integrated Moving Average (ARIMA) Model. In practice, many time series data show non-stationary behavior and such data are made stationary by applying finite differencing of the data points. When a time series data exhibit seasonal behavior, the ARIMA model is usually not able to capture the behavior along the seasonal part of the series, hence, the tendency for wrong order selection for the non-seasonal component. Identification of relevant models and inclusion of suitable seasonal variables is therefore necessary when a time series data exhibit periodic patterns. The SARIMA model, therefore, has the advantage of capturing both seasonal and non-seasonal components. The general expression for the order of a SARIMA model is: ARMA (p, d, q) (P, D, Q) S and can be expressed using the backshift operator as:

$$\phi(B)\Phi(B^S)(1-B)^d(1-B^S)^D Y_t = \theta(B)\Theta(B^S)\varepsilon_t \tag{3}$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^P \tag{4}$$

$$\Phi(B^S) = 1 - \phi_1 B^S - \phi_2 B^{2S} - \cdots - \phi_p B^{PS} \tag{5}$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{6}$$

$$\Theta(B^S) = 1 + \Theta_1 B^S + \Theta_1 B^{2S} + \cdots + \Theta_q B^{qS} \tag{7}$$

Where Y_t represents the time series data at period t
B denotes the backshift operator
ε_t is a sequence of i. i. d variables with mean zero and

variance σ^2, s is the seasonal order Φ_i and Φ_j are the non-seasonal and seasonal AR parameters respectively

Θ_i and Θ_j are respectively non-seasonal and seasonal MA parameters.

P, d and q denote the non-seasonal AR, I and MA orders respectively and P, D and Q respectively represent the seasonal AR, I and MA orders respectively.

2.5. SETAR Model

Self-Excited Threshold Autoregressive (SETAR) model is a class of the Threshold Autoregressive (TAR) model proposed by Tong [6] and further studied by Tong and Lim [7], and later by Tong in his study of threshold model [8] and in study non-linear time series [9]. The SETAR model is a set of different linear AR models, changing according to the value of the threshold variable (s) which is the lagged values of the series. The process is linear in each regime, but the movement from one regime to the other makes the entire process nonlinear. The two regime version of the SETAR model of order p is given as by [10]:

$$y_t = \begin{cases} \emptyset_0^{(1)} + \sum_{i=1}^{p(1)} \emptyset_i^{(1)} y_{t-i} + \varepsilon_t^{(1)} \; if \; y_{t-d} \le r \\ \emptyset_0^{(2)} + \sum_{i=1}^{p(2)} \emptyset_i^{(2)} y_{t-i} + \varepsilon_t^{(2)} \; if \; y_{t-d} > r \end{cases} \quad (8)$$

The where ϕ_1^i and ϕ_2^i are the coefficient in lower and higher regime respectively which needs to be estimated; r is the threshold value; $p^{(1)}$ and $p^{(2)}$ are the order of the linear AR model in low and high regime respectively. In this work, the order of the AR model in both regimes are equal y_{t-d} is the threshold variable that governs the transition between the two regimes with d being the delay parameter which is a positive integer (d < p); $\{\varepsilon_t^{(1)}\}$ and $\{\varepsilon_t^{(1)}\}$ are sequence of independently and identically distributed random variables with zero mean and constant variance (i.e. i. i. d $(0,\sigma_\varepsilon^2)$). In this study, we consider two regime SETAR model which can be written in its simplest form as SETAR (2; p, d). The properties of the general SETAR model are hard to obtain and little is known about the condition under which the SETAR models generate time series that are stationary [11]. Such conditions have only been established for the first-order SETAR model. For effective model selection, we follow the procedure discussed in [11]. The approach of SETAR modeling starts with AR (p) model specification and linearity against the SETAR model, SETAR model identification, estimation and evaluation of the selected model and then forecasting which is precisely discussed as follows.

2.6. Linearity Test

To apply the SETAR model to an observable time series, the series must first be nonlinear. That is the existence of nonlinear behavior in the series must first be checked. In testing for the linearity in the series, we first have to specify an appropriate linear AR (p) model for the series under consideration. The choice of the maximum lag order is based on the autoregressive lag order that minimize the AIC value, [11]. After determining the linear AR (p) model we then test for linearity using a well-known linearity test such as Keenan Test. Keenan test [12] was is applied to detect nonlinearity in an observable time series. The test is considered as a special case of the RESET test [13]. The avoidance of multi-collinearity makes it special. The Keenan test for nonlinearity analogous to Turkey's one degree of freedom for non-additivity test is motivated by approximating a nonlinear stationary time series by a second-order Volterra expansion which is given by:

$$y_t = u + \sum_{u=-\infty}^{\infty} \theta_u \varepsilon_{t-u} + \sum_{v=-\infty}^{\infty} \sum_{u=-\infty}^{\infty} \theta_{uv} \, \varepsilon_{t-u} \varepsilon_{t-v} \quad (9)$$

Where $\{\varepsilon t \; -\infty < t < \infty\}$, is a sequence of independent and identically distributed with zero mean random variable. The process $\{yt\}$ is linear if the double sum of the right-hand side the equation does not exist. Thus we can test the linearity of the time series by testing whether or not the double sum of the equation does not exist. That is, the test requires that one distinguish between linearity versus a second-order Voltera expansion, by examining $\theta uv = 0$ as well as the coefficients on higher orders. Keenan's test is equivalent to testing if $\eta = 0$ in the multiple regression model [14] (with the constant 1 being absorb in to θ_0):

$$y_t = \theta_0 + \phi_1 y_{t-1} + \cdots + \phi_m y_{t-m} + \eta \hat{y}_t^2 + \varepsilon_t \quad (10)$$

The Keenan's test statistic for the null hypothesis of linearity $(Ho{:}\eta = 0)$ is given as

$$\widehat{F} = \frac{\eta^2 (n-2m-2}{RSS-\eta^2} \quad (11)$$

Where

m=lag order of the linear autoregressive process

n=same size considered

RSS=the residual sum of squares from the AR (m) process

When the null hypothesis is satisfied, $\hat{F}$ is approximately F-distributed with 1 and $n - 2m - 2$ degree of freedom. The null hypothesis of linearity is rejected if the p–value associated with is small ($p - value \; <$) or when the value of F^ is greater than the selected critical value of the F-distribution with 1 and *n – 2m – 2* degrees of freedom.

2.7. The Box-Jenkins Methodology

Named after George Box and Gwilym Jenkins, it applies four steps in developing and application of the model developed i.e. model identification, estimation, diagnostic checking and model identification that best fits past values data [15]. The main steps in setting up the model are as follows:

2.7.1. Model Identification

For identification of parameters of SARIMA, we use sample ACF and PACF, which is obtained from training data that should match with the corresponding theoretical or actual values [15]. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can also be utilized in the identification of parameters [16]. For SARIMA the model that we obtain the smallest AIC, BIC and Akaike Information Criterion corrected (AICc) is the best mode:

1. Akaike's Information Criteria (AIC) is given as

AIC=-2 ln (maximum likelihood) + 2k Where k is the number of parameters in the model. AIC will be high if the number of parameters are high.

2. Bayesian Information Criteria (BIC) given as

BIC=- 2 ln (maximum likelihood) + k ln (n) Where n is the number of observations in the given stationary time series data and, k is the number of parameter.

3. Akaike Information Criterion corrected (AICc), which is given as

$$AICc = AIC + \frac{2K(K+1)}{n-k-1} \quad (12)$$

An optimal model is one that minimizes AIC and BIC criteria. The difference between the BIC and the AIC is the greater penalty imposed for the number of parameters by the former than the latter. Both criteria are correct depending on the goal and set of assumptions. For the best results, both criteria are applied. The graphical procedure is mostly applied in identifying d which involves plotting the data over time and the corresponding (Partial) autocorrelation function. For a non- stationary model it is expected the ACF not to decrease to zero or to exhibit a very slow decay. A time-series to be considered stationary it has to have a constant mean, variance and covariance statistical characteristics over time. The sample autocorrelation function is zero for lags beyond q if the model under consideration is an MA (q) model. To determine the order of autoregressive models a different function is needed since AR (p) does not turn into zero after a certain number of lags of ACF since the model attenuates instead of a cutoff. A function of such form is described as a correlation between Yt and Yt-k, excluding the effect of the intervening variables. To determine the order of AR (p) model we apply partial autocorrelation function. For an AR (p) model, the PACF drops off to zero after the pth lag [17]. In practice, identifying p and q using the ACF and PACF involves a trial and error approach, with more or less subjectivity in interpreting these functions.

2.7.2. Estimation and Information Criteria

This is the next step after the order of the model has been identified which entails parameter estimation of the models. The parameter of the SARIMA model is estimated from the observed time series through the use of either linear least squares method, Maximum likelihood estimation, and Method of moments.

The method of moments is easy to calculate as compared to the maximum likelihood but it is not efficient than maximum likelihood method. Unlike other approaches, maximum likelihood estimation offers a unified approach for parameters estimation of SARIMA.

Also, it provides a standard way to deal with models of stochastic time Processes. Observation in time series are interrelated, the likelihood approach through the use of probability density function is obtained by: Assuming the error follows white noise, i.e. $\varepsilon \sim N(0; \sigma^2)$ then the joint probability distribution function is given $f(\varepsilon_1, \varepsilon_2, \ldots. \varepsilon_n) = f(\varepsilon_1) f(\varepsilon_2) \ldots \ldots f(\varepsilon_n)$

Instead of $f(Y1;\ Y2, \ldots\ldots Yn)$ due to dependency between time series observation which will not be written as a multiplication of marginal probability density functions. A stationary general ARIMA (p, q) process, therefore, can be given as [18]:

$$\widehat{Y_t} = \beta_1 \hat{Y}_{t-1} + \beta_2 \hat{Y}_{t-2} + \cdots + \beta_p \hat{Y}_{t-p} + \varepsilon_t - \theta_1 \varepsilon_t - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (13)$$

Where $\widehat{Y_t} = Y_t - \mu$

The probability distribution function of errors is given by:

$$f(\varepsilon_1, \varepsilon_2, \ldots. \varepsilon_n \mid \mu, \beta, \theta, \sigma_e^2) = 2\Pi\sigma_e^2)^{\frac{-n}{2}} \left\{\frac{-1}{2\sigma_e^2}\right\} \sum_{t=1}^{n} \sigma_t^2 \quad (14)$$

2.7.3. Diagnostic Checking

The stage its primary objective is to check the goodness of fit of the model identified through the iterative process [14]. It's important if the model can be improved to ensure it makes meaningful inferences. Adequacy of the model is assessed through checking if it satisfies underlying assumptions after the parameters have been estimated. For the model to be considered appropriate it should extract all relevant information. For example, residuals obtained should be small and uncorrelated having zero mean and constant variance. Diagnostic checking in the Box-Jenkins methodology primarily involves testing the statistical properties of the error terms (normality assumption, weak white noise assumption) if they are satisfied.

1. Residual Analysis: Residual is the difference between the observed value and the predicted value. Residuals obtained should nearly attain white noise properties. If the properties are met then the model identified is appropriate and parameters under estimation are close to true values. If the model doesn't meet these properties then it needs improvement.
2. Normality and independence: Through the application of histograms and quantile-quantile (Q-Q) plot, normality assumption can be checked for residuals while the run test can be employed to check independence.
3. Ljung-Box Test: The test that is used to determine the presence or absence of auto-correlation in a time series up to a certain lag. The test statistic is given by [19]:

$$Q(K) = n(n+2) \sum_{i=1}^{k} \frac{r_j^2}{n-j} \quad (15)$$

Where r_j is the j_t residual autocorrelation while n is the total number of data points or number of residual and k is the total number of lags tested. The decision criteria is that null hypothesis is rejected if Q (k) is greater than chi-square table value.

4. Residual autocorrelation and Partial autocorrelation Function: The residuals of
5. ACF and PACF should not be forecastable, that is the terms of the residual ACF and residual PACF should all approximately lie between the 95% confidence limit. If this is not the case, there are elements of residuals

which can be forecastable.

2.7.4. Diagnostic Checking

Forecasting involves the application of the model identified to historical data to predict a variable of interest, the procedure requires routine calculations to make use of a large number of events [20]. Through forecasting, we are able to achieve our main objective in dealing with modeling exercise that is able to predict the value of the random variable in the future from the currently existed one and get information in advance. The best forecast is achieved when we obtain a minimum mean square error, whose forecast is given by:

$$\hat{Y}_t(l) = Ezzz\left(\hat{Y}_{t+1}\middle|Y_1, Y_2, \dots \dots Y_t\right) \quad (16)$$

Where $\hat{Y}_t$ is the minimum mean square error forecast and Y_1, Y_2..... Y_t is the observed time series data The Choice of the model may rely on the goodness of fit of the information criteria or the residual mean square error. The decision on the criteria to apply will depend on the main objective of the model if the objective is forecasting future value using current and past values then model selection criteria can be based on forecast error. The comparison of the forecast error measures help us to know how much we should rely on the chosen prediction method is based on the following statistics.

1. Mean Percentage Error (MPE), which is given by

$$MPE = \left(\frac{1}{K}\sum_{I=1}^{K}\frac{e_1}{Y_{n+1}}\right) \quad (17)$$

2. Mean Square Error

$$MSE = \left(\frac{1}{K}\sum_{i=1}^{K} e_i^2\right) \quad (18)$$

3. Mean Absolute Error

$$MAE = \frac{1}{K}\sum_{i=1}^{K}|e_i| \quad (19)$$

4. Mean Absolute Percentage Error

$$\left(\frac{1}{K}\sum_{i=1}^{K}\left|\frac{e_i}{Y_{n+1}}\right|\right) \quad (20)$$

The best model for forecasting is the one that results into smallest MPE, MSE, MAE and MAPE.

2.8. Diebold-Mariano Test

To compare the forecasting accuracy SETAR and SARIMA models, lower values of mean square errors of one forecast in comparison to the alternative do not necessarily translate into the superiority of this forecast. In order to verify whether there is a significant difference in the forecasting accuracy of any two competing models, the Diebold and Mariano test [21] of equal forecasting accuracy will be used to assess whether the differences in the mean square errors of competing forecasts are statistically significant. The test statistic follows the standard normal distribution and tests the null hypothesis of equal forecast accuracy against the alternative.

$$s_1 = [\hat{v}(\bar{d})]^{\frac{-1}{2d}} \quad (21)$$

Where $\bar{d}$ is the mean of the coefficient of d_t, which is the difference between the sets of squared forecast errors from two competing models,

$$d_t = \ell_{1t}^2 - \ell_{2t}^2 \quad (22)$$

$\hat{v}(\bar{d})$ is an estimate of the variance of $\bar{d}$

3. Results and Discussion

3.1. Data Overview

The maximum and minimum values of the cases for the entire study period were 148,272 and 8,632 respectively. Moreover, the average pneumonia cases were 66,906.85. The coefficients of variation (CV) for the pneumonia cases were 55.98%. Pneumonia cases recorded for the entire period was found to be positively skewed. The nature of trend characterizing the pneumonia cases overtime was investigated using the linear, quadratic, log-linear and log-quadratic trend models as shown in Table 1.

Table 1. *Trend analysis of Pneumonia case.*

Model	AIC	BIC
Linear	1275.256	1281.75
Quadratic	1165.406	1175.147
Log-linear	281.328	287.822
Log-quadratic	209.793*	219.534*

*: Means best based on the selection criteria.

The log quadratic trend model was observed as the best since it had the least AIC and BIC. The parameters of the log-quadratic trend models for the pneumonia cases were estimated as shown in Table 2. All the parameters were highly significant at the 5% level of significance. It was also shown that the estimated log-quadratic model for the cases trends downwards and is quadratic in logarithm form. It, therefore, indicates that the presence of trend was the major cause of the variation in the pneumonia cases. Thus, the estimated log-quadratic trend model for pneumonia cases is given by;

$$\ln pnc = 2.325 + 0.019t - 0.000tt^2 \quad (24)$$

Table 2. *Estimated parameters of the Log-quadratic trend.*

Variable	Coefficient	standard error	T- statisti	P-value
Constant	2.3536	0.0917	25.6495	0.000**
Time	0.0191	0.0022	8.6069	0.000**
$(Time)^2$	0.0001	0.0001	9.4009	0.000**

AIC=209.7925.
BIC=219.5336.

3.2. Fitting the SARIMA Model

A visual inspection of the ACF plot of the pneumonia cases showed a slow decay in the ACF suggesting non-stationarity of the series. The PACF plot also revealed very dominant significant spikes at lag 1 as shown in Figure 1.

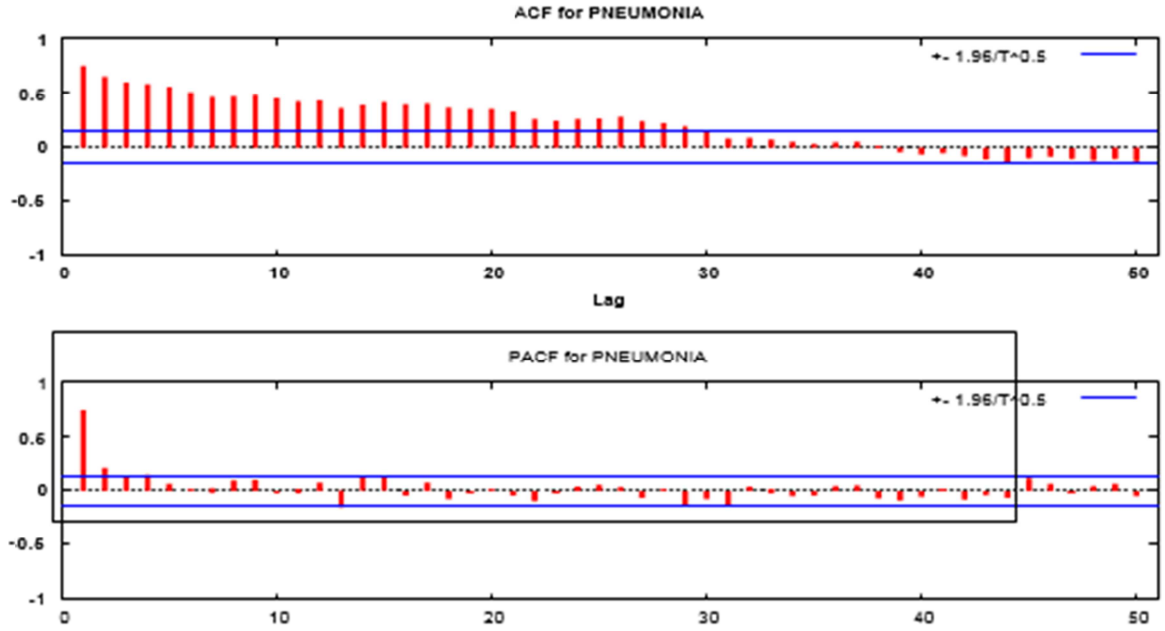

Figure 1. *ACF and PACF plot of Pneumonia cases.*

To further confirm the non-stationarity of the series, the KPSS and ADF test for unit root was carried out on the original data. Using the KPSS test, the results in Table 3:

Table 3. *KPSS test of Pneumonia cases.*

Test	Test Statistic	Critical value
KPSS	0.708775	0.464

The ADF test also confirms the existence of unit root with only a constant term and a constant with the quadratic trend. This affirmed the presence of unit root in the series since the p-value was greater than the 0.05 level of significance as illustrated in Table 4.

Table 4. *ADF test of Pneumonia cases.*

Test	Constant		Constant+quadratic Trend	
	Test statistic	P-vale	Test statistic	P-vale
ADF	-1.4342	0.567	-1.4314	0.852

The series was transformed logarithmically to stabilize the variance. The transformed series was then differenced and then tested for stationarity. The KPSS and ADF tests for the pneumonia cases revealed that the transformed differenced series were now stationary since the p-value for the ADF test is less than the 5% significance level and the test statistic being less than the critical value in the case of the KPSS test as shown in Tables 5 and 6 respectively.

Table 5. *KPSS test of differenced series.*

Test	Test statistic	Critical value
KPSS	0.0341	0.464

Table 6. *ADF test of differenced series.*

Test	Constant		Constant+quadratic Trend	
	Test statistic	P-vale	Test statistic	P-vale
ADF	-5.4783	0.000	-5.4829	0.000

After obtaining the order of integration of the Pneumonia cases, the order of the Autoregressive and Moving Average components was determined based on the ACF and PACF plots [15]. The ACF plot in Figure 2 shows significant spikes at lag 1, 6, 12 and 13. The PACF plot also has significant spikes at lag 1, 6, 13 and 19. Using the lower significant lags of both the ACF and PACF, tentative SARIMA models were developed as shown in Table 7. Among these possible models SARIMA (1, 1, 1) (0, 0, 1) 12 was adjudged the best since it had the least AIC, AICc and BIC values as compared to the other models.

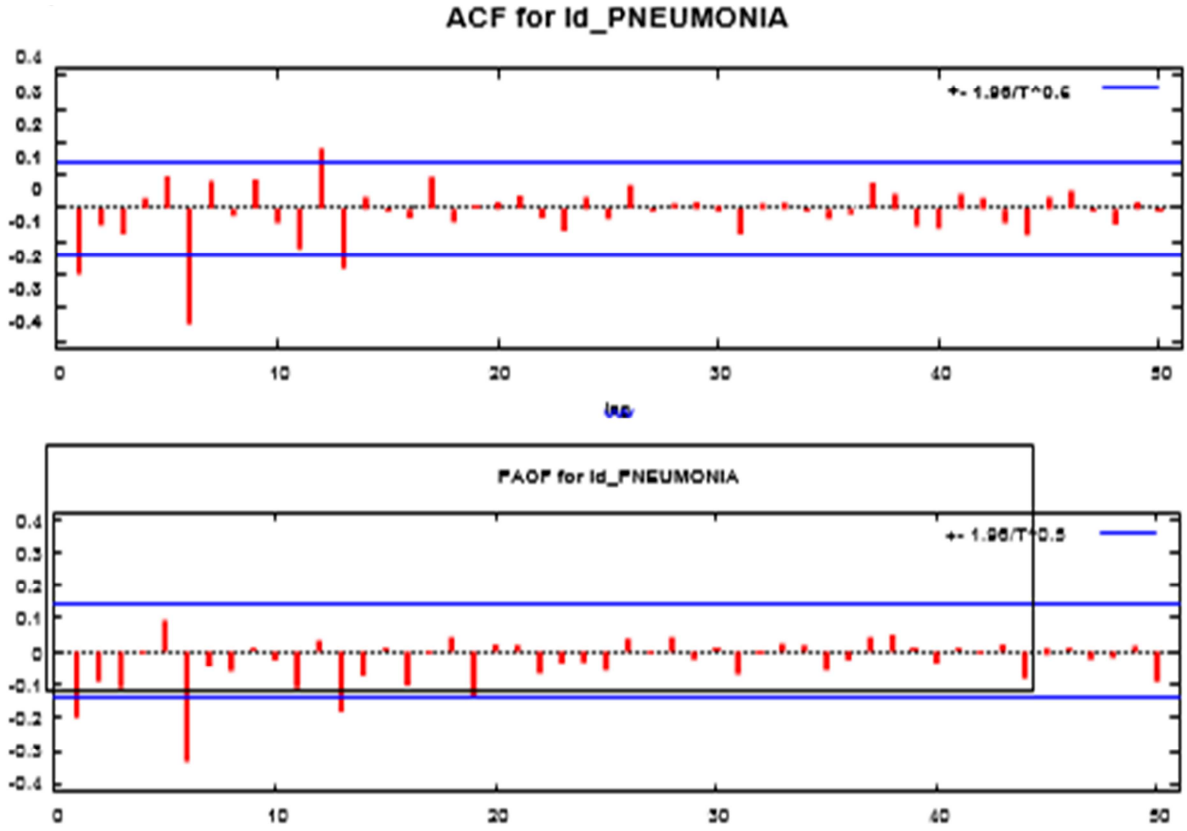

Figure 2. *ACF and PACF plot of differenced series.*

The SARIMA (1, 1, 1) (0, 0, 1) 12 model is the best as per results obtained and can be expressed in terms of backshift operator as;

$$(1 - 0.755B)(1 - B)\ln pnc = (1 - 0.959B)(1 + 0.169B^{12})\varepsilon_t \quad (25)$$

From Table 7 that the p-values of the parameters of the selected model for the Autoregressive and Moving Average components were highly significant at the 5% level of significance. The model thus appears to be the best model among the suggested models.

Table 7. *Estimates of parameters for SARIMA (1, 1, 1) (0, 0, 1) 12.*

Variable	Coefficient	Standard error	Z-statistic	P-value
θ_1	-0.859552	0.0327374	-29.3106	0.00001
Θ_1	0.168451	0.0764908	2.2153	0.0267
ϕ_1	0.694616	0.0689403	10.9459	0.00001

To ensure that the fitted model is adequate, Ljung-Box test was performed. It revealed that the model was free from serial correlation and conditional heteroscedasticity at lag 12, 24, 36 and 48 respectively since the p-values of the test statistics were insignificant at the 5% significance level. This implies that the residuals of the model were uncorrelated, thus have zero mean and constant variance over time; hence are white noise series. It can, therefore, be concluded that the selected model, SARIMA (1, 1, 1) (0, 0 1) 12 is the best model since it satisfies all the diagnostic conditions.

Table 8*. Residuals diagnostic test for SARIMA (1, 1, 1) (0, 0, 1) 1.*

Lag	Ljung-Box Test	
	Test statistic	P-value
12	20.8664	0.05237
24	28.2171	0.251
36	31.3243	0.6905
48	38.693	0.8288

3.3. Fitting the SETAR Model

The 2 regime Self Excited Threshold Autoregressive (SETAR) model approach was used to model and forecast the pneumonia cases. To model a time series with the SETAR model, the series must be non-linear. To test for non- linearity in the series we first specify the linear AR (p), model. Using AIC, we found the AR (4) model for the series. The choice of the AR (4) lag order is based on the Autoregressive lag order that gives the minimum AIC value based on the significant PACF lag orders. After we determined the linear AR model we employ the Keenan1-degree test to test for linearity against the alternative of nonlinearity for the Keenan test. This linearity test depends on the linear AR model selected. Table 9 below summarizes the results from the Keenan1-degree test. From the results, in the Keenan1-degree test, we reject the null

hypothesis of linearity since the P-value is less than the 5% significant level.

Table 9. *Linearity test.*

Test	Test statistic	P-value	Decision
Keenan1-degree	6.36	0.02	Linearity Rejected

After checking if data is nonlinear, we proceed to obtain the SETAR model that best fits the data. We do this by determining the Autoregressive lag order P in each regime and the threshold variable where d represents the delay parameter. We choose the model with P lag order for both regimes and threshold variables with the minimal AIC value by performing a grid search on all possible combinations of SETAR models that can be fitted to the data.

After performing a grid search on all possible combinations of SETAR models that can be fitted to the data, SETAR (2; 4, 3) model with a threshold variable y_{t-3} could be appropriate to explain the nonlinearity in the data. This model has a minimum AIC value which is presented in Table 10.

Table 10. *AIC for the selected SETAR Model.*

Model	AIC	BIC
SETAR (2: 4, 3)	-165.42	91.07

After we have found that SETAR (2; 4, 3) model with threshold variabley_{t-3}) as the best model that fits the data well since it has the minimum value for AIC. Further assessment of the forecast-ability of the model was done. The corresponding model for SETAR (2: 4, 3) with a threshold variable y_{t-3} that governs the transitions between the two regimes with delay parameter 3 and threshold value 1.255 is given by:

$$y = \begin{cases} 0.295 + 0.727y_{t-1} + 0.008y_{t-2} - 0.142y_{t-3} + 0.087y_{t-3} & if\ y_{t-3} \leq 1.255 \\ -0.069 + 0.484y_{t-1} + 0.085y_{t-2} + 0.300y_{t-3} & if\ y_{t-3} \geq 1.255 \end{cases} \quad (26)$$

After the parameters of the SETAR model have been estimated, we check the residuals of the model for the best fit. That is we check for the non-existence of serial autocorrelation, zero mean and constant variance of the residuals. We used the ARCH-LM test to check for a constant variance of the residuals. Ljung-Box test was also used to check for serial correlation. From the results, as shown in Table 11, we fail to reject the null hypothesis of the two tests for the SETAR (2; 4, 3) model since their P-values were greater than the 5% significant level.

Table 11. *Residuals diagnostic test for SETAR (2; 4, 3).*

ARCH-LM			Ljung Box Test	
lag	Test Statistic	p-value	Test Statistic	p-value
12	17.9482	0.1085	14.6184	0.263
24	37.9226	0.2782	32.8826	0.1066
36	38.0165	0.3357	41.1826	0.2542
48	46.2422	0.6276	50.6098	0.3709

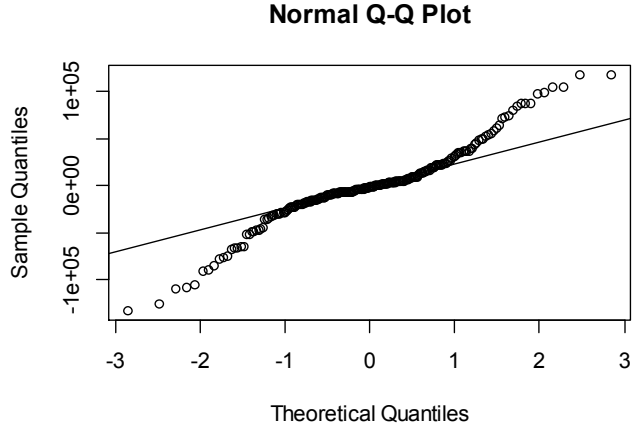

Figure 3. *Normal Q-Q plot.*

3.4. Comparative Analysis of the Models

The residuals have been assumed to be normally distributed throughout the analysis. Quantile-Quantile plots (QQ) plots are an effective tool for assessing the normality of residuals. From the plot in Figure 3 it can be easily observed that the Q-Q plot is approximately normally distributed.

If models satisfy all the assumptions, we can conclude that the models are adequate and can be used to predict the pneumonia cases. Hence, there is the need to compare the forecasting accuracy of the SARIMA (1, 1, 1) (0, 0, 1) 12 model with SETAR (2; 4, 3) model. From Table 12, it can be revealed that most accuracy tests support SETAR (2; 4, 3) model which has the minimum value of BIC, AIC, MSE, RMSE, and MAPE respectively:

Table 12. *Forecast accuracy test of models.*

Model	BIC	AIC	MSE	RMSE	MAPE
SARIMA (1, 1, 1) (0, 0, 1) 12	96.08	-233.77	0.0797	0.128	8.735
SETAR (2; 4, 3)	90.06*	-768*	0.000245*	0.01566*	0.09025*

Though the nonlinear SETAR model outperforms the linear SARIMA model as suggested by the forecast measures, it is interesting to know whether there is a significant difference in the forecast from the two models. Using the approach of Diebold and Mariano test, we test the null hypothesis that there is no difference between the forecast

accuracy from the two models against the alternative hypothesis that the selected SETAR provide better forecast accuracy as compared to the selected SARIMA model. The results from the test as presented in Table 13 fail to reject the null hypothesis of equal forecast accuracy at 5% level of significance and conclude that the forecast results from both models are the same.

Table 13. *Diebold-Mariano test.*

Test statistic	P-value
0.9856	0.3256

The developed models were cross-validated using the chi-square goodness of fit test. The results, as shown in Table 14 revealed that there is no significant difference between the observed pneumonia cases and their forecasted values. This can be seen from the insignificant chi-square statistic obtained for the results of both models. This indicates that the fitted models produce values that depict the behavior of the pneumonia cases over time even though the values of the observed and expected are not exactly the same:

Table 14. *Chi-square Goodness of Fit Test of the Models.*

Model	Chi-squared Statistic	p-value
SARIMA	0.9705	0.9142
SETAR	0.1819	0.9961

It can, therefore, be concluded that both models are good for predicting the pneumonia cases since there is no significant difference in their forecasting accuracy. The two models were therefore used to predict the cases of pneumonia. The predicted values for SARIMA (1, 1, 1) (0, 0, 1) 12 model indicates that pneumonia cases are increasing while SETAR (2; 4, 3) model gives a constant pattern of the cases over the forecast period. The predicted values for the models fall within the confidence interval. Hence, we say both models are adequate to be used for predicting pneumonia cases. The indication that the confidence interval becomes wider as the number of forecast increases suggests that the data was highly deterministic as evidence from the predicted values

4. Conclusion and Recommendation

In this study, the monthly number of patients with pneumonia cases, from January 1999 to December 2018 was studied. Before fitting the model to the pneumonia cases, the monthly characteristics of the series were examined. The careful examination of the series revealed that pneumonia cases were decreasing at a constant quadratic rate. The two models developed for predicting the monthly pneumonia cases were both adequate for representing the series as evident from all the diagnostics and model comparison techniques employed in the study. However, based on the forecast assessment from the linear SARIMA and the non-linear SETAR model, the forecast measures suggest that the non-linear SETAR model outperforms the linear SARIMA model. Also, the forecast performance of the non-linear SETAR models is superior to that of the linear SARIMA model in predicting pneumonia cases in Kenya. Predicted Pneumonia cases were made beyond the period under consideration based on the developed models. The Ministry of Health (MOH), and other stakeholders in the health sector can also predict pneumonia cases based on the developed models. There is, however, the need for continuous monitoring of the forecasting more reliable. Based on the findings of this research work, the following recommendations can be made;

i. The results revealed that the non-linear SETAR Model outperforms the linear SARIMA Model in predicting pneumonia cases in the region. It is therefore recommended that this study should be carried out in other regions to monitor the performance of the two models in predicting Pneumonia cases.
ii. The log-quadratic trend model depicts decreasing levels in the number of pneumonia cases for a unit change in time. These decreasing levels do not warrant public health workers to suggest that pneumonia cases are not prevalent in the region. It is rather recommended that the MoH should collaborate with health personnel to provide intensive education on some of the dangers of the disease and the need to seek early treatment in any nearby health facility because there can be a reverse trend of the cases.
iii. This study compared the non-linear SETAR model and the linear SARIMA model in predicting pneumonia cases in Kenya. It is therefore recommended that further studies should be carried out by comparing the non-linear SETAR model with other linear models to see which one would outperform the other since the non-linear SETAR model is the best model in this study.
iv. It is also recommended that the MoH advise the heads of its various institutions in the country to make data on pneumonia cases available. This will make it possible for researchers to study and predict pneumonia cases ahead of time for policy formulation and implementation to avert future loss of lives.

References

[1] Black, R., Sazawal, S., Clad, W., (2003). Effect of pneumonia case management on mortality in neonates, infants, and preschool children: a meta-analysis of community-based trials Lancet Infectious Diseases, 3: 547-556.

[2] Zar, H. J., Madhi, S. A., Aston, S. J., Gordon, S. B., (2013). Pneumonia in low and middle income countries: progress and challenges, Thorax, 68: 1052–1056.

[3] WHO (2013). Report on Child Health. Accessed date: 01/26/13.

[4] Dickey, D. A., and W. A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association 74: 427–431.

[5] Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). Time series analysis. New Jersey: Prentice Hall.

[6] Tong, H., (1978). On a threshold model. In Pattern Recognition and Signal Processing, edited by Chen, C. H. Amsterdam: Kluwer.

[7] J. Tong, H., and Lim, K. S. L., (1980). Threshold autoregression limit cycles and cyclical data. Journal of the Royal Statistical Society Series, B 42: 245-292.

[8] Tong, H., (1983). Threshold Models in Non-linear Time Series Analysis, Springer: New York.

[9] Tong, H., (1990). Non-linear Time Series: A Dynamical Systems Approach, Oxford: Oxford University Press.

[10] Boero, G., and Marrocu, E., (2004). The performance of SETAR models: A regime conditional evaluation of point, interval and density forecasts. International Journal of Forecasting, 20: 305-320.

[11] Frances, H. P., and Van Dijk, D., (2000). Nonlinear Time Series in Empirical Finance. Cambridge: Cambridge University Press.

[12] . Keenan, D. M., (1985). A Tukey non-additivity-type test for Time Series Nonlinearity, Biometrika, 72: 39-44.

[13] Ramsey, J. B., (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. Journal of the Royal Statistical Society Series B 31: 350–371.

[14] Cryer, J. D., and Chan, K. S., (2008). Time Series Analysis with Applications in R. 2 EdSpringer Science+Business Media, LLC, NY, USA.

[15] Box, G. E. P., a n d Jenkins, G. M., (1976). Time Series Analysis Forecasting and Control. Holden – Day, San-Francisco.

[16] J. M. Kihoro, R. O. Otieno, C. Wafula, (2004), "Seasonal Time Series Forecasting: A Comparative Study of ARIMA and ANN Models", African Journal of Science and Technology (AJST) Science and Engineering Series Vol. 5, No. 2, pages: 41-49.

[17] Durbin, J. (1960). The fitting of time series models. Review of the International Institute of Statistics 28: 233-244.

[18] Wei, William, W. S. (2006). Time series analysis: Univariate and Multivariate Methods 2nd ed. Pearson Addison Wesley.

[19] Ljung, G. M., and Box. G. E. P., (1978). On a Measure of Lack of Fit in Time Series Models. Biometrika, 65: 297–303.

[20] G. P. Zhang, (2003), "Time series forecasting using a hybrid ARIMA and neural network model", Neurocomputing 50, pages: 159–175.

[21] Diebold, F. X., and Mariano, R., S., (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics, 13: 253–263.