UNIVERSITY OF CAPE COAST

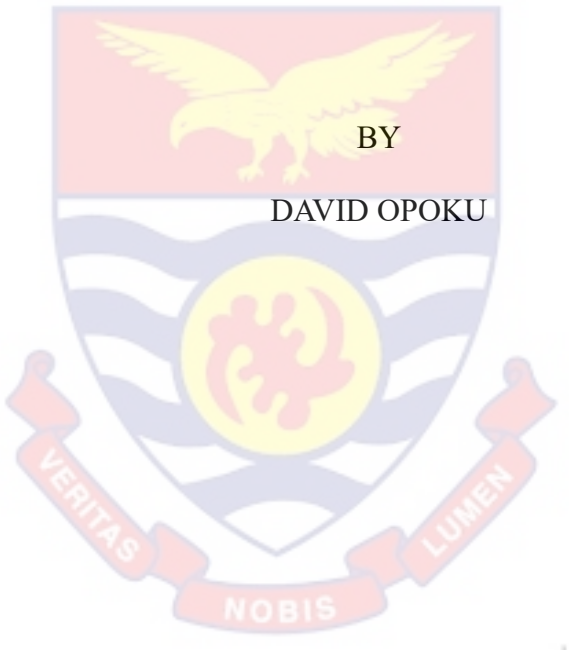
TIME SERIES MODELLING OF TUBERCULOSIS DISEASE IN GHANA: THE BOX-JENKINS APPROACH

DAVID OPOKU

2024

UNIVERSITY OF CAPE COAST

TIME SERIES MODELLING OF TUBERCULOSIS DISEASE IN GHANA: THE BOX-JENKINS APPROACH



BY

DAVID OPOKU

Thesis submitted to the Department of Statistics of the School of Physical Sciences, College of Agriculture and Natural Sciences, University of Cape Coast, in partial fulfilment of the requirements for the award of Master of Philosophy degree in Statistics

DECEMBER, 2024

DECLARATION

Candidate's Declaration

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

Candidate's Signature ……. Date …….

Name: David Opoku

Supervisor's Declaration

I hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Supervisor's Signature …… Date ……..

Name: Dr. Francis Eyiah-Bediako

ABSTRACT

Tuberculosis (TB) disease, in spite being preventable and treatable, is still ranked among the world's major causes of infectious disease-related mortality, killing almost 1.5 million people in 2020 alone. In order to put control measures in place, numerous research have been carried out globally to determine the disease's trend and seasonal variations. “Box-Jenkins method” was employed to analyze monthly TB cases recorded in Ghanaian health facilities between January 2014 and June 2024. The main objective was to find a suitable and economical model to forecast the incidence of the disease. Ten candidate models were identified by the study. The overall best one was chosen using parameter estimation techniques, such as AIC and BIC values. The findings showed seasonal fluctuation and an upward trend in the 126-point TB data. The data was made stationary after first difference and first seasonal difference. SARIMA $(1,1,3)\ (3,1,3)_{12}$ was identified as the most effective model for predicting Tuberculosis disease in Ghana following a number of diagnostic tests. The model gave a prediction of monthly TB cases from July 2024 to June 2026. The study concluded that the disease in Ghana showed an upward trend with some seasonal variation but no signs of random fluctuations. The model equation for SARIMA $(1,1,3)\ (3,1,3)_{12}$, which was suitable for TB data in Ghana, was;

$$X_t = -0.307X_{t-12} + 0.795X_{t-24} - 0.250X_{t-36} - 2.413W_{t-1} + 1.842W_{t-2}$$

$$-0.427W_{t-3} + 0.305W_{t-12} - 0.794W_{t-24} + 0.250W_{t-36} + \text{Ø}_t$$

KEY WORDS

Autoregressive

Diagnostics

Forecasting

Modelling

Moving Average

ACKNOWLEDGEMENTS

DEDICATION

To my wife, Diana Opoku Frimpong; my children, Louisa, Precious and Desmond; and in memory of my late Mummy, Hannah Afrakoma Opoku.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
AICc	Akaike Information Criterion Second-order
AIDS	Acquired Immune Deficiency Syndrome
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BIC	Bayesian Information Criterion
COVID	Corona Virus Disease
DHIMS	District Health Information Management System
HIV	Human Immunodeficiency Virus
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SDGs	Sustainable Development Goals
TB	Tuberculosis

UHC	Universal Health Coverage
UN	United Nations
WHO	World Health Organization

CHAPTER ONE

INTRODUCTION

Overview

The chapter outlined the introductory aspect of the study. It provided a thorough summary of the background of Tuberculosis disease. There was a thorough discussion of the problem statement, study objectives, purpose, significance, delimitations and limitations, definition of terms, and study organization.

Background to the Study

One disease that is known to be preventive and curable is Tuberculosis (TB). This bacterial disease primarily affects the lungs (CDC, 2015). A bacterium known as Mycobacterium Tuberculosis is the primary cause of the disease. People who have the disease can spread the bacterium through their coughs, sneezes, and spitting (Dodd, 2016). It should be mentioned that TB might be lethal if treatment is not received for an extended length of time (WHO, 2021). Drugs are the primary treatment for the illness.

According to World Health Organization (2021), “Tuberculosis caused the death of around 1.5 million people in the year 2020, making it one of the main causes of infectious disease-related mortality worldwide, and that every year, more than 10 million people still get Tuberculosis.” According to Auld, Kasmar, Dowdy, Mathema, Gandhi, Churchyard, and Rustomjee (2017), “every member state of the World Health Organization (WHO) and the United Nations (UN) has agreed that swift action is required to halt the global Tuberculosis epidemic by 2030,". Dodd (2016), observed that approximately 25% of people worldwide are thought to have Tuberculosis (TB).

The first two years after infection are when the risk of developing Tuberculosis (TB) is highest (WHO, 2022). It reaches a height of roughly 5% before dropping off quickly. To ensure that those who require treatment for Tuberculosis (TB) or another sickness receive it, Universal Health Coverage (UHC) is necessary (WHO, 2022). A multisectoral approach to treating HIV infection, poverty, malnutrition, smoking, Diabetes, and Tuberculosis is necessary to lower the number of individuals who get the disease, get sick, and pass away from it (WHO, 2015). Nowadays, Tuberculosis kills one person and generates less than 10 infections per 100,000 people annually in many nations (WHO, 2022). Floyd, Glaziou, and Zumla (2018) contend that new vaccines and other technological developments are required to rapidly bring the yearly worldwide case count down to the levels already attained in these low-burden nations. Financial and human resources are frequently limited in TB countries with high caseloads that use broad-based strategies that address the illness as a single epidemic (Auld *et al*., 2017). Despite this, contemporary molecular techniques have demonstrated that transmission of the disease can occur through several channels, which require different interventions to deal with it (McBryde *et al*., 2021).

In countries where there are records of high number of cases, prevention strategies that focus on transmission may be very helpful in reducing Tuberculosis (WHO, 2021). Blankson (2012) reported that "more than 20% of total family income of households are spent on direct medical expenses, non-medical charges, and indirect costs like lost income". According to WHO (2020), certain communities are more susceptible to contracting tuberculosis (TB) or becoming infected and developing the disease, even though everyone can get it. HIV-positive individuals, healthcare workers, youngsters, and residents of communal settings

such as jails, prisons, refugee camps, and assisted living facilities are a few examples of these susceptible groups (WHO operational handbook on Tuberculosis, 2020). Because tuberculosis (TB) is expensive and difficult to treat in both the community and individuals, especially when it comes to drug-resistant strains, prevention is crucial. Stopping the spread of Mycobacterium Tuberculosis in healthcare facilities, social settings, workplaces, and TB patients' residences is crucial, according to the WHO consolidated guidelines on Tuberculosis (2020).

Even though Ghana is no longer recording high Tuberculosis cases, the country still accounts for almost 1% of all TB cases that go undiagnosed globally (WHO Global TB Report, 2021). According to Bonsu, Afutu, Hanson-Nortey, and Ahiabu (2017), "the actual number of tuberculosis cases recorded in the year 2013 was 290 cases per 100,000 individuals, which was four times higher than the previously predicted rate, and that majority of these cases were disregarded". As stated in the WHO Global TB Report (2021), Ghana recorded 44,000 tuberculosis cases in 2020. Statistically, this figure amounts to a yearly rate of 143 cases per 100,000 persons (WHO Global TB Report, 2021). However, according to Bonsu, Hanson-Nortey, Afutu, Kulevome, Dzata, and Chimzizi (2020), less than 70% of the cases were reported, with only 12,700 instances being reported. Ghana missed about 65% of TB cases in 2019 (Bonsu *et al*., 2020).

Adèr, Mellenbergh (2008) reaffirms that statistical models can be used to estimate, assess, and compare control and prevention measures as well as the interaction of biological, environmental, and social factors that affect disease transmission. Results from a number of research indicate that modeling-based TB control has advanced (WHO, 2013). The World Health Assembly established new post-2015 global targets for Tuberculosis control. The strategy was that member

states should ensure a reduction in TB cases and deaths between 2015 and 2025 (WHO, 2013). With the global targets for 2030, the current post-2015 TB targets demands that member states ensure that number of Tuberculosis cases goes down by 80% and deaths reduced by 90% over the 2015 levels (WHO, 2015). China has claimed that by 2025, the latter goals seem even more reachable (Xiaolin, Xiulei, Jia, John, Rachel, Guanyang, Hongmei, Fang, & Zhimin, 2014). China chose relevant projects that could be carried out over the next ten years, taking into consideration the previously provided data and using model-based analysis to help define acceptable targets for Tuberculosis prevention (Xiaolin, Xiulei, Jia, John, Rachel, Guanyang, Hongmei, Fang, & Zhimin, 2014). According to a Chinese study by Glaziou *et al*. (2015), modeling can be used to account for data uncertainty and synthesize relevant information from several sources.

A group of policy experts, field specialists, and modelers can also help put modeling conclusions into practice and policy. Abu-Raddad, Sabatelli, Achterberg, Sugimoto, Longini & Dye (2009) examined potential advantages of using mathematical model to control Tuberculosis in the Asian Regions. The study concluded that "the number of Tuberculosis cases can be reduced by 94% with a two-month pharmacological treatment plan in conjunction with preventative latent therapy" (Abu-Raddad *et al*., 2009).

Statement of the Problem

According to WHO (2022), Tuberculosis affects a lot of people, and it leads to death of several people around the world. Despite improvements in recent decades, there are still significant systemic gaps in the struggle to completely eliminate Tuberculosis (TB), particularly in areas where resources to control it is limited (WHO, 2021). According to WHO (2021), "efforts to prevent and cure

Tuberculosis will have prevented about 74 million deaths from the disease globally between 2000 and 2021". However, in 2021 alone, more people developed the disease, out of which several died (Global tuberculosis report, 2022).

There have been attempts to end worldwide Tuberculosis disease by United Nation member countries between 2015 and 2030, as contained in the Sustainable Development Goals (SDG report, 2015). In comparison with the 2015 baseline year, the World Health Assembly in 2014 proposed a strategy called "WHO End TB strategy" where member countries were to ensure that the number of Tuberculosis cases are reduced by 80% and Tuberculosis deaths reduced by 90% by the year 2030 (WHO, 2015). "WHO End TB Strategy" was not significantly met, despite a 19% global decrease in TB-related fatalities between 2015 and 2022" (WHO, 2023).

According to WHO (2023), Ghana's TB death rate (all forms of TB, excluding HIV coinfection) was 36/100,000 in 2021 from 37/100,000 in 2015, a 3.0% decrease in TB-related mortality between 2015 and 2022. These numbers made it abundantly evident that Ghana would have difficulty meeting its goal of lowering the country's TB death rate by 90% by 2030 in order to comply with the "WHO End TB" plan. According to Dodd (2016), mathematical modeling is one of the appropriate ways of managing the Tuberculosis disease. According to modeling, in order to expedite the eradication of Tuberculosis, new medications, diagnostic tests, and vaccinations will need to be developed (Abu-Raddad, 2009).

Numerous studies have been conducted to find out how the disease behave during a specific time period and whether the number of cases could be predicted into the future. In a study conducted in another region of Africa, Ade, Békou, Adjobimey, Adjibode, and Ade (2016) employed the Box-Jenkins technique and

concluded that Tuberculosis cases were identified and reported differently in Benin, depending on the season, with the highest cases reported in the first quarter of 2011. There have been studies to attempt to determine the behaviour of the disease in certain regions and facilities in Ghana. Aryee, *et al*. (2018) in a bid to know TB patients who patronized "Korle Bu Teaching Hospital" in Ghana, concluded on ARIMA (1, 0, 1) to forecast the disease. In their work, Gyasi-Agyei *et al*. (2014) also discovered ARMA (1, 0) to be the best model for the TB data in the Ashanti Region of Ghana. In a similar vein, Adetunde (2009) examined a dynamical behaviour of TB disease in the "Upper East Region" of Ghana and came to conclusion that the danger of instability of the model's disease-free equilibrium increased with population density.

There is one important observation from these investigations. Only a facility or a region of Ghana was taken into account in these investigations and that no study has predicted the prevalence of Tuberculosis disease across the entire territory of Ghana. Additionally, most of these investigations did not also consider the seasonal fluctuations of the disease. Therefore, this study was carried out to create suitable model for forecasting TB disease and to examine its seasonality throughout Ghana. With the 2030 timeline in mind, forecasting the disease's pattern, was going to impact policies and programmes put into place for the reduction target.

Purpose of the Study

The main goal of this study was to use TB data from January 2014 to June 2024 to create a suitable model for the disease in Ghana. The result was be used to forecast the anticipated number of cases of the disease over a two-year span. Time series modeling is one of the practical approaches to disease control, according to previous studies. Thus far, time series models have been very useful in determining

which individuals should be prioritized for isolation or vaccination in order to lower the number of infections.

Member countries of the United Nations were to ensure that the number of Tuberculosis cases are reduced by 80% and Tuberculosis deaths reduced by 90% by the year 2030 as contained in the "WHO End TB strategy". As a signatory to the United Nations General Assembly (UNGA), Ghana must be involved in accomplishing this goal. The prediction of the disease's progress will impact the policy orientation of Ghana's TB control initiatives. This study offers the chance to develop a model that would offer a more thorough comprehension of Tuberculosis control. Last but not the least, this study was to serve as a baseline for related research initiatives in other regions of Africa and beyond.

Research Objectives

General objective

To forecast the number of TB cases in Ghana using Box-Jenkins modeling approach.

Specific objectives

1. To use Tuberculosis data from January 2014 to June 2024 to create an appropriate model for the disease in Ghana.
2. To predict the number of Tuberculosis disease in Ghana over the period of two years.

Significance of the Study

In 2020, TB killed 1.5 million people, making it one of the main causes of infectious disease-related mortality (WHO, 2021). Knowing the number of cases of the disease over time is required to help obtain a better understanding of the

behavior of Tuberculosis transmission. Additionally, it will affect the direction of policy toward preventive and control measures.

Forecasting the trajectory of the disease will give decision-makers and medical professionals baseline data to evaluate current disease-fighting strategies.

Delimitations

This study focused on using the Box-Jenkins' ARIMA technique to develop a model, appropriate enough to forecast Tuberculosis disease in Ghana. The idea was to obtain a suitable model, among other several candidate models, that was good enough to forecast the disease.

Limitations

The parameter estimation stage of the "Box-Jenkins" method constituted a limitation. The need to select a model from the diverse "possible models" may lead to an error, which will eventually affect the forecast. There may be an attempt to add more variables, which may overfit the data.

Definition of Terms

Autoregressive model:

Future value of the dependent variable of a time series is a function of its past value.

Moving Average model:

Future value of the dependent variable of a time series is a function of its past errors.

Sustainable Development Goals:

A call to eradicate poverty and injustice, safeguard the environment, and guarantee that everyone has access to justice, wealth, and good health.

Parsimony:

As long as the models fit the data comparably well, the parsimony principle in statistics favours simpler models with fewer parameters over more sophisticated models with more parameters. A statistical model that fits the data similarly well should have fewer parameters than a more complex model with more parameters, according to the parsimony principle.

Seasonal variation:

If a time series shows regular, predictable, and consistent annual fluctuations, it is considered seasonal. Any predictable trend or variation that occurs repeatedly over the course of a year is referred to as seasonal. A time series is deemed seasonal if it exhibits regular, predictable, and consistent annual changes. Seasonality is any recurring, predictable trend or fluctuation that takes place over the course of a year.

Organization of the study

This thesis was broken up into five sections. The first chapter provided an introduction to the whole thesis. Presented were the study's background, problem statement, purpose, research objectives, significance, delimitation, limitations, organizational structure, and definitions of all relevant terminology. An overview of relevant work on TB time series modelling was presented in Chapter Two. Literature on the pattern of Tuberculosis disease and time series modelling of the disease was reviewed. In Chapter Three, the research method was introduced. The ARIMA model, ACF and PACF plots, the stationary time series, the Augmented

Dickey-Fuller test, model identification, estimation, and diagnosis, and forecasting were all covered. Chapter Four of the thesis presented and discussed the data analysis and findings from the analysis. An overview of major findings, conclusions and recommendations was given in Chapter Five

CHAPTER TWO

LITERATURE REVIEW

Introduction

This chapter summarized pertinent research that has been done in the field in Ghana and other countries, giving a broad overview of the study of modeling tuberculosis disease in Ghana. The subthemes include the pattern of tuberculosis disease, and time series modeling and forecasting of the disease.

TB is one of the most common infectious disease-related causes of death, accounting for about 1 in 5 million deaths globally (WHO, 2021). Investigating its pathways of transmission, both globally and specifically in Ghana, is therefore crucial. A more comprehensive examination of the models employed by other researchers and how they could be adjusted to assist the eradication is necessary to successfully achieve this. Information for this literature study was taken from peer-reviewed journal articles, books, policy documents, yearly reports, and student theses.

Tuberculosis Disease Pattern

A specific kind of bacteria called "Mycobacterium Tuberculosis" causes tuberculosis (CDC, 2015. The bacteria spreads through the air when infected persons cough, sneeze, or spit. Nearly 25% of people globally are thought to be infected with the TB bacteria (Dodd, 2016). Eventually, five to ten percent of TB patients will get sick and show symptoms. Infected individuals who are not yet ill are unable to spread the infection to others. TB is typically treated with medications, mainly antibiotics, even though it can be lethal if untreated (CDC, 2012).

An observation was made by Luquero, Sanchez-Padilla, Simon-Soria, and Eiros (2008) in a study to ascertain the prevalence of Tuberculosis in Spain. They

came to the conclusion that time-series studies should be the first step in creating a predictive model.

According to Cruz-Ferro (2007), the incidence of Tuberculosis (TB) in Galicia dropped by an average of 7.0% year, from 72.3 cases per 100,000 population in 1996 to 37.7 cases per 100,000 population in 2005. Over a ten-year span, from 1996 to 2005, this study detailed the epidemiological evolution and features of Tuberculosis.

Lin (2014) studied the behaviour of "Multidrug-resistant (MDR) TB" in Taiwan and discovered that there is a 1% chance that more than 50% of the population has "MDR TB". In addition to offering valuable insights into the interplay between seasonal TB dynamics and environmental factors, the model can forecast seasonal number of TB cases linked to the probability of "MDR TB" infection.

The "Directly Observed Therapy Short course" (DOTS) approach decreased Albania's TB incidence from 17 per 100,000 people in 2001 to 12 per 100,000 people in 2008, claimed (Hafizi, Tafaj, Bardhi, and Dilko, 2009). Furthermore, from 42% in 2001 to 75% in 2007, the estimated case detection for smear-positive cases increased.

According to a study by Douglas, Strachan, and Maxwell (1996) to find the seasonal variation of TB in the UK, the pattern of tuberculosis notifications showed a summer peak with an amplitude of 10%, which is significantly different from other respiratory disorders that show a winter peak and a summer trough. According to the study's findings, there is no explanation for the unusual seasonality of tuberculosis.

In general, the time series on PTB diagnosis exhibited seasonality and a declining tendency, peaking in March and falling in December, according to a Portuguese study by (Bras, Gomes, Filipe, Sousa, and Nunes, 2014). Males and adults between the ages of 25 and 54, as well as high-incidence locations, consistently showed a larger mean seasonal amplitude. Trend and seasonal persistence were found to be predicted by SARIMA models through precise fitting and forecasting of the time series.

According to a study by Willis, Winston, Heilig, Cain & Walter (2012), "TB is a seasonal disease that peaks in the spring and troughs in late fall in the United States". According to the study, latitude-dependent factors might not have a major effect on seasonality in the US. These include less sun exposure throughout the winter and its possible impact on vitamin D levels' seasonality.

According to Manabe (2019), "the number of tuberculosis cases in Japan among adults over 15 exhibits seasonality and rises from June to September, and those monthly cases varied according to age and sex.

Padberg, Bätzing-Feigenbaum (2015) discovered that the association between extra-pulmonary tuberculosis and age, sex, and seasons differs depending on the diseased organ. Additionally, they discovered significant differences in the way TB organs present according to age, sex, and season, which could point to different pathophysiological pathways of disease progression.

Borgdorff, Sebek, Geskus, Kremer, and Kalisvaart (2011) investigated the distribution of TB incubation times using a molecular epidemiological approach. According to the study, people who got the condition within 15 years had an 83% probability of being sick within five years, a 62% chance within two years, and a 45% Kaplan-Meier likelihood within a year. Molecular epidemiological study has

made it possible to describe the TB incubation period more accurately and to discover risk variables for shorter incubation periods than was previously feasible.

A study by Blankson (2012) estimated the impact of tuberculosis (TB) on household wellbeing and the economic burden in Ghana's Western Region. The costs were estimated using the human capital method. Multiple regression analysis and the Wells-Riley model were used to find out the effect of tuberculosis on household welfare and the chance of transmission within households. The findings demonstrated that TB significantly reduces welfare and household income. The research claims that TB uses a significant amount of the public health system's resources and has a significant financial impact on impacted households. To help TB-affected households manage the heavy financial burden and to ensure that patients receive all the support and treatment they need, safety nets or income insurance should be established.

Time Series Modeling and Forecasting of Tuberculosis Disease

Mathematical link between random and non-random variables is a common definition of statistical modeling (Brian Z., 2024). For the creation of forecasts, data-driven decisions, and scientific discoveries, statistical modeling is essential (Brian Z., 2024). It is possible to distinguish between acceptable and unacceptable judgments based on quantitative evidence using statistical modeling (Brian Z., 2024).

The number of TB cases reported at “Korle Bu Teaching Hospital” in Ghana was estimated by (Aryee et al., 2018). According to the study, “there was no obvious increasing or decreasing trend in the TB data and that stationarity was attained by the log-transformed data, which showed rather constant movements

around the series mean". For the TB data, the most effective model was ARMA (1,1).

Cao et al. (2013) conducted a study in China to create a suitable model for tuberculosis outbreak prediction and possibility of seasonal variation. The study compared a hybrid model that incorporated a "generalized regression neural network model" and a "seasonal autoregressive integrated moving average (SARIMA) model" and concluded that the hybrid model outperformed the SARIMA model in predicting the number of cases of tuberculosis in China.

Adetunde (2009) used "mathematical model" to study the dynamical behavior of tuberculosis disease in the Upper East region of Ghana. The model system's equilibrium points were identified as endemic equilibrium and disease-free equilibrium. According to studies employing computer simulation and stability theory, the population regulates the rate of tuberculosis infection. Consequently, as population density rises, so does the risk of disease-free equilibrium instability (Adetunde, 2009).

According to a study conducted by Gyasi-Agyei et al. (2014), "The best models for tuberculosis occurrence in the Ashanti region of Ghana were found to be ARMA (1, 0) or AR (1), which are stochastic time series linear models". Additionally, the study forecasted that between April 2013 and April 2015, the TB epidemic in the Ashanti Region would not change (Gyasi-Agyei et al. 2014).

Chowdhury et al., 2013 conducted a study to come up with a model to predict the number of tuberculosis cases in "Rural West Bengal". The study concluded that SARIMA (1,2,0) was the best predictor of the univariate model that could forecast the number of tuberculosis in the country.

A study that employed time series analysis to forecast the number of smear-positive TB cases in Iran found that SARIMA $(0, 1, 1)\ (0, 1, 1)_{12}$ was the best model for predicting the disease (Mahmood K., & Nasehi, 2015).

In Delhi, India, Kumar, Singh, Adhikary, Daral, Khokhar, & Singh, (2014) aimed to determine the recurrent fluctuations in tuberculosis. The study concluded that "Winter's multiplicative model was the most effective at forecasting the time series data with 69.8% variability, according to a univariate model created by a skilled SPSS modeler". The forecast's drop was a sign of seasonality.

A study was conducted to learn about the seasonal occurrence of Tuberculosis in Jiangsu Province, China. The study observed "BPNN model" to be performing better than the "ARIMA model" at forecasting the seasonal incidence and pattern of pulmonary Tuberculosis in China (Liu et al., 2019).

Olanrewaju et al. (2020) looked at Tuberculosis disease in "Minna Niger State, Nigeria" and observed that "ARIMA (2, 1, 3)" was the most appropriate model to predict the number of Tuberculosis cases in the country, and that the number of recorded cases of Tuberculosis decreased by 7% every year over the time in question due to government, medical experts, and individual action.

The seasonal occurrence and change in cases per time prediction of Tuberculosis frequency data were examined in the Eastern Cape of South Africa using a blended model. This study, which was conducted by Azeez, Obaromi, Odeyemi, & Ndege (2016), observed that although both models could potentially predict the occurrence of tuberculosis, the study concluded that their combined performance was better. The "SARIMA-NNAR model" indicated that the seasonal occurrence of TB cases will be somewhat greater than the individual (Azeez *et al*., 2016).

Chapter Summary

This chapter of the study outlined theoretical and empirical evidence of the behaviour of Tuberculosis disease. The chapter also examined other writers' studies on the seasonal variations in tuberculosis disease. The bulk of the examined studies used multiple regression techniques, including the Wells-Riley models, ARIMA, ARMA, SARIMA, SARIMA-NNAR, and SARIMA, to forecast the trend of tuberculosis disease over time. The literature review states that studies on modeling tuberculosis disease have been carried out in Ghana at the regional and facility levels. However, no time series modeling of tuberculosis disease have been carried out in Ghana as a whole. To help with Ghana's TB situation, this study forecasted and modeled the country's TB disease using the Box-Jenkins approach.

CHAPTER THREE

RESEARCH METHODS

Introduction

A statistical model is a mathematical depiction of a collection of statistical assumptions related to the generation of sample data and, to a lesser degree, analogous data from a broader population (Box & Jenkins, 1970). Probabilities are particularly referred to as "probabilistic models" in this context. All statistical estimators and hypothetical tests are produced by statistical models. Lastly, statistical models serve as the basis for statistical inference.

Statistical models differ from other types of mathematical models in that they are not deterministic. As a result, some variables in a statistical model that is formally defined may be unpredictable and have probability distributions rather than fixed values. Selecting a statistical model that effectively captures a particular data-generating process can occasionally be challenging, necessitating knowledge of the process as well as pertinent statistical calculations (Box et al., 1970).

Examining the TB disease in Ghana and forecasting its transmission trend over a 24-month period were the goals of this study.

Data for the study

The study used 126 data points, made up of all monthly TB cases reported by Ghanaian healthcare facilities between January 2014 and June 2024. This data was collected as secondary data.

Time series modeling

This is an ordered set of data points collected typically to forecast the future, in which time is frequently the independent variable. It contains techniques for understanding modeling and forecasting (Hipel & McLeod, 1994).

Forecasting collects and evaluates historical data and usually, the outcomes of predictions are utilized to inform crucial strategic choices and preventive measures (Zhang, 2007a). The appropriate model must be matched with a time series. Over the past few decades, researchers have put a lot of effort into developing and improving suitable time series forecasting models (Zhang, 2003).

Four elements typically impact a time series: seasonal, cyclical, irregular, and trend (Hipel & McLeod, 1994). The term "trend" describes a time series' overall propensity to rise, fall, or remain constant over an extended period of time. Seasonal variation, which is typically brought on by weather and temperature as well as customs and traditional behaviors, is the element that accounts for variations within a year or within the season. The cyclical component describes the medium-term alterations brought forth by situations that repeat in cycles. Unexpected factors that are irregular and do not repeat in a certain cycle are the source of erratic or arbitrary fluctuations (Hipel & McLeod, 1994).

The factors include floods, earthquakes and wars. Several statistical methods are used to quantify random fluctuations in a time series. They include method of variance and standard deviation, moving average model, autocorrelation function (ACF), spectral analysis, decomposition method, statistical test, among other methods.

Two distinct model types are frequently employed, depending on the effects of the four elements mentioned.

Additive Model

$$Y_t = T(t) + S(t) + C(t) + I(t) \quad (1)$$

According to the additive model, the four elements exist independently of one another.

Multiplicative Model

$$Y_t = T(t) \times S(t) \times C(t) \times I(t) \quad (2)$$

The multiplicative model assumes that the four constituent parts can influence one another and do not exist independently.

Examples of time series

White noise

A simple time series could be made up of a number of uncorrelated random variables, $\{w_t\}$, with zero mean $\mu = 0$ and finite variance σ_w^2, and it is expressed as; $W_t \sim wn(0, \sigma_w^2)$.

Gaussian White Noise

One particular useful kind of white noise is called Gaussian white noise, which is expressed as follows: W_t are self-existing normal random variables with zero mean, $\mu = 0$ and finite variance, σ_w^2. $W_t \sim iid(0, \sigma_w^2)$. Because they would enable the application of traditional statistical techniques to characterize the contentious behaviors of any time series in terms of the white noise model, white noise time series are extremely significant. "White noise" is the term used to describe the assumption that every item in a series is selected at random from a population with constant variance and zero mean. This rarely happens in practice.

Stationary time series

When “mean”, “variance”, and “autocorrelation” all show constant trends across time the data is said to be stationary. Forecasting can be difficult at times because time series are non-deterministic, meaning that future events cannot be confidently foreseen. It's interesting to note, however, that the time series exhibit a rather consistent pattern over a long period of time rather than an exact i.i.d. For

instance, based on what transpired today, it is plausible to predict that there is a high possibility of rainfall in that particular city tomorrow (Cochrane, 1997).

There are two kinds of stationarity that have been widely employed. They consist of weakly stationary and strong stationary. The time series $\{X_t, t \in Z\}$ is said to be weakly stationary if these three conditions are met;

$$E[X_t^2] < \infty, \forall_t \in Z;$$

$$E[X_t] = \mu, \forall_t \in Z;$$

$$\gamma x\{s,t\} = \gamma x\{s+h, t+h\}, \forall_{s,t,h} \in Z$$

A weakly stationary time series $\{X_t\}$, must have three essential properties: finite variance, a constant first moment, and the second moment $\gamma \times \{s,t\}$ depends only on $\{t-s\}$ and not on s or t. One crucial difference between strong and weak stationarity is that they do not imply one another. However, a process that is weakly stationary but has a normal distribution is also highly stationary, claims (Cochrane, 1997). Mathematical tests such as Dickey and Fuller's are commonly used to identify stationarity in time series data (Cochrane, 1997).

According to Box (1970) & Hipel (1994), "mathematical concept of stationarity was created to facilitate the theoretical and practical development of stochastic systems". To develop a model that is good for prediction, the underlying data must be stationary. However, that's not always the case. Studies have shown that when historical observations are more, the data would exhibit non-stationary characteristics (Hipel, 1994). Nonetheless, it makes sense to mimic the time series for a comparatively brief period of time using a stationary stochastic process. Time series that show sporadic seasonal trends or cycles are usually not stationary. In these cases, differencing and power transformations are commonly used to remove the trend and render the series stationary (Box & Jenkins, 1970).

When designing a model, parsimony principle should be considered, which states that the researcher should always select the model with the fewest number of parameters to guarantee that these parameters will adequately represent the underlying time series data (Chatfield, 1996). One aspect of this technique is the selection of the most straight forward explanation when multiple adequate and competing explanations are available (Hipel, 1994). It is often known that the more complicated the model, the more options there are to depart from the actual model assumptions. The likelihood of overfitting also increases with an increase in model parameters. An overfitted time series model may adequately describe the observed data, but it may not be suitable for future forecasting. Parsimony is often used as a guiding idea to overcome the issue of potential overfitting, which reduces a model's ability to produce accurate predictions. In summary, while developing time series forecasts, real attention should be given to selecting the best cost-effective model from all available possibilities (Chatfield, 1996).

Time series differencing

Non-stationary time series forecasting is recognized to be challenging. However, stationarity can be achieved after making mathematical corrections. One transformation method for turning non-stationary time series stationary is differencing, which looks at the differences between subsequent observations:

$$y'_t = y_t - y_{t-1} \quad (3)$$

$$y_t - y_{t-1} = \varepsilon_t$$

Re-arranging;

$$y_t = y_{t-1} + \varepsilon_t \quad (4)$$

ε_t is called “white noise” and equation (4) is called “Random Walk”.

“Random walks” typically exhibit the following traits:

- Prolonged periods of apparent upward or downward rise
- Abrupt and unpredictable direction changes.

The predictions of a random walk model are equal to the most recent observation since future movements are uncertain and have an equal chance of being up or down. Fast forward, "random walk" serves as the foundation for naïve forecasts.

The dissimilarities can have a non-zero mean in a model that is roughly related. Hence;

$$y_t = c + y_{t-1} + \varepsilon_t \quad (5)$$

y_t will tend to travel downward if "c" is not positive, and increases if "c" is positive. Sometimes the dissimilar data might not appear to be stationary, and in order to create a stationary series, a second difference may be needed:

$$y_t'' = y_t' - y_{t-1}'$$

And,

$$y_t'' = y_t - 2y_{t-1} + y_{t-2} \quad (6)$$

In practice, going beyond second-order differencing is nearly never required. If the time series shows a seasonal tendency, seasonal differencing will be required to make it stationary. "A seasonal difference" is the difference between one observation and the previous observation taken during the same period.

$$y_t' = y_t - y_{t-m} \quad (7)$$

A suitable model for the original data is as follows if seasonally different data seem to be white noise:

$$y_t = y_{t-m} + \varepsilon_t \quad (8)$$

$$y_t'' = y_t' - y_{t-1}'$$

$$y_t'' = (y_t - y_{t-m}) - (y_{t-1} - y_{t-m-1})$$

$$y_t'' = y_t - y_{t-1} - y_{t-m} + y_{t-m-1} \quad (9)$$

Applying both seasonal and first differences will get the same result regardless of which is done first. However, "seasonal differencing" first if the data has a significant seasonal tendency. This is due to the possibility that the resultant series will sometimes be stationary, which would eliminate the need for a following first difference. If first differencing is done initially, seasonality will still exist. Understanding the differences is essential when employing differencing. The way one observation differs from the next is the initial difference. Seasonal variation is the change from one year to the next. Introducing the "Backshift operator";

$$\begin{aligned} Y_t' &= y_t - y_{t-1} \\ &= y_t - By_t \\ &= (1 - B)y_t \end{aligned} \quad (10)$$

Take note that (1−B) is used as the starting difference. Because the operator may be handled using standard algebraic methods, backshift notation is especially useful when merging differences. Specifically, phrases that involve B can be multiplied collectively.

"The Kwiatkowski Phillips Schmidt Shin (KPSS)" and "Augmented Dickey Fuller (ADF)" are tests that can be used to find out if data is stationary or not. This recognized statistical method can be used to find indications of non-stationarity in the system that creates the data.

This is the basic equation:

$X_t = \emptyset_1 X_{t-1} + Y_t;\ t = 1, 2, 3, \cdots$ and $\{Y_t\}$ is a stationary process.

The following criteria determine whether the time series $\{X_t\}$ is stationary or not:

$\{X_t\}$ is non-stationary if $|\emptyset_1| = 1$,

$\{X_t\}$ is stationary if $|\emptyset_1| < 1$.

The hypothesis in the equation is examined by the ADF test;

$$\Delta X_t = \beta_0 + \beta_1 X_{t-1} + \sum_{i=1}^{p} \Delta X_{t-1} + w_t \quad (11)$$

H_0: X_t is not stationary

H_1: X_t is stationary.

Autocorrelation for stationary time series

The "autocovariance function" is defined as:

$$\gamma x\{t+h,t\} = Cov\{X_{t+h}, X_t\} \quad (12)$$

$$= Cov\{X_{h,} X_0\}$$

$$= \gamma\{h,0\}$$

$$= \gamma_h$$

In terms of Autocovariance;

$$\gamma(h) = Cov\{X_{t+h}, X_t\}$$

$$= E[(X_{t+h} - \mu)(X_t - \mu)]$$

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}}$$

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \quad (13)$$

Partial Autocorrelation for stationary time series

"The Partial Autocorrelation Function" is defined as;

For $h = 1, 2, \ldots$ is explained as;

$$\emptyset_{11} = \text{Corr}(X_{t+1}, X_t) = \rho 1$$

$$\emptyset_{hh} = \text{Corr}(X_{t+h} - \widehat{X}_{t+h,}\ X_t - \widehat{X}_t),\ h \geq 2$$

Where $\widehat{X}_{t+h}$ and $\widehat{X}_t$ are defined as;

$$\widehat{X}_{t+h} = \beta_1 X_{t+h-1} + \beta_2 X_{t+h-2} + \cdots + \beta_{h-1} X_{t+1}$$

$$X_t = \beta_1 X_{t+1} + \beta_2 X_{t+2} + \cdots + \beta_{h-1} X_{t+h-1}$$

If X_t is Gaussian, then $\emptyset_{hh}$ is really conditional correlation

$$\emptyset_{hh} = \text{Corr}(X_t, X_{t+h} | X_{t+1}, X_{t+2}, \cdots, X_{t+h-1}) \quad (14)$$

Autoregressive (AR) Models

"X_t", the current value of the series, can be described as a linear combination of p previous values, $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ with a random error in the same series. This is the foundation of autoregressive models. The following is the expression for the autoregressive model:

$$X_t = c + \sum_{i=1}^{p} \emptyset_i X_{t-i} + w_t \quad (15)$$

White noise, or W_t, is represented by $W_t \sim wn(0, \sigma_w^2)$ as X_t remains stationary. Generally speaking, the following characteristics of AR(p) models should be noted:

- When $h > p$, theoretical partial autocorrelation function (PACF) is zero:

$$\begin{aligned} \emptyset_{hh} &= \text{corr}(X_{t+h} - \widehat{X}_{t+h}, X_t - \widehat{X}_t) \\ &= \text{corr}(w_{t+h}, X_t - \widehat{X}_t \\ &= 0 \end{aligned}$$

- When $h \leq p$, $\emptyset_{pp}$ is not zero and $\emptyset_{11}, \emptyset_{22}, \cdots \emptyset_{h-1,\ h-1}$ are not always zero.

Because of these characteristics, the PACF is often the most effective method for locating an "AR model of order p".

Moving Average (MA) Models

Instead of using past values of the predicted variable as the autoregressive model does, a moving average model uses past prediction errors in a regression-like model. The moving average model is described as follows:

$$X_t = w_t + \sum_{j=1}^{q} \theta_j w_{t-j} \quad (16)$$

Similar to autoregressive models, the variance of the error component W_t will only change the scale of the series, not its patterns. Any stationary AR(p) model can be expressed as an MA (∞) model. This is called causality.

An AR (1) model is used as an example to illustrate causality:

$$X_t = \emptyset_1 X_{t-1} + w_t$$

$$= \emptyset_1(\emptyset_1 X_{t-2} + w_{t-1}) + w_t$$

$$= \emptyset_1{}^2 X_{t-2} + \emptyset_1 w_{t-1} + w_t$$

$$= \emptyset_1{}^3 X_{t-3} + \emptyset_1{}^2 w_{t-2} + \emptyset_1 w_{t-1} + w_t \text{ and so on.}$$

$X_t = w_t + \emptyset_1 w_{t-1} + \emptyset_1{}^2 w_{t-2} + \emptyset_1{}^3 X_{t-3} + \cdots$, an MA (∞) process.

However, if we limit the MA parameters in any manner, the outcome remains unchanged. The MA model is thus described as being invertible.

The MA (1) model can be used to demonstrate the invertibility concept.

$$X_t = w_t + \theta_1 w_{t-1}$$

$$w_t = \sum_{j=0}^{\infty} (-\theta)^j X_{t-j}$$

When $|\theta| > 1$, the weights increase as lags increase, so the more distant the observations the bigger their influence on the current error. When $|\theta| = 1$, the weights are static in size, and the distant observations have the same effect as the recent observations. As neither of these conditions make much sense, we need $|\theta| < 1$, so the most recent observations have higher weight than observations from the more distant past. Thus, the process is invertible when $|\theta| < 1$

The invertibility constraints for other models are similar to the stationarity constraints. Since weights rise as lags increase for $|\theta| > 1$, the more distant the data, the more they cause inaccuracy.

Autoregressive Moving Average (ARMA) models

Autoregressive moving average models are created by combining autoregressive and moving average models. It has the following definition:

$$X_t = w_t + \sum_{i=1}^{p} \emptyset_i X_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j} \quad (17)$$

It is crucial to remember that the time series $\{X_t, t = 0, \pm 1, \pm 2, \ldots\}$ is stationary and it is ARMA (p, q).

Techniques for bringing the non-stationary series to a stationary condition are part of the integrated component.

The ARIMA (p, d, q) model is defined as follows:

$$X_t = c + \emptyset_1 X'_{t-1} + \emptyset_2 X'_{t-2} + \cdots + \emptyset_p X'_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q} \quad (18)$$

Table 1: Established ARIMA Models

Type of model	Special ARIMA model
White Noise	ARIMA (0, 0, 0)
Random Walk	ARIMA (0, 1, 0) with no constant
Random Walk with drift	ARIMA (0, 1, 0) with a constant
Autoregression	ARIMA (p, 0, 0)
Moving Average	ARIMA (0, 0, q)

Source: Box & Jenkins (1970)

To make things easier, equation (18) can be expressed in terms of the back shift operator, as illustrated in equation (19):

$$\left(1 - \emptyset_1 B - \emptyset_p B^p\right)(1 - B)^d X_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q) w_t \quad (19)$$

AP(p) process Differences (d) MA(q) process

Forecasting requires an understanding of how ARIMA models behave. The long-term predictions that these models provide are significantly impacted by the constant c.

- The long-term projections will be zero if c = 0 and d = 0.
- The long-term forecasts will shift to a non-zero constant if c = 0 and d = 1.
- The long-term forecasts will show a straight line if c = 0 and d = 2.
- The long-term forecasts will be based on the mean if c ≠ 0 and d = 0.
- The long term forecasts will show a straight line if $c \neq 0$ and $d = 1$.
- The long term forecasts will exhibit a quadratic trend if $c \neq 0$ and $d = 2$.

Furthermore, the prediction intervals are influenced by the value of d; the higher the value of d, the faster the prediction intervals increase in size. All of the prediction intervals will be almost the same since the long-term forecast standard deviation for d = 0 will be equal to the standard deviation of the historical data. Along with a few additional parameter constraints, cyclic forecasts can only be generated if P > 2.

Seasonal behavior is present in an AR (2) model if $\emptyset_1^2 + 4\emptyset_2 < 0$. In that case, the average period of the season is $\frac{2\pi}{\arccos\left(-\emptyset_1(1-\emptyset_2)/(4\emptyset_2)\right)}$.

Seasonal ARIMA Models

ARIMA models can also be used to model a wide range of seasonal data. A seasonal ARIMA model is produced by including additional seasonal components in the ARIMA models we have already discussed. This is how it is written:

ARIMA	(p, d, q)	$(P, D, Q)_m$
	↑	↑
	Non-seasonal part of the model	Seasonal part of the model

The total number of observations made annually is the parameter m.

When the backshift notation is used, the periodic ARIMA model can be expressed as follows:

$$(1 - \emptyset_1 B)(1 - \emptyset_1 B^4)(1 - B)(1 - B^4)X_t = (1 + \theta_1 B)(1 + \theta_1 B^4)w_t \quad (20)$$

ACF and PACF of Seasonal ARIMA Models

The periodic occurrence component of an AR or MA model is displayed by the seasonal lags of the PACF and ACF. According to an ARIMA $(0,0,0)$ $(0,0,1)_{12}$ model, the ACF will show a spike at lag 12 but no further significant spikes, whereas the seasonal lags of the PACF that is, at lags 12, 24 will show exponential decline. Whereas the PACF will show a single significant spike at lag 12, the ACF's periodic lags will show an exponential drop. For an ARIMA $(0,0,0)$ $(1,0,0)_{12}$ model, this is also true.

Finding the appropriate periodic orders for a seasonal ARIMA model should only focus on the periodic lags. The modeling method includes selecting the seasonal AR and MA terms as well as the non-periodic components of the model.

The Box-Jenkins Modelling

"The Box-Jenkins method" is an analytical and forecasting technique. The three stages of the "Box-Jenkins approach" can be used to examine the data. Among the steps that comprise this process are identification, diagnosis, estimation, and prediction (Box & Jenkins, 1970). The Box-Jenkins technique is an iterative procedure; until a good and well-diagnosed model is produced, stages 1 through 3-from identification to diagnostic verification are frequently repeated. It is crucial to remember that the methodology is based on the assumption that the underlying data is generated by a stationary, linear process.

Identification of "candidate models"

The first identification stage of the Box-Jenkins approach aids in identifying the proper sequences for the autoregressive (AR), differencing (I), and moving average (MA) components of a given time series. Finding "p", "d", and "q" for the given time series is made simpler by this step. Determining if the time series is stationary and whether any significant seasonality needs to be addressed are the two most crucial goals at this level.

Stationarity check

The process of ensuring that time series statistics such as mean, variance, covariance, and standard deviation do not change over time is known as a stationarity check. By differentiating, data that is not already stationary is made so. A time series' stationarity can be ascertained in a number of ways.

- To identify any anomalies, such as seasonality or trend, the time series data is plotted for visual inspection.
- The autocorrelation function (rh) shown against time lags (h) is called a correlogram. The correlogram is a frequently used method for analyzing the unpredictability of a dataset. If time-lag separations are random, then autocorrelations for all of them should be near zero.
- To determine whether the time series is stationary, further rigorous tests are conducted. If it is not random, at least one autocorrelation will be distinctly non-zero. The two most often used tests are the "Augmented-Dickey Fuller (ADF)" and "Kwiatkowski Phillips Schmidt Shin (KPSS)".

ACF and PACF plots for model identification

The ACF plot and the closely related PACF plot can sometimes determine appropriate values for "p" and "q". Keep in mind that an ACF plot shows the autocorrelations that quantify the relationship between X_t and X_{t-k} over a range of values of k. When X_t and X_{t-1} are correlated, then $X_{t-1}$and X_{t-2} must also be correlated. However, X_t and X_{t-2} might be related just because they have a relationship to X_{t-1}, not because X_{t-2} has any new information that could be used to forecast X_t.

The use of partial autocorrelations helps overcome this problem. After the effects of lags 1, 2, 3, …, and k−1 are eliminated, these measure the correlation between X_t. and X_{t-k}. The initial partial autocorrelation is identical to the first autocorrelation, though, because there is nothing to remove from them. Each partial autocorrelation can be estimated using the final coefficient of an autoregressive model. Specifically, in an AR_k model, $\emptyset_k$ is equal to α_k, the k^{th} partial autocorrelation coefficient. If the data is from an ARIMA (p, d, 0) or ARIMA (0, d, q) model, the value of p or q can be determined using the ACF and PACF plots.

The data may be appropriate for an ARIMA (p, d, 0) model if the ACF and PACF graphs of the differenced data exhibit the following patterns:

- The Autocorrelation Function shows a peak at lag p and no additional spikes;
- ACF shows a significant peak at lag p but no further spikes.

The data may fit an ARIMA (0, d, q) model if the ACF and PACF graphs of the differenced data show the following patterns:

- The PACF is exponentially decaying or sinusoidal;
- The ACF shows a significant spike at lag q, but none passes lag q.

One significant discovery is that we may ignore one prominent spike in each plot as long as it is barely beyond the limits and outside of the first few lags. After all, there is only a one in twenty chance that a spike will become significant by accident (Box, 1970).

Model estimation

We must compute the parameters c, $\emptyset_1$, …, $\emptyset_p$, θ_1, …, θ_q after determining the candidate model and knowing the inputs of "p", "d", and "q". The Maximum Likelihood Estimation (MLE) is used in this sense.

First, the probability density function needs to be defined in order to generate the Maximum Likelihood Estimation:

$$f(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{\theta_2\sqrt{2\pi}}} \exp\left[-\left(\frac{x_i-\theta_1}{2\theta_2}\right)^2; \right. \quad (21)$$

$$\text{For } -\propto < \theta_1 < \propto \text{ and } 0 < \theta_2 < \propto$$

The Likelihood function then becomes;

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2) = \theta_2^{-\frac{n}{2}} \exp\left[\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right]$$

$$\log L(\theta_1, \theta_2) = -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum(x_i-\theta_1)^2}{2\theta_2} \quad (22)$$

In actuality, ARIMA models are harder to compute than regression models, and because they use distinct optimization algorithms and estimation methodologies, results from other statistical software will differ slightly. In practice, the statistical software will show the log likelihood of the data, which is the probability, represented as a logarithm, that the data observed originates from the model estimated. As previously demonstrated;

$$\log L(\theta_1, \theta_2) = -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum(x_i-\theta_1)^2}{2\theta_2} \quad (23)$$

The "Bayesian Information Criterion (BIC)" and "Akaike's Information Criterion (AIC)" values can also be used to determine c, $\emptyset_1$, …, $\emptyset_p$, θ_1, …, θ_q. The order of an ARIMA model can also be determined using the Akaike's Information Criterion:

$$AIC = -2\log(L) + 2(p + q + k + 1) \quad (24)$$

L is the likelihood of the data, and k = 1 if $c \neq 0$ and k = 0 if c = 0.

The number of parameters in the model, including the variance of the residuals, σ^2, is represented by the value of (p+q+k+1).

The adjusted AIC for ARIMA models can be written as follows:

$$AICc = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2} \quad (25)$$

Thus, one way to express the Bayesian Information Criterion is as follows:

$$BIC = AIC + [\log(T) - 2](p + q + k + 1) \quad (26)$$

To produce acceptable models, the AIC, AICc, or BIC should be kept to a minimum. These criteria are only helpful in selecting the values of "p" and "q", not in figuring out how to properly arrange differencing (d) in a model. Because differencing alters the data used to compute the likelihood, it is impossible to compare the AIC values of models with different differencing orders. Therefore, before utilizing the AICc to ascertain p and q, we must use alternate techniques to ascertain d. The model under consideration for selection, assuming all other factors remain constant, should satisfy the following conditions:

- Be the one with most coefficients that are statistically significant.
- Be the one with the least volatility, that is the lowest estimated error variance.

- Be the one with the highest log-likelihood statistic.
- Akaike Information Criterion and Bayessian Information Criterion should have the lowest values.

Model diagnostics

Every observation in the data can be predicted by using all of the previously collected data. The y_t forecast resulting from the observations $y_1, \ldots, y_{t-1}$ is represented by these fitted values, which we call $\acute{Y}_t|t\text{-}1$. Because it is so often, we may enter $\acute{Y}_t$ rather than $\acute{Y}_t|t\text{-}1$ and leave out a part of the subscript. Fitted values always include a single-step forecast. Fitted values are often not reliable forecasts since any parameters employed in the prediction approach are calculated using all available data in the time series, including future observations. For example, the average approach is used to obtain the fitted values;

$$\hat{y}_t = \hat{c} \tag{27}$$

All available observations are used to derive the drift parameter. The fixed values in this instance are created by

$$\hat{y}_t = y_{t-1} + \hat{c} \tag{28}$$

Where $\hat{c} = (y_T - y_1/(t-1)$.

It is necessary to estimate a parameter from the data for both equations 27 and 28. The "hat" above the c indicates that this is an estimate. The fitted values are not precise forecasts when c is determined from observations made over time t. Fitted values, however, are precise predictions in these cases since naïve or seasonal naïve forecasts don't employ any parameters. After fitting a time series model, what's left over are called "residuals."

$$e_t = y_t - \hat{y}_t \tag{29}$$

When determining if a model has accurately depicted the data, residuals are useful. A successful forecasting method will yield residuals with the following properties:

- The mean of residuals does not change over time
- Non-correlation exists among the residuals.

These conditions can be met by several different methodologies that forecast the same data gathered. Keeping an eye on these characteristics is essential to determine whether a forecasting method makes use of all the pertinent data, even though it is not a wise strategy to choose one. If either of these characteristics is not met, the forecasting algorithm can be changed. Prediction intervals are calculated using the following two characteristics. It is not always possible to improve a forecasting method that does not meet these criteria.

Apart from ACF and Histogram plots, autocorrelation test is used in evaluating the total number of r_k values collectively rather than as separate numbers. As you may recall, the autocorrelation for lag k is r_k. We are unintentionally performing many hypothesis tests. When enough of these tests are run, it is likely that at least one of them will result in a false positive, which would cause us to assume that the residuals still show some autocorrelation when in reality they do not.

One "portmanteau" test, the “Box-Pierce test”, which is predicated on the following statistic can be used:

$$Q = T\sum_{k=1}^{l} r^2{}_k \qquad (30)$$

Where T is the number of observations and ℓ is the greatest lag that is considered. If all of the r_k are close to zero, Q will be quite little. Large r_k numbers, whether positive or negative, will result in a large Q. With m being the seasonality

period, we advise using ℓ=2m for seasonal data and ℓ=10 for non-seasonal data. However, as the test is not good when ℓ is enormous, use ℓ=T/5 if these numbers are greater than T/5. The Ljung-Box test is regarded as a related (and more accurate) test because of the following:

$$Q *= T(T+2)\sum_{k=1}^{l}(T-k)^{-1}\,r^{2}{}_{k} \qquad (31)$$

Forecasting

The following three steps can be used to compute point forecasts. It is necessary to expand the ARIMA equation so that yt is on the left and all other terms are on the right. To rephrase the equation, substitute T+h for t. On the right-hand side of the formula, substitute the correct residuals for historical errors, zero for potential future errors, and the projections for future observations.

Beginning with h = 1, these steps are repeated for h = 2, 3, …, and so on, until all forecasts have been calculated. The technique is simplest to understand when illustrated using an example. We will illustrate it using the ARIMA (3,1,1) model fitted in the previous section. The following is one approach to express the model:

$$(1-\hat{\emptyset}_1B-\hat{\emptyset}_2B^2-\hat{\emptyset}_3B^3)(1-B)y_t=(1+\hat{\theta}_1B)\varepsilon_t \qquad (32)$$

Expanding the left-hand side, we receive;

$$[1-(1+\hat{\emptyset}_1)B+(\hat{\emptyset}_1-\hat{\emptyset}_2)\quad B^2+(\hat{\emptyset}_2-\hat{\emptyset}_3)\quad B^3+\hat{\emptyset}_3B^4]y_t=(1+\hat{\theta}_1B)\varepsilon_t \qquad (33)$$

The backshift operator is used to get;

$$y_t-(1+\hat{\emptyset}_1)y_{t-1}+(\hat{\emptyset}_1-\hat{\emptyset}_2)y_{t-2}+(\hat{\emptyset}_2-\hat{\emptyset}_3)y_{t-3}+\hat{\emptyset}_3y_{t-4}=\varepsilon_t+\hat{\theta}_1\varepsilon_{t-1}$$

rearranging on the left side with y_t;

$$y_t=(1+\hat{\emptyset}_1)y_{t-1}-(\hat{\emptyset}_1-\hat{\emptyset}_2)y_{t-2}-(\hat{\emptyset}_2-\hat{\emptyset}_3)y_{t-3}-\hat{\emptyset}_3y_{t-4}+\varepsilon_t+\hat{\theta}_1\varepsilon_{t-1} \qquad (34)$$

In equation (33) for the second step, we substitute T+1 for t:

$$y_{t+1} = (1 + \hat{\emptyset}_1)y_T - (\hat{\emptyset}_1 - \hat{\emptyset}_2)y_{T-1} - (\hat{\emptyset}_2 - \hat{\emptyset}_3)y_{T-2} - \hat{\emptyset}_3 y_{T-3} + \varepsilon_{T+1} + \hat{\theta}_1 \varepsilon_T \quad (35)$$

Substituting ϵ_{T+1} with zero, and ϵ_T, with the final residual e_T:

$$y_{T+1|T} = (1 + \hat{\emptyset}_1)y_T - (\hat{\emptyset}_1 - \hat{\emptyset}_2)y_{T-1} - (\hat{\emptyset}_2 - \hat{\emptyset}_3)y_{T-2} - \hat{\emptyset}_3 y_{T-3} + \hat{\theta}_1 e_T \quad (36)$$

By substituting T+2 for t in (36), a prediction of y_{T+2} is obtained; except for y_{T+1}, which we substitute with $y\hat{}_{T+1|T}$, and ε_{T+2} and ε_{T+1}, which we substitute with zero:

$$\hat{y}_{T+2|T} = (1 + \hat{\emptyset}_1)\hat{y}_{T+1|T} - (\hat{\emptyset}_1 - \hat{\emptyset}_2)y_T - (\hat{\emptyset}_2 - \hat{\emptyset}_3)y_{T-1} - \hat{\emptyset}_3 y_{T-2} \quad (37)$$

This is how the procedure goes on for all subsequent times. Any number of point forecasts can be produced in this manner.

The result is true for all ARIMA models, regardless of their ordering and parameters. The multi-step forecast interval model for ARIMA (0, 0, q) can be expressed as follows:

$$y_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

The anticipated forecast variance can therefore be expressed as follows:

$$\hat{\sigma}_h = \hat{\sigma}^2{}_h \left[1 + \sum_{i=1}^{h-1} \hat{\theta}_i^{\,2}\right], \text{ h= 2, 3, . . .}$$

Where $\hat{\theta}_i = 0$ for $i > q$, and a 95% prediction interval is given by $\hat{y}_{T+h|T} \pm 1.96\sqrt{\hat{\sigma}h}$.

As was previously shown, an AR (1) model can be expressed as an MA (∞) model. Prediction intervals for AR (1) models can be derived using the same outcome as for MA(q) models due to its comparability. To compute the prediction intervals for ARIMA models, the residuals are assumed to be uncorrelated and normally distributed. The prediction intervals may not be accurate if either of these assumptions is incorrect. Therefore, it is crucial to plot the residuals' ACF and histogram in order to validate the assumptions before generating prediction intervals. The prediction intervals of ARIMA models usually increase with the

length of the forecast horizon. Stationary models will converge at d = 0, producing prediction intervals for longer time horizons.

Future prediction intervals will continue to grow for d ≥ 1. Like the majority of them, ARIMA-based prediction intervals are frequently unduly small. This occurs as a result of only the error variation being taken into account. Furthermore, the computation has not taken into account the variation in model order and parameter estimations. Additionally, the calculation assumes that the modeled historical trends will continue over the course of the prediction period (Brockwell & Davis, 2016).

When evaluating forecast accuracy, one should consider how well a model performs on new data that was not used to fit the model. When choosing models, it is normal practice to separate the available data into training and test data. The test data is used to evaluate the accuracy of the forecasting approach, while the training data is used to estimate any parameters. Since the test data is not used to produce the forecasts, it should provide a reliable signal of how well the model is likely to forecast on new data. The size of the test set is typically 20% of the whole sample; however, this percentage depends on the sample size and the required prediction time. Notice that;

- A perfect fit is always possible when a model has enough parameters.
- Missing a systematic trend in the data is just as bad as overfitting a model to the data.

The unexpected part of an observation called “errors” is stated as;

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$$

When it comes to scale-dependent errors, the two most popular approaches are squared errors and absolute errors, which are defined as follows:

- The definition of mean absolute error (MAE), as stated by Hamzacebi (2008) and Zhang (2007), is MAE=mean (| et |).

The researcher can make informed conclusions about the model's predicted values thanks to the inherent qualities of this metric. Here are a handful of them:

- ✓ The mean absolute non-similarities between the initial and forecasted values are computed.
- ✓ It determines the extent of the overall forecasting error.
- ✓ Errors, both positive and negative, exacerbate each other.
- ✓ To get appropriate forecast, the obtained MAE should be as low as possible.
- ✓ The measurement scale and data processing also affect MAE.

  Extreme forecast errors aren't panelized in MAE.

- "Root mean squared error (RMSE)" is defined as follows by Park (1999) and Zhang (2007): RMSE = $\sqrt{}$ (mean (e_t^2)).

Some of its attributes are as follows:

- ✓ RMSE gives a broad idea of the errors that happen when forecasting.
- ✓ The opposite signed errors have no effect on one another
- ✓ Significant individual errors do, in fact, have a meaningful effect on the overall forecast inaccuracy, as the "MSE" emphasizes.
- ✓ RMSE does not reveal the overall error's direction.
- ✓ Changes in size and data manipulations may affect RMSE.
- ✓ Data transformations and size changes can have an impact on RMSE.

Although being more difficult to comprehend, the RMSE is also widely used.

- The definition of mean absolute percentage error (MAPE), as stated by Hamzacebi (2008) and Park (1999), is MAPE=mean (| pt |).

But there are certain disadvantages to this approach. If any y_t is close to zero, it has extreme values; if $y_t = 0$ for any t during the time of interest, it is endless or undefined. Its foundation is a meaningful zero in the measurement unit. Furthermore, errors that are negative are penalized more harshly than those that are good. As a result, Armstrong (1978) identified symmetric Mean Absolute Percentage Error (sMAPE) and described it as follows:

$$\text{sMAPE} = \text{mean}\,(200|y_t - \hat{y}_t)/(y_t + \hat{y}_t).$$

The following are its working conditions:

- ✓ Data transformation has an effect on it regardless of the measurement scale.
- ✓ The direction of the error is not shown.
- ✓ High variances are not panelized by MAPE.
- ✓ It represents the proportion of average absolute error that occurred.
- ✓ The opposing signed errors, which do not eliminate one another, are used to symbolize it.

According to Hyndman & Koehler (2006), "scaled errors" can be used in place of percentage errors, and it is defined as:

$$q_j = \frac{e_j}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|}$$

Chapter Summary

This chapter demonstrated how to use the Box-Jenkins time series modeling approach to predict the transmission of tuberculosis in Ghana. The first topics covered were time series ideas, basic time series models, and how to tell if a data is stationary. The chapter also includes a thorough description of Box-Jenkins time series modeling, which estimates an ARIMA model using a three-stage iterative approach. The processes for determining model parameters, diagnosing the model,

and discovering possible models were all carefully looked at. A variety of theories that require rigorous statistical testing were examined where suitable. Additionally, a detailed explanation of forecasting measures and the forecasting stage, the fourth step of the Box-Jenkins technique, was provided.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

A monthly total number of TB cases reported in Ghanaian medical facilities was the source of data for the study. This secondary data is stored in the Ministry of Health's District Health Information Management System (DHIMS). 126 data points between January 2014 and June 2024 were picked for the study. The analysis was conducted using a number of statistical software packages, including Stata, SPSS, and Microsoft Excel. A time series was created using the TB data. A visual analysis of the stationarity of the plot was conducted. To conduct additional stationarity checks, an autocorrelation plot was employed. After transforming the data, the stationarity status was once more determined. Using the ACF and PACF graphs, possible models were found. Parameter estimation procedures enabled the proper model to be chosen. Before being used for forecasting, the selected model was put through "Ljung-Box Q test" for residuals. The Box-Jenkins method was applied, and the various steps were part of the study. A detailed discussion followed a comparison of the findings of the analysis with previous research.

Results

Tuberculosis time series plot

Figure 1 shows a time series plot of TB data from January 2014 to June 2024. The information was compiled from secondary data on monthly TB cases recorded in Ghanaian healthcare facilities.

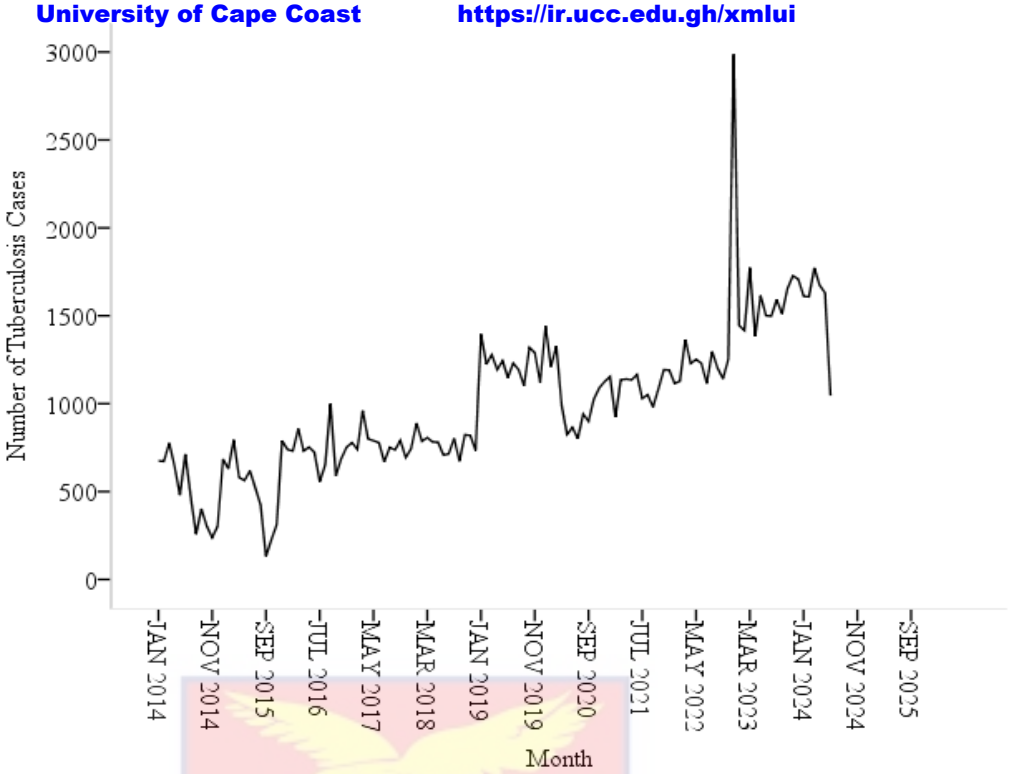


Figure 1: Time Series Plot of Tuberculosis Disease in Ghana

Monthly cases of tuberculosis recorded in Ghanaian healthcare facilities between January 2014 and June 2024 were included in the data as shown in figure 1. There were 126 observations in the series overall.

The series appears to be growing over time based on visual inspection. The mean of the series is seen rising with time, indicating an upward tend. A close examination of the plot reveals seasonal changes of some kind as well. The series has a peak and a trough, which repeats itself every calendar year. As a result, the series seems to be on an upward trend multiplicative seasonal cycle. The series is therefore changing with time. Creation of autocorrelation function and partial autocorrelation function plots could support this observation. These are presented in Figures 2 and 3.

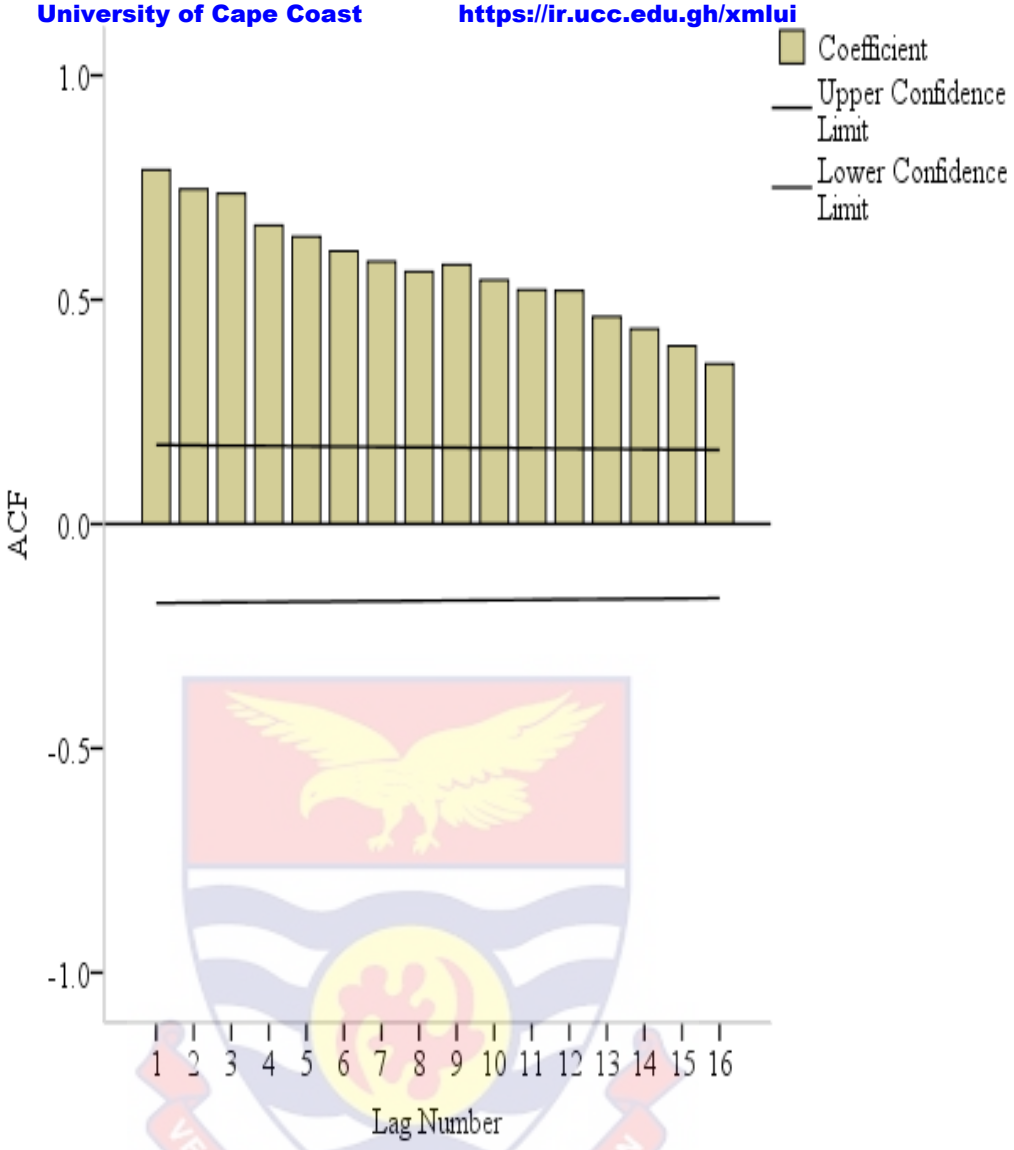


Figure 2: Sample ACF Plot for Monthly TB Disease in Ghana.

As lag rises, sample ACF values become noticeably larger and decrease very slowly. The series was still displaying relevance even at lag 15. Box (1970) defined a non-stationary series as one that decays slowly and shows large meaningful values at increasing lags. This validates the previous visual examination by demonstrating that the tuberculosis data is not stationary.

Figure 3 presents plot of Partial Autocorrelation Function, to also test the stationarity status of the TB time series.

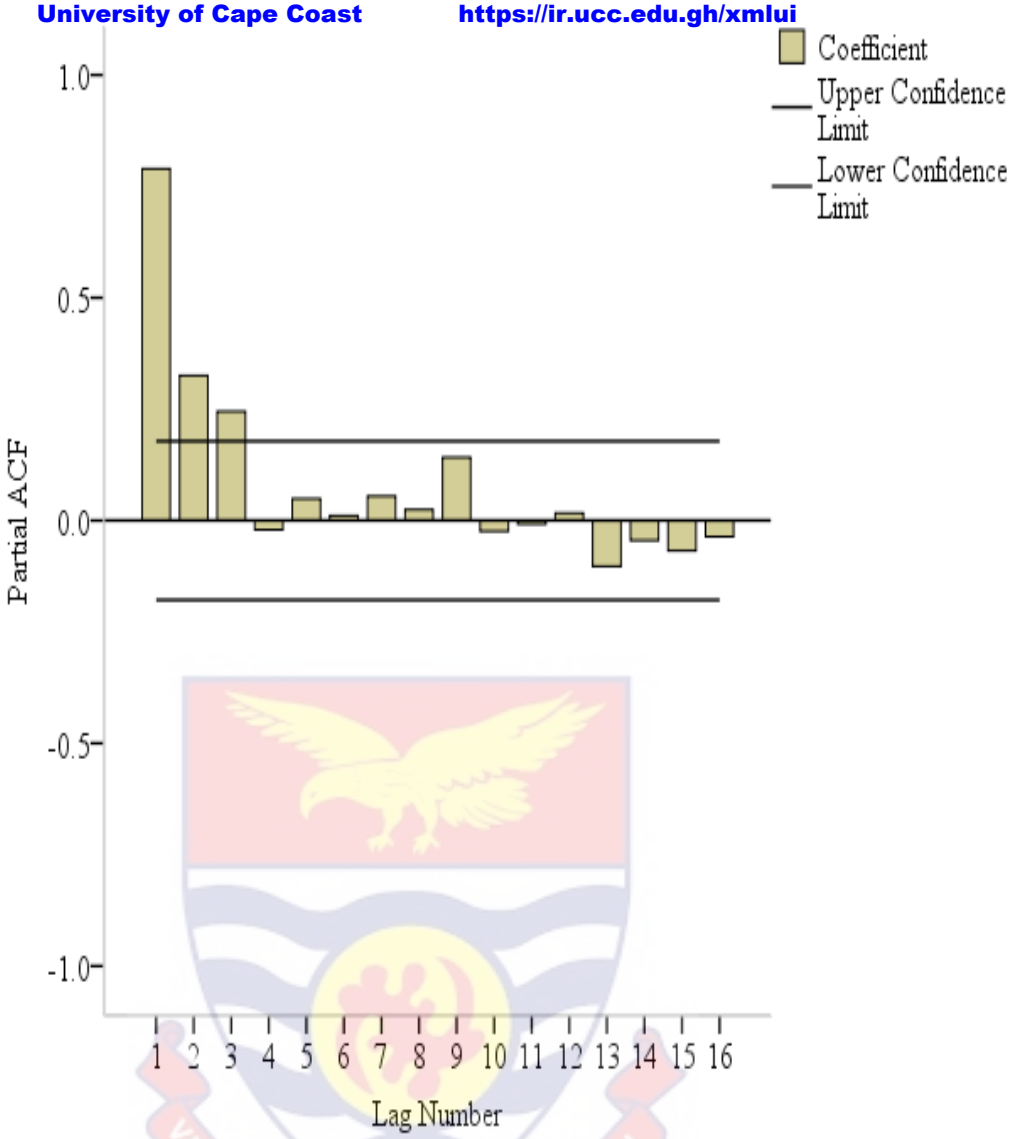


Figure 3: Sample PACF Plot for Monthly TB Disease in Ghana.

Additionally, the sample PACF decays slowly and exhibits substantial significance as the lag grows. In addition to the TB time series plot, ACF, and PACF plots, which have shown the series to be non-stationary, a more formal test known as the "Augmented Dickey-Fuller (ADF)" test helped to further clarify the issue. Table 2 presents the findings.

Table 2: Augmented Dickey-Fuller Test of TB Series

Item	Value
Test statistic	-2.505
P-value	0.1144
# of lags used	1
# of observations used	124
Critical value (1%)	-3.502
Critical value (5%)	-2.888
Critical value (10%)	-2.578

Source: Researcher, 2024

The Augmented Dickey-Fuller test actually test the following hypothesis;

H_0: The TB data is not stationary

H_1: The TB data is stationary

The test statistic is greater than the 5% "critical value". As a result, we lacked adequate justification to reject the null hypothesis, H_0, and so declare the series non-stationary at this point. The p value lends further credence to this. The non-stationarity of the TB series has been established beyond reasonable doubt by these three sources.

Transformation to achieve stationarity of Tuberculosis time series

Once the non-stationarity of the series was confirmed, an attempt was made to make it stationary. This is because forecasting, hypothesis testing, and prediction cannot be done with a non-stationary time series (Box, 1970). Transformations were used to make the data stationary. Sometimes taking first difference of the

series can make the series stationary by removing both the trend and the seasonality. In some other cases, a seasonal diffrencing is required to deal with the seasonality. There may be instances too where both first difference and seasonal difference need to be taken before the series becomes stationary.

At this point,first difference was taken to remove the trend. Figure 4 illustrated this.

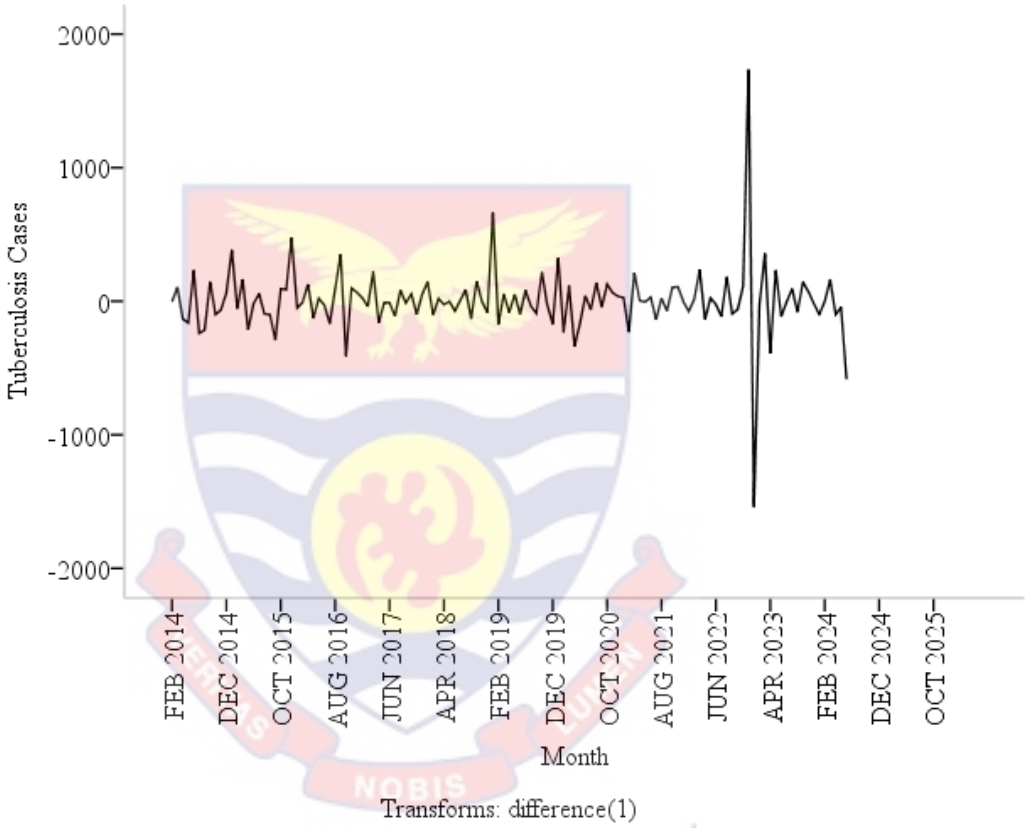


Figure 4: First Difference of Tuberculosis Time Series

The plot in figure 4 shows that the mean of the series is constant over time, indicating that the trend has been removed. However, there is evidence of up and down movement of the plot which is repititive yearly. This is suggestive of the

presence of seasonality. Seasonal difference was then taken in addition to the first difference. Figure 5 shows the first difference and seasonal difference of the data.

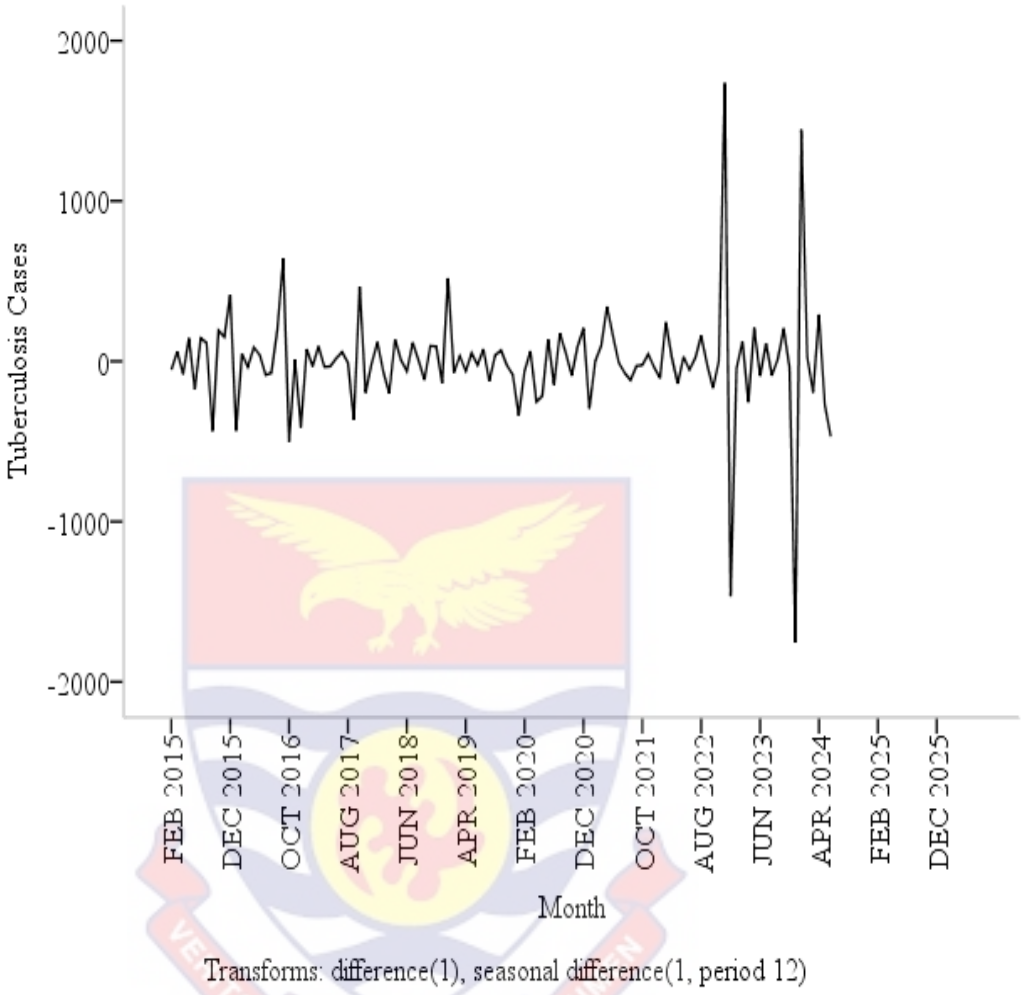


Figure 5: First Difference, First Seasonal Difference of TB Time Series

Following the first difference and the first seasonal difference, the data is now comparatively stationary. The Augmented Dickey-Fuller (ADF) test, a formal unit root test for the differenced TB data, is shown in Table 3.

The following hypothesis was examined by the ADF:

H_0: The TB data is not stationary

H_1: The TB data is stationary

Table 3: Augmented Dickey-Fuller Test of Differenced TB Series

Item	Value
Test statistic	-12.496
P-value	0.0000
# of lags used	1
# of observations used	123
Critical value (1%)	-3.502
Critical value (5%)	-2.888
Critical value (10%)	-2.578

Source: Researcher, 2024

Table 3 shows that the absolute value of the test statistic, -12.496, is more than the absolute value of 5% critical value of -2.888. With this, we have enough reason to reject the null hypothesis and declare the TB series to be stationary. The TB series is currently stationary after the first difference and first seasonal difference.

Identification of candidate models

Finding a suitable ARIMA model for our TB series is now our goal. Plotting the ACF and PACF allowed for this. Figures 6 and 7 display the two plots. The prior assertion that the series is stationary is supported by the ACF and PACF plots, which show a small number of significant spikes at different lags and the remaining spikes that gradually decays to zero.

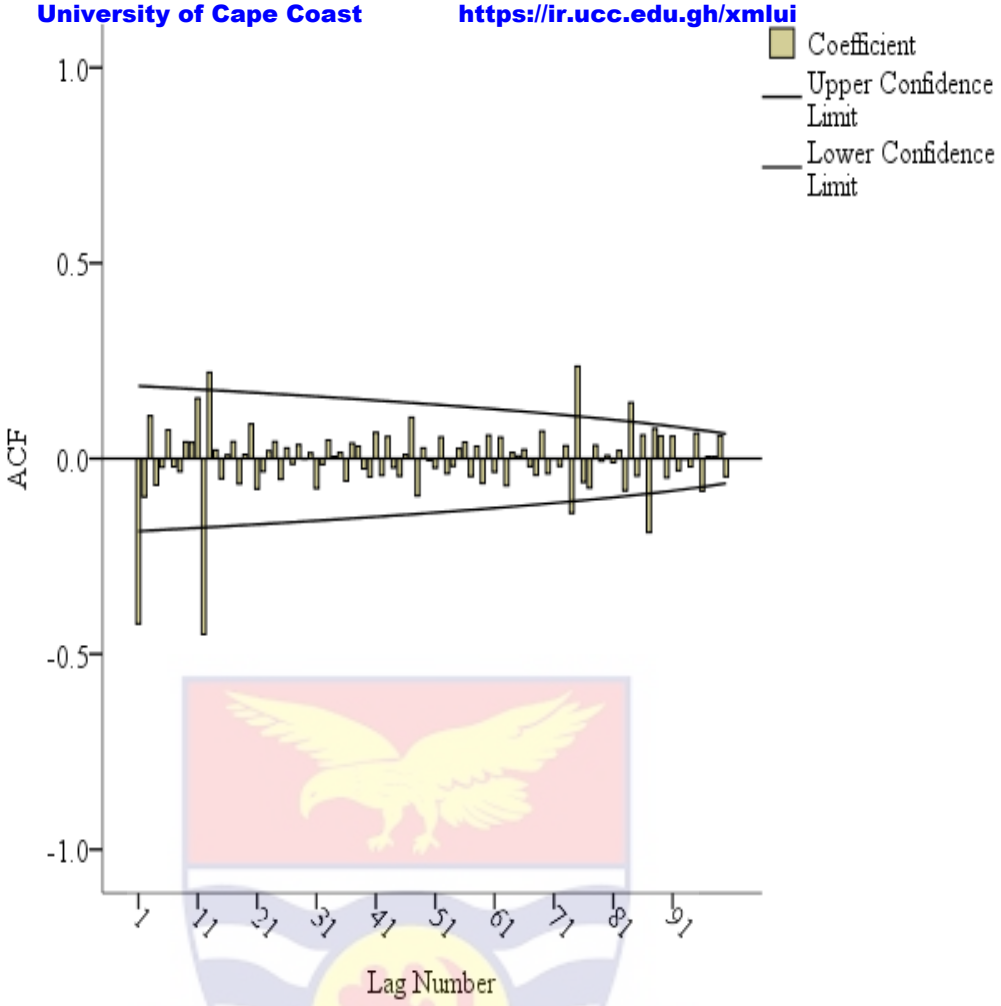


Figure 6: Sample ACF Plot for TB Time Series after Transformation

The seasonal component and the non-seasonal component are the two parts of the ACF plot shown in Figure 6. The ACF shows a substantial increase at lag 13 in the non-seasonal component, which points to a non-seasonal MA (1) component. Significant spikes in the ACF at lags 1 and 12 indicate a seasonal MA (2) root in the seasonal component. Consequently, we discovered SARIMA $(0,1,1)\ (0,1,2)_{12}$ model using the ACF plot, which shows a seasonal difference and first difference as well as non-seasonal MA (1) and seasonal MA (2) components.

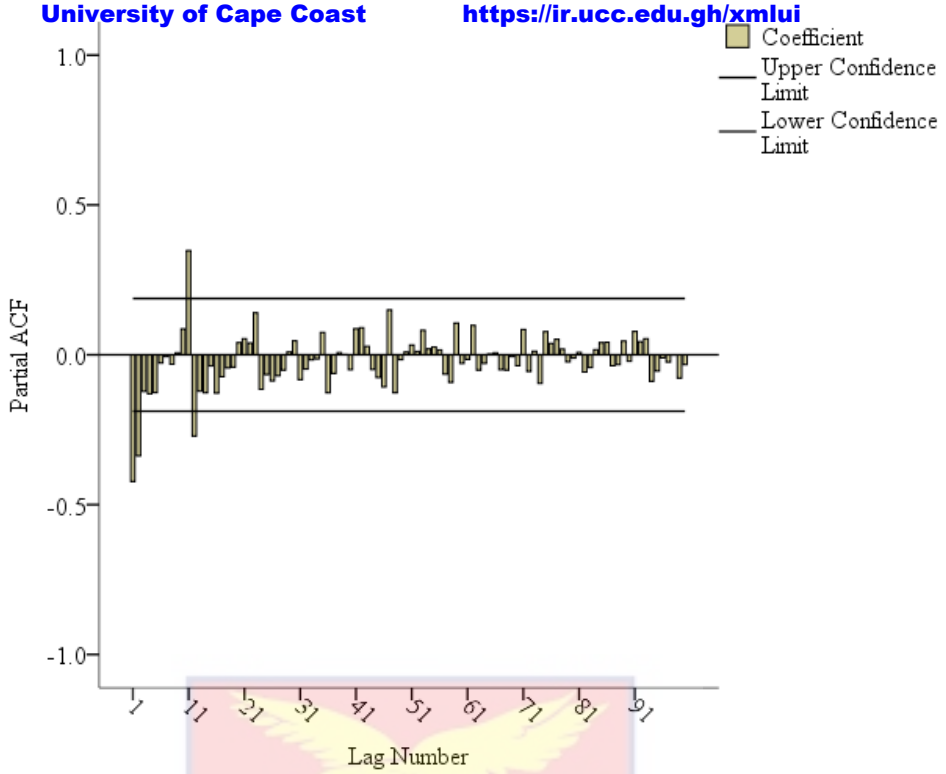


Figure 7: Sample PACF Plot for TB Time Series after Transformation.

The seasonal component and the non-seasonal component are the two parts of the PACF plot shown in Figure 7. Once more, a non-seasonal AR (1) component is suggested by the notable jump in the PACF at lag 11 in the non-seasonal component. Relevant increases at lags 1, 2, and 12 in the PACF indicate a seasonal AR (3) component. As a result, using the PACF plot, we were able to identify an ARIMA (1,1,0) $(3,1,0)_{12}$ model that had non-seasonal AR (1) and seasonal AR (3) elements after first and seasonal differences have been applied. SARIMA (3,1,0) $(1,1,0)_{12}$ and SARIMA (0,1,1) $(0,1,1)_{12}$ were other potential models. Thus, SARIMA (3,1,0) $(1,1,0)_{12}$, SARIMA (1,1,0) $(3,1,0)_{12}$, SARIMA (1,1,3) $(3,1,3)_{12}$, SARIMA (2,1,0) $(2,1,0)_{12}$, SARIMA (0,1,2) $(0,1,2)_{12}$, SARIMA (1,1,0) $(1,1,0)_{12}$, SARIMA (0,1,1) $(0,1,1)_{12}$, SARIMA (3,1,0) $(3,1,0)_{12}$, SARIMA (0,1,3) $(0,1,3)_{12}$ and SARIMA (0,1,1) $(0,1,3)_{12}$ were finally selected as the candidate models.

Parameters of the candidate models

We are now considering ten different candidate models and they include: SARIMA (3,1,0) $(1,1,0)_{12}$, SARIMA (1,1,0) $(3,1,0)_{12}$, SARIMA (1,1,3) $(3,1,3)_{12}$, SARIMA (2,1,0) $(2,1,0)_{12}$, SARIMA (0,1,2) $(0,1,2)_{12}$, SARIMA (1,1,0) $(1,1,0)_{12}$, SARIMA (0,1,1) $(0,1,1)_{12}$, SARIMA (3,1,0) $(3,1,0)_{12}$, SARIMA (0,1,3) $(0,1,3)_{12}$ and SARIMA (0,1,1) $(0,1,3)_{12}$. All of these potential candidate models are estimated over time. This will help us identify a parsimonious and stationary model that works well with the data. The model that meets these four requirements will be the best one:

- The model that has the most number of coefficients that are statistically significant.
- The one having the largest log likelihood statistic.
- Lowest AIC and BIC values
- One with the smallest estimate of the error variance, which is determined by the sigma square value.

The features of the ten identified candidate models are shown in Tables 4 through 13. Table 14 provided a summary of these. This will enable us to identify the optimal model that meets the previously mentioned requirements.

Table 4 presents characteristics of the model SARIMA (3,1,0) $(1,1,0)_{12}$.

Table 4: Characteristics of SARIMA (3,1,0) (1,1,0)$_{12}$ Model

Model	Coef.	Std. Err.	P-value	95% confidence interval	
SARIMA (3,1,0) (1,1,0)$_{12}$					
Constant	1.006	15.453	0.948	-29.28	31.29
ARMA					
AR (1)	-1.389	0.082	0.000	-1.55	-1.23
AR (2)	-1.061	0.111	0.000	-1.28	-0.84
AR (3)	-0.428	0.093	0.000	-0.61	-0.25
SARIMA					
AR (1)$_{12}$	-0.001	0.040	0.990	-0.08	0.08
Sigma SQ	565.3				
Log likelihood	-869.8				
AIC	1751.7				
BIC	1767.9				

Source: Researcher, 2024

The non-seasonal AR (1), AR (2) and AR (3) models were statistically significant. The constant was however statistically not significant. The seasonal AR (1) model was statistically not significant.

Table 5 presents characteristics of the model SARIMA (1,1,0) (3,1,0)$_{12}$.

Table 5: Characteristics of SARIMA (1,1,0) $(3,1,0)_{12}$ Model

Model	Coef.	Std. Err.	P-value	95% confidence interval	
SARIMA (3,1,0) $(1,1,0)_{12}$					
Constant	3.056	45.633	0.947	-86.38	92.49
ARMA					
AR (1)	-0.730	0.036	0.000	-0.80	-0.66
SARIMA					
AR $(1)_{12}$	-0.002	0.042	0.954	-0.09	0.08
AR $(2)_{12}$	0.000	0.086	0.998	-0.17	0.17
AR $(3)_{12}$	-0.000	0.0134	1.000	711.06	837.34
Sigma SQ	565.3				
Log likelihood	-869.8				
AIC	1810.2				
BIC	1826.4				

Source: Researcher, 2024

The non-seasonal AR (1) model was statistically significant. The constant was however statistically not significant. The seasonal AR (1), AR (2) and AR (3) models were all statistically not significant. The implication is that the model will not have a constant term. Also, SARIMA (1,1,0) $(3,1,0)_{12}$ model is basically an AR(1) model.

Table 6 shows characteristics of the model SARIMA (1,1,3) $(3,1,3)_{12}$.

Table 6: Characteristics of SARIMA (1,1,3) $(3,1,3)_{12}$ Model

Model	Coef.	Std. Err.	P-value	95% confidence interval	
SARIMA (1,1,3) $(3,1,3)_{12}$					
Constant	-0.014	0.468	0.976	-0.90	0.93
ARMA					
AR (1)	-0.068	0.082	0.406	-0.23	0.09
MA (1)	-2.413	0.000	0.000	-2.41	-2.41
MA (2)	1.842	0.000	0.000	1.84	1.84
MA (3)	-0.427	0.001	0.000	-0.43	-0.42
SARIMA					
AR $(1)_{12}$	-0.307	0.002	0.000	-0.31	-0.31
AR $(2)_{12}$	0.795	0.009	0.000	0.79	0.79
AR $(3)_{12}$	-0.250	0.007	0.000	-0.25	-0.25
MA $(1)_{12}$	0.305	0.002	0.000	0.31	0.31
MA $(2)_{12}$	-0.794	0.003	0.000	-0.79	-0.79
MA $(3)_{12}$	0.250	0.002	0.000	0.25	0.25
Sigma SQ	304.2				
Log likelihood	-820.0				
AIC	1656.8				
BIC	1678.5				

Source: Researcher, 2024

The non-seasonal AR (1) was statistically not significant. However, non-seasonal MA (1), MA (2) and MA (3) models were statistically significant. The

constant was statistically not significant. The seasonal AR (1), AR (2), AR (3), MA (1), MA (2) and MA (3) were all statistically significant.

Table 7 presents characteristics SARIMA (2,1,0) $(2,1,0)_{12}$.

Table 7: Characteristics of SARIMA (2,1,0) $(2,1,0)_{12}$ Model

Model	Coeff.	Std. Err.	P-value	95% confidence interval	
SARIMA (2,1,0) $(2,1,0)_{12}$					
Constant	1.516	25.369	0.952	-48.21	51.24
ARIMA					
AR (1)	-1.132	0.071	0.000	-1.27	-0.99
AR (2)	-0.564	0.064	0.000	-0.69	-0.44
SARIMA					
AR $(1)_{12}$	-0.001	0.104	0.984	-0.09	0.09
AR $(2)_{12}$	0.004	0.104	1.00	-0.20	0.20
Sigma SQ	635.4				
Log likelihood	-880.2				
AIC	1772.3				
BIC	1788.6				

Source: Researcher, 2024

Both non-seasonal AR (1) and AR (2) models were statistically significant. The constant was statistically not significant. Seasonal AR (1) and AR (2) were also statistically not significant. This implies that SARIMA (2,1,0) $(2,1,0)_{12}$ model will not exhibit a constant term, and that it is basically a non-seasonal model since the seasonal components were statistically not significant.

Table 8 shows characteristics of the model SARIMA $(0,1,2)\ (0,1,2)_{12}$. This will help ascertain if it is the most appropriate model for TB disease in Ghana.

Table 8: Characteristics of SARIMA $(0,1,2)\ (0,1,2)_{12}$ Model

Model	Coef.	Std. Err.	P-value	95% confidence interval	
SARIMA $(0,1,2)\ (0,1,2)_{12}$					
Constant	0.003	0.209	0.990	-0.41	0.41
ARIMA					
MA (1)	-1.991	0.429	0.000	-2.83	-1.15
MA (2)	0.997	0.415	0.016	0.18	1.81
SARIMA					
$MA\ (1)_{12}$	-0.001	0.038	0.997	-0.08	0.07
$MA\ (2)_{12}$	0.002	0.036	0.994	-0.07	0.07
Sigma SQ	374.9				
Log likelihood	-833.6				
AIC	1679.3				
BIC	1695.5				

Source: Researcher, 2024

The non-seasonal MA (1) and MA (2) models were statistically significant. The constant was statistically not significant. Seasonal MA (1) and MA (2) were also statistically not significant.

Table 9 shows characteristics of the model SARIMA $(1,1,0)\ (1,1,0)_{12}$.

Table 9: Characteristics of SARIMA (1,1,0) $(1,1,0)_{12}$ Model

Model	Coef.	Std. Err.	P-value	95% confidence interval	
SARIMA (1,1,0) $(1,1,0)_{12}$					
Constant	3.052	44.601	0.945	-84.36	90.47
ARIMA					
AR (1)	-0.730	0.034	0.000	-0.79	-0.66
SARIMA					
AR $(1)_{12}$	-0.002	0.042	0.955	-0.86	0.08
Sigma SQ	774.2				
Log likelihood	-899.1				
AIC	1806.2				
BIC	1817.0				

Source: Researcher, 2024

The non-seasonal AR (1) model was statistically significant. The constant was statistically not significant. The seasonal AR (1) model was also statistically not significant. Since the non-seasonal AR(1) was statistically not significant, SARIMA (1,1,0) $(1,1,0)_{12}$ model is basically an AR(1) model.

Table 10 presents characteristics of the model SARIMA (0,1,1) $(0,1,1)_{12}$. These characteristics will help ascertain whether or not the model is appropriate for the TB disease in Ghana.

Table 10: Characteristics of SARIMA (0,1,1) $(0,1,1)_{12}$ Model

Model	Coeff.	Std. Err.	P-value	95% confidence interval	
SARIMA (0,1,1) $(0,1,1)_{12}$					
Constant	-0.356	1.619	0.826	-3.53	2.82
ARIMA					
MA (1)	-0.999	0.605	0.099	-2.18	0.18
SARIMA					
MA $(1)_{12}$	-0.003	0.044	0.950	-0.09	0.09
Sigma SQ	625.4				
Log likelihood	-881.3				
AIC	1770.7				
BIC	1781.5				

Source: Researcher, 2024

The models did not show statistical significance. The constant was also statistically not significant. This implies that SARIMA (0,1,1) $(0,1,1)_{12}$ Model is statistically not ideal for the TB disease in Ghana.

Table 11 presents characteristics of the model SARIMA (3,1,0) $(3,1,0)_{12}$. These characteristics will help ascertain whether or not the model is appropriate for the TB disease in Ghana.

Table 11: Characteristics of SARIMA (3,1,0) $(3,1,0)_{12}$ Model

Model	Coeff.	Std. Err.	P-value	95% confidence interval	
SARIMA (3,1,0) $(3,1,0)_{12}$					
Constant	1.017	15.798	0.949	-29.95	31.98
ARIMA					
AR (1)	-1.389	0.087	0.000	-1.56	-1.22
AR (2)	-1.061	0.127	0.000	-1.31	-0.81
AR (3)	-0.428	0.103	0.000	-0.63	-0.23
SARIMA					
AR $(1)_{12}$	-0.001	0.043	0.990	-0.09	0.08
AR (2)12	0.000	0.111	1.000	0.22	0.08
AR (3)12	0.005	0.156	1.000	-0.31	0.31
Sigma SQ	565.3				
Log likelihood	-869.8				
AIC	1755.7				
BIC	1777.3				

Source: Researcher, 2024

All the non-seasonal AR models, that is, AR (1), AR (2) and AR (3) were statistically significant. On the other hand, all the seasonal AR models, as well as the constant were statistically not significant. This is typically non-seasonal AR model.

Table 12 presents characteristics of the model SARIMA (0,1,3) $(0,1,3)_{12}$.

Table 12: Characteristics of SARIMA (0,1,3) (0,1,3)$_{12}$ Model

Model	Coeff.	Std. Err.	P-value	95% confidence interval	
SARIMA (0,1,3) (0,1,3)$_{12}$					
Constant	0.002	0.025	0.935	-0.05	0.05
ARIMA					
MA (1)	-2.607	0.409	0.000	-3.41	-1.81
MA (2)	2.219	0.636	0.000	0.97	3.47
MA (3)	-0.612	0.232	0.008	-1.06	-0.16
SARIMA					
MA (1)$_{12}$	-0.001	0.045	0.988	-0.09	0.09
MA (2)$_{12}$	0.000	0.095	1.000	-0.19	0.19
MA (3)$_{12}$	0.006	0.139	1.000	-0.27	0.27
Sigma SQ	317.9				
Log likelihood	-820.4				
AIC	1658.1				
BIC	1682.4				

Source: Researcher, 2024

The table showed that all the non-seasonal MA models, that is, MA (1), MA (2) and MA (3) were statistically significant. On the other hand, all the seasonal MA models, that, MA (1)$_{12}$, MA (2)$_{12}$ and MA (3)$_{12}$, as well as the constant were statistically not significant.

Table 13 presents characteristics of the model SARIMA (0,1,1) (0,1,3)$_{12}$.

Table 13: Characteristics of SARIMA (0,1,1) $(0,1,3)_{12}$ Model

Model	Coeff.	Std. Err.	P-value	95% confidence interval	
SARIMA (0,1,1) $(0,1,3)_{12}$					
Constant	-0.346	1.609	0.830	-3.50	2.81
ARIMA					
MA (1)	-0.998	0.604	0.098	-2.18	0.19
SARIMA					
MA $(1)_{12}$	-0.032	0.042	0.937	-0.09	0.08
MA $(2)_{12}$	0.001	0.041	0.977	-0.08	0.08
MA $(3)_{12}$	0.001	0.065	0.993	-0.13	0.13
Sigma SQ	624.9				
Log likelihood	-881.3				
AIC	1774.6				
BIC	1790.9				

Source: Researcher, 2024

All the non-seasonal and seasonal MA models, as well as the constant term did not show significance. This model is therefore not statistically suitable to forecast TB disease in Ghana.

Table 14 provides a summary of all the ten models. The model that passed the estimation criteria was chosen as the most appropriate model for the TB disease in Ghana. Such a model was subjected to diagnostic test to ensure that it is fit for purpose.

Table 14: Summary of Characteristics of the Candidate Models

Model	Sig. Coef.	Sigma SQ	Likelihood	AIC	BIC
A:SARIMA $(1,1,3)(3,1,3)_{12}$	9/10	304.24	-820.02	1656.88	1678.56
B: SARIMA $(0,1,1)(0,1,3)_{12}$	0/4	624.852	-881.315	1774.63	1790.89
C: SARIMA $(0,1,3)(0,1,3)_{12}$	3/6	317.853	-820.442	1658.05	1682.44
D: SARIMA $(3,1,0)(3,1,0)_{12}$	3/6	565.310	-869.847	1755.69	1777.37
E: SARIMA $(0,1,1)(0,1,1)_{12}$	0/2	625.423	-881.332	1770.66	1781.50
F: SARIMA $(1,1,0)\ (1,1,0)_{12}$	1/2	774.241	-899.09	1806.17	1817.01
G:SARIMA $(0,1,2)(0,1,2)_{12}$	1/4	374.988	-833.636	1679.27	1695.53
H:SARIMA $(2,1,0)(2,1,0)_{12}$	2/4	635.41	-880.17	1772.3	1788.6

Source: Researcher, 2024

Table 14 cont'd: Summary of Characteristics of the Candidate Models

Model	Sig. Coef.	Sigma SQ	Likelihood	AIC	BIC
I:SARIMA $(1,1,0)(3,1,0)_{12}$	1/4	565.323	-869.847	1810.17	1826.43
J:SARIMA $(3,1,0)(1,1,0)_{12}$	3/4	565.323	-869.847	1751.69	1767.95
BEST MODEL	A	A	A	A	A

Source: Researcher, 2024

Each model was mapped with its characteristics as shown in the previous tables, and in accordance with the criteria discussed earlier. The table revealed that the most suitable model was SARIMA (1,1,3) $(3,1,3)_{12}$. This is because it satisfied all the 5 criteria. The non-seasonal moving average (MA) component has coefficients -2.413, 1.842 and -0.427. The seasonal autoregressive component has coefficients -0.037, 0.795 and -0.250. The seasonal moving average component has coefficients 0.305, -0.794 and 0.250. The model equation for the SARIMA (1,1,3) $(3,1,3)_{12}$, which has been identified as appropriate model for forecasting TB disease in Ghana is stated as follows:

$$X_t = -0.307X_{t-12} + 0.795X_{t-24} - 0.250X_{t-36} - 2.413W_{t-1} + 1.842W_{t-2}$$

$$-0.427W_{t-3} + 0.305W_{t-12} - 0.794W_{t-24} + 0.250W_{t-36} + \text{Ꞷ}_t$$

Diagnostic Test of SARIMA (1,1,3) $(3,1,3)_{12}$ Model

The next step was to put the selected candidate model, SARIMA (1,1,3) $(3,1,3)_{12}$, through a diagnostic test to determine whether it qualified as a stable univariate process. The following requirements should be met:

- The residual plot should be around the mean.
- The Portmanteau test is used to demonstrate that the model's residuals are white noise, or independently distributed.

  H_0: Residuals are White Noise

  H_1: Residuals are not White Noise
- The predicted ARIMA process need to be invertible and covariant stationary, resulting in an autocorrelogram that is flat.

Table 15 shows "Descriptive Statistics" of SARIMA (1,1,3) $(3,1,3)_{12}$.

Table 15: Descriptive Statistics of Residuals of SARIMA (1,1,3) $(3,1,3)_{12}$

Variable	Mean	Std. dev.	Min	Max
Error	2.968	266.500	-1542	1734

Source: Researcher, 2024

Figure 8 is a plot of residuals to ascertain whether or not a valuable information has been left in the model. If it is proven that some valuable information have been left out in the model, then the selected model cannot be used for forecasting. In this case, a different model is selected and then subjected to diagnostic analysis before it is accepted as a good model to forecast the disease.

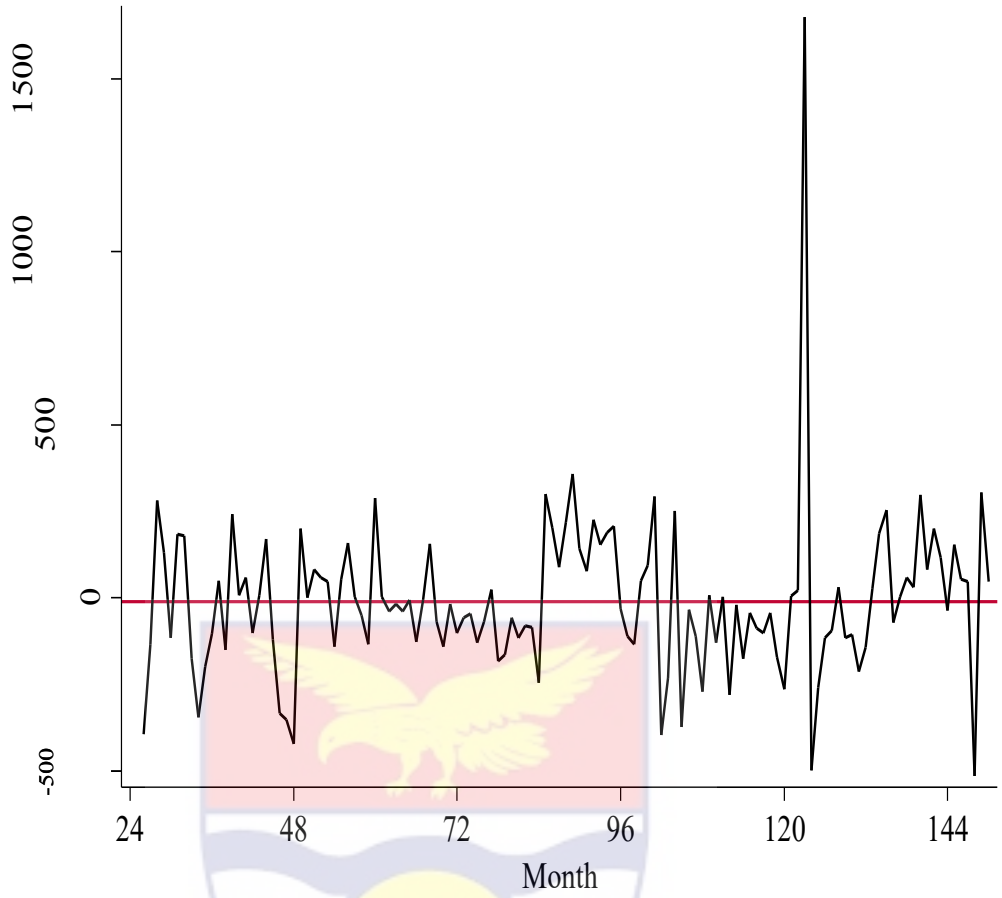


Figure 8: Plot of Residuals of SARIMA (1,1,3) $(3,1,3)_{12}$.

A clearer look at the plot indicates that the residuals revolve almost around zero. This is one of the criteria to indicate that no valuable information has been left unaccounted for, and that the selected model is fit for purpose. Autocorrelation plot of residuals and Ljung-Box Q test of the residuals will further confirm the suitability of the model to forecast.

Figure 9 is the autocorrelation plot of residuals to ascertain that the selected model is fit for purpose, and that it is most appropriate to forecast TB disease in Ghana.

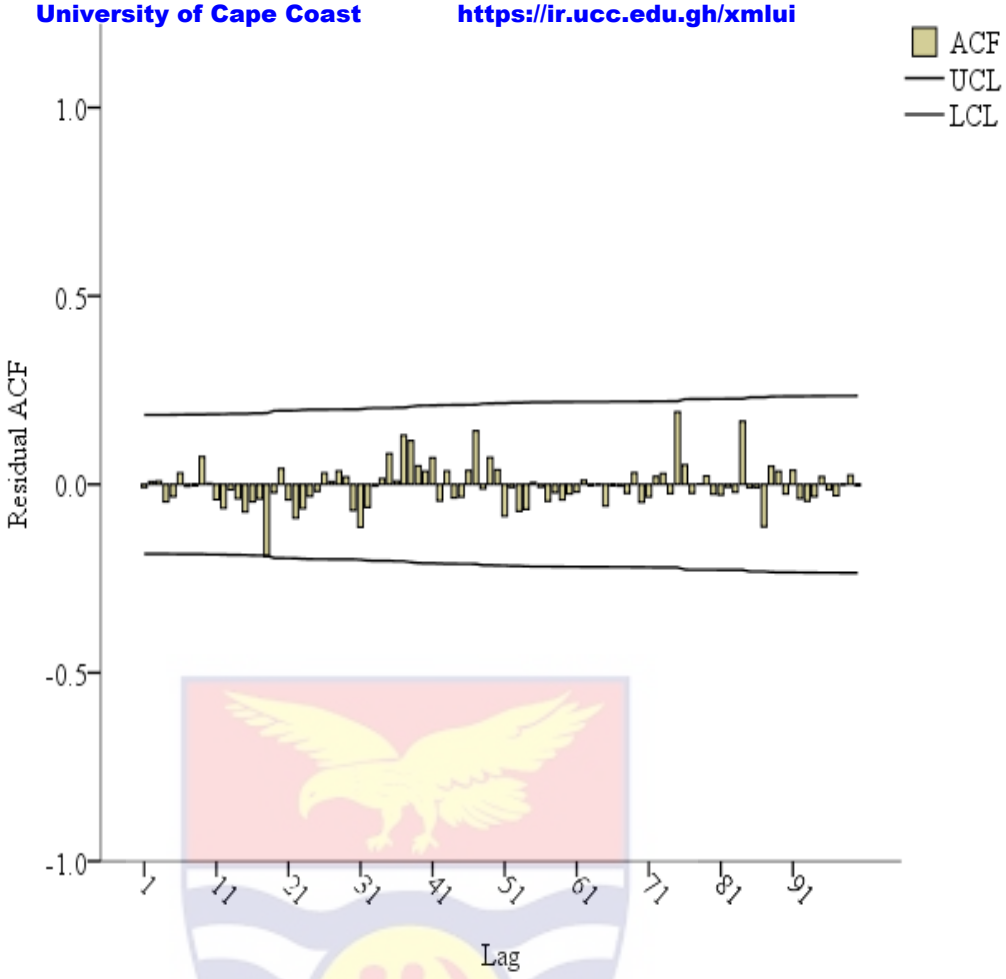


Figure 9: ACF Plot of Residuals of SARIMA (1,1,3) $(3,1,3)_{12}$.

It could be observed that the plot is flat and that all the autoregressive roots (non-seasonal and seasonal) are inside the 95% confidence band. This also means that the estimated model, SARIMA (1,1,3) $(3,1,3)_{12}$ is covariant stationary, hence the causality assumption is upheld.

Figure 10 is the partial autocorrelation plot of residuals to also ascertain that the selected model is fit for purpose.

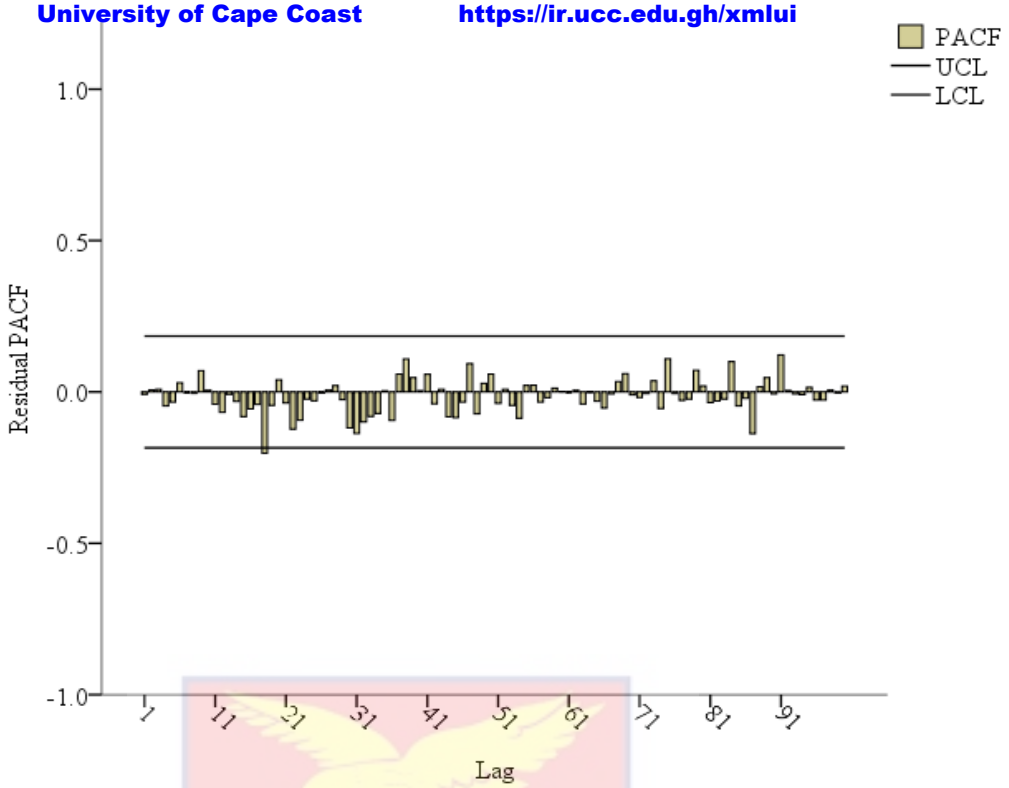


Figure 10: PACF Plot of Residuals of SARIMA (1,1,3) $(3,1,3)_{12}$.

The plot shows that all the moving average roots (non-seasonal and seasonal) are inside the 95% confidence band. This also means that the estimated model, SARIMA (1,1,3) $(3,1,3)_{12}$ is invertible, hence the invertibility assumption is upheld.

A more formal test to confirm the suitability of the model was conducted. Table 16 is result of Ljung-Box Q test which was constructed to ascertain the suitability of SARIMA (1,1,3) $(3,1,3)_{12}$ to forecast TB disease in Ghana.

Table 16: Ljung-Box Test of Residuals

Q test statistic	30.565
Prob > $chi^2(40)$	0.8589

Source: Researcher, 2024

The Ljung-Box test of residuals produced p-value greater than the 5% critical value. There is therefore no reasonable evidence to reject the null hypothesis, and conclude that residuals of the model are white noise.

It is important to note that all three diagnostic steps have shown that the selected model is good for forecasting Tuberculosis disease in Ghana. Normality and linearity are important assumptions in both univariate and multivariate analysis, and as such, they should be satisfied, according to (Hair, Black, Babin, and Anderson, 2010). To make sure the model SARIMA $(1,1,3)\ (3,1,3)_{12}$ is appropriate for forecasting, these normality tests were also conducted on it. Figure 11 shows Histogram plot of residuals. The purpose is to check if the residuals of the model is normally distributed, which is often a requirement for the validity of a statistical model (Hair et al., 2010). It reveals how the errors or deviations from the model's predictions are spread out across all observations. There is a limitation in this sense. For very small sample sizes, a histogram might not be the best way to judge the distribution (Hair et al., 2010). Because of that, the researcher also used the normal probability plot (normal P-P plot) to further check for the validity of the chosen model.

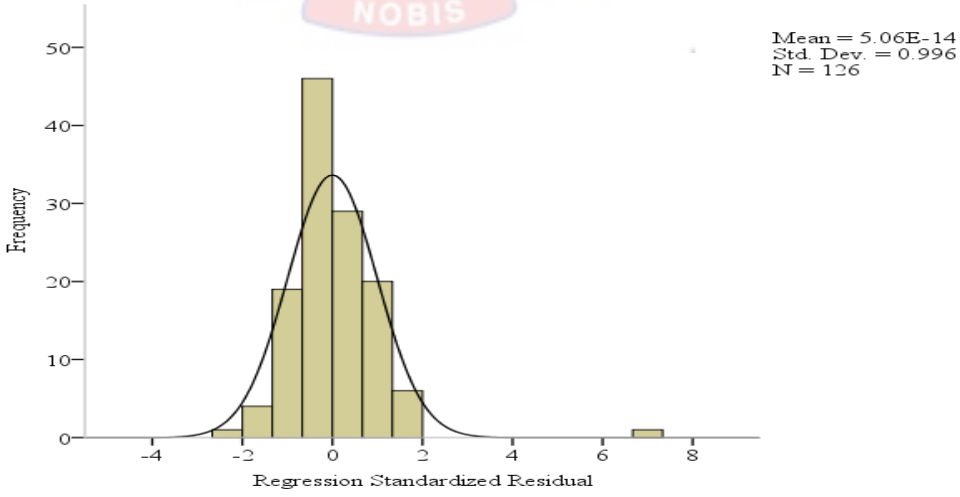


Figure 11: Histogram Plot of Residuals.

Normality can be confirmed using the "Normal P-P" plot and the normalized residual histogram plot. The fact that the bell-shaped symmetrical curve in figure 11 has lower values at the edges and higher scores in the middle of the histogram plot supports the normality assumptions. The few isolated bars are outliers in the data, which may not have strong effect on the validity of the model selected.

Figure 12 shows the Normal P-P plot to also test for the normality assumptions. This plot is to identify outliers, skewness, kurtosis, and the need for transformations. It is basically a scatter plot that shows the relationship between a data value and its precicted z-score (Hair et al., 2010).

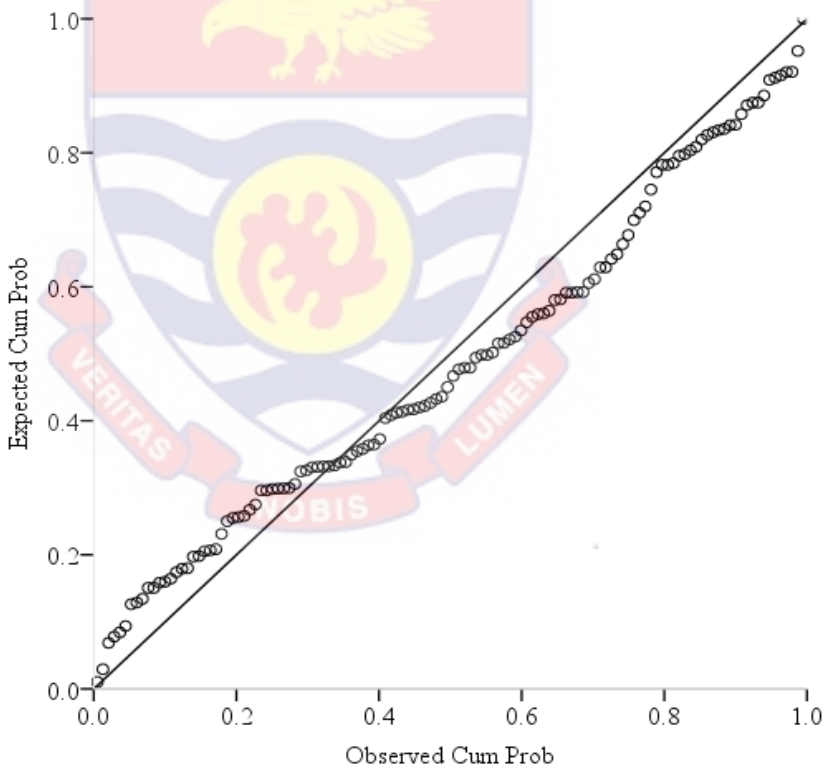


Figure 12: Normal P-P Plot.

The normal distribution is shown by the straight diagonal line of the plot, which contrasts with the standardized residual. With a few instances that nearly

deviate from the diagonal line, it is evident that the residuals are nearly along it. This suggests that the normality assumption is not broken. The normal P-P plot has shown normal distribution of the data. This suggests that the model's assumptions are met, and the model is performing well.

The scatter plot of residuals, which will also check for the normality assumptions, is shown in Figure 13. This will also help identify possible outliers, detect non-linear patterns and check for heteroscedasticity, by looking for a random distribution of points around zero. A random pattern of points around the horizontal line at zero suggests a good model fit, while any systematic pattern or shape indicates the model is not appropriate (Hair et al., 2010).

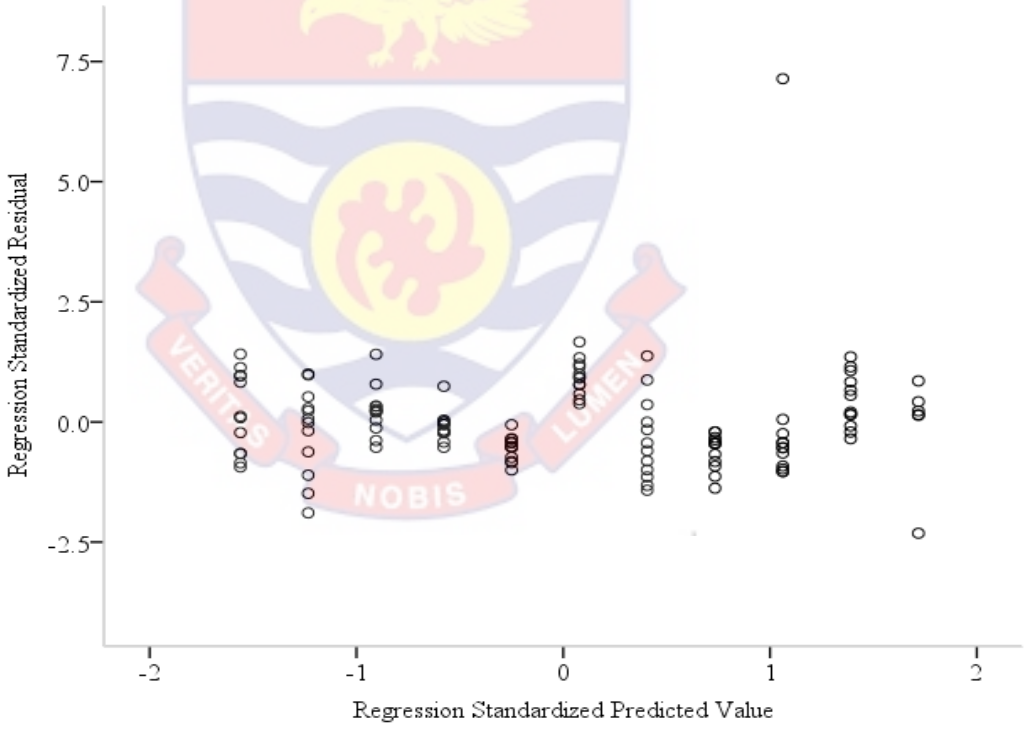


Figure 13: Scatter Plot of Residuals

The scatter plot in figure 13 also confirms the non-violation of the normality assumptions. Apart from few outliers, majority of the scores were concentrated at the center.

Table 17 presents "Test of Equality of Error Variances", which is another way of testing for normality. The "test of Equality of Error Variances" test hypothesis that:

H_0: The error variance of dependent variable is equal across groups

H_1: The error variance of dependent variable is not equal across groups

Table 17: Levene's Test of Equality of Error Variances

Number	Item	Value
1	F statistic	0.738
2	df 1	11
3	df 2	114
4	P-value	0.701

Source: Researcher, 2024

As shown in table 17, the p-value greater than the 5% critical value, suggests that equality of error variance condition has been satisfied and consequently satisfying the normality assumption.

Table 18 also presents the normality of residuals.to identify skewness and kurtosis in the model.

Table 18: Normality of the Distribution of Residuals

Descriptives	Statistic	Std. Error
Mean	0.0000	0.08508
95% Confidence Interval for mean lower bound	-0.1684	
95% Confidence Interval for mean upper bound	0.1684	
Median	-0.1087	
Variance	0.912	
Std. Deviation	0.95499	
Minimum	-2.02	
Maximum	4.40	
Range	6.42	
Interquartile Range	1.32	
Skewness	0.851	0.216
Kurtosis	2.390	0.428

Source: Researcher, 2024

As shown in table 18, the value for skewness and that for kurtosis are both less than 1 standard error, which suggested that the values are not relevantly different from the expected values of zero for normal distribution.

Figure 14 shows a “Normal Q-Q” plot of standardized residual for Tuberculosis to also test for the normality assumption of the model. This ia a plot of the quartiles of tha actual standardized residuals against the theoretical quantiles of a perfect normal distribution. If the data points form a straight line that aligns with the expected diagonal line, the normality assumption is met, indicating the

model's residuals are well-behaved and likely derived from a normal distribution. Deviations from this line suggests violations of the normality assumption (Hair et al., 2010).

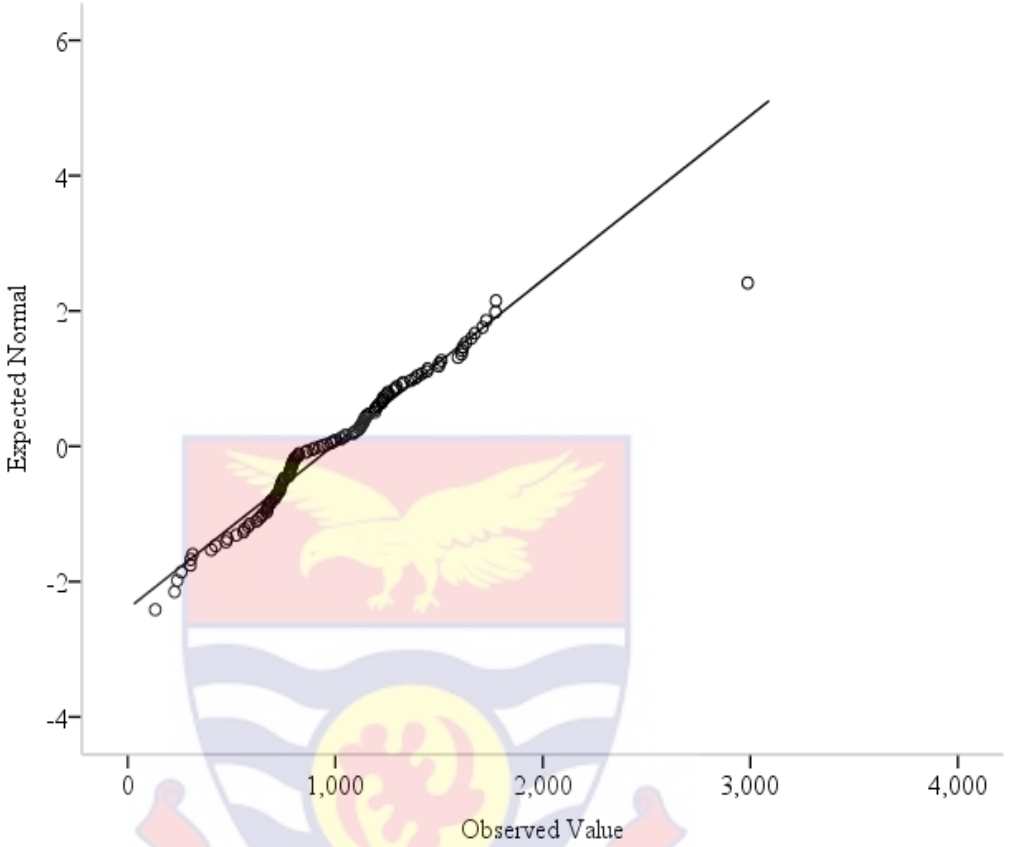


Figure 14: Normal Q-Q Plot.

The residuals on the Normal Q-Q plot revolve tightly around the normal distribution, which is shown by the straight diagonal line. Consequently, there is no violation of the normality criterion.

It was noted that the model for this project satisfied every normality assumption.

Forecasting

Figure 15 shows forecasting of Tuberculosis disease in Ghana from July 2024 to June 2026 based on the developed model, SARIMA $(1,1,3)\ (3,1,3)_{12}$.

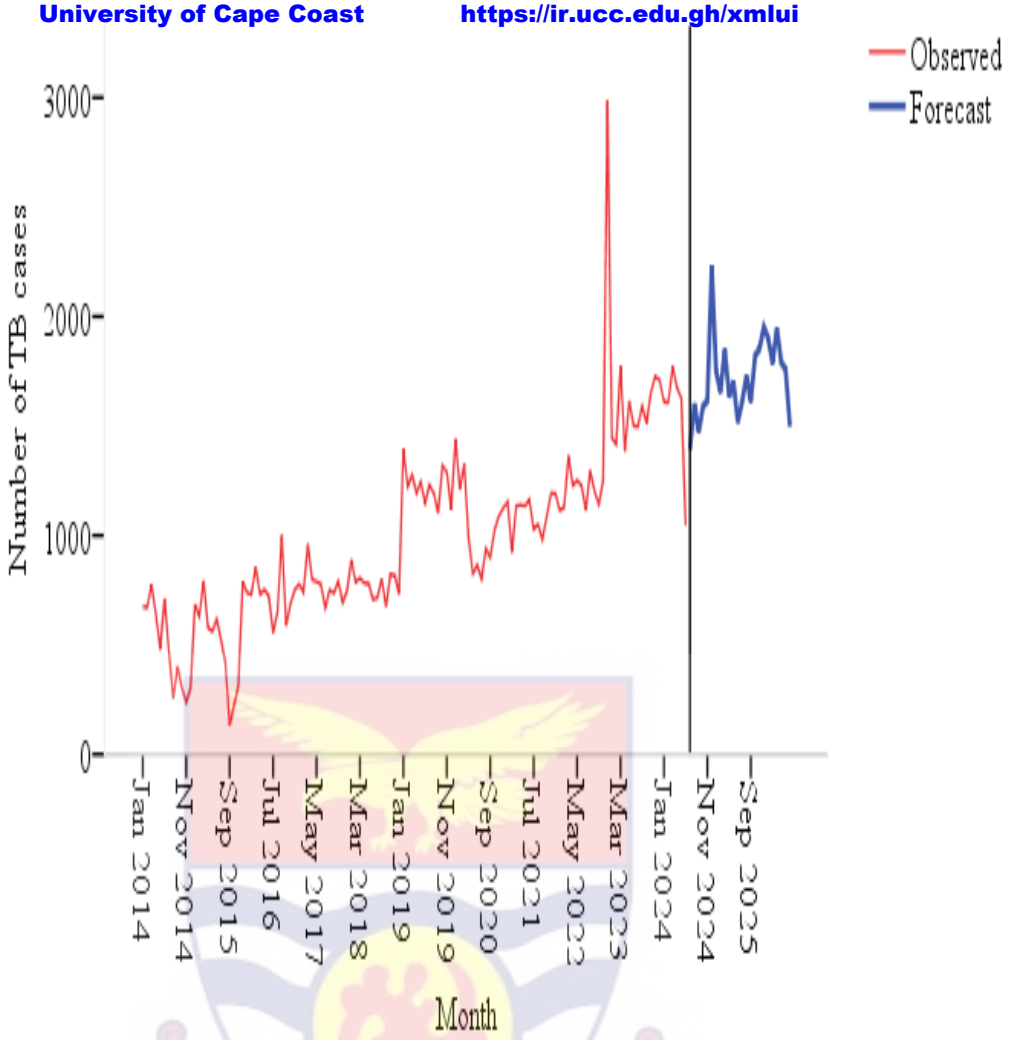


Figure 15: Forecasting of Tuberculosis Disease in Ghana.

According to the forecast, the disease peaks in December and declines in June each calendar year.

Table 19 shows forecasting accuracy of the models. Table 20 also shows forecasted values of Tuberculosis disease for the two-year period under study, from July, 2024 to June, 2026.

Table 19: Forecasting Accuracy of the Models

Model	R-squared	RMSE	MAPE	MAE
SARIMA (1,1,3) $(3,1,3)_{12}$	0.667	238.448	236.676	133.871
SARIMA (0,1,1) $(0,1,3)_{12}$	0.647	238.489	241.465	137.829
SARIMA (0,1,3) $(0,1,3)_{12}$	0.649	239.926	237.251	137.995
SARIMA (3,1,0) $(3,1,0)_{12}$	0.631	246.181	235.507	145.448
SARIMA (0,1,1) $(0,1,1)_{12}$	0.644	239.250	239.855	135.641
SARIMA (1,1,0) $(1,1,0)_{12}$	0.556	264.998	290.635	160.563
SARIMA (0,1,2) $(0,1,2)_{12}$	0.645	239.068	240.101	136.508
SARIMA (2,1,0) $(2,1,0)_{12}$	0.624	246.307	238.327	147.104
SARIMA (1,1,0) $(3,1,0)_{12}$	0.561	266.122	290.282	156.354
SARIMA (3,1,0) $(1,1,0)_{12}$	0.620	247.619	237.193	149.195

Source: Researcher, 2024

The developed model, SARIMA (1,1,3) $(3,1,3)_{12}$ exhibited the highest R-squared value and the lowest values for MAPE, MAE and RMSE according to the table. This implies that the disease was correctly predicted by the developed model.

Table 20 presented the TB forecast for the period of July, 2024 to June, 2026 in accordance with the model developed.

Table 20: Forecasted values of TB Disease from July 2024 to June 2026

Month	Point forecast	95% Confidence Interval	
		Lower bound	Upper bound
Jul 2024	1391	938	1844
Aug 2024	1602	1123	2082
Sept 2024	1471	979	1963
Oct 2024	1591	1075	2107
Nov 2024	1616	1084	2147
Dec 2024	2235	1684	2786
Jan 2025	1752	1185	2319
Feb 2025	1653	1068	2237
Mar 2025	1852	1252	2452
Apr 2025	1634	1018	2250
May 2025	1705	1074	2336
June 2025	1518	872	2164
Jul 2025	1610	926	2294
Aug 2025	1732	1027	2437
Sept 2025	1611	888	2333
Oct 2025	1823	1081	2565
Nov 2025	1852	1093	2612
Dec 2025	1953	1175	2730
Jan 2026	1902	1108	2696
Feb 2026	1785	974	2596

Source: Researcher, 2024

Table 20 cont'd: Forecasted values of TB Disease from July 2024 to June 2026

Month	Point forecast	95% Confidence Interval	
		Lower bound	Upper bound
Mar 2026	1949	1121	2776
Apr 2026	1790	946	2634
May 2026	1761	902	2621
June 2026	1498	622	2373

Source: Researcher, 2024

According to the model, the point forecast indicates that the TB disease peaks in December and declines in June in a calendar year, with the model equation;

$$X_t = -0.307X_{t-12} + 0.795X_{t-24} - 0.250X_{t-36} - 2.413W_{t-1} + 1.842W_{t-2} - 0.427W_{t-3} + 0.305W_{t-12} - 0.794W_{t-24} + 0.250W_{t-36} + \text{Ɛ}_t$$

Discussions

The study used the Box-Jenkins approach to model Tuberculosis disease in Ghana and forecasted the trend of the disease from July 2024 to June, 2026. A plot of the series was observed to be increasing with time and also showed seasonal variation. According to Box (1970), a series that exhibits large significant values at increasing lags and decays slowly is described as non-stationary series. There was clear evidence that the data was not stationary. The first difference and the first seasonal difference ensured a stationary series. There were fewer sample ACF and PACF values as lag increases, which decayed very fast, indicating a stationary series at this point. This was further supported by the Augmented Dickey-Fuller test, which resulted in a test statistic of -2.888 and a p-value of 0.000.

The ACF and PACF showed two components namely, seasonal component and non-seasonal component. Based on the ACF and PACF plots, 10 candidate models were identified. The models were SARIMA (3,1,0) $(1,1,0)_{12}$, SARIMA (1,1,0) $(3,1,0)_{12,}$ SARIMA (1,1,3) $(3,1,3)_{12,}$ SARIMA (2,1,0) $(2,1,0)_{12}$, SARIMA (0,1,2) $(0,1,2)_{12}$, SARIMA (1,1,0) $(1,1,0)_{12}$, SARIMA (0,1,1) $(0,1,1)_{12}$, SARIMA (3,1,0) $(3,1,0)_{12}$, SARIMA (0,1,3) $(0,1,3)_{12}$ and SARIMA (0,1,1) $(0,1,3)_{12}$. Finally, SARIMA (1,1,3) $(3,1,3)_{12}$ was selected as the most suitable model, after passing the estimation criteria. It was the model that had the highest number of significant coefficients, the lowest variability (given by the sigma squared value), the highest log likelihood statistic, lowest AIC" lowest BIC.

This selected model was taken through diagnostic test to ascertain its appropriateness to forecast the disease. The plot of residuals to ascertain whether or not there is valuable information left in the model was observed to be revolving around the mean of 2.968. The autocorrelation plot of residuals showed both AR roots and MA roots inside the 95% confidence band. This meant that the estimated model, SARIMA (1,1,3) $(3,1,3)_{12}$ was covariant stationary and invertible as well. Ljung-Box Q test resulted in a test statistic of 30.565 and probability value of 0.8589, confirming the suitability of the model. The suitability of the model was also ascertained through the Histogram of residuals, Normal P-P plot of regression standardized residuals, Scatter plot of residuals, Levene's Test of Equality of Error variances, Normality of distribution of residuals, and Normal Q-Q plot of standardized residuals.

The non-seasonal ARIMA model is the most widely used and well-liked technique for TB disease prediction in Ghana and other sub-Saharan African countries, per the literature search done for this study. SARIMA (1,1,3) $(3,1,3)_{12}$

model obtained in this study to forecast TB disease in Ghana is a deviation from findings of similar studies conducted in some parts of Ghana. In order to investigate the incidence of TB patients in the hospital's chest clinic, Aryee *et al*. (2018) used the Box-Jenkins ARIMA technique to examine monthly TB cases reported at the Korle-Bu Teaching Hospital (KBTH) in Ghana. They concluded that the appropriate model for the TB data was either ARMA (1, 1) or ARIMA (1, 0, 1), with no seasonal variation. Gyasi-Agyei & Obeng-Denteh (2014) modeled TB disease in Ghana's Ashanti Region using the total number of TB cases. The study could unfortunately not be able to uncover any seasonal changes. They concluded that the best model for the data in the region was either ARIMA (1,0) or AR (1).

In contrast, the same literature search found that the seasonal ARIMA model is the most effective and popular method for TB disease prediction in highly industrialized African countries as well as in Europe, America, and Asia. SARIMA $(1,1,3)\ (3,1,3)_{12}$ model obtained in this study to forecast TB disease in Ghana is consistent with findings of similar studies conducted elsewhere. According to a Portuguese study by Bras, Gomes, Filipe, Sousa, and Nunes (2014), the time series generally showed a seasonality and declining tendency. It was found that by accurately fitting and forecasting the time series, SARIMA models can predict trend and seasonal persistence. In the United States, tuberculosis is a disease that shows seasonal variation, peaks in June and declines in November, according to a study by (Willis, Winston, Heilig, Cain, and Walter, 2012). According to a different study by Manabe, Jin & Kudo (2019), the prevalence of tuberculosis (TB) in Japan varies by age and sex and peaks in the summer and fall. Wang, Tian, & Wang (2018) developed periodic ARIMA $(0,1,1)\ (0,1,1)_{12}$ with springtime peaks to represent TB. Azeez *et al*. (2016) conducted a study in South Africa that showed a seasonal shift,

indicating the Eastern Cape's recurring TB incidence. In Cape Town, South Africa, it was determined that the best model for tuberculosis disease was SARIMA (3,0,1) $(0,1,2)_{12}$.

Under the circumstances indicated above, developing countries in Africa and other regions have concluded on non-seasonal ARIMA models for tuberculosis, whereas highly industrialized countries in Europe, America, Asia, and Africa were better suited for seasonal ARIMA. Seasonal ARIMA (1,1,3) $(3,1,3)_{12}$ has been found to be the best model for TB disease in Ghana, with a peak in December and a drop in June. This study may be one of the first to discover seasonal patterns in TB data in Ghana, which is a low middle-income country in Africa. One possible reason is Ghana's weather conditions recently displaying summer and winter cycles. The southern hemisphere of Ghana has winter period from June to August and summer from December to February. Because of this, it was believed that tuberculosis disease, which is a droplet infection, could peak around December. Investigations into additional factors that may have contributed to this observation are necessary.

Chapter Summary

The Box-Jenkins technique was used to model the TB disease in Ghana, and the results were presented in this chapter. Analysis of monthly TB cases was conducted from January 2014 to June 2024. Using Stata, SPSS, and Excel software, the time series analysis was carried out. Forecasting was preceded by stationarity assessment, model identification, parameter estimation, and model diagnostic tests. The chapter also included a thorough analysis of the data and connected them to findings from other studies.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview

In this last chapter, the time series modeling of TB data reported in Ghanaian healthcare facilities was explained. It was acknowledged that tuberculosis disease exhibits a seasonal trend and that in order to prevent the disease's spread, control measures could be implemented during peak periods. An overview of Ghana's time series modeling of tuberculosis disease was given in this chapter. This topic was thoroughly discussed in the preceding chapters. Furthermore, the chapter concludes the thesis and offers some recommendations.

Summary

The purpose of this study was to use the Box-Jenkins technique to model TB disease in Ghana. The study's 126 data points came from monthly reports of tuberculosis cases in Ghanaian healthcare facilities between January 2014 and June 2024. The primary goal was to analyze the trend and seasonality of TB disease in Ghana and create a suitable model for its prediction. The TB time series' stationarity condition was examined. The disease was found to show seasonal variation and an upward trend. After taking the first difference and the first seasonal difference, the series became stationary.

Autocorrelation function and Partial autocorrelation function plots identified ten candidate models. The candidate models identified were SARIMA $(3,1,0)\ (1,1,0)_{12}$, SARIMA $(1,1,0)\ (3,1,0)_{12}$, SARIMA $(1,1,3)\ (3,1,3)_{12}$, SARIMA $(2,1,0)\ (2,1,0)_{12}$, SARIMA $(0,1,2)\ (0,1,2)_{12}$, SARIMA $(1,1,0)\ (1,1,0)_{12}$, SARIMA $(0,1,1)\ (0,1,1)_{12}$, SARIMA $(3,1,0)\ (3,1,0)_{12}$, SARIMA $(0,1,3)\ (0,1,3)_{12}$ and SARIMA $(0,1,1)\ (0,1,3)_{12}$. Parameters of all the candidate models were estimated

using the Maximum Likelihood Estimation (MLE) procedures. Five criteria were used to obtain the best model: the model that has the highest number of coefficients that are statistically significant, the lowest estimated error variance, the highest log likelihood statistic and the lowest AIC and BIC values. After going through the stated parameter estimation process, which is the second stage of the Box-Jenkins technique, SARIMA $(1,1,3)\ (3,1,3)_{12}$ was identified as the most suitable model for forecasting TB disease in Ghana. The selected model was passed through diagnostic test to ascertain its suitability to accurately forecast the disease. After the diagnostic analysis, it was observed that residuals of the series were white noise, the series was covariant stationary and invertible, and the residual plot also revolved aroud the mean of the series. With this, the model was declared fit for forecasting with the following model equation;

$$X_t = -0.307X_{t-12} + 0.795X_{t-24} - 0.250X_{t-36} - 2.413W_{t-1} + 1.842W_{t-2}$$

$$-0.427W_{t-3} + 0.305W_{t-12} - 0.794W_{t-24} + 0.250W_{t-36} + Ꞷ_t$$

Results of this study show seasonal variation of the disease. For this reason, control strategies of the disease should be enhanced during the peak periods.

Conclusions

The univariate data of monthly Tuberculosis cases in Ghana exhibited an upward trend with some seasonal variation. The 126 data points that were examined, which covered the period from January 2014 to June 2024, showed no signs of random fluctuations. The appropriate model for TB data in Ghana was SARIMA $(1,1,3)\ (3,1,3)_{12}$. Consequently, the model equation to forecast TB cases each month in Ghana was;

$$X_t = -0.307X_{t-12} + 0.795X_{t-24} - 0.250X_{t-36} - 2.413W_{t-1} + 1.842W_{t-2}$$

$$-0.427W_{t-3} + 0.305W_{t-12} - 0.794W_{t-24} + 0.250W_{t-36} + Ꞷ_t$$

This implies the non-seasonal moving average (MA) component has coefficients -2.413, 1.842 and -0.427. The seasonal autoregressive component has coefficients -0.037, 0.795 and -0.250. The seasonal moving average component has coefficients 0.305, -0.794 and 0.250, with no significant constant term. There was seasonal variation in the monthly forecasted values of TB from July 2024 to June 2026, with peak in December and a trough in June. This calls for further investigations to ascertain factors contributing to this seasonal pattern, so that control measures could be implemented effectively.

Recommendations

Given the disease's seasonal pattern, the Ministry of Health of Ghana should shift the celebration of world TB day celebration from March to December, when the disease is at its peak. This will enable interested parties to focus more attention on the reduction efforts.

Findings of this study is a call to action for the Ministry of Health of Ghana to provide world-class service to control the spread of the disease so as to achieve the sustainable development goal 3 of the United Nations by 2030. The "End TB" strategy, which is a component of the sustainable development goal 3, calls for member countries of the United Nations to achieve 90% reduction in TB-related mortality and 80% reduction in the incidence rate by 2030.

Findings of the study can be used for future investigations into the variables (factors) influencing the disease's seasonal pattern.

REFERENCES

Abu-Raddad, L. A., Lorenzo, S., Jerusha, T. A., Sugimoto, J. D., Longini, I. M., Christopher, & M. E. H. (2009). Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. *National Library of Medicine*, 106(33), 13980-5. https://doi.org/10.1073/pnas.0901720106.

Ade, S., Békou, W., Adjobimey, M., Adjibode, O., Ade, G. H. A. (2016). Tuberculosis case finding in Benin, 2000 – 2014 and beyond : a retrospective cohort and time series study. *Tuberc Res Treat*, 1, 25-48. https://doi.org/10.1155/2016/3205843.

Adèr, H. J., & Mellenbergh, G. J. (2008). *Advising on Research Methods: A consultant's companion* (2nd ed.). Netherlands: Johannes van Kessel Publishing, 271–304.

Adetunde, I. (2009). The mathematical models of the dynamical behaviour of Tuberculosis disease in the Upper East Region of the northern part of Ghana: A case study of Bawku. *ResearchGate*, *1*, 15–20.

Armstrong, J. S. (1978). *Long-range forecasting: From crystal ball to computer.* New Jersey, U.S.A. : John Wiley & Sons, 348.

Aryee, G., Kwarteng, E., Essuman, R., Nkansa-Agyei, A., Kudzawu, S., Djagbletey, R., Owusu Darkwa, E., & Forson, A. (2018). Estimating the incidence of tuberculosis cases reported at a tertiary hospital in Ghana: A time series model approach. *BMC Public Health*, 1, 1–8. https://doi.org/10.1186/s12889-018-6221-z.

Auld, S. C., Kasmar, A. G., Dowdy, D. W., Barun, M., Gandhi, N. R., Churchyard, G. J., Rustomjee, R. N. S. (2017). Research roadmap for tuberculosis transmission science: Where do we go from here and how will we know

when we are there? *The Journal of Infectious Diseases*, 6, 662– 668.

Azeez, A., Obaromi, D., Odeyemi, A., & Ndege, J. (2016). Seasonality and trend forecasting of Tuberculosis prevalence data in Eastern Cape , South Africa , using a hybrid model. *International journal of environmental research and public health*, 8, 71-757. https://doi.org/10.3390/ijerph13080757

Blankson, H. K. (2012). Economic burden of tuberculosis (TB) in the western region of Ghana. *Thesis*, 48-61.

Bonsu, F., A., Hanson-Nortey, N., Afutu F. K., Kulevome D. K., Dzata, F, A., Chimzizi, R., & Addo, K. O. (2020). *The National Tuberculosis Health Sector Strategic Plan for Ghana 2015–2020*. Accra, Ghana: Ministry of Health, 18-20.

Bonsu, F. A., Hanson-Nortey, N., & Ahiabu, M. A. A. (2017). Satisfaction of tuberculosis patients with health services in Ghana: Views of healthcare professionals. *International Journal of Health Care Quality Assurance*, 6, 545–553. https://doi.org/10.1108/IJHCQA-10-2016-0146.

Borgdorff, W. M., Sebek, M., Geskus B. R., Kremer, K., & Kalisvaart, N. (2011). The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. *National Library of Medicine*, 4, 964-70. https://doi.org/10.1093/ije/dyr058.

Box, G.E.P., & Jenkins, G.M. (1970). *Time series analysis, forecasting and control*. San Francisco, U.S.A: Holden-Day, 26, 118-415.

Bras, A. L., Gomes, D., Filipe, P. A., Sousa, B., & Nunes, C. (2014). Trends, seasonality and forecasts of pulmonary tuberculosis in Portugal. *The International Journal of Tuberculosis and Lung Disease*, 10, 1202-1210. https://doi.org/https://doi.org/10.5588/ijtld.14.0158

Brian Z. (2024). *What is statistical modeling*. California, U. S. A. : Coursera, 1. https://coursera.org/share/5d88ed590baf536c394701f56aa25353

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. New York, USA: Springer.

Cao, S., Wang, F., Tam, W., Tse, L. A., Kim, J. H., Liu, J., & Lu, Z. (2013). A hybrid seasonal prediction model for tuberculosis incidence in China. *BMC Medical Informatics and Decision Making*, 1. https://doi.org/10.1186/1472-6947-13-56

CDC (2015). *Health, United States, 2015*. Atlanta, U.S.A : DHHS Publication, 1, 19-461

Chatfield, C. (1996). Model uncertainty and forecast accuracy. *Journal of Forecasting*, *15*, 495–508.

Chowdhury, R., Mukherjee, A., Naska, S., Adhikary, M., & Lahiri, S. (2013). Seasonality of tuberculosis in rural West Bengal: A time series analysis. *International Journal of Health & Allied Sciences*, 2, 95. https://doi.org/10.4103/2278-344x.115684.

Cochrane, J. H. (1997). Time series for macroeconomics and finance. *Graduate School of Business, University of Chicago*, 773, 702-3059.

Cruz-Ferro, E. E. (2007). Epidemiology of tuberculosis in Galicia, Spain, 1996-2005. *The International Journal of Tuberculosis and Lung Disease*, 10, 1073-9.

Dodd, P. J., & Houben, R. M. (2016). The global burden of latent tuberculosis infection: A re-estimation using mathematical modelling. *National Library of Medicine*, 10, 18-35. https://doi.org/10.1371/journal.pmed.1002152.

Douglas, A. S., Strachan, D. P., & Maxwell, J. D. (1996). Seasonality of tuberculosis: The reverse of other respiratory diseases in the UK. *Scientific Research*, 9, 944–946. https://doi.org/10.1136/thx.51.9.944

Floyd, K., Glaziou, P., Korenromp, E. L., Sismanidis, C., Bierrenbach, A. L., Brian, G., Atunb, R., & Raviglione, M. (2015). Lives saved by tuberculosis control and prospects for achieving the 2015 global target for reducing tuberculosis mortality. *Bulletin of the World Health Organization*, 6, 12-22. https://doi.org/10.2471/BLT.11.087510.

Gyasi-Agyei, K., Gyasi-Agyei, A., & Obeng-Denteh, W. (2014). Mathematical modeling of the epidemiology of tuberculosis in the Ashanti Region of Ghana. *British Journal of Mathematics & Computer Science*, 3, 375–393. https://doi.org/10.9734/bjmcs/2014/5571.

Hafizi, A. H., Tafaj, S., Bardhi, D., Dilko, E. A. A. (2009). TB situation in Albania, 2001-2008. *Pneumologia*, 4, 104-7.

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate Data Analysis*. Upper Saddle River, U. S. A. : Prentice-Hall, Inc.

Hamzacebi, C. (2008). Improving artificial nural networks' performance in seasonal time series forecasting. *Information Sciences*, *178*, 4550–4559.

Hipel, K. W., & McLeod, A. I. (1994). *Time series modelling of water resources and environmental systems.* Amsterdam, Netherlands: Elsevier Science Publishers. https://doi.org/10.1016/0022-169495, 90010-1

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 4, 679–688.

Katherine, F., Philippe, G., & Alimuddin, Z. M. R. (2018). The global tuberculosis epidemic and progress in care, prevention, and research: An overview in

year 3 of the End TB era. *National Library of Medicine*, 5, 299–314. https://doi.org/10.1016/S2213-2600(18)30057-2.

Kumar, V., Singh, A., Adhikary, M., Daral, S., Khokhar, A., & Singh, S. (2014). Seasonality of tuberculosis in Delhi , India : A time series analysis. *National Library of Medicine*, 1, 1–5. https://doi.org/10.1155/2014/514093

Lin, Y. (2014). Seasonal dynamics of tuberculosis epidemics and implications for multidrug-resistant infection risk assessment. *National Library of Medicine*, 2, 358–370. https://doi.org/10.1017/S0950268813001040.

Liu, Q., Li, Z., Ji, Y., Martinez, L., Zia, U. H., Javaid, A., & Wang, J. (2019). Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu Province of China using advanced statistical time-series analyses. *National Center for Biotechnology Information*, 12, 2311–2322 https://doi.org/10.2147/IDR.S207809.

Luquero, F. J., Sanchez-Padilla, E., Simon-Soria, F., Eiros, J. M., & Golub, J. E. (2008). Trend and seasonality of tuberculosis in Spain, 1996-2004. *International Journal of Tuberculosis and Lung Disease*, 2:221-4.

Mahmood M., Khanjani, N., & Nasehi, M. (2015). Predicting the incidence of smear positive tuberculosis cases in Iran using time series analysis. *National Center for Biotechnology Information*, 11, 1526–1534.

Manabe, T. M., Jin, T., & Kudo, K. (2019). Seasonality of newly notified pulmonary tuberculosis in Japan, 2007–2015. *BMC Infectious Diseases*, 19, *497*.

McBryde, E. S., Debebe, S., Doan, T. N., Traucer J. M., & Denholm, J. T., (2021). Geospatial clustering and modelling provide policy guidance to distribute funding for active TB case finding in Ethiopia. *Science Direct*, 36, 104-214 *36*(https://doi.org/10.1016/j.epidem.2021.100470).

Olanrewaju, S. O., Ojo, E. O., & Oguntade, E. S. (2020). Time series analysis on reported cases of tuberculosis in Minna Niger State Nigeria. *Open Journal of Statistics*, 3, 412–430. https://doi.org/10.4236/ojs.2020.103027

Padberg, I., & Bätzing-Feigenbaum, J. S. D. (2015). Association of extra-pulmonary tuberculosis with age, sex and season differs depending on the affected organ. *National Library of Medicine*, 6, 723-8. https://doi.org/10.5588/ijtld.14.0735.

Park, H. (1999). A nural network ensemble method with jiltered training data for time series forecasting. *Information Sciences*, *180*, 329–546.

Park, H. (1999). Forecasting three-month treasury bills using ARIMA and GARCH models.. *International journal of information technology and management information systems*, 3, 1-7

SDG (2015). *Global Sustainable Development Report*. New York, U.S.A. : United Nations Publications.

Wang, H., Tian, C. W., Wang, W. M. (2018). Time-series analysis of tuberculosis from 2005 to 2017 in China. *Epidemiol Infect*, 46, 935–9.

WHO (2013). *Global tuberculosis report*. Geneva, Switzerland: Bulletin of the World Health Organization. https://www.who.int.

WHO (2015). *Global tuberculosis report*. Geneva, Switzerland: Bulletin of the World Health Organization. https://www.who.int.

WHO (2020). *WHO consolidated guidelines on tuberculosis*. Geneva, Switzerland: Bulletin of the World Health Organization. https://www.who.int.

WHO (2020). *WHO operational handbook on tuberculosis: module 4: treatment: drug-resistant tuberculosis treatment*. Geneva, Switzerland: Bulletin of the World Health Organization. https://www.who.int.

WHO (2021). *Global tuberculosis report*. Geneva, Switzerland: Bulletin of the World Health Organization. https://www.who.int.

WHO (2022a). *Global tuberculosis report*. Geneva, Switzerland: Bulletin of the World Health Organization. https://www.who.int.

WHO (2022b). *Global tuberculosis report*. Geneva, Switzerland: Bulletin of the World Health Organization. https://www.who.int/teams/global-tuberculosis-programme/tb-reports

WHO (2023). *Global Tuberculosis Report*. Geneva, Switzerland: Bulletin of the World Health Organization. https://www.who.int/teams/global-tuberculosis-programme/tb-reports.

Willis, M. D., Winston, C. A., Heilig, C. M., Cain, K. P., Walter, N. D. (2012). Seasonality of tuberculosis in the United States , 1993 – 2008. *Clin Infect Dis.*, 11, 1553–60.

Xiaolin, W., Xiulei, Z., Jia, Y., John, W., Rachel, B., Guanyang, Z., Hongmei, Z., Fang, L., & Zhimin, L. B. C. Z. (2014). Changes in pulmonary tuberculosis prevalence: evidence from the 2010 population survey in a populous province of China. *BMC Infectious Diseases*, 1, 21.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175.

Zhang, G. P. (2007). A nural network ensemble method with jiltered training data for time series forecasting. *Information Sciences*, *177*, 5329–5346.

APPENDICES

APPENDIX A

MONTHLY TB CASES (JAN 2014 - JUN 2015)

Month	Total PTB	EPTB	PTB New	PTB Relapse
January 2014	675	44	631	60
February 2014	672	95	577	47
March 2014	776	65	711	49
April 2014	642	44	598	45
May 2014	480	37	443	46
June 2014	712	64	648	65
July 2014	473	26	447	23
August 2014	258	24	234	21
September 2014	402	24	378	40
October 2014	303	25	278	21
November 2014	238	11	227	17
December 2014	302	8	294	16
January 2015	685	60	613	35
February 2015	630	54	568	20
March 2015	794	63	712	30
April 2015	580	57	511	36
May 2015	562	46	500	32
June 2015	618	49	561	26

Source: Ministry of Health, Ghana TB Report (Jan, 2014 - Jun, 2015)

APPENDIX B

MONTHLY TB CASES (JUL 2015 - DEC 2016)

Month	Total PTB	EPTB	PTB New	PTB Relapse
July 2015	523	58	456	21
August 2015	423	51	369	37
September 2015	132	30	101	8
October 2015	225	8	213	12
November 2015	312	23	281	19
December 2015	789	45	728	32
January 2016	739	52	677	47
February 2016	731	63	655	24
March 2016	858	125	713	47
April 2016	731	89	634	48
May 2016	753	105	631	32
June 2016	724	106	611	42
July 2016	557	70	478	29
August 2016	649	104	539	33
September 2016	1,000	199	757	30
October 2016	589	75	500	28
November 2016	687	95	581	34
December 2016	753	84	664	47

Source: Ministry of Health, Ghana TB Report (Jul, 2015 - Dec, 2016)

APPENDIX C

MONTHLY TB CASES (JAN 2017 - JUN 2018)

Month	Total PTB	EPTB	PTB New	PTB Relapse
January 2017	778	90	674	42
February 2017	740	102	625	45
March 2017	960	88	855	49
April 2017	799	103	667	37
May 2017	789	92	678	34
June 2017	778	109	644	43
July 2017	669	104	551	32
August 2017	751	113	623	39
September 2017	737	91	629	32
October 2017	791	80	705	45
November 2017	694	103	583	26
December 2017	745	68	665	43
January 2018	889	98	773	31
February 2018	786	84	692	46
March 2018	807	126	650	50
April 2018	782	107	661	33
May 2018	781	97	651	24
June 2018	708	102	591	37

Source: Ministry of Health, Ghana TB Report (Jan, 2017- June, 2018)

APPENDIX D

MONTHLY TB CASES (JUL 2018 - DEC 2019)

Month	Total PTB	EPTB	PTB New	PTB Relapse
July 2018	715	100	600	46
August 2018	802	91	698	48
September 2018	673	75	585	31
October 2018	823	90	710	37
November 2018	819	97	705	30
December 2018	734	101	616	30
January 2019	1,396	95	886	38
February 2019	1,224	91	758	33
March 2019	1,278	105	779	37
April 2019	1,193	104	760	27
May 2019	1,244	87	755	44
June 2019	1,148	92	669	28
July 2019	1,230	111	715	44
August 2019	1,193	92	729	31
September 2019	1,102	85	616	32
October 2019	1,320	113	764	40
November 2019	1,290	98	774	46
December 2019	1,120	80	692	34

Source: Ministry of Health, Ghana TB Report (Jul, 2018 – Dec, 2019)

APPENDIX E

MONTHLY TB CASES (JAN 2020 - JUN 2021)

Month	Total PTB	EPTB	PTB New	PTB Relapse
January 2020	1,442	97	866	49
February 2020	1,210	91	732	41
March 2020	1,327	93	794	42
April 2020	990	76	584	38
May 2020	825	58	486	33
June 2020	865	82	546	30
July 2020	802	68	533	23
August 2020	941	67	597	31
September 2020	899	67	604	25
October 2020	1,027	76	685	31
November 2020	1,090	66	719	46
December 2020	1,126	84	747	35
January 2021	1,154	82	711	45
February 2021	924	68	605	24
March 2021	1,136	94	775	29
April 2021	1,140	67	738	32
May 2021	1,135	54	778	45
June 2021	1,165	91	767	49

Source: Ministry of Health, Ghana TB Report (Jan, 2020- Jun, 2021)

APPENDIX F

MONTHLY TB CASES (JUL 2021 - DEC 2022)

Month	Total PTB	EPTB	PTB New	PTB Relapse
July 2021	1,029	56	693	31
August 2021	1,051	75	630	31
September 2021	980	71	589	29
October 2021	1,085	85	669	49
November 2021	1,193	74	826	35
December 2021	1,191	71	830	30
January 2022	1,115	54	725	46
February 2022	1,128	58	715	47
March 2022	1,364	68	886	40
April 2022	1,229	83	856	40
May 2022	1,252	63	833	46
June 2022	1,229	69	844	26
July 2022	1,115	70	714	33
August 2022	1,296	82	833	41
September 2022	1,203	53	859	40
October 2022	1,142	50	789	52
November 2022	1,253	54	865	38
December 2022	2,987	102	2,154	78

Source: Ministry of Health, Ghana TB Report (July, 2021- Dec, 2022)

APPENDIX G

MONTHLY TB CASES (JAN 2023 - JUN 2024)

Month	Total PTB	EPTB	PTB New	PTB Relapse
January 2023	1,445	62	992	60
February 2023	1,416	58	984	52
March 2023	1,774	89	1,172	74
April 2023	1,384	66	945	63
May 2023	1,615	83	1,099	50
June 2023	1,501	75	1,045	42
July 2023	1,498	66	1,035	54
August 2023	1,591	77	1,107	57
September 2023	1,511	62	1,034	42
October 2023	1,655	79	1,127	48
November 2023	1,727	70	1,189	77
December 2023	1,709	43	1,126	66
January 2024	1,611	68	1,085	47
February 2024	1,608	84	1,046	61
March 2024	1,772	87	1,196	54
April 2024	1,673	62	1,078	61
May 2024	1,629	87	1,069	69
June 2024	1,046	53	715	26

Source: Ministry of Health, Ghana TB Report (Jan, 2023- Jun, 2024)