A TIME SERIES ANALYSIS OF PATTERNS IN TB CASE NOTIFICATION IN NORTH HEALTH ZONE IN MALAWI

MASTER OF SCIENCE (BIOSTATISTICS) THESIS

MATHIAS DUNCAN NGWIRA

UNIVERSITY OF MALAWI

AUGUST, 2023

A TIME SERIES ANALYSIS OF PATTERNS IN TB CASE NOTIFICATION IN NORTH HEALTH ZONE IN MALAWI

MASTER OF SCIENCE (BIOSTATISTICS) THESIS

By

MATHIAS DUNCAN NGWIRA

BSc. (Statistics and Geography)-UNIMA

Thesis submitted to the Department of Mathematical Sciences, Faculty of Science, in Partial fulfilment of the requirements for the degree of Master of Science (Biostatistics)

University Of Malawi

August, 2023

DECLARATION

I, the undersigned hereby declare that this thesis is my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used, acknowledgments have been made.

MATHIAS DUNCAN NGWIRA

Full Legal Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and effort and has been submitted with our approval.

Signature: _____Date: _____

Jupiter Simbeye, PhD

Supervisor

DEDICATION

To my beloved wife Janet, for your support and encouragement; to my parents and siblings for your endless support.

ACKNOWLEDGEMENTS

First and more importantly, I thank the almighty God for his never-ending love and blessings in getting this work done. To him be the glory.

ABSTRACT

Seasonal Autoregressive Moving Average (SARIMA) models are an extension of ARIMA models that specifically address the presence of seasonality in time series data. By incorporating both non-seasonal and seasonal components, SARIMA models capture long-term trends, lagged values within seasons. The SARIMA model was applied to TB data obtained in the north health zone of Malawi from January 2013 to September, 2020. The Box-Jenkin seasonal ARIMA approach was used to identify the best model for forecasting. We used the **auto.arima** function in R to identify the best model to predict future trends in TB case notifications. The winter multiplicative method of exponential smoothing was used to forecast future trends in TB case notifications. For model selection, we used the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Quarterly TB case notifications were analyzed, stratifying the data by disease site, HIV status, sex, and patient age group. A cyclic pattern of TB case notifications was observed, with peaks during the rainy season and at the end of the cold season. The best model for predicting future trends in TB case notifications was determined to be SARIMA (0, 1, 2) $(1, 0, 0)_4$ (the lower AIC and BIC values, 240.81 and 246.41, respectively) Additionally, a higher proportion of TB incidence was found among males across all age groups. The study's findings indicate an increasing trend in predicted TB incidence in the near future, accompanied by a seasonal pattern. Forecasting of PTB incidence between the years 2021 and 2024 showed a slightly increasing trend. The implications of this study highlight the importance of health education, timely medical care seeking, and proactive service planning to accommodate higher service utilization during high TB risk periods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND ABBREVIATIONS

ACF	Autocorrelation Function
AIC	Akaike Information Criteria
AICc	Corrected Akaike Information Criteria
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BIC	Bayesian Information Criteria
CDR	Case Detection Rate
CI	Confidence Interval
DHS	Demographic Health Survey
DOTS	Directly Observed Treatment Short-course
GAMs	Generalized Additive Models
HIV	Human Immunodeficiency Virus
HMIS	Hospital Management Information System
SARIMA	Seasonal Autoregressive Integrated Moving Average
MA	Moving Average
MDR-TB	Multiple Drug Resistant Tuberculosis

MHMIS	Malawi Hospital Management Information System
MoH	Ministry of Health
NTP	National Tuberculosis Programme
PACF	Partial Autocorrelation Function
SA	Seasonal Amplitude
SDGs	Sustainable Development Goals
SP	Strategic Plan
TB	Tuberculosis
WHO	World Health Organisation

CHAPTER ONE

INTRODUCTION

Tuberculosis (TB) is a global public health concern. Surveillance programs present invaluable epidemiological information regarding their temporal evolution, particularly for pulmonary tuberculosis (PTB), the most common form of TB and the one that presents the greatest challenge in public health. Trends and seasonal variations have been demonstrated in a number of studies in different countries, with reported peaks in late winter and early summer or spring (Gashu, 2018).

According to Mohammad (2012), predictions of future events and conditions are called forecasts and the act of making such predictions is called forecasting. Forecasting is very important as predictions of future events must be incorporated into the decision-making process. In forecasting events that will occur in the future, information concerning events that have occurred in the past must be relied on.

In order to prepare forecasts, past data needs to be analysed to identify a pattern that can be used to describe them. Then, this pattern is extrapolated or extended into the future. This forecasting technique rests on the assumption that the pattern that has been identified will continue to make good predictions. If the data pattern that has been identified does not persist into the future, then this indicates that the forecasting technique used is likely to produce inaccurate predictions (Bowerman and O'Connel, 1993).

Most time-based forecasting problems involve the use of time series data. In this study, time series will be used to prepare forecasts. A time series is formed from measurements of a variable taken at regular intervals over time. It is a stochastic process that amounts to a sequence of random variables (Box & Jenkins, 1976). The TB case notifications fall under the category of time series. According to Box & Jenkins (1976) time series can be used in the

application of forecasting of future values of a time series from current and past values and could be used for forecasting TB case notifications.

1.1. Background of the Study

The study on modelling seasonal patterns in TB case notification aims to understand and predict the temporal variations in tuberculosis (TB) incidence throughout the year. TB is still one of the leading infections causing deaths, killing at least 2 million people every year (WHO, 2018). In 2018, an estimated 7.2 million people developed active TB, resulting in 1.2 million TB-related deaths (WHO, 2018). However, only 6.9 million cases were notified, indicating a gap in the number of cases that were not officially notified. TB remains a significant global health concern, and its transmission dynamics often exhibit seasonal patterns. These patterns can be influenced by various factors, such as climatic conditions, socioeconomic factors, healthcare access, and population mobility (Choi, Seo, Choi, Kim, & Youn, 2013). Understanding the seasonal variations in TB cases can help public health authorities develop targeted interventions, allocate resources effectively, and improve disease control strategies.

Previous research has suggested that TB incidence rates often fluctuate seasonally, although the specific patterns and underlying drivers can vary across different regions and populations. Several studies, conducted by Bras et al. (2014) and Zhang et al. (2020) in China, Moosazadeh & Amiresmaili. (2018) in Iran, (Kirolos, et al., 2021) in Malawi, and Gashu, (2018) in Ethiopia, have observed seasonal variation in TB case notification. Some of these studies have linked TB transmission to climatic factors, such as temperature and humidity (Gashu, 2018), Wubuli *et al.* (2017), Yang *et al.* (2014), while others have identified socioeconomic factors such as poverty Bohena, *et al* (2019), Mososazadeh et al. (2014), and human behaviour such as delay in diagnosis or delay in seeking health care (Fares 2011) as key contributors. Nyirenda (2006) attributed seasonal variation to overcrowding. The sex and age of an individual are also regarded as determinants of TB case notification Soetens *et al.* (2013). The present study aims to identify a suitable time series model to investigate seasonal patterns in TB notification and forecast future trends in TB case notification in the north health zone of Malawi. The findings will contribute to the development of more effective and targeted interventions to combat TB and reduce its burden on public health systems.

1.2. Problem Statement

While seasonal patterns in diseases such as malaria, influenza, and meningitis are well acknowledged in Malawi, this remains subtle for diseases such as TB. Seasonal patterns in TB case notification have been documented in other countries in Asia, Europe, and other parts of Africa, e.g., Ethiopia and Nigeria (Roderick, 2016). The patterns of seasonal peaks and troughs in TB numbers reported in such studies appear to vary by country and hemisphere. The reasons for such variations are currently not well understood, and it is likely that there are several interrelated factors. Seasonal variations affect the health system's functioning, including TB services, but there is little evidence about seasonal variation in TB case notification in tropical countries, including Malawi. Understanding the epidemiology of TB in the country in terms of seasonal variation would make significant contributions to designing high-yield case-finding strategies. A few studies done in Ethiopia (Gashu, 2018), Zimbabwe (Takarinda, Harries, & Mutasa-Appolo, 2020), the Republic of South Africa, Morocco (Ottmani, Obermeyer, Bencheikh, & Mahjour, 2021), and Asia (Zhang, et al., 2020) have tried to indicate the seasonal variation in TB, but their findings are inconsistent and limited in scope. This study sought to fill this knowledge gap using TB data reported in the health zone under study.

Analysis of routinely collected Hospital Management Information System (HMIS) data in the context of a TB disease programme, involves an investigation of changes in rates over time, followed by attempts to understand their underlying causes (WHO, 2018). Proper understanding and possible prediction of patterns in TB case notification would aid in focused programming of TB control and prevention. There is rich data that is currently routinely collected within the health system that is underutilized. This study, therefore, will focus on modelling the available data in order to come up with the best time-series model to investigate seasonal patterns in TB case notification and, in turn, use the developed model to predict future seasonal patterns in TB case notification.

Predicting the incidence of TB plays an important role in planning health control strategies for the future, developing intervention programs and allocating resources where they are needed most.

1.3. Study Objectives

The purpose of this study was to predict future trends in the incidence of TB in the north health zone of Malawi. To meet this purpose, the following were the specific objectives of this study;

1.3.1. To identify a suitable time series model to investigate seasonal patterns in TB case notification in the north health zone in Malawi.

1.3.2. To use the identified time series model to predict future seasonal patterns in TB case notification.

1.3.3. To suggest potential causes for the identified patterns and also suggest what interventions could be put in place in view of this.

1.4. Research Questions

The main interest in conducting this study is in specific questions that address specific objectives of this study. In order to answer the research objectives, we define the following research questions;

1.4.1 What is a suitable time series model that can investigate seasonal patterns in TB case notification in the north health zone of Malawi?

1.4.2 From the identified time series model, what are the predicted future seasonal patterns in TB case notification?

1.4.3 What are the potential causes of the identified patterns, and what interventions could be put in place in view of this?

1.5. Hypothesis

The following are the null and alternative hypotheses for the study;

1.5.1. Null Hypothesis

Tb case notification data show seasonality

1.5.2. Alternative Hypothesis

TB case notifications do not show a seasonal pattern.

1.6. Significance of the Study

Time-series analysis is widely employed in public health research to better describe data and/or make inferences that take into consideration the correlation between time-adjacent observations. TB research is no exception. The findings from this study will add to the body of knowledge about the seasonality of TB case notification and other factors affecting trends in TB incidence. According to Gashu (2018), knowledge about seasonality and other factors affecting trends in TB incidence will help in predicting future TB incidence epidemics and hence help in planning for service requirements, assessing health needs, and manage the disease by using the predictions as reference information.

The study has also suggested the potential causes or risk factors associated with the identified pattern and, hence, suggested interventions that could be put in place as a means to combat the spread of the disease in the study area.

CHAPTER TWO

LITERATURE REVIEW

This chapter provides a summary of important determinants of patterns in TB case notification. The chapter further reviews and provides a critique of previous studies that have been conducted with the aim of modelling seasonal patterns in TB case notification.

2.1 Tuberculosis Situation Analysis in Malawi

TB remains a major public health problem in Malawi and is among the top ten killer disease in the country. There have been good achievements over the past two decades in tuberculosis control through DOTS strategy. However, the results of the TB prevalence survey conducted in 2014 showed that the TB disease burden in Malawi is significantly higher than initially estimated by the WHO. The TB prevalence among urban populations was more than double the national average. In 2018, the WHO estimated that 181 new and relapse TB cases occurred per 100, 000 population. This translated to about 30, 000 new and relapse cases of TB occurring in 2018 (Global TB Report 2019). During the same year, nearly 16, 000 new and relapse TB cases were reported to the National TB control Program representing about 53% of incidence cases.

Provisional results from the National TB Prevalence Survey completed in 2014 showed a higher TB prevalence of 1014/100,000 compared to the previous estimated prevalence of 373/100,000 by the WHO. According to Nyirenda (2020), in 2014, a total of 17,723 new and relapse TB cases were identified, a decline from 19,539 reported in 2013. Treatment success rate for smear positive case for evaluated 2013 cohort was at 86%. It was further observed that MDR-TB was an emerging issue in Malawi, with a prevalence of 0.4 percent among new and 4.8 percent among previously treated TB patient populations, respectively.

Zumla, Chakaya, Centis, Mwaba, & D'Ambrosio (2015) further reported that HIV remains an important risk factor for developing active TB disease in Malawi: 52 percent of people with TB are also infected with HIV. Ninety five percent of registered TB patients know their status, and 92 percent of those infected are on antiretroviral therapy during the period of their TB treatment.

For years, the NTP has collected data such as sex and age for smear positive TB cases only because of their importance to public health as the source of the majority of TB infections in the community. However, from 2002 onwards, such information started being collected on patients with other forms of TB. The study done by Nyirenda, (2006) found that the attack rates (new cases per 100,000 population) were highest in people between the ages of 25 and 44 years. The age group of 25 – 34 contributed about 40% of all smear positive TB cases while 20% of the cases were from 15 – 24 and 35 – 44 age groups. Thus, accounting for 80% of all cases to be between the ages of 15 and 44 years. Between the ages of 0 and 24 there are more females with smear positive TB than males, the distribution being equal in the ages of 25 and 34 and more men than women after the age of 34. The same distribution followed the same pattern as the HIV seroprevalence among the Malawian population.

In general, the ratio of men to women among TB patients in Malawi from NTP data is 1.1. This implies that there are no significant gender differences among TB patients (Boeree, Harries, & Godschalk, 2000). A similar study done in northern Malawi has shown that in the HIV era, the ratio has decreased from 1.3 to 0.8. The spatial distribution of TB cases in Malawi depends on the size of each catchment population. The distribution of diagnostic services and differences in health seeking behaviour among different populations may also be contributing factors. In general, the southern region districts of Malawi contribute about 60% of all known TB cases in the country.

2.2 Modelling Approach and Evaluation

Time series modelling was performed to investigate seasonal pattern in TB case notification in the health zone under study and hence predict future seasonal patterns in TB case notification using key stages specified in the preceding section. When investigating a time series, one of

the first things to check before building an ARIMA model is that the series is stationary. That is, it needs to be determined that the time series is constant in its mean and variance and that the mean and variance are not dependent on time. Below are the necessities of the assumption of stationarity;

i) Stationarity means that the statistical properties of a time series (or rather, the process generating it) do not change over time.

ii) Stationarity is important because many useful analytical tools, and statistical tests, and models rely on it.

iii) Standard techniques are largely invalid where the data is non-stationary.

iv) Sometimes autocorrelation may result because the time series are non-stationary.

v) Non-stationary time series regressions may also result in spurious regression, i.e. cases where the regression equation show significant relationship between two variables when there should not be any such relationship.

As such, the ability to determine if a time series is stationary is important. Rather than deciding between two strict options, this usually means being able to ascertain, with high probability, that a series is generated by a stationary process. Furthermore, the Kolmogorov-Smirnov test was employed to examine the normality of the data.

In this study, we looked at a couple of methods for checking the stationarity of the time series. This was done so that if the time series is provided with seasonality, trend, or a change point in mean or variance, then the influences need to be removed or accounted for. The first method that was applied was the auto-correlation function (ACF) and partial auto-correlation function (PACF) plots. ACF tells us how correlated points are with each other based on how many steps they are separated by. It is used to determine how past and future data points are related in a time series. Its values range from -1 to 1. When the ACF plot crosses the blue dashed line, this means that the values are correlated, hence non-stationary. For a stationary signal, because we expect no dependence with time, we would expect the ACF to go to 0 for each time lag. PACF, on the other hand, could be defined as the degree of association between two variables while

adjusting the effects of one or more additional variables. PACF is used to find the correlation of the residuals (which remain after removing the effects that are already explained by the earlier lag(s)) with the next lag value, hence 'partial' and not 'complete' as we removed already found variations before we find the next correlation. So, if there is any hidden information in the residuals that can be modelled by the next lag, we might set a good correlation, and we will keep that next lag as a feature while modelling. Keeping in mind that while modelling we do not want to keep too many features that are correlated, as that can create multicollinearity issues. Hence, there is a need to retain only the relevant features.

The Augmented Dickey-Fuller (ADF) Test is another method that was used to test for stationarity. This is another method used to determine more objectively if the data is stationary or not. According to Fuller (1976), an ADF tests the null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test is used, but it is usually stationary or trend-seasonality. ADF is an augmented version of the Dickey-Fuller test for a larger and more complicated set of time series models. The ADF test statistic, used in the test, is a negative number. The more negative the number, the stronger the rejection of the null hypothesis. Since the null hypothesis assumes the presence of a unit root, the P-value obtained by the test should be less than the significance level (usually 0.05) to reject the null hypothesis.

The raw original data was plotted to check the time series pattern. After the pattern for the time series data was recognised, proper models were fitted to the data. The models were then compared for best fit to the data. The AIC and BIC values were used to choose the best fitted model among the fitted models. As a general rule, the model with the smallest values of AIC or BIC was chosen to be the best model (Koehler & Murphree, 2008).

2.3 Time-Series Approaches and Models to be Considered in this Study

In this study, the open-source statistical analysis package R (The R Foundation for Statistical Computing, version 5.0.2, 2019) was utilized for estimating the models. The following provides an overview of the time-series approaches and models that will be considered for modelling seasonal patterns in TB case notifications, along with their corresponding estimation strategies:

2.3.1 Autoregressive Models

An autoregressive model is when a value from a time series is regressed on previous values from that same time series. In this model, we forecast the variables of interest using a linear combination of past values of the variables. Thus, an autoregressive model of order p can be written as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} \ldots + \beta_p Y_{t-p} + \epsilon_t$$

where ϵ_t is a white noise. A series is called white noise if it is purely random in nature hence it has zero mean and a constant variance. The scatter plot for such a series across time indicates no pattern and hence forecasting the future values of such a series is not possible.

An AR model is like a multiple regression but with lagged values of Y_t as predictors. This kind of model is referred to as an $AR_{(P)}$ model, an autoregressive model of order p. With this kind of model, the assumption is that the past values have an effect on the current values.

2.3.2 The Moving Average (MA) Model

In time series analysis, the moving average model (MA model), also known as moving-average process, specifies that the output variable depends linearly on the current and various past values of a stochastic (imperfectly predictable) term. It is a most common approach for modelling univariate time-series.

A common representation of a moving average model where it depends on q of its past value is called MA (q) model and is represented as:

$$Y_t = \beta_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} \ldots + \phi_q \varepsilon_{t-q}$$

Where ε_t are the error terms and are assumed to be white noise processes with mean zero and a constant variance.

2.3.3 Autoregressive Moving Average (ARMA) Model

These are a kind of model where the time-series may be represented as a mix of both AR and MA models referred as $ARMA_{pq}$. The general form for such a time-series model, which depends on p number of parameters of its own past values and q past values of white noise disturbances, takes the following form;

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} \ldots + \phi_q \varepsilon_{t-q}$$

where ϵ_t is a white noise.

2.3.4 Autoregressive Integrated Moving Average (ARIMA) Model

One of the commonly used prediction models is the ARIMA model, which is a time series analysis tool proposed by George Box and Gwilym Jenkins in the 1970s (Box & Jenkins, 1970). The ARIMA model regards the data sequence formed by the prediction object over time as a random sequence. This model is easy to construct, only requires intrinsic variables, and has relatively high prediction accuracy. The ARIMA model has been widely used in the prediction of such diseases such as malaria (Anwar M. Y., Lewnard, Parikh, & Pitzer, 2016), influenza (He & Tao, 2018), homorrhagic fever (Li, Guo, & Han, 2012) or hand, foot and mouth disease (Liu, Luan, Yin, Zhu, & Lu, 2016). The ARIMA model is one of the most classical methods of time series analysis which was first proposed by Box-Jenkins in 1976 (Lin, Ezzati, Chang, & Murray, 2009). It is represented as a Moving Average (MA) model combined with an AR model to fit the temporal dependence structure of a time series using the shift and lag of historical information. According to Box, (2015) ARIMA models consist of three sections in the order of auto-regression (p), the degree of difference (d) and the order of moving average (q). In epidemiological and many other studies, this model has widely been used to predict the incidence of infectious diseases such as dengue fever, influenza, hepatitis, etc. In practice many time-series are non-stationary and so one cannot apply stationary AR, MA or ARMA processes directly. Before constructing the ARIMA model, one firstly needs to identify the stationarity state of the observed data in the series, of which the mean value remains constant. If the observed data is not stationary, it is then transformed into a stationary time-series by taking a suitable difference. If the original data series is differenced *d* times before fitting an ARMA (*p, q*) process, then the model for the original undifferenced series is said to be an ARIMA (*p, d,*

q) process where the letter 'I' in the acronym stands for *integrated* and d denotes the number of differences taken.

2.3.5 Development of the SARIMA Model

The seasonal ARIMA model (SARIMA) is an expanded form of ARIMA, which allows for seasonal factors to be reflected Bras, Gomevaluationes, Filipe, de Sousa, & Nunes (2014). Time series seasonality is an unvarying pattern that recurs over S period of time until the pattern changes over again. The SARIMA model integrated both non-seasonality and seasonality factors in a generative model. In the SARIMA model, seasonality in AR and MA terms predict Y_t using data values and errors at time interval that are multiples of S (Moghram & Rahman, 1989). The SARIMA model is given by:

$$SARIMA(p,d,q) \times (P,D,Q)^S$$

Where $p = AR$ order in non- seasonality, d = difference in non-seasonality, $q = MA$ order in non- seasonality,$P = AR$ order in seasonality, D = difference in seasonality, Q= MA order in seasonality, and S = recurrence of time periods in the seasonality pattern. The general SARIMA model has the following form

$$\Phi(\beta^S)\varphi(\beta)(Y_t - \mu) = \Theta(\beta^S)\theta(\beta)\varepsilon_t$$

The non-seasonality components are;

$$AR: \varphi(\beta) = 1 - \varphi_1(\beta) - \cdots - \varphi_p\beta^p$$

$$MA: \theta(\beta) = 1 + \theta_1(\beta) + \cdots + \theta_q\beta^q$$

The seasonality components are;

$$AR: \Phi(\beta^S) = 1 - \Phi_1 - \Phi_1\beta^S - \cdots - \Phi_P\beta^{PS}$$

$$MA: \Theta(\beta^S) = 1 + \Theta_1\beta^S + \cdots + \Theta_Q\beta^{QS}$$

In the equations, β represents the backward shift operator, ε_t stands for estimated residual error at t for $\mu = 0$ and variance is constant and Y_t represents the observed values at $t(t = 1,2,3, \dots, k)$ $\boldsymbol{\phi}$ is a vector of the AR coefficients, $\theta$is a vector of the MA coefficients, Φ is a vector of the seasonal AR coefficients, and Θ is a vector of the seasonal MA coefficients.

In the SARIMA model, seasonal subtraction of appropriate order is used to remove non-stationary data from the series. A first order seasonal difference is the deviation between a value and the corresponding value from the previous year and it is expressed as:$Y_t = X_t - X_{t-s}$ for quarterly time series (S) = 4.

2.3.6 Autoregressive Fractionally Integrated Moving Average (AFRIMA) models

Autoregressive Fractionally Integrated Moving Average (AFRIMA) models are time series models that generalize ARIMA models by allowing non-integer values of the differencing parameter. These models are useful in modelling time series with long memory, that is, in which deviations from the long-run mean decay more slowly than an exponential decay. The acronyms "ARFIMA" or "FARIMA" are often used, although it is also conventional to simply extend the "ARIMA (*p*, *d*, *q*)" notation for models, by simply allowing the order of differencing, *d*, to take fractional values. The general formula for the AFRIMA model is given below;

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k = \sum_{k=0}^{\infty} \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d+1-k)} (-B)^k$$

2.3.7 Exponential Smoothing Technique

Exponential smoothing was first suggested in the statistical literature without reference to previous work by Robert Goodell Brown in 1956 and then expanded by Charles C. Holt in 1957. Exponential smoothing is a technique used to detect significant changes in data by considering the most recent data. Also known as averaging, this method is used in making short-term forecasts. The simplest form of an exponential smoothing formula is given by:

$$F_t = \alpha A_{t-1} + (1 - \alpha) F_{t-1}$$

Here,

F_t = smoothed statistic; A_{t-1} = previous smoothed statistic; α = smoothing factor of data; $0 < \alpha < 1$ and t = time period

If the value of the smoothing factor is larger, then the level of smoothing will reduce. Value of α close to 1 has less of a smoothing effect and give greater weight to recent changes in the data, while the value of α closer to zero has a greater smoothing effect and are less responsive to recent changes.

There is no official accurate procedure for choosing α. The statistician's judgment is used to choose an appropriate factor sometimes. Otherwise, a statistical technique may be used to optimize the value of α.

Exponential smoothing is best used for forecasts that are short-term and in the absence of seasonal or cyclical variations. As a result, forecasts aren't accurate when data with cyclical or seasonal variations are present. As such, this kind of averaging won't work well if there is a trend in the series.

Methods like this are only accurate when a reasonable amount of continuity can between the past and future can be assumed. As such, it's best suited for short-term forecasting as it assumes future patterns and trends will look like current patterns and trends. While this kind of assumption may sound reasonable in the short term, it creates problems the further the forecast goes.

CHAPTER THREE

MATERIALS AND METHODS

This chapter provides details on the study design, specifically on the study setting, data sources, data collection methods and instruments used, use of the data in this research, data analysis approach, assumptions, and their basis. The chapter further presents the study variables to be investigated, the model fitting procedure, and forecasting.

3.1 Study Design

This was a retrospective study conducted within a hospital setting and involved patients diagnosed with TB. The data used for the study was collected from January 1, 2013 to September 30, 2020, encompassing all health facilities located in the north health zone of Malawi. The study relied on secondary data obtained from hospital records, specifically focussing on TB case notifications. All forms of TB cases were included in the study, and data was gathered from various healthcare institutions and facilities that offer DOTS services. We performed a time series analysis to investigate our hypothesis regarding seasonality in TB case notifications.

3.2 Study Setting

Malawi is a low-income country located in southern Africa and has a land area of 118, 000 square kilometres. It shares borders with Zambia to the west, Mozambique to the east, and Tanzania to the north and northeast. The country is divided into three administrative regions: Northern, Central, and Southern regions. To facilitate operations, the Ministry of Health (MoH) has established five health zones, with Southern and Central regions each divided into two zones. The five health zones of Malawi are as follows: North, Central East, Central West, South East and South West.

For the purpose of this study, data was collected from the North Zone of Malawi, with its headquarters located in Mzuzu City. The north health zone comprises six administrative districts: Chitipa, Karonga, Rumphi, Nkhata-Bay, Mzimba and Likoma. Mzimba district is further divided into two health districts namely Mzimba north and Mzimba South. All TB cases notified at a facility level are reported to their respective district hospital through the district TB coordinator who in turn reports the TB notified cases to the TB zonal Coordinator who is based at the zone headquarters (Mzuzu Central Hospital). This study used the data aggregated at the zone headquarters.

Based on census, birth, and death data, the northern region has an estimated total human population of 2,289,780 persons and covers a land area of 27, 130 square kilometres, making it the smallest region both by population and area. According to (National Statistical Office, 2018) data, on population density (measurement of average number of persons per square kilometre), the northern region has the least population density of the three administrative regions with a population density of 84 person per square kilometre. Rumphi district had the lowest population density of 50 persons per square kilometre followed by Chitipa with 54 persons per square kilometre. National Statistical Office, (2018) further reported that Likoma district had the highest population density of 726 persons per square kilometres. Its capital city is Mzuzu. Starting in the north and going clockwise, the Northern Region borders Tanzania, Lake Malawi, Malawi's central region, and Zambia. Mzuzu Central Hospital in Mzuzu city is the main referral hospital serving the region, with district hospitals, rural hospitals and smaller health centres and facilities in the wider areas of the region.

3.3 Data Sources

This study used two different kinds of data; TB case notification data and a population dataset. Cumulative data on quarterly TB case notification was obtained from the health facility register for the period 2013–2020. TB case notification was standardized per 100,000 population per year. For the purpose of this study, the case definition used are further defined in Table 1.

Secondly, yearly district and national population data based on the national demographic health survey (DHS) and census was extracted from the population projections report, which was released by NSO. The population projections report is one of the many reports NSO has produced, including the Census Preliminary Report, released in November 2008, and the Main

Census Report, released in 2009 and 2019. The population projections report presents the projected absolute numbers and age-sex differentials of population in Malawi until 2050 and the national level and until 2030 at the district level. The analytical results were based on data from the 2008 and 2018 Population and Housing Census that was conducted by NSO. According to NSO, the planning and organizational structure put in place ensured high household coverage.

The population projection data was used in calculating TB case notification rates as aggregated by age, age group, HIV status, and district where the TB case was reported.

Table 1: Reference TB case definition

	Clinically diagnosed TB case	A patient who does not fulfil the criteria for bacteriological confirmation but has been diagnosed with active TB.
By site	Pulmonary TB patient	Refers to any bacteriologically confirmed or clinically diagnosed case of TB involving the lung parenchyma or the tracheobronchial tree.
	Extra pulmonary TB patient	Refers to any bacteriologically confirmed or clinically diagnosed patient with TB involving organs other than the lungs, e.g. pleura, lymph nodes, abdomen, genitourinary tract, skin, joints, bones and meninges.
By history of previous treatment	New TB patient	A patient who has never had treatment for TB or who has taken anti-TB drugs for less than one month.
	Relapse patients	A patient who has previously been treated for TB, was declared *cured* or *treatment completed* at the end of their most recent course of treatment, and who is now diagnosed with a recurrent episode of TB.
By HIV status	HIV-positive TB patient	Refers to any bacteriologically confirmed or clinically diagnosed case of TB who has a positive HIV test result from the time of TB diagnosis or other documented evidence of enrolment in HIV care.

3.4 Study Variables

The study utilized variables from health facility records for its analysis. These variables included the age category of the patient, the patient's sex, the notification period of the case, the HIV status of the patient, the name of the health facility, and the district where the case was reported. The TB case notifications were categorized into three age groups: individuals aged below 25 years, those between 25 and 44 years, and those above 45 years. Additionally, the data were stratified based on the sex and HIV status of the study participants.

Table 2: Description of all the variables used in this study

Ser. #	Variable code	Description of the variable	Categories
1	Age-cat	Age category of the TB patients in years	0 - 24 years; 25 – 44 years; 45+ years
2	Sex	Gender of the TB patient	1 = Male 2 = Female
3	HIV Status	HIV status of the TB patient	Positive Negative Unknown
4	Facility	Name of the health facility where patients reported their TB status	All the health facilities where the data was collected
5	District	Name of the district where the health facility is located	Chitipa, Karonga, Rumphi, Mzimba North, Mzimba South, NKhata-Bay, Likoma
6	Quarter	Quarter of the year	Q1 = First quarter Q2 = Second quarter Q3 = Third quarter Q4 = Fourth quarter
7	Year	The year in which the TB case was reported to a particular health facility	2013 – 2020
8	Count	Total number of TB cases reported	
9	Popln	Human population	

3.5 Data Collection Procedure

The data for this study was obtained from patient records following their completion of laboratory tests and clinical diagnosis for TB.

Table 3: Summary of study sample size

District	Sample health facilities	Number of TB cases
Chitipa	9	987
Karonga	11	1, 821
Rumphi	8	929
Mzimba North	12	5, 447
Mzimba South	14	1, 890
Nkhata-Bay	9	1, 212
Likoma	1	38
Total	**64**	**12, 324**

3.6 TB Case Notification Data as Time Series

A time series dataset consists of a sequential of data points recorded at specific time intervals. These intervals can be regular, such as, hourly, daily, weekly, monthly, quarterly or annual. The primary objective of analysing time series data is to identify and understand the underlying components, including trend, seasonality, cyclic patterns, and irregular or random fluctuations (Saadettin, 2022). By examining these components, we can describe the behaviour of the time series and make forecasts based on its historical and current values.

Time series analysis has widespread in various fields, including statistics, epidemiology, econometrics, mathematical finance, weather forecasting, earthquakes prediction and many more (Saadettin, 2022). In the case of this study, the TB case notification data for the north health zone of Malawi qualifies as a time series since it was collected and reported at regular intervals of quarterly throughout the entire study period. This characteristic has led us to employ time series models to forecast future patterns in TB case notifications within the study area.

3.7 Data Analysis and Procedures

Data analysis is an important stage in research that is used to transform, remodel and revise data in order to reach to a certain conclusion for a given situation or problem. For this study, data entry and merging were performed using Microsoft Excel 2013. Exploratory analysis and the generation of descriptive statistics to summarize information were conducted using Microsoft Excel's Pivot Tables. Recognising the significance of the data analysis stage, this study has divided it into sub-stages, outlined below, to facilitate the interpretation of results and ensure a comprehensive analysis.

3.8 Descriptive Analysis

Descriptive statistics were employed to characterize the demographic and health attributes of the study participants who reported their TB cases to health facilities in the north health zone of Malawi between 2013 and 2020. Frequency distribution tables, charts and graphs were used to provide a detailed description of the study participants. The time unit utilised in this study was quarterly, dividing the year into three-month periods: the first quarter (Q1; January to March), second quarter (Q2; April to June), third quarter (Q3; July to September), and fourth quarter (Q4; October to December). This resulted in a total of 31 seasons spanning from January 2013 to September 2020.

Additionally, yearly TB case notification rates were computed for all the six districts using the following formula:

$$TB\ yearly\ case\ notification\ rate\ = \frac{Number\ of\ cases\ per\ given\ year}{Total\ year\ population} \times 100,000$$

In addition, time series graphs were sketched to examine stationarity and non-stationarity of the mean, variance and seasonality/periodicity or trend of the data. P-value < 0.05 was considered to be statistically significant. For the data to be considered stationary, the following requirements were satisfied; constant mean and variance, constant autocorrelation structure and

the data not containing periodic components. If the data is not stationary, proper differencing was applied to make them stationary.

3.9 Estimation of Model Parameters

There are several methods such as the method of moments, maximum likelihood, and least squares that can be employed to estimate the parameters in the tentatively identified model. However, unlike the regression models, most ARIMA models are nonlinear models and require the use of a nonlinear model fitting procedure. This is usually automatically performed by sophisticated software packages such as Minitab, SAS, and R.

Let $\theta = \left(\emptyset_{1,\ldots,}\emptyset_{p,}\theta_1, \ldots, \theta_q\sigma^2\right)'$ denote the vector of population parameter.

Suppose that we have observed a sample of size T then

$$X = (X_1, \ldots, X_T)$$

Let the joint probability density function be (p.d.f.) of these data be denoted

$$f\left(X_T, X_{T-1\,,\ldots,}X_1; \theta\right)$$

The likelihood function is the joint density treated as a function of the parameters θ given the data x;

$$L(\theta IX) = f\left(X_T, X_{T-1\,,\ldots,}X_1 : \theta\right)$$

The maximum likelihood estimator (MLE) is

$$\hat{\theta}_{MLE} = \arg max_{\theta\in\emptyset}\, L(\theta|X)$$

where $\emptyset$ is the parameter space.

For simplifying calculations, it is customary to work with the natural logarithm of L which is a function referred to as the log-likelihood and is given by the following formula;

$$Log\ L\ (\theta 1X) = l\ (\theta|X)$$

Since the logarithm is a monotone transformation the values that maximize $L(\theta|X)$ are the same as those that maximize$l(\theta|X)$, that is

$$\hat{\theta}_{MLE} = \arg max_{\theta\in\emptyset}\, L(\theta|X) = \arg max_{\theta\in\emptyset}\, l(\theta|X)$$

But the log-likelihood is computationally more convenient. Now, we assume that the derivative of $l(\theta|X)$ (w.r. θ) exists and is continuous for all θ.

The necessary condition for maximizing $l\ (\theta|X)$ is

$$\frac{\delta l(\theta|X)}{\delta\theta} = 0$$

Which is called likelihood equation and hence the maximum likelihood estimate will be the solution to

$$\frac{\delta l(\theta|X)}{\delta\theta} = 0$$

3.10 Model Comparison and Selection

Model selection is a crucial step in the forecasting process as it involves choosing a model that could plausibly generate the observed time series and is suitable for producing accurate forecasts and prediction intervals. Since the exact model that completely describes a system is typically unknown, the objective of model selection, as stated by Leeb & Pötscher (2005), is to identify a model that optimizes a particular process.

The primary goal of model selection is to compare different competing models and select the one that best describes the system under investigation. The ultimate aim is to choose the model that exhibits the best predictive ability on average. In the context of time series analysis, model selection becomes crucial because researchers often encounter multiple competing models that may adequately fit the data (Höge, Wöhling, & Nowak, 2018).

Modelling, by nature, involves approximating reality, and therefore, model selection aims to reject models that deviate significantly from reality and select the one that closely aligns with it (Shibata, 1989). Ongbali (2018) emphasizes that the main purpose of model selection is to evaluate the performance of various models and identify the most suitable one for a specific dataset. Failing to consider proper model selection procedures can lead to misleading conclusions in statistical reasoning (Leeb & Pötscher, 2005).

In this study, we used the Box-Jenkin SARIMA and exponential smoothing approaches to identify the best model to forecast future patterns in the TB case notification rate in the north health zone of Malawi. We used **auto.arima** function in R to identify the best model to predict future trends of TB case notifications. Previous research suggest that TB case notification exhibit seasonal pattern (Bodena, Ataro, & Tesfa, 2019; Liu, Zhao, & Zhou, 2010; Fares, 2011). Therefore, when modeling seasonal patterns in TB case notifications, the SARIMA model has been widely employed by researchers. It has been found that this method provides an appropriate model for forecasting future trends in TB case notifications. The SARIMA model has already been specified in the preceding sub-section. Below are several statistics commonly utilized in the model selection process by researchers, which this study has also adopted. It is generally preferred to select models with smaller values based on the chosen criterion, as outlined below.

In addition, the holdout approach was employed to generate forecasts for the holdout set, which represented a future period not used during the model training. The holdout approach offers a valuable means of assessing the model's performance on unseen data, allowing for an understanding of its generalization capabilities beyond the training period. It also helps identify any potential issues of overfitting or underfitting, providing insights into the model's ability to capture the underlying trends and dynamics of the time series.

3.10.1 Akaike Information Criterion (AIC)

For maximum likelihood or empirical Bayesian, one can use the Akaike Information Criterion (Cui & George, 2008). The Akaike, (1973, 1974) information criteria was developed as estimators of the expected Kullback-Lieber discrepancy between the model generating the data and a fitted candidate model (Cui & George, 2008). The AIC is one of the statistics used to select the best model. It is defined as:

$$AIC = 2k - 2In(L)$$

Where k is the number of parameters fitted in the statistical model, and L is the maximised value of the likelihood function for the estimated model. The smaller values indicate more parsimonious models and as such, models with the lowest/minimum AIC is chosen. The term 2k is a penalty to be paid for overfitting and this discourages adding too many variables in the models which always leads to a smaller likelihood. This provides the trade-off between over fitting and optimum model fit.

3.10.2 Bayesian Information Criterion (BIC)

The BIC is a model selection criterion in statistics, introduced by Schwarz in 1978. It is derived from the empirical log-likelihood function and it does not necessitate the specifications of prior distributions. This characteristic makes the BIC particularly advantageous in situations where setting priors is challenging or impractical (Schwarz, 1978). The BIC is closely related to the Akaike Information Criterion (AIC), and both criteria take into account the balance between model fit and complexity. They penalize models with a larger number of parameters, aiming to strike a balance between goodness of fit and model simplicity. Therefore, the BIC is defined as follows;

$$BIC = -2\,In(L) + kIn(n)$$

Where L is the maximised value of the likelihood function of the model, k is the number of free parameters in the model and n is the number of observations in the time series or simply the sample size.

3.11Examining the Seasonality of the Time Series

Time series models are different from Multiple and Poisson Regression models in that time series models do not contain the cause-effect relationship. They use mathematical equation(s) to find time pattern in series of historical data. These equations are then used to project into the future the historical time pattern in the data. There are three types of patterns in time series; trend, seasonal and cyclic. A trend pattern exists when there is a long-term increase or decrease in the series. These trends may be linear, exponential, or different one can change direction

during the time. Seasonal exists when data is influenced by seasonal factors, such as a day, a week, month, or a quarter of the year. A seasonal pattern exists of a fixed known period. A cyclic pattern occurs when data rise and fall, but this does not happen within the fixed time. In addition to the three patterns of time series data, there also exists errors or residuals. The process of extracting these components/types of patterns from the time series data is what is called decomposition (Cleveland & Tiao, 1976).

To examine seasonality, seasonal decomposition among time series analysis methods was used to calculate the seasonal index. There are two approaches for decomposing time series models, there are multiplicative and additive approaches. The multiplicative approach posits that the variance in the results can be explained by the product of the trend factor, circular factor, seasonal factor, and error. The additive approach posits that the variance can be explained by the sum of the four factors of trend, seasonality, residuals and circular. Of these two models, the additive seasonal approach is taken if the seasonal variation is consistent despite the increase in time series, while the multiplicative seasonal approach is taken if seasonal variance increases or decreases depending on the increase or decrease of the time series. Because the data of the current study did not have a consistent seasonal variation across the flow of time, the multiplicative approach was taken.

3.12 Model Diagnostics

Typically, the goodness of fit of a statistical model to a set of data is judged by comparing the observed values with the corresponding predicted values obtained from the fitted model. If the fitted model is appropriate, then the residuals should behave in a manner that is consistent with the model. One of the tests that was used in the diagnostics of this study was the Ljung-Box test. Ljung-Box test is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero (Brockwell & Davis, 2002). The null and alternative hypotheses for the Ljung-Box test are defined as follows:

H_0: the data are independently distributed (i.e. correlations in the population from which the sample is drawn are zero, so that any observed correlation in the data results from randomness of the sampling process).

H_1: the data are not independently distributed; they exhibit serial correlation.

The test statistic for Ljung-Box test is given by:

$$\boldsymbol{Q = n(n+2)\sum_{k=1}^{h}\frac{\rho_k^2}{n-k}}$$

Where n is the sample size, ρ_k^1 is the sample autocorrelation at lag k, and h is the number of lags being tested. Under the null hypothesis, Q asymptotically follows Chi-square distribution.

We reject the null hypothesis and say that the model shows lack of fit if

$$Q = x_{1-\alpha,h}^2$$

During the diagnostic testing, we wanted to check the error terms of the chosen time series model to be used in making the forecasts. The P-value for the Ljung-Box statistics should be greater than the chosen significant level (0.05) and hence we fail to reject the null hypothesis, therefore we can make a conclusion that the mean error terms of our model is zero.

The relationship of the error terms was checked using the ACF plots. This was done by checking the ACF of the residuals chart. Our wish was that the ACF values of the error terms be non-significant. If the error terms were significant, that means they are correlated, hence our time series analysis model does not fully explain the relationship between independent and dependent. Thus, the error terms are correlated. Also, if the error terms of the ACF plots are not significant, that means the error terms are random and not correlated which is a good thing because our chosen time series model can fully explain the relationship between independent and dependent. The diagnostic results showed that our chosen model was suitable to fully explain the relationship between the dependent and independent, hence it was possible to use our chosen model in forecasting future incidence of TB in the north health zone.

3.13 Model Forecasting

To forecast the TB case notifications for the future, seasonal ARIMA (0, 1, 2) $(1, 0, 0)_4$ model was utilized. The Ljung-Box goodness-of-fit test was also used to ascertain the significance of the fitted seasonal pattern. The Ljung-Box test was correctly specified and there was no outlier in the data.

Four models were compared for the suitability to be used in the forecasting future incidence of TB in the north health zone. After examining different models, ultimately the seasonal ARIMA (0, 1, 2) $(1, 0, 0)_4$ model was chosen to be the best model suitable to be used for forecasting future seasonal patterns in TB case notification in the north health zone in Malawi. The AIC and BIC values were used in determining which models was the most suitable to predict future incidence of TB. The study done by Zhang et al. (2020) on predicting TB prevalence in China used the ARIMA model to fit the changes of the incidence and also to predict the incidence in the future. This study will investigate the performance of various forecasting methods including MA, ARIMA, Holt-Winter's, and SARIMA for monthly TB data forecasting. Na et al. (2004) shows that the ARIMA model can be used to appropriately fit the changes of the incidence of PTB in Sichuan province of China. The exponential smoothing model was also applied to make short-term predictions of TB case notification rates in the study area.

3.14 Ethical Consideration

Confidentiality of TB patients has been ensured as each TB patient's name has not been used anywhere in the analysis. The letter of permission was written to the north health zone in Malawi to access to their quarterly reports. An authorisation to access TB case notification data was granted by the Hospital's Director – Mzuzu Central Hospital, which is the main referral health facility in the health zone under study.

CHAPTER FOUR

RESULTS AND INTERPRETATION

This chapter presents results of the study from the analysis of the TB case notification data collected from the hospital records in the north health zone of Malawi. The first part provides descriptive results of the data analysis, the second part provides the results of the time series modelling of the TB case notifications and the last part provides the predicted future patterns in TB case notifications. Furthermore, this chapter also provides interpretation to key findings of the study.

4.1 Descriptive Results of TB Case Notifications

The North Health zone of Malawi has got more than 65 health facilities where TB screening is done. This study, however, has considered 64 health facilities where TB screening has been done during the study period. Other health facilities where TB screening is also done have not been included in this study due to their failure to report TB cases to their respective reporting lines. Of these 64 health facilities, there are referral hospitals, district hospitals, community/rural hospitals, health centres and prison hospitals. In terms of ownership of the health facilities, some are public/government facilities, private profit facilities as well as Christian Health Association of Malawi (CHAM) facilities. Between 1st January, 2013 and 30th September, 2020, about 12, 324 newly and active diagnosed cases of TB were recorded in the North health zone of Malawi. Of these 12173 TB cases notified during the study period, 4800 (39%) were female and 7, 524 (61%) were male. This means that more males are diagnosed of TB in the north health zone as compared to women. Thus, gender of a person could be regarded as one factor associated with prevalence of TB. In terms of age category, 2, 594 (21%) were from the 0 – 24 years age group, 5, 695 (46%) were between the ages of 24 and 44 while the age group 45 years and above were represented by 4, 035 (33%) TB patients.

The data was also analysed to find out the proportion of TB patients who were HIV-positive. Of the 12,324 study participants, 10, 812 (87.7%) knew their HIV status. Of the 10,812 patients who had their blood tested for HIV, it was observed that 5, 461 (44.31%) were HIV-negative while 5, 351 (43.44%) were HIV-positive and 1, 512 (12.27%) didn't know their HIV status during data collection. This means that the TB and HIV co-infection among the tested TB patients in our study was 43.4% which shows that there is a very high burden of TB and HIV co-infection incidence in the health zone under study. The results of this study further showed that in the north zone there were more male TB patients who are HIV positive across all the age groups.

In terms of the period of the year, first quarter had a highest TB case notification with 3207 (26.02%). Second, third and fourth quarters were represented by 3, 026 (24.55%), 3191 (25.89%) and 2, 900 (23.53%) respectively. Table 4 below summarizes the demographic details/characteristics of the TB cases notified during the study period.

Table 4: Socio-demographic and clinical characteristics of the study participants in the north health zone of Malawi, January 2013 – September 2020.

Characteristic	Number (N = 12324)	Percentage
Age category in years		
0 – 24 years	2594	21.05
25 – 44 years	5695	46.21
45 years and above	4035	32.74
Sex		
Male	7524	61.05
Female	4800	38.95
HIV Status		
Positive	5351	43.42
Negative	5461	44.31
Unknown	1512	12.27
Period (quarter of the year)		
First quarter	3207	26.02
Second quarter	3026	24.55
Third quarter	3191	25.89
Fourth quarter	2900	23.53

4.2 TB Case Notification Rates

From figure 1 below we can observe that the number of TB case notification rate was almost constant from the year 2013 to 2015 followed by a slight decrease in 2016. This was followed by a slightly increase in the year 2017 and 2018 and then there was a sharp rise in the TB case notifications in 2019 followed by a sudden decrease in the year 2020. A sharp increase in the case notification rate in 2019 may be due to community mass campaign about care seeking behaviours which improved and influenced the population's understanding and behaviour about timely health seeking behaviours. A sharp decrease in the TB case notification rate in the year 2020 may have been due to the Covid-19 pandemic. During this period, there was apathy from the general population in seeking health care and that most health facilities were not working normally. Based on annual TB case notifications, the notification rate decreased from

80.58 to 48.90 cases per 100,000 population per year between 2013 and 2020. According to (National Statistical Office, 2018), the population for the north health zone region increased from 1, 826, 802 persons in 2013 to 2, 247, 493 persons in 2020. Therefore, TB case notification rates increased to a plateau of 96 cases per 100 000 population in the year 2019 to of 48 cases per 100 000 population in 2020. In addition, the TB case notification rate from 2013 to 2020 trended slightly downward but were still significantly high.

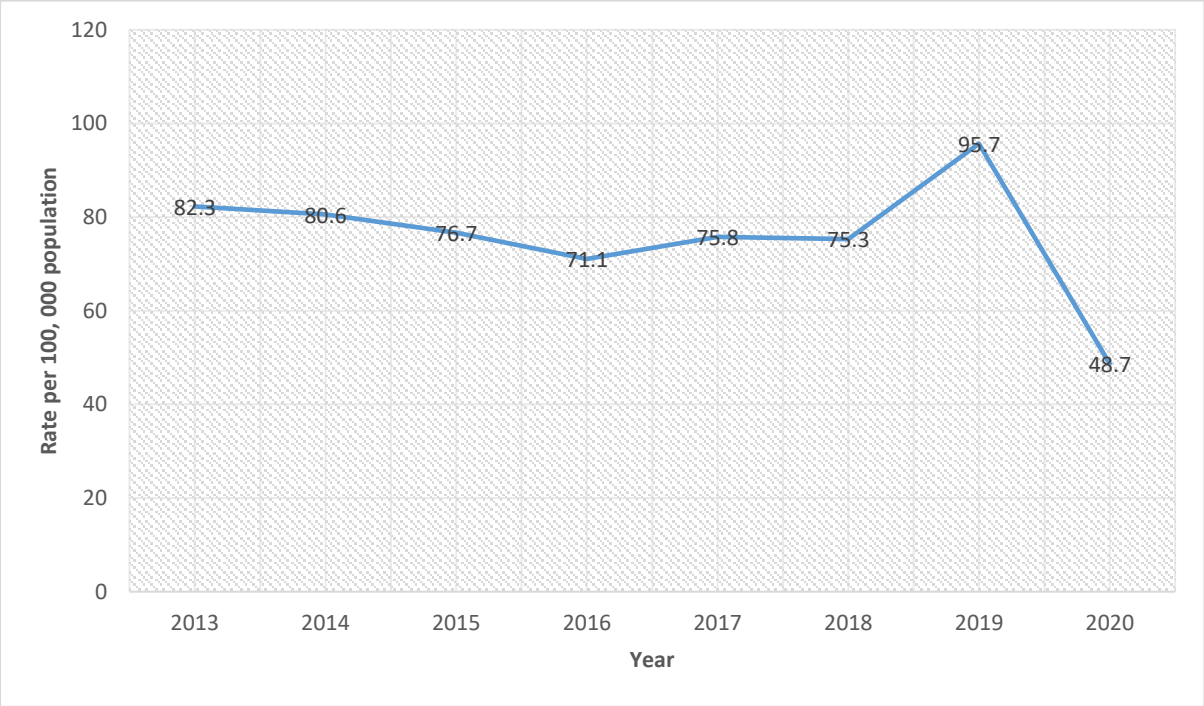

Figure 1: Graph of TB case notifications rates per given year (2013 -2020)

4.2.1. TB Case Notification Rate as Stratified by Age Group, Sex, HIV Status, Year and District

The TB case notification rate was stratified by age group, sex, place of notification, and HIV status. Age was categorized into three groups: <25 years (young-aged group), 25-44 years (middle-aged group), and above 44 years (aged group). Seasonal trends were observed in the middle-aged and aged groups. Similar seasonal variations in the TB case notification rates were seen in both the middle-aged and aged groups. However, certain peaks in the TB case notification rates seen in the aged group were not observed in both the young-aged and middle-

aged groups. Furthermore, the young-aged group had the lowest TB case notification rates among all the age groups. A study conducted by the National Tuberculosis Commission in 2018 reported that age-specific notification rates were highest among individuals aged 65 years and above. This contrasted with the 2017 study, where individuals aged between 35-44 years had the highest notification rates. These results are shown in Figure 2 below.

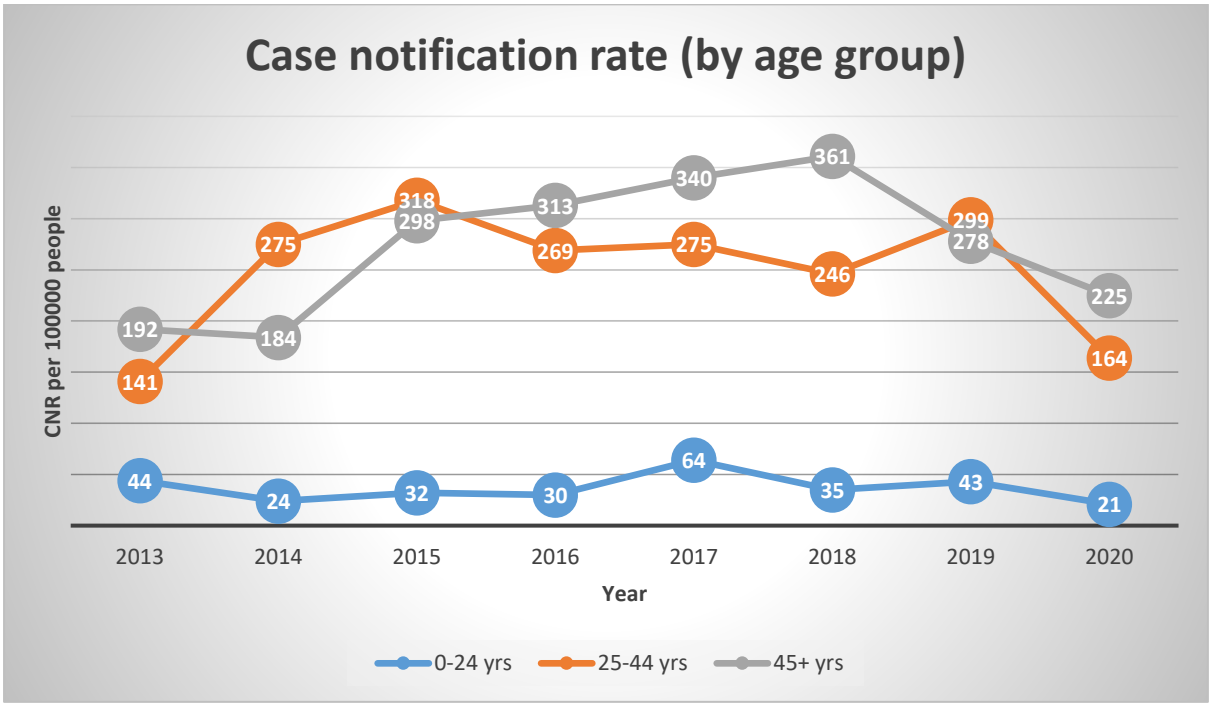

Figure 2: TB case notification rate (CNR) per 100, 000 population as stratified by the Age group

A similar pattern of seasonal variation was seen in both female and male case notifications indicating that sex of an individual is not likely to be a significant factor that influences TB seasonality. However, the male gender has a higher TB case notification rate across all the years of the study period. The highest record was observed in the male gender in 2017 (398 cases per 100,000 persons) while the lowest TB case notification rate was observed in female gender in 2020 (57 cases per 100,000 persons), as shown in the figure 3 below. This is consistent with findings of the national prevalence survey which showed that men had higher prevalence compared to women.

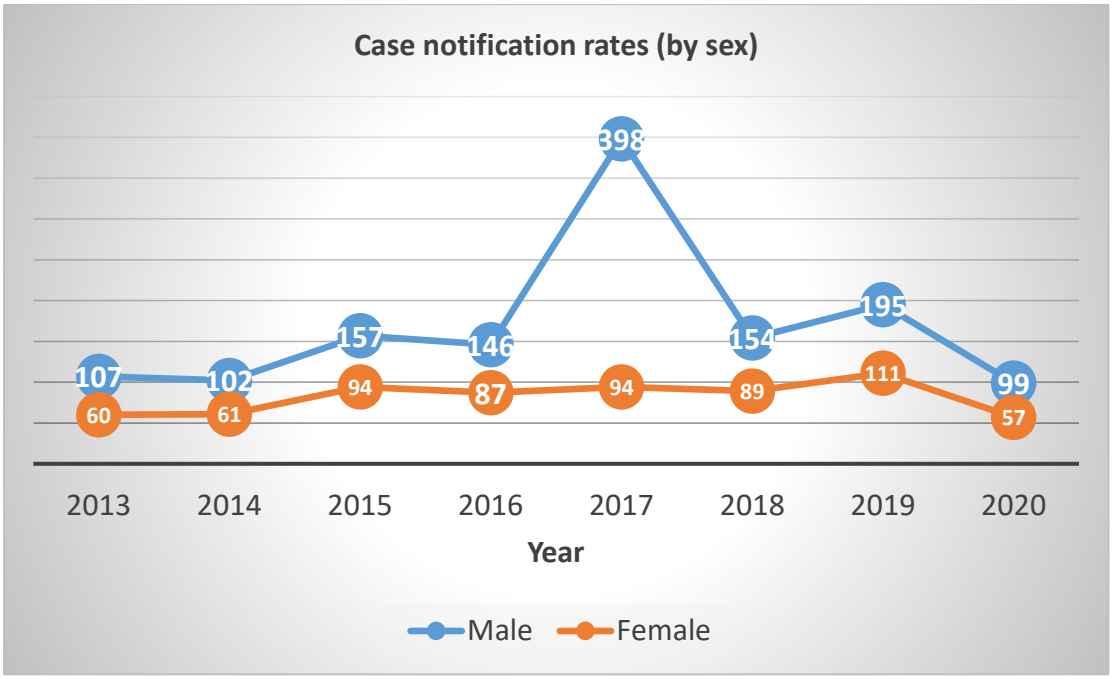

Figure 3: TB case notification rate (CNR) per 100, 000 population (by Sex)

In terms of the HIV status of TB patients, we observed some form of a zigzag pattern with lows and highs throughout the study period. In the HIV-positive TB patients, the TB case notification rate was lowest in the second quarter of 2018 with 7 cases per 100 000 population and hit the maximum point in the third quarter of 2019 with 13 cases per 100 000 population. For the HIV-negative TB patients, the TB case notification rate hit its all-time lowest point in the third quarter of 2016 with 6 cases per 100, 000 population and reached its maximum point in the second quarter of 2019 with 15 cases per 100, 000 population. Overall, the TB case notification rates for both HIV-positive and HIV-negative TB patients were almost constant from the year 2013 to 2018 followed by a sharp increase in 2019 and a sharp decrease in 2020 as evidenced from figure 4 below.

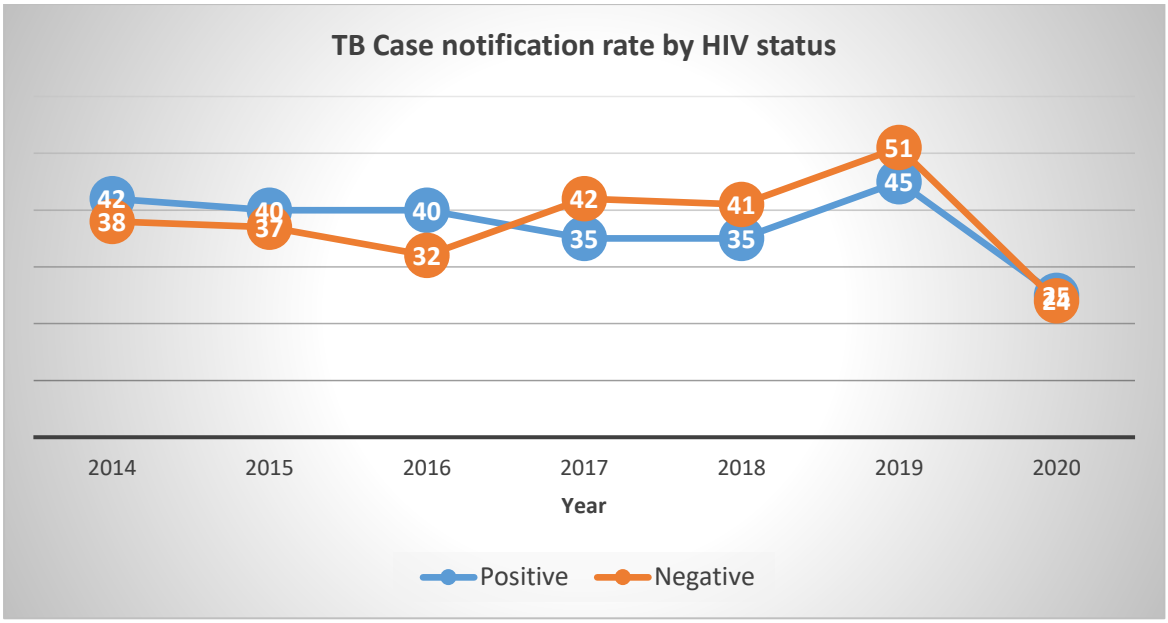

Figure 4: TB case notification rate (CNR) per 100, 000 population (by HIV Status)

In terms of district where the patients reported their cases, Mzimba North and South were combined as one district in the scenario because population projections for the same were combined. Figure 5 shows that Mzimba district had the highest TB case notification rates of all the district under study throughout the entire study period. There was a decline in the case notification rate between 2013 and 2018 with 110 and 85 cases per 100 000 population respectively. The district hit its highest point in terms on TB case notification rate in 2019 with 132 cases per 100 000 population followed by the lowest record in case notification rate in 2020 with 58 cases per 100 000 population. The lowest case notification rate was observed in Chitipa district in 2016 with 9 cases per 100 000 population. The rest of the districts registered their lowest record in case notification rate in the year 2020 with Rumphi registering the lowest record with 5 cases per 100 000 population.

The high total TB notification rate as well as EPTB in Mzimba district can be explained by the cold weather around the main health facilities of the districts (Mzuzu central hospital, Mzuzu health centre, St. John's hospital as well as Ekwendeni mission hospital). Low temperatures lead to low vitamin D which significantly increases the incidence of smear and sputum positive tuberculosis. In addition, the high TB case notification rate in the district can be explained by the capacity to diagnose TB in individuals owing to availability of expertise as well as diagnostics in this district that allow for diagnosis and treatment of TB in individuals at all

ages. Mass campaigns can also be used to explain the variation of TB case notification rates across the districts whereby there are more campaigns in urban areas (where the major health facilities of Mzimba districts are located) than in rural settings. At this point, heterogeneity of TB burden in districts of the health zone might not strongly account for inter-district variation due to unavailability of sub national disease burden estimates.

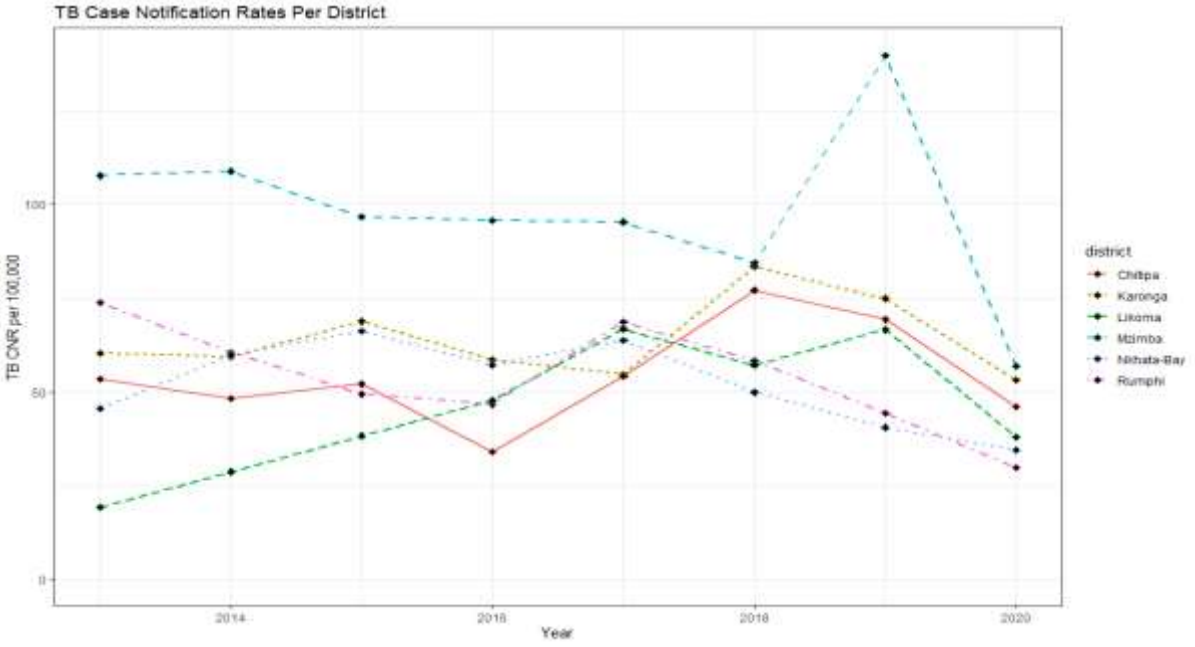

Figure 5: TB case notification rate (CNR) per 100, 000 population (by District)

4.3. Building the ARIMA model

The focus of this research was on comparing forecasts in time series analysis of TB case notification data. Prior to model fitting, a time series plot was created to assess the behaviour of the data over an 8-year period (see Figure 6). From Figure 6, it is evident that the TB case notification data exhibit non-stationarity, with a varying mean and fluctuating variance. Additionally, there appears to be a cyclical pattern present in the data. Notably, significant peaks and troughs are observed, which do not occur at regular intervals, and the time gaps between the troughs and peaks are irregular. Figure 6 represents the original dataset before applying transformations such as differencing and calculating the moving average for all notified TB cases during the study period.

The graph illustrates that between 2013 and 2018, the number of TB cases reported ranged from approximately 310 to 450 per season. There was a sharp increase in TB case notifications from 2018 to 2019, followed by a notable decrease in 2020. Given these observations, it was necessary to examine the data for seasonality and describe the nature of the seasonal patterns.

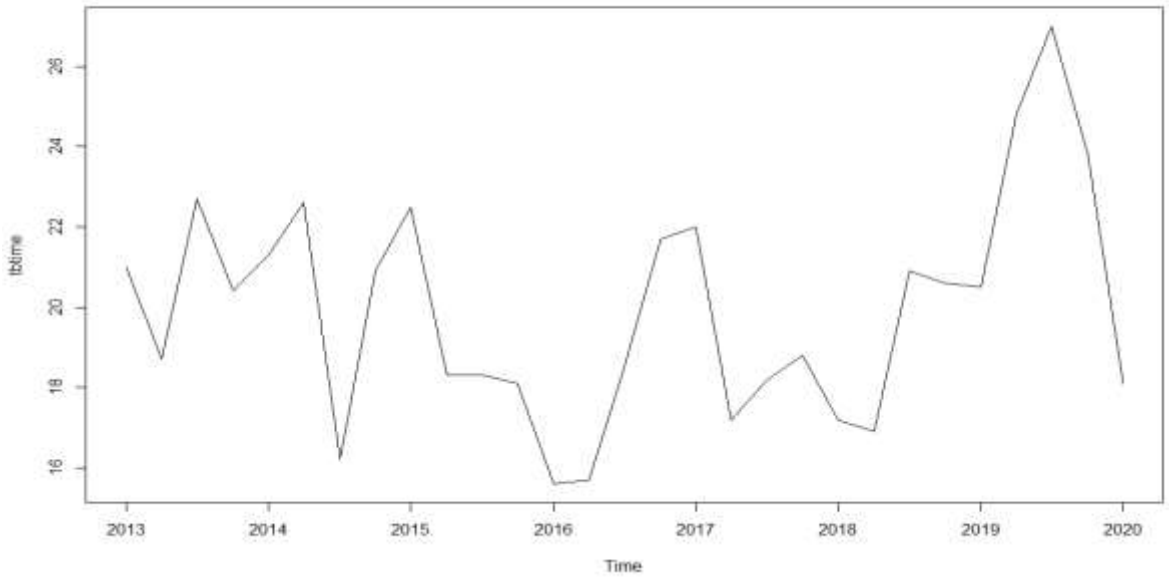

Figure 6: Quarterly TB case notification rates from January 2013 to September 2020

4.3.1. Stabilising the variance by transforming the data

Since the graph above shows ups and downs in the data, this prompted us to manipulate the data by smoothing out the ups and down by a way of taking the moving average of number of TB cases notified during the study period. Data transforms are intended to remove noise and improve the signal in time series forecasting. Since our data set contains a time period of four quarters, we took a four quarterly moving average on the original data.

4.3.2. Results of the analysed stationarity of the transformed time series data

Two methods were used to test for stationarity i.e., ACF-PACF graphs and the ADF unit root test. The ACF and PACF plots were used to identify which time series model to use in the analysis. From the ACF plot (figure 7), we can see that some of the lines spike through the blue dashed lines indicating that our data is non-stationary. Similarly, the PACF plot shows that two

of the vertical lines spike through the blue-dashed lines indicating the non-stationarity of our time series data. So, both the ACF and the PACF plots display correlation between a series and its lags explained by the previous lags.

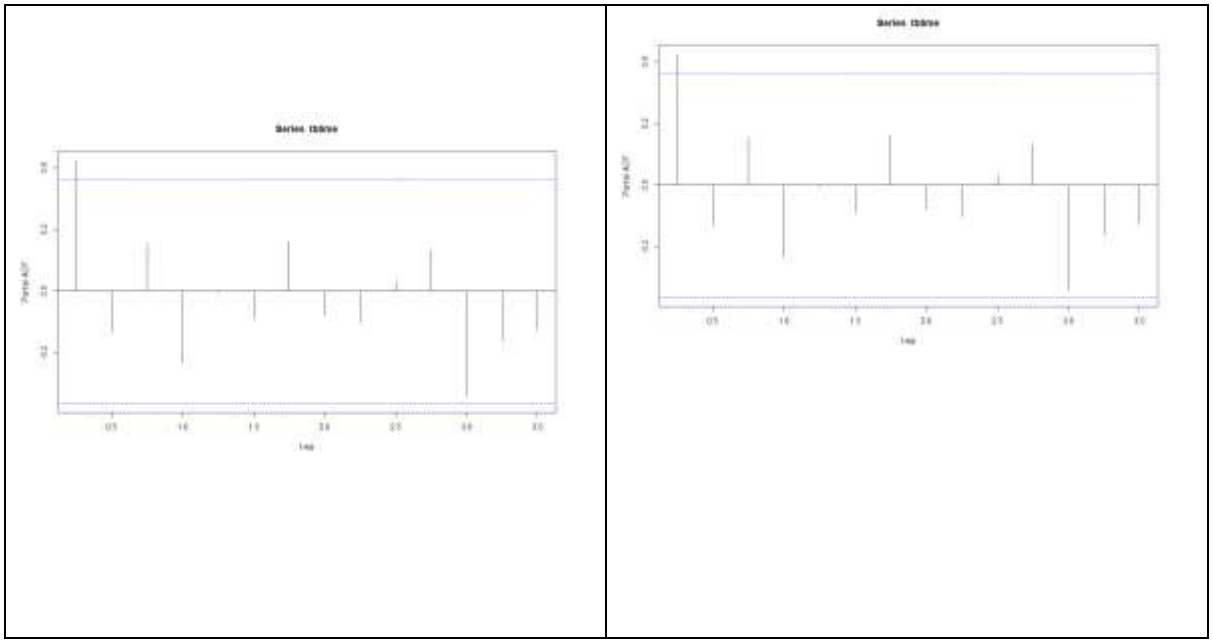

Figure 7: ACF and PACF graph for the un-differenced time series data

The ADF test results showed that our data is not stationary (Dickey-Fuller = -2.4675, Lag order = 3, P-value = 0.3934). Therefore, we fail to reject the null hypothesis at the 5% significance level, suggesting that there is insufficient evidence to conclude that the time series is stationary. This corresponds to the ACF AND PACF graphs as shown above. As such, this prompted us to differentiate the data to make it stationary before testing our time series models.

4.3.3. Differencing the data

We performed first-order differencing on our data to eliminate the seasonal component and bring the data points closer together. This approach was chosen so as to obtain more accurate data compared to higher-order differencing, which would have resulted in wider gaps between data points, posing a challenge for forecasting and reducing accuracy. Subsequently, an Augmented Dickey-Fuller (ADF) test was conducted to assess the stationarity of the differenced data. The ADF test results (Dickey-Fuller = -5.7044, Lag order = 3, P-value = 0.01) indicated that our differenced data achieved stationarity. Since the P-value is less than the

chosen significance level of 0.05, we reject the null hypothesis of non-stationarity, conclude that the time series is indeed stationary. This suggests that the differencing process successfully resulted in achieving data stationarity.

During the model identification process, we explored various potential models using the "**auto.arima**" function from the "forecast" package in R software. The selection method of the best-fitted model was based evaluating the AIC, AICc (Corrected Alkaike Information Criterion), and BIC, with minimum values. Typically, the model with the lowest AIC (or AICc) is considered a strong candidate for the best-fitted model, rather than sorely relying on the BIC value. In Table 5 below, suggested ARIMA models have been presented with their corresponding AICc and AIC information criteria.

Table 5: AIC values for suggested ARIMA models of TB case notification rates time series data by using stepwise selection.

ARIMA(2,1,2)(1,0,1)[4] with drift : Inf
ARIMA(0,1,0) with drift : 270.9173
ARIMA(1,1,0)(1,0,0)[4] with drift : 246.3509
ARIMA(0,1,1)(0,0,1)[4] with drift : 244.1917
ARIMA(0,1,0) : 268.918
ARIMA(0,1,1) with drift : 250.9219
ARIMA(0,1,1)(1,0,1)[4] with drift : 244.1787
ARIMA(0,1,1)(1,0,0)[4] with drift : 242.5038
ARIMA(0,1,1)(2,0,0)[4] with drift : 244.1948
ARIMA(0,1,1)(2,0,1)[4] with drift : 246.173
ARIMA(0,1,0)(1,0,0)[4] with drift : 260.5784
ARIMA(1,1,1)(1,0,0)[4] with drift : 242.4019
ARIMA(1,1,1) with drift : 251.0265
ARIMA(1,1,1)(2,0,0)[4] with drift : 243.7547
ARIMA(1,1,1)(1,0,1)[4] with drift : 243.8249
ARIMA(1,1,1)(0,0,1)[4] with drift : 244.742
ARIMA(1,1,1)(2,0,1)[4] with drift : 245.753
ARIMA(2,1,1)(1,0,0)[4] with drift : 244.3371

ARIMA(1,1,2)(1,0,0)[4] with drift	: 244.1503
ARIMA(0,1,2)(1,0,0)[4] with drift	: 242.1959
ARIMA(0,1,2) with drift	: 250.5047
ARIMA(0,1,2)(2,0,0)[4] with drift	: 243.5275
ARIMA(0,1,2)(1,0,1)[4] with drift	: 243.6065
ARIMA(0,1,2)(0,0,1)[4] with drift	: 244.6192
ARIMA(0,1,2)(2,0,1)[4] with drift	: 245.5176
ARIMA(0,1,3)(1,0,0)[4] with drift	: 244.155
ARIMA(1,1,3)(1,0,0)[4] with drift	: Inf
ARIMA(0,1,2)(1,0,0)[4]	: 240.8076
ARIMA(0,1,2)	: Inf
ARIMA(0,1,2)(2,0,0)[4]	: 242.4969
ARIMA(0,1,2)(1,0,1)[4]	: 242.5549
ARIMA(0,1,2)(0,0,1)[4]	: 243.5017
ARIMA(0,1,2)(2,0,1)[4]	: 244.3993
ARIMA(0,1,1)(1,0,0)[4]	: 241.2273
ARIMA(1,1,2)(1,0,0)[4]	: 242.7822
ARIMA(0,1,3)(1,0,0)[4]	: 242.7876
ARIMA(1,1,1)(1,0,0)[4]	: 240.8967
ARIMA(1,1,3)(1,0,0)[4]	: 244.7773

Best model: ARIMA $(0, 1, 2)\ (1, 0, 0)_4$

As shown in the Table 5 above, the best model under the stepwise method among the other models has been chosen as ARIMA $(0, 1, 2)\ (1, 0, 0)_4$ model with drift and having the smallest AIC value of 240.8076. All other models which have greater AIC values have been provided only for comparison purposes. After noting the best model based on AIC, we estimated the significance of parameters and our results are shown in Table 6 as follows;

Table 6: Estimates of parameters from the ARIMA (0, 1, 2).

	MA1	MA2	SAR1
Coefficients	1.0770	0.2886	-0.7055
Standard error	0.2307	0.1994	0.1614
Sigma squared estimate	131.9		
Log likelihood	-116.4		
AIC = 240.81	AICc = 242.41	BIC = 246.41	

In order to select the best model to be used in the prediction, three competing ARIMA models were further tested in order to select the model with the best predictive ability. Table 7 below provides the estimates from the competing models. Furthermore, ACF and PACF plots were sketched for the identified best models together with the competing models. From the ACF & PACF plots and the estimates from the competing models, it is obvious that our ARIMA (0, 1, 2) model remains our best model to be used in the prediction of future trends of TB cases.

Table 7: Estimates of parameters from the competing models

Model	AIC	AICc	BIC
ARIMA (1, 1, 0)	255.58	265.03	258.39
ARIMA (1, 1, 3)	251.86	254.36	258.86
ARIMA (1, 1, 1)	250.50	252.10	256.11

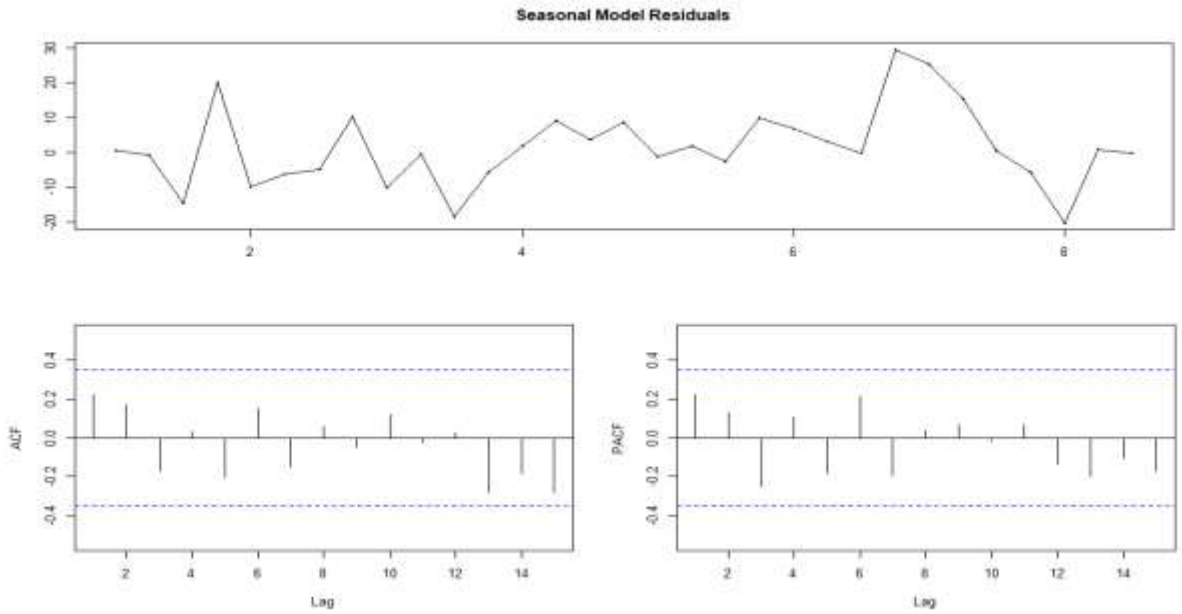

Figure 8: Time plot, ACF and PACF plot for the ARIMA (1, 1, 1) model residual

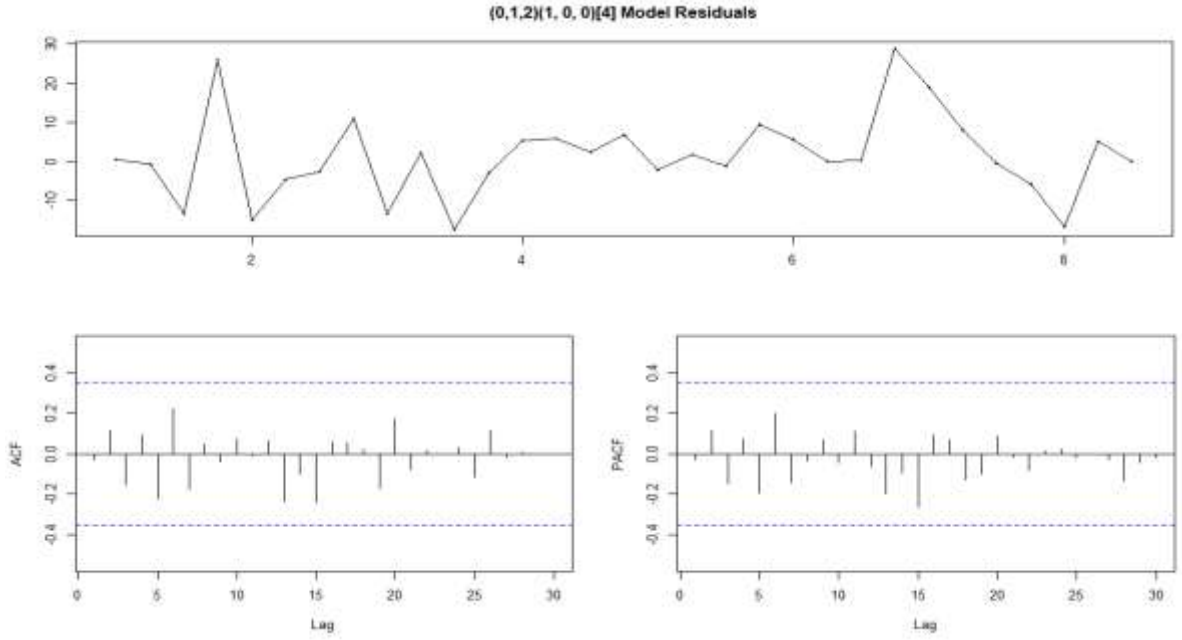

Figure 9: Time plot, ACF and PACF plot for the ARIMA (0, 1, 2) (1, 0, 0)$_4$ model residual

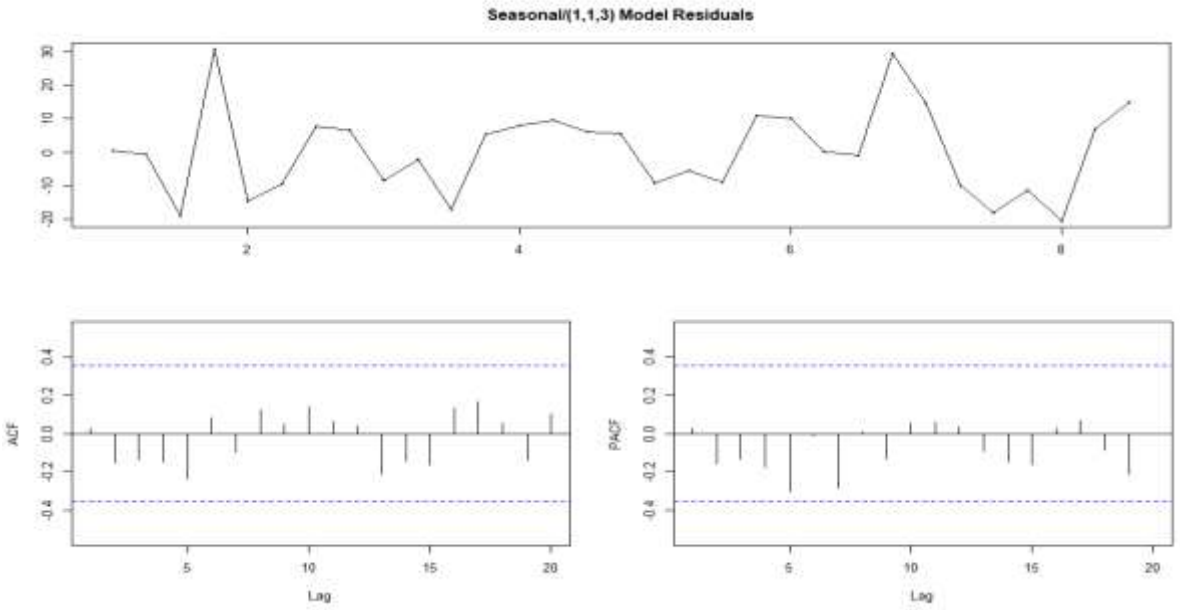

Figure 10: Time plot, ACF and PACF plot for differenced seasonal ARIMA (1, 1, 3) model residual.

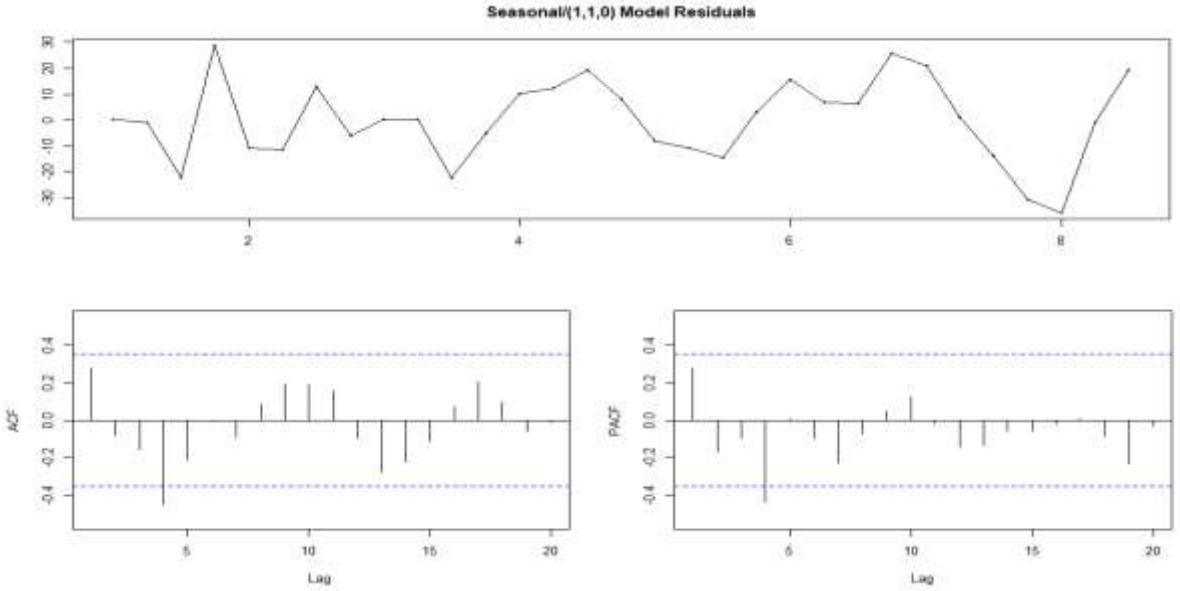

Figure 11: Time plot, ACF and PACF plot for differenced seasonal ARIMA (1, 1, 0) model residual

4.4. The Ljung-Box test results for the randomness of the residuals

After examining the aforementioned figures, the next step was to select the best-fitting model for predicting future trends in TB case notifications in the north health zone of Malawi. To accomplish this, we proceeded with the examination of residuals diagnostics, which is essential to determine whether the residuals exhibit a white noise process. A white noise process is a crucial assumption for a reliable ARIMA model, characterized by a zero mean, constant variance, and no serial correlation. In this stage, we specifically focused on the Ljung-Box test results to ensure that the residuals did not possess any remaining autocorrelation. The null and alternative hypotheses for the Ljung-Box test are gives as follows;

H_0: The residuals are random (independently distributed – the model does not show lack of fit)

H_1: The residuals are not random (not independently distributed, displaying serial correlation – the model does show a lack of fit).

Table 8: Ljung - Box Goodness-of-fit Test Results of ARIMA (0, 1, 2) model

Seasonal lags	X-Squared statistics	P-Values
1	0.0347	0.8523
2	0.5145	0.7732
3	1.3645	0.7139
4	1.6734	0.7955
5	3.6375	0.6027
14	11.335	0.6595

4.4.1. The Ljung-Box Test Results for the Randomness of Residuals from ARIMA (0, 1, 2) Model

We applied the Ljung-Box test to the residuals from an ARIMA (0, 1, 2) model fit to determine whether residuals are random. In this analysis, the Ljung-Box test results showed that the first 14 lag autocorrelations among the residuals are zero (p-value = 0.6595) indicating that the residuals are random and that the model provides an adequate fit to the data. Therefore, according to the results in the Table 8 above, we failed to reject the null hypothesis and

conclude that the mean of error terms of our model is zero and that our selected time series model fits the data well. In addition, the Box-Pierce test results also showed that the model fits the data well (p-value = 0.8591).

In addition, the autocorrelation plot of residuals (ACF Residuals) from the ARIMA (0, 1, 2) model was generated (Figure 12). The autocorrelation plot shows that for the first 14 lags, all sample autocorrelation except those at lag 0 and lag 11 fall inside the 95% confidence band indicating the residual appear to be random.

Furthermore, the result is also verified by looking at the correlogram of the residuals as shown in the ACF and PACF plots below. From the ACF and PACF plots, we can see that the spikes are within the significance limit and mean of the residuals seem to be randomly distributed around zero. Thus, the residuals appear to be white noise.

4.4.2. The Ljung-Box Test Results for the Randomness of Residuals from ARIMA (1, 1, 3) Model

Similar to the result for the ARIMA (0,1,2) model, ACF plot for the residuals (figure 14) shows that for the first 14 lags, all sample autocorrelations fall inside the 95% confidence bounds indicating the residuals appear to be random. The Box-Ljung test was also applied to the residuals from the ARIMA (1, 1, 3) model. The Ljung-Box test results showed that the first 14 lag autocorrelations among the residuals are zero (p-value = 0.6788) indicating that the residuals are random and that the model provides an adequate fit to the data.

4.4.3. The Ljung-Box Test Results for the Randomness of Residuals from ARIMA (1, 1, 0) Model

The residual model analysis was also performed on ARIMA (1, 1, 0) model to check whether it was appropriate or not. From Figure 15 we observed that all the lags except lag 4 were within the 95% confidence band. We further observed that ACF plot showed that the residuals formed a seasonal pattern. The Ljung-Box test indicated that there was at least one non-zero autocorrelation among the first 14 lags. We conclude that there is not enough evidence to claim that the residuals are random (p-value = 0.022).

4.4.4. The Ljung-Box Test Results for the Randomness of Residuals from ARIMA (1, 1, 0) Model

We conducted the Ljung-Box test on the residuals obtained from fitting an ARIMA (1, 1, 1) model to examine whether the residuals exhibit randomness. The Ljung-Box test results revealed the presence of at least one non-zero autocorrelation among the first 14 lags. Based on these results, we conclude that there is insufficient evidence to support the claim that the residuals for ARIMA (1, 1, 1) model are truly random (p-value = 0.059).

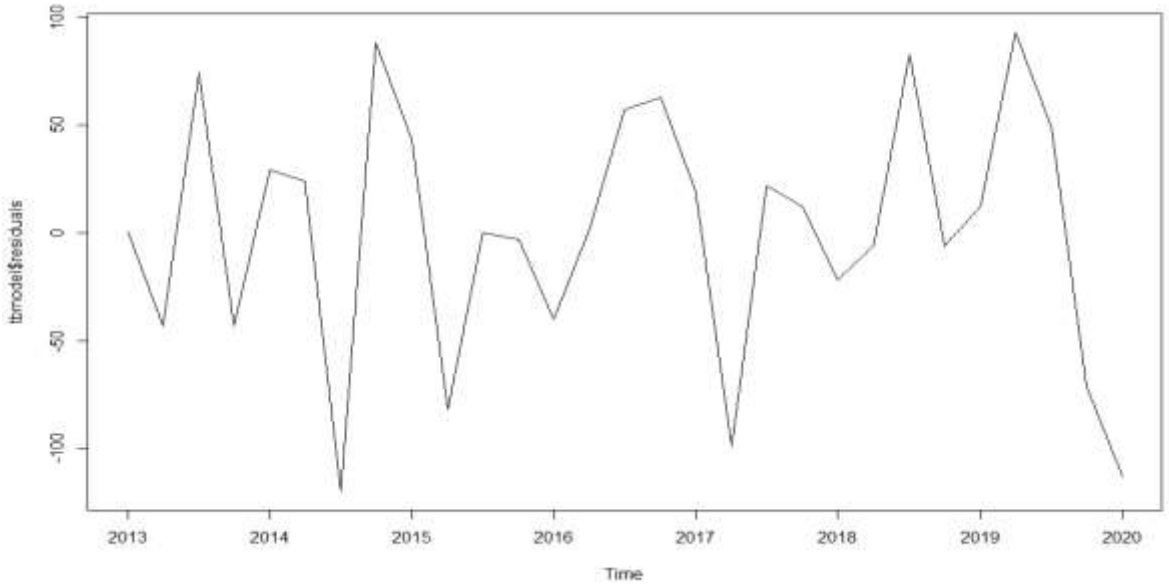

Figure 12: The residual ARIMA (0, 1, 2) graph

After conducting diagnostic checks, including the Ljung-Box test, we determined that that the ARIMA (0, 1, 2) model was the best fit for our data (Sigma squared = 3478, Log likelihood = -153.89, AIC = 309.78 and AICc = 309.93 and BICC = 311.11). To assess the diagnostics, we examined the ACF and PACF plots as shown in figure 21. The ACF and PACF plots indicated that our time series data achieved stationarity. In the ACF plot, I can be observed that all spikes, except for one, fall withing the dashed blue dotted lines, indicating stationarity. Similarly, the PACF plots shows that all spikes are within the confidence band, further confirming the stationarity of our time series data.

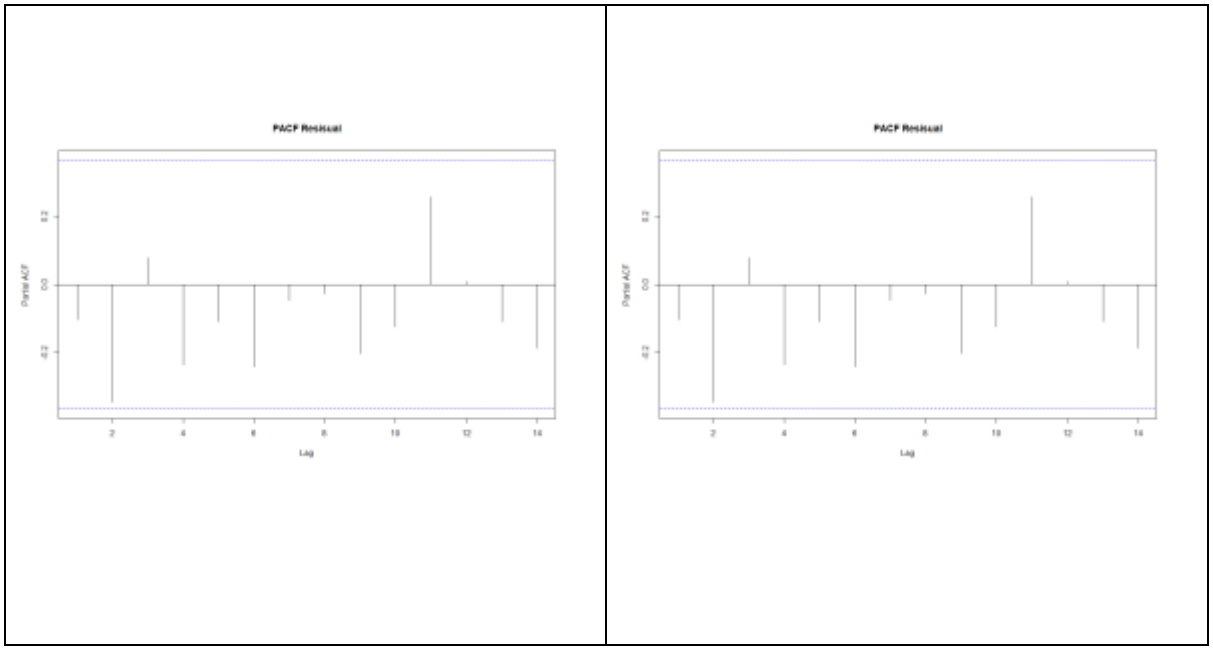

Figure 13: Autocorrelation plots of the residuals from ARIMA (0, 1, 2) model

4.5. Forecasted TB Case Notifications for the Next 12 Seasons

We furthermore compared the four competing models by plotting their graphs which were analysed to see which model best predicts the seasonality of TB case notifications. From the Figure 22 below, it is noted that ARIMA (0, 1, 1) (1, 0, 0) [4] has the best ability to predict future trends in the TB case notifications. The other models fail to qualify because they just show a straight line into the future which could not reflect on reality of the future patterns in the TB case notifications.

Based on the two forecasting graphs below, it is evident that the exponential smoothing method produced forecasts that lagged behind the actual trend. In contrast, the three ARIMA models displayed a straight line of the graph, suggesting that they projected a relatively constant trend for future TB incidences. These models indicate a consistent pattern without significant fluctuation. However, the seasonal ARIMA model, identified as the optimal model, exhibited a wave-like pattern in the forecast area of the graph. This model accounted for seasonal fluctuations and projected a pattern similar to the observed historical data.

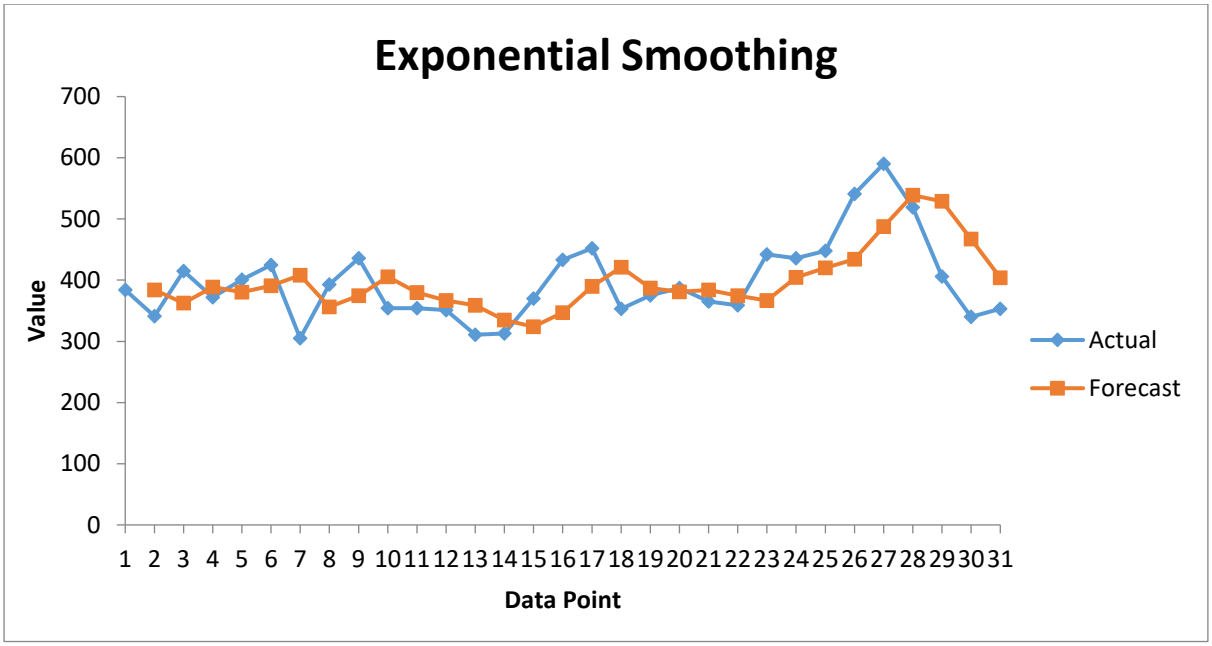

Figure 14: Forecast from the Exponential smoothing method

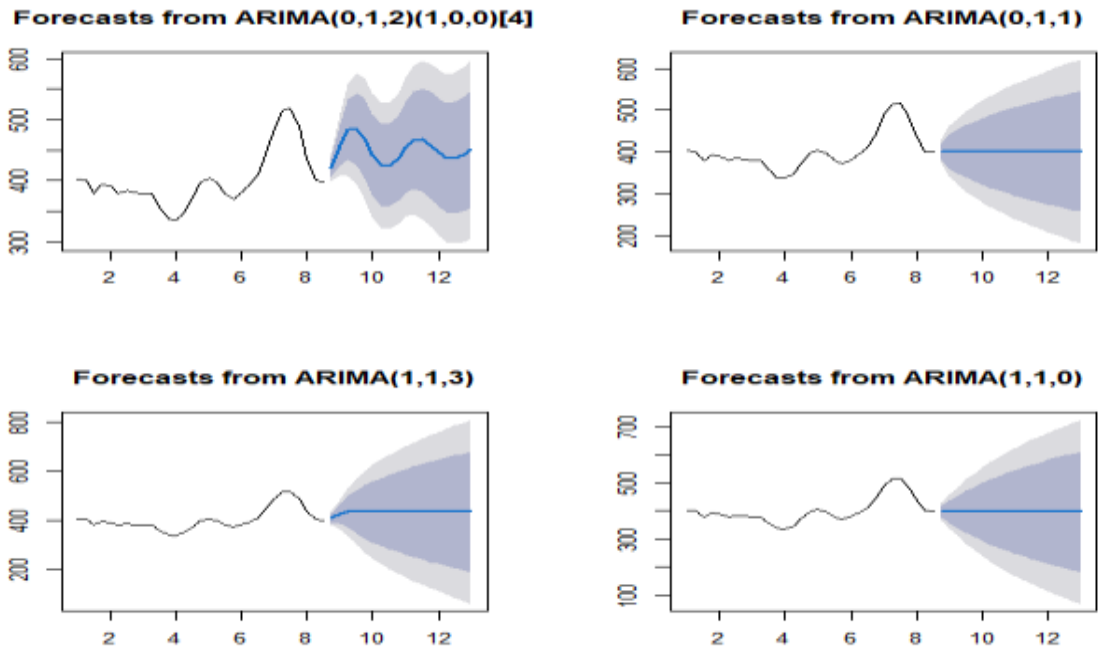

Figure 15: Forecasts from the four competing ARIMA models

We applied the Winter's Multiplicative method to the seasonal ARIMA (0, 1, 2) (1, 0, 0)$_4$ model to predict future trends in TB case notifications in the north health zone of Malawi at 95% confidence interval (CI). The predictions were done for the next 12 quarters (fourth

quarter of 2020 to third quarter of 2027). Figure 5 below shows the graphical presentation of the forecasted TB case notifications for the north health zone in Malawi. The predictions were done on the assumptions that there shall be no any extra interventions done by the government and other stakeholders that would otherwise have an influence on the number of TB case notified in the health zone. We assumed that the current prevailing interventions will continue for the next few years.

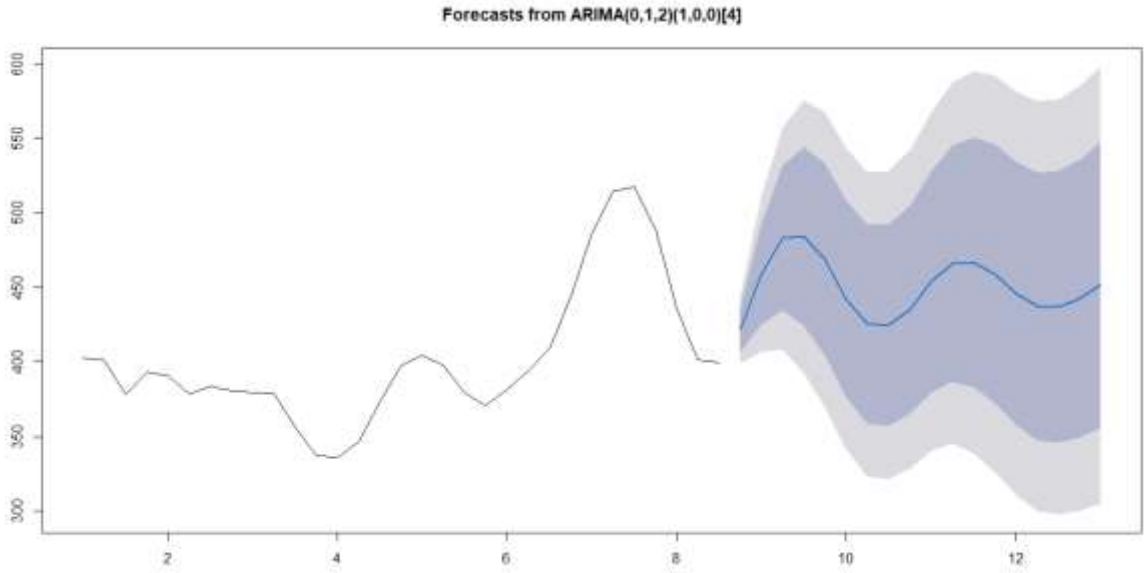

Figure 16: The predicted/forecasted number of TB case notification for the north health zone of Malawi from October 2020 to December 2027

Table 9 below, provides a forecasted trend analysis/ pattern of TB case notification. In summary, the results show that the forecast follow the recent trend in the data with some form of seasonality in it. The rapidly and largely increased prediction intervals indicate that the TB incidence may possibly start increasing or decreasing at any period of time and in contrast, the point forecast tend upwards during the first four quarters of our prediction time and then change its course by following a downwards trend, and the prediction intervals allow for the data to trend upwards and downwards during the forecast period. Other prediction studies with non-time series methods on tuberculosis data have been conducted as well, such as a study in Spain (Rios, Garcia, Sanchez, & Perez, 2000) with mathematical modelling on the registered cases of tuberculosis from 1971 until 1996, which has predicted the pattern for tuberculosis incidence and showed increases in the incidence. The incidence of pulmonary tuberculosis in Iran has a

seasonal trend (Rafei, Pasha, & Jamshidi, 2012) and a study from the Mazandaran Province of Iran has reached similar results (Moosazadeh, et al., 2014).

Table 9: Predicted future seasonal patterns in TB case notification for the north health zone of Malawi from October 2020 to September 2023.

Point of forecast	Predicted number of TB cases notifications	95 % Confidence Interval (CI)	
		Lower value	Upper value
2020 Q4	421.24	398.7291	443.7551
2021 Q1	458.89	406.9944	510.7867
2021 Q2	483.00	408.6401	557.3603
2021 Q3	484.25	392.7887	575.7163
2022 Q4	469.00	370.1925	567.8024
2023 Q1	442.44	341.5747	543.2995
2023 Q2	425.43	323.3534	527.5028
2023 Q3	424.54	321.2719	567.6491
2023 Q4	435.31	328.5908	542.0228
2024 Q1	454.04	340.4404	567.6491
2024 Q2	466.04	344.8585	587.2301
2024 Q3	466.66	338.3474	594.9877
2024 Q4	459.08	326.2534	591.9877
2025 Q1	445.86	310.5567	581.1547
2025 Q2	437.39	300.0671	574.7134
2025 Q3	436.95	297.6326	576.2685
2025 Q4	442.31	299.9889	584.6250
2026 Q1	451.63	305.0566	598.2093
2026 Q2	457.61	306.5083	608.7021
2026 Q3	457.92	302.4292	613.4016
2026 Q4	454.14	295.2389	613.0342
2027 Q1	447.56	286.1218	608.9926
2027 Q2	443.34	279.6036	607.0841
2027 Q3	443.13	277.1119	609.1381
2027 Q4	445.79	277.0619	614.5200

4.5.1. Comparison of Competing Models for Predicting Future Seasonal Patterns in TB Case Notification in the

In our analysis, we compared several competing models to predict future seasonal patterns in TB case notifications in the North Health Zone of Malawi from October 2020 to September 2023. Among the models considered, one model demonstrated a strong fit to the data, capturing the underlying patterns and variations observed in the historical TB case notifications.

However, it is important to note that three other models indicated a different outcome. According to these models, the predicted pattern of TB case notifications would remain constant over the specified time period, without any noticeable changes or seasonal fluctuations. This divergence in the models' predictions highlights the complexity of forecasting and the inherent uncertainty involved in capturing the dynamics of TB case notifications. While one model suggests a dynamic pattern with varying seasonal trends, the other models indicate a more stable and consistent pattern throughout the forecasted period as shown in Table 10 below.

Table 10: Comparison of Competing Models for Predicting Future Seasonal Patterns in TB Case Notification in the North Health Zone of Malawi from October 2020 to September 2023

Point of forecast	ARIMA (0,1,2) $(1,0,0)_4$	ARIMA (1, 1, 0)	ARIMA (1, 1, 3)	ARIMA (1, 1, 1)
2020 Q4	421.24	398.52	406.69	400.92
2021 Q1	458.89	397.83	423.92	400.92
2021 Q2	483.00	397.40	433.52	400.92
2021 Q3	484.25	397.14	433.03	400.92
2022 Q4	469.00	396.98	433.06	400.92
2023 Q1	442.44	396.77	433.06	400.92
2023 Q2	425.43	396.75	433.06	400.92
2023 Q3	424.54	396.73	433.06	400.92
2023 Q4	435.31	396.72	433.06	400.92
2024 Q1	454.04	396.72	433.06	400.92
2024 Q2	466.04	396.72	433.06	400.92
2024 Q3	466.66	396.72	433.06	400.92
2024 Q4	459.08	396.72	433.06	400.92
2025 Q1	445.86	396.72	433.06	400.92
2025 Q2	437.39	396.72	433.06	400.92
2025 Q3	436.95	396.72	433.06	400.92
2025 Q4	442.31	396.72	433.06	400.92
2026 Q1	451.63	396.72	433.06	400.92
2026 Q2	457.61	396.72	433.06	400.92
2026 Q3	457.92	396.72	433.06	400.92
2026 Q4	454.14	396.72	433.06	400.92
2027 Q1	447.56	396.72	433.06	400.92
2027 Q2	443.34	396.72	433.06	400.92
2027 Q3	443.13	396.72	433.06	400.92
2027 Q4	445.79	396.72	433.06	400.92

CHAPTER FIVE

DISCUSSION OF THE RESULTS

This chapter discusses findings of the study. It compares the findings of the study to those findings from previous studies, while noting interesting findings, agreements, contradictions or inconsistencies.

5.1 The Most Suitable Model to Predict Future Trends in TB

A seasonal ARIMA model was developed to forecast the future quarterly incidence of TB cases in north health zone of Malawi. The results of our analysis indicate that the SARIMA (0, 1, 2) $(1, 0, 0)_4$ model provides superior forecasts for the TB data. We evaluated the model's performance using metrics such as AIC, AICc, BIC, and by examining the residuals for white noise pattern. In a related study conducted by Permanasari *et al*., they evaluated the performance of six different forecasting methods, including linear regression, moving average, decomposition, ARIMA, Neural Network and Holt-Winter's, for monthly tuberculosis data prediction. The study concluded that the most suitable model of their data was the ARIMA model (Permanasari, Rambli, & Dominic, 2011).

In another study by Zhang et al., two models, ARIMA and (GRNN)-ARIMA, were investigated for predicting tuberculosis incidence. The time series of tuberculosis exhibited a gradual secular decline and a striking seasonal variation. Among several plausible ARIMA models, the ARIMA (2, 1, 0) × (0, 1, 1) [12] model was selected. The author reported that the hybrid model had lower mean absolute error and mean absolute percentage error compared to the ARIMA model (Zhang, et al., 2013). The above studies demonstrate the effectiveness of ARIMA models in forecasting TB incidence and highlight the importance of selecting appropriate models to capture the underlying patterns and dynamics of the data.

The ARIMA model assumes that there is a certain relationship between the future state of the target object and the historical data of the past and the present (Yang, Duan, Wang, Zhang, & Jiang, 2014). According to the seasonal fluctuations of the target sequence, the ARIMA model can be divided into a seasonal model or a non-seasonal model. This model overcomes the limitation of the requirement for a prior assumption about the development mode of the time series. The process of identification, estimation, and diagnosis is repeated until the optimized model is obtained (Box & Jenkins, 1976). The ARIMA model is widely used in many types of time series analysis and is by far the most versatile time series prediction method. Anwar et al used the ARIMA (4,1,1) $(1,0,1)_{12}$ model to predict future malaria incidence in Afghanistan. Li et al used the ARIMA (0,1,1) $(2,1,0)_{12}$ model to forecast the incidence of haemorrhagic fever with renal syndrome in Hebei Province, China. Mahmood et al used the ARIMA (0,1,1) $(0,1,1)_{12}$ model to predict the incidence of smear-positive TB cases in Iran (Moosazadeh, Khanjani, Nasehi, & Bahrampour, 2015). However, the ARIMA model is only suitable for a short-term prediction and can only capture the linear relationship in the incidence trend. As the occurrence of TB is affected by many known and unknown factors, the incidence trend tends to exhibit nonlinear characteristics, which cannot be effectively solved through the ARIMA model.

The Box-Jenkins approach described and implemented in this study is well-established and widely used method in time series analysis. It provides a solid foundation for understanding the autoregressive, moving average, and differencing components of the time series data. However, it is important to acknowledge that there are alternative approaches, such as generalised additive models and their extensions that could have been explored to incorporate seasonality.

5.2 Pattern in TB Case Notification in North Health Zone

A detected pattern of peaks and troughs in TB case notifications in our study is consistent with the pattern detected in other studies for TB in different countries (Khaliq, Syeda, & Chaudhry, 2015; Liu, Luan, Yin, Zhu, & Lu, 2016; Yang, Duan, Wang, Zhang, & Jiang, 2014; and Bras, Gomes, Filipe, de Sousa, & Nunes, 2014). First quarter (rainy season) was the dominant quarter seconded by third quarter. Fourth quarter has the least number of TB cases during the study

period. We can postulate that high cases of TB reported during the third quarter, which lies between the cold season and the hot season. It is possible that this was so because of a number of reasons; Firstly, poor ventilated rooms crowded with people could increase the chances of transmission among the infections source and the contractors in cold seasons. In the north health zone of Malawi, cold season is from May to August, and the coldest months are June and July in most of the year. During this cold period, people are more likely to stay indoors and close the windows because of low temperatures. We thus suggest that TB transmission is rampant during this period due to lack of fresh air.

Secondly, it has been observed that there are delays in seeking medical healthcare during the coldest months, specifically in June and July. (Yang, Duan, Wang, Zhang, & Jiang, 2014) suggested that individuals may be less inclined to seek medical help when the weather is very cold unless absolutely necessary. Consequently, people tend to seek medical assistance more readily when the weather becomes somewhat warmer, typically between August and September. During this warmer period, the consultation rate between patients and healthcare providers tends to be relatively high.

According to (Yang, et al., 2014) delays in seeking healthcare contribute to diagnostic delays, which can increase the risk of disease transmission due to the extended communicable period of tuberculosis (TB) during cold seasons. Moreover, it is believed that in winter, symptoms of TB may not immediately manifest after infection. However, as summer approaches and the temperature increases, the bacterium starts to proliferate and grow, leading to the onset of noticeable symptoms of the infection (Smith, 2003).

Our study further discovered that the number of TB cases was high during the first quarter of the year which coincides with the rainy season in the zone. Similar findings were observed in Cameroon (Anyangwe, et al., 2006) and Korea (Choi, Seo, et al., 2013), where TB cases increased during the rainy season. The authors attributed it to two potential factors: vitamin D deficiency due to reduced sunlight exposure and higher risks of indoor infection due to humid and cold weather. Combined with high incidence of other seasonal respiratory infections, these factors may result in worsening symptoms, potentially leading to a peak in TB notifications during the rainy season in January to March, first quarter (Murray, 2012).

5.3 Forecasted Incidence of TB Case Notification

The study observed is a significant declining trend in TB case notification starting from the fourth quarter of 2022, which is likely associated with the scale-up of antiretroviral therapy (ART) and increased access to ART services. The high prevalence of HIV in Malawi suggests that TB incidence in the HIV-positive population plays a crucial role in community TB prevalence and transmission (Lawn, Kranzer, & Wood, 2009). The study suggests that the reduction in TB risk due to the "ART protective effect" in this susceptible population may contribute to the declining trend in TB incidence and case notification. Factors such as high mortality prior to TB case detection, improving socio-economic status, and isoniazid preventive treatment for HIV-positive individuals may confound the observed trend (Lawn, Kranzer, & Wood, 2009).

Results of the present study match a study in Spain that also predicted increases in TB cases and emphasizes the suitability of time series analysis models, particularly the seasonal ARIMA model, for examining trends and predicting TB incidence (Kam, Sung, & Park, 2010). However, the study's forecasts suggest that there will be no apparent improvement in the high burden of TB in the near future in the north health zone of Malawi. This contradicts global trends reported by the WHO, which indicate a decline in TB burden worldwide.

In this study, the forecasts show that there will be no apparent improvement in the high burden incidence of TB in the north health zone of Malawi in the near future. The overall predicted outcomes indicate that the reported quarterly TB incidence cases will slightly increase in the nearest future in the north health zone. Nevertheless, our findings are inconsistent with WHO (2018i) which highlighted that the disease burden caused by TB is falling globally, in all WHO regions, and in most countries, but not fast enough to reach the first (2020) milestones of the End TB strategy. Our findings revealed that progress in TB control in the north health zone needs to be more intensified and adequate interventions (e.g., the introduction of new vaccine, advanced diagnostic techniques, etc) are urgently needed by the government of Malawi through the MoH.

The apparent reason to a continued high incidence of TB in the north health zone is due to an improved case detection rate (CDR) in the zone during the most recent past years. The other reason could be the improvement in recording and reporting of detected TB cases following

the introduction of DOTS without a real increase in TB case detection rate (Obermeyer, Abbott-Klafter, Christopher, & Murray, 2008). Nevertheless, the increased trend might also be due to a true increase in TB incidence cases fuelled by the powerful interaction between HIV and tuberculosis (Corbett, et al., 2003). The TB and HIV co-infection among tested TB patients in our study was 43.4%. It might also be due to the notification of large backlog of TB cases that resulted from improved TB diagnostic access in the health zone.

5.4 Effects of Social-Demographic Factors on TB Case Notification

In this study, it was observed that there were more reported cases of tuberculosis (TB) among males than females across all age groups, with a female-to-male ratio of 1.57. The reasons behind this higher TB case notification in males compared to females in the north health zone of Malawi are not fully understood. However, some potential factors contributing to this disparity could be the time taken before seeking medical care and differences in access to healthcare. Men may be slower to report or get diagnosed with TB due to work or other considerations leading to delay in seeking medical assistance. Several studies have attempted to explain the differential TB infection rates between men and women, focusing on biological factors. Some studies suggest that men may be biologically more vulnerable to pulmonary TB (Neyrolles & Quintana-Murci, 2009). This finding is consistent with other studies conducted in Bangladesh, Malawi, and South Africa, which argue that TB is more challenging to diagnose in women (Begum, et al., 2001; Boeree & Harries, 2000; Austin, et al., 2004).

Research indicates that women with pulmonary TB may exhibit a different immune response to the disease compared to men (Long, 2001), resulting in different symptoms, signs, and outcomes. This can make it harder to detect TB in women, as they may not test positive on microscopic examination of sputum. Additionally, one study found that TB lung lesions might be less severe in women compared to men, potentially leading to milder symptoms and more challenging diagnosis in women (Long, 2002).

This study also observed that TB rates are higher amongst males than females. This finding contradicts to a study done in regions of Pakistan bordering Afghanistan where more women than men were detected with TB (WHO, Tuberculosis in Women Factsheet, 2014). However, the reasons for higher TB rates among women in these regions are poorly understood. Although the Afghanistan National Strategic Plan for Tuberculosis Control attributes these rates to early

marriage and short intervals between pregnancies (Islamic Republic of Afghanistan, 2013), the lack of comparative data from countries with similar early marriage and high birth rates makes it difficult to determine whether this explanation holds true. In countries with high HIV prevalence, the numbers of women notified with TB are exceeding those of men.

Our present study confirmed the discovery done by other researchers that people living with HIV are more likely to have TB as compared to people without HIV. This finding is comparable to that of a study done in Ethiopia (Tesfaye, et al., 2018). But this finding is not in line with a study from Sub-Saharan Africa (Nagua, Aboud, & Mwiru, 2017). The difference might be due to different levels of CD4 count and advanced WHO clinical stage that may determine the immunity of individuals living with HIV. With a low CD4 count (below 200), a person's body becomes vulnerable to opportunistic infections such as the TB. Those HIV patients with a CD4 count of above 200 might have some form of immunity to fight off infectious diseases than their counterpart.

5.5 The influence of Setting on TB Case Notifications

The study identified district variations in TB case notification rates. Mzimba district consistently had the highest TB case notification rates throughout the study period. The cold weather in the district, along with better diagnostic capacity, was suggested as potential factors contributing to the high rates. Interestingly, this study identified living in urban areas as a factor associated with TB, which contradicts the results from similar studies conducted in Pakistan (Khaliq, Syeda, & Chaudhry, 2015) and Ethiopia (Yeshi et al., 2018). The heterogeneity of TB burden across districts emphasizes the need for targeted interventions and tailored strategies to address the specific challenges faced by each district.

If the inequality in TB case notification is indeed a result of limited access to healthcare, it may hinder the effectiveness of the DOTS strategy in achieving the global targets set by the World Health Organization (WHO) of a 70% TB case detection rate and an 85% treatment success rate. These targets aim to interrupt TB transmission, reduce mortality, and prevent the emergence of drug resistance (Keshavje & Farmer, 2012). Therefore, addressing the underlying factors contributing to the inequality in TB case notification is crucial to ensure the success of TB control efforts in the north health zone and similar settings.

5.6 The Choice for ARIMA Models

In the study, the decision to utilize ARIMA models was justified based on several considerations, taking into account the availability of alternative modelling approaches such as non-linear regime-changing models and generalized linear models.

Firstly, ARIMA models are widely recognized and extensively used in time series analysis due to their ability to capture and forecast linear dependencies and trends within the data. They have a solid theoretical foundation and have proven to be effective in capturing the autocorrelation and seasonality patterns often observed in time series data (Anwar M. , Lewnard, Parikh, & Pitzer, 2016). By incorporating autoregressive (AR), differencing (I), and moving average (MA) components, ARIMA models can adequately capture the temporal dynamics of the data and make accurate predictions.

Secondly, while non-linear regime-changing models and generalized linear models have their merits, they may introduce additional complexity and assumptions that may not be appropriate for the specific characteristics of the data under study. Non-linear regime-changing models are useful for capturing abrupt changes or shifts in the underlying data-generating process, but their implementation requires identifying and estimating specific breakpoints or thresholds, which may not be well-suited for all datasets (Goutee, Ismail, & Pham, 2017). Similarly, according to (Chuang, Mazumdar, Park, Tang, & Nicolich, 2011) generalized linear models are valuable when dealing with non-normal or non-Gaussian data, but they may not adequately capture the sequential dependencies and temporal patterns present in time series data. GLMs with time factors can provide valuable insights into how count outcomes change over time, identify significant temporal patterns, and examine the impact of time-related predictors on the response variable (Arnold, Davies, Mbotwa, & Gilthorpe, 2020).

Considering the objectives of the study and the nature of the data, the simplicity and interpretability of ARIMA models, coupled with their ability to capture linear dependencies and trends, were deemed sufficient for the analysis. It was crucial to choose a modelling approach that aligned well with the specific characteristics and objectives of our study, and in this case, the ARIMA models were considered a suitable choice for effectively capturing and forecasting the temporal dynamics observed in the time series data.

CHAPTER SIX

CONCLUSIONS, RECOMMENDATIONS, LIMITATIONS AND FUTURE DIRECTION OF RESEARCH

This chapter summarises the findings of the study. The first section presents conclusions drawn from the findings; the second section provides some limitations to the study; the third section makes some recommendations, while the last one provides future direction of the research.

6.1 Conclusions

The following constitutes the study conclusions;

1) In this study it has been observed that across all the age groups, more cases were reported among males than females TB patients (F: M ratio 1.57). At present, it is not fully understood why TB case notification has been observed to be higher in males than females in the north health zone of Malawi, but we can only speculate the reasons why. However, this is an important epidemiological finding from the point of view of public healthcare.

2) Based on the study design and data orientation, the seasonal ARIMA model is the most appropriate model in predicting pattern in TB case notification, this is confirmed from the goodness-of-fit diagnostic's results which confirmed the appropriateness of the ARIMA model. In addition, the selected forecast model, SARIMA (0, 1, 2) $(1, 0, 0)_4$, satisfies all necessary assumptions (no serial correlation, constant variance and normality) and is better in all aspects than the other comparable models which have spikes at both the ACF and PACF plots. Therefore, having satisfied all model assumptions, ARIMA (0, 1, 2) $(1, 0, 0)_4$ can be regarded as the best-fitted model for forecasting quarterly TB case notification in north health zone of Malawi.

3) The results from this study also indicate that there was a cyclic pattern in the TB case notification in the zone with peaks during the rainy season and at the end of the cold season.

4) The results further show that number of TB case notification will follow a seasonal pattern for the next three years with an increasing trend as indicated by the forecasted number of TB case notifications.

6.2 Recommendations

The key recommendations based on the research findings and conclusions include the following;

1) While the research results conclude that the TB prevalence in the north health zone of Malawi will not increase remarkable in the forthcoming years, there is a high probability that TB case notification is on the rise in the north health zone of Malawi and it will continue rising. This being the case, it is highly recommended that the government through the MoH put in some control measures/strategies in the north health zone focusing on minimising the increase of TB cases in the study area. We recommend that in both rural and urban settings/areas, clinicians need to educate people on health issues. In addition to that we recommend that healthcare facilities should be improved for timely diagnosis, treatment and prevention of the disease.

2) We suggest that the results from this study can be used to plan service needs in the north zone and in Malawi as general by anticipating higher service use during the rainy season (January to March) and period after the cold season (July to September) – as high cases were recorded during these two periods. Local health services can also use these data to address potential service issues which contribute to delayed diagnosis of TB, for example after the cold season, or due to changes in clinical service provision at certain times of the year.

3) HIV/TB collaboration initiatives should be intensified in the north health zone in order to reduce TB-related morbidity and mortality.

4) For (1) (3) to materialize, the Malawi government and its partners (especially those in non-governmental sector) should provide adequate funding for TB surveillance and control programmes. Without such funding, the north health zone will continue to suffer from the TB scourge and yet there is room for reducing it significantly if adequate funds are availed.

5) Further research is necessary to explore alternative methodologies and incorporate additional predictors to improve the accuracy and reliability of TB case notification analysis and forecasting.

6.3 Limitations of the Study

The following are concluded to the research limitations that may be looked into in further analysis of this study;

1) Malawi has three distinct seasons of cool dry (May to August), hot dry (September to November) and hot wet season (December to April) seasons. These seasonal patterns also apply to the north health zone of Malawi. However, the reporting of TB case notification data in the study area is done on a quarterly basis, which does not align with the weather seasons of the region. This overlap of quarterly reporting with weather season makes it challenging to observe clear and distinct seasonal patterns in the original TB case notification data.

2) In our study, we primarily used conventional forecasting methods and focused on predicting the incidence of an infectious disease within the available data. However, to improve prediction accuracy and broaden our analysis, it is crucial to explore alternative models such as time-varying models and generalized linear models. Additionally, incorporating out-of-sample predictions can offer valuable insights and enhance the accuracy of predictions for a wider range of infectious diseases. Further investigation into these alternative approaches would be valuable for advancing our understanding and forecasting capabilities in the field of infectious diseases.

3) Furthermore, incorporating additional predictors such as demographics, climatic conditions, and environmental factors, could further enhance the predictive capacity of

the model and provide a more comprehensive analysis of the underlying patterns. The diverse geographical and climatic conditions in Malawi may have an effect in the progression of TB during different seasons across various geographic districts and regions and as such, it is important to note that the findings of this study cannot be extrapolated to the entire nation as a whole.

4) Additionally, Quarterly population sizes for the health zone were not available to compare the quarterly TB case notification rates and quarterly incidence of TB cases. This data could have helped us compare the peaks and troughs of both the TB case notification rate and the seasonality of TB incidences.

The identified limitations present potential avenues for future research in the field of TB case notification. Firstly, investigating the impact of departures from normality, tail properties, volatility clustering, and the influence of time-varying predictors on TB case notification could provide valuable insights. Future studies can explore robust techniques or transformations to address non-normality and develop models that appropriately account for heavy or fat tails in the data distribution. Furthermore, incorporating specialized models like GARCH that capture volatility clustering could enhance forecasting accuracy and risk management strategies in the context of TB.

Secondly, the dynamic effects of time-varying predictors on TB case notification warrant further investigation. Future research could focus on developing models that effectively capture the nonlinearity and dynamic nature of these predictors. Such models are the generalized additive models (GAMs) which would capture non-linear and non-parametric relationships between predictors and the response variable. Nevertheless, our study aimed and predicting future trends in TB case notifications in the north health zone of Malawi hence, the GAMs models were not implored in this study.

By addressing these limitations in future research, we can improve the rigor and reliability of TB case notification analysis, leading to more accurate insights and informing targeted interventions and policies. Ultimately, these advancements could also contribute to more effective TB control and prevention efforts.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control, AC-19*, 716-723.

Anwar, M. Y., Lewnard, J. A., Parikh, S., & Pitzer, V. E. (2016). Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trendsin incidence. *Malar J, 15*(1), 566.

Anwar, M., Lewnard, J., Parikh, S., & Pitzer, V. (2016). Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. *Malaria Journal 15, 15*. doi:https://doi.org/10.1186/s12936-016-1602-1

Anyangwe, A. I., Akenji, T. N., Mbacham, W. F., Penlap, V. N., & Titanji, V. P. (2006). Seasonal variation and prevalence of tuberculosis among health seekers in the South Western Cameroon. *East Afr Med J, 83*, 588 – 595.

Arnold, K., Davies, V., Mbotwa, J., & Gilthorpe, M. (2020). Reflection on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *International Journal of Epidemiology*, 2074–2082. doi:https://doi.org/10.1093/ije/dyaa049

Austin, J. F., Dick, J. M., & et al. (2004). Gender Disparity Amongst TB Suspects and New TB Patients According to Data Recorded at the South African Institute of Medical Research Laboratory for the Western Cape Region of South Africa. *International Journal of Tuberculosis and Lung Disease, 8*(4), 435 – 439.

Azeez, A., Obaromi, D., Odeyemi, A., Ndege, J., & Muntabayi, R. (2016). Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape South Africa, using a hybrid model. *Int J Environ Res Public Health, 13*.

Begum, V., Colombani, P., & Das Gupta, S. (2001). Tuberculosis and patient gender in Bangladesh: sex differences in diagnosis and treatment outcome. *Int J Tuberc Lung Dis, 5*, 604 - 670.

Bodena, D., Ataro, Z., & Tesfa, T. (2019). Trend Analysis and Seasonality Of Tuberculosis Among Patients At The Hiwot Fana Specialized University Hospital, Eastern Ethiopia: A Retrospective Study. Risk management and healthcare policy. *Risk Management and health care policy, 12*, 297–305. doi:https://doi.org/10.2147/RMHP.S228659

Boeree, M. J., & Harries, A. (2000). Gender Differences in Relation to Sputum Submission and Smear Positive Pulmonary Tuberculosis in Malawi. *International Journal of Tuberculosis and Lung Disease, 4*(9), 882 – 884.

Boeree, M. J., Harries, A. D., & Godschalk, P. (2000). Gender differences and rates of sputum submission and smear- positive pulmonary tuberculosis in Malawi. *Int J Tuberc Lung Dis*, 882–884.

Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control.* SanFrancisco: Holden Day.

Box, G. E., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control* (2nd ed.). San Francisco: Holden-Day.

Bras, A. L., Gomes, D., Filipe, P. A., de Sousa, B., & Nunes, C. (2014). Trends, seasonality and forecasts of pulmonary tuberculosis in Portugal. . *Int J Tuberc Lung Dis, 10*(18), 1202–1210. doi:https://doi.org/10.5588/ijtld.14.0158

Brockwell, P., & Davis, R. (2002). *Introduction to Time Series and Forecasting (https://book.* Springer. Retrieved September 12, 2020, from https://books.google.com/books?id=VHB4OSAmwcU&pg=PA35

Choi, C. J., Seo, M., Choi, W. S., Kim, K. S., & Youn, S. A. (2013). Relationship between serum 25-hydroxyvitamin D and lung function among Korean adults in Korea National Health and Nutrition Examination Survey (KNHANES),. *J Clin Endocrinol Metab, 98,* 1703 - 1710.

Chuang, Y., Mazumdar, S., Park, T., Tang, G., & Nicolich, J. (2011). Generalized linear mixed models in time series studies of air pollution. *Atmospheric Polution Research,* 428 - 435.

Cleveland, W. P., & Tiao, G. (1976). Decomposition of seasonal time series: a model for the census X-11 program. *J Am Stat Asso*, 581-587.

Corbett E, L., Watt, C. J., Walker, N., Maher, D., & Williams, B. G. (2003). The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med, 163*, 1009 - 1021.

Crampin, A. C., & Glynn, J. R. (2004). Tuberculosis and gender: exploring the patterns in a case control study in Malawi. *Int J Tuberc Lung Dis*, 194–203.

Cui, W., & George, E. (2008). Empirical Bayes vs. Fully Bayes variable selection. *Journal of Statistical Planning and Inference, 4*(138), 888-900.

Dara, M., Acosta, C. D., Melchers, N. ,., Al-Darraji, H. A., Chorgoliani, D., Reyes, H., . . . Migliori, G. B. (2015). Tuberculosis control in prisons: current situation and research gaps. *Int J Infect Dis*, 111-117.

Fares, A. (2011). Seasonality of tuberculosis. *J Glob Infect Dis, 3*(1), 46–55. doi:https://doi.org/10.4103/0974-777X

Frah, E. A., & Alkhalifa, A. (2016). Tuberculosis Cases in Sudan; Forecasting Incidents 2014 - 2023 Using Box-Jenkins ARIMA Model. *American Journal of Mathematics and Statistics, 6*(3), 108 - 114. Retrieved from http://dx.doi.org/10.5923/j.ajms.20160603.04

Fuller, W. A. (1976). *Introduction to Statistical Time Series.* New York: John Wiley and Sons.

Gashu, Z. J. (2018). Seasonal patterns of tuberculosis case notification in the tropics of Africa: A six-year trend analysis in Ethiopia. *11*(13). doi:https://doi.org/10.1371/journal.pone.0207552

Goutee, S., Ismail, A., & Pham, H. (2017). Regime-switching stochastic volatility model: estimation and calibration to VIX options. *Applied Mathematical Finance*, 38 - 75.

Halim, S., Intan, R., & Dewi, L. P. (2019). Fuzzy linear reression for tuberculosis case notification rate prediction in Surabaya. Proceedings of the International Conference on Advanced Informattion Science and System. 1 - 5. doi:https://doi.org/10.1145/3373477.3373492

He, Z., & Tao, H. (2018). Epidemiology and ARIMA model of positive-rate ofinfluenza viruses among children in Wuhan, China: a nine-year retrospectivestudy. *Int J Infect Dis, 74*, 61 - 70.

Höge, M., Wöhling, T., & Nowak, W. (2018). A Primer for Model Selection: The Decisive Role of Model Complexity. *54*(3), 1688-1715. doi:https://doi.org/10.1002/2017WR021902

Holmes, C. B., Housler, H., & Nunn, P. (1998). A review of sex differences in the Epidemiology of tuberculosis. *Int J Tuberc Lung Dis*, 96–104.

Kam, H. J., Sung, J. O., & Park, R. W. (2010). Prediction of Daily Patient Numbers for a Regional Emergency Medical Centre using Time Series Analysis. *Healthcare Inform Res, 16*(3), 158 - 165.

Kanyerere, T. &. (2005). Delays in TB hospital diagnosis – a major threat for the HIV/AIDS situation in a society: A study of TB as an opportunistic infection in southern Malawi. *Norwegian Journal of Geography*, 55 - 64. doi:10.1080/00291950510020529

Kemp, J., Boxshall, M., Nhlema, B., Salaniponi, F. M., & Squire, S. B. (2006, February 20). Is there a relationship between poverty and lack of access to the TB DOTS programme in urban Malawi? 580-585. Lilongwe, Malawi, Malawi. Retrieved from http://www.equi-tb.org.uk/docs

Keshavje, S., & Farmer, P. E. (2012). Tuberculosis Drug Resistance and the History of Modern Medicine. *NEJ M , 367*, 931 – 936.

Khaliq, A., Syeda, A. B., & Chaudhry, M. N. (2015). Seasonality and trend analysis of tuberculosis in Lahore, Pakistan from 2006 to 2013. *J Epidemiol Glob Health*, 397 – 403.

Kirolos, A., Thindwa, D., Khundi, M., Burke, R. M., Henrion, M. Y., Nakamura, R., . . . MacPherson, P. (2021). Tuberculosis case notifications in Malawi have strong seasonal and weather-related trends. *Sci. Rep*. doi:https://doi.org/10.1038/s41598-021-84124-w

Koehler, A., & Murphree, E. (2008). A Comparison of the Akaike and Schwarz Criteria for Selecting Model Order. *Journal of the Royal Statistical Society Series, 37*, 187–195.

Lawn, S. D., Kranzer, K., & Wood, R. (2009). Antiretroviral therapy for control of the HIV-associated tuberculosis epidemic in resource limited settings. *Clin Chest Med, 30*, 685 – 699.

Leeb, h., & Pötscher, B. M. (2005). Model Selection and Inference. *Facts and Fiction [J]*, 21 - 59.

Li, Q., Guo, N. N., & Han, Z. Y. (2012). Application of an autoregressiveintegrated moving average model for predicting the incidence ofhemorrhagic fever with renal syndrome. *Am J Trop Med Hyg*, 364 – 370. doi:doi:10.4269/ajtmh.2012.11-0472

Lin, H. H., Ezzati, M., Chang, H. Y., & Murray, M. (2009). Association between tobacco smoking and active tuberculosis in Taiwan Prospective cohort study. *Am J Respir Crit Care Med*, 475 - 480.

Liu, L., Luan, R. S., Yin, F., Zhu, X. P., & Lu, Q. (2016). Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model. *Epidemiol Infect, 144*(1), 144–151. doi:doi:10.1017/S095026881500114

Liu, L., Zhao, X. Q., & Zhou, Y. (2010). A tuberculosis model with seasonality. *BullMath Biol*, 931–952. doi:doi:10.1007/s11538-009-9477-8

Long, N. (2002). Difference in Symptoms Suggesting Pulmonary Tuberculosis Among Men and Women. *Journal of Clinical Epidemiology, 55*(2), 115 – 120.

Long, N. E. (2001). Fear and Social Isolation as Consequences of Tuberculosis in Vietnam: A Gender Analysis. *the Lancet, 58*(1), 69 - 81.

Millet, J., Moreno, A., Fina, L., del Baño, L., Orcau, A., de Olalla, P., & Caylà, J. (2013). Factors that influence current tuberculosis epidemiology. *European spine journal, 22*(4), 539 - 548. Retrieved from https://doi.org/10.1007/s00586-012-2334-8

Millet, J., Moreno, A., Fina, L., del Baño, L., Orcau, A., de Olalla, P., & Caylà, J. (2013). Factors that influence current tuberculosis epidemiology. *European spine journal, 4*(22), 539-548. doi:https://doi.org/10.1007/s00586-012-2334-8

Moghram, I., & Rahman, S. (1989). Analysis and evaluation of five short-term load forecasting techniques. 1484–1491.

MoH. (2015). *Malawi Standard Treatment Guidelines.* Lilongwe: Ministry of Health.

Moosazadeh, M., & Amiresmaili, M. (2018). Challenges in case finding of tuberculosis control program in Iran: A qualitative study. *Bangladesh Journal of Medical Science, 17*(3), 462-469. doi:https://doi.org/10.3329/bjms.v17i3.37002

Moosazadeh, M., Khanjani, N., Bahrampour, A., & Nasehi, M. (2014). Does tuberculosis have a seasonal pattern among migrant population entering Iran? *PMID, 2*(4), 181-185. doi:https://doi.org/10.15171/ijhpm.2014.43

Moosazadeh, M., Khanjani, N., Nasehi, M., & Bahrampour, A. (2015). A. Predicting the incidence of smear positive tuberculosis cases in Iran using time series analysis. *Public Health, 44*, 1526–1534.

Moosazadeh, M., Nasehi, M., Bahrampour, A., Khanjani, N., Sharafi, S., & Ahmadi, S. (2014). Forecasting tuberculosis incidence in iran using box-jenkins models. Retrieved from https://doi.org/10.5812/ircmj.11779

Murray, E. L. (2012). Rainfall, household crowding, and acute respiratory infections in the tropics. *Epidemiol. Infect, 140*, 78 – 86. doi:https ://doi.org/10.1017/s0950268811000252

Nagayama, N., & Ohmori, M. (2006). Seasonality in various forms of tuberculosis. *nt J Tuberc Lung Dis, 10*, 1117-1122.

Nagua, T. J., Aboud, S., & Mwiru, R. (2017). Tuberculosis associated mortality in a prospective cohort in Sub Saharan Africa: association with HIV and antiretroviral therapy. *Int J Infect Dis, 2017;2017*(56), 39 - 44. doi:doi:10.1016/j.ijid

National Statistical Office. (2018, November 19). *2018 Malawi Population and Housing Census.* Retrieved from National Statistical Office: http://www.nsomalawi.mw/images/stories/data_on_line/demography/census_2018/2018%20Population%20and%20Housing%20Census%20Preliminary%20Report.pdf

Neyrolles, O., & Quintana-Murci, L. (2009). Sexual Inequality in Tuberculosis. *Plos Medicine*. Retrieved from http://www.plosmedicine.org/article/info%3adoi%2f10.1371%2fjournal.pmed.1000199#pmed.1000199-who1.

Nyirenda, T. (2006). Epidemiology of Tuberculosis in Malawi. *Malawi Med J, 18*(3), 147–159.

Obermeyer, Z., Abbott-Klafter, J., Christopher, J., & Murray, L. (2008). Has the DOTS Strategy Improved Case Finding or Treatment Success? An Empirical Assessment. *PLoS ONE, 3*(3), 1721 - 1727. Retrieved from https://www.plosone.org.

Ongbali, S. O., Igboanugo, A. C., Afolalu, S. A., Udo, M. O., & Okokpujie, I. P. (2018). Model Selection Process in Time Series Analysis of Production System with Random Output. *Journal of the American Statistical Association*, 199 - 209. doi:doi:10.1088/1757-899X/413/1/012057

Ongbali, S., Igboanugo, A., Afolalu, A., Udo, M., & Okokpujie, I. (2018). Model Selection Process in Time Series Analysis of Production System with Random Output. *IOP Conference Series: Materials Science and Engineering*.

Ottmani, S., Obermeyer, Z., Bencheikh, N., & Mahjour, J. (2021). Knowledge, attitudes and beliefs about tuberculosis in Urban Morocco. *East Mediterr Health J. 2008, 14*(2), pp. 298 - 304.

Permanasari, A. E., Rambli, D. R., & Dominic, P. D. (2011). Performance of Univariate Forecasting on Seasonal Diseases: The Case of Tuberculosis. *Adv Exp Med Biol, Adv Exp Med Biol*, 1791- 179.

Rafei, A., Pasha, E., & Jamshidi, O. R. (2012). Tuberculosis surveillance using a hidden markov model. *Iran J Public Health, 41*(10), 87-96.

Ricks, P. M., Cain, K. P., Oeltmann, J. E., Kammerer, J. S., & Moonan, P. K. (2011). Estimating the burden of tuberculosis among foreign-born persons acquired prior to entering the U.S., 2005–2009. *PLoS One, 6*, e27405.

Rios, M., Garcia, J. M., Sanchez, J. A., & Perez, D. (2000). A statistical analysis of the seasonality in pul-monary tuberculosis. *Eur J Epidemiol, 16*, 483 - 488.

Saadettin, A. (2022). Time Series Analysis and Some Applications in Medical Research. *Journal of Mathematics and Statistics Studies, 3*, 31-36. doi:10.32996/jmss.2022.3.2.3

Schwarz, G. (1978). Estimatig the dimension of a model. *Annuals of statistics, 6*(2), 461-464.

Sharma, G., Bloss, E., Heiling, C. M., & Click, E. S. (2016). Tuberculosis Caused by Mycobacterium, 2004 - 2013. *Emerging Infectious Diseases, 22*(3), 396 - 403.

Shibata, R. (1989). *Statistical Aspects of Model Selection. Working Paper [C]*. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Smith, I. (2003). Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev, 16,* 463 – 496.

Soetens, L. C., Boshuizen, H. C., & Altes, H. K. (2013). Contribution of Seasonality in Transmission of Mycobacterium tuberculosis to Seasonality in Tuberculosis Disease: A Simulation Study. *American Journal of Epidemiology, 178*(8), 1281– 1288.

Takarinda, K. C., Harries, A. D., & Mutasa-Appolo, T. (2020). Trend analysis of tuberculosis case notification with scale-up of antiretriviral therapy and roll-out of isoniazid preventive therapy in Zimbabwe, 2000 - 2018. *BMJ Open 2020, e034721*. doi:doi:10.1136/bmjopen-2019-034721

Tesfaye, B., Animut, A., Alemu, G., Abriham, Z., Cheru, T., & Bekalu, K. (2018). The twin epidemics: prevalence of TB/HIV coinfection and its associated factors in Ethiopia: a systematic review and meta-analysis. *PLoS One, 10*(13), 1 – 18.

Uplekar, M., Atre, S., Wells, A. W., Weil, D., Lopez, R., Migliori, G. G., & Raviglione, M. (2016). Mandatory tuberculosis case notification in high tuberculosis-incidence countries:policy and practice. *Eur Respir J,* 1571–1581. doi:DOI: 10.1183/13993003.00956-2016

WHO. (2013). *Islamic Republic of Afghanistan, National Tuberculosis Control Program, 'National Strategic Plan for Tuberculosis Control 2014–2018'.* Kabul: National Tuberculosis Control Program.

WHO. (2014). *Tuberculosis in Women Factsheet.* Geneva: UNAIDS. Retrieved from http://www.who.int/tb/publications/tb_women_factsheet

WHO. (2018i). *Global Tuberculosis Report - 2018, WHO.* Geneva, Switzerland.

WHO. (2019). *Global Tuberculosis Report 2019.* Geneva: World Health Organisation.

Wubuli, A., Li, Y., Xeu, F., Yao, X., Upur, H., & Wushour, Q. (2017). Seasonality of active tuberculosis notification from 2005 to 2014 in Xinjian, China. *PLoS ONE, VII*(12). Retrieved from https://doi.org/10.13771/journal.pone.0180226

Wubuli, A., Xue, F., Jiang, D., Yao, X., Upur, H., & Wushouer, Q. (2015). Socio-Demographic Predictors and Distribution of Pulmonary Tuberculosis (TB) in Xinjiang,China: A Spatial Analysis. *PubMed/NCBI, 10*.

Yang, X., Duan, Q., Wang, J., Zhang, Z., & Jiang, G. (2014). Seasonal Variation of Newly Notified Pulmonary Tuberculosis Cases from 2004 to 2013 in Wuhan, China. *PloS One, 9*(10), 254-261. doi:https://doi.org/10.1371/journal.pone.0108369

Yeshi, M., Abdurahaman, S., Genet, M., & Fenta, D. G. (2018). Assessment of extra pulmonary tuberculosis using gene xpert MTB/RIF assay and fluorescent microscopy and its risk factors at Dessie Referral Hospital, Northeast Ethiopia. *Biomed Res Int*, 1 – 10.

Zhang, G., Huang, S., Duan, Q., Shu, W., Hou, Y., & Zhu, S. (2013). Application of a hybrid model for predicting the incidence of tubercu-losis in Hubei, China. *PLoS One, 6;8*(11), e80969.

Zhang, Y. Q., Li, X., Li, W., Jiang, J., Zhang, Y., Xu, J., . . . Sun, J. S. (2020). Analysis and prediction of tuberculosis registration rates in Henan Province, China: an exponential smoothing model study. *Infectious Disease of Poverty, 9*(123). doi:https://doi.org/10.1186/s40249-020-00742-y

Zumla, A., Chakaya, J., Centis, R., Mwaba, P., & D'Ambrosio, L. (2015). Tuberculosis treatment and management-an update on treatment regimens, trial, new drugs, and adjunt therapies. *The Lacet Respiratory Medicine, 3*(3), 220-234. doi:http://dx.doi.org/10.1016/S2213-2600(15)00063-6

Appendix

TB case finding report (TB Registration sites only)														
Name of Health Facility		Zone	NORTH	Type HF	**Referral Hospital**	**Ownership**	Public/Govt							
					District Hospital		CHAM							
District		**Reporting Period**	**Jan -Mar**		Health center		Private profit							
			Apr-June		Prison		other specify							
Year			**Jul-Sept**		Rural Hospital									
			Oct -Dec											

Age Category	**New smear positive**		**New MTB detected (Expert)**		**New clinically diagnosed pulmonary TB cases**		**New EPTB**		**Relapse (bact confirmed)**		**Relapse (clinical pulmonary)**		**Relapse (EPTB)**	
	M	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
0-4														
5_14														
15-24														
25-34														
35-44														
45-54														
55-64														
65+>														
Total														

TB_HIV Reporting form

TB HIV Reporting form							
Name of Reg. Centre		**District**		Quarter	**1.Jan-Mar** **Apri-Jun** **3. Jul-Sept** **4. Oct-Dec**	Year	
Name of TB Coordinator		**Zone**	**NORTH**				

Sex	New and relapse TB cases notified	Total with HIV test result documented	Total Tested HIV Positive	Started CPT	Started ART before TB Treatment	Started ART While on Treatment
Male						
Female						
Total						

A Time Series Analysis of Patterns in TB Case Notification in North Health Zone in Malawi.

Mathias D. Ngwira and Jupiter Simbeye[1]

ABSTRACT

Introduction: Tuberculosis (TB) is a respiratory infectious disease which shows seasonality. Seasonal variations in TB notifications has been reported in different parts of the world, suggesting that various environmental and demographic factors are involved in seasonality

Objectives: this study aimed to identify a suitable time series model to investigate seasonal patterns in the notification of TB, additionally predicting/forecasting future trends of TB case notification in north health zone of Malawi

Study design: This was a hospital-based retrospective time series study conducted among patients diagnosed for TB by using data recorded from January 1, 2013 to September 31, 2020, in the north health zone of Malawi

Methods: This study used the TB data obtained from hospital records covering the period from January 2013 to September 2020. Data of 12,173 (4,800 were women) TB patients were analysed. The data were entered into Microsoft Excel 2013 and analysed using an open source statistical analysis package R-studio. The time series model was created by quarterly TB incidence data from 2013 – 2020. We investigated and found that SARIMA $(0, 1, 2)\ (1, 0, 0)_4$ is suitable for the given data. Quarterly incidences of TB and 95% confidence interval (CI) from 2021 to 2027 were predicted.

Results: We detected a pattern of peaks and troughs (cyclic) in TB case notification in our study. This cyclic pattern of TB case notification showed some peaks in the number of TB cases reported during the first quarter 3, 207 (26.02%) - rainy season; and third quarter 3, 191 (25.89%) - end of cold season. A decreasing trend in TB case notification was observed from 2013 to 2018 followed by a sharp increase in 2019 and a drop in 2020. The predicted outcome

[1] Department of Mathematical Sciences, University of Malawi, P.O. Box 280, Zomba, Malawi; duncanngwira@gmail.com (M.D.N.); jsimbeye@unima.ac.mw (J.S.)

indicate that the reported quarterly Tb incidence cases will slightly increase in the nearest future in north health zone.

Conclusion: The results showed that the number of TB case notification will follow a seasonal pattern for the next few years with an increasing trend as indicated by the forecasted number of TB case notifications. The findings suggest that progress in TB control in north health zone needs to be more intensified and adequate interventions are urgently needed.

INTRODUCTION

Tuberculosis (TB) is deadly infectious disease mainly caused by the *Mycobacterium tuberculosis* and the disease usually affects the lungs, although it can affect almost any part of the body [1]. TB is still one of the leading infections causing deaths which kills at least 2 million people every year. The disease occurs when the bacteria overcome immune defences, multiply and become large enough in number to cause tissue damage [2]. The risk of infection depends on the concentration of the expelled bacilli from the patient, the level of ventilation in households and the duration of exposure of the uninfected individual to the patient.

Globally, about one-third of the world's population has dormant M. tuberculosis and hence is at the risk of getting an illness. That hinders the socio-economic development of a country, as 75% of people with TB are within the economically productive age group of 15-54 years [3].

In 2018, 7.2 million people were estimated to develop active TB with 1.2 million TB deaths [4]. However, only 6.9 million cases were notified leaving a gap of cases that were not notified. TB cases detected in the public sector health facilities generally get notified through routine reporting systems. However, a large proportion of cases that are detected and treated in the private sector do not get notified in many settings. Thus under-notification remains a major issue especially in countries with high incidence and a large private sector.

TB case-notification policies and practices are well established in low-incidence countries [5], and mandatory notification is often recommended as a policy or practice in programme reviews in high-incidence countries. However, there is little documentation available from high-

incidence countries on either the status of these policies or the issues and challenges with implementing them. Understanding the current situation would be a first step to identifying opportunities and ways to make mandatory TB case notification operational in all settings.

While seasonal patterns in diseases such as malaria, influenza and meningitis are well acknowledged in Malawi, this remains subtle for diseases such as TB. Seasonal patterns in TB case notification have been documented in other countries in Asia, Europe and other parts of Africa, e.g., Ethiopia [6] and Nigeria [5]. The patterns of seasonal peaks and troughs in TB numbers reported in such studies appear to vary by country and hemisphere. The reasons for such variations are currently not well understood and it is likely that there are several inter-related factors. A few studies done in Ethiopia, Zimbabwe, Republic of South Africa and Morocco [7 - 10] have tried to indicate the seasonal variation in TB, but their findings are inconsistent and limited in scope. This study sought to fill this gap in knowledge using TB data reported in the health zone under study.

The purpose was to identify a suitable time-series model to investigate seasonal patterns in TB case notification in the north health zone of Malawi and use the identified model to predict future trends in the incidence of TB in the north health zone of Malawi, which would provide some reference point for TB programming. Many models such as Markov chain models, autoregressive integrated moving average class models (ARIMA), general regression models, grey models and the exponential smoothing have been proposed, which can be used to forecast infectious diseases [11]. For better forecasting performance, a comparison of two models to forecast TB case notifications was studied. The findings from this study will add to the body of knowledge about the seasonality of TB case notification and other factors affecting trends in TB incidence. Knowledge about seasonality and other factors affecting trends in TB incidence will help in predicting future TB incidence epidemics and hence help in planning for service requirements, assess health needs and manage the disease by using the predictions as a reference information. The study has also suggested the potential causes or risk factors associated with the identified pattern and hence suggested interventions that could be put in place as a means to combat the spread of the disease in the study area.

MATERIALS AND METHODS

Study Design

This was a hospital-based retrospective study conducted among patients diagnosed for TB by using data recorded from January 1, 2013 to September 31, 2020 (constituting 32 epidemiological quarters), in all the 64 health facilities in the north health zone of Malawi. The study used secondary data from the hospital records on TB case notifications. All forms of TB cases registered during the study period in all health institutions/facilities that provide DOTS services [21]. We carried out a time series analysis to test our hypothesis that TB case notifications show seasonality.

Study setting

Malawi is a low-income country located in Southern Africa and has a land area of 118, 000 square kilometres. It shares boundaries with Zambia to the west, Mozambique to the east and Tanzania to the north and north-east. The country is divided into three administrative regions namely Northern, Central and Southern regions. For operation purposes, the Ministry of Health (MoH) has created 5 health zones with Southern and Central administrative regions each divided into two zones. The five health zones of Malawi are as follows: North, Central East, Central West, South East and South West. This study used data collected from the North zone of Malawi whose headquarters is in Mzuzu City. The north health zone comprises six administrative districts namely; Chitipa, Karonga, Rumphi, Nkhata-Bay, Mzimba and Likoma.

Data Analysis and Procedure

Data entry and merging was done in Microsoft Excel 2013. Exploratory analysis and generation of descriptive statistics for summarizing information was done using Microsoft Excel - Pivot Tables. Considering the importance of the data analysis stage, this study has partitioned the stage into sub-stages as indicated below for easy interpretation of the results. Box-Jenkins time series approach put forward as Autoregressive Moving Average (ARIMA) model and the Exponential smoothing methods were used in the modelling.

The Box-Jenkins methodology comprised of model identification, Parameter Estimation, model diagnostics and forecasting [12]. Time series of the data was plotted for the period 2013 to 2020 to identify various time series components in the data. The original data was transformed by differencing and then re-plotted. Stationarity was assessed and confirmed using Augmented Dickey-Fuller (ADF) test on the transformed data. The series was judged stationary with the p-value of the ADF test $\leq$ 5% level of significance. An Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were plotted to obtain the order p and q of Autoregressive (AR) and moving average (MA) respectively. Upon obtaining the order of AR and MA terms, the model was obtained.

Development of the Model.

This study was centred on forecasting time-series analysis of TB incidence data. Prior to model fitting, a stationarity check was done to determine that the time-series is constant in its mean and variance and that the mean and variance are not dependent on time. Secondly, a time-series plot was sketched to evaluate the behavioural pattern in the data over a period of seven years (**figure 1).** A multiplicative decomposition of the TB time-series was done to describe the seasonality components and trends. From the graph, we observed that the TB occurrence data had a cyclic pattern of movement. Firstly, we looked at ARIMA model to assess the TB data. The Seasonal ARIMA and the Exponential smoothing models were used in analysing the trend of the time series data independently of the seasonal components and predicting the quarterly TB incidence in north health zone in Malawi.

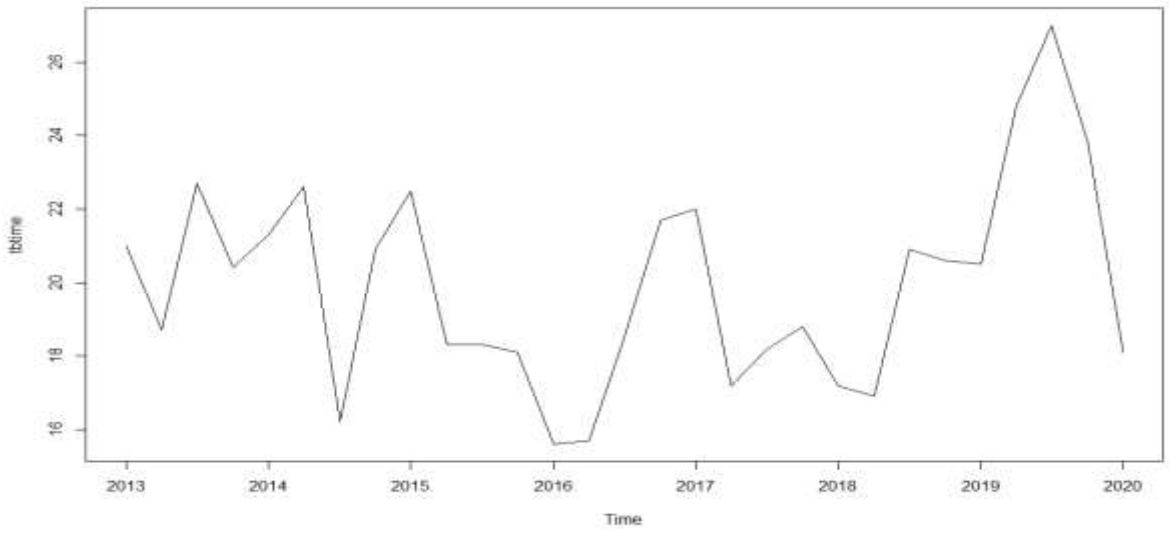

Figure 1. Quarterly TB case notification rates from January 2013 to September 2020

Development of the SARIMA Model

The seasonal ARIMA model (SARIMA) is an expanded form of ARIMA, which allows for seasonal factors to be reflected [13]. Time series seasonality is an unvarying pattern that recurs over S period of time until the pattern changes over again. The SARIMA model integrated both non-seasonality and seasonality factors in a generative model. In the SARIMA model, seasonality in AR and MA terms predict Y_t (TB case notifications) using data values and errors at time interval that are multiples of S. The SARIMA model is given as:

$$SARIMA(p,d,q) \times (P,D,Q)^S$$

Where $p = AR$ order in non- seasonality, d = difference in non-seasonality, $q = MA$ order in non- seasonality,$P = AR$ order in seasonality, D = difference in seasonality, Q= MA order in seasonality, and S = recurrence of time periods in the seasonality pattern. The general SARIMA model has the following form

$$\Phi(\beta^S)\varphi(\beta)(Y_t - \mu) = \Theta(\beta^S)\theta(\beta)\varepsilon_t$$

The non-seasonality components are;

$$AR: \varphi(\beta) = 1 - \varphi_1(\beta) - \cdots - \varphi_p\beta^p$$

$$MA: \theta(\beta) = 1 + \theta_1(\beta) + \cdots + \theta_q\beta^q$$

The seasonality components are;

$$AR: \Phi(\beta^S) = 1 - \Phi_1 - \Phi_1\beta^S - \cdots - \Phi_P\beta^{PS}$$

$$MA: \Theta(\beta^S) = 1 + \Theta_1\beta^S + \cdots + \Theta_Q\beta^{QS}$$

In the equations, β represents the backward shift operator, ε_t stands for estimated residual error at t for $\mu = 0$ and variance is constant and Y_t represents the TB notifications data at $t(t = 1,2,3, \ldots, k)$ $\boldsymbol{\phi}$ is a vector of the AR coefficients, $\theta$is a vector of the MA coefficients, Φ is a vector of the seasonal AR coefficients, and Θ is a vector of the seasonal MA coefficients.

In the SARIMA model, seasonal subtraction of appropriate order is used to remove non-stationary data from the series. A first order seasonal difference is the deviation between a value and the corresponding value from the previous year and it is expressed as:$Y_t = X_t - X_{t-s}$ for quarterly time series (S) = 4.

Development of an Exponential Smoothing Model

Exponential smoothing was first suggested in the statistical literature without reference to previous work by [14] in 1956 and then expanded by [15] in 1957. Exponential smoothing is a technique used to detect significant changes in data by considering the most recent data. Also known as averaging, this method is used in making short-term forecasts. The simplest form of an exponential smoothing formula is given by:

$$F_t = \alpha A_{t-1} + (1 - \alpha) F_{t-1}$$

Here; F_t = smoothed statistic; A_{t-1} = previous smoothed statistic; α = smoothing factor of data; $0 < \alpha < 1$ and t = time period

If the value of the smoothing factor is larger, then the level of smoothing will reduce. Value of α close to 1 has less of a smoothing effect and give greater weight to recent changes in the data, while the value of α closer to zero has a greater smoothing effect and are less responsive to recent changes.

Exponential smoothing is best used for forecasts that are short-term and in the absence of seasonal or cyclical variations. As a result, forecasts aren't accurate when data with cyclical or seasonal variations are present. As such, this kind of averaging won't work well if there is a trend in the series.

RESULTS OF THE STUDY

TB incidence data from January 2013 to September 2020 was used to perform the time-series model fit. ACF and PACF plots were used to determine the key parameters (p, P, d, D, q, Q) of seasonal ARIMA model. In this study, we used the Box-Jenkin SARIMA and exponential smoothing approaches to identify the best model to forecast future patterns in the TB case notification rate in the north health zone of Malawi. We used ***auto.arima*** function in R to identify the best model to predict future trends of TB case notifications. In this study, we used two principal model selection methods of Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC). The model with the smallest values of AIC and BIC were regarded as the best model to predict future incidences of TB in the north health zone in Malawi.

TB Case Notification Rate as Stratified by Age Groups, Sex, HIV Status, Year and District

The results of the study revealed that the number of TB case notification rate was almost constant from 2013 to 2015 followed by a slight decrease in 2016. We observed a slightly sharp

increase of TB case notification rates between 2017 and 2019 followed by a sharp decrease in 2020 which could be explained due to Covid-19 pandemic whereby there was apathy from the general population in seeking health care. The highest case notification rate was observed in 2019 and could be due to mass campaigns about care seeking behaviour which improved and influenced people's understanding about timely health seeking behaviours.

The results further revealed that males had higher notification rates than females and the age category 0 – 24 years has the lowest notification rates than others. In terms of HIV status, there was no variation of notification rates for both positive and negative TB patients across the study period. Furthermore, the results showed that Mzimba district had the highest notification rates of the rest of the district. This could be explained due to cold weather in the district. A decrease in vitamin D and sunlight significantly increases the incidence of smear and sputum positive tuberculosis [16].

Stationarity of the Transformed Data

Our original data was transformed by taking the difference of the data in order to make it stationary and allow us to apply ARIMA and Exponential smoothing on the data. It was found that the transformed series achieved stationarity. The Augmented Dickey-Fuller test statistically confirmed the stationarity of the series (ADF = - 5.7044, p-value = 0.01). From the plots of ACF and PACF (figure 4)

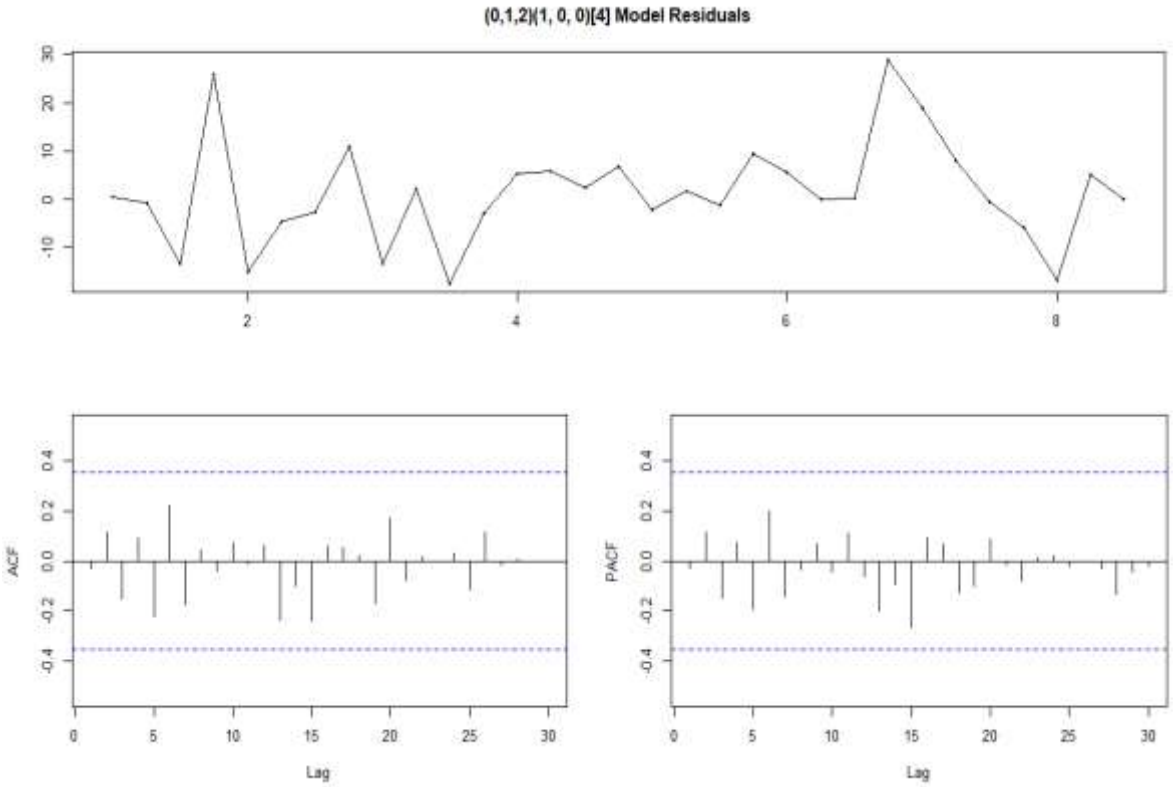

Figure 4. Time plot, ACF and PACF plot for the ARIMA (0, 1, 2) (1, 0, 0)4 model residual

Analysis of the Competing Models

In order to select the best model to be used in the predictions, three competing ARIMA models were further tested and compared with the Exponential smoothing model in order to select the model with the best predictive ability. Table 1 below shows the estimates of parameters from the three competing ARIMA models.

Table 1: Estimates of parameters from the three competing ARIMA models.

Model	AIC	AICc	BIC
ARIMA (1, 1, 0)	255.58	265.03	258.39
ARIMA (1, 1, 3)	254.36	254.36	258.86
ARIMA (1, 1, 1)	252.50	252.10	256.11

Exponential smoothing was tested and compared to the other three competing ARIMA models. Figure 3 below shows a graphical presentation of the Exponential smoothing model. From the graph below, it is shown that the exponential smoothing method produced forecasts that lagged behind the actual trend.

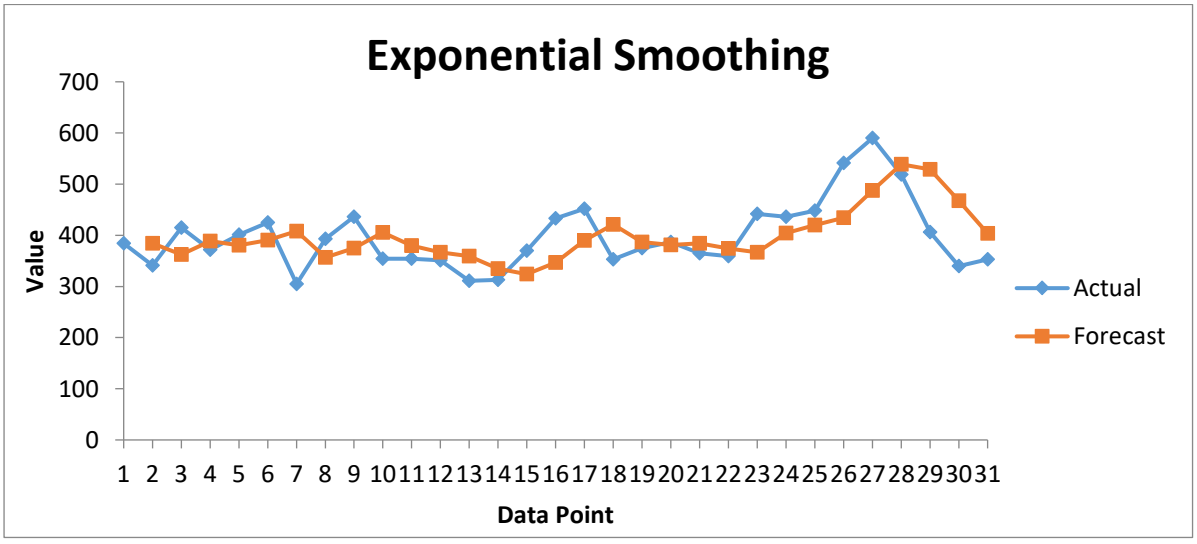

Figure 3. Forecast from the Exponential smoothing method

Figure 4 below shows the graphical presentation of the three competing ARIMA model and the seasonal ARIMA model. From figure 4 below, it was noted that ARIMA (0, 1, 2) $(1, 0, 0)_4$ has the best ability to predict future trends in the TB case notifications. The other models fail

to qualify because they just show a straight line into the future which could not reflect on reality of the future patterns in the TB case notifications.

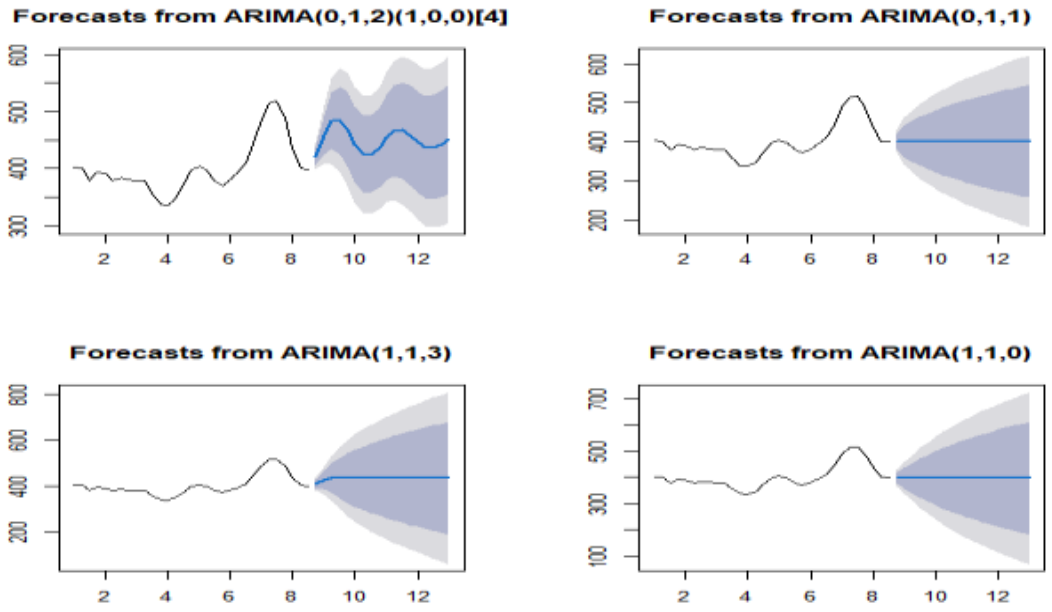

Figure 4. Forecasts from the four competing ARIMA models

Predicted Trends of TB case Notifications

We applied the Winter's Multiplicative method to the seasonal ARIMA (0, 1, 2) $(1, 0, 0)_4$ model to predict future trends in TB case notifications in the north health zone of Malawi at 95% confidence interval (CI). The predictions were done for the next 12 quarters (fourth quarter of 2020 to third quarter of 2027). Figure 5 below shows the graphical presentation of the forecasted TB case notifications for the north health zone in Malawi. The predictions were done on the assumptions that the shall be no any extra interventions done by the government and other stakeholders that would otherwise have an influence on the number of TB case notified in the health zone. We assumed that the current prevailing interventions will continue for the next few years.

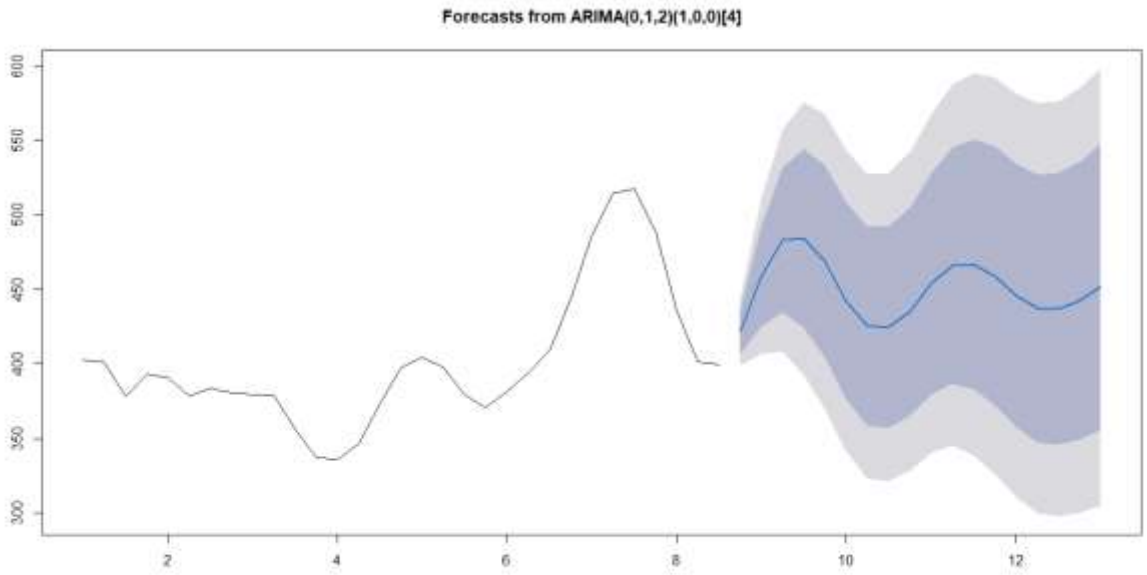

Figure 5. The graphical presentation of the predicted/forecasted number of TB case notification for the north health zone of Malawi from October 2020 to December 2027

Table two below presents the comparison of the predicted incidences of TB in the north health zone of Malawi from 2020 to 2027.

Table 2: Comparison of Competing Models for Predicting Future Seasonal Patterns in TB Case Notification in the North Health Zone of Malawi from October 2020 to September 2023

Point of forecast	**ARIMA (0,1,2) $(1,0,0)_4$**	**ARIMA (1, 1, 0)**	**ARIMA (1, 1, 3)**	**ARIMA (1, 1, 1)**
2020 Q4	421.24	398.52	406.69	400.92
2021 Q1	458.89	397.83	423.92	400.92

2021 Q2	483.00	397.40	433.52	400.92
2021 Q3	484.25	397.14	433.03	400.92
2022 Q4	469.00	396.98	433.06	400.92
2023 Q1	442.44	396.77	433.06	400.92
2023 Q2	425.43	396.75	433.06	400.92
2023 Q3	424.54	396.73	433.06	400.92
2023 Q4	435.31	396.72	433.06	400.92
2024 Q1	454.04	396.72	433.06	400.92
2024 Q2	466.04	396.72	433.06	400.92
2024 Q3	466.66	396.72	433.06	400.92
2024 Q4	459.08	396.72	433.06	400.92
2025 Q1	445.86	396.72	433.06	400.92
2025 Q2	437.39	396.72	433.06	400.92
2025 Q3	436.95	396.72	433.06	400.92
2025 Q4	442.31	396.72	433.06	400.92
2026 Q1	451.63	396.72	433.06	400.92
2026 Q2	457.61	396.72	433.06	400.92

2026 Q3	457.92	396.72	433.06	400.92
2026 Q4	454.14	396.72	433.06	400.92
2027 Q1	447.56	396.72	433.06	400.92
2027 Q2	443.34	396.72	433.06	400.92
2027 Q3	443.13	396.72	433.06	400.92
2027 Q4	445.79	396.72	433.06	400.92

DISCUSSION

Tuberculosis (TB) is a disease that continues to be a major public health problem amongst the top ten disease causes of mortality. Millions of people continue to fall sick with TB each year, particularly in developing countries [17, 18]. In this study, a seasonal ARIMA model was developed to forecast future quarterly incidence of TB cases in north health zone of Malawi. From our results we can clearly see that the SARIMA $(0, 1, 2)\ (1, 0, 0)_4$ provided a better forecast of the TB the data. The model was appraised by AIC, AICc, BIC and the white noise residuals. [19] investigated the performance of six different forecasting methods, including linear regression, moving average, decomposition, ARIMA, Neural Network and Holt-Winter's for monthly tuberculosis data prediction and the results from the study showed that the most appropriate model was ARIMA [20]. [21] investigated two models of ARIMA and (GRNN)-ARIMA in prediction of tuberculosis incidence. The time series of tuberculosis shows a gradual secular decline and a striking seasonal variation.

The predicted outcome indicate that the reported quarterly Tb incidence cases will slightly increase in the nearest future in north health zone. The findings suggest that progress in TB

control in north health zone needs to be more intensified and adequate interventions are urgently needed.

The first quarter (rainy season) was the dominant quarter seconded by third quarter. Fourth quarter has the least number of TB cases during the study period. We can postulate that high cases of TB reported during the third quarter, which lies between the cold season and the hot season. It is possible that this was so because of a number of reasons; Firstly, poor ventilated rooms crowded with people could increase the chances of transmission among the infections source and the contractors in cold seasons. In the north health zone of Malawi, cold season is from May to August, and the coldest months are June and July in most of the year. During this cold period, people are more likely to stay indoors and close the windows because of low temperatures. We therefore suspect that TB transmission is rampant during this period due to lack of fresh air.

Secondly, delays to seek medical health care in the coldest months (June and July) could possibly increase the risk of transmission. [21] postulated that people could not feel comfortable to seek medical help when it is very cold unless otherwise. This being the case, people tend to seek medical help when the weather is somewhat warmer (in our scenario this could be between August and September). Thus, the consultation rate between patients and health care givers during this warmer period is relatively high. According to [22] delay in health care seeking contributes to diagnosis delay which may increase the risk of disease transmission (due to the longer communicable period of TB) in cold seasons. It is further believed that in winter, symptoms of TB may possibly not be seen immediately after becoming infected, but when summer approaches and the temperature increases, the bacterium starts proliferating and growing with positive symptoms of the infection [21].

CONCLUSION

TB continues to be a serious threat to public health in north health zone in Malawi. The study identified SARIMA (0, 1, 2) $(1, 0, 0)_4$ as the best model to forecast future trends of TB case notifications in the north health zone of Malawi. The model indicates that the TB prevalence in north health zone will not increase remarkably in the forthcoming years; it is essential to

effect better TB incidence control measures in the health zone under study. The observed pattern of TB notifications showed that there was a cyclic pattern in the TB case notification in the zone with peaks during the rainy season and at the end of the cold season. The predicted case notification indicate that the number of TB case notification will follow a seasonal pattern for the next three years with an increasing trend as indicated by the forecasted number of TB case notifications. The TB prevalence seasonality from the model also indicate a great necessity for TB interventions, focused on reducing infectious disease transmission with co-infection with HIV and other concomitant diseases and also public events and over clouded places.

References

1. Sharma, G., Bloss, E., Heiling, C. M., & Click, E. S. (2016). Tuberculosis Caused by Mycobacterium, 2004 - 2013. *Emerging Infectious Diseases, 22*(3), 396 - 403.

2. Ricks, P. M., Cain, K. P., Oeltmann, J. E., Kammerer, J. S., & Moonan, P. K. (2011). Estimating the burden of tuberculosis among foreign-born persons acquired prior to entering the U.S., 2005–2009. *PLoS One, 6,* e27405.

3. Moosazadeh, M., Khanjani, N., Nasehi, M., & Bahrampour, A. (2015). A. Predicting the incidence of smear positive tuberculosis cases in Iran using time series analysis. *Public Health, 44*, 1526–1534.

4. WHO. (2018i). *Global Tuberculosis Report - 2018, WHO.* Geneva, Switzerland.

5. Uplekar, M., Atre, S., Wells, A. W., Weil, D., Lopez, R., Migliori, G. G., & Raviglione, M. (2016). Mandatory tuberculosis case notification in high tuberculosis-incidence countries:policy and practice. *Eur Respir J,* 1571–1581. doi:DOI: 10.1183/13993003.00956-2016.

6. Nagua, T. J., Aboud, S., & Mwiru, R. (2017). Tuberculosis associated mortality in a prospective cohort in Sub Saharan Africa: association with HIV and antiretroviral therapy. *Int J Infect Dis, 2017;2017*(56), 39 - 44. doi:doi:10.1016/j.ijid

7. Gashu, Z. J. (2018). Seasonal patterns of tuberculosis case notification in the tropics of Africa: A six-year trend analysis in Ethiopia. *11* (13) . doi:https://doi.org/10.1371/journal.pone.0207552

8. Takarinda, K. C., Harries, A.D. & Mutasa-Apollo, T. (2020). Trend analysis of tuberculosis case notifications with scale-up of antiretroviral therapy and roll-out of isoniazid preventive therapy in Zimbabwe, 2000 – 2018. BMJ Open 2020; 10: e034721. doi: 10.1136/bmjopen-2019-034721

9. Azeez, A., Obaromi, D., Odeyemi, A., Ndege, J., & Muntabayi, R. (2016). Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape South Africa, using a hybrid model. *Int J Environ Res Public Health, 13*.

10. Ottmani, S. Obermeyer, Z. Bencheikh, N & Mahjour, J. (2021). Knowledge, attitudes and beliefs about tuberculosis in urban Morocco. East Mediterr Health J. 2008; 14 (2): 298 – 304.

11. Halim, S., Intan, R., & Dewi, L. P. (2019). Fuzzy linear reression for tuberculosis case notification rate prediction in Surabaya. Proceedings of the International Conference on Advanced Informattion Science and System. 1 - 5. doi:https://doi.org/10.1145/3373477.3373492.

12. He, Z., & Tao, H. (2018). Epidemiology and ARIMA model of positive-rate ofinfluenza viruses among children in Wuhan, China: a nine-year retrospectivestudy. *Int J Infect Dis, 74*, 61 - 70.

13. Brass, A.L., Gomes, D., Filipe, P.A., de Sousa, B. & Nunes, C. (2014). Trend, seasonality and forecasts of pulmonary tuberculosis in Portugal. Int J Tuberc Lung Dis. 2014 Oct;18 (10):1202-10. doi: 10.5588/ijtld.14.0158. PMID: 25216834.

14. Brown R.G. (1956). *Smoothing, forecasting and prediction of discrete time series.* Englewoods Cliffs, N.J. Prentince-Hall.

15. Charles C. Holt (2004), *Forecasting Trends and Seasonals by Exponentially Weighted Averages, International Journal of Forecasting*, 20 (1), 5 – 10.

16. Abate D, Bineyam T, Mohammed A, Sibhatu B. Epidemiology of anti-tuberculosis drug resistance patterns and trends in tuberculosis referral hospital in Addis Ababa, Ethiopia. BMC Infect Dis. 2012; 5 (462):1–6.

17. Permanasari, A. E., Rambli, D. R., & Dominic, P. D. (2011). Performance of Univariate Forecasting on Seasonal Diseases: The Case of Tuberculosis. *Adv Exp Med Biol, Adv Exp Med Biol*, 1791- 179.

18. Yang, X., Duan, Q., Wang, J., Zhang, Z., & Jiang, G. (2014). Seasonal Variation of Newly Notified Pulmonary Tuberculosis Cases from 2004 to 2013 in Wuhan, China. *PloS One, 9*(10), 254-261. doi:https://doi.org/10.1371/journal.pone.0108369

19. Smith, I. (2003). Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev, 16*, 463 – 496.

20. Global tuberculosis report 2018. World Health Organization.

21. Frah, E. A., & Alkhalifa, A. (2016). Tuberculosis Cases in Sudan; Forecasting Incidents 2014 - 2023 Using Box-Jenkins ARIMA Model. *American Journal of Mathematics and Statistics, 6*(3),108-114.

22. Nyirenda, T. (2006). Epidemiology of Tuberculosis in Malawi. *Malawi Med J, 18*(3), 147–159.