UNIVERSITY FOR DEVELOPMENT STUDIES

COMPARATIVE ANALYSIS OF SARIMA AND SETAR MODELS IN PREDICTING PNEUMONIA CASES IN THE NORTHERN REGION OF GHANA

MOHAMMED BADAWI

2016

UNIVERSITY FOR DEVELOPMENT STUDIES

COMPARATIVE ANALYSIS OF SARIMA AND SETAR MODELS IN PREDICTING PNEUMONIA CASES IN NORTHERN REGION

BY

MOHAMMED BADAWI (B.Sc. Statistics)

(UDS/MAS/0040/14)

THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS, FACULTY OF MATHEMATICAL SCIENCES, UNIVERSITY FOR DEVELOPMENT STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE DEGREE IN APPLIED STATISTICS

MAY, 2016

DECLARATION

Student

I hereby declare that this thesis is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere.

Candidate‘s Signature:…….. Date:.……..…

Name: Mohammed Badawi

Supervisor

I hereby declare that the preparation and presentation of the thesis was supervised in accordance with the guidelines on supervision of the laid down by the University for Development Studies:

Supervisor‘s Signature:…….. Date:……...

Name: Dr. Albert Luguterah

ABSTRACT

Acute respiratory infection (ARI) range, in spectrum, from mild colds and coughs to life-threatening pneumonias. ARI particularly pneumonia is the major cause of morbidity and mortality among young children under five in developing countries with Ghana not an exception. In this study, we compared the linear SARIMA Model and the nonlinear two regime SETAR Model in predicting pneumonia cases in northern region of Ghana. Data on monthly pneumonia cases obtained from the Tamale Teaching Hospital database was modeled using SARIMA and SETAR models. The results revealed that SARIMA $(1, 1, 1)(0, 0, 1)_{12}$ model was the best SARIMA model for the pneumonia cases. This model has the least AIC of 83.50, AICc of 83.71 and BIC of 96.08. Also, SETAR $(2; 4, 3)$ model was identified as the best SETAR model. This model has an AIC of -165.42 and BIC of 90.06 as the least among the possible models based on the grid search for the best model. Diagnostic checks of both models with the Ljung-Box test and ARCH-LM test revealed that both models were free from higher-order serial correlation and conditional heteroscedasticity respectively. Based on the forecast assessment from the linear SARIMA and the nonlinear SETAR models, the forecast measures suggest that the nonlinear SETAR model outperform the linear SARIMA model. A comparative analysis of the forecasting accuracy of these models with the Diebold-Mariano test revealed that there were no significant difference in the forecasting performance of the two models. The two models were therefore proposed for predicting Pneumonia cases in the region.

ACKNOWLEDGEMENT

The completion of this research has been possible through the help of many individuals who supported me in different stages of this study. I would like to express my deepest appreciation to my supervisors Dr. Albert Luguterah. Who despite his heavy schedule has rendered me immeasurable supports by reviewed the manuscript. His comments and suggestions immensely enriched the content of this work. This acknowledgement would not be complete without thanking Mr. Nasiru Suleman, Lecturer, Department of Statistics for ensuring professional presentation of the work, sound advice, his brotherly love, courage, help and friendship during this period of my study. His contribution to the success of this academic document is untouchable and may the Almighty God reward him with wisdom and understanding uncountable.

I am also grateful to all the lecturers and entire staffs of the University.

I have also benefited from the help of many other individuals during my studies at the University, especially Amadu Yakubu, Shei Baba, Benson Lamidi Majeed and Sulemana Salam.

DEDICATION

This work is dedicated to my dear parents Hajia Azaratu and Alhaji Mohammed Abonbeni and also to my late grandfather Alhaji Bawa Nayugu of blessed memory.

TABLE OF CONTENT

CONTENT **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AAT	Average Atmosphere Temperature
ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
AICc	Akaike Information Criterion Corrected
AIDS	Acquired Immunodeficiency Syndrome
ALRI	Acute lower Respiratory Infection
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroscedasticity
ARCH-LM	Autoregressive Conditional Heteroscedasticity Lagrange Multiplier
ARI	Acute Respiratory Infection
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
ART	Antiretroviral Therapy
BIC	Bayesian Information Criterion
CDC	Centre for Disease Control
Cox	Coxsackier

CPI	Consumer Price Index
CV	Coefficient of Variation
DF	Dickey-Fuller
DHF	Dengue Haemorrhagic Fever
DLM	Dynamic Linear Model
EV	Enter Virus
GARCH	Generalised ARCH
GDP	Gross Domestic Product
GHS	Ghana Health Service
GLM	Generalised Linear Model
GNP	Gross Net Product
GRNN	Generalised Regression Neural Network
GSM	Generalised Seasonal Model
HEV	Human Antivirus
HFMD	Hand Foot Mouth Disease
HIb	Haemophilus Influenza b
HIV	Human Immunodeficiency Virus

ICU	Intensive Care Unit
IDR	Indonesian Rupiah
KNBS	Kenya National Bureau Statistics
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LCL	Lower Confidence Limit
LSE	Least Squares Estimate
LSTAR	Linear Smooth Transition Autoregressive
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percent Error
Max	Maximum
Min	Minimum
MOH	Ministry of Health
MPE	Mean Percent Error
MSE	Mean Square Error
MSETARX	Multivariate Self-excited Threshold Autoregressive with Exogenous Input
MYR	Malaysian Ringgit

OLS	Ordinary Least Square
PACF	Partial Autocorrelation Function
PAG	Pediatric Association of Ghana
PNC	Number of Pneumonia Cases
RCTS	Randomized Control Trials
RH	Relative Humidity
RT-PCR	Reverse-Transcriptase Polymerase Chain Reaction
RMSE	Root Mean Square Error
RSV	Respiratory Syncytial Virus
SARIMA	Seasonal Autoregressive Integrated Moving Average
SETAR	Self-Excited Threshold Autoregressive
THB	Thai Babt
TL	Turkish Lira
TTH	Tamale Teaching Hospital
UCL	Upper Confidence Limit
UNICEF	United Nations Children's Fund
URTI	Upper Respiratory Tract Infection
WHO	World Health Organization

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Acute respiratory infection (ARI) range, in spectrum, from mild colds and coughs to life threatening pneumonias. ARI particularly pneumonia is the major cause of morbidity and mortality among young children under five in developing countries, accounting for more than 95 percent of all new cases worldwide. According to Black *et al*. (2003), between 11 million and 20 million children with pneumonia will require hospitalisation and more than 2 million will die. Pneumonia accounts for approximately 1.9 million deaths globally in children under five each year (Black *et* al., 2003 and Williams *et* al., 2002).

This is worrying as most pneumonia deaths should be preventable. Further, pneumonia remains the most common reason for adult hospitalisation in Sub-Saharan Africa, with an estimated 4 million episodes and 200,000 deaths 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 (Zar *et al*., 2013). Most of these deaths (81%) occurred in children under 2 years of age. The incidence and the severity of childhood pneumonia was higher in Africa and South East Asia, which account for 30% and 39% respectively of global burden of severe causes.

Pneumonia is a form of acute respiratory infection that affects the lungs and is the leading cause of deaths in children worldwide (Black *et al*., 2003). Pneumonia can be caused by bacteria, viruses, or fungi. *Streptococcus* pneumonia is the most common cause of bacterial

pneumonia in children whilst, *Haemophilus influenza* type b (Hib) is the second most common cause of bacterial pneumonia. Respiratory *syncytial* is the most common cause of pneumonia. Majority of pediatric pneumonia cases in developing countries are due to bacterial causes. In particular, the most common organism is *Streptococcus* pneumonia, which has been identified in 30%-50% of pneumonia cases. The second most common bacteria causing pneumonia is Hib, followed by *Staphylococcus aureus* and *Klebsiella* pneumonia. Viral pathogens such as Respiratory Syncytial Virus (RSV) influenza A and B, *parainfluenza,* human *metapneumovirus,* and adenovirus can also lead to ARI. Poll and Opal (2009) in their study of pathogenesis, treatment and prevention of *pneumoccoca* pneumonia has found that severe pneumonia is the leading cause of community-acquired pneumonia worldwide. It is a bacterium that is a common resident in the upper respiratory tract of humans causing many individuals to be carriers. It is estimated that 10% of adults and 20%-40% of healthy children are carriers, allowing the organism to be maintained in human populations. It is transmitted through direct contact with respiratory secretions, most commonly of those in the same household. Although pneumonias are not considered highly contagious, large community-wide outbreaks can still occur, especially in urban settings or areas with high population density. It commonly affects patients with Acquired Immunodeficiency Syndrome (AIDS) and other immune compromised states and the elderly as well as children under five.

Pneumococcal pneumonia usually begins as a mild upper airway infection similar to a viral respiratory infection. If the *pneumococcal* bacteria enter the lower airways despite responses to prevent its descent such as coughing, mucous clearance, and local immune defenses, pneumonia will develop abruptly (Poll and Opal, 2009). The initial symptoms of *pneumococcal* pneumonia include fever, chills, fatigue, cough, and shortness of breath. The

cough associated with *pneumococcal* pneumonia will then become purulent with blood-tinged sputum and will be associated with chest pain. If left untreated, this disease can progress to acute respiratory failure, septic shock or death within days of onset. Severe pneumonia can also have clinical manifestations of meningitis, sepsis, pericarditis, and endocarditis. The more severe manifestations often occur in those with immune compromised states including the elderly, neonates, Human Immunodeficiency Virus (HIV) positive individuals, and asplenic individuals. Hib is the second most common bacteria- causing pneumonia worldwide. Hib pneumonia caused about 7.9 million cases worldwide in the year 2000 and accounted for 16% of total pneumonia deaths in children (Izadnegahdar *et al*., 2013). In a study of disease caused by Hib in children younger than five years, Hib's total mortality in children age 1-59 months was high at 371,000 deaths yearly. Combined with Severe pneumonia, these two pathogens are directly responsible for as many deaths worldwide as HIV/AIDs, malaria, and tuberculosis combined (Rudan and Campbell, 2009).

In addition to causing pneumonia, Hib has other severe and life-threatening manifestations of meningitis and epiglottis (Watt *et al*., 2009). Similarly to Severe pneumonia, Hib also resides in the nose and throat mucosa and can be spread through respiratory droplets (Hall, 2010). Hib pneumonia and pneumococcal pneumonia are both forms of lobar community-acquired pneumonias and present clinically in the same way.

The most common viral cause of pneumonia is RSV which caused 33.8 million cases of Acute Lower Respiratory Infections (ALRI) in children under five in 2005 (Hall, 2010). This estimate corresponds to 22% of all ALRI and 3%-9% of ALRI-related deaths. RSV occurs worldwide, yet has the greatest burden of disease in developing countries; 96% of ALRI are caused by RSV and 99% of fatal cases of RSV were in developing countries in 2005 (Hall,

2015). RSV has been attributed to 15%-40% of children hospitalised for pneumonia or bronchiolitis in developing countries (Singh and Aneja, 2011). Age group primarily affected by RSV pneumonia is between 2 and 12 months of age. Specifically, the infants most at risk are those with underlying conditions including prematurity, lung disease, malnutrition, or congenital heart disease. RSV is a seasonal virus and occurs in cold seasons or wet seasons in temperate and Mediterranean climates, respectively. In islands in which there is perennial high rainfall, RSV seasonality can be difficult to predict (Weber *et al*., 1998). RSV bronchiolitis begins as an upper respiratory tract infection (URTI) with rhinitis, cough, and fever. These signs precede involvement of the lower respiratory tract in which patients begin to have shortness of breath, difficulty feeding, and respiratory distress (Simoes, 1999). In RSV bronchiolitis, unlike typical bacterial pneumonia, patients often also have a wheeze, prolonged expiratory phase, and air trapping, similar to asthmatics. RSV can also cause a pneumonia-like syndrome in one-third of cases, which requires longer respiratory support than bronchiolitis, and most often occurs in infants with underlying diseases. In a study in Pakistan, greater than 25% of patients with RSV also had a co-infection with severe pneumonia or Hib (Singh and Aneja, 2011). Co-infection with bacteria is predicted account for a large part of the mortality caused by RSV (Simoes, 1999). Treatment in RSV bronchiolitis and pneumonia is mostly supportive with oxygen therapy and mechanical ventilation as needed (Simoes, 1999). In addition, since airway inflammation is a major factor in mortality due to RSV, the use of a long-acting beta adrenergic inhaler or inhaled racemic epinephrine are possible treatment options, although they have had mixed results in clinical studies. In developing countries, anti-viral treatment with ribavirin is rarely used due to its high cost and lack of consensus on its efficacy in RSV.

Generally, pneumonia can be treated with antibiotics. Pneumonia is frequently an associated cause of mortality in children with other underlying conditions. Co-morbid conditions especially malnutrition, measles or HIV increase severity and risk mortality from pneumonia (Black *et al.,* 2003). The World Health Organisation (WHO) 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. According to Paediatric Association of Ghana (PAG, 2010) pneumonia has been rated as one of the leading causes of under-five mortality and morbidity in Ghana, with twenty two percent of children dying from it. Due to the high incidence of pneumonia death revealed on existing literature especially under five which can have consequence on population growth and productivity in future, there is the need for stringent measures to curb the trend. This study therefore seek to compare both linear and non-linear time series models in predicting pneumonia cases in Northern Region of Ghana.

1.2 Problem Statement

Pneumonia kills more children under five than malaria, measles and AIDS combined (WHO, 2013). It is the single major killer of children under five in developing world. In spite of this, relatively few global resources are dedicated to solving the problem and have received far less attention. Each year, more than two million children under five die of pneumonia in the developing world, compared to an estimated 800,000 children who die from malaria and around 300,000 children under five who die from AIDS (PAG, 2010). In the last decade there have been several advances and new interventions resulting in a substantial reduction in pneumonia incidence and improve outcomes. These include more widespread use of case management strategies, development and implementation of polysaccharide protein conjugate

vaccines, better prevention of HIV transmission and uptake of effective Antiretroviral Therapy (ART) of HIV – infected adults and children. As a result, there has been a considerable reduction in pneumonia mortality in children under five years of age from 1.7 million cases globally in 2000 to 1.3 million cases in 2011 (Zar *et al.* 2013). Nevertheless, pneumonia remains the major cause of deaths in children worldwide beyond the neonatal period. Recognising the symptoms of pneumonia is the first step in reducing deaths among children under five. Caregivers or mothers play a critical role in recognising pneumonia's symptoms and immediately seeking appropriate care for their sick children. Indeed it is critical that caretakers or mothers understand the importance of this disease and the risk it poses to their children's health. Yet, even though pneumonia is the leading killer of children in the developing world, most mothers and caregivers are ignorant about the danger signs and risk factors for pneumonia. Only about one in five caregivers know the danger signs of pneumonia (UNICEF, 2004).

The increasing trend of this risk factor could be much worrying to the economic development of any state and world as a whole. Governments and other health institutions need to devote billions of dollars into 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 linear and non-linear time series models in predicting pneumonia cases in the Northern Region of Ghana.

1.3 General Objective

The main objective of this study is to develop appropriate SARIMA and SETAR models for predicting pneumonia cases in Northern Region of Ghana.

1.4 Specific Objectives

i. To investigate the trend characterising pneumonia cases in Northern Region.

ii. To compare the performance of the SARIMA and SETAR models in predicting pneumonia cases in Northern Region.

iii. To predict future pneumonia cases beyond the period under consideration.

1.5 Significance of the Study

The findings of the study could be used by the Ministry of Health (MOH) to initiate policies to reduce the incidence of this condition in the Northern Region.

In addition, the finding of the study would give an overview about the pattern of pneumonia cases in Ghana, therefore providing basis for further research about this disease in Ghana as a whole. Finally the findings of this study would contribute significantly to existing literature in health studies.

1.6 Structure of the Thesis

The thesis is organised into five chapters. Chapter one contains the introduction of the research work. Chapter two comprises of literature review. Chapter three outlines the methodology employed in this research, whiles chapter four presents the analysis and discussion of results. Chapter five is devoted to conclusion and recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

This chapter reviews empirical works done on pneumonia cases and some relevant time series methods that have been used in this study. The chapter is divided into two main headings namely; empirical researches on pneumonia, empirical researches on time series models.

2.1 Empirical researches on Pneumonia

A number of researches have been carried out on pneumonia using different methodologies.

UNICEF (2004) estimated that more than 150 million episodes of pneumonia occur every year among children under five in developing countries, accounting for more than 95 percent of all new cases worldwide. It has also been estimated that 26 percent of neonatal deaths are caused by severe infections during the neonatal period and a significant proportion of these infection are caused by pneumonia.

Also, Theodoratou (2010) conducted a systematic review of the literature assessing the effect of pneumonia case management on mortality from childhood pneumonia. The review covered the following interventions: community case management with antibiotic treatment, and hospital treatment with antibiotics, oxygen, zinc and vitamin A. Pneumonia mortality outcomes were sought where available data were also recorded on secondary outcomes. They summarized results from randomized controlled trials (RCTs), cluster RCTs, quasi-experimental studies and observational studies across outcome

measures using standard meta-analysis methods. Their results estimate that community case management of pneumonia could result in a 70% reduction in mortality from pneumonia in 0–5-year-old children. In contrast, treatment of pneumonia episodes with zinc and vitamin A is ineffective in reducing pneumonia mortality. There was insufficient evidence to make a quantitative estimate of the effect of hospital case management on pneumonia mortality based on the published data.

Again, Meyer *et al*. (2007) evaluate the impact of an intervention to reduce the duration of antibiotic treatment for pneumonia in a neurosurgical intensive care unit (ICU). The usage of antibiotics and the resultant costs were examined using interrupted time series analysis while resistance and device-associated infection rates were also described. Their result shows that intervention was associated with significant decrease from 949.8 to 626.7 after the intervention. This was mainly due to reduced consumption of second-generation cepha-losporins, imidazole and carbapenems. Similarly, total antibiotic costs showed a significant decrease from 13.16 before to 7.31 after the intervention. The incidence of patients dying with pneumonia did not change significantly.

Also, Victer *et al.* (1994) in a case study conducted in Brazil confirmed that in a number of community based studies, boys appears more frequently affected by pneumonia than girls. However, in clinical studies possibility of gender bias in seeking care cannot be ruled out, which may show male prepondence.

Moreover, Shan (1994) in a study of risk factors for severe pneumonia in children have shown that parental education is a significant risk factor for severe pneumonia. Again, WHO (2013) reported that 6.3 million children under five years old died worldwide,

nearly 17,000 every day. Almost 75% of all child deaths were attributed to only six conditions: neonatal causes, pneumonia, diarrhoea, malaria, measles, and HIV/AIDS.

Furthermore, Walker *et al.* (2013) in their study of global burden of childhood pneumonia and diarrhoea found that pneumonia is the leading cause of death in children under five, with 1.3 million deaths out of a total of 120 million episodes of pneumonia in the year 2011.

Also, Singh and Aneja (2011) in their study of epidemiology and aetiology of childhood pneumonia established that an approximately 11-20 million cases (7%-13%) of pneumonia are severe type and that require hospitalization.

Next, Rudan *et al.* (2008) in their study of epidemiology of pneumonia in developing world found that although paediatric pneumonia affects children worldwide, the majority of cases occur in developing countries. In the year 2000, the estimated incidence of pneumonia in children under 5 years old in developing countries was 0.29 episodes per child-year compared to 0.05 episodes per child-year in developed countries.

Again, Singh and Aneja (2011) in their study of pneumonia management in the developing world established that majority of paediatric pneumonia cases in developing countries are due to bacterial causes. In particular, the most common organism is *Streptococcus* pneumonia, which has been identified in 30%-50% of pneumonia cases. The second most common bacteria causing pneumonia is *Haemophilis influenzae* type b *(*Hib*),* followed by *Staphylococcus aureus* and *Klebsiella* pneumonia. Viral pathogens such as RSV, influenza A and B, *parainfluenza,* human *metapneumovirus,* and *adenovirus* can also lead to acute respiratory infections.

Moreover, Poll and Opal (2009) in their study of pathogenesis, treatment, and prevention of *pneumococcal* pneumonia has found that severe pneumonia is the leading cause of community-acquired pneumonia worldwide. It is a bacterium that is a common resident in the

upper respiratory tract of humans causing many individuals to be carriers. It is estimated that 10% of adults and 20% - 40% of healthy children are carriers, allowing the organism to be maintained in human populations. It is transmitted through direct contact with respiratory secretions, most commonly of those in the same household. Although pneumonia are not considered highly contagious, large community-wide outbreaks can still occur, especially in urban settings or areas with high population density. It commonly affects patients with *AID*S and other immune compromised states and the elderly as well as children under five.

Also, Izadnegahgar *et al*. (2013) in their study of childhood pneumonia in developing countries found that 13.8 million cases of pneumococcal pneumonia in year 2000, accounting for 41% of all pneumonia deaths in children.

Again, Watt *et al*. (2009) found in their study of burden of disease caused by *Haemophilus Influenza* type b in children younger than five years that Hib's total mortality in children age 1-59 months is high at 371,000 deaths yearly.

Furthermore, Weber *et al*. (1998) in a study of RSV in Tropical and Developing Countries found that age group primarily affected by RSV pneumonia is between 2 and 12 months of age. Specifically, the infants most at risk are those with underlying conditions including prematurity, lung disease, malnutrition, or congenital heart disease.

Next, Wingerter *et al.* (2012) study application of the WHO criteria to predict radiologic pneumonia in United States (US) based paediatric emergency department. The WHO criteria, used for diagnosis of pneumonia, requires cough or difficulty breathing as well as age-specific *tachypnea.* Age-specific t*achypne*a is defined as greater than 60 breaths/minute for children less than 2 months old, greater than 50 breaths/minute for children 2-11months of age, and greater than 40 breaths/minute for children 1-5 years of age. WHO identifies three categories of

pneumonia: very severe, severe, or not severe. These categories are determined following the WHO algorithm as follows: “very severe” pneumonia contains one of the following signs or symptoms: central cyanosis, inability to breastfeed or drink, severe respiratory distress, or convulsions, lethargy, or unconsciousness. “Severe” pneumonia is defined as not meeting criteria for “very severe” pneumonia and having one of the following: retractions, nasal flaring, or grunting. Lastly, “not severe” pneumonia is defined by the WHO as meeting the general criteria for pneumonia, but not meeting the criteria for “very severe” or “severe” pneumonia.

Further, Madhi *et al*. (2008) in their study of vaccines to prevent and improve child survival found that, historically, pertussis and measles were significant causes of childhood pneumonia. The pertussis vaccination, which has been available since the 1950s and is included in most immunization programs worldwide, is estimated to have prevented 38.3 million cases and 607,000 deaths.

Also, Van *et al.* (2003) in their study of influenza vaccination in the year 2000, had found that measles vaccination has also had a significant effect; measles deaths declined from 2.5 million to 873,000 from 1980 to 1999. More recently, vaccines have been developed for severe pneumonia, Hib, and influenza.

Again, in a birth cohort study in Cape Town, South Africa, by Campbell and Nair (2015), it was shown that the majority of the pneumonia burden among children is within the first 2 years of life. The results of this study indicated that severe pneumonia accounts for the most pneumonia deaths in the first 6 months of life. This relationship of increased pneumonia cases in younger ages has also been demonstrated by Monto (1994). This study also evaluated the effect of gender and revealed a male: female incidence ratio of 2:1, hypothesising a “biological frailty” in males compared to females or differences in care- seeking by gender.

In addition, a longitudinal cohort study in Pakistan by Khan *et al*. (2009) noted similar increased incidences of childhood pneumonia in younger children and males.

Another study by Khan *et al.* (2009) conducted in the Himalayas showed that high altitude is significantly associated with increased pneumonia cases. High altitude is likely to contribute to pneumonia due to lung physiologic compensatory mechanisms such as increased ventilation, increased cardiac output, and a shift in the oxygen-haemoglobin affinity curve. These compensations are delayed in infants who take 3-4 years to adapt fully.

Again, Schanzer *et al.* (2013) study statistical estimates of respiratory admissions attributed to seasonal and pandemic influenza pneumonia for Canada. The discharge abstracts of persons admitted with any respiratory condition were extracted from the Canadian Discharge Abstract Database, for April 2003–March 2010. Stratified, weekly admissions were modelled as a function of viral activity, seasonality, and trend using Poisson regression models. The results that an estimated 1 out of every 6 admissions attributable to seasonal influenza (2003–April 2009) were coded to J10 (influenza virus identified). During the 2009 pandemic (May–March 2010), the influenza virus was identified in 1 of 6 admissions attributed to the pandemic strain. Compared with previous H1N1 seasons (2007 ⁄ 08, 2008 ⁄ 09), the influenza-attributed hospitalization rate for persons less than 65 years was approximately six times higher during the 2009 H1N1 pandemic, whereas for persons 75 years or older, the pandemic rate was approximately five fold lower.

2.2 Empirical researches on Time series Methods.

Williamson and Hudson (1999) in their study of monitoring system for detecting aberrations in public health surveillance successfully identified and fitted the Seasonal Autoregressive Integrated Moving Average (SARIMA) model for hepatitis A,

hepatitis B, hepatitis non-A-non-B, le-gionellosis, malaria, meningococcal infections, and tuberculosis and pneumonia. Sharmin and Rayhan (2011) modelled an infectious disease to provide an early signal of infectious disease epidemics by analysing the disease dynamics. A two-stage monitoring system was applied, which consists of univariate Box-Jenkins model or Autoregressive Integrated Moving Average (ARIMA) model and subsequent tracking signals from several statistical process-control charts. The analyses were illustrated on January 2000–August 2009 national measles data. Their results of this empirical study revealed that the most adequate model for the occurrences of measles in Bangladesh was the SARIMA (3, 1, 0)(0, 1, 1)$_{12}$ model, and the statistical process-control charts detected no measles epidemics during September 2007–August 2009. The two-stage monitoring system performed well to capture the measles dynamics in Bangladesh without detection of an epidemic because of high measles-vaccination coverage.

Also, Feng *et al*. (2014) modelled outbreaks of hand-foot-mouth disease (HFMD) in Mainland China. Samples obtained from children hospitalized with HFMD in Zhengzhou, Henan, China, were examined for the existence of pathogens with reverse-transcriptase polymerase chain reaction (RT-PCR) from 2008 to 2012. SARIMA models for the weekly number of HFMD, Human Entero virus 71 (HEV71) and Coxsackie VirusA16 (CoxA16) associated HFMD were developed and validated. Cross correlation between the number of HFMD hospitalizations and climatic variables was computed to identify significant variables to be included as external factors. Time series modelling was carried out using multivariate SARIMA models when there was significant predictor meteorological variable. Their result showed that out of 2932 samples from the patients hospitalised with HFMD, 748 were detected with HEV71, 527 with CoxA16

and 787 with other Enter virus (EV) from January 2008 to June 2012. Average Atmospheric Temperature (AAT) lagged at 2 or3 weeks were identified as significant predictors for the number of HFMD and the pathogens. SARIMA $(0, 1, 0)(1, 0, 0)_{52}$ associated with AAT at lag 2 weeks and SARIMA $(0, 1, 1)(1, 1, 0)_{52}$ Lag 3 weeks were developed and validated for description and predication the weekly number of HFMD, HEV71-associated HFMD, and Cox A16-associated HFMD hospitalisations.

Again, Briet *et al.* (2008) study forecasting models, with an aim to developing a forecasting system which could assist in the efficient allocation of resources for malaria control in Sri Lanka. Exponentially Weighted Moving Average models, ARIMA models and SARIMA models were compared on monthly time series of district malaria cases for their ability to predict the number of malaria cases one to four months ahead. The addition of covariates such as the number of malaria cases in neighbouring districts or rainfall were assessed for their ability to improve prediction of selected SARIMA models. They found that the best model for forecasting and the forecasting error varied strongly among the districts. The addition of rainfall as a covariate improved prediction of selected SARIMA models modestly in some districts but worsened prediction in other districts. Improvement by adding rainfall was more frequent at larger forecasting horizons. Their research point to the fact that heterogeneity patterns of malaria in Sri Lanka requires regionally specific prediction models. Prediction error was large at a minimum of 22% (for one of the districts) for one month ahead predictions. The modest improvement made in short term prediction by adding rainfall as a covariate to these prediction models may not be sufficient to merit investing in a forecasting system for which rainfall data are routinely processed.

Next, Ekezie *et al*. (2014) conducted a research to model and forecast malaria mortality rate

using SARIMA models at Aboh Mabaise general hospital, Imo state Nigeria. They used the Box- Jenkins methodology to build ARIMA model for malaria mortality rate for the period. They concluded that the forecasted results revealed a decreasing pattern of malaria mortality in the last quarter of the year 2014.

Also, Nobre *et al*. (2001) studied data collected through a national public health surveillance system in the United States to evaluate and compare the performances of SARIMA and a Dynamic Linear Model (DLM) for estimating case occurrence of two noticeable diseases (malaria and hepatitis A). Their comparison found out that the two forecasting modelling techniques (SARIMA and DLM) are comparable when long historical data are available (at least 52 reporting periods).The residuals for both predictor models showed that they were adequate tools for use in epidemiological surveillance.

Again, Permanasari *et al*. (2009) analysed data set on human Salmonellosis occurrences in United States which comprises of fourteen years of monthly data obtained from a study published by Centres for Disease Control and Prevention (CDC). Several SARIMA models were developed to forecast the occurrence of the disease. The models were validated using the diagnostic test to obtain the appropriate model. Their result showed that the SARIMA (9, 0, 1 4)$(12, 1, 24)_{12}$ was the fittest model. Moreover, Wongkoon *et al*. (2008) studied the incidence of Dengue Haemorrhagic Fever (DHF) in Northern Thailand. SARIMA models were applied to analyse 2003 to 2006 data. The forecasted data were validated with data collected from January 2007 to September 2007. Their results showed that SARIMA model was suitable for predicting the number of DHF incidence in Northern Thailand.

Furthermore, Chen and Trajkovic (2004) analysed data collected from a deployed network and used clustering techniques to characterise patterns of individual users' behaviour. A

network traffic prediction approach was then developed based on user clusters. Based on the identified user clusters, they used the SARIMA model to forecast the network traffic by aggregating the predicted traffic of each user cluster. The predicted network traffic shows good agreement with the collected traffic data.

Also, Yan *et al*. (2010) developed and evaluated an innovative hybrid model, which combines the SARIMA and the Generalised Regression Neural Network (GRNN) models, for bacillary dysentery forecasting in Yichang City of China. The model was applied to monthly data of bacillary dysentery from 2000-2007. Their test results showed that hybrid SARIMA-GRNN model outperformed the SARIMA model with lower mean square error, mean absolute error and mean absolute percentage error when simulation and forecasting performance are compared.

Again, Li (2009) analysed the measured temperature of Stockholm from 1756 to 2007 by using Generalised Linear Model (GLM) and ARIMA models. He forecasted the monthly temperature of 2008 and compared with the true values. His conclusion was that the SARIMA model for the series fits the data better than the GLM. Moreover, Shekhar (2004) researched into recursive methods as applied to SARIMA $(1, 0, 1)(0, 1, 1)_{12}$ model parameter estimation. His results established the stability and consistency of the SARIMA model and concluded that the parameters did not show a highly variable pattern with time. Also, the model was insensitive to minor fluctuations in the parameters.

Next, Akter and Rahman (2010) studied milk supply of a dairy cooperative in the UK using Holt-Winter's seasonal model and SARIMA model. Their results showed that longer series produces better forecasts than a shorter series and the generated forecasts had error of less than 3 per cent. Also, Xiang (2008) used SARIMA model to study climate change by considering

temperature measurement of Stockholm from 1756 to 2007. The result indicated that even the strongest outlier has weak effect and there exist a stable structure in the temperature data. Moreover, Yantei *et al.* (2005) fitted multiplicative seasonal ARIMA models to measured GSM traffic traces of China Mobile of Tianjin network. Their experiments showed that the forecasting values and the actual values had a relation.

Furthermore, Fannoh *et al.* (2014) used SARIMA approach to model Liberia's monthly inflation rates which showed that SARIMA model was appropriate for modelling the inflation rates.

Again, Otu *et al*. (2014) used Box-Jenkins methodology to build SARIMA model for Nigeria's monthly inflation. SARIMA $(1, 1, 1)(0, 0, 1)_{12}$ model was developed and used to forecast monthly inflation for the year 2014.

Also, Saz (2011) studied the efficacy of SARIMA models in the view of forecasting the inflation rates in Turkish economy. Tests were performed on the seasonality in the Turkish inflation rate. They used the systematic modelling approach in conjunction with the step-wise selection procedure of the HK algorithm to find the best model of inflation in Turkey.

Next, Kibunja *et al*. (2014) considered a univariate time series model to forecast precipitation in Mt. Kenya region. SARIMA model was fitted on to the data and the model which exhibited the least AIC and BIC values was selected. The model passed residual normality test and the forecasting evaluation statistics.

Again, Gikungu *et al.* (2015) developed a SARIMA model to forecast Kenya's inflation rate using quarterly data for the period 1981 to 2013 obtained from Kenya National Bureau Statistics (KNBS). SARIMA $(0, 1, 0)(0, 0, 1)_4$ was identified as the best model. More so,

Sumer *et al*. (2009) developed ARIMA, SARIMA and regression models with seasonal latent variable in forecasting electricity demand of the data from 'Kayseri and Vicinity Electricity Joint- Stock Company'. Their results show that the regression model with seasonal latent variable was more efficient than ARIMA and SARIMA.

Also, Chikobvu and Sigauke (2012) developed SARIMA and regression with SARIMA errors (regression-SARIMA) models to predict daily peak electricity demand in South Africa using data for the period 1996 to 2009. The performance of the developed models was evaluated by comparing them with winter's triple exponential smoothing model. Empirical results from the study show that the SARIMA model produces more accurate short-term forecasts.

Moreover, Shulze and Prinz (2009) applied SARIMA model and Holt-Winters exponential Smoothing approach to forecast container transhipment in Germany, the results show that SARIMA approach yields slightly better values of modelling the container throughout than the exponential smoothing approach.

Again, Oduro *et al.* (2012) conducted a study on application to microwave transmission of Yeji-Salaga (Ghana). They applied the SARIMA model to analyse the monthly data. The results showed that SARIMA $(1, 1, 1)(0, 1, 2)_{12}$ was the best fitted model.

Also, Pufnik and Kunovac (2006) provided a method of forecasting the Croatia's Consumers Price Index (CPI) by using univariate SARIMA models and forecasting future values of the variables from past behaviour of the series. Their paper attempts to examine whether separate modelling and aggregating of the sub-indices improves the final forecast of the all items index.

Again, Akhter (2013) forecasted the short-term inflation rate of Bangladesh using the monthly CPI from January 2000 to December 2012. The paper employs the SARIMA models proposed by Box *et al*. (1994). Because of the presence of structural break in the CPI, the study truncates

the series and used data from September 2009 to December 2012. The forecasted result suggests an increasing pattern and high rates of inflation over the forecasted period of 2013.

Furthermore, Nasiru and Sarpong (2012) employed an empirical approach to modelling monthly inflation data in Ghana using the SARIMA model. Their result showed that SARIMA $(3, 1, 3)(2, 1, 1)_{12}$ model was appropriate for modelling Ghana's inflation rate. Diagnostic test of the model residuals with the ARCH LM test and Durbin Watson test indicates the absence of autocorrelations and ARCH effect in the residuals. The forecast results inferred that Ghana was likely to experience single digit inflation values in 2012.

Also, Myriam *et al*. (2011) conducted a time series analysis of dengue incidence in Guadeloupe, French West Indies to forecast using climate variables as predictors. They used the Box-Jenkins approach to fit a SARIMA to model the incidence of the dengue from 2000 to 2006 using clinical suspected cases. They concluded that temperature improves dengue outbreak forecasts better than humidity and rainfall. The SARIMA model used climate data as independent variables and incorporated it into an early and a reliably monitoring system of dengue outbreak. Next, Omane-Adjepong *et al*. (2013) examined the most appropriate short-term forecasting method for Ghana's inflation. The monthly dataset used was divided into two sets, with the first set used for modelling and forecasting, while the second set was used as test. SARIMA and Holt-Winters approaches were used to obtain short-term out of sample forecast. From the results, they concluded that an out of sample forecast from an estimated SARIMA $(2,1,2)(0,0,1)_{12}$ model far supersedes any of the Holt-Winters' approach with respect to forecast accuracy.

Moreover, Ismail *et al.* (2009) used a rule-based forecasting approach for forecasting peak load electricity demand. They concluded that rule-based forecasting increases the forecast accuracy

when compared to the traditional SARIMA model and that improvement depends on conditions of the data, knowledge, development and validation.

Again, Jamaludin *et al.* (2015) study Relative Humidity (RH) of 13 stations all over the peninsular Malaysia for the period of 1968 to 2009. They model and make future trend predictions using RH data by applying the Box-Jenkins methodology to build SARIMA model for monthly RH data. The SARIMA model for each station was developed. These models were used to forecast 30 months upcoming RH data. Also, Amos (2010) examined financial time series with special application to modelling inflation data for South Africa. The data spanned from January 1994 to December 2008. The study considered two families of time series namely ARIMA with extension to the SARIMA model and the Autoregressive Conditional Heteroscedastic (ARCH) with extensions to the Generalised ARCH (GARCH) model. The study concluded that the SARIMA $(1, 1, 0)(0, 1, 1)_{12}$ was the best fitting model from the ARIMA family of models while the GARCH (1, 1) was chosen to be the best fit from the ARCH-GARCH models. Furthermore, a comparison of the two selected models based on the goodness of fit and the forecasting power of the two models was carried out. It was established that the GARCH (1, 1) model was superior to the SARIMA $(1, 1, 0)(0, 1, 1)_{12}$ model according to both criteria as the data was characterised by changing mean and variance.

Also, Watiert *et al.* (1994) fitted a Self-excited Threshold Autoregressive (SETAR) model to epidemiological time series of reported cases of Salmonella typhimurium in France. The fitted 'full' model was compared with a Simple Autoregressive (AR) model. They compared the full model with the 'restricted' model. Their result favour modelling by a SETAR process instead of an AR process. Thus the time series of infections due to Salmonella typhimurium exhibits a type of non-linearity which can be accounted for by a threshold model.

Again, Modarres and Ouarda (2012) modelled the heteroscedasticity in the residuals of the SARIMA model using a GARCH model. The model is applied to two monthly rainfall time series from humid and arid regions. The effect of Box–Cox transformation and seasonal differencing on the remaining seasonal heteroscedasticity in the residuals of the SARIMA model was also investigated. It was shown that the seasonal heteroscedasticity in the residuals of the SARIMA model can be removed using Box–Cox transformation along with seasonal differencing for the humid region rainfall. Therefore, the GARCH modelling approach was necessary to capture the heteroscedasticity remaining in the residuals of the SARIMA model. However, the evaluation criteria do not necessarily show that, the GARCH model improves the performance of the SARIMA model.

Also, Kahraman and Aydiner (2013) study daily exchange rate of dollar (USD) and gold prices series in Turkish Lara (*TL)* using multivariate self-exciting threshold autoregressive model application. Gold prices series has been taken as indicator variable and multivariate SETAR model has been created. Then, predictions have been obtained from the model to evaluate performance of the model. Accordingly, the model was said to be suitable to make predictions. According to this obtained multivariate SETAR model, the prices of gold and dollar affect each other in Turkey market and they can be modelled together.

Again, Clements and Smith (1997) investigated the multi-period forecast performance of a number of empirical SETAR models that have been proposed in the literature for modelling exchange rates and Gross Net Product (GNP), amongst other variables. They took each of the empirical SETAR models in turn as the Domestic Gross Product (GDP) to ensure that the 'non-linearity' characterises the future, and compare the forecast performance of SETAR and linear

autoregressive models on a number of quantitative and qualitative criteria. Their results indicate that non-linear models have an edge in certain states of nature.

Further, Gharleghi *et al*. (2014) used a nonlinearity test and a structural change test to detect the nonlinearity and the break date in three ASEAN currencies, namely the Indonesian Rupiah (IDR), the Malaysian Ringgit (MYR) and the Thai Baht (THB). Their study finds that the null hypothesis of linearity was rejected and evidence of structural breaks exist in the exchange rates series. Therefore, the decision to use the self-exciting threshold autoregressive (SETAR) model in their present study was justified. Their results showed that the SETAR model, as a regime switching model, could explain abrupt changes in a time series.

Also, Boero and Marrocu (2003) analysed the out-of-sample performance of SETAR models relative to a linear AR and GARCH model using daily data for the Euro effective exchange rate. The evaluation was conducted on point, interval and density forecasts, unconditionally, over the whole forecast period, and conditional on specific regimes. Their results show that overall the GARCH model was better in capturing the distributional features of the series and predicting higher-order moments than the SETAR model.

Next, Addo (2014) study a multivariate Self–Exciting Threshold Autoregressive with exogenous input (MSETARX) models and present an estimation procedure for the parameters. The conditions for stationarity of the nonlinear MSETARX models was provided. The efficiency of an adaptive parameter estimation algorithm and Least Squares Estimate (LSE) algorithm for this class of models was then provided via simulations.

More so, Aidoo (2010) forecast the performance of Ghana Inflation rates using SARIMA models and SETAR model. Based on the in-sample forecast assessment from the linear SARIMA and the nonlinear SETAR models, the forecast measure Mean Absolute Error (MAE)

and Residuals Mean Square Error (RMSE) suggest that the nonlinear SETAR model outperform the linear SARIMA model. Also, using multi-step-ahead forecast method he predicted and compared the out-of-sample forecast of the linear SARIMA and the nonlinear SETAR models over the forecast horizon of 12 months. The results as suggested by MAE and RMSE, the forecast performance of the nonlinear SETAR models was superior to that of the linear SARIMA model in forecasting Ghana inflation rate.

Furthermore, Acatrinei and Caraiani (2011) investigate the existence of nonlinear patterns in the dynamics of the main stock index returns in Romania. They use daily closing data of the BET stock index series from 2004 to early 2010. Based on several tests for nonlinearity they reject the null hypothesis of linearity. They also use several types of threshold models and compare their fitness and forecasting performance with basic AR models. They found that the LSTAR and SETAR models fit best the data; however, they cannot outperform the simpler AR models in forecasting. Their results suggest that although there are nonlinear features in data, the threshold models are not complex enough to reveal the data complexity.

Finally, Basikhasteh *et al.* (2014) study the demand of gold prices due to economic crisis in Iran. They used AR model, AR-IGARCH model, SETAR and STAR models for forecasting and these methods were applied for modelling a monthly log return time series of gold price from August 2007 to November 2013 (price in Iranian Rial against 1 gram of gold). Their result shows that the time series is nonlinear and SETAR (2; 1, 3) model yields the best result for 2012.

2. 3 Conclusion

The chapter dealt with reviewing of literature that is relevant to the study. Reviewing of the literature has exposed us to the diverse techniques that researchers have employed in predicting pneumonia cases. However, in most study of the condition, little time series models have been used, more specially the linear and nonlinear models. Based on the research gap found, this study therefore employed the Seasonal Autoregressive Integrated Moving Average (SARIMA) models and Self Excited Threshold Autoregressive (SETAR) models in predicting pneumonia cases in Northern Region of Ghana.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter focused on the source of data collected for the research and the various statistical techniques employed in analysing the data in order to meet the desired objectives of the study. The chapter is divided into thirteen main headings namely; data and source, trend analysis, unit root test, Autoregressive model ($AR(p)$), Moving Average model ($MA(q)$), Autoregressive Moving average ($ARMA(p, q)$) model, Autoregressive Integrated Moving Average (ARIMA(P, d, q)) model, Seasonal Autoregressive Integrated Moving Average ($SARIMA(p, d, q)(P,D,Q))_{12}$ model, Self-Excited Threshold Autoregressive (SETAR) model, linearity test, model selection criteria, model diagnostics and Diebold-Mariano test.

3.1 Data and Source

This study used secondary data on monthly pneumonia inpatients cases from January 2000 to October 2015 obtained from the Tamale Teaching Hospital (TTH) data base. The data for this research was analysed with R, Gretl, and Minitab statistical packages.

3.2 Trend Analysis

The trend of a series reflects the long term growth of the time series over time. A time series variable may exhibit different type of trends; the linear, linear constant growth, quadratic and

quadratic constant growth among others. This study evaluated the above different types of trend models for the disease under consideration. If the trend in the time series is a linear function of time *t*, then

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t \quad (3.1)$$

where, Y_t are the observations of the time series, t is a time dummy ($t = 1,2, \ldots, n-1, n$) and ε_t is a random error component. When the series exhibit quadratic trends, the model is given as;

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t \quad (3.2)$$

For a polynomial of order *k*

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_k t^k + \varepsilon_t \quad (3.3)$$

If the trend is characterised by a constant growth rate, then the equation is

$$Y_t = \beta_0 e^{\beta_1 t} \varepsilon_t \quad (3.4)$$

It logarithmic form can be written as

$$\ln Y_t = \ln \beta_0 + \beta_1 t + \ln \varepsilon_t \quad (3.5)$$

If the constant growth rate is quadratic, then its logarithmic form is given as

$$\ln Y_t = \ln \beta_0 + \beta_1 t + \beta_2 t^2 + \ln \varepsilon_t \quad (3.6)$$

The coefficients appearing in the equations (3.1) to (3.6) above are obtained by applying the principles of Ordinary Least Squares (OLS).

3.3 Unit Root Test

A very essential aspect of time series analysis is to ensure that the data is weakly stationary. A variable is said to be covariance or weakly stationary if the first two moments of the series; the mean and the autocovariance are finite and are time invariant. That is, the expected value of the time series does not depend on time and the autocovariance function, $cov(y_t, y_{t+k})$ at any lag say k, remain constant over time. Stationarity condition ensures that the properties of the estimated parameters from the model are standard. When this condition is assured, then the estimated model can be used for forecasting. To check for stationarity, we sometimes test for the existence or nonexistence of what we call unit root. 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. In testing for a unit root in a given series the features of the series must be known. When the series contains both seasonal and non-seasonal behaviour, the test of stationarity must be conducted on both components. The presence or absence of unit roots is imperative to identifying the nature of the processes that generate the time series data and to investigate the order of integration of the series. This is because, contemporary econometrics has indicated that, regression analysis using non-stationary time series variables produce spurious regression since standard results of ordinary least squares do not hold.

Many methods have been suggested for testing for stationarity of a time series data. These include both graphical and quantitative methods. The graphical approach includes observing the Autocorrelation function (ACF) plots. A strong and slow dying ACF will suggest deviation from stationarity. For the purpose of this study, in addition to the ACF, other quantitative techniques for testing for unit root were employed. These are discussed below.

3.3.1 Augmented Dickey-Fuller (ADF) Test

The ADF test proposed by Dickey and Fuller (1979) was employed in the study to determine whether the disease involved contained a unit root (non-stationary) or has stationary covariance. The ADF test proposed by Dickey and Fuller was an extension of the Dickey-Fuller (DF) test based on the assumption that, the series follows a random walk. Given an autoregressive process of order one, AR (1), below

$$Y_t = \phi Y_{t-1} + \varepsilon_t \tag{3.7}$$

where ε_t denotes a serially uncorrelated white noise sequence with a mean of zero and constant variance. If $\phi = 1$, equation (3.7) becomes a random walk model without drift, which is known as a non-stationary process. The fundamental concept of the ADF test is to simply regress Y_t on its lagged value Y_{t-1} and find out if the estimated ϕ is statistically equal to one or not. Subtracting Y_{t-1} from both sides of equation (3.7) result in

$$\Delta Y_t = \delta Y_{t-1} + \varepsilon_t \tag{3.8}$$

where $\delta = \phi - 1$ and $\Delta Y_t = Y_t - Y_{t-1.}$ We further test for the null hypothesis of $\delta = 0$ against the alternative $\delta \neq 0$. If $\delta = 0$, then $\phi = 1$, meaning that the series have a unit root. Under the null hypothesis $\delta = 0$, the *t*-value of the estimated coefficient of Y_{t-1} does not have an asymptotic normal distribution (Erdogdu, 2007).

The decision to reject the null hypothesis or otherwise is based on the DF critical values of the τ-statistic. Since errors of the DF test usually show evidence of serial correlation rather on the assumption that the error terms were uncorrelated. In order to overcome this problem, the ADF test includes the lags of the first difference series in the regression equation to make the error term white noise and therefore the regression equation is presented in the following form.

$$\Delta Y_t = \delta Y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + \varepsilon_t \quad (3.9)$$

For the inclusion of intercept as well as time trend *t*, the model becomes.

$$\Delta Y_t = \alpha + \beta t + \delta Y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + \varepsilon_t \quad (3.10)$$

where α is a constant, β the coefficient on time trend series, $\sum_{i=1}^{p} \gamma_i \Delta Y_{t-i}$ is the sum of the lagged values of the dependent variable ΔY_t and p is the lag order of the autoregressive process. The parameter of interest in the ADF test is δ. For $\delta = 0$, the series contains unit root and hence non-stationary. The choice of the starting augmentation order depends on; data periodicity, significance of γ_i estimates and white noise residuals. After preliminary estimation, non-significant parameter augmentation can be dropped in order to enjoy more efficient estimates. The test statistic for the ADF test is given by

$$\mathrm{F}_{\tau} = \frac{\hat{\delta}}{\mathrm{SE}(\hat{\delta})} \quad (3.11)$$

where $\mathrm{SE}(\hat{\delta})$ 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.

3.3.2 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

KPSS proposed by Kwiatkowski *et al*. (1992) is another complementary test for determining the order of integration of a series Y_t by testing the null hypothesis that, the data generating process is stationary ($H_0: Y_t \sim \mathbf{I}(\mathbf{0})$) against the alternative that it is non-stationary ($H_1: Y_t \sim \mathbf{I}(\mathbf{1})$). Kwiatkowski *et al*. (1992) derived a test for this pair of hypotheses. The test assumes that if there is no linear trend term, the point of departure is a data generating process of the form

$$Y_t = X_t + \varepsilon_t \tag{3.12}$$

where X_t is a random walk, $X_t = X_{t-1} + v_t$, $v_t \sim \text{iid}(0, \sigma_v^2)$ and ε_t is a white noise process. In this context, the foregoing pair of hypotheses is equivalent to the pair;

$H_0: \sigma_v^2 = 0.$

$H_1: \sigma_v^2 > 0.$

If H_0 holds, Y_t is composed of a constant and the stationary process ε_t; thus, Y_t is also stationary. Kwiatkowski *et al*. (1992) therefore proposed the following test statistic

$$\text{KPSS} = \frac{1}{T^2} \sum_{t=1}^{T} \frac{S_t^2}{\hat{\sigma}_\infty^2} \tag{3.13}$$

where T is the number of observations, $S_t = \sum_{j=1}^{t} \widehat{\omega_j}$ with $\widehat{\omega_j} = Y_t - \bar{Y}$ and $\hat{\sigma}_\infty^2$ is an estimator of

$$\sigma_\infty^2 = \lim_{T \to \infty} T^{-1} Var(\sum_{t=1}^{T} \varepsilon_t)$$

That is, $\hat{\sigma}_{\infty}^{2}$ is an estimator of the long-run variance of the process ε_t. If Y_t is a stationary process, S_t is integrated of order one (**I (1)**) and the quantity in the denominator of the KPSS statistic is an estimator of its variance, which has a stochastic limit. The term in the denominator ensures that overall; the limiting distribution is free of unknown nuisance parameters. If, however, Y_t is integrated of order one (**I (1)**), the numerator will grow without bounds, causing the statistic to become large for large sample sizes. The null hypothesis of stationarity is rejected for large values of KPSS.

3.4 Autoregressive Model of Order (AR (p))

A time series Y_t is said to be an autoregressive process of order *p*, if it is a measured linear sum of the past *p* values plus a random error. The general AR model of order *p* is given by:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \qquad (3.14)$$

where the *Y's* and ε_t are respectively the original series and random error at time period *t*, ϕ_i are the AR parameters to be estimated with $i = 1, 2, \ldots, p$ and *p* is the order of the AR model. Thus the value at time *t* depends linearly on the last *p* values and the model looks like a regression model; hence the term autoregression. Using the backward shift operator *B* such that;

$BY_{t-1} = BY_t$ and $B^2 Y_t = Y_{t-2}$, the AR(p) model may be written more simply in the form.

$$\phi(B) Y_t = \varepsilon_t \qquad (3.15)$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is a polynomial in *B* of order *p*. The AR (*p*) time series is said to be stationary if the roots of the polynomial:

$m^p - \phi_1 m^{p-1} - \phi_2 m^{p-2} - \ldots - \phi_p$ are less than one in absolute terms.

3.5 Moving Average Model of Order (MA (q))

A time series Y_t is said to be a moving average process of order q if it is a weighted linear sum of the last q random shocks. That is, the current values of the series depend on its past shocks. The general MA model is given by:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (3.16)$$

where q is the order of the model, θ_j are the model parameters to be estimated and $j = 1,2, \ldots, q$. The random shocks are assumed to be a white noise process, that is a sequence of independent and identically distributed (i.i.d) random variables with zero mean and a constant variance σ^2. Irrespective of the values of the weights, an MA process is always stationary. The MA(q) model can be expressed in terms of the backshift operator as:

$$Y_t = \theta(B)\varepsilon_t \quad (3.17)$$

where $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ is a polynomial in B of order q. Generally, the random shocks are assumed to follow the traditional normal distribution. The MA(q) process is invertible if the characteristic roots of the polynomial $m^q + \theta_1 m^{q-1} + \theta_2 m^{q-2} + \cdots + \theta_q = 0$ are less than one in absolute terms

3.6 Autoregressive Moving Average (ARMA) Model

Autoregressive (AR) and Moving Average (MA) models can be combined together to form a precise and useful class of time series models, known as the ARMA (p, q) models, where p and

q are the orders of the AR and MA processes respectively (Box *et al.*, 1994). Basically, an ARMA (p, q) model is given as:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \qquad (3.18)$$

where ϕ_i and θ_j are parameters of the autoregressive and moving average components respectively, $i = 1, 2 \ldots p$ and $j = 1, 2, \ldots, q$.

It can also be express as: $\phi(B)Y_t = \theta(B)\varepsilon_t$, where $\phi(B)$ and $\theta(B)$ are polynomials in B of finite order p, q respectively. 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.

3.7 Autoregressive Integrated Moving Average (ARIMA) Model

The ARIMA model is a generalisation of the ARMA model that is defined to incorporate an integrated component to carter for time series data that are non-stationary in nature. In practice many time series data show non-stationary behaviour and such data are made stationary by applying finite differencing of the data points. The backshift operator for the ARIMA (p, d, q) model is expressed as:

$$\phi(B)(1 - B)^d Y_t = \theta(B)\varepsilon_t \qquad (3.19)$$

where p, d and q are integers greater than or equal to zero and denote the order of the autoregressive, integration and moving average parts of the model respectively. The integer d handles the level of differencing.

3.8 Seasonal ARIMA (SARIMA) Model

When a time series data exhibit seasonal behaviour, the ARIMA model is usually not able to capture the behaviour 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 ARIMA$(p, d, q)(P, D, Q)_s$ and can be expressed using the backshift operator as:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B^s)\varepsilon_t \quad (3.20)$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \quad (3.21)$$

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps} \quad (3.22)$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q \quad (3.23)$$

$$\Theta(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs} \quad (3.24)$$

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.

3.9 SETAR Model

Self-Excited Threshold Autoregressive (SETAR) model is a class of the Threshold Autoregressive (TAR) model proposed by Tong (1978) and further studied in Tong and Lim (1980), Tong (1983,1990). 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 (Boero and Marrocu, 2004):

$$y_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{P^{(1)}} \phi_i^{(1)} y_{t-i} + \varepsilon_t^{(1)} & if\, y_{t-d} \le \tau \\ \phi_0^{(2)} + \sum_{i=1}^{P^{(2)}} \phi_i^{(2)} y_{t-i} + \varepsilon_t^{(2)} & if\, y_{t-d} > \tau \end{cases} \qquad (3.25)$$

where $\phi_i^{(1)}$ and $\phi_i^{(2)}$ are the coefficient in lower and higher regime respectively which needs to be estimated; τ 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 ($d < p$); $\left\{\varepsilon_t^{(1)}\right\}$ and $\left\{\varepsilon_t^{(2)}\right\}$ 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 considers 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 (Dijk, 2000). Such conditions has only been established for first-order SETAR model. For effective model selection, we follow the procedure discussed in Franses and van Dijk (2000). The approach of SETAR modeling start with AR (p) model specification and linearity against SETAR model, SETAR model identification, estimation and evaluation of the selected model and then forecasting which is precisely discussed as follows.

3.10 Linearity Test

In order to apply the SETAR model to an observable time series, the series must first be nonlinear in nature. That is the existence of nonlinear behaviour 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 minimise the AIC value, Franses and van Djik (2000). After determining the linear AR (p) model we then test for linearity using a well-known linearity test such as Keenan test and Tsay F-test which are discussed below.

3.10.1 Keenan Test

Keenan test was introduced by Keenan (1985) to detect nonlinearity in an observable time

series. The test is considered as a special case of the RESET test proposed by Ramsey (1969). The avoidance of multicollinearity makes it special. The Keenan test for nonlinearity analogous to Tukey's one degree of freedom for nonadditivity test is motivated by approximating a nonlinear stationary time series by a second-order Volterra expansion which is given by:

$$y_t = u + \sum_{u=-\infty}^{\infty} \theta_u \varepsilon_{t-u} + \sum_{v=-\infty}^{\infty} \sum_{u=-\infty}^{\infty} \theta_{uv} \varepsilon_{t-u} \varepsilon_{t-v} \qquad (3.26)$$

where $\{\varepsilon_t, -\infty < t < \infty\}$ is a sequence of independent and identically distributed with zero-mean random variable. The process $\{y_t\}$ is linear if the double sum of the right-hand side of (3.26) does not exist. Thus we can test the linearity of the time series by testing whether or not the double sum of (3.26) does not exist. That is, the test requires that one distinguish between linearity versus a second-order Voltera expansion, by examining θ_{uv}=0 as well as the coefficients on higher orders. Cryer and Chan (2008) has shown that the Keenan`s test is equivalent to testing if $\eta = 0$ in the multiple regression model (with the constant 1 being absorb in to θ_0):

$$y_t = \theta_0 + \phi_1 y_{t-1} + \cdots + \phi_m y_{t-m} + \eta \hat{y}_t^2 + \varepsilon_t \qquad (3.27)$$

The Keenan's test statistic for the null hypothesis of linearity ($H_0: \eta = 0$) is given as:

$$\hat{F} = \frac{\eta^2 (n - 2m - 2)}{RSS - \eta^2} \qquad (3.28)$$

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 $\hat{F}$ is small ($p - value < \alpha$) or when the value of $\hat{F}$ is greater than the selected critical value of the *F*-distribution with 1 and $n - 2m - 2$ degrees of freedom.

3.10.2 Tsay`s F- test

Tsay (1989) introduce a test for detecting nonlinearity in an observable time series. The test is been considered as more general nonlinear alternative and also a combined version of the nonlinear test of Keenan (1985). The test is further based on arranged autoregression and predictive residuals. For the arranged regression approach, the linear AR (*p*) model is considered in the null against the alternative hypothesis of nonlinear threshold model. For an AR (*p*) regression with *n* observation as $y_t = (1, y_{t-1}, \dots, y_{t-p})\beta + a_t$ for $t = p + 1, \dots, n$ where β is a $(p + 1)$ dimensional vector of coefficients and a_t is the noise. Reference is made to $(y, 1, y_{t-1}, \dots, y_{t-p})$ in a case of data for the AR (*p*) model. Then an arranged auto regression is an autoregression with cases rearranged based on the values of a particular regressor. When we Consider a two regime TAR (2*; p, d*) model with *n* observations, the threshold variable y_{t-d} may assume values $(y_h, \dots, y_{n-d})$, where $h = max\ \{1, p + 1 - d\}$. Let π_i, be the time index of the i^{th} smallest observation of $(y_h, \dots, y_{n-d})$. Then the arranged autoregression with the first *s* cases in the first regime and the rest in the second regime is given by:

$$y_t = \begin{cases} \Phi_0^{(1)} + \sum_{v=1}^{p} \Phi_v^{(1)} y_{\pi_i+d-v} + a_{\pi_i+d}^{(1)} & if\ i \le s \\ \Phi_0^{(2)} + \sum_{v=1}^{P^{(2)}} \Phi_v^{(2)} y_{\pi_i+d-v} + a_{\pi_i+d}^{(2)} & if\ i > s \end{cases} \qquad (3.29)$$

where s satisfies $y_{\pi_s} < \tau_1 \leq y_{\pi_{s+1}}$.The arranged autoregression provides a means by which the observations are separated into two groups such that if the true model is indeed Threshold Autoregressive (2;p,d) process, the observations in a group follow the same linear

autoregressive model. Accordingly the separation of the observation does not require knowing the precise value of τ_1 and only the number of observation in each group depends on τ_1. But since the threshold value is unknown, however the sequential least square estimates $\widehat{\Phi}_v^{(1)}$ are consistent for $\Phi_v^{(1)}$ if there is sufficiently large number of observations in the first regime. For the arranged autoregression, let $\hat{\beta}_m$ be the vector of least squares estimates based on the first *m* cases P_m, the associated $X'X$ inverse matrix, and x_{m+1} the vector regressor of the next observation to enter the autoregression $y_{d+\pi_{m+1}}$ Then the recursive least squares estimates can be computed efficiently by:

$$\hat{\beta}_{m+1} = \hat{\beta}_m + K_{m+1}\left[y_{d+\pi_{m+1}} - x'_{m+1}\,\hat{\beta}_m \right] \tag{3.30}$$

$$D_{m+1} = 1.0 + x'_{m+1} P_m x_{m+1} \tag{3.31}$$

$$K_{m+1} = P_m x_{m+1} / D_{m+1}, \tag{3.32}$$

and

$$P_{m+1} = \left(1 - P_m \frac{x_{m+1} x'_{m+1}}{D_{m+1}}\right) P_m \tag{3.33}$$

and the predictive and standardized predictive residuals is given by:

$$\hat{a}_{d+\pi_{m+1}} = y_{d+\pi_{m+1}} - x'_{m+1}\,\widehat{\beta_m} \tag{3.34}$$

and

$$\hat{e}_{d+\pi_{m+1}} = y_{d+\pi_{m+1}} / \sqrt{D_{m+1}} \tag{3.35}$$

for fixed *p* and *d*, the effective number of observations in arranged autoregression is *n – d –h*

+1. Assuming the recursive autoregressions begin with *b* observation so that there are *n* – *d* –*b* - *h* predictive residuals available. We do the least squares regression

$$\hat{e}_{\pi_{1+d}} = \omega_0 + \sum_{v=1}^{p} \omega_v \, y_{\pi_i+d-v} + \varepsilon_{\pi_i+d} \tag{3.36}$$

For $i = b + 1,\ldots,n - d - h + 1$, and compute the associated *F* statistic under the null hypothesis of linear AR (*p*).

$$\hat{F}(p,d) = \frac{(\sum \hat{e}_t^2 - \sum \hat{\varepsilon}_t^2)/(p+1)}{\sum \hat{\varepsilon}_t^2/(n-d-b-p-h)}, \tag{3.37}$$

where the $\hat{\varepsilon}_t$ is the square residual of equation (3.36) and the argument (*p*, *d*) of $\hat{F}$ is used to signify the dependence of the *F*-ratio on *p* and *d*. Supposed that y_t, is a linear stationary autoregressive process of order *p*, then for large *n* the statistic $\hat{F}(p,d)$ follows an asymptotic *F* distribution with *p* + 1 and *n* – *d* – *b* – *p* – *h* degrees of freedom. The null hypothesis of linearity is rejected if the *p*-value of $\hat{F}(p,d)$ is greater than the selected critical value of the *F*-distribution with *p*+1 and *n* – *d* – *b* – *p* - *h* degrees of freedom.

3.11 Criterion for Model Selection

In order to obtain the most adequate model that best describes a time series data, it is important for model selection criteria to be carried out. This is because there is the possibility of two or more models to compete in the selection of the best model for the study. The Akaike Information Criterion (AIC), the Akaike Information Criterion corrected (AICc) and the Bayesian Information Criterion (BIC) are the model selection criteria that were employed in this study to select the most adequate model. The information criteria include a penalty that is

an increasing function of the number of parameters. The penalty discourages overfiting, that is, increasing the number of parameters improves the goodness of fit. The best model is the one

with the smallest AIC, AICc or BIC values, given a set of possible models. The AIC, AICc, and BIC are given by

$$\text{AIC} = 2k - 2In(L) \tag{3.38}$$

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1} \tag{3.39}$$

$$\text{BIC} = \log(\sigma_e^2) + \frac{k}{n}\log(n) \tag{3.40}$$

where;

k represents the number of parameters in the model

L denotes the maximised value of the likelihood function

n is the number of observations in the data

σ_e^2 is the error variance.

However, in the case of the Self-Excited Threshold Autoregressive (SETAR) approach for modelling, the AIC and BIC for the AR model in the two regimes as defined by Tong (1990) is given by:

$$\text{AIC}(p_1, p_2) = n_1 \ln \hat{\sigma}_1^2 + n_2 \ln \hat{\sigma}_2^2 + 2(p_1 + 1) + 2(p_2 + 1) \tag{3.41}$$

$$BIC(p_1, p_2) = n_1 \ln \hat{\sigma}_1^2 + n_2 \ln \hat{\sigma}_2^2 + (p_1 + 1) \ln n_1 + (p_2 + 1) \ln n_2 \qquad (3.42)$$

where

$n_j, j = 1,2$ is the number of observations in the j^{th} regimes and

$\hat{\sigma}_j^2, j = 1,2$ is the variance of the residuals in the j^{th} regimes

p_1 and p_2 are the selected lags order in regime 1 and 2 respectively for which the information criterion is minimised.

3.12 Model Diagnostics

After a model has been built, it is important to diagnose the model in order to ensure that it truly follows the real time series observations. When these checks are done the model can be used to make meaningful inferences. The Ljung-Box and ARCH-LM tests were employed in this study to diagnose the developed models as discussed as follows.

3.12.1 Ljung-Box Test

Serial correlation poses a major challenge to researchers when fitting time series models. For this reason, Ljung and Box (1978) proposed a test that is used to determine the presence or absence of serial correlation in a time series up to a given order say *k*. The test assumes that the residuals do not contain serial correlation up to order *k.* It tests the overall residuals based on a given number of lags. Hence it is considered as a portmanteau test. The test procedure is as follows;

H_0: The data do not contain serial correlation up to order *k*.

H_A: The data contains serial correlation up to order *k*.

The test statistic is given by;

$$Q_h = n(n+2)\sum_{k=1}^{h}\frac{r_k^2}{n-k} \tag{3.43}$$

where;

r_k^2 represent the residual autocorrelation at lag *k*

n is the number of residuals

h is the number of lags being tested

We reject the null hypothesis if Q_h is greater than the chi-square table value. The model is therefore considered adequate when the *p*-value associated with Q_h is large; otherwise the whole estimation process has to be repeated again in order to get the most adequate model.

3.12.2 ARCH-LM Test

When the variance of the residuals is not constant, the issue of conditional heteroscedasticity is one of the key problems that a researcher is likely to encounter when fitting models. To ensure that the fitted model is adequate, the assumption of constant variance must be achieved. The ARCH-LM test proposed by Engle (1982) was used to test for the presence of conditional heteroscedasticity in the model residuals. The test procedure is as follows;

H_0: There is no heteroscedasticity in the model residuals

H_1: There is heteroscedasticity in the model residuals

The test statistic is

$$\mathrm{LM} = nR^2 \tag{3.44}$$

where n is the number of observations and R^2 is the coefficient of determination of the auxiliary residual regression.

$$e_t^2 = \beta_0 + \beta_1 e_{t-1}^2 + \beta_2 e_{t-2}^2 + \ldots + \beta_q e_{t-q}^2 + v_t \tag{3.45}$$

where e_t is the residual. The null hypothesis is rejected when the *p*-value is less than the level of significance and is concluded that there is heteroscedasticity.

3.13 Diebold-Mariano Test

Sometimes when comparing the forecasting accuracy of two 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 significant difference in the forecasting accuracy of any two competing models, the Diebold and Mariano (1995) test of equal forecasting accuracy was 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.

$$\mathrm{S}_1 = \quad \left[\hat{\mathbb{V}}(\bar{\mathrm{d}})\right]^{\frac{-1}{2}} \bar{\mathrm{d}} \tag{3.46}$$

Where $\bar{\mathrm{d}}$ is the mean of the coefficient of $\mathrm{d_t}$, which is the difference between the sets of squared forecast errors from two competing models, $\mathrm{d_t} = \mathrm{e}_{1\mathrm{t}}^2 - \mathrm{e}_{2\mathrm{t}}^2$.

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

CHAPTER FOUR

ANALYSIS AND DISCUSSION OF RESULTS

4.0 Introduction

This chapter deals with the analysis and discussion of the results obtained from the study. The chapter is sub-divided into three main headings namely; preliminary analysis, further analysis and discussion of results.

4.1 Preliminary Analysis

This section explains the descriptive statistics of the data on monthly pneumonia cases.

The maximum and minimum values of the cases for the entire study period were 35.00 and 0.00 respectively. Moreover, the average pneumonia cases was 17.77. The coefficients of variation (CV) for the pneumonia cases was 38.74%. Furthermore, the pneumonia cases recorded for the entire period was found to be negatively skewed and leptokurtic in nature, that is the cases were closely distributed around their mean value as shown in Table 4.1.

Table 4.1: Descriptive Statistics of Pneumonia Cases

Variable	Mean	Min	Max	CV (%)	Skewness	Kurtosis
Pneumonia cases	17.774	0.000	35.000	38.740	-0.150	0.080

An investigation of the pneumonia cases for the various months revealed that, the highest average pneumonia (19.4) case occurred in the month of March as shown in Table 4.2. The least average pneumonia (16.63) occurred in the month of May. In terms of the minimum and maximum pneumonia cases, the months of January, February, March, April, May and June have the minimum pneumonia cases and the month of September recorded the maximum pneumonia case. The month of June has the largest variability followed by May as shown by their coefficient of variations in Table 4.2. Again, it was observed that pneumonia cases were positively skewed for the months of June, July, September, October and November while it was negatively skewed for the rest of the months. Moreover, the months of January, February, March, April, May, were found to be leptokurtic, indicating that pneumonia cases were closely distributed around their mean value, while those of the remaining months were platykurtic in nature, also demonstrating cases been widely distributed around their mean value.

Table 4.2: Monthly descriptive statistics of Pneumonia Cases

Month	Mean	Min	Max	CV (%)	Skewness	Kurtosis
January	17.810	0.000	28.000	40.910	-0.580	1.050
February	18.310	0.000	32.000	41.620	-0.410	1.280
March	19.440	0.000	29.000	35.490	-1.420	3.270
April	16.810	0.000	28.000	42.790	-0.720	0.520
May	16.630	0.000	30.000	43.030	-0.450	1.030
June	16.690	0.000	31.000	53.320	0.050	-0.520
July	16.750	8.000	30.000	35.720	0.550	-0.220
August	17.560	9.000	26.000	34.160	-0.090	-1.610
September	19.190	9.000	35.000	41.700	0.510	-0.500
October	17.630	9.000	32.000	38.420	0.700	-0.220
November	17.730	7.000	30.000	38.100	0.410	-0.480
December	18.800	11.000	25.000	25.130	-0.410	-0.670

A time series plots of the pneumonia cases depicts that the series fluctuates with time in an increasing and decreasing manner as shown in Figures 4.1.

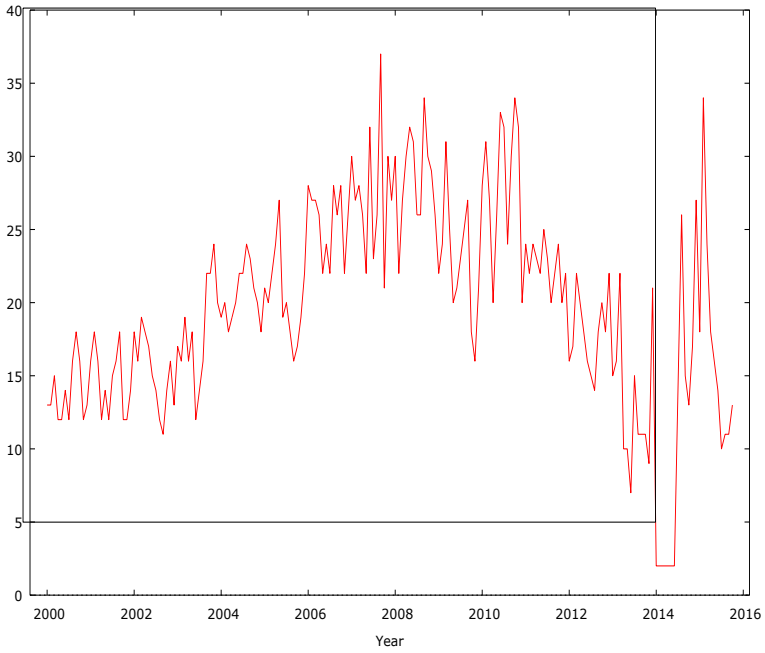

Figure 4.1: Time series plot of Pneumonia cases

The nature of trend characterising the pneumonia cases overtime was investigated using the linear, quadratic, log-linear and log-quadratic trend models as shown in Table 4.3. The log-quadratic trend model was observed as the best since it had the least AIC and BIC values.

Table 4.3: Trend analysis of Pneumonia cases

Model	AIC	BIC
Linear	1275.256	1281.750
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 parameters of the log-quadratic trend models for the pneumonia cases were estimated as shown in Table 4.4. 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;

$$\ln Pnc = 2.3536 + 0.0191t - 0.0001t^2 \quad (4.1)$$

where, t=Time and Pnc means pneumonia cases.

Table 4.4: Estimated parameters of the Log-quadratic trend Model

Variable	Coefficient	Standard error	T-statistic	*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			

**: Means significant at the 5% significance level

4.2 Further Analysis

4.2.1 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 Figures 4.2.

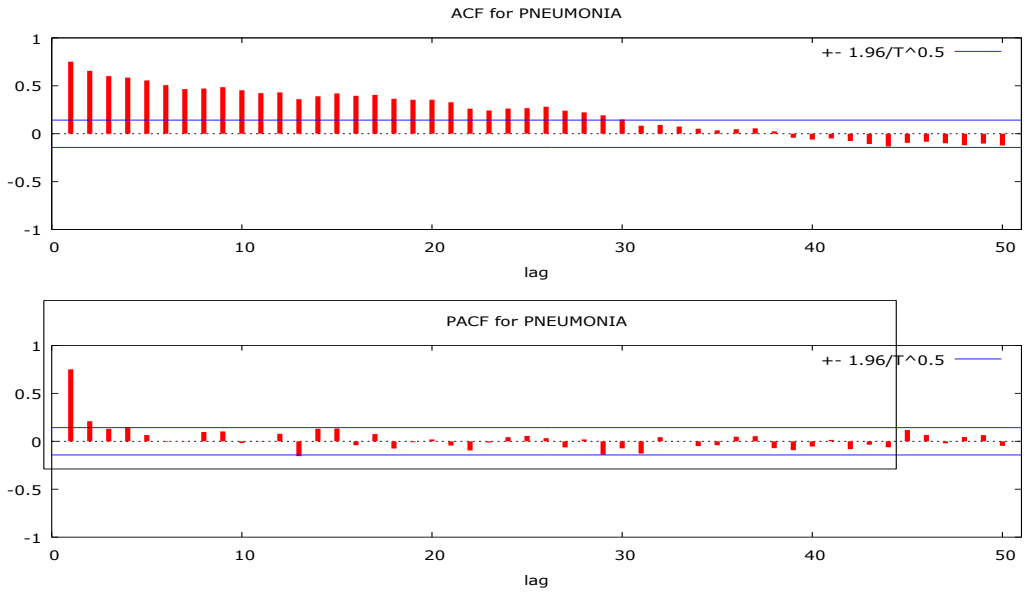

Figure 4.2: ACF and PACF plot of Pneumonia cases

To further confirm the non-stationarity of the series, the KPSS and ADF test for unit root were carried out on the original data. Using the KPSS test, the results in Tables 4.5 revealed that the calculated value was greater than the critical value at 5% level of significance. The null hypothesis of stationarity was therefore rejected indicating the series was not stationary.

Table 4.5: 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 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 Tables 4.6.

Table 4.6: ADF test of Pneumonia cases

Test	Constant		Constant+Quadratic Trend	
	Test Statistic	P-value	Test Statistic	P-value
ADF	-1.4342	0.567	-1.4314	0.852

The series was transformed logarithmically in order to stabilise 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 Table 4.7 and 4.8 respectively.

Table 4.7: KPSS test of differenced series

Test	Test Statistic	Critical value
KPSS	0.0341	0.464

Table 4.8: ADF test of differenced series

Test	Constant		Constant + Quadratic Trend	
	Test Statistic	P-value	Test Statistic	p-value
ADF	-5.4783	0.0000	-5.4829	0.0000

After obtaining the order of integration of the Pneumonia cases, the order of the Autoregressive and Moving Average components was determined by using Box and Jenkins (1976) approach based on the ACF and PACF plots. The ACF plot in Figure 4.3 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 4.9. 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.

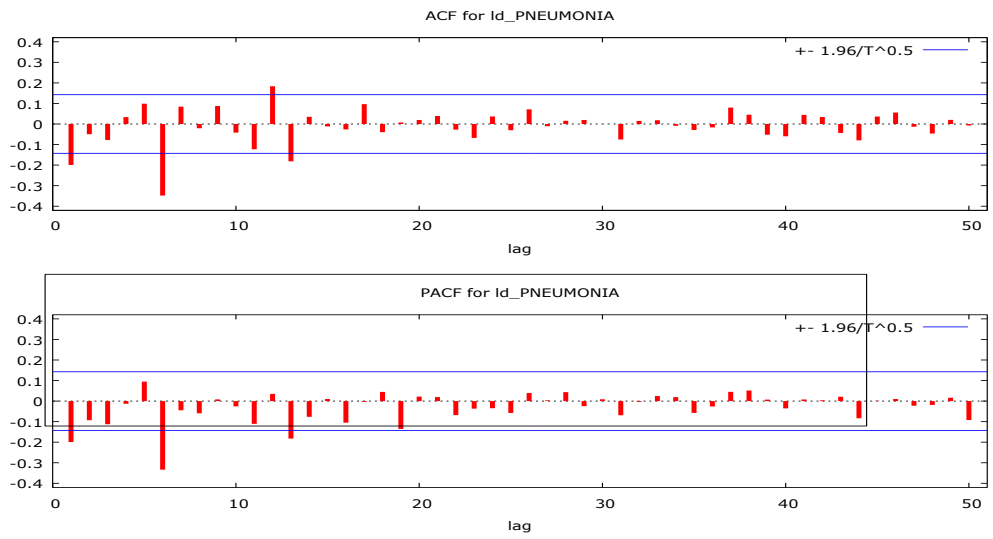

Figure 4.3: ACF and PACF plot of differenced series

Table 4.9 Tentative SARIMA models

Model	AIC	AIC_C	BIC
SARIMA (1, 1, 0)(0, 0, 1)$_{12}$	95.55	95.62	102.03
SARIMA (1, 1, 1)(0, 0, 1)$_{12}$	86.35	86.48	96.48
SARIMA (0, 1, 1)(0, 0, 1)$_{12}$	93.83	93.89	100.31
SARIMA (1, 1, 0)(0, 0, 1)$_{12}$	93.83	93.96	103.56
SARIMA (1, 1, 1)(0, 0, 1)$_{12}$	83.5*	83.71*	96.08*
SARIMA (0, 1, 1)(0, 0, 1)$_{12}$	92.76	92.89	102.48

*: Means best based on the selection criteria

Using the method of maximum likelihood, the estimated parameters of our derived model are shown in Table 4.10. The SARIMA (1, 1, 1)(0, 0, 1)$_{12}$ model can be expressed in terms of backshift operator as;

$$(1 - 0.755B)(1 - B)\ln pnc = (1 - 0.959B)(1 + 0.169B^{12})\varepsilon_t \qquad (4.2)$$

It is observed from Table 4.10 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 4.10: Estimates of parameters for SARIMA (1, 1, 1)(0, 0, 1)$_{12}$

Variable	Coefficient	Standard error	Z-statistic	P-value
θ_1	-0.959552	0.0327374	-29.3106	0.00001
Θ_1	0.169451	0.0764908	2.2153	0.0267
ϕ_1	0.754616	0.0689403	10.9459	0.00001

The selection of the best model among competing models to fit a data in time series analysis depends largely on the performance of the residuals of the model. One of the assumptions of SARIMA model is that for a good model, the residual must follow a white noise process. That is the residuals have zero mean, constant variance and uncorrelated. It was observed from the diagnostic plot in Figure 4.4 that the standardised residuals of the model have zero mean and constant variance. Also, the ACF of the residuals depicts that the autocorrelation of the residuals are all zero that is they are uncorrelated. In addition, the Ljung-Box statistic clearly shows that the *p*-values of the test statistic exceed the 5% level of significance for all lag orders which implies that there is no significant departure from white noise for the residuals.

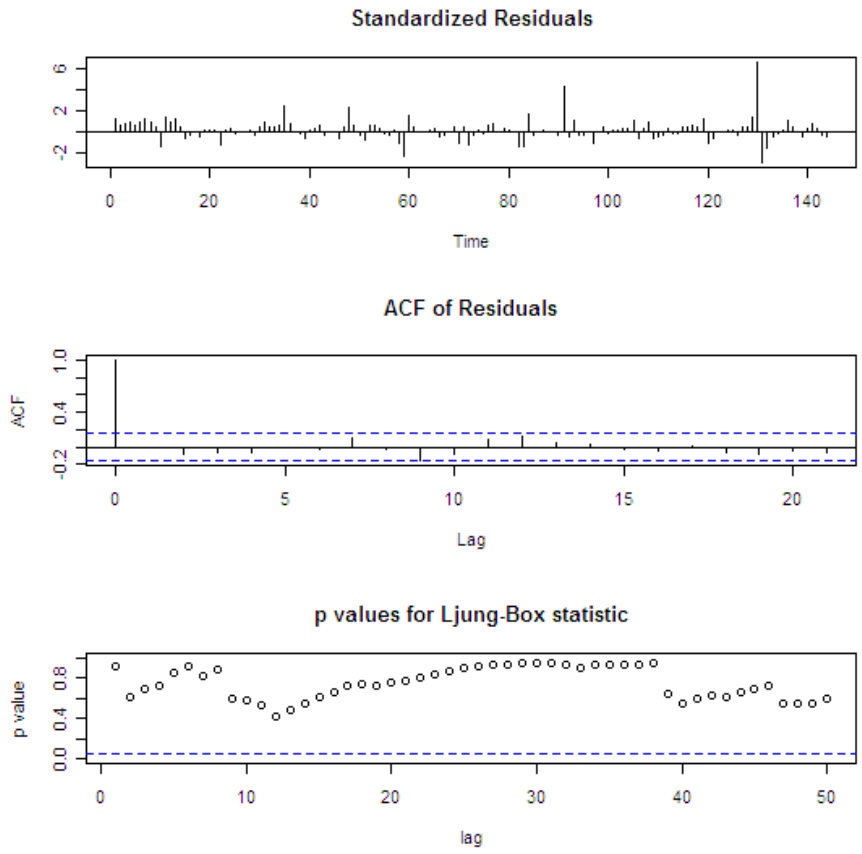

Figure 4.4: Diagnostic plot of SARIMA (1, 1, 1)(0, 0, 1)$_{12}$

To ensure that the fitted model is adequate, both the Ljung-Box test and ARCH-LM test were performed. The Ljung-Box test and ARCH-LM test as shown in Table 4.11 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 all the test statistics were insignificant at the 5% significance level. This implies that the residuals of the model was uncorrelated, thus have zero mean and constant variance overtime; 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 4.11 Residuals diagnostic test for SARIMA (1, 1, 1)(0, 0, 1)$_{12}$

	Ljung-Box Test		ARCH-LM Test	
Lag	**Test statistic**	**p-value**	**Test statistic**	**p-value**
12	20.8664	0.05237	35.1563	0.4422
24	28.2171	0.251	41.1187	0.6161
36	31.3243	0.6905	40.288	0.2862
48	38.693	0.8288	63.6418	0.6471

4.2.2 Fitting the SETAR model

In this section the 2 regime Self Excited Threshold Autoregressive (SETAR) model approach was used to model and forecast the pneumonia cases. In the modelling cycle, the approach presented in Franses and Dijk (2000) was adopted in fitting the model.

4.2.2.1 Linearity Test

In order to model a time series with SETAR model, the series must be nonlinear, hence we have to test for the existence of nonlinearity in the pneumonia cases. To test for nonlinearity in the series we first specifies linear AR (*p*) model. Using Akaike information criterion, we found 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 Tsay *F*-test and the Keenan1-degree test to test for linearity against the alternative of nonlinearity for the Keenan test. The *F*-test of Tsay has the alternative of threshold-type nonlinearity. Both linearity tests depend on the linear AR (4) model selected. Table 4.12 below summarizes the results from the Tsay and 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. Also in the Tsay test, we reject the null hypothesis of no threshold nonlinearity since the P-value is less than 5% significant level.

From both test about the nonlinearity of our data we conclude that the pneumonia cases is nonlinear and it can be well explained by the regime switching model as compare to the simple linear model.

Table 4.12: Linearity test

Test	Test statistic	*P*-value	Decision
Keenan 1-degree	6.24	0.01	Linearity rejected
Tsay	1.83	0.02	No threshold nonlinearity rejected

After confirming that the data is nonlinear, we then identify the specific SETAR model that best fit the data. We do this by determining the Autoregressive lag order P in each regime and the threshold variable y_{t-d} where *d* represent the delay parameter. We choose the model with *P* lag order for both regimes and y_{t-d} threshold variable with the minimal AIC value by performing a grid search on all possible combinations of SETAR models that can be fitted to the data. The grid search of the possible models combinations are illustrated in Table 4.13.

Table 4.13: Grid search for the best model

Lag	ML(mL)	MH(mH)	Threshold delay(d)	Threshold variable	Pooled-AIC	Model
1	1	1	1	1.176	-154.732	SETAR(2, 1,1)
2	1	1	2	1.176	-162.393	SETAR(2, 2,2)
3	1	1	1	1.255	-160.534	SETAR(2, 3,1)
4	1	1	3	1.255*	-165.422*	SETAR(2, 4,3)*
5	1	1	1	1.279	-154.073	SETAR(2, 5,1)
6	1	1	2	1.176	-151.940	SETAR(2, 6,2)
7	1	1	1	1.204	-151.108	SETAR(2, 7,1)
8	1	1	2	1.204	-149.700	SETAR(2, 8,2)
9	1	1	1	1.230	-148.341	SETAR(2, 9,1)
10	1	1	2	1.279	-145.793	SETAR(2, 10,2)

*: Means best based on the selection criteria

After performing a grid search on all possible combination 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 have a minimum AIC value which is presented in Table 4.14.

Table 4.14: AIC for the selected SETAR Model

Model	AIC	BIC
SETAR(2: 4, 3)	-165.42	90.06

After we have found that SETAR (2; 4, 3) model with threshold variable y_{t-3} as the best model that fit the data well since it has the minimum value for AIC. Further assessment on the forecast ability of the model was done following the approach of Franses and Dijk (2000). Table 4.15 below present the estimated parameters of the selected SETAR model with the corresponding threshold value. The corresponding model for SETAR (2: 4, 3) with 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;

$$y_t = \begin{cases} 0.295 + 0.727y_{t-1} + 0.008\, y_{t-2} - 0.142y_{t-3} + 0.087y_{t-4} & if\ y_{t-3} \leq 1.255 \\ -0.069 + 0.484y_{t-1} + 0.085y_{t-2} + 0.300y_{t-3} + 0.171y_{t-4} & if\ y_{t-3} > 1.255 \end{cases} \quad (4.3)$$

Table 4.15: Estimates of parameters for SETAR (2; 4, 3)

	Low Regime			High Regime		
Coefficient	**Estimate**	**Std error**	**t-value**	**Estimate**	**Std error**	**t-value**
Constant	0.294926	0.079414	3.7138	-0.069301	0.251735	-0.2753
ϕ_1	0.726973	0.082488	8.8130	0.484075	0.164587	2.9412
ϕ_2	0.077171	0.101976	0.7568	0.085256	0.187513	0.4547
ϕ_3	-0.141584	0.117673	-1.2032	0.300137	0.176860	1.6970
ϕ_4	0.086578	0.110297	0.7849	0.170647	0.188192	0.9068
Threshold value		1.255				
Proportion		43.55%			56.45%	

After the parameters of the SETAR model have been estimated, we check the residuals of the model for best fit. That is we check for nonexistence of serial autocorrelation, zero mean and constant variance of the residuals. We used the ARCH-LM test to check for constant variance of the residuals. Ljung-Box test was also used to check for serial correlation. From the results as shown in Table 4.16, we fail to reject the null hypothesis of the two test for SETAR (2; 4, 3) model since their *P*-values were greater than the 5% significant level.

Table 4.16 Residuals diagnostic test for SETAR (2; 4, 3)

	ARCH-LM		Ljung Box Test	
lag	**Test Staistic**	**p-value**	**Test Statistic**	**p-value**
12	18.2422	0.1085	14.6184	0.263
24	38.9226	0.2782	32.8826	0.1066
36	39.0165	0.3357	41.1826	0.2542
48	44.2422	0.6276	50.6098	0.3709

4.2.3 Comparative Analysis of the SARIMA and SETAR Models

The main task of this research work is to compare the forecast performance between the linear SARIMA model and the non-linear SETAR model. Once the selected models from both linear and nonlinear have been shown to satisfy all the model 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 4.17, it can be revealed that most accuracy tests supports SETAR (2; 4, 3) model which has the minimum value of BIC, AIC, MSE, RMSE and MAPE respectively.

Table 4.17: 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*

*: Means best based on the measure of accuracy

Though the nonlinear SETAR model outperform the linear SARIMA model as suggested by the forecast measures, it is interesting to know whether there is significant difference in forecast from the two models. Using the approach of Diebold and Mariano (1995), 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 compare to the selected SARIMA model. The results from the test as presented in Table 4.18 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 4.18: 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 4.19 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 depicts the behaviour of the pneumonia cases over time even though the values of the observed and expected are not exactly the same.

Table 4.19: 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 as shown in Tables A1 and A2 respectively in the appendix. 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

4.3 Discussion of Results

The descriptive statistics of the pneumonia cases depict a leptokurtic in nature which gives information about how the cases were closely distributed around their mean value. The monthly distribution of the pneumonia cases clearly showed that the highest numbers of cases were recorded in the month of March. This may be attributed to the beginning of the change of dry season to rainy season which provides non favourable condition for those whose genetic makeup is non-resistance to change of weather and also provides room for bacterial and viral infections. An investigation was carried out on the data set to determine whether or not a trend exists in the data set. The results clearly showed that there was a log-quadratic trend in the data set. The log-quadratic trend model indicates that pneumonia cases was decreasing at a certain constant quadratic rate of about 0.00001.

A unit root test conducted to investigate the stationarity of the pneumonia cases clearly revealed that the data was not stationary. This was affirmed by the time series plot, ACF and PACF plots of pneumonia cases. The series was then transformed logarithmically and first differenced. From the results, the ADF test revealed that the transformed first differenced series was stationary. The stationary series was then used to investigate the comparative analysis of the linear SARIMA model and nonlinear SETAR model in predicting pneumonia cases in the region. Further, before fitting the SETAR model to the pneumonia cases the series must be nonlinear, hence, the existence of nonlinearity in the pneumonia cases was tested by the use of Tsay *F*-test and the Keenan1-degree test. The results shows that Keenan1-degree test reject the null hypothesis of linearity since the p-value was less than the 5% significant level. Also in the Tsay test, the null hypothesis of no threshold nonlinearity was rejected since the p-value was less than 5% significant level. Hence we conclude that the data follows a threshold nonlinearity.

Forecast values are of importance for decision making and policy formulation. As described by Box and Jenkins (1976), forecasting provide basis for economic and business planning, inventory and production control and optimization of industrial processes. Obtaining a good model that produce best forecast is the core point of every researcher hence, it was imperative to forecast the incidence of pneumonia cases. This will serve as a guiding tool to the Government of Ghana, individuals and stakeholders in the health sector in strengthening existing control measures and also implement new ones with the aim of reducing the cases of the disease to the barest minimum or possibly eliminate it. Two forecasting models were developed to aid in the monthly prediction of the pneumonia cases. The two models were the SARIMA $(1, 1, 1)(0, 0, 1)_{12}$ model and the SETAR (2; 4, 3) model.

The diagnostic checks carried out on these models proved that both models were adequate for predicting the monthly pneumonia cases in the region. To identify which of these models has the best forecasting accuracy measures, a comparison of the forecasting accuracy measures of these models were made. The SETAR (2; 4, 3) model with the least information criteria happens to have the least measures of accuracy. Therefore, a Diebold-Mariano test was performed to check whether there were significant difference in the forecasting performance of the two models. The results of the test revealed that there is no significant difference in the forecasting performance of the models. Further assessment on cross validation of the models with chi-square revealed that the models were fit for forecasting. Hence, it was concluded that both models were good for forecasting pneumonia cases. A twenty six months forecast with these models revealed an increasing pattern as in the case of the SARIMA model and a constant trend in terms of the SETAR model. This increasing trend for pneumonia cases as in the case of the SARIMA model as evidence from the forecast result could be worrying to the health development of the state as it could lead to loss of lives. Whiles, the constant trend as portrays by the forecast values of the SETAR model should not be taken for granted because it might not be sustainable in the long run.

4.4 Conclusion

This chapter dwelled on the analyses and discussion of the results obtained. It presented the major findings of the study in detailed and concise manner.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.0 Introduction

This chapter presents the conclusion and recommendations of the study.

5.1 Conclusion

In this study, the monthly number of inpatients pneumonia cases, from January 2000 to October 2015 was studied. Before fitting model to the pneumonia cases, the monthly characteristics of the series were examined. The carefull examination of the series revealed that pneumonia cases was 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 nonlinear SETAR model, the forecast measures suggest that the nonlinear SETAR model outperform the linear SARIMA model. Also, the forecast performance of the nonlinear SETAR models is superior to that of the linear SARIMA model in predicting pneumonia cases in Northern Region of Ghana. Predicted Pneumonia cases were made beyond the period under consideration based on the developed models. The Ghana Health Service (GHS), 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 performance of these models, strengthening and maintenance of the existing health systems in order to make the use of these models more reliable.

5.2 Recommendations

Based on the findings of this research work, the following recommendations were made;

i. The results revealed that the nonlinear 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 in the country.

ii. The log-quadratic trend model depicts a decreasing levels in the number of pneumonia cases for a unit change in time. This decreasing levels does 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 as in the case of the recent outbreak of pneumococcal meningitis in some parts of the country.

iii. This study compared the nonlinear SETAR model and the linear SARIMA model in predicting pneumonia cases in the region. It is therefore recommended that further studies should be carried out by comparing the nonlinear SETAR model with other linear model to see which one would outperforms the other since the nonlinear SETAR model is the best model in this study.

iv. It is also recommended that the MoH and GHS 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 advert future loss of lives as in the case of the recent pneumococcal meningitis in the country.

REFERENCES

Addo, P. M., (2014). Multivariate Self-Exciting Threshold Autoregressive Models with exogenous Input. Unpublished PhD in Economics, Department of Economics, University of Paris, Sorbonne.

Aidoo, E., (2010). Modelling and forecasting inflation rate in Ghana: An application of SARIMA models. (Unpublished Thesis). School of Technology and Business Studies, Hogskolen, Dalarma.

Akaike, H., (1974). A New Look at the Statistical Model Identification. *IEEE Translation on Automatic Control,* AC-**19**:716-723.

Akter, S., and Rahman, S., (2010). Agribusiness Forecasting with Univariate Time Series Modelling Techniques: The Case of a Dairy Cooperative in the UK. *Journal of Farm Management*, **13**(11):747-764.

Akhter, T., (2013). Short-term Forecasting of Inflation in Bangladesh with Seasonal ARIMA Processes. Munich Personal Archive **Paper No** 43729.

Amos, C., (2010). Time Series Modelling with Applications to South African Inflation Data. Unpublished master's thesis. University of Kwazulu Natal.

Basikhasteh, M., mohamad, A. M., and Maryam, S., (2014). Time series forecasting model for Iranian gold price. *Indian Journal of Science Research*, **7** (1): 1227-1232.

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.

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.

Box, G. E. P., and Jenkins, G. M., (1976). Time Series Analysis Forecasting and Control. Holden – Day, San-Francisco.

Briet, O., Penelope, V., Dissanayake, M., Gunawardena, G., Galappaththy, N., and Priyanie H., (2008). Models for short term malaria prediction in Sri Lanka. *Malaria Journal,* **7**: 7-76.

Caraiani, M., and Acatrinei, P., (2011). Modelling and forecasting the dynamics in Romanian stock market indices using threshold models. *Romanian Journal of Economic Forecasting*-**2/2011**.

Clements, M., and Smith, J., (1997). The Performance of Alternative Forecasting Methods for SETAR Models. *International Journal of Forecasting*, **13:** 463-475.

Chen, H., and Trajkovic, L., (2004). Prediction of Traffic in a Public Safety Network. School of Computing Science. Simon Fraser University. Burnaby. Canada.

Campbell, H., and Nair, H., (2015). Child pneumonia at a time of epidemiological transition. *The Lancet Global Health*, **3**: 65-66.

Chikobvu, D., and Sigauke, C., (2012). Regression-SARIMA modelling of daily peak electricity demand in South Africa. *Journal of energy in Southern Africa*, **3**: 1-23.

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

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

Dickey, D. A., and Fuller, W. A., (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit-root. *Journal of American Statistical Association*, **74** (366): 427-431.

Engle, R. F., (1982). Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation. *Econometrica,* **50**: 987–1008.

Ekezie, D. D., Opara, J., and Okenwe, I., (2014). Modelling and Forecasting Malaria Mortality Rate using SARIMA Models. A Case Study of Aboh Mbaise General Hospital, Imo State Nigeria. *Science Journal of Applied Mathematics and Statistics*, **2**(1): 31-41.

Erdogdu, E., (2007). Electricity Demand Analysis Using Cointegration ARIMA Modelling. A Case Study of Turkey. *Energy Policy*, **35**: 1129-1146.

Fannoh, R., Orwa, G. O., and Mungatu, J. K., (2014). Modeling the inflation rates in Liberia SARIM approach. *International Journal of Science and Research,* **3**(6): 1360 – 1367.

Feng, H., Guangcai, D., Rongguang, Z., and Weidong, Z., (2014). Time Series Analysis of Hand-Foot-Mouth Disease Hospitalization in Zhengzhou: Establishment of Forecasting Models Using Climate Variables as Predictors. *Journal of Infectious Diseases,* **9**(1): 8000-8791.

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

Gharleghi, B., Abu, H., Shaari, M. D., Nr, T., (2014). Application of the Threshold Model for Modelling and Forecasting of Exchange Rate in Selected ASEAN Countries. Faculty of Economics and Management, University of Kebangsaan, Malaysia. Sains Malaysiana, **43**(10): 1609–1622.

Gikungu, S. W., Waititu, A. G., Kihoro, J. M., (2015). Forecasting Inflation Rate in Kenya Using SARIMA Model. *American Journal of Theoretical and Applied Statistics*, **4** (1): 15-18.

Hall, C. B., (2010). Respiratory syncytial virus in young children. *The Lancet,* **375**:1500-150.

Ismail, Z., Suhartono, Y.A., and Efendi, R., (2009). Intervention model for analyzing the impact of terrorism to tourism industry. *Journal of Mathematics and Statistics*, **5**(4): 322–329.

Izadnegahdar, R., Cohen, L., Klugman, K., Qazi, S., (2013). Childhood pneumonia in developing countries. *Lancet Respiratory Medicine*, **1**:574-584.

Jamaludin, A. R., Yusof, F., and Ibrahim, L. K., (2015). Temporal Dynamics of Trend in Relative Humidity with RH-SARIMA Model. *Modern Applied Science,* **15**: 3-9.

Kahraman, U. M., Aydiner, O., (2013). An application of the dollar and gold prices in Turkey with multivariable SETAR model. *Dumlupinar University Journal of Social Science*, **16:**133-148.

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

Khan, A. J., Hussain, H., Omer, S. B., Chaudry, S., (2009). High incidence of childhood pneumonia at high altitudes in Pakistan: a longitudinal cohort study. *Bulletin of the World Health Organization***, 87**:193-199.

Kibunja, H., Kihoro, J., and Orwa, G., (2014). Forecasting Precipitation Using SARIMA Model: A Case Study of Mt. Kenya Region. International Institute for Science, Technology and Education, **4**(11): 50-58.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y., (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit-root. How sure are we that Economic Time Series have a unit-root? *Journal of Econometrics*, **54**: 159-178.

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

Li, X., (2009). Applying GLM Model and ARIMA Model to the Analysis of Monthly Temperature of Stockholm. D-level Essay in Statistics. Department of Economics and Society. Dalarna University. Sweden.

Madhi, S. A., Levine, O. S., Hajjeh, R., Mansoor, O. D., (2008). Vaccines to prevent pneumonia and improve child survival. *Bulletin of the World Health Organization*, **86**: 365-370.

Meyer, E., Juergen, B., Christian, S., Egid, S., Barbara, S. B., Petra, G., and Frank, S., (2007). Modified guidelines impact on antibiotic use and costs: duration of treatment for pneumonia in a neurosurgical ICU is reduced. *Journal of Antimicrobial Chemotherapy,* **59**: 1148–1154.

Monto, A. S., (1994). Studies of the community and family: acute respiratory illness and infection. *Epidemiological Review*, **16:** 351-373.

Modarres, R., and Ouarda, T. B. M. J. (2013). Modelling heteroscedasticty of stream flow time series. *Hydrological Sciences Journal*, **58**(1): 54-64.

Myriam, G., Philippe, Q., Joel, G., Sylvie, C., Guy, L., (2011). Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors. *Journal of infectious disease*, **2334:** 111-166.

Nasiru, S., and Sarpong, S., (2012). Empirical Approach to Modelling and Forecasting Inflation in Ghana. *Current Research Journal of Economic Theory, **4***(3): 83-87.

Nobre, F. F., Monteiro, A. B. S., Telles, P. R., and Williamson, D., (2001). Dynamic Linear Model and SARIMA: A Comparison of their Forecasting Performance in Epidemiology, *Statistics in Medicine,* **20**(20): 3051-3069.

ODuro, F. k., Harris, E., and Darkwah, K. F., (2012). SARIMA Time Series Model Application to Microwave Transmission of Yeji- Salaga (Ghana) Line-of-Sight Link. *International Journal of Applied Science and Technology*, **2**(9): 41-51.

Omane-Adjepong, M., Oduro, F. T., and Oduro, S. D., (2013). Determining the Better Approach for Short Term Forecasting of Ghana's Inflation: Seasonal ARIMA Verses Holt-Winters. *International Journal of Business, Humanities and Technology,* **3**(1): 69-79.

Otua, O., Osuji, G. A., Opara, J., Ifeyinwa, M. H., and Iheagwara, A. I., (2014). Application of SARIMA models in modelling and forecasting Nigeria's inflation rates. *American Journal of Applied Mathematics and Statistics*, **2**(7): 16-28.

Poll, T., Opal, S. M., (2009). Pathogenesis, treatment, and prevention of pneumococcal pneumonia. *The Lancet*, **374**:1543-1556.

Permanasari, A. E., Rambli, D. R. A., and Dominic, P. D. D., (2009)**.** Prediction of Zoonosis Incidence in Human using Seasonal AutoRegressive Integrated Moving Average (SARIM A). *International Journal of Computer Science and Information Security*, **5**(1): 103-110.

Pufnik, A., and Kunovac, D., (2006). Short-Term Forecasting of Inflation in Croatia with Seasonal ARIMA Processes. Croatia National Bank, working paper***, W-16*.**

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.

Rudan, I., Boschi, C., Biloglav, Z., Mulholland, K., Campbell H., (2008). Epidemiology and etiology of childhood pneumonia. *Bull World Health Organ,* **86:** 408–416.

Rudan, I., Campbell, H., (2009). The deadly toll of severe pneumonia and Haemophillus influenzae type b. *The Lancet,* **374**:854-856.

Schulze, P. M., and Prinz, A., (2009). Forecasting container transshipment in Germany. *Journal of Applied Economics,* **41**: 2809–2815.

Saz, S., (2011). The Efficacy of SARIMA Models for Forecasting Inflation Rates in Developing Countries: The Case for Turkey. *International Research Journal of Finance and Economics,* **62**: 112-142.

Sharmin, S., and Rayhan, I., (2011). Modelling of Infectious Diseases for Providing Signal of Epidemics: A Measles Case Study in Bangladesh. Institute of Statistical Research and Training, University of Dhaka, Ramna. Bangladesh, **29**(6): 567-573.

Shan, F., (1994). Risk factors for severe pneumonia in children in Goroka Hospital, Papua New Guinea. *Lancet*, **2**: 537–541.

Simoes, E. A. F., (1999). Respiratory syncytial virus infection. *The Lancet,* **354**: 847-852.

Singh, V. Aneja, S., (2011). Pneumonia management in the developing world. *Pediatric Respiratory Reviews*, **12**: 52-59.

Schanzer, D., Allison, M., Kathleen, M., (2013) Statistical estimates of respiratory admissions attributable to seasonal and pandemic influenza for Canada. Influenza and Other Respiratory Viruses, **7**(4): 799–808.

Sumer, K. K., Goktas, O., and Hepsag, A., (2009). The application of seasonal latent variable in forecasting electricity demand as an alternative method. *Energy Policy*, **37**(4): 1317-1322.

Theodoratou, E., (2013). The effect of case management on childhood pneumonia mortality in developing countries. *International Journal of Epidemiology*, **39**(1): 155–171.

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

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.

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

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

Tsay, R. S., (1989). Testing and modelling threshold autoregressive processes. *Journal of the American Statistical Association,* **84**(405): 231-240.

UNICEF (2004). Pneumonia and Diarrhea. Accessed date: 21 Jun 2013.

Van, G. A., Palache, A. M., Forleo, F., (2003). Influenza vaccination in 2000: recommendations and vaccine use in 50 developed and rapidly developing countries. *Vaccine,* **21**:1780-1785.

Victer, L. Y., Jeremy, A., Shigeki, F., (1994). Pneumonia Due to Pseudomonas Aeruginosa Epidemiology, Clinical Diagnosis, and Source. *Chest Journal, **139**:* 909-919.

Walker, C. L., Rudan, I., Liu, L., (2013). Global burden of childhood pneumonia and diarrhoea. *Lancet,* **381**:1405–1416.

Williamson, G. D., and Hudson, G. A., (1999). Monitoring system for detecting aberrations in public health surveillance reports. *Stat Med*, **18**: 3283-3298.

Watier, L., Richardson, S., and Hubert, B., (1994). Salmonella enteritis infections in France and the United States: characterization by a deterministic model. *Journal of public Health,* **83**:1694-1700.

Watt, J. P., Wolfson, L. J., Brien, O., Henkle, E., (2009). Burden of disease caused by *Haemophilus influenzae* type b in children younger than 5 years: global estimates. *The Lancet*, **374**: 903-911.

Weber, M. W., Mulholland, E. K., Greenwood, B. M., (1998). Respiratory syncytial virus infection in tropical and developing countries. *Tropical Medicine and International Health,* **3**: 268-280.

WHO (2013). Report on Child Health. Accessed date: 01/26/15.

Wingerter, S. L., Bachur, R. G., Monteaux, M. C., and Neuman, M. I., (2012). Application of the World Health Organization criteria to predict radiologic pneumonia in a US- based pediatric emergency department. *The Pediatric Infectious Disease Journal***, 31:** 561-564.

Williams, B. G., Gouws, E., Boschi-Pinto, C., Bryce, J., Dye, C., (2002). Estimates of worldwide distribution of child deaths from acute respiratory infections. Lancet Infectious Disease, **2**: 25–32.

Wongkoon, S., Pollar, M., Jaroensutasinee, M., and Jaroensutasinee, K., (2008). Predicting DHF Incidence in Northern Thailand using Time Series Analysis Technique. *International Journal of Biological and Life Sciences*, **4**(3): 117-121.

Xiang, J., (2008). Applying ARIMA Model to the Analysis of Monthly Temperature of Stockholm. Unpublished Master's thesis in Statistics. Department of Economics and Society. Dalarna University. Sweden.

Yan, W., Xu, Y., Yang, W., and Zhou, Y., (2010). A Hybrid for Short-Term Bacillary Dysentery Prediction in Yichang City, China. *Japanese Journal of Infectious Disease*, **63**: 264-270.

Yantai, S., Minfang. Y., Oliver, Y., Jiankun, L., and Huifang, F., (2005). Wireless traffic modelling and Prediction using seasonal ARIMA models. The Institute of Electronics, Information and Communication Engineers (IEICE) Transaction on Communications, **88**(10): 3992-3999.

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.

APPENDIX

Table A1: Forecast values for SARIMA (1, 1, 1)(0, 0, 1)$_{12}$

Year	Month	Forecast	LCL	UCL
2015	November	14	13.89	14.39
2015	December	15	14.85	15.49
2015	January	16	15.8	16.52
2016	February	17	16.83	17.6
2016	March	20	17.8	20.59
2016	April	21	20.76	21.58
2016	May	22	21.75	22.58
2016	June	24	22.73	24.57
2016	July	25	24.68	25.53
2016	August	26	25.67	26.53
2016	September	27	26.67	27.53
2016	October	28	27.69	28.56
2016	November	29	28.68	29.57
2016	December	30	29.67	30.58
2016	January	31	30.67	31.58
2017	February	33	31.66	33.59
2017	March	34	33.65	34.59
2017	April	35	34.66	35.59
2017	May	36	35.65	36.6
2017	June	37	36.65	37.61
2017	July	38	37.64	38.61
2017	August	39	38.64	39.6
2017	September	40	39.64	40.61
2017	October	42	40.64	42.61
2017	November	43	42.64	43.62
2017	December	44	43.63	44.62

Table A2: Forecast values for SETAR (2; 4, 3)

Year	Month	Forecast
2015	November	3.08
2015	December	3.14
2016	January	3.16
2016	February	3.18
2016	March	3.19
2016	April	3.2
2016	May	3.21
2016	June	3.22
2016	July	3.22
2016	August	3.22
2016	September	3.22
2016	October	3.23
2016	November	3.23
2016	December	3.23
2017	January	3.24
2017	February	3.24
2017	March	3.24
2017	April	3.24
2017	May	3.24
2017	June	3.24
2017	July	3.24
2017	August	3.24
2017	September	3.23
2017	October	3.24
2017	November	3.24
2017	December	3.24

Table A3: Data on Monthly Pneumonia Cases

Year	Month	Pnc	Year	Month	Pnc
2000	January	11	2003	January	12
2000	February	11	2003	February	10
2000	March	13	2003	March	9
2000	April	10	2003	April	12
2000	May	10	2003	May	14
2000	June	12	2003	June	11
2000	July	10	2003	July	15
2000	August	14	2003	August	14
2000	September	16	2003	September	17
2000	October	14	2003	October	14
2000	November	10	2003	November	16
2000	December	11	2003	December	10
2001	January	14	2004	January	12
2001	February	16	2004	February	14
2001	March	14	2004	March	20
2001	April	10	2004	April	20
2001	May	12	2004	May	22
2001	June	10	2004	June	18
2001	July	13	2004	July	17
2001	August	14	2004	August	18
2001	September	16	2004	September	16

2001	October	10	2004	October	17
2001	November	10	2004	November	18
2001	December	12	2004	December	20
2002	January	16	2005	January	20
2002	February	14	2005	February	22
2002	March	17	2005	March	21
2002	April	16	2005	April	19
2002	May	15	2005	May	18
2002	June	13	2005	June	16
2002	July	19	2005	July	21
2002	August	18	2005	August	24
2002	September	20	2005	September	35
2002	October	22	2005	October	19
2002	November	25	2005	November	28
2002	December	17	2005	December	25
2006	January	18	2009	January	28
2006	February	16	2009	February	20
2006	March	14	2009	March	25
2006	April	15	2009	April	28
2006	May	17	2009	May	30
2006	June	20	2009	June	29
2006	July	26	2009	July	24
2006	August	25	2009	August	24

2006	September	25	2009	September	32
2006	October	24	2009	October	28
2006	November	20	2009	November	27
2006	December	22	2009	December	24
2007	January	20	2010	January	20
2007	February	26	2010	February	22
2007	March	24	2010	March	26
2007	April	26	2010	April	29
2007	May	20	2010	May	25
2007	June	24	2010	June	18
2007	July	28	2010	July	24
2007	August	25	2010	August	31
2007	September	26	2010	September	30
2007	October	24	2010	October	22
2007	November	20	2010	November	28
2007	December	30	2010	December	32
2008	January	30	2011	January	8
2008	February	18	2011	February	5
2008	March	22	2011	March	13
2008	April	20	2011	April	9
2008	May	22	2011	May	9
2008	June	21	2011	June	9
2008	July	20	2011	July	7

2008	August	23	2011	August	19
2008	September	21	2011	September	0
2008	October	18	2011	October	0
2008	November	20	2011	November	0
2008	December	22	2011	December	0
2012	January	18	2014	January	0
2012	February	20	2014	February	0
2012	March	14	2014	March	12
2012	April	15	2014	April	24
2012	May	20	2014	May	13
2012	June	18	2014	June	11
2012	July	16	2014	July	15
2012	August	14	2014	August	25
2012	September	13	2014	September	16
2012	October	12	2014	October	32
2012	November	16	2014	November	22
2012	December	18	2014	December	16
2013	January	16	2015	January	32
2013	February	20	2015	February	22
2013	March	13	2015	March	16
2013	April	14	2015	April	14
2013	May	20	2015	May	12
2013	June	8	2015	June	8

2013	July	12	2015	July	9
2013	August	24	2015	August	9
2013	September	13	2015	September	11
2013	October	11			
2013	November	15			
2013	December	25			

Source: Tamale Teaching Hospital, 2015